**UNIT-I**

**Business intelligence:** Effective and timely decisions, Data, information and knowledge, the role of mathematical models, Business intelligence architectures, Ethics and business intelligence

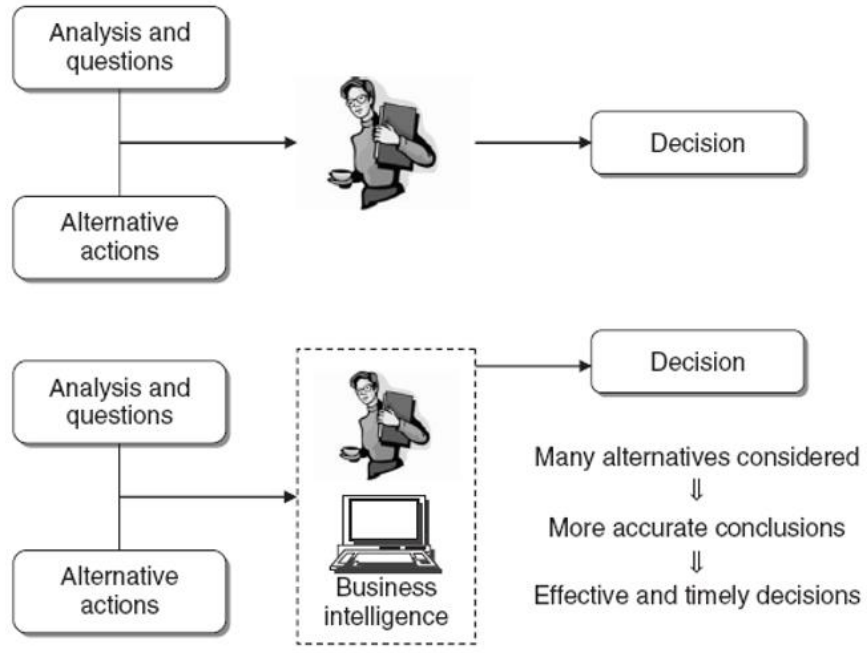
**Decision support systems:** Definition of system, Representation of the decision-making process, Evolution of information systems, Definition of decision support system, Development of a decision support system

## Effective and timely decisions:

The main purpose of business intelligence systems is to provide knowledge workers with tools and methodologies that allow them to make *effective* and *timely* decisions.

**Effective decisions.** The application of rigorous analytical methods allows decision makers to rely on information and knowledge which are more dependable

**Timely decisions.** The ability to rapidly react to the actions of competitors and to new market conditions is a critical factor in the success or even the survival of a company.




*Benefits of a Business Intelligence System*

## Data, information and knowledge

**Data:**
For a retailer data refer to primary entities such as customers, points of sale and items, while sales receipts represent the commercial transactions.
Data is unprocessed facts and figures without any added interpretation or analysis. "The price of crude oil is $80 per barrel."

**Information:**
Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain
Information is data that has been interpreted so that it has meaning for the user. "The price of crude oil has risen from $70 to $80 per barrel" gives meaning to the data and so is said to be information to someone who tracks oil prices.

**Knowledge:**
Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions.
Knowledge is a combination of information, experience and insight that may benefit the individual or the organization. "When crude oil prices go up by $10 per barrel, it's likely that petrol prices will rise by 2p per litre" is knowledge.

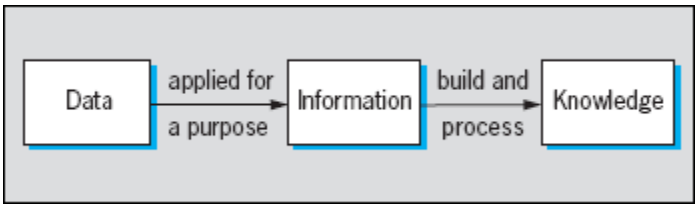



**Figure: From data to information to knowledge**

Several public and private enterprises and organizations have developed in recent years formal and systematic mechanisms to gather, store and share their wealth of knowledge, which is now perceived as an invaluable intangible asset. The activity of providing support to knowledge workers through the integration of decision-making processes and enabling information technologies is usually referred to as ***knowledge management***.

## The role of mathematical models:

A business intelligence system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms. In some instances, this activity may reduce to calculations of totals and percentages, graphicall`y represented by simple histograms, whereas more elaborate analyses require the development of advanced optimization and learning models.

A business intelligence system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms.

- **First, the objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined.**
- **Mathematical models are then developed by exploiting the relationships among system control variables, parameters and evaluation metrics.**

- **Finally, *what-if analyses are carried out to evaluate the effects on* the performance determined by variations in the control variables and changes in the parameters.**

## Business intelligence architectures:

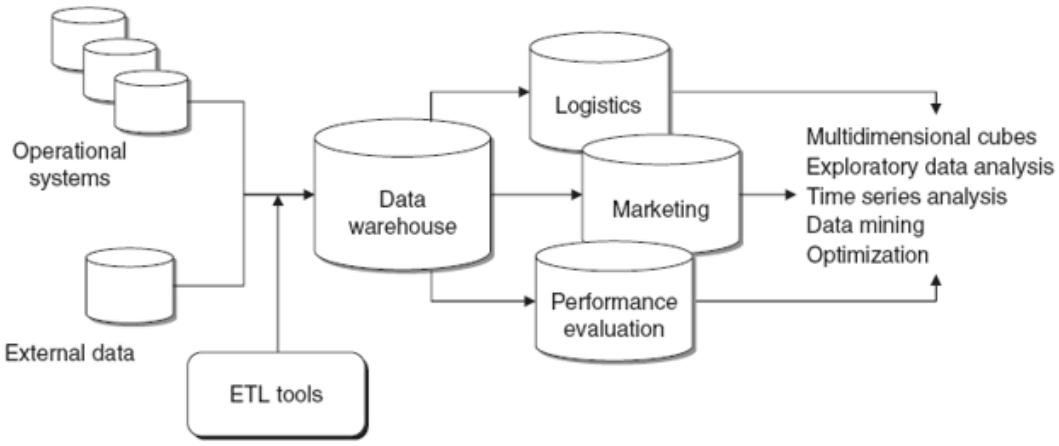
The architecture of a business intelligence system, includes three major components.

**Data sources.** In a first stage, it is necessary to gather and integrate the data stored in the various primary and secondary sources, which are heterogeneous in origin and type. The sources consist for the most part of data belonging to operational systems, but may also include unstructured documents, such as emails and data received from external providers. Generally speaking, a major effort is required to unify and integrate the different data sources.

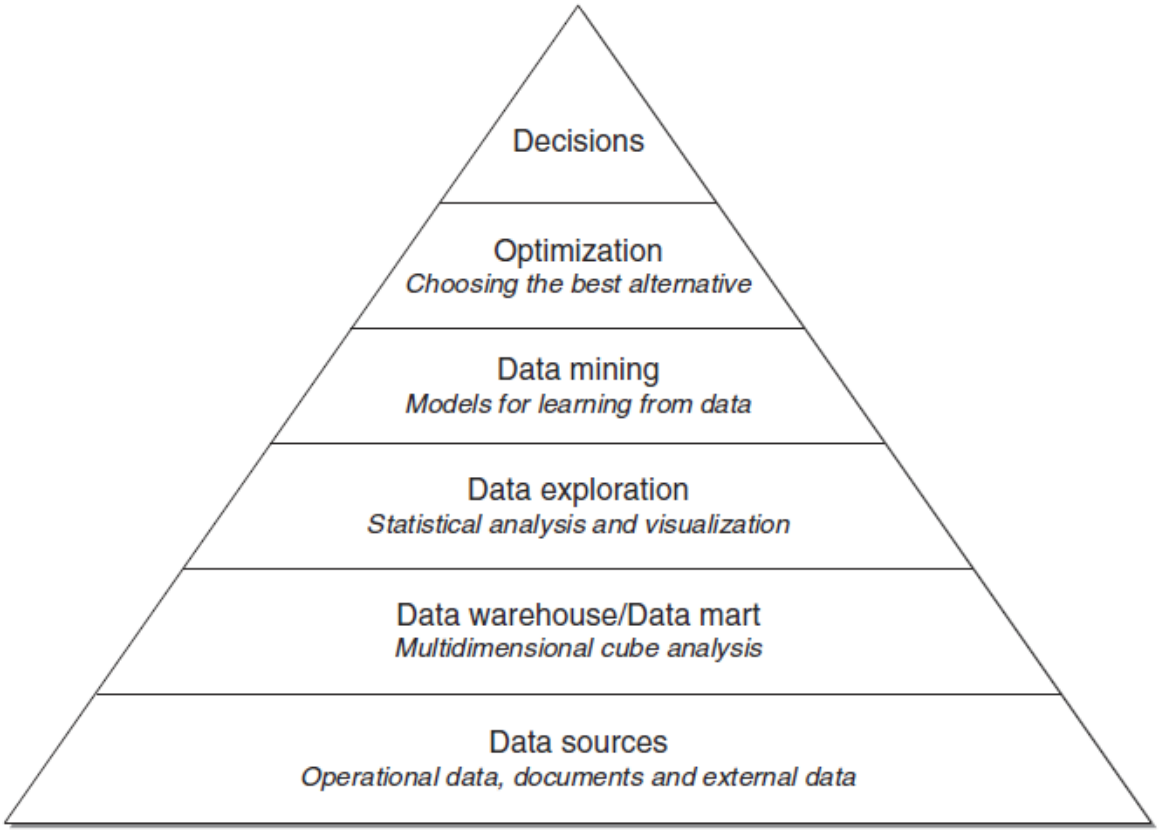
**Data warehouses and data marts.** Using extraction and transformation tools known as *extract, transform, load* (ETL), the data originating from the different sources are stored in databases intended to support business intelligence analyses. These databases are usually referred to as *data warehouses* and *data marts*.

**Business intelligence methodologies.** Data are finally extracted and used to feed mathematical models and analysis methodologies intended to support decision makers. In a business intelligence system, several decision support applications may be implemented.

- Multidimensional cube analysis;
- Exploratory data analysis;




Business Intelligence Architectures




*The main components of a business intelligence system*

1. **Data sources.**
   The sources consist for the most part of data belonging to operational systems, may also include unstructured documents, such as emails and data received from external providers
2. **Data warehouses and data marts**
   Using extraction and transformation tools known as *extract, transform, load (ETL), the data* originating *from the different* sources are stored in databases intended to support business intelligence analyses.

   **Business intelligence methodologies.**

   Data are finally extracted and used to feed mathematical models and analysis methodologies intended to support decision makers. In a business intelligence system, several decision support applications may be implemented:

   - multidimensional cube analysis
   - exploratory data analysis
   - time series analysis
   - inductive learning models for data mining

- optimization models

3. **Data exploration:**

   ***Passive Business Intelligence Analysis*** *consists of*

   1. **Query and Reporting Systems**
   2. **Statistical Methods.**

   These are referred to as passive methodologies because decision makers are requested to generate prior hypotheses or define data extraction criteria, and then use the analysis tools to find answers and confirm their original insight.

   **For instance, consider the sales manager of a company who notices that revenues in a given geographic area have dropped for a specific group of customers. Hence, she might want to bear out her hypothesis by using extraction and visualization tools, and then apply a statistical test to verify that her conclusions are adequately supported by data.**

4. **Data Mining:**

   The fourth level includes *active* business intelligence methodologies, whose purpose is the extraction of information and knowledge from data.

   These include mathematical models for pattern recognition, machine learning and data mining techniques. Unlike the tools described at the previous level of the pyramid, the models of an active kind do not require decision makers to formulate any prior hypothesis to be later verified. Their purpose is instead to expand the decision makers' knowledge.

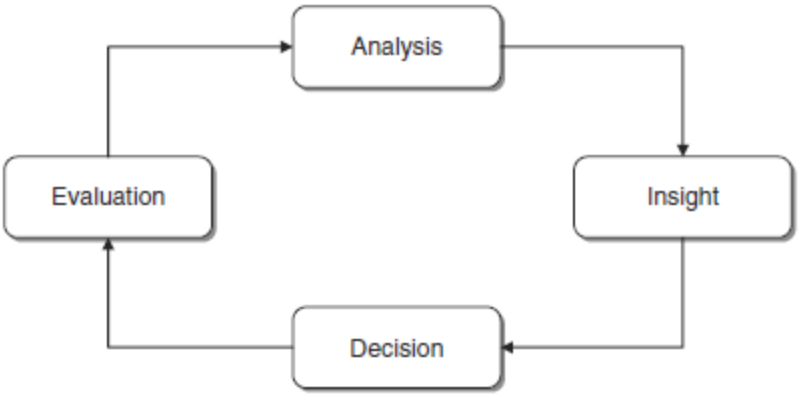
5. **Optimization.**

   By moving up one level in the pyramid we find optimization models that allow us to determine the best solution out of a set of alternative actions, which is usually fairly extensive and sometimes even infinite.

6. **Decision**

   Finally, the top of the pyramid corresponds to the choice and the actual adoption of a specific decision, and in some way represents the natural conclusion of the decision-making process. Even when business intelligence methodologies are available and successfully adopted, the choice of a decision pertains to the decision makers, who may also take advantage of informal and unstructured information available to adapt and modify the recommendations and the conclusions achieved through the use of mathematical models.

**Cycle of a business intelligence analysis:**




*Cycle of a business intelligence analysis*

**Analysis.**

During the analysis phase, it is necessary to recognize and accurately spell out the problem at hand. Decision makers must then create a mental representation of the phenomenon being analyzed, by identifying the critical factors that are perceived as the most relevant. The availability of business intelligence methodologies may help already in this stage, by permitting decision makers to rapidly develop various paths of investigation. For instance, the exploration of data cubes in a multidimensional analysis, according to different logical views, allows decision makers to modify their hypotheses flexibly and rapidly, until they reach an interpretation scheme that they deem satisfactory. Thus, the first phase in the business intelligence cycle leads decision makers to ask several questions and to obtain quick responses in an interactive way.

**Insight.**

The second phase allows decision makers to better and more deeply understand the problem at hand, often at a causal level. For instance, if the analysis carried out in the first phase shows that a large number of customers are discontinuing an insurance policy upon yearly expiration, in the second phase it will be necessary to identify the profile and characteristics shared by such customers. **The information obtained through the analysis phase is then transformed into knowledge during the insight phase.** On the one hand, the extraction of knowledge may occur due to the intuition of the decision makers and therefore be based on their experience and possibly on unstructured information available to them. On the other hand, inductive learning models may also prove very useful during this stage of analysis, particularly when applied to structured data.

**Decision.**

During the third phase, knowledge obtained as a result of the insight phase is converted into decisions and subsequently into actions. The availability of business intelligence methodologies allows the analysis and insight phases to be executed more rapidly so that more effective and timely decisions can be made that better suit the strategic priorities of a given organization. This leads to an overall reduction in the execution time of the *analysis–decision–action– revision* cycle, and thus to a decision-making process of better quality.

**Evaluation.**

Finally, the fourth phase of the business intelligence cycle involves performance measurement and evaluation. Extensive metrics should then be devised that are not exclusively limited to the financial aspects but also take into account the major performance indicators defined for the different company departments.

## Enabling factors in business intelligence projects`

Some factors are more critical than others to the success of a business intelligence project: *technologies*, *analytics* and *human resources*.

**Technologies.**

Hardware and software technologies are significant enabling factors that have facilitated the development of business intelligence systems within enterprises and complex organizations. On the one hand, the computing capabilities of microprocessors have increased on average by 100% every 18 months during the last two decades, and prices have fallen. This trend has enabled the use of advanced algorithms which are required to employ inductive learning methods and optimization models, keeping the processing times within a reasonable range. Moreover, it permits the adoption of state-of-the-art graphical visualization techniques, featuring real-time animations. A further relevant enabling factor derives from the exponential increase in the capacity of mass storage devices, again at decreasing costs, enabling any organization to store terabytes of data for business intelligence systems. And network connectivity, in the form of *Extranets* or *Intranets*, has played a primary role in the diffusion within organizations of information and knowledge extracted from business intelligence systems. Finally, the easy integration of hardware and software purchased by different suppliers, or developed internally by an organization, is a further relevant factor affecting the diffusion of data analysis tools.

**Analytics.**

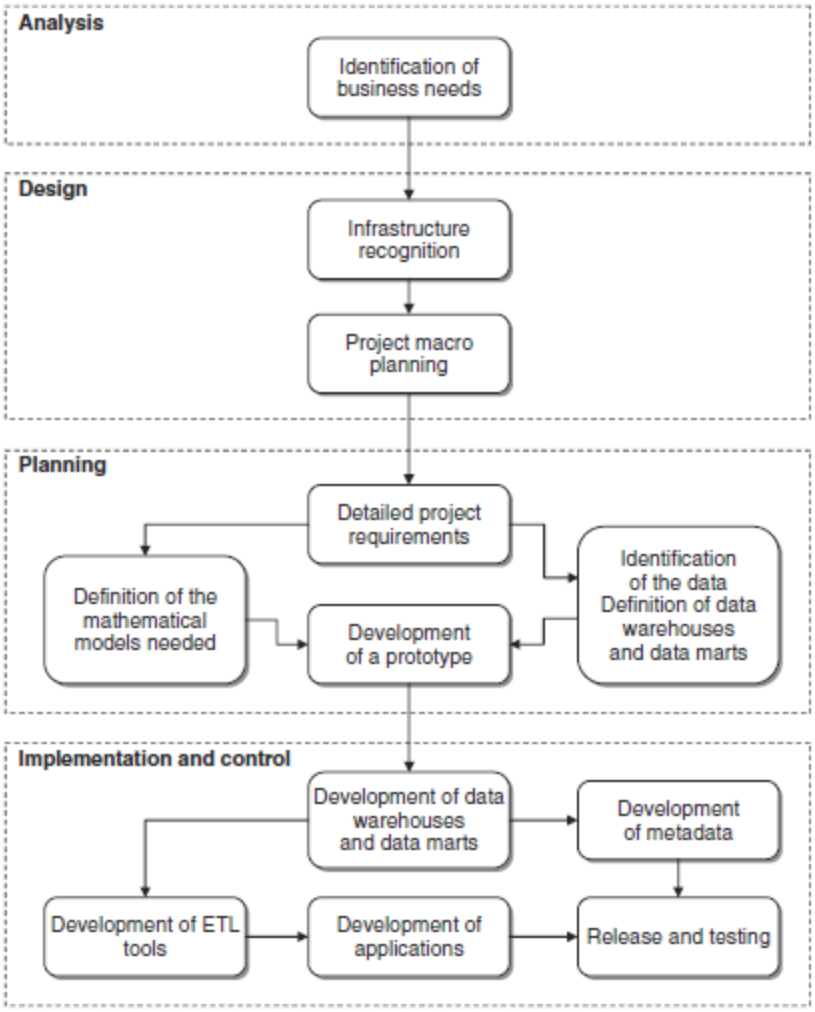
As stated above, mathematical models and analytical methodologies play a key role in information enhancement and knowledge extraction from the data available inside most organizations. The mere visualization of the data according to timely and flexible logical views, as described in Chapter 3, plays a relevant role in facilitating the decision-making process, but still represents a passive form of support. Therefore, it is necessary to apply more advanced models of inductive learning and optimization in order to achieve active forms of support for the decision-making process.

**Human resources.**

The human assets of an organization are built up by the competencies of those who operate within its boundaries, whether as individuals or collectively. The overall knowledge possessed and shared by these individuals constitutes the *organizational culture*. The ability of knowledge workers to acquire information and then translate it into practical actions is one of the major assets of any organization, and has a major impact on the quality of the decision-making process. If a given enterprise has implemented an advanced business intelligence system, there still remains much scope to emphasize the personal skills of its knowledge workers, who are required to perform the analyses and to interpret the results, to work out creative solutions and to devise effective action plans. All the available analytical tools being equal, a company employing human resources endowed with a greater mental agility and willing to accept changes in the decision-making style will be at an advantage over its competitors.

## Development of a business intelligence system:

The development of a business intelligence system can be assimilated to a project, with a specific final objective, expected development times and costs, and the usage and coordination of the resources needed to perform planned




*Phases in the development of a business intelligence system*

**Analysis:**

During the first phase, the needs of the organization relative to the development of a business intelligence system should be carefully identified.

This preliminary phase is generally conducted through a series of interviews of knowledge workers performing different roles and activities within the organization. It is necessary to clearly describe the general objectives and priorities of the project, as well as to set out the costs and benefits deriving from the development of the business intelligence system.

**Design:**

The second phase includes two sub-phases and is aimed at deriving a provisional plan of the overall architecture, taking into account any development in the near future and the evolution of the system in the mid-term. First, it is necessary to make an assessment of the existing information infrastructures. Moreover, the main decision-making processes that are to be supported by the business intelligence system should be examined, in order to adequately determine the information requirements. Later on, using classical project management methodologies, the project plan will be laid down, identifying development phases, priorities, expected execution times and costs, together with the required roles and resources.

**Planning:**

The planning stage includes a sub-phase where the functions of the business intelligence system are defined and described in greater detail. Subsequently, existing data as well as other data that might be retrieved externally are assessed. This allows the information structures of the business intelligence architecture, which consist of a central data warehouse and possibly some satellite data marts, to be designed. Simultaneously with the recognition of the available data, the mathematical models to be adopted should be defined, ensuring the availability of the data required to feed each model and verifying that the efficiency of the algorithms to be utilized will be adequate for the magnitude of the resulting problems. Finally, it is appropriate to create a system prototype, at low cost and with limited capabilities, in order to uncover beforehand any discrepancy between actual needs and project specifications.

**Implementation and control:**

The last phase consists of five main sub-phases. First, the data warehouse and each specific data mart are developed. These represent the information infrastructures that will feed the business intelligence system. In order to explain the meaning of the data contained in the data warehouse and the transformations applied in advance to the primary data, a *metadata* archive should be created, as described in Chapter 3. Moreover, ETL procedures are set `out to extract and transform the data existing in the primary sources, loading them into the data warehouse and the data marts. The next step is aimed at developing the core business intelligence applications that allow the planned analyses to be carried out. Finally, the system is released for test and usage.

Definition of system, Representation of the decision-making process, Evolution of information systems, Definition of decision support system, Development of a decision support system

## Decision support systems:

Definition of system:

# Decision Support and Business Intelligence Systems

## Chapter 1:

## Decision Support Systems and Business Intelligence

# Learning Objectives

- Understand today's turbulent business environment and describe how organizations survive and even excel in such an environment (solving problems and exploiting opportunities)
- Understand the need for computerized support of managerial decision making
- Understand an early framework for managerial decision making
- Learn the conceptual foundations of the decision support systems (DSS)

# Learning Objectives – cont.

- Describe the business intelligence (BI) methodology and concepts and relate them to DSS
- Describe the concept of work systems and its relationship to decision support
- List the major tools of computerized decision support
- Understand the major issues in implementing computerized support systems

# Opening Vignette:

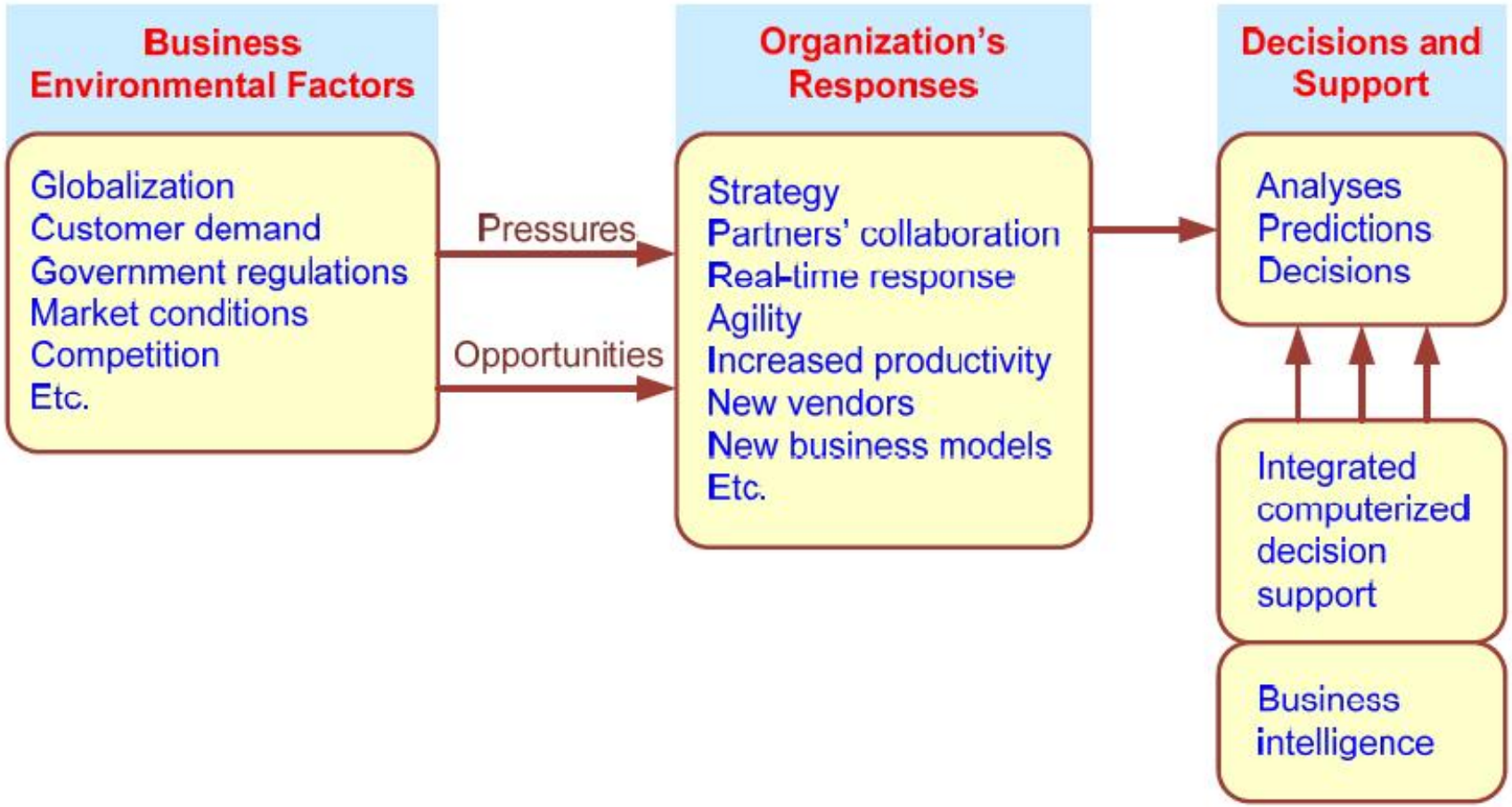
"Norfolk Southern Uses BI for Decision Support to Reach Success"

- Company background
- Problem
- Proposed solution
- Results
- Answer and discuss the case questions

# Changing Business Environment

- Companies are moving aggressively to computerized support of their operations => Business Intelligence
- Business Pressures–Responses–Support Model
  - Business pressures result of today's competitive business climate
  - Responses to counter the pressures
  - Support to better facilitate the process

# Business Pressures–Responses–Support Model



# The Business Environment

- The environment in which organizations operate today is becoming more and more complex, creating:
  - opportunities, and
  - problems
  - Example: globalization
- Business environment factors:
  - markets, consumer demands, technology, and societal...

# Business Environment Factors

| FACTOR | DESCRIPTION |
|---|---|
| **Markets** | Strong competition |
| | Expanding global markets |
| | Blooming electronic markets on the Internet |
| | Innovative marketing methods |
| | Opportunities for outsourcing with IT support |
| | Need for real-time, on-demand transactions |
| **Consumer demand** | Desire for customization |
| | Desire for quality, diversity of products, and speed of delivery |
| | Customers getting powerful and less loyal |
| **Technology** | More innovations, new products, and new services |
| | Increasing obsolescence rate |
| | Increasing information overload |
| | Social networking, Web 2.0 and beyond |
| **Societal** | Growing government regulations and deregulation |
| | Workforce more diversified, older, and composed of more women |
| | Prime concerns of homeland security and terrorist attacks |
| | Necessity of Sarbanes-Oxley Act and other reporting-related legislation |
| | Increasing social responsibility of companies |
| | Greater emphasis on sustainability |

# Organizational Responses

- Be Reactive, Anticipative, Adaptive, and Proactive
- Managers may take actions, such as
  - Employ strategic planning
  - Use new and innovative business models
  - Restructure business processes
  - Participate in business alliances
  - Improve corporate information systems
  - Improve partnership relationships
  - Encourage innovation and creativity ...cont...>

# Managers actions, continued

- Improve customer service and relationships
- Move to electronic commerce (e-commerce)
- Move to make-to-order production and on-demand manufacturing and services
- Use new IT to improve communication, data access (discovery of information), and collaboration
- Respond quickly to competitors' actions (e.g., in pricing, promotions, new products and services)
- Automate many tasks of white-collar employees
- Automate certain decision processes
- Improve decision making by employing analytics

# Closing the Strategy Gap

- One of the major objectives of computerized decision support is to facilitate closing the gap between the current performance of an organization and its desired performance, as expressed in its mission, objectives, and goals, and the strategy to achieve them

# Managerial Decision Making

- Management is a process by which organizational goals are achieved by using resources
  - Inputs: resources
  - Output: attainment of goals
  - Measure of success: outputs / inputs
- Management ] Decision Making
- Decision making: selecting the best solution from two or more alternatives

# Mintzberg's 10 Managerial Roles

**Interpersonal**

1. Figurehead
2. Leader
3. Liaison

**Informational**

4. Monitor
5. Disseminator
6. Spokesperson

**Decisional**

7. Entrepreneur
8. Disturbance handler
9. Resource allocator
10. Negotiator

# Decision Making Process

- Managers usually make decisions by following a four-step process (a.k.a. the scientific approach)
  1. Define the problem (or opportunity)
  2. Construct a model that describes the real-world problem
  3. Identify possible solutions to the modeled problem and evaluate the solutions
  4. Compare, choose, and recommend a potential solution to the problem

# Decision making is difficult, because

- Technology, information systems, advanced search engines, and globalization result in more and more alternatives from which to choose
- Government regulations and the need for compliance, political instability and terrorism, competition, and changing consumer demands produce more uncertainty, making it more difficult to predict consequences and the future
- Other factors are the need to make rapid decisions, the frequent and unpredictable changes that make trial-and-error learning difficult, and the potential costs of making mistakes

# Why Use Computerized DSS

- Computerized DSS can facilitate decision via:
  - Speedy computations
  - Improved communication and collaboration
  - Increased productivity of group members
  - Improved data management
  - Overcoming cognitive limits
  - Quality support; agility support
  - Using Web; anywhere, anytime support

# A Decision Support Framework

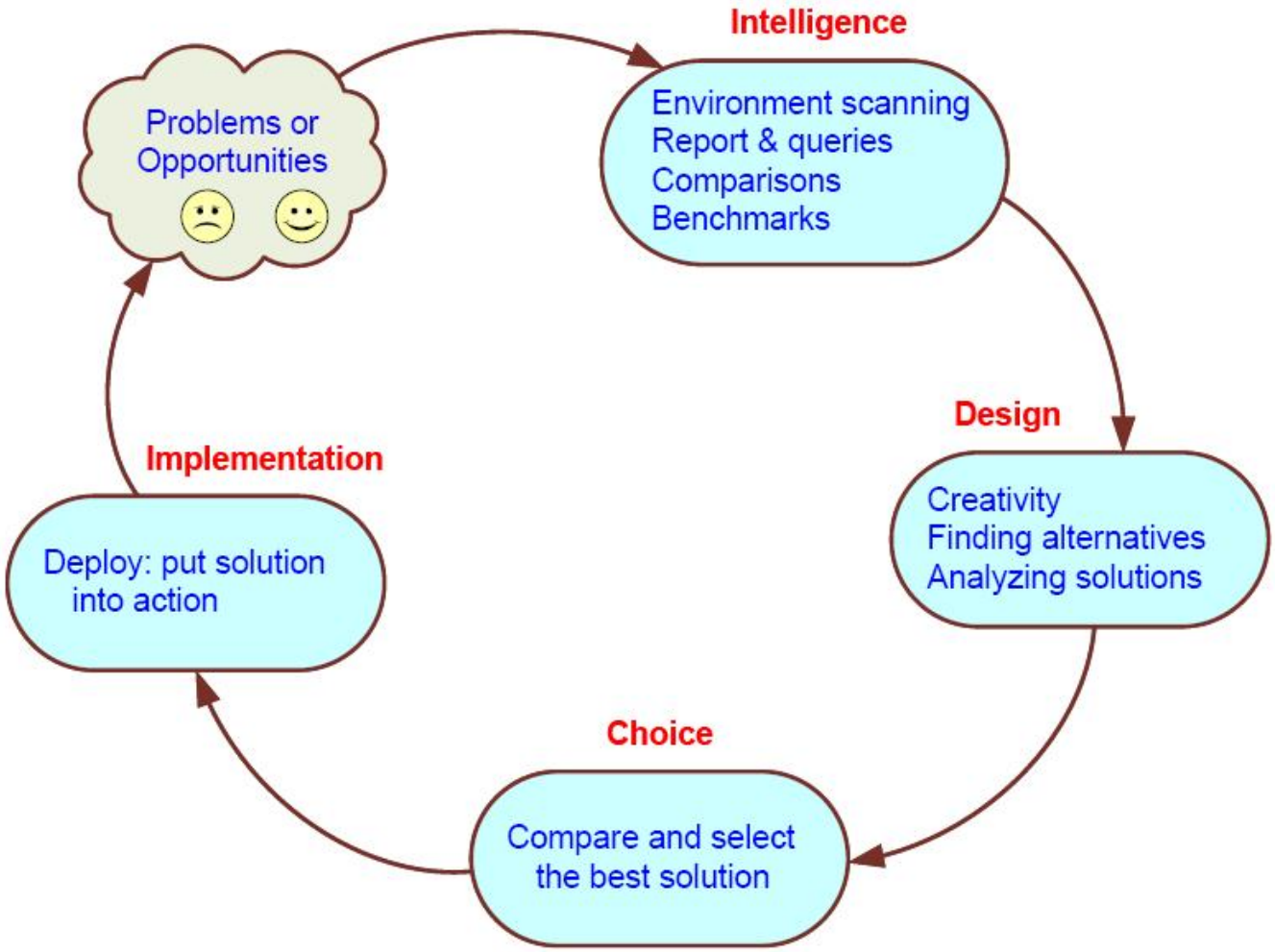
## (by Gory and Scott-Morten, 1971)

| | Type of Control | | |
|---|---|---|---|
| **Type of Decision** | **Operational Control** | **Managerial Control** | **Strategic Planning** |
| **Structured** | 1 Accounts receivable<br>Accounts payable<br>Order entry | 2 Budget analysis<br>Short-term forecasting<br>Personnel reports<br>Make-or-buy | 3 Financial management<br>Investment portfolio<br>Warehouse location<br>Distribution systems |
| **Semistructured** | 4 Production scheduling<br>Inventory control | 5 Credit evaluation<br>Budget preparation<br>Plant layout<br>Project scheduling<br>Reward system design<br>Inventory categorization | 6 Building a new plant<br>Mergers & acquisitions<br>New product planning<br>Compensation planning<br>Quality assurance<br>HR policies<br>Inventory planning |
| **Unstructured** | 7 Buying software<br>Approving loans<br>Operating a help desk<br>Selecting a cover for a magazine | 8 Negotiating<br>Recruiting an executive<br>Buying hardware<br>Lobbying | 9 R & D planning<br>New tech. development<br>Social responsibility planning |

# A Decision Support Framework – cont.

- Degree of Structuredness (Simon, 1977)
  - Decision are classified as
    - Highly structured (a.k.a. programmed)
    - Semi-structured
    - Highly unstructured (i.e., non-programmed)
- Types of Control (Anthony, 1965)
  - Strategic planning (top-level, long-range)
  - Management control (tactical planning)
  - Operational control

# Simon's Decision-Making Process



# Computer Support for Structured Decisions

- Structured problems: encountered repeatedly, have a high level of structure
- It is possible to abstract, analyze, and classify them into specific categories
  - e.g., make-or-buy decisions, capital budgeting, resource allocation, distribution, procurement, and inventory control
- For each category a solution approach is developed => Management Science

# Management Science Approach

- Also referred to as Operation Research
- In solving problems, managers should follow the five-step MS approach
  1. Define the problem
  2. Classify the problem into a standard category (*)
  3. Construct a model that describes the real-world problem
  4. Identify possible solutions to the modeled problem and evaluate the solutions
  5. Compare, choose, and recommend a potential solution to the problem

# Automated Decision Making

- A relatively new approach to supporting decision making
- Applies to highly structures decisions
- Automated decision systems (ADS)

  (or decision automation systems)
- An ADS is a rule-based system that provides a solution to a repetitive managerial problem in a specific area
  - e.g., simple-loan approval system

# Automated Decision Making

- ADS initially appeared in the airline industry called revenue (or yield) management (or revenue optimization) systems
  - dynamically price tickets based on actual demand
- Today, many service industries use similar pricing models
- ADS are driven by business rules!

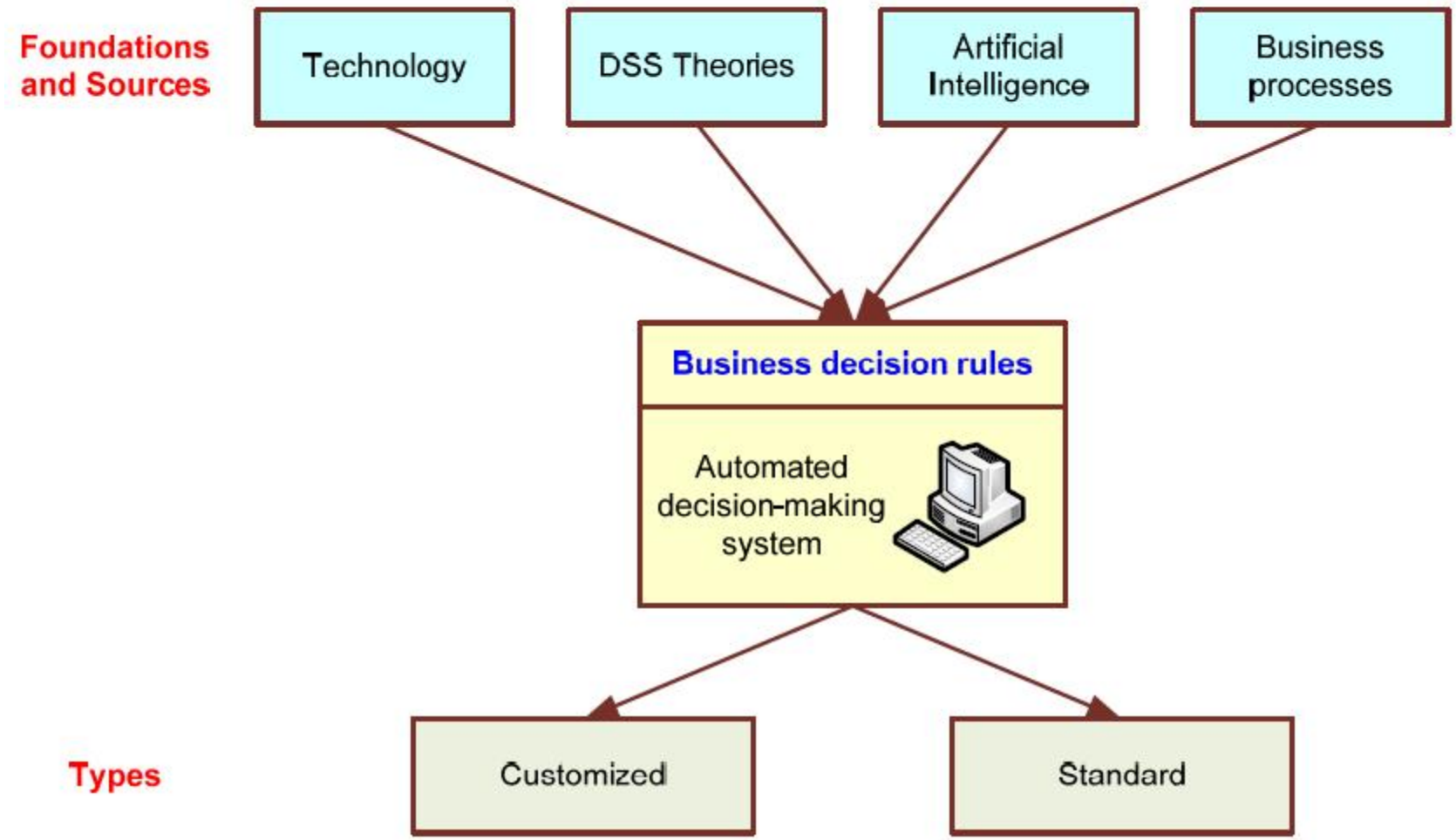
# Computer Support for Unstructured Decisions

- Unstructured problems can be only partially supported by standard computerized quantitative methods
- They often require customized solutions
- They benefit from data and information
- Intuition and judgment may play a role
- Computerized communication and collaboration technologies along with knowledge management is often used

# Computer Support for Semi-structured Problems

- Solving semi-structured problems may involve a combination of standard solution procedures and human judgment
- MS handles the structured parts while DSS deals with the unstructured parts
- With proper data and information, a range of alternative solutions, along with their potential impacts

# Automated Decision-Making Framework



# Concept of Decision Support Systems

## Classical Definitions of DSS

- Interactive computer-based systems, which help decision makers utilize data and models to solve unstructured problems" - Gorry and Scott-Morton, 1971

- Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semistructured problems - Keen and Scott-Morton, 1978

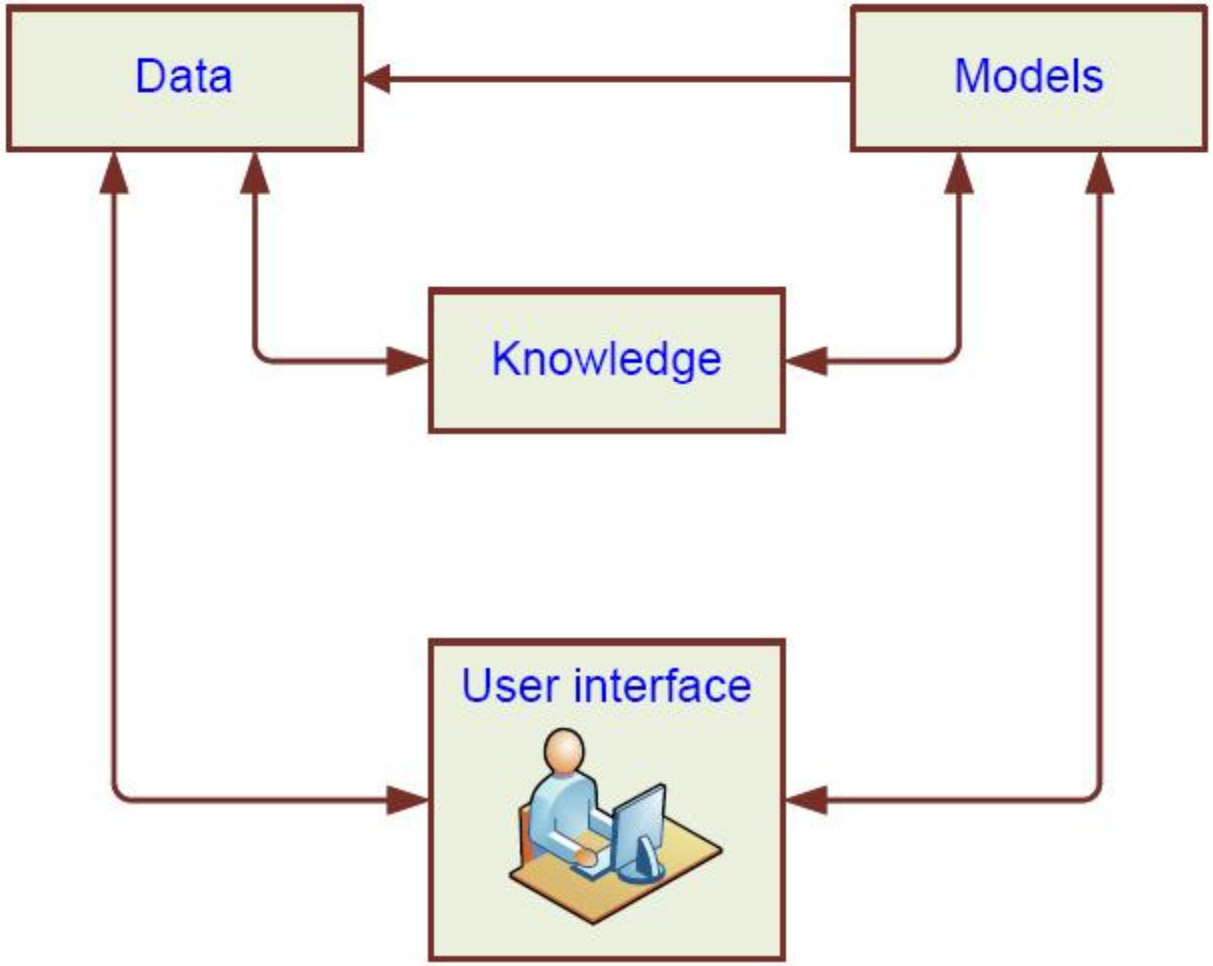
# DSS as an Umbrella Term

- The term DSS can be used as an umbrella term to describe any computerized system that supports decision making in an organization
  - E.g., an organization wide knowledge management system; a decision support system specific to an organizational function (marketing, finance, accounting, manufacturing, planning, SCM, etc.)

# DSS as a Specific Application

- In a narrow sense DSS refers to a process for building customized applications for unstructured or semi-structured problems
- Components of the DSS Architecture
  - Data, Model, Knowledge/Intelligence, User, Interface (API and/or user interface)
  - DSS often is created by putting together loosely coupled instances of these components

# High-Level Architecture of a DSS



# Types of DSS

- Two major types:
  - Model-oriented DSS
  - Data-oriented DSS
- Evolution of DSS into Business Intelligence
  - Use of DSS moved from specialist to managers, and then whomever, whenever, wherever
  - Enabling tools like OLAP, data warehousing, data mining, intelligent systems, delivered via Web technology have collectively led to the term "business intelligence" (BI) and "business analytics"

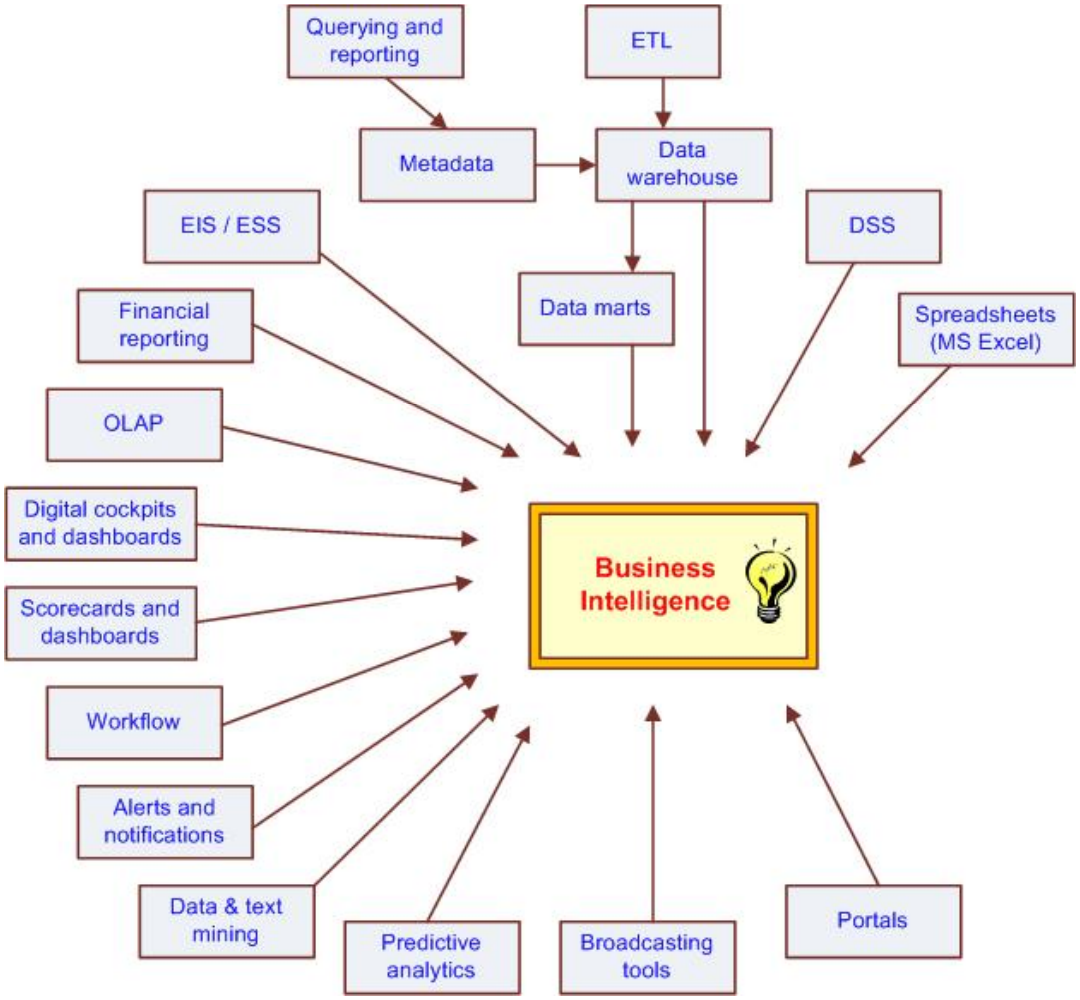
# Business Intelligence (BI)

- BI is an umbrella term that combines architectures, tools, databases, analytical tools, applications, and methodologies
- Like DSS, BI a content-free expression, so it means different things to different people
- BI's major objective is to enable easy access to data (and models) to provide business managers with the ability to conduct analysis
- BI helps transform data, to information (and knowledge), to decisions and finally to action

# A Brief History of BI

- The term BI was coined by the Gartner Group in the mid-1990s
- However, the concept is much older
  - 1970s - MIS reporting - static/periodic reports
  - 1980s - Executive Information Systems (EIS)
  - 1990s - OLAP, dynamic, multidimensional, ad-hoc reporting -> coining of the term “BI”
  - 2005+ Inclusion of AI and Data/Text Mining capabilities; Web-based Portals/Dashboards
  - 2010s - yet to be seen

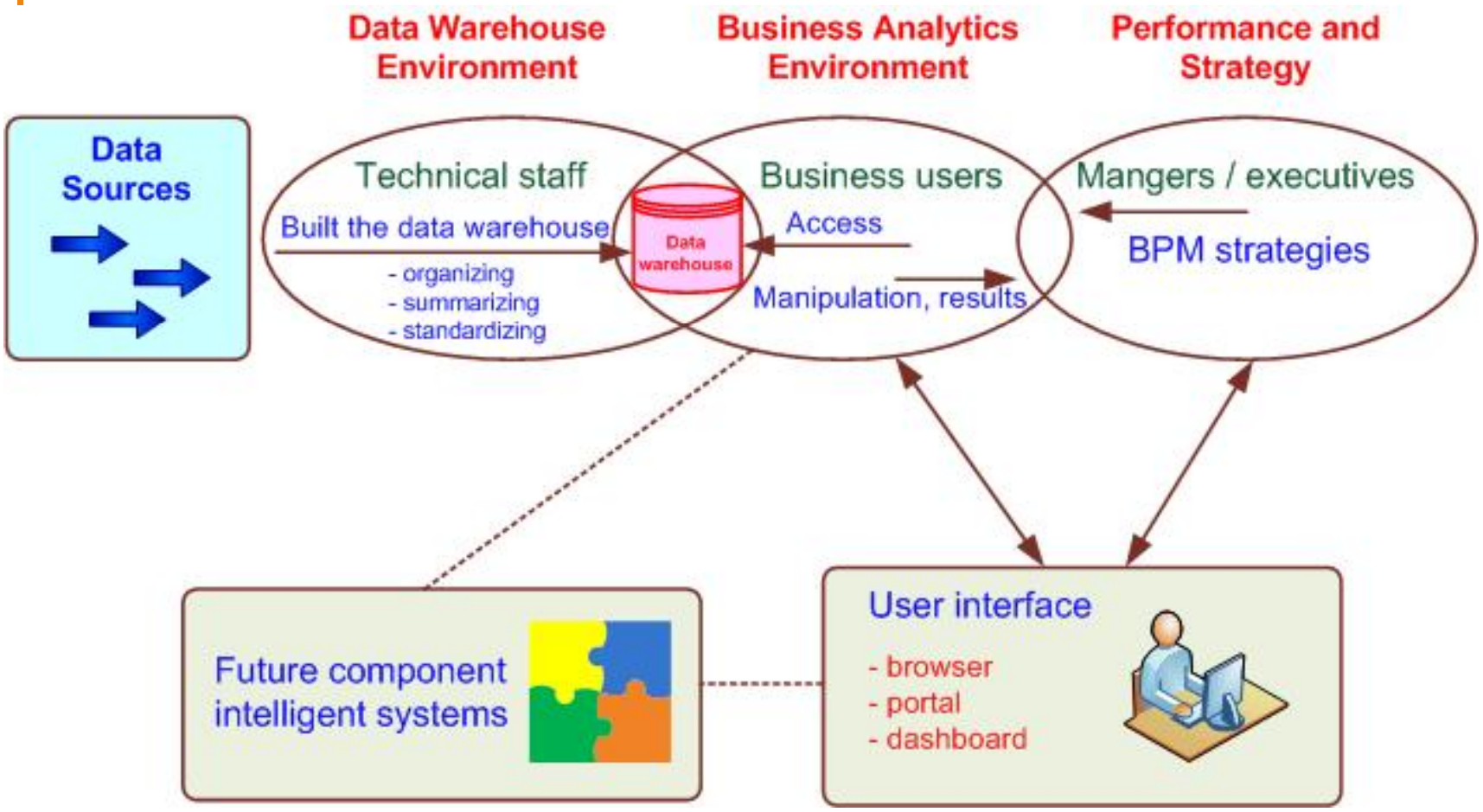
# The Evolution of BI Capabilities



# The Architecture of BI

- A BI system has four major components
  - a data warehouse, with its source data
  - business analytics, a collection of tools for manipulating, mining, and analyzing the data in the data warehouse;
  - business performance management (BPM) for monitoring and analyzing performance
  - a user interface (e.g., dashboard)

# A High-Level Architecture of BI



# Components in a BI Architecture

- The data warehouse is a large repository of well-organized historical data
- Business analytics are the tools that allow transformation of data into information and knowledge
- Business performance management (BPM) allows monitoring, measuring, and comparing key performance indicators
- User interface (e.g., dashboards) allows access and easy manipulation of other BI components

# Styles of BI

- MicroStrategy, Corp. distinguishes five styles of BI and offers tools for each
  1. report delivery and alerting
  2. enterprise reporting (using dashboards and scorecards)
  3. cube analysis (also known as slice-and-dice analysis)
  4. ad-hoc queries
  5. statistics and data mining

# The Benefits of BI

- The ability to provide accurate information when needed, including a real-time view of the corporate performance and its parts
- A survey by Thompson (2004)
  - Faster, more accurate reporting (81%)
  - Improved decision making (78%)
  - Improved customer service (56%)
  - Increased revenue (49%)
- See Table 1.3 for a list of BI analytic applications, the business questions they answer and the business value they bring

# The DSS–BI Connection

- First, their architectures are very similar because BI evolved from DSS
- Second, DSS directly support specific decision making, while BI provides accurate and timely information, and indirectly support decision making
- Third, BI has an executive and strategy orientation, especially in its BPM and dashboard components, while DSS, in contrast, is oriented toward analysts

# The DSS–BI Connection – cont.

- Fourth, most BI systems are constructed with commercially available tools and components, while DSS is often built from scratch
- Fifth, DSS methodologies and even some tools were developed mostly in the academic world, while BI methodologies and tools were developed mostly by software companies
- Sixth, many of the tools that BI uses are also considered DSS tools (e.g., data mining and predictive analysis are core tools in both)

# The DSS–BI Connection – cont.

- Although some people equate DSS with BI, these systems are not, at present, the same
  - some people believe that DSS is a part of BI—one of its analytical tools
  - others think that BI is a special case of DSS that deals mostly with reporting, communication, and collaboration (a form of data-oriented DSS)
  - BI is a result of a continuous revolution and, as such, DSS is one of BI's original elements
  - In this book, we separate DSS from BI
- MSS = BI and/or DSS

# A Work System View of Decision Support (Alter, 2004)

- drop the word “systems” from DSS
- focus on “decision support”

  “use of any plausible computerized or noncomputerized means for improving decision making in a particular repetitive or nonrepetitive business situation in a particular organization”

- Work system: a system in which human participants and/or machines perform a business process, using information, technology, and other resources, to produce products and/or services for internal or external customers

# Elements of a Work System

1. Business process. Variations in the process rationale, sequence of steps, or methods used for performing particular steps
2. Participants. Better training, better skills, higher levels of commitment, or better real-time or delayed feedback
3. Information. Better information quality, information availability, or information presentation
4. Technology. Better data storage and retrieval, models, algorithms, statistical or graphical capabilities, or computer interaction

-->

# Elements of a Work System – cont.

5. Product and services. Better ways to evaluate potential decisions
6. Customers. Better ways to involve customers in the decision process and to obtain greater clarity about their needs
7. Infrastructure. More effective use of shared infrastructure, which might lead to improvements
8. Environment. Better methods for incorporating concerns from the surrounding environment
9. Strategy. A fundamentally different operational strategy for the work system

# Major Tool Categories for MSS

| TOOL CATEGORY | TOOLS AND THEIR ACRONYMS |
|---|---|
| Data management | Databases and database management system (DBMS)<br>Extraction, transformation, and load (ETL) systems<br>Data warehouses (DW), real-time DW, and data marts |
| Reporting status tracking | Online analytical processing (OLAP)<br>Executive information systems (EIS) |
| Visualization | Geographical information systems (GIS)<br>Dashboards, Information portals<br>Multidimensional presentations |
| Business analytics | Optimization, Web analytics<br>Data mining, Web mining, and text mining |
| Strategy and performance management | Business performance management (BPM)/<br>Corporate performance management (CPM)<br>Business activity management (BAM)<br>Dashboards and Scorecards |
| Communication and collaboration | Group decision support systems (GDSS)<br>Group support systems (GSS)<br>Collaborative information portals and systems |
| Social networking<br>Knowledge management | Web 2.0, Expert locating systems<br>Knowledge management systems (KMS) |
| Intelligent systems | Expert systems (ES)<br>Artificial neural networks (ANN)<br>Fuzzy logic, Genetic algorithms, Intelligent agents |
| Enterprise systems | Enterprise resource planning (ERP),<br>Customer Relationship Management (CRM), and<br>Supply-Chain Management (SCM) |

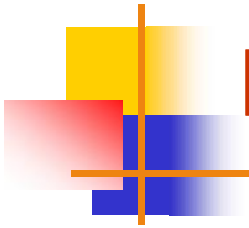
Source: Table 1.4

# Hybrid (Integrated) Support Systems

- The objective of computerized decision support, regardless of its name or nature, is to assist management in solving managerial or organizational problems (and assess opportunities and strategies) faster and better than possible without computers
- Every type of tool has certain capabilities and limitations. By integrating several tools, we can improve decision support because one tool can provide advantages where another is weak
- The trend is therefore towards developing hybrid (integrated) support system

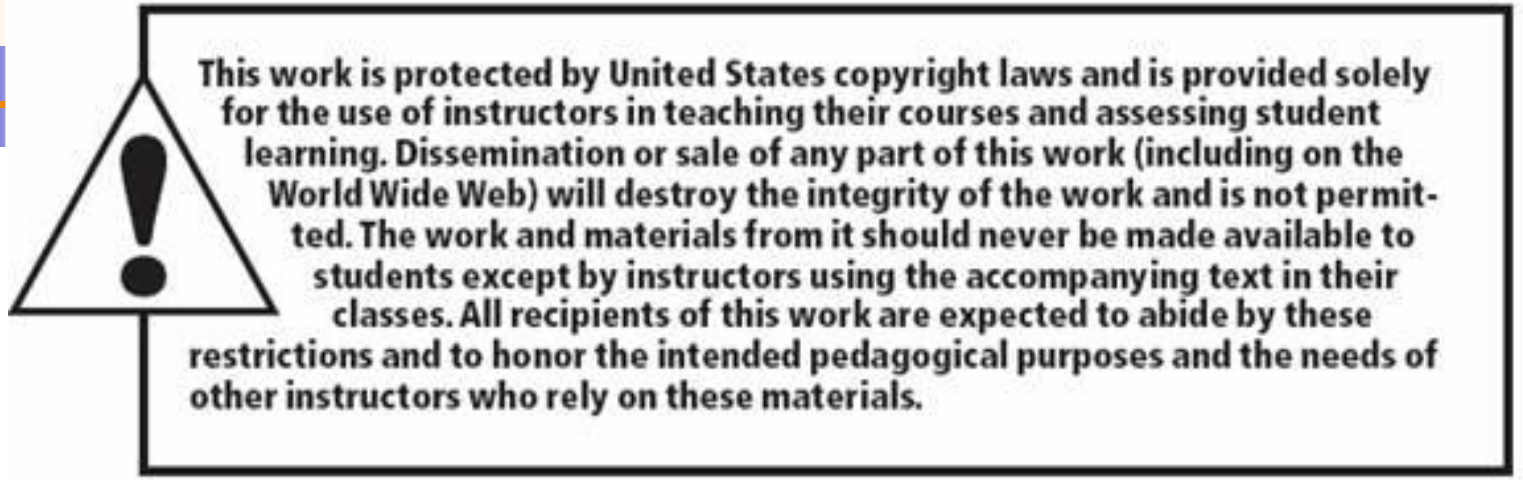
# Hybrid (Integrated) Support Systems

- Type of integration
  - Use each tool independently to solve different aspects of the problem
  - Use several loosely integrated tools. This mainly involves transferring data from one tool to another for further processing
  - Use several tightly integrated tools. From the user's standpoint, the tool appears as a unified system
- In addition to performing different tasks in the problem-solving process, tools can support each other



# End of the Chapter

- Questions / Comments...



# Chapter 5
# BI Definitions and Concepts

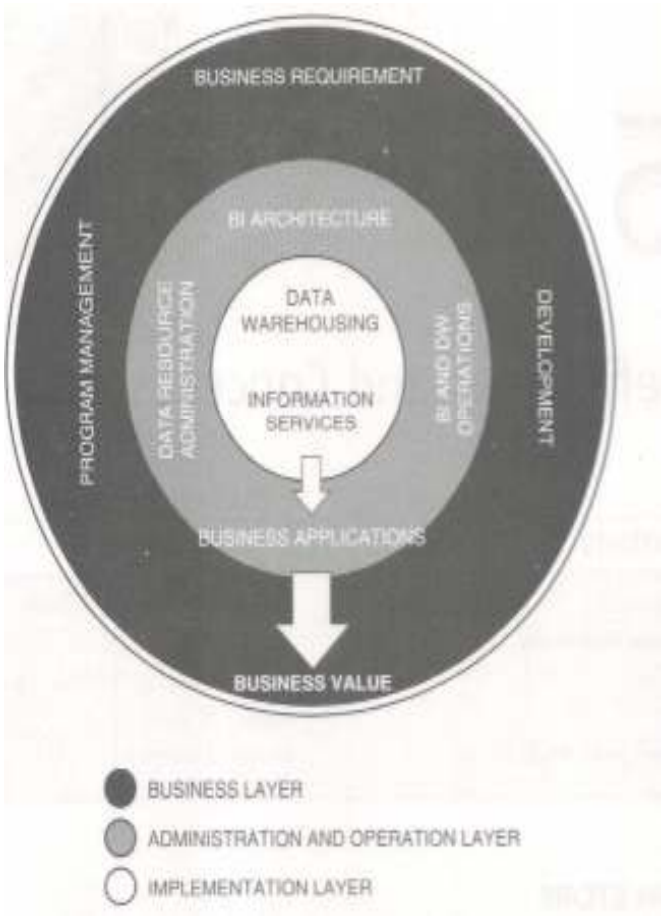
**Content of this presentation has been taken from Book**

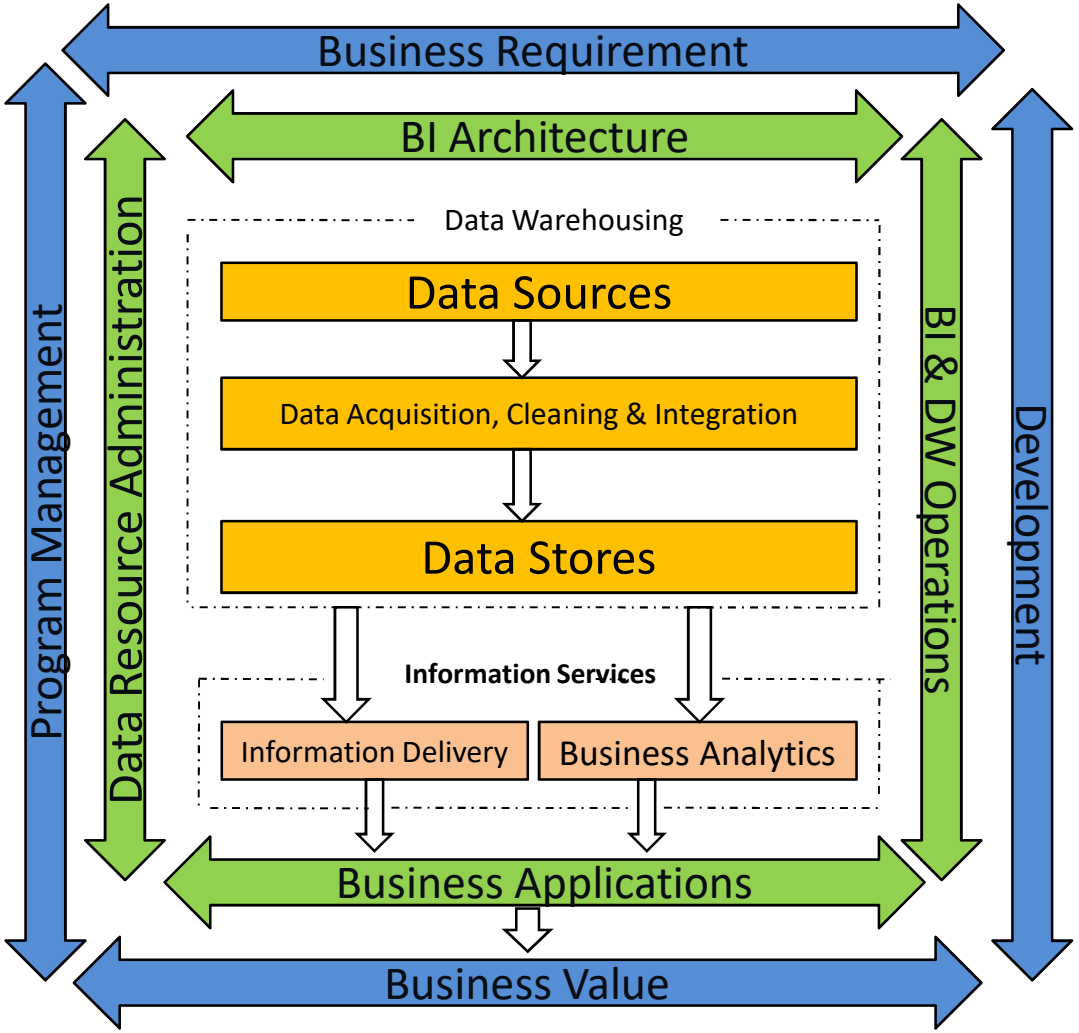
# "Fundamentals of Business Analytics"

RN Prasad and Seema Acharya

Published by Wiley India Pvt. Ltd.

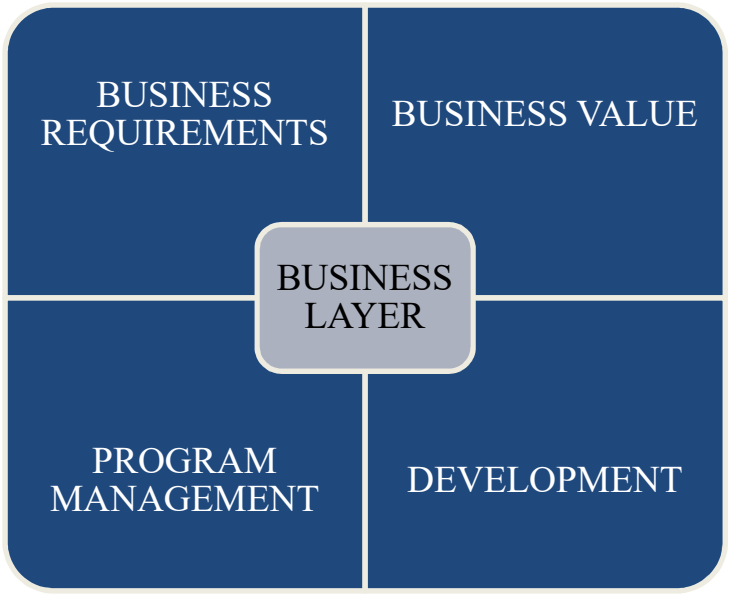
# BI Component Framework



## Business Layer

This layer consists of four components –




1. Business requirements
   - Business drivers
   - Business Goals
   - Business Strategies

2. Business Value
   - Return on Investment
   - Return on Assest
   - Total Cost of Ownership
   - Total Value of Ownership

3. Program Management

4. Development

## Business Layer – Business Requirements

**Business requirements:** The requirements are a product of three steps of a process that includes:

- *Business drivers* - the impulses that initiate the need to act.

  Examples: changing workforce, changing labor laws, changing economy, changing technology, etc.
- *Business goals*- the targets to be achieved in response to the business drivers.

  Examples: increased productivity, improved market share, improved profit margins, improved customer satisfaction, cost reduction, etc.
- *Business strategies*- the planned course of action that will help achieve the set goals.

  Examples: outsourcing, global delivery model, partnerships, customer retention programs, employee retention programs, competitive pricing, etc.

# Business Layer- Business Value

When a strategy is implemented against certain business goals, then certain costs (monetary, time, effort, information produced by data integration and analysis, application of knowledge from past experience, etc.) are involved.

However, the final output of the process should create such value for the business whose ratio to the costs involved should be a feasible ratio.

The business value can be measured in the terms of ROI (Return on Investment), ROA (Return on Assets), TCO (Total Cost of Ownership), TVO(Total Value of Ownership), etc. Let us understand these terms with the help of a few examples –

**Return on Investment (ROI):** We take the example of "Digicom", a digital electrocompany which has an online community platform that allows their prospective clients to engage with their users. "Digicom" has been using social media (mainly Twitter and Facebook) to help get new clients and to increase the number of prospects/leads. They attribute 10% of their daily revenue to social media. Now, that is an ROI from social media!

**Return on Asset (ROA):** Suppose a company, "Electronics Today", has a net income of $l million and has total assets of $5 million. Then, its ROA is 20%. So, ROA is the earning from invested capital (assets).

**Total Cost of Ownership (TCO):** Let us understand TCO in the context of a vehicle. TCO defines the cost of owning a vehicle from the time of purchase by the owner, through its operation and maintenance to the time it leaves the possession of the owner.

**Total Value of Ownership (TVO):** TVO has replaced the simple concept of Owner's Equity in some companies. It could include a variety of subcategories such as stock, undistributed dividends, retained earnings or profit, or excess capital contributed.

In its simplest form, the basic accounting equation containing TVO as a component is

Assets = Liabilities + Owner's Equity, or if you like TVO

## Business Layer- Program Management

This component of the business layer ensures that people, projects, and priorities work in a manner in which individual processes are compatible with each other so as to ensure seamless integration and smooth functioning of the entire program. It should attend to each of the following:

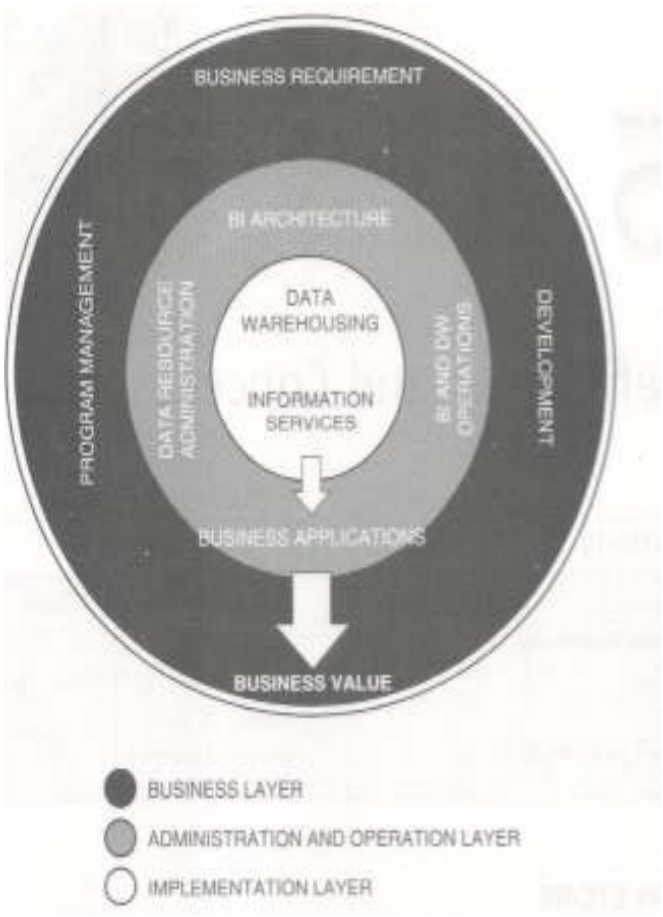
- Business priorities
- Mission and goals
- Strategies and risks
- Multiple projects
- Dependencies
- Cost and value
- Business rules
- Infrastructure

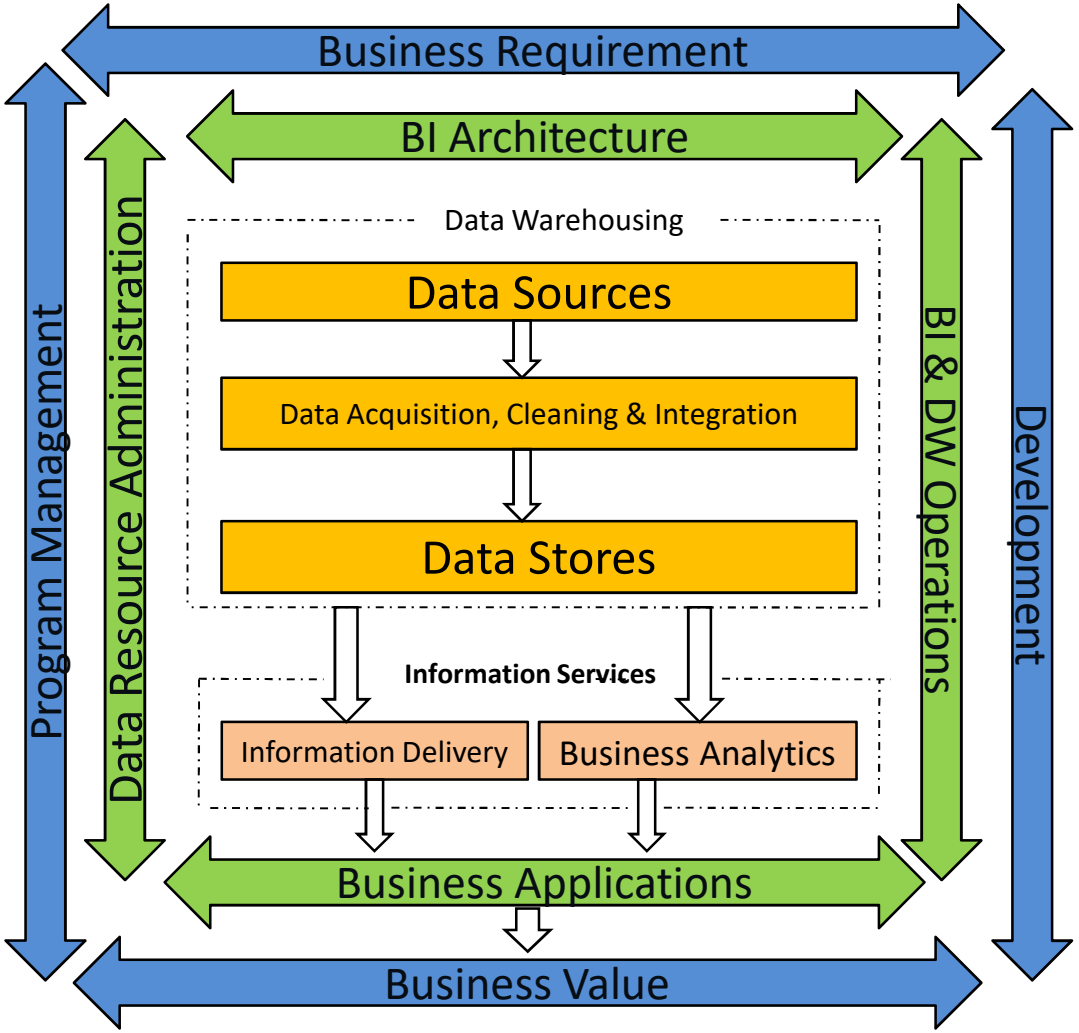
## Business Layer- Development

The process of development consists of

- *database/data-warehouse development* (consisting of ETL, data profiling, data cleansing and database tools),
- *data integration system development* (consists of data integration tools and data quality tools)
- *business analytics development* (about processes and various technologies used).

# BI Component Framework

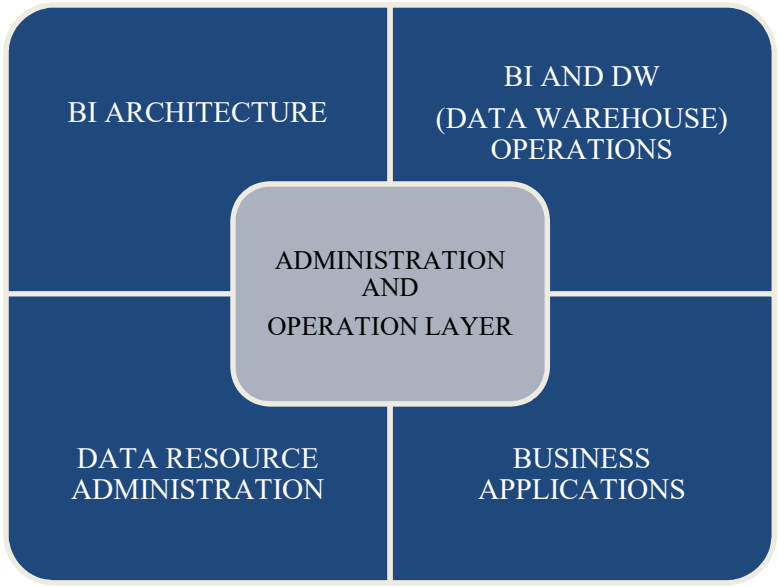


# BI Component - Administration and Operation Layer

**This layer consists of four components-**

1. BI Architecture
    a. Data
    b. Integration
    c. Information
    d. Technology
    e. Organization
2. BI and DW Operations
    a. Backup and restore
    b. Security
    c. Configuration and Management
    d. Database Management
3. Data Resource Management
    a. Data Governance
    b. Metadata management
4. Business Applications



# BI Component - Administration and Operations Layer - BI Architecture

| | |
|---|---|
| **DATA** | • Should follow design standards<br>• Must have a logically apt data model<br>• Metadata should be of high standard |
| **INTEGRATION** | • Performed according to business semantics and rules<br>• During integration, certain processing standards have to be followed<br>• Data must be consistent |
| **INFORMATION** | • Information derived from data that has been integrated should be usable, findable and as per the requirements |
| **TECHNOLOGY** | • Technology used for deriving information must be accessible<br>• Also, it should have a good user-interface<br>• Should support analysis, decision support, data and storage management |
| **ORGANIZATION** | • Consists of different roles and responsibilities, like management, development, support and usage roles |

## BI Component - Administration and Operations Layer – BI and DW Operations

Data Warehouse (DW) administration requires the usage of various tools to monitor the performance and usage of the warehouse, and perform administrative tasks on it. Some of these tools would be:

- Backup and restore
- Security
- Configuration management
- Database management

**Data resource administration:** Involves *data governance* and *metadata management*.

***Data governance*** is a technique for controlling data quality, which is used to assess, improve, manage and maintain information. It helps to define standards that are required to maintain data quality. The distribution of roles for governance of data is as follows:

- Data ownership
- Data stewardship
- Data custodianship

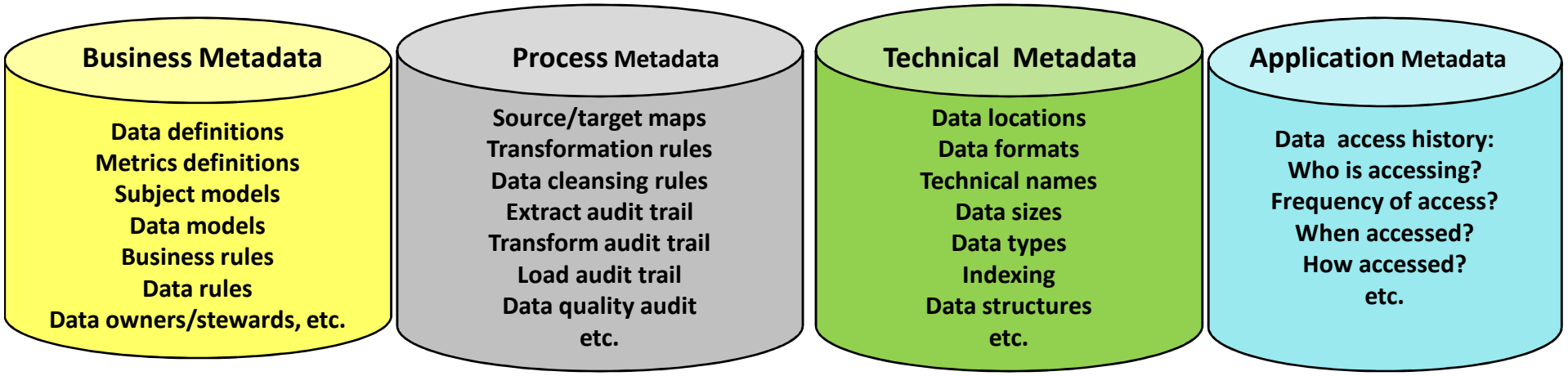
# BI Component - Administration and Operations Layer– BI and DW Operations

***Metadata management:*** Metadata is data about data.

Consider CD/DVD of music. There is the date of recording, the name of the artist, the genre of music, the songs in the album, copyright information, etc. All this information constitutes the metadata for the CD/DVD of music. In the context of a camera, the data is the photographic image. The metadata then is the date and time when the was taken. In simple words, metadata is data about data. When used in the context of a data warehouse, it is the data that defines the warehouse objects. Few examples of metadata are timestamp at which the data was extracted, the data sources from where metadata has been extracted, and the missing fields/columns that have been added by data cleaning or integration processes. Metadata management involves tracking, assessment, and maintenance of metadata.

**Metadata can be divided into four groups:**

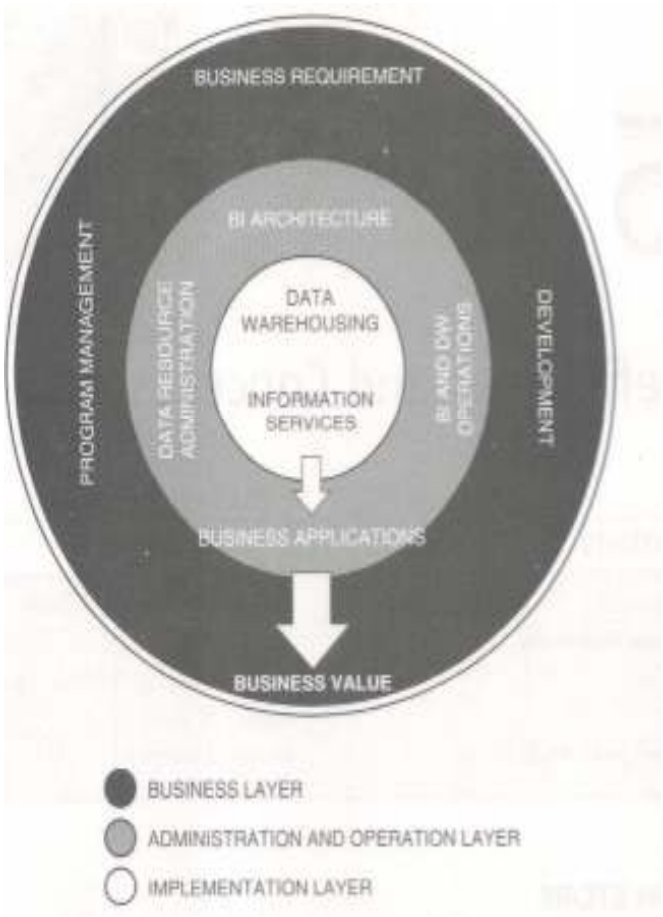
- Business metadata
- Process metadata
- Technical metadata
- Application metadata

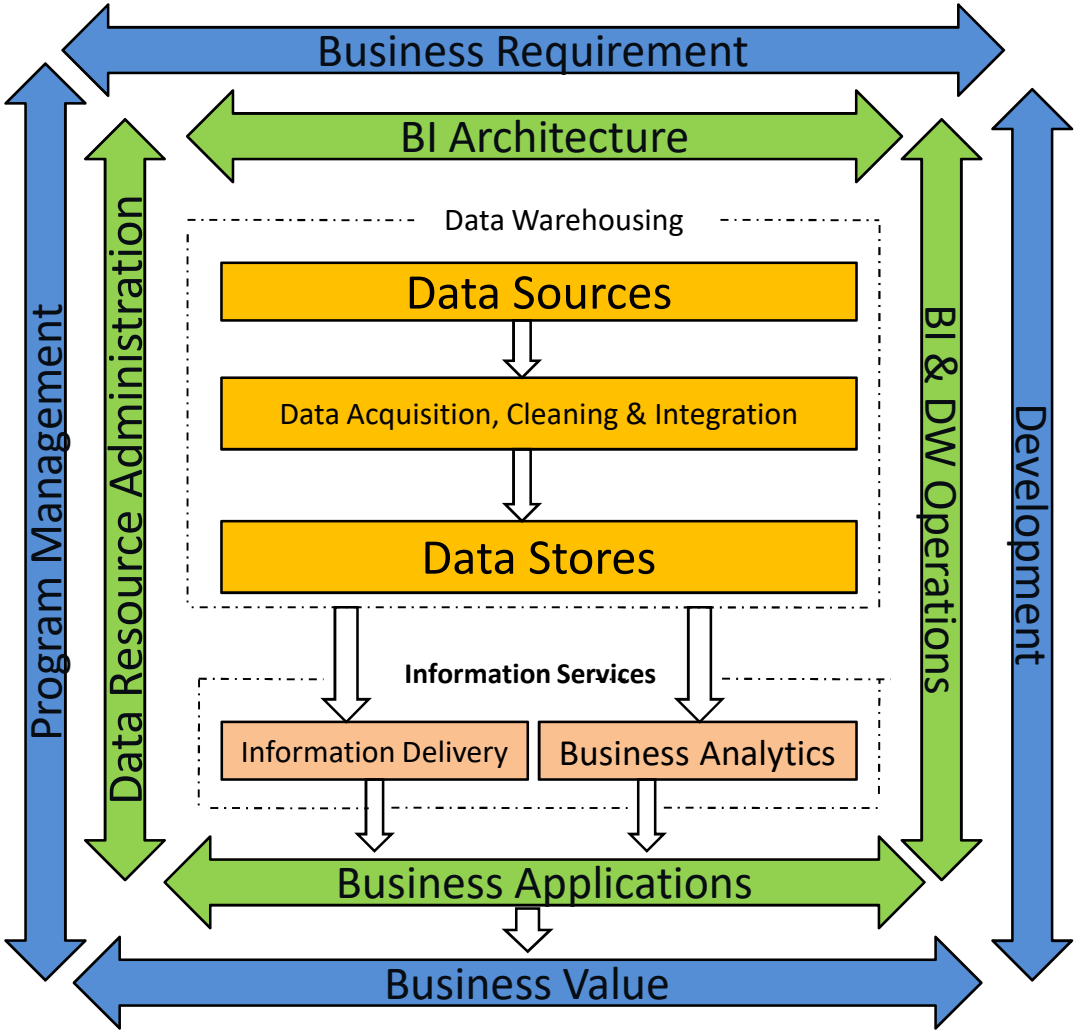


## BI Component - Administration and Operations Layer – Business Applications

The application of technology to produce value for the business refers to the generation of information or intelligence from data assets like data warehouses/data marts. Using BI tools, we can generate strategic, financial, customer, or risk intelligence. This information can be obtained through various BI applications, such as DSS (decision support system), EIS (executive information system), OLAP(On-line analytical processing), data mining and discovery, etc.

# BI Component Framework



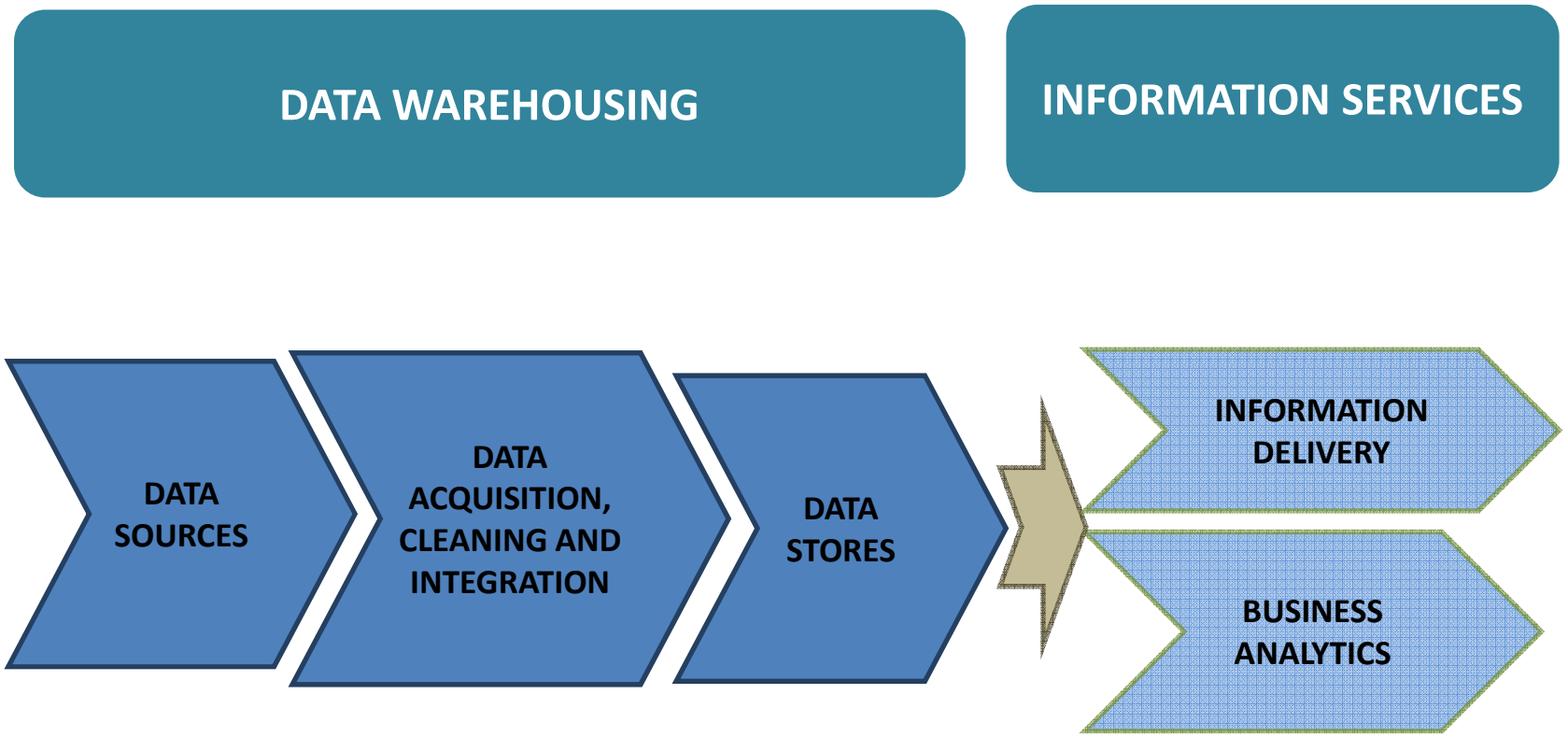



# BI Component – Implementation Layer

The implementation layer of the BI component framework consists of technical components that are required for data capture, transformation and cleaning, data into information, and finally delivering that information to leverage business goals and produce value for the organization.

1. Data Warehousing
    1. Data Sources
    2. Data Acquisition, Cleaning, and Integration
    3. Data Stores
2. Information Services
    1. Information Delivery
    2. Business Analytics

# BI Component - Implementation Layer



# BI Component – Implementation Layer – Data Warehousing

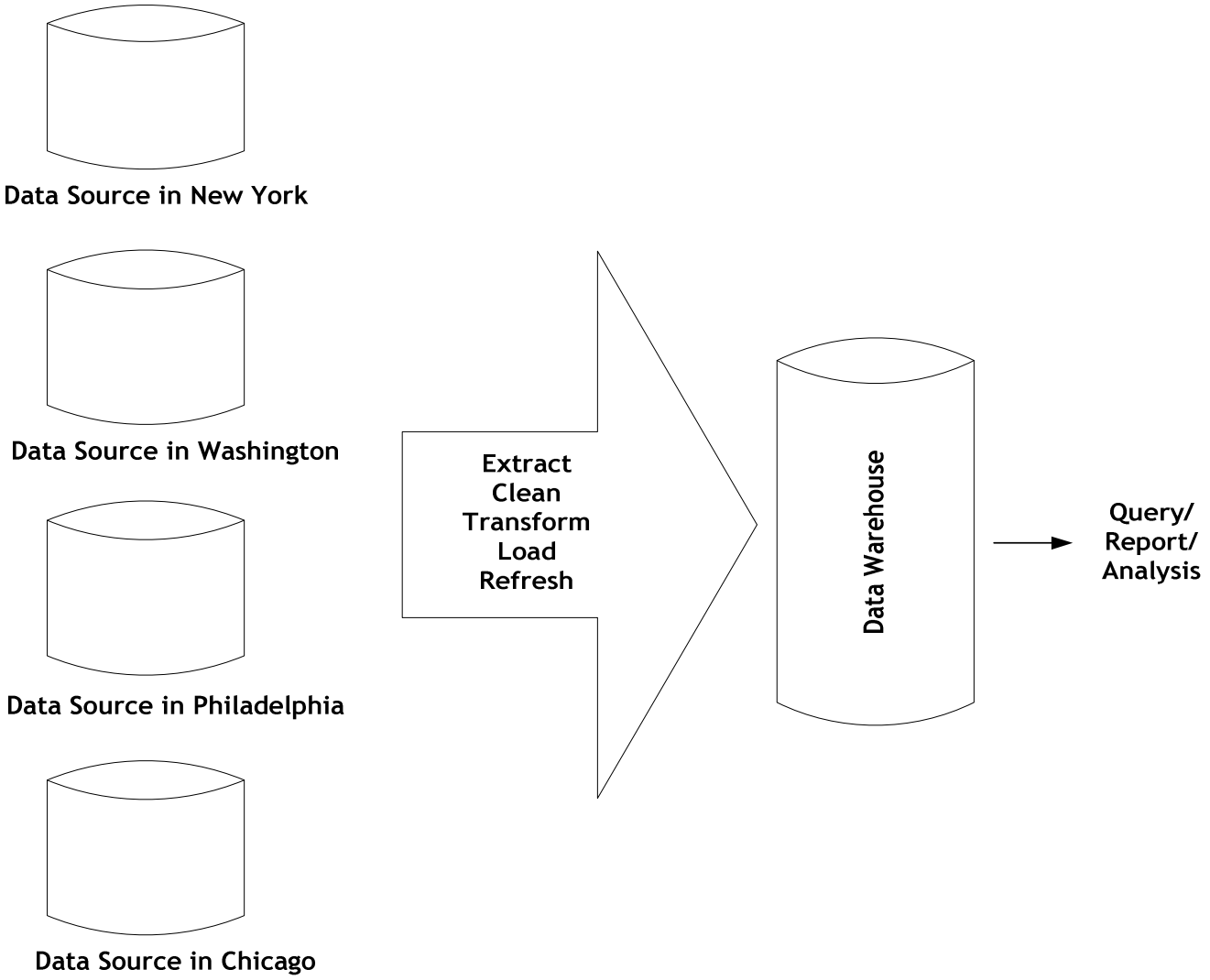
It is the process which prepares the basic repository of data (called data warehouse) that becomes the data source where we extract information from.

**Date Warehouse:** A data warehouse is a data store. It is structured on the dimensional model schema, which is optimized for data retrieval rather than update.

Data warehousing must play the following five distinct roles:

- Intake
- Integration
- Distribution
- Delivery
- Access

# Implementation Layer



## BI Component – Implementation Layer – Information Services

- It is not only the process of producing information; rather, it involves ensuring that the information produced is aligned with business requirements and can be acted upon to produce value for the company.

- Information is delivered in the form of KPI's, reports, charts, dashboards or scorecards, etc., or in the form of analytics.

- Data mining is a practice used to increase the body of knowledge.

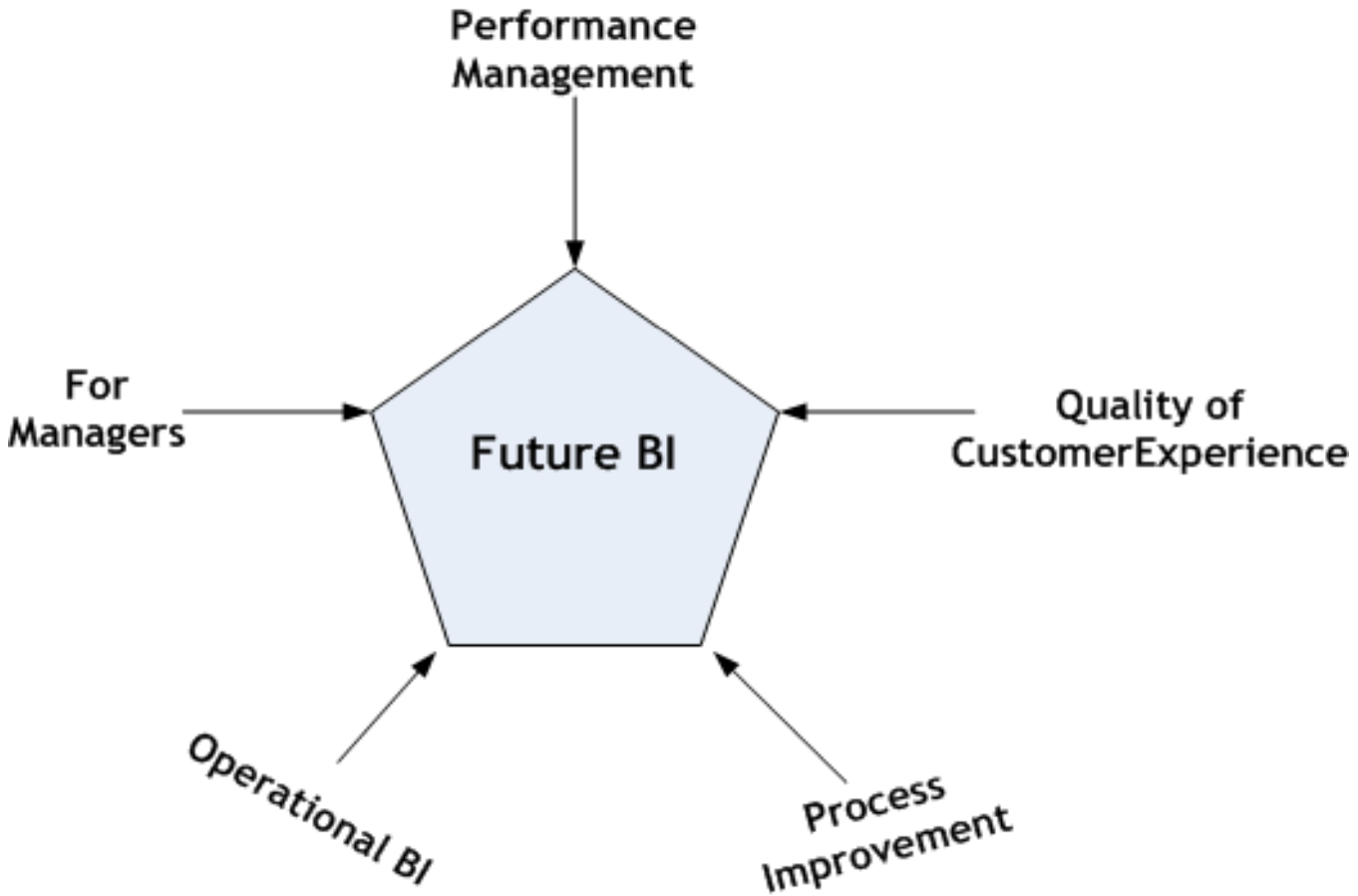
- Applied analytics is generally used to drive action and produce outcomes.

## Answer a Quick Question

**Is BI only for managers?**

## Who is BI for?

It is a misnomer to believe that BI is only for managers or the executive class. True, it is used more often by them. But does that mean that BI can be used only for management and control? Thus, the answer is: NO!



## Types of BI Users

| Type of user | Casual users/ Information consumers | Power users/Information producers |
|---|---|---|
| Example of such users | Executives, managers, customers, suppliers, field/operation workers, etc. | SAS, SPSS developers, administrators, business analysts, analytical modelers, IT professionals, etc. |
| Usage | Information consumers | Information producers |
| Data Access | Tailor made to suit the needs of their respective role | Ad hoc/exploratory |
| Tools | Pre-defined reports/dashboards | Advanced Analytical/ Authoring tools |
| Sources | Data warehouse/Data Marts | Data Warehouse/Data Marts (both internal and external) |

## BI Applications

BI applications can be divided into:

- **Technology solutions**
  - DSS
  - EIS
  - OLAP
  - Managed Query and Reporting
  - Data Mining
- **Business Solutions**
  - Performance Analysis
  - Customer Analysis
  - Market Place Analysis
  - Productivity Analysis
  - Sales Channel Analysis
  - Behavioral Analysis
  - Supply Chain Analysis

# BI Roles and Responsibilities

| Program Roles | Project Roles |
|---|---|
| | Business Manager |
| BI Program Manager | BI Business Specialist |
| BI Data Architect | BI Project Manager |
| BI ETL Architect | Business Requirements Analyst |
| BI Technical Architect | Decision Support Analyst |
| Metadata Manager | BI Designer |
| BI Administrator | ETL Specialist |
| | Data Administrator |

# BI DW Best Practices

**The list of best practices is adapted from an article TDWI's FlashPoint e-newsletter of April 10, 2003.**

- Practice "User First" Design
- Create New Value
- Attend to Human Impacts
- Focus on Information and Analytics
- Practice Active Data Stewardship
- Manage BI as a long term investment
- Reach out with BI/DW solutions
- Make BI a business Initiative
- Measure Results
- Attend to strategic Positioning

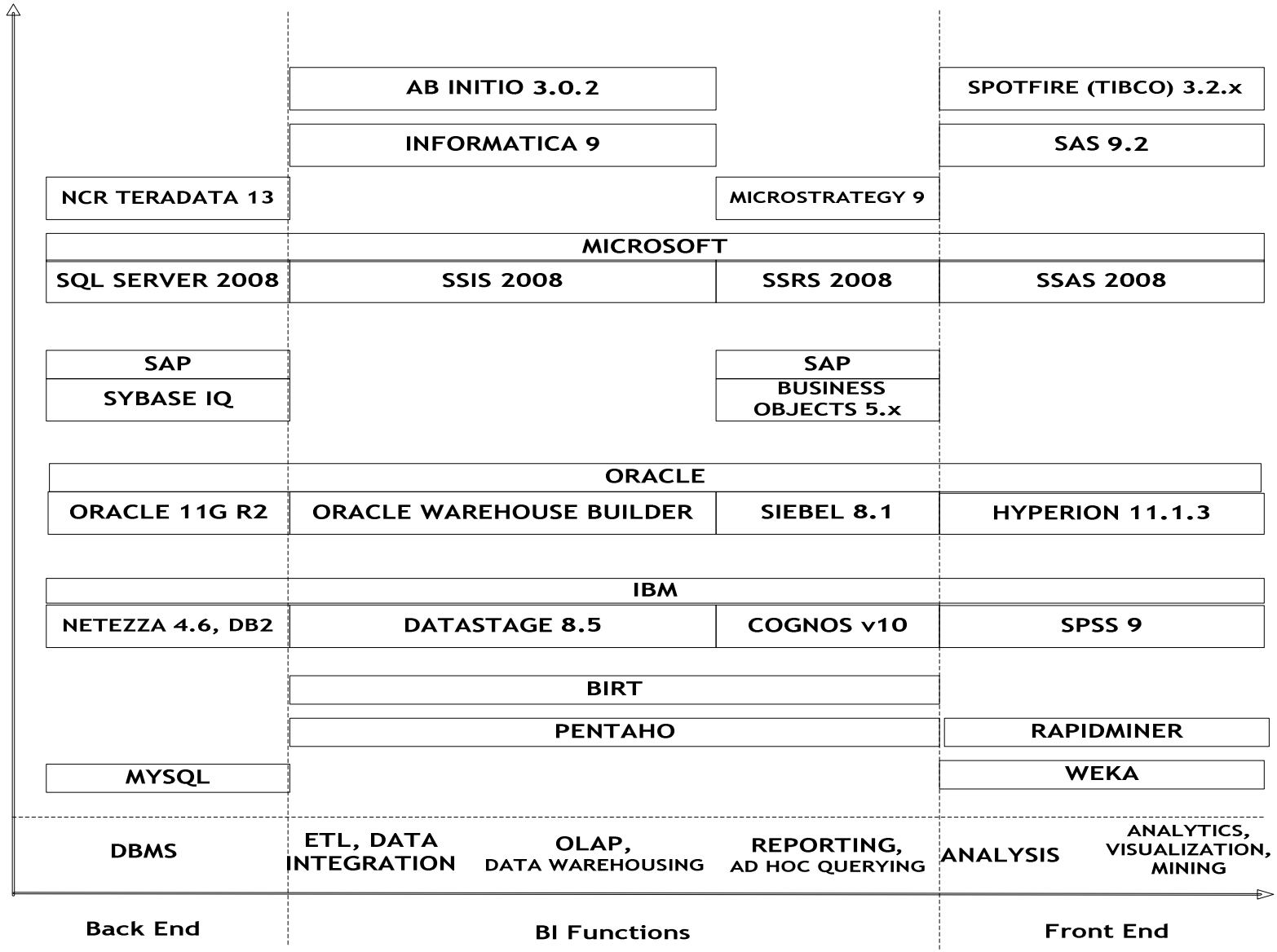
## Do It exercise

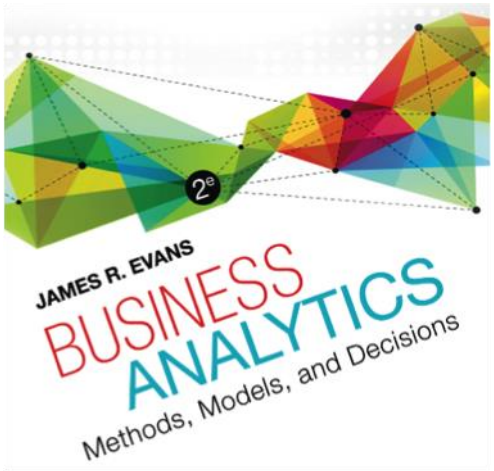
**Visit www.tdwi.org to read more about BI DW best practices**

## Open Source BI Tools

| | |
|---|---|
| **RDBMS** | MySQL, Firebird |
| **ETL Tools** | Pentaho Data Integration (formerly called Kettle), SpagoBI |
| **Analysis Tools** | Weka, RapidMiner, SpagoBI |
| **Reporting Tools/Ad Hoc Querying/Visualization** | Pentaho, BIRT, Actuate, Jaspersoft |

# Popular BI Tools






# Chapter 1 Introduction to Business Analytics


Modified and Shortened

# Contents

- **Introduction to Analytics**
- Tools
- Data
- Models
- Problem solving with analytics

# Business Analytics

**Analytics** is the use of:

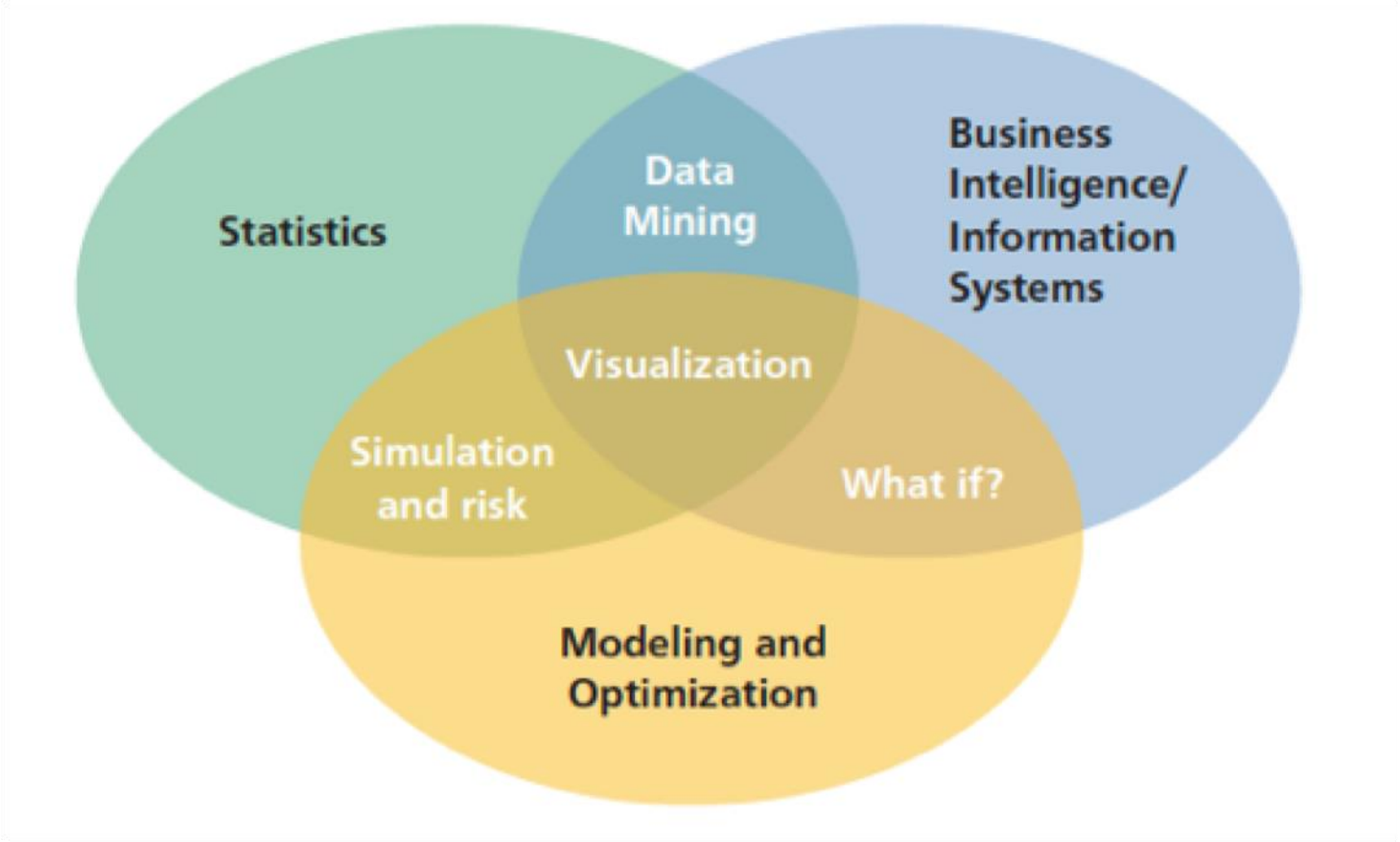
- data,
- information technology,
- statistical analysis,
- quantitative methods, and
- mathematical or computer-based models

to help managers gain improved insight about their business operations and **make better, fact-based decisions.**

# Examples of Applications

- **Pricing**
  - setting prices for consumer and industrial goods, government contracts, and maintenance contracts
- **Customer segmentation**
  - identifying and targeting key customer groups in retail, insurance, and credit card industries
- **Merchandising**
  - determining brands to buy, quantities, and allocations
- **Location**
  - finding the best location for bank branches and ATMs, or where to service industrial equipment
- **Social Media**
  - understand trends and customer perceptions; assist marketing managers and product designers

# A Visual Perspective of Business Analytics



# Impacts and Challenges

- **Benefits**
  - …reduced costs, better risk management, faster decisions, better productivity and enhanced bottom-line performance such as profitability and customer satisfaction.

- **Challenges**
  - …lack of understanding of how to use analytics, competing business priorities, insufficient analytical skills, difficulty in getting good data and sharing information, and not understanding the benefits versus perceived costs of analytics studies.

Privacy?

# Scope of Business Analytics

- **Descriptive analytics**: the use of data to understand past and current business performance and make informed decisions

- **Predictive analytics**: predict the future by examining historical data, detecting patterns or relationships in these data, and then extrapolating these relationships forward in time.

- **Prescriptive analytics**: identify the best alternatives to minimize or maximize some objective

# Example 1.1: Retail Markdown Decisions

- Most department **stores clear seasonal inventory** by reducing prices.
- *Key question*: When to reduce the **price** and by how much to maximize revenue?

- Potential applications of analytics:
  - Descriptive analytics: examine historical data for similar products (prices, units sold, advertising, …)
  - Predictive analytics: predict sales based on price
  - Prescriptive analytics: find the best sets of pricing and advertising to maximize sales revenue

# Contents

- Introduction to Analytics
- **Tools**
- Data
- Models
- Problem solving with analytics

# Tools

- Database queries and analysis
- Spreadsheets
- Data visualization
- Dashboards to report key performance measures
- Data and Statistical methods
- Data Mining basics (predictive models)

In this course

- Simulation
- Forecasting
- Scenario and “what-if” analyses
- Optimization
- Text Mining
- Social media, web, and text analytics

# Software Support






- **SQL** various databases
- **Excel** Spreadsheets
- Tableau Software Simple drag and drop tools for visualizing data from spreadsheets and other databases.
- IBM Cognos Express An integrated business intelligence and planning solution designed to meet the needs of midsize companies, provides reporting, analysis, dashboard, scorecard, planning, budgeting and forecasting capabilities.
- SAS / SPSS / **Rapid Miner** Predictive modeling and data mining, visualization, forecasting, optimization and model management, statistical analysis, text analytics, and more using visual workflows.
- R / Python Advanced programing-based data preparation, analytics and visualization.

# Contents

- Introduction to Analytics
- Tools
- **Data**
- Models
- Problem solving with analytics

# Data for Business Analytics

- **Data**: numerical or textual facts and figures that are collected through some type of measurement process.

- **Information**: result of analyzing data; that is, extracting **meaning** from data to support evaluation and decision making.

# Examples of Data Sources and Uses

- Internal
  - Annual reports
  - Accounting audits
  - Financial profitability analysis
  - Operations management performance
  - Human resource measurements
- External
  - Economic trends
  - Marketing research
- New developments: Web behavior – Social Media – Mobile - IOT
  - page views, visitor's country, time of view, length of time, origin and destination paths, products they searched for and viewed, products purchased, what reviews they read, and many others.

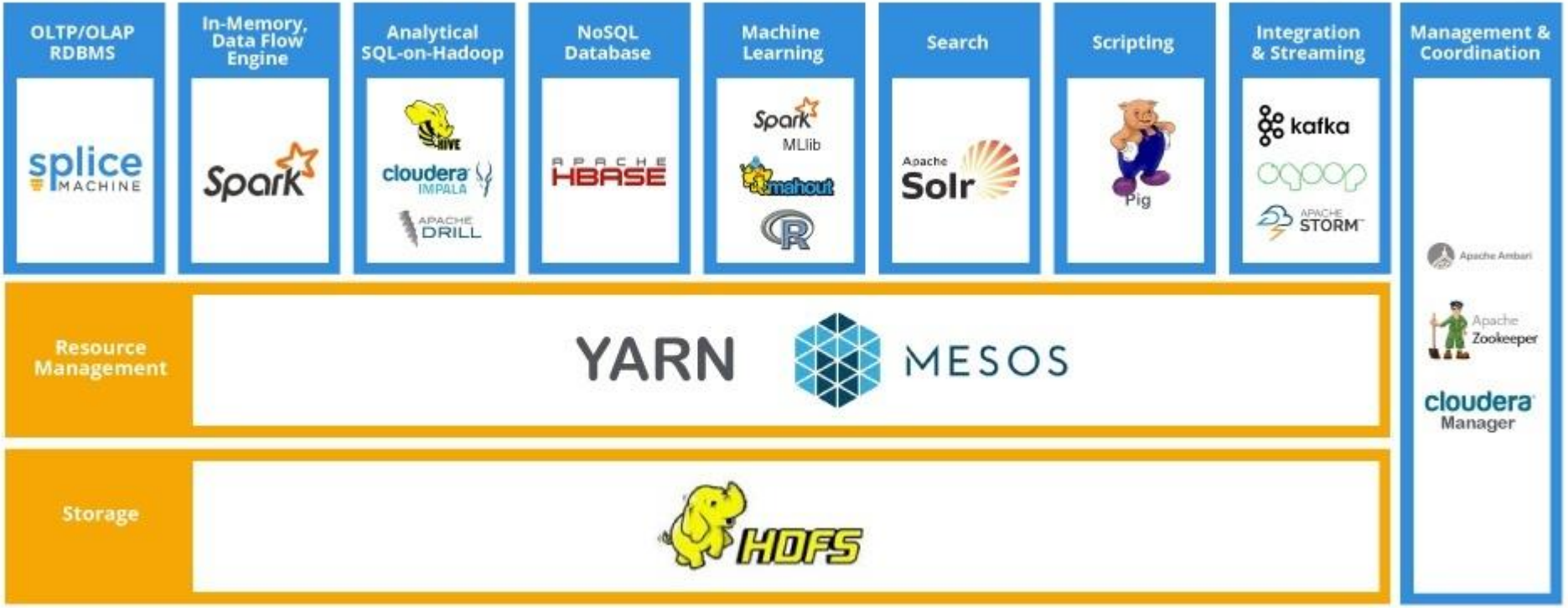
# Big Data

- **Big data** to refer to massive amounts of business data from a wide variety of sources, much of which is available in real time, and much of which is uncertain or unpredictable. IBM calls these characteristics **volume**, **variety**, **velocity,** and **veracity**.

"*The effective use of big data has the potential to transform economies, delivering a new wave of productivity growth and consumer surplus. Using big data will become a key basis of competition for existing companies, and will create new competitors who are able to attract employees that have the critical skills for a big data world.*" - McKinsey Global Institute, 2011

# Big Data

- Apache Hadoop Ecosystem for Big Data



# Data Sets and Databases

- **Database** - a collection of related tables containing records on people, places, or things.
  - In a database table the columns correspond to each individual element of data (called *fields*, or *attributes*), and the rows represent records of related data elements.

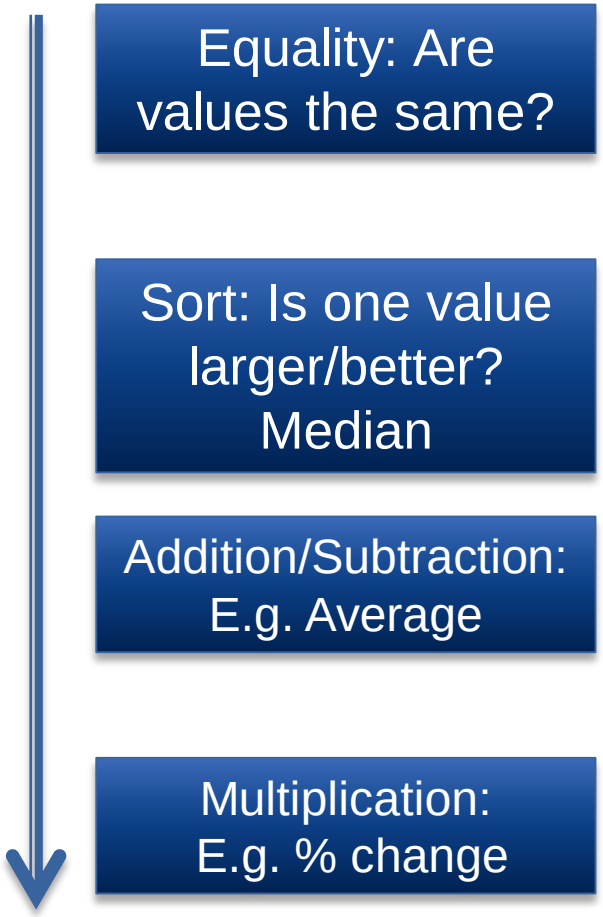

- **Data set** - a collection of data (often a single "spread sheet" or data mining table).
  - Examples: Marketing survey responses, a table of historical stock prices, and a collection of measurements of dimensions of a manufactured item.

# Types of Data

- **Discrete** - derived from **counting** something.
  - For example, a delivery is either on time or not; an order is complete or incomplete; or an invoice can have one, two, three, or any number of errors. Some discrete metrics would be the proportion of on-time deliveries; the number of incomplete orders each day, and the number of errors per invoice.
- **Continuous** based on a **continuous scale of measurement.**
  - Any metrics involving dollars, length, time, volume, or weight, for example, are continuous.

# Measurement Scales

- **Categorical (nominal) data** - sorted into categories according to specified characteristics.
- **Ordinal data** - can be ordered or ranked according to some relationship to one another.
- **Interval data** - ordinal but have constant differences between observations and have arbitrary zero points.
- **Ratio data** - continuous and have a natural zero.



# Example 1.3: Classifying Data Elements

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **Purchase Orders** | | | | | | | | | |
| 2 | | | | | | | | | | |
| 3 | **Supplier** | **Order No.** | **Item No.** | **Item Description** | **Item Cost** | **Quantity** | **Cost per order** | **A/P Terms (Months)** | **Order Date** | **Arrival Date** |
| 4 | Hulkey Fasteners | Aug11001 | 1122 | Airframe fasteners | $ 4.25 | 19,500 | $ 82,875.00 | 30 | 08/05/11 | 08/13/11 |
| 5 | Alum Sheeting | Aug11002 | 1243 | Airframe fasteners | $ 4.25 | 10,000 | $ 42,500.00 | 30 | 08/08/11 | 08/14/11 |
| 6 | Fast-Tie Aerospace | Aug11003 | 5462 | Shielded Cable/ft. | $ 1.05 | 23,000 | $ 24,150.00 | 30 | 08/10/11 | 08/15/11 |
| 7 | Fast-Tie Aerospace | Aug11004 | 5462 | Shielded Cable/ft. | $ 1.05 | 21,500 | $ 22,575.00 | 30 | 08/15/11 | 08/22/11 |
| 8 | Steelpin Inc. | Aug11005 | 5319 | Shielded Cable/ft. | $ 1.10 | 17,500 | $ 19,250.00 | 30 | 08/20/11 | 08/31/11 |
| 9 | Fast-Tie Aerospace | Aug11006 | 5462 | Shielded Cable/ft. | $ 1.05 | 22,500 | $ 23,625.00 | 30 | 08/20/11 | 08/26/11 |
| 10 | Steelpin Inc. | Aug11007 | 4312 | Bolt-nut package | $ 3.75 | 4,250 | $ 15,937.50 | 30 | 08/25/11 | 09/01/11 |
| 11 | Durrable Products | Aug11008 | 7258 | Pressure Gauge | $ 90.00 | 100 | $ 9,000.00 | 45 | 08/25/11 | 08/28/11 |
| 12 | Fast-Tie Aerospace | Aug11009 | 6321 | O-Ring | $ 2.45 | 1,300 | $ 3,185.00 | 30 | 08/25/11 | 09/04/11 |
| 13 | Fast-Tie Aerospace | Aug11010 | 5462 | Shielded Cable/ft. | $ 1.05 | 22,500 | $ 23,625.00 | 30 | 08/25/11 | 09/02/11 |
| 14 | Steelpin Inc. | Aug11011 | 5319 | Shielded Cable/ft. | $ 1.10 | 18,100 | $ 19,910.00 | 30 | 08/25/11 | 09/05/11 |
| 15 | Hulkey Fasteners | Aug11012 | 3166 | Electrical Connector | $ 1.25 | 5,600 | $ 7,000.00 | 30 | 08/25/11 | 08/29/11 |

Categorical
Ordinal
Categorical
Categorical
Ratio
Ratio
Ratio
Ratio
Interval
Interval

# Data Reliability and Validity

- **Reliability** - data are accurate and consistent.
- **Validity** - data measures what it is supposed to measure.

- Examples:
  - A tire pressure gage that consistently reads several pounds of pressure below the true value is **not reliable**, although it is valid because it does measure tire pressure.
  - The number of calls to a customer service desk might be counted correctly each day (and thus is a reliable measure) but **not valid** if it is used to assess customer dissatisfaction, as many calls may be simple queries.
  - A survey question that asks a customer to rate the quality of the food in a restaurant may be **neither reliable** (because different customers may have conflicting perceptions) **nor valid** (if the intent is to measure customer satisfaction, as satisfaction generally includes other elements of service besides food).

# Contents

- Introduction to Analytics
- Tools
- Data
- **Models**
- Problem solving with analytics

# Models in Business Analytics

- **Model** - an abstraction or representation of a real system, idea, or object.
  - Often a **simplification** of the real thing.
  - Captures the **most important features.**
  - Can be a written or verbal description, a visual representation, a mathematical formula, or a spreadsheet.

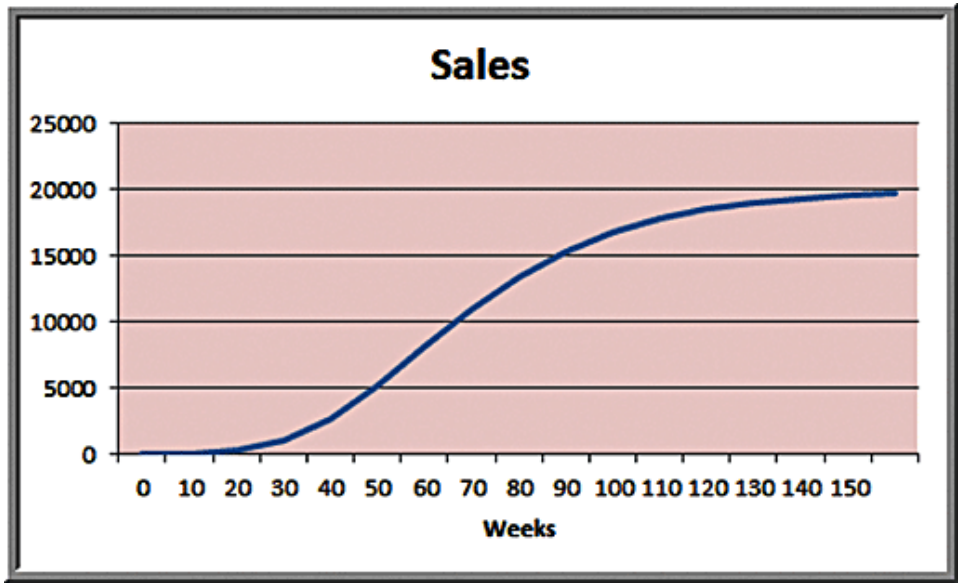
# Example 1.4: Three Forms of a Model

The sales of a new product, such as a first-generation iPad or 3D television, often follow a common pattern.

1. **Verbal description**: The rate of sales starts small as early adopters begin to evaluate a new product and then begins to grow at an increasing rate over time as positive customer feedback spreads. Eventually, the market begins to become saturated and the rate of sales begins to decrease.

# Example 1.4 (continued)

**2. Visual model**: A sketch of sales as an S-shaped curve over time



# Example 1.4 (continued)

**3. Mathematical model**:

$$S = ae^{be^{ct}}$$

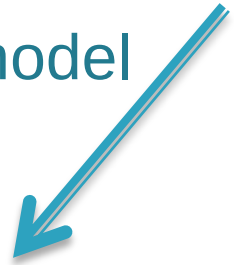
where

- $S$ is sales,
- $t$ is time,
- $e$ is the base of natural logarithms, and
- $a$, $b$ and $c$ are constants that need to be estimated.

# From Data to Model

| Week | Price ($) | Coupon (0,1) | Advertising ($) | Store 1 Sales (Units) | Store 2 Sales (Units) | Store 3 Sales (Units) |
|---|---|---|---|---|---|---|
| 1 | $6.99 | 0 | $0 | 501 | 510 | 481 |
| 2 | $6.99 | 0 | $150 | 772 | 748 | 775 |
| 3 | $6,99 | 1 | $0 | 554 | 528 | 506 |
| 4 | $6.99 | 1 | $150 | 838 | 785 | 834 |
| 5 | $6.49 | 0 | $0 | 521 | 519 | 500 |
| 6 | $6.49 | 0 | $150 | 723 | 790 | 723 |
| 7 | $6.49 | 1 | $0 | 510 | 556 | 520 |
| 8 | $6.49 | 1 | $150 | 818 | 773 | 800 |
| 9 | $7.59 | 0 | $0 | 479 | 491 | 486 |
| 10 | $7.59 | 0 | $150 | 825 | 822 | 757 |
| 11 | $7.59 | 1 | $0 | 533 | 513 | 540 |
| 12 | $7.59 | 1 | $150 | 839 | 791 | 832 |
| 13 | $5.49 | 0 | $0 | 484 | 480 | 508 |
| 14 | $5.49 | 0 | $150 | 686 | 683 | 708 |
| 15 | $5.49 | 1 | $0 | 543 | 531 | 530 |
| 16 | $5.49 | 1 | $150 | 767 | 743 | 779 |

Learn model



**Model:** Sales = 500 – 0.05(price) + 30(coupons) + 0.08(advertising) + 0.25(price)(advertising)

If the price is $6.99, no coupons are offered, and no advertising is done (the experiment corresponding to week 1), the model estimates sales as

Sales = 500 - 0.05 × $6.99 + 30 × 0 + 0.08 × 0 + 0.25 × $6.99 × 0 = 500 units

**How do we find this model?**

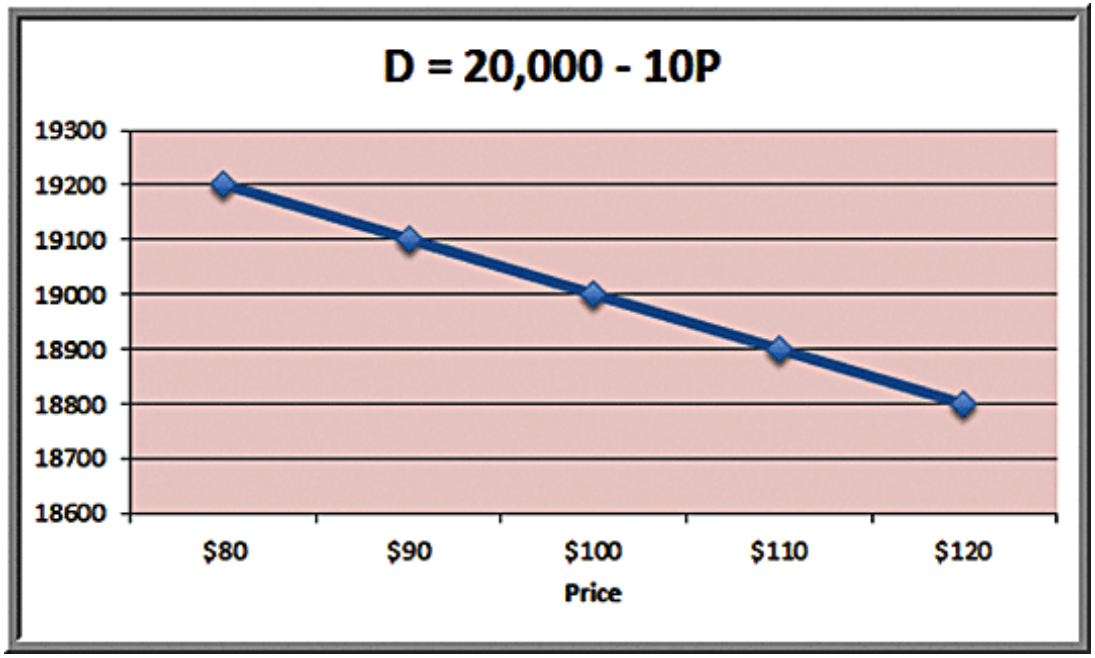
In this case: linear regression

# Model Assumptions

- Assumptions are made to
  - To **simplify** a model and make it more tractable; that is, able to be easily analyzed or solved.
  - To **add prior knowledge** about the relationship between variables.
- The task of the modeler is to select or build an appropriate model that best represents the behavior of the real situation.
- Example: economic theory tells us that demand for a product is negatively related to its price. Thus, as prices increase, demand falls, and vice versa.

# Example 1.9: A Linear Demand Prediction Model

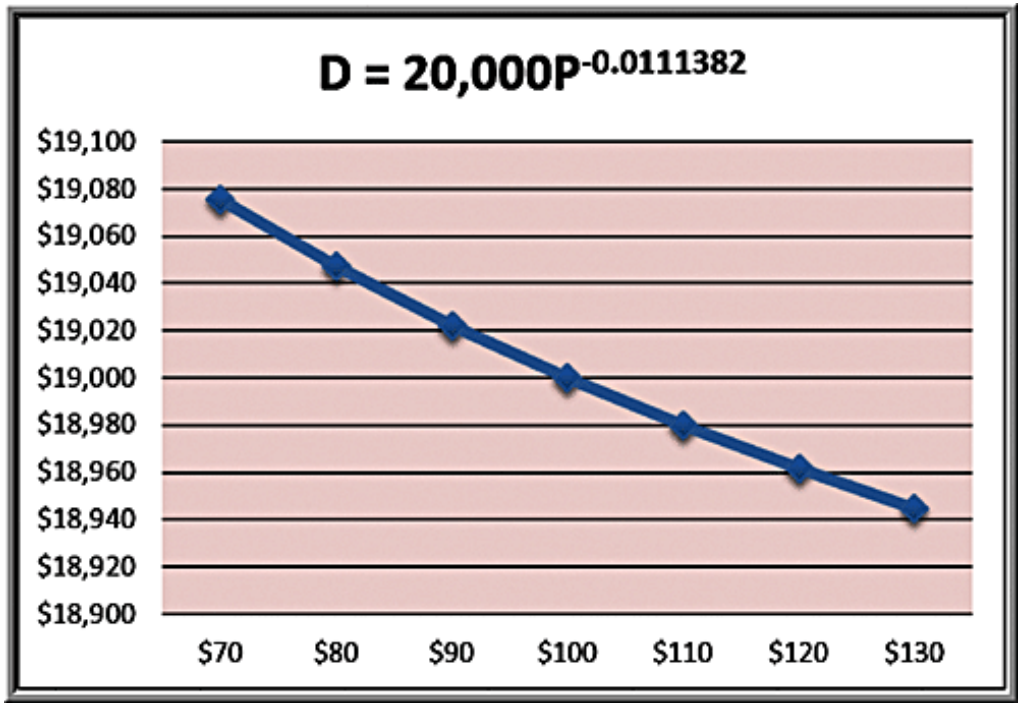
As price increases, demand falls.




**Issues:** Demand can become negative + empirical data has a poor fit.

# Example 1.10 A Nonlinear Demand Prediction Model

Assumes price elasticity is constant (constant ratio of % change in demand to % change in price)



# Uncertainty and Risk

- **Uncertainty** is **imperfect knowledge** (of what will happen in the future).
- **Risk** is the potential of (gaining or) losing something of value. It is the **consequence of actions** taken under uncertainty.

Often measured using standard deviation of variables. (=Deviation risk measure)

*"To try to eliminate risk in business enterprise is futile. Risk is inherent in the commitment of present resources to future expectations. Indeed, economic progress can be defined as the ability to take greater risks. The attempt to eliminate risks, even the attempt to minimize them, can only make them irrational and unbearable. It can only result in the greatest risk of all: rigidity."*

– Peter Drucker

# Prescriptive Decision Models

- **Prescriptive decision models** help decision makers identify the best solution.

- **Optimization** - finding values of decision variables that minimize (or maximize) something such as cost (or profit).
  - **Objective function** - the equation that minimizes (or maximizes) the quantity of interest.
  - **Constraints** - limitations or restrictions.
  - **Optimal solution** - values of the decision variables at the minimum (or maximum) point.

# Example 1.11: A Prescriptive Pricing Model

- A firm wishes to determine the best pricing for one of its products in order to maximize profit.

- Analysts determined the following **predictive model**:
  Sales = -2.9485(price) + 3240.9
  Total revenue = (price)(sales)
  Cost = 10(Sales) + 5000

- Identify the price that maximizes profit, subject to any constraints that might exist.

**max. Profit**
**s.t. Sales >= 0**
**Sales is integer**

# Contents

- Introduction to Analytics
- Tools
- Data
- Models
- **Problem solving with analytics**

# Problem Solving With Analytics

1. Recognize a problem
2. Define the problem
3. Structure the problem
4. Analyze the problem
5. Interpret results and make a decision
6. Implement the solution

Focus of the remainder of this course

# Recognize a Problem

Problems exist when there is a gap between what is happening and **what we think should be happening.**

- For example, costs are too high compared with competitors.

# Define the Problem

- Clearly defining the problem is not a trivial task.

- Complexity increases when the following occur:
  - large number of courses of action
  - the problem belongs to a group and not an individual
  - competing objectives
  - external groups are affected
  - problem owner and problem solver are not the same person
  - time limitations exist

- **What is part of the problem? What not?**

# Structure the Problem

- Stating **goals** and objectives
- Characterizing the possible decisions
- Identifying any **constraints** or restrictions

# Analyze the Problem

- Analytics plays a major role.

- Analysis involves some sort of experimentation or solution process, such as evaluating different scenarios, analyzing risks associated with various decision alternatives, finding a solution that meets certain goals, or **determining an optimal solution**.

# Interpret Results and Make a Decision

- What do the results found by the model mean for the application?

- Models cannot capture every detail of the real problem. Managers must understand the **limitations of models** and their underlying assumptions and often incorporate judgment into making a decision.

# Implement the Solution

- Translate the results of the model back to the real world.

- Requires providing adequate resources, motivating employees, eliminating resistance to change, modifying organizational policies, and developing trust.

# How to do an analytics project? CRISP-DM Reference Model

- **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
- De facto standard for conducting data mining and knowledge discovery projects.
- Defines tasks and outputs.
- Now developed by IBM as the Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM).
- SAS has SEMMA and most consulting companies use their own process.

Columbia University

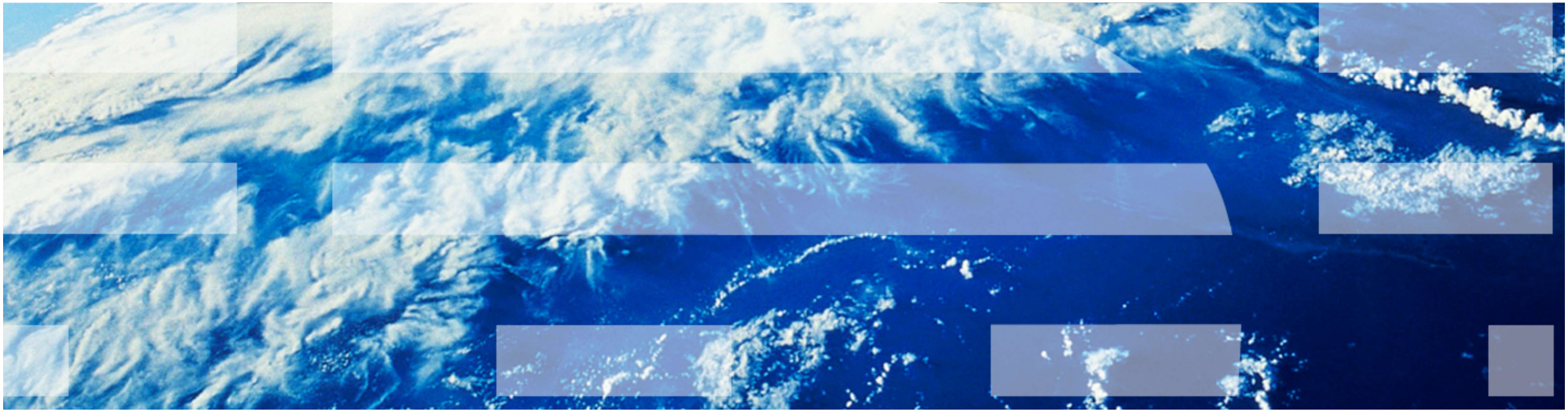
# EECS E6893 Big Data Analytics Lecture 1:

## *Overview of Big Data Analytics*

**Ching-Yung Lin**, Ph.D.

Adjunct Professor, Depts. of Electrical Engineering and Computer Science

IEEE Fellow



September 10th, 2021

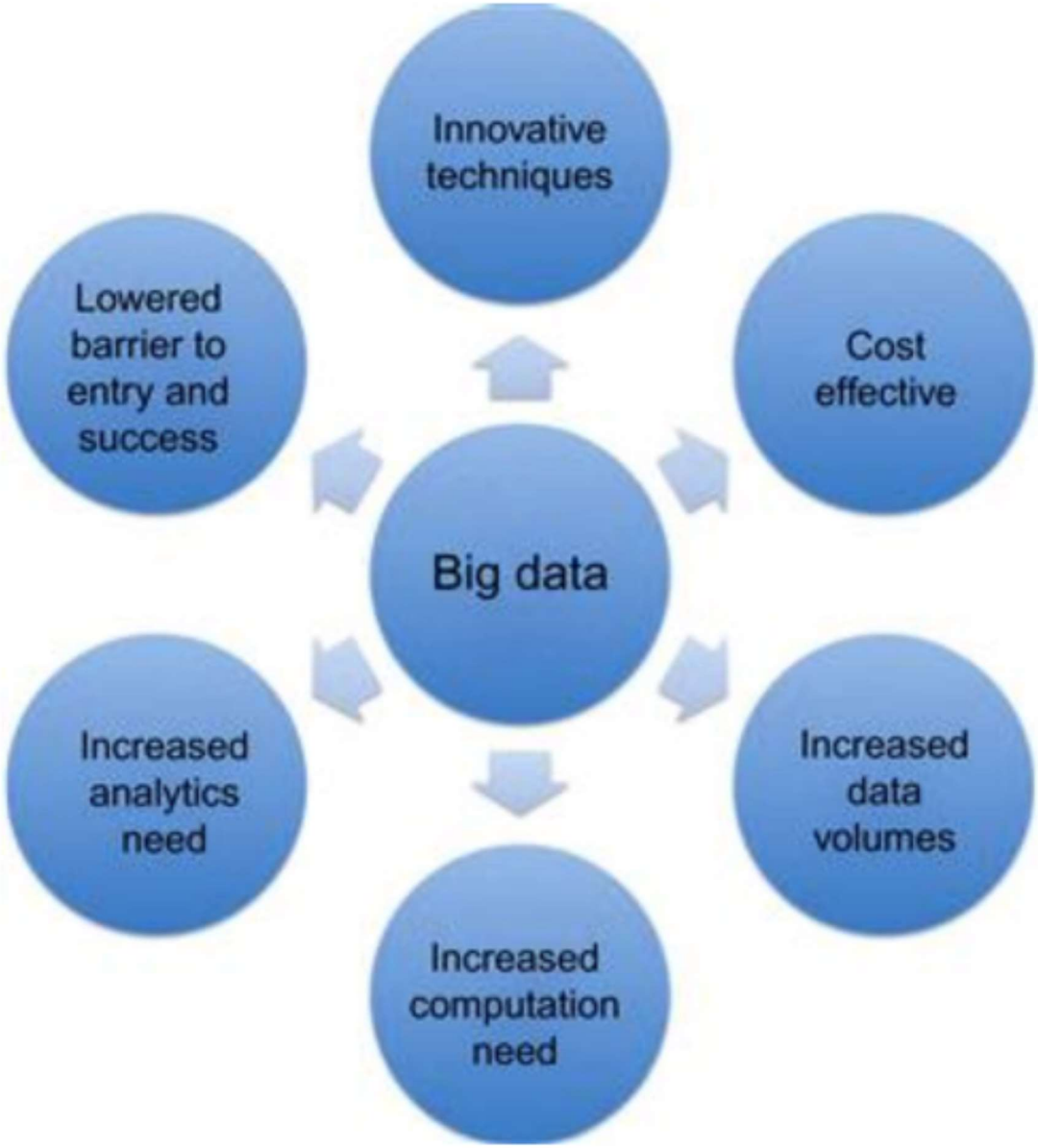
## Definition and Characteristics of Big Data

"*Big data is high-**volume**, high-**velocity** and high-**variety** information assets that demand **cost-effective**, **innovative** forms of information processing for **enhanced insight and decision making**.*" -- Gartner

which was derived from:

"*While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: **volumes, velocity and variety**. In 2001/02, IT organizations much compile a variety of approaches to have at their disposal for dealing each.*" – Doug Laney

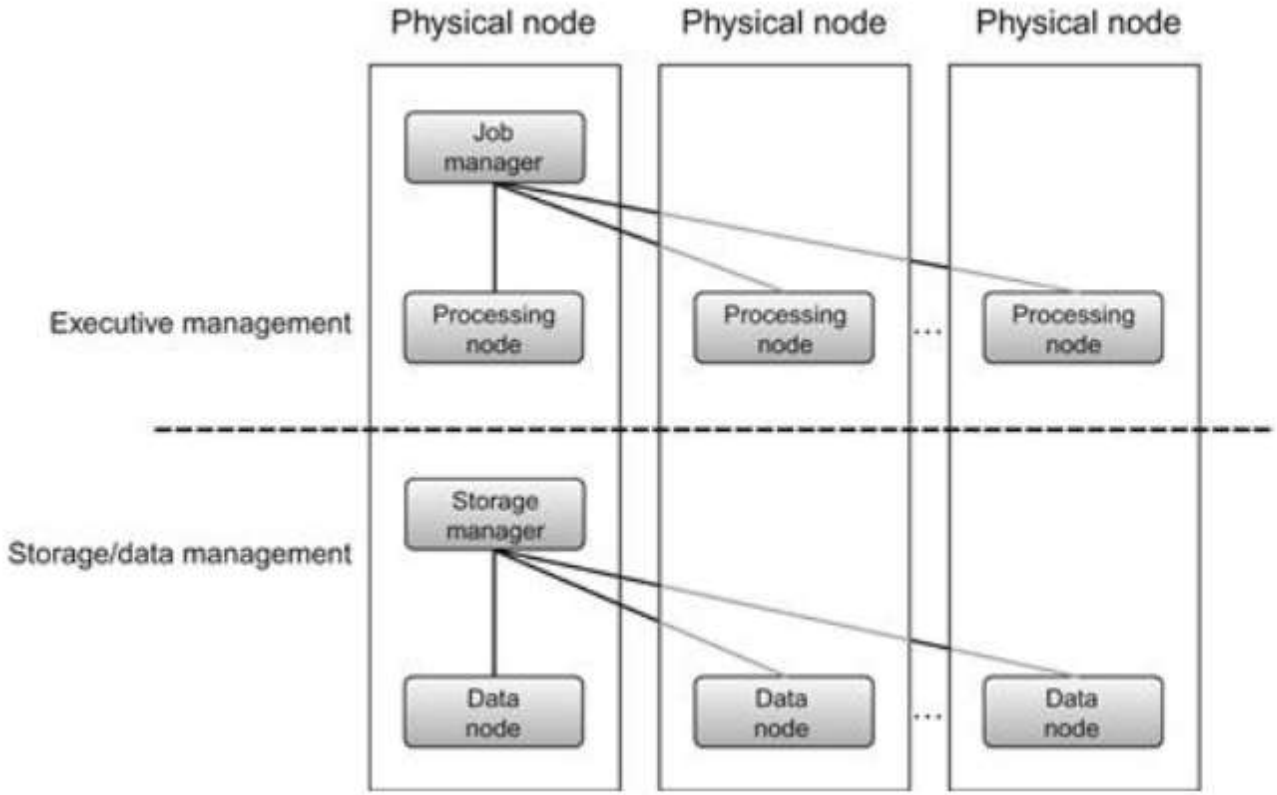
# What made Big Data needed?




"Big Data Analytics", David Loshin

# Key Computing Resources for Big Data

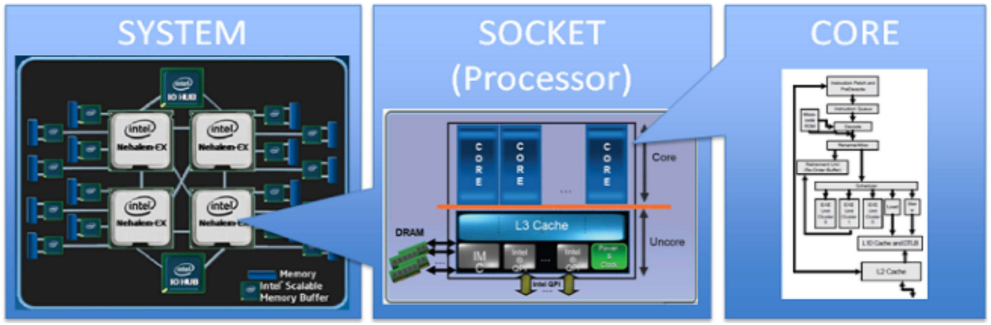
- Processing capability:  CPU, processor, or node.
- Memory
- Storage
- Network

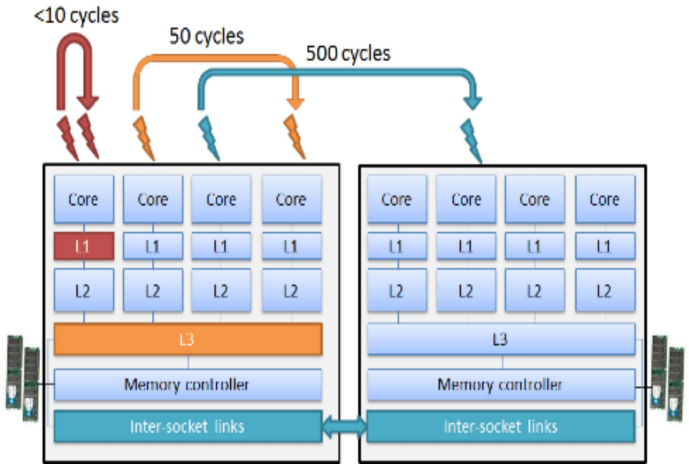



"Big Data Analytics", David Loshin

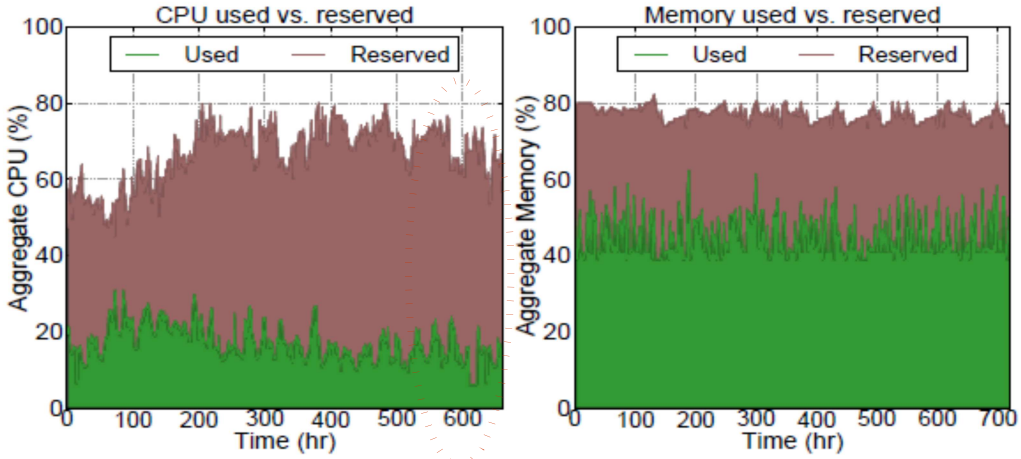
# Scalability — Scale Up & Scale Out

- Scale out
  - Use more resources to distribute workload in parallel
  - Higher data access latency is typically incurred
- Scale up
  - Efficiently use the resources
  - Architecture-aware algorithm design




Example: Resource utilization for a large production cluster at Twitter data center




www.stanford.edu/~cdel/2014.asplos.quasar.pdf

- For independent data ==> scale up may not have obvious advantage than scale out
- For linked data ==> utilizing scale up as much as possible before scale out

# Contrasting Approaches in Adopting High-Performance Capabilities

| Aspect | Typical Scenario | Big Data |
|---|---|---|
| Application development | Applications that take advantage of massive parallelism developed by specialized developers skilled in high-performance computing, performance optimization, and code tuning | A simplified application execution model encompassing a distributed file system, application programming model, distributed database, and program scheduling is packaged within Hadoop, an open source framework for reliable, scalable, distributed, and parallel computing |
| Platform | Uses high-cost massively parallel processing (MPP) computers, utilizing high-bandwidth networks, and massive I/O devices | Innovative methods of creating scalable and yet elastic virtualized platforms take advantage of clusters of commodity hardware components (either cycle harvesting from local resources or through cloud-based utility computing services) coupled with open source tools and technology |
| Data management | Limited to file-based or relational database management systems (RDBMS) using standard row-oriented data layouts | Alternate models for data management (often referred to as NoSQL or "Not Only SQL") provide a variety of methods for managing information to best suit specific business process needs, such as in-memory data management (for rapid access), columnar layouts to speed query response, and graph databases (for social network analytics) |
| Resources | Requires large capital investment in purchasing high-end hardware to be installed and managed in-house | The ability to deploy systems like Hadoop on virtualized platforms allows small and medium businesses to utilize cloud-based environments that, from both a cost accounting and a practical perspective, are much friendlier to the bottom line |

"Big Data Analytics", David Loshin

# Techniques towards Big Data

- Massive Parallelism
- Huge Data Volumes Storage
- Data Distribution
- High-Speed Networks
- High-Performance Computing
- Task and Thread Management
- Data Mining and Analytics
- Data Retrieval
- Machine Learning
- Data Visualization

➔ Techniques exist for years to decades. Why is Big Data hot now?

# Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

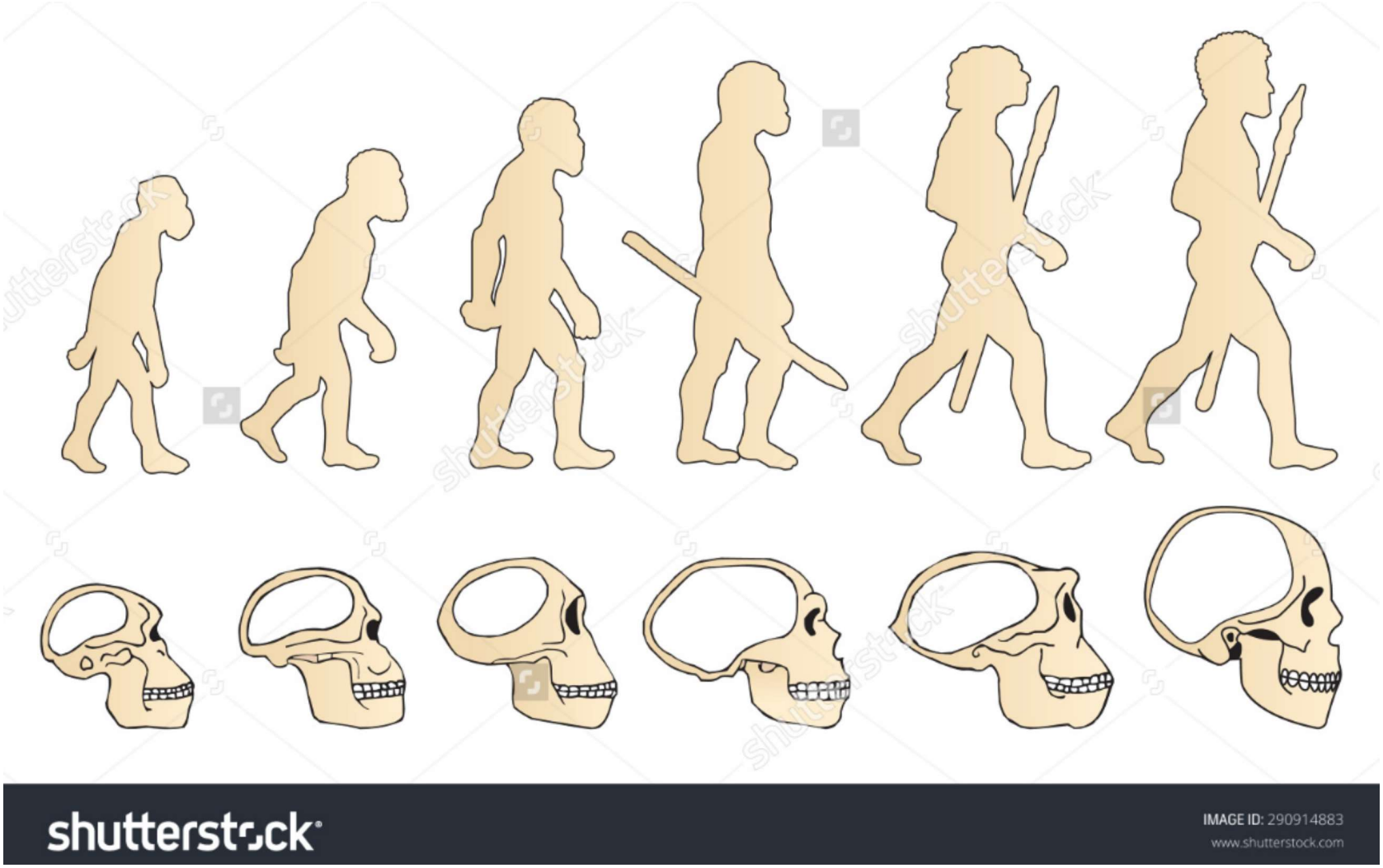
# Why Big Data now?

- More data are being collected and stored
- Open source code
- Commodity hardware / Cloud

→
- High-Volume
- High-Velocity
- High-Variety

→ Artificial Intelligence

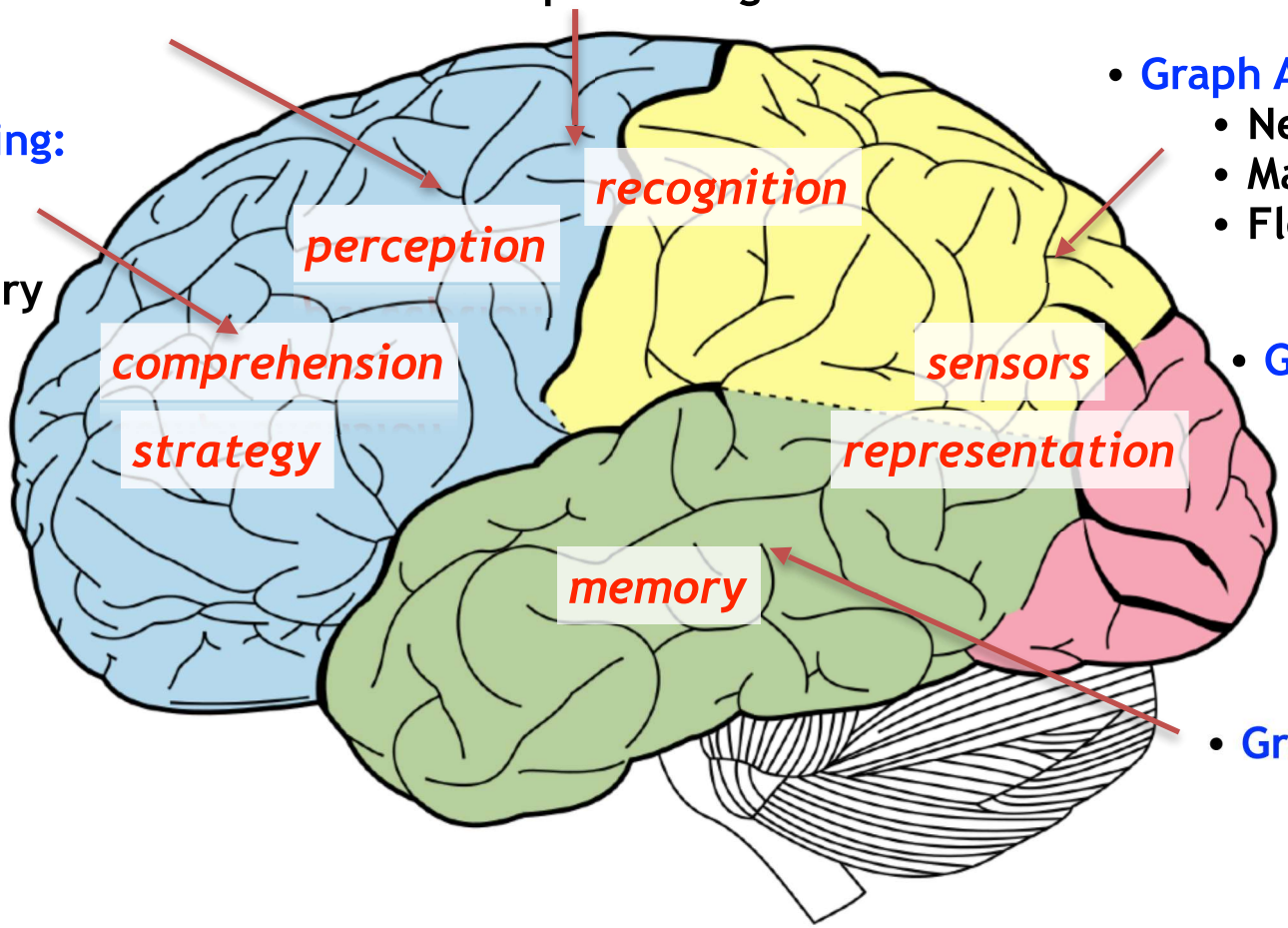

https://www.youtube.com/watch?v=BV8qFeZxZPE



# Human brain is a graph/network of 100B nodes and 700T edges.

- **Machine Cognition:**
  - Robot Cognition Tools
  - Feeling
- **Machine Reasoning:**
  - Bayesian Networks
  - Game Theory Tools

- **Machine Learning:**
  - Machine Learning Tools
  - Deep Learning Tools




- **Graph Analytics:**
  - Network Analysis
  - Matching and Search
  - Flow Prediction
- **Graph Visualization:**
  - Dynamic Graph
  - Big Graph
- **Graph Database:**
  - Large-Scale Native Store

# Course Main Thrust 1: Apache Hadoop and Big Data




The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

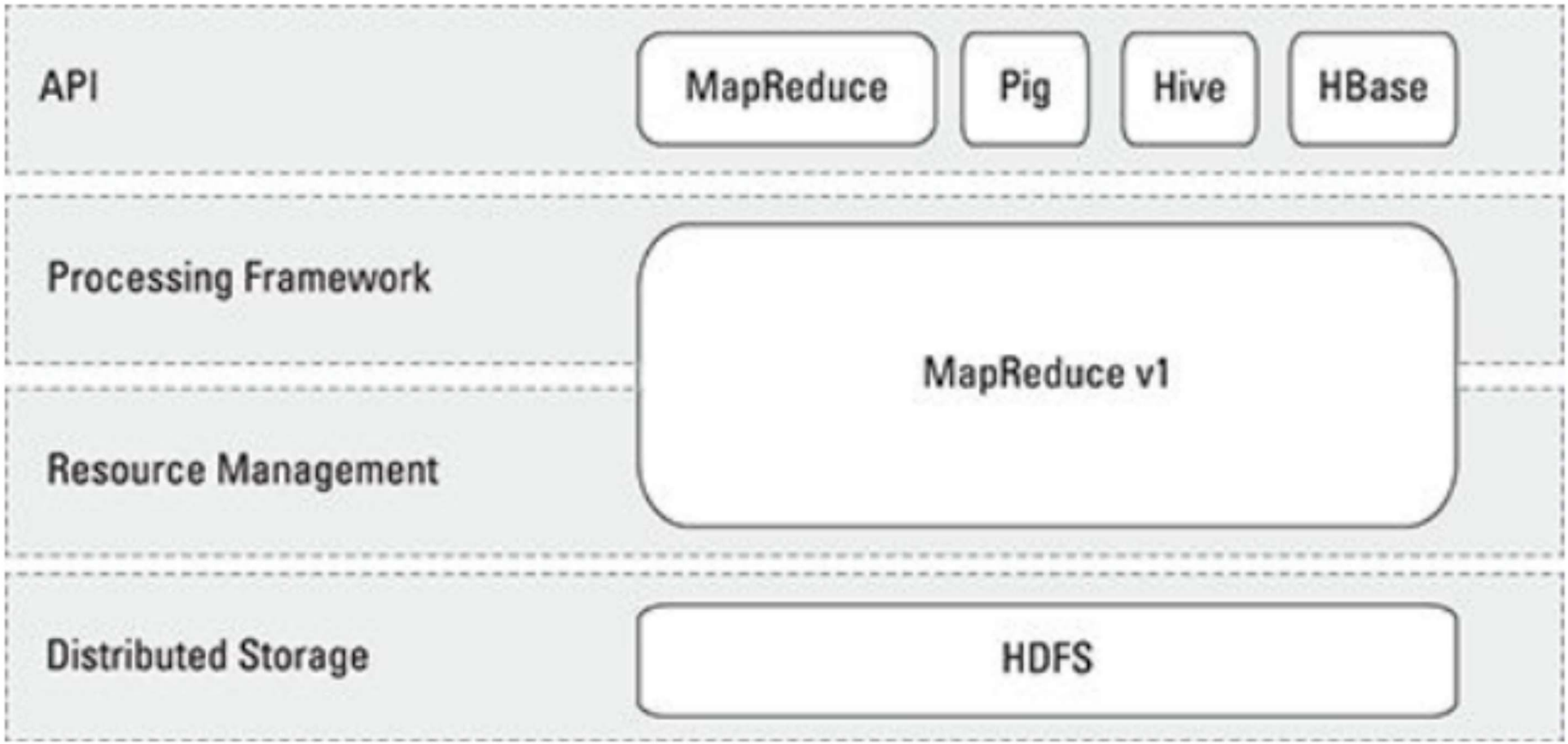
The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

http://hadoop.apache.org

# Four distinctive layers of Hadoop



# Course Main Thrust 2: Apache Spark and ML

Columbia University

Apache Spark™ Lightning-fast unified analytics engine

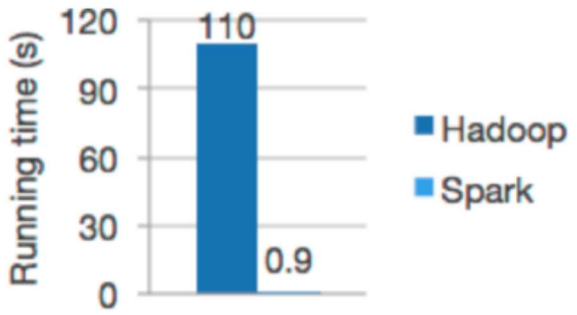
Download | Libraries | Documentation | Examples | Community | Developers

**Apache Spark™** is a unified analytics engine for large-scale data processing.

## Speed

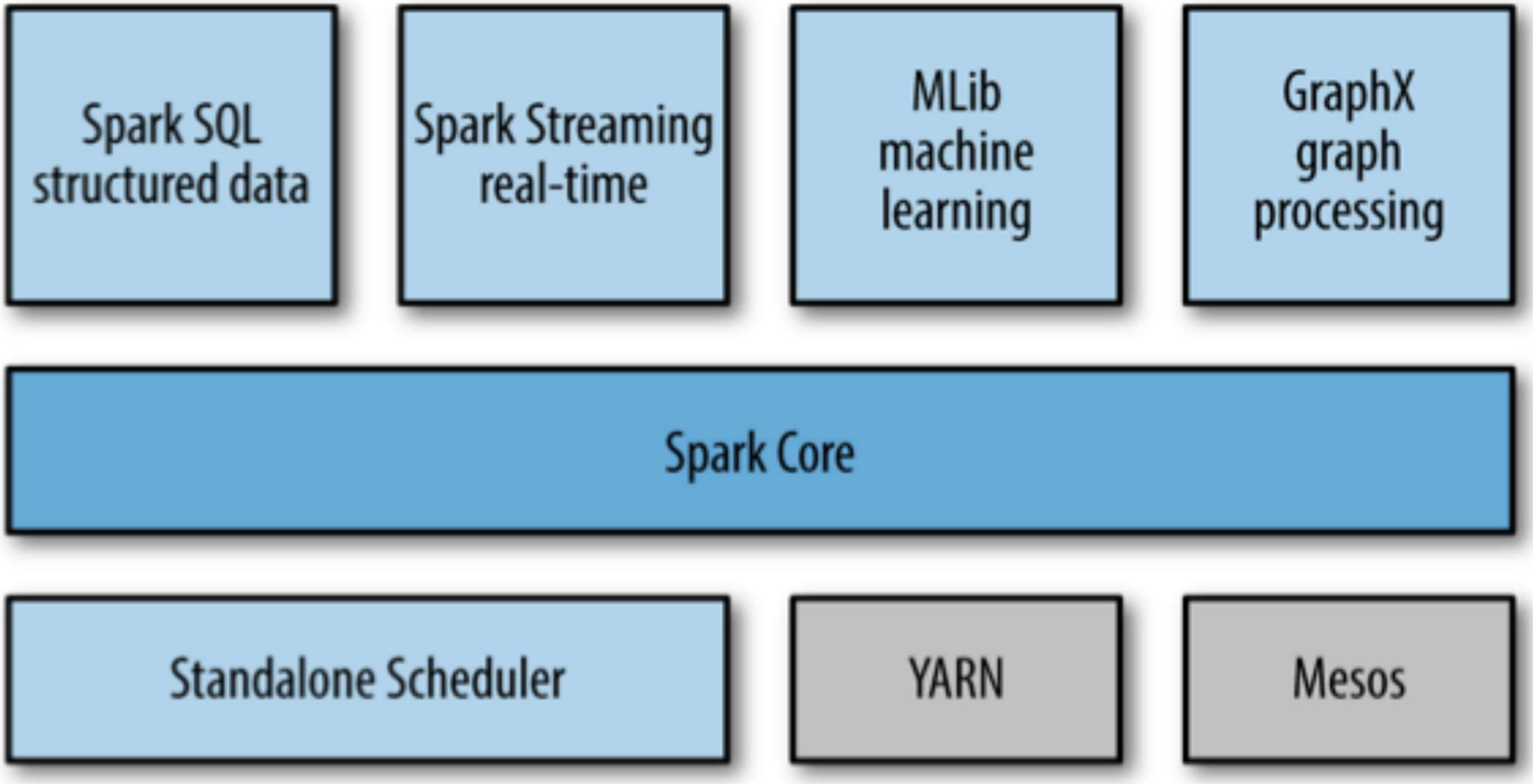
Run workloads 100x faster.

Apache Spark achieves high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.



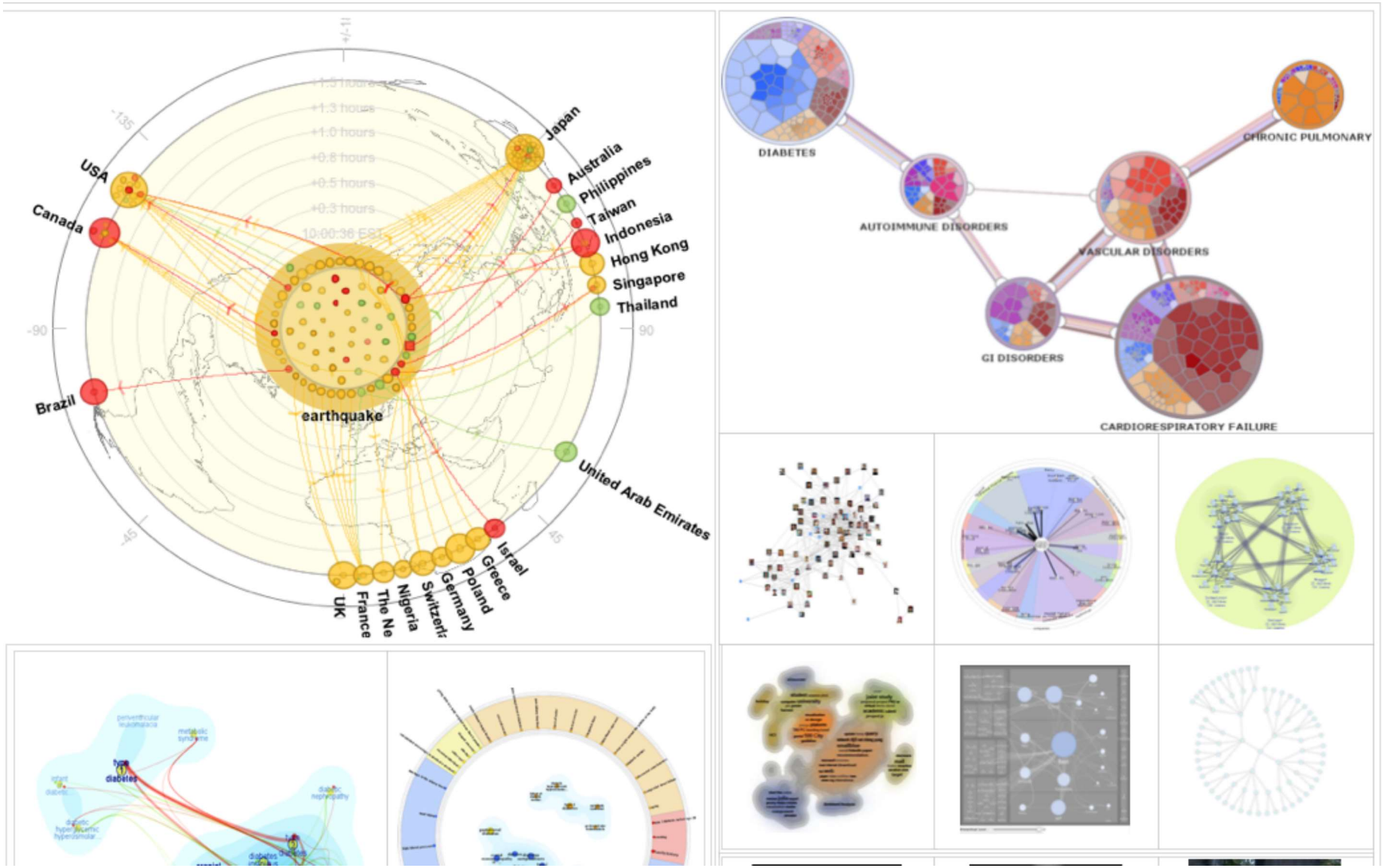

Logistic regression in Hadoop and Spark

# Main Spark Stack

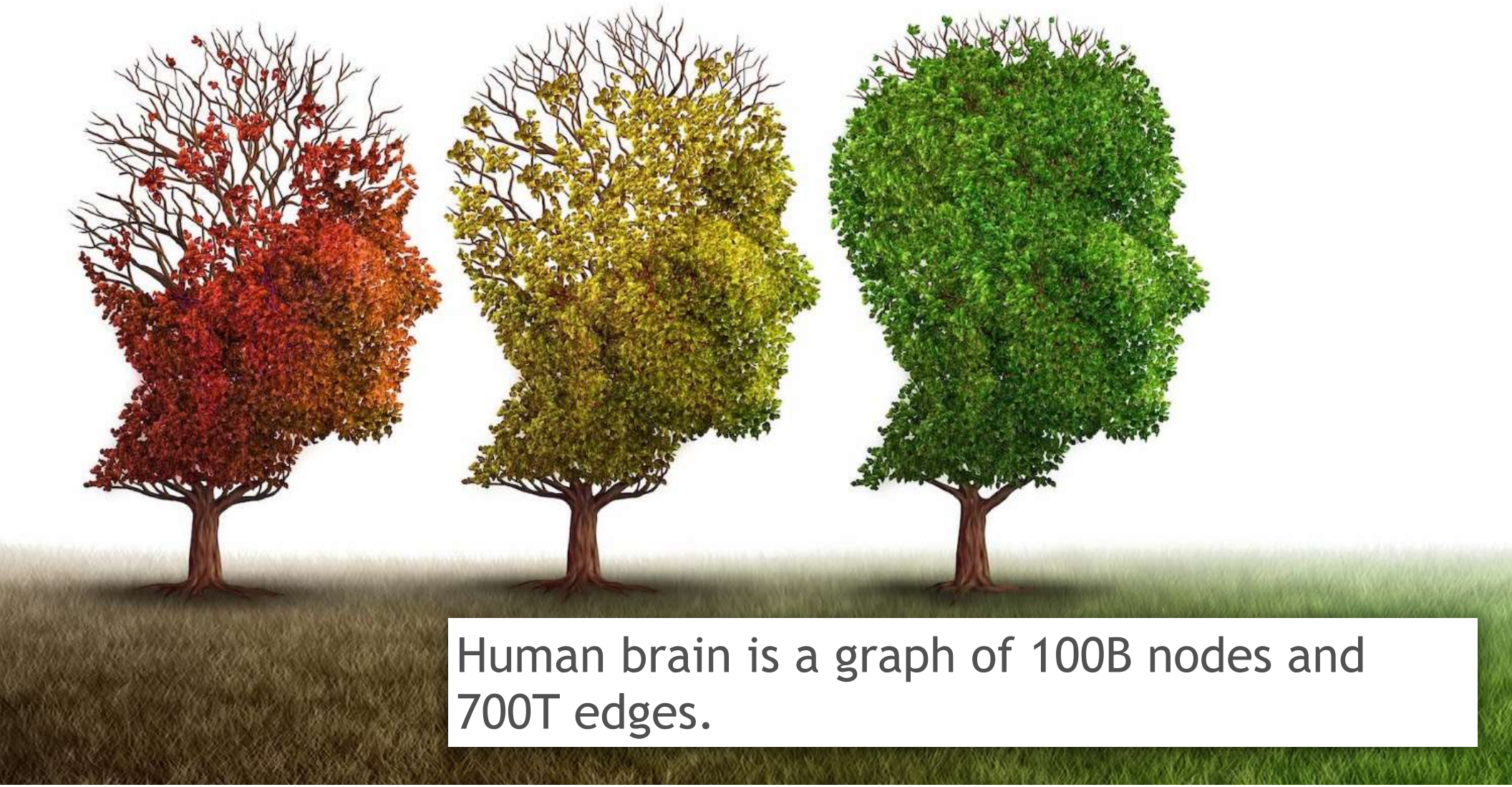


# Course Main Thrust 3: Streaming Big Data Analytics

# Course Main Thrust 4: Big Data Visualization



# Course Main Thrust 5: Linked Big Data Analysis

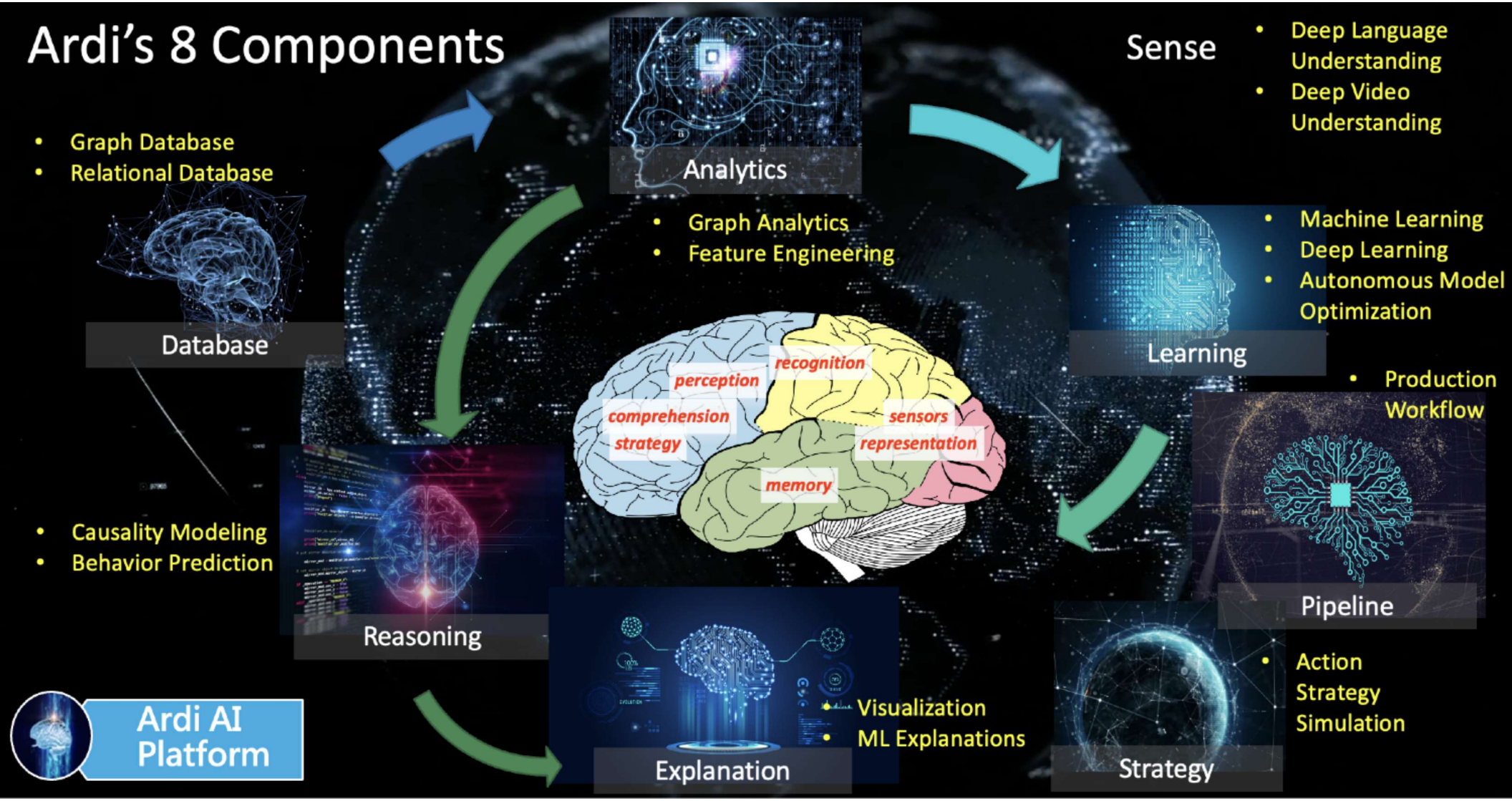


Human brain is a graph of 100B nodes and 700T edges.

# Course Main Thrust 6: Big Data System and AI Solutions

- **Big Data Pipeline**
- **Big Data and AI for Finance**
- **Big Data and AI for Healthcare**

# Big Data AI Platform Example: Graphen Ardi



# Why you want to take this class

- **Key Differentiator of this class:** Focusing on building a full-spectrum understanding of the latest Big Data Analytics technologies and using them to build real industry real-world solutions.

- **Sapphire Big Data Analytics Open Source Applications:** Create a Big Data open source toolsets for various industries (and disciplines)


- **Dataset and Use Cases:** Welcome!!

# Course Grading

- **5 Homeworks: 50%**

  -- **Individual work**; Language Requirement: Python, JavaScript; Get familiar with Linux

  -- **Report (including description of the work, discussions, experiments, etc) and source code**

  - **HW #0: Big Data Environment Setup and Testing**
  - **HW #1: Big Data Analytics and Machine Learning**
  - **HW #2: Streaming Big Data Analytics**
  - **HW #3: Big Data Analytics Visualization**
  - **HW #4: Linked Big Data Analytics**

# Course Grading

- **Final Project: 50%**

  **-- Teamwork: 2 - 3 students per team (on campus); 1 - 3 students per team for CVN**

  - **Proposal** (slides — short presentation in the class)
  - **Progress Presentation** (slides — short presentation in the class)
  - **Progress Report** (report)
  - **Final Report** (paper, up to 10 pages)
  - **Workshop Presentation** (Oral and Demo)
  - **Open Source Codes**
  - **Video Presentation** (on YouTube)

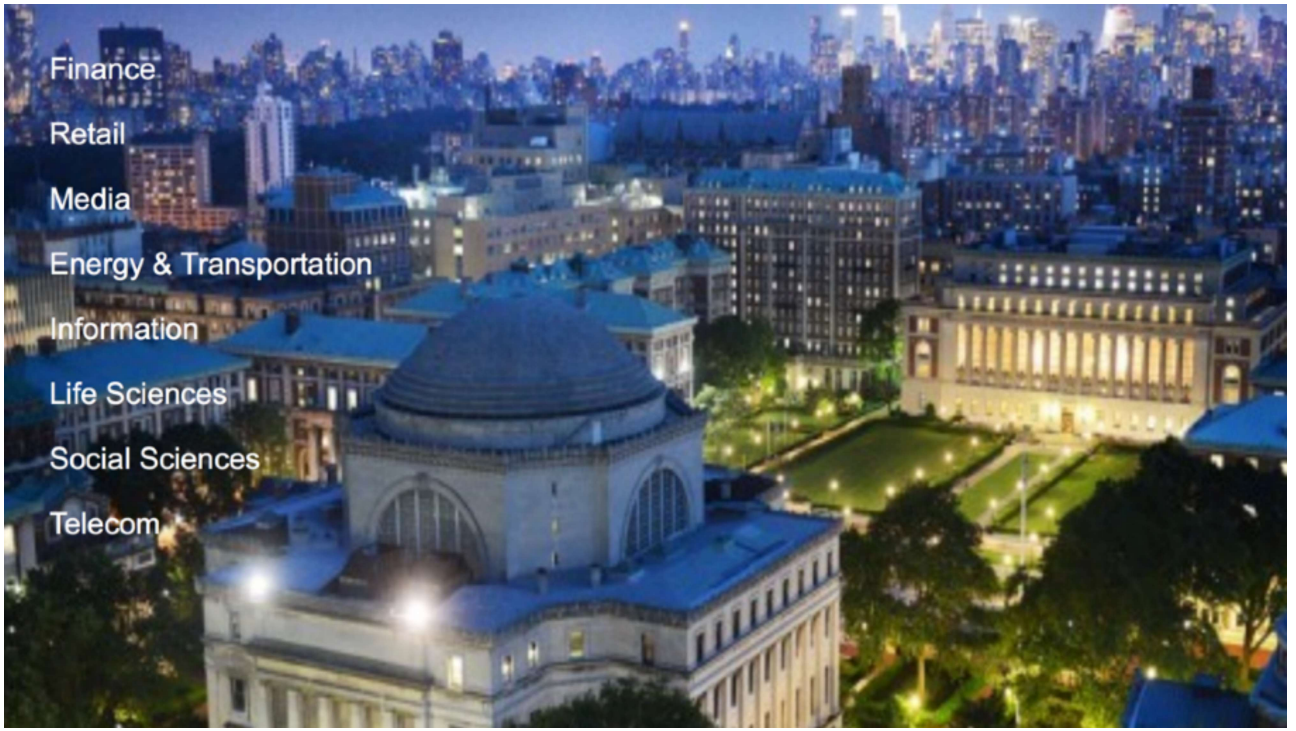
# Course Information

- Website:

  http://www.ee.columbia.edu/~cylin/course/bigdata/

- Textbook:

  -- None, but reference book(s) and/or articles/papers will be provided each lecture.



# Course Outline

| Class Date | Class Number | Topics Covered |
|---|---|---|
| 09/10/21 | 1 | Introduction of Big Data Analytics |
| 09/17/21 | 2 | Big Data Platforms and Data Storage |
| 09/24/21 | 3 | Big Data Analytics Algorithms I |
| 10/01/21 | 4 | Big Data Analytics Algorithms II |
| 10/08/21 | 5 | Real-Time Stream Analysis |
| 10/15/21 | 6 | Big Data Visualization I |
| 10/22/21 | 7 | Big Data Visualization II |
| 10/29/21 | 8 | Linked Big Data Analysis and Graph Computing I |
| 11/05/21 | 9 | Final Project Proposal Presentation |
| 11/12/21 | 10 | Linked Big Data Analysis and Graph Computing II |
| 11/19/21 | 11 | Final Project Progress Presentation |
| 11/26/21 | | Thanksgiving Holiday |
| 12/03/21 | 12 | Big Data Analytics Applications -- AI Finance |
| 12/10/21 | 13 | Big Data Analytics Applications -- AI Medical |
| 12/17/21 | 14 | Big Data Analytics Workshop |

# Assignments and Submissions

| Class Date | Assignment | Due |
|---|---|---|
| 09/10/21 | HW #0 Big Data Environment Setup and Testing [tutorial] | |
| 09/17/21 | | |
| 09/24/21 | HW #1 Big Data Analytics and Machine Learning [assignment][tutorial] | HW #0 |
| 10/01/21 | | |
| 10/08/21 | HW #2 Streaming Big Data Analytics [assignment] [tutorial] | HW #1 |
| 10/15/21 | | |
| 10/22/21 | HW #3 Big Data Visualization [assignment][tutorial] | HW #2 |
| 10/29/21 | | |
| 11/05/21 | | HW#3 & Proposal Slides |
| 11/12/21 | HW #4 Linked Big Data Analytics [assignment][tutorial] | |
| 11/19/21 | | Progress Slides |
| 11/26/21 | | |
| 12/03/21 | | HW #4 & Progress Report |
| 12/10/21 | | |
| 12/17/21 | | Final Project Slides and Other Materials |

# Other Issues

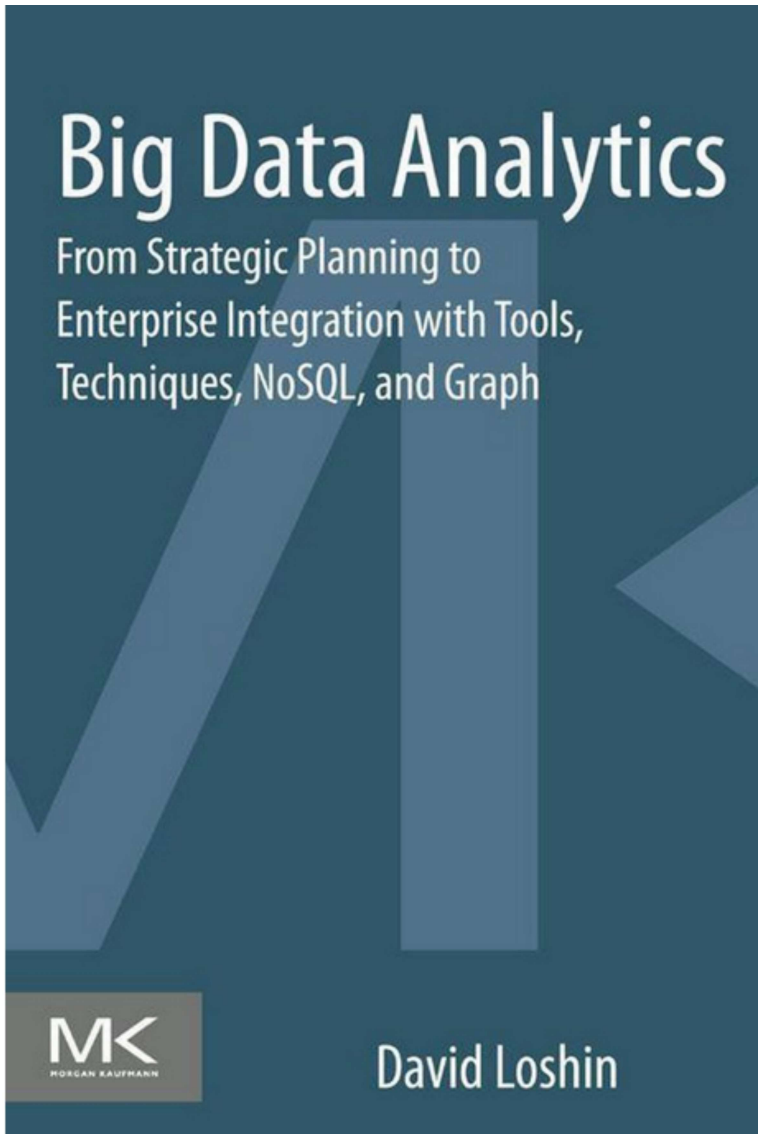
- Professor Lin:
  - Office Hours:

    After the class or by appointment

  - Contact: c.lin@columbia.edu

- TA (CA/IA/Grader) —
  - Cong Han (ch3212): Tue 4-6pm
  - Yvonne Lee (yl4573) : Wed 4-6 pm
  - Guoshiwen Han (gh2567): Mon 9-11am
  - Yiwen Fang (yf2560): Thu 5:30-7:30pm (may change to 3-5pm; please see the course website)

# Reading Reference for Lecture 1




Chapter 1: Market and Business Drivers for Big Data Analysis

Chapter 2: Business Problems Suited to Big Data Analytics

Chapter 3: Achieving Organizational Alignment for Big Data Analytics

Chapter 4: Developing a Strategy for Integrating Big Data Analytics into the Enterprise

Chapter 5: Data Governance for Big Data Analytics: Considerations for Data Policies and Processes

Chapter 6: Introduction to High-Performance Appliances for Big Data Management

Chapter 7: Big Data Tools and Techniques

Chapter 8: Developing Big Data Applications

Chapter 9: NoSQL Data Management for Big Data

Chapter 10: Using Graph Analytics for Big Data

Chapter 11: Developing the Big Data Roadmap

# 5 Example Big Data Use Case Categories

**Big Data Exploration**
Find, visualize, understand all big data to improve decision making

**Enhanced 360º View of the Customer**
Extend existing customer views (MDM, CRM, etc) by incorporating additional internal and external information sources

**Security/Intelligence Extension**
Lower risk, detect fraud and monitor cyber security in real-time

**Operations Analysis**
Analyze a variety of machine data for improved business results

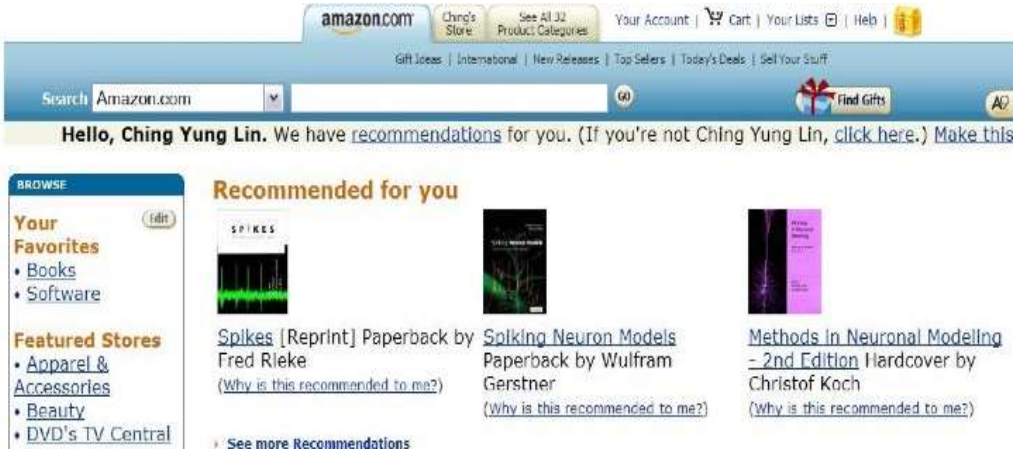
**Data Warehouse Augmentation**
Integrate big data and data warehouse capabilities to increase operational efficiency

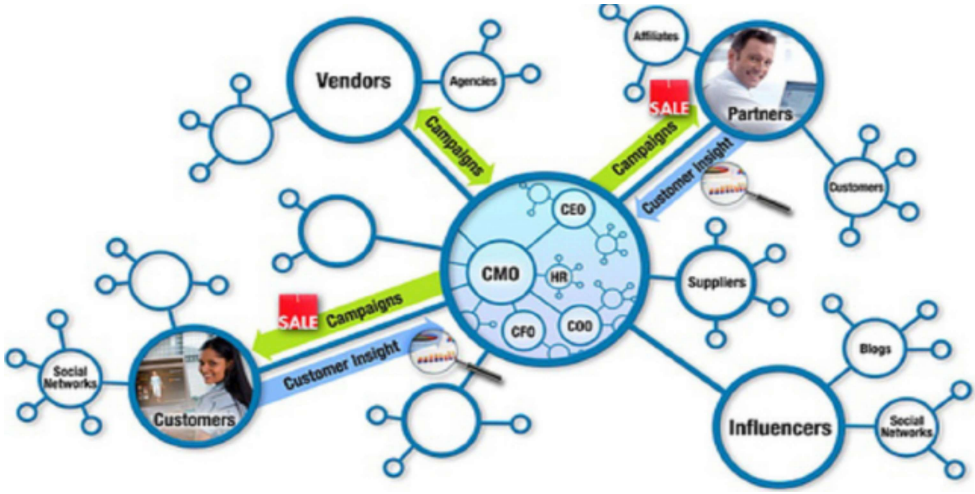
# Big Data Examples -- Application Use Cases

1. Expertise Location
2. Recommendation
3. Commerce
4. Financial Analysis
5. Social Media Monitoring
6. Telco Customer Analysis
7. Healthcare Analysis
8. Data Exploration and Visualization
9. Personalized Search
10. Anomaly Detection
11. Fraud Detection
12. Cybersecurity
13. Sensor Monitoring (Smarter another Planet)
14. Cellular Network Monitoring
15. Cloud Monitoring
16. Code Life Cycle Management
17. Traffic Navigation
18. Image and Video Semantic Understanding
19. Genomic Medicine
20. Brain Network Analysis
21. Data Curation
22. Near Earth Object Analysis

# Category 1: 360° View

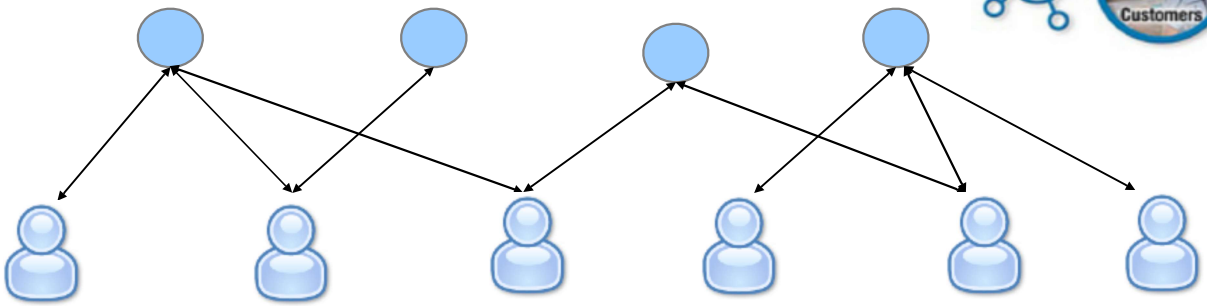
*Recommendation*




item

user



*Enhancing:*

Coremetrics An IBM Company

unica

Graph Visualizations

Communities | Graph Search | Network Info Flow | Bayesian Networks

Centralities | Graph Query | Shortest Paths | Latent Net Inference

Ego Net Features | Graph Matching | Graph Sampling | Markov Networks

Middleware and Database

# Use Case 1: Social Network Analysis in Enterprise for Productivity

Columbia University

**Production Live System used by IBM GBS since 2009 – verified ~$100M contribution**

15,000 contributors in 76 countries; 92,000 annual unique IBM users

25,000,000+ emails & SameTime messages (incl. Content features)

1,000,000+ Learning clicks; 14M KnowledgeView, SalesOne, …, access d

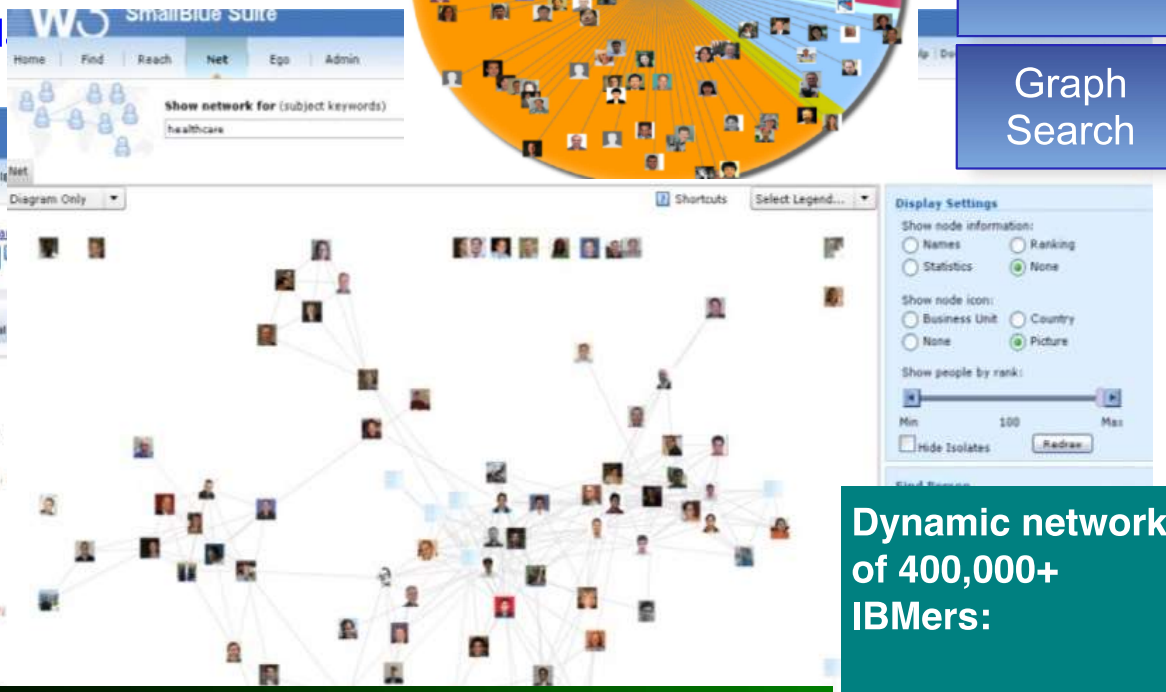
1,000,000+ Lotus Connections (blogs, file sharing, bookmark) data

200,000 people's consulting project & earning d

Shortest Paths

Centralities

Graph Search

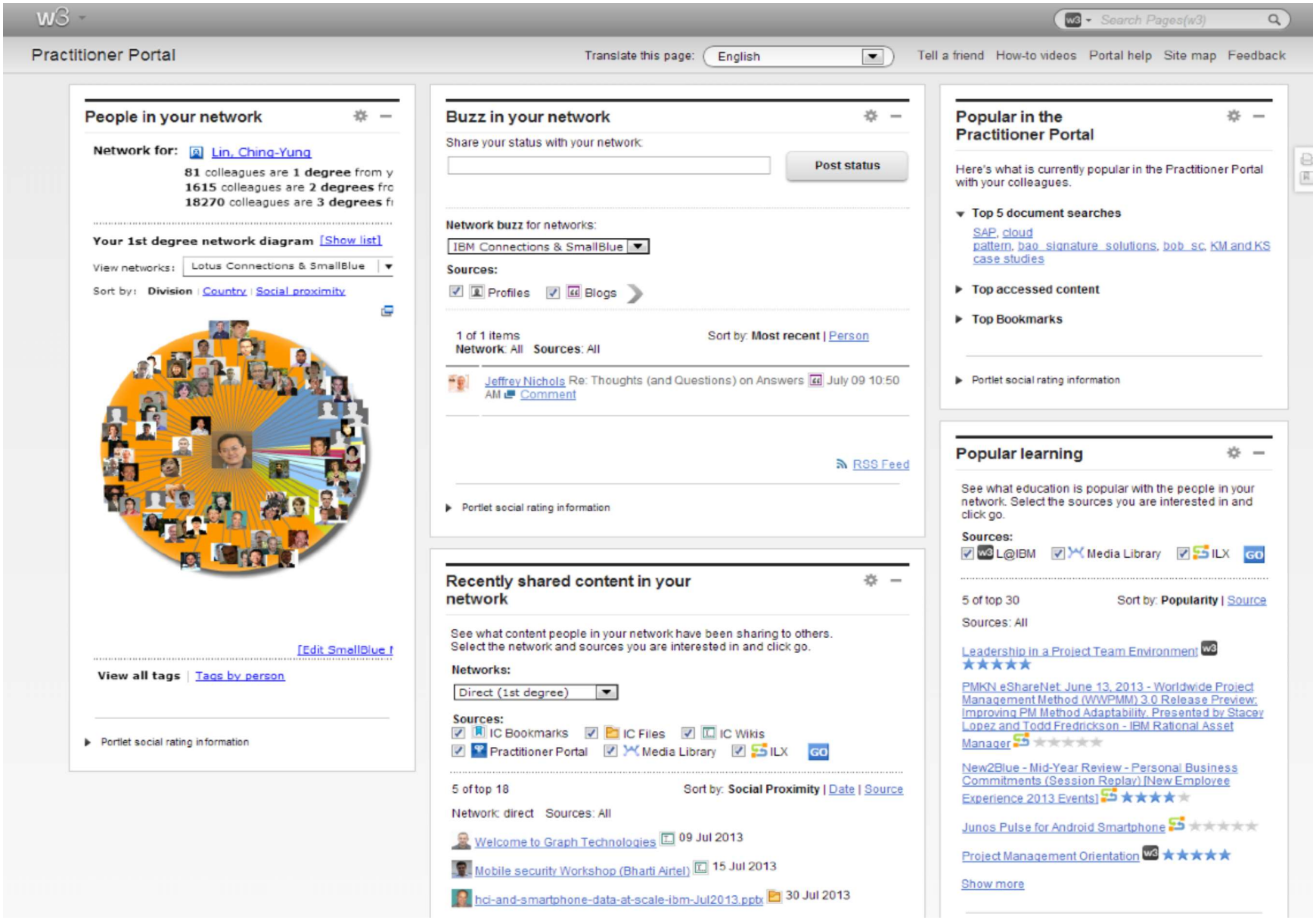




- On BusinessWeek four times, including being the Top Story of Week, April 2009
- Help IBM earned the 2012 Most Admired Knowledge Enterprise Award
- Wharton School study: $7,010 gain per user per year using the tool
- In 2012, contributing about 1/3 of GBS Practitioner Portal $228.5 million savings an
- APQC (WW leader in Knowledge Practice) April 2013:
  *"The Industry Leader and Best Practice in Expertise Location"*

**Dynamic networks of 400,000+ IBMers:**

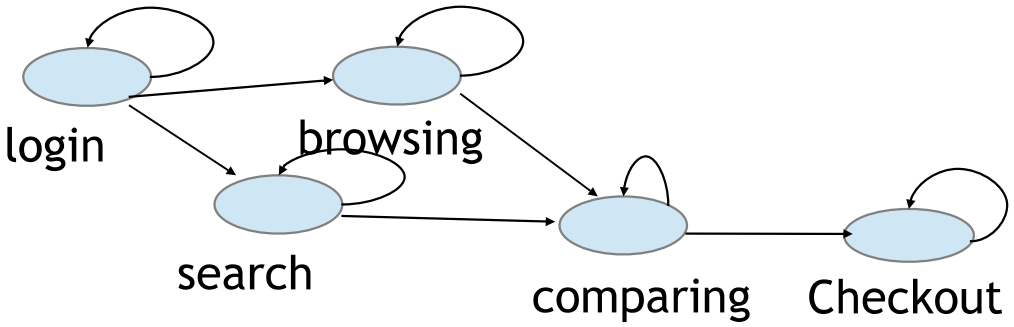
Shortest Paths
Social Capital
Bridges
Hubs
Expertise Search
Graph Search
Graph Recomm.

# Use Case 2: Personalized Recommendation



# Use Case 3: Customer Behavior Sequence Analytics

Markov Network | Latent Network | Bayesian Network

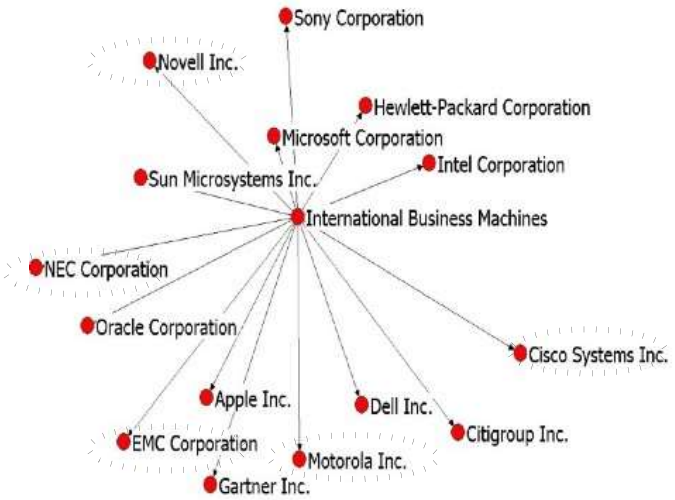




- Behavior Pattern Detection
- Help Needed Detection

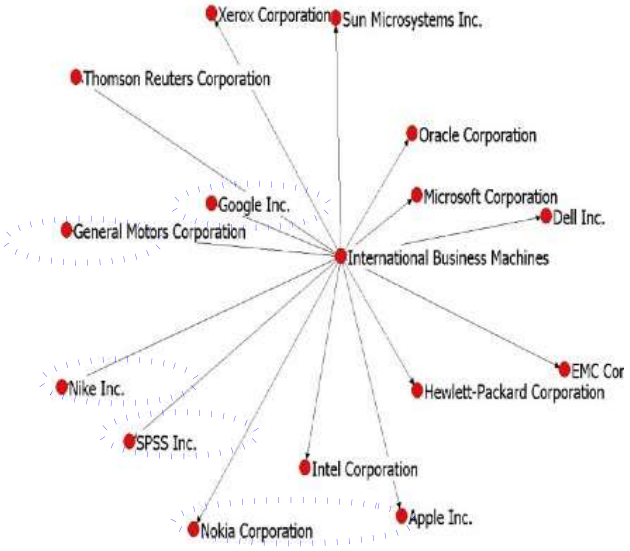
# Use Case 4: Graph Analytics for Financial Analysis

**Goal:** *Injecting Network Graph Effects for Financial Analysis. Estimating company performance considering correlated companies, network properties and evolutions, causal parameter analysis, etc.*

- IBM 2003




- IBM 2009

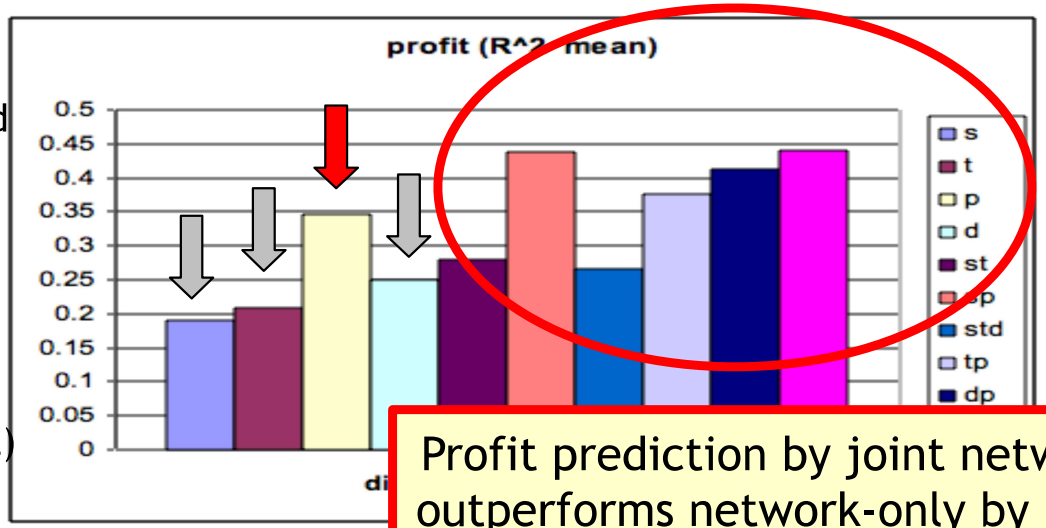



- Data Source:
  - Relationships among 7594 companies, data mining from NYT 1981 ~ 2009

Targets: 20 Fortune companies' normalized Profits

Goal: Learn from previous 5 years, and predict next year

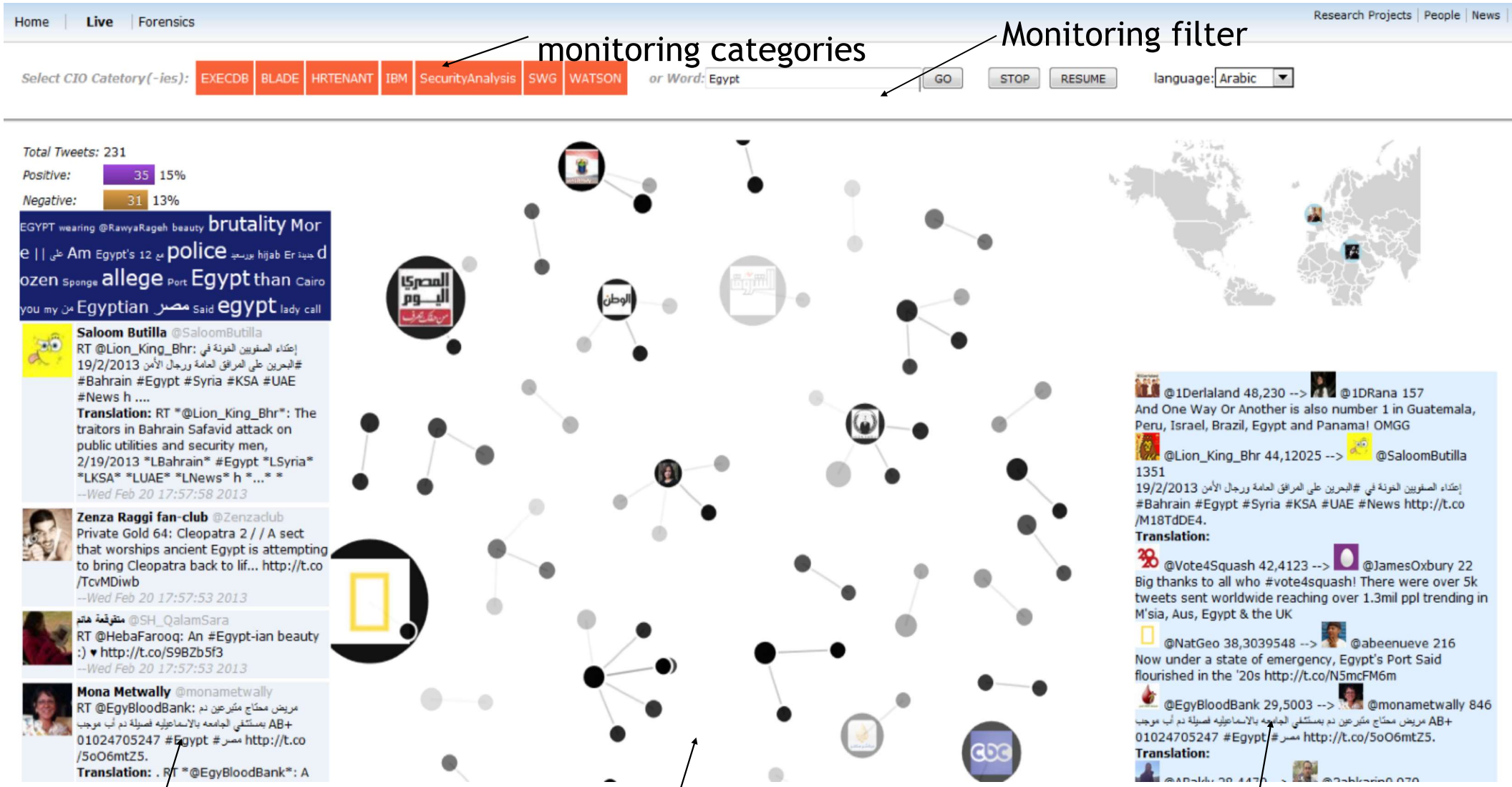
Model: Support Vector Regression (RBF kernel)




**Network feature:**
s (current year network feature),
t (temporal network feature),
d (delta value of network feature)
**Financial feature:**
p (historical profits and

Profit prediction by joint network and financial analysis outperforms network-only by 130% and financial-only by 33%.

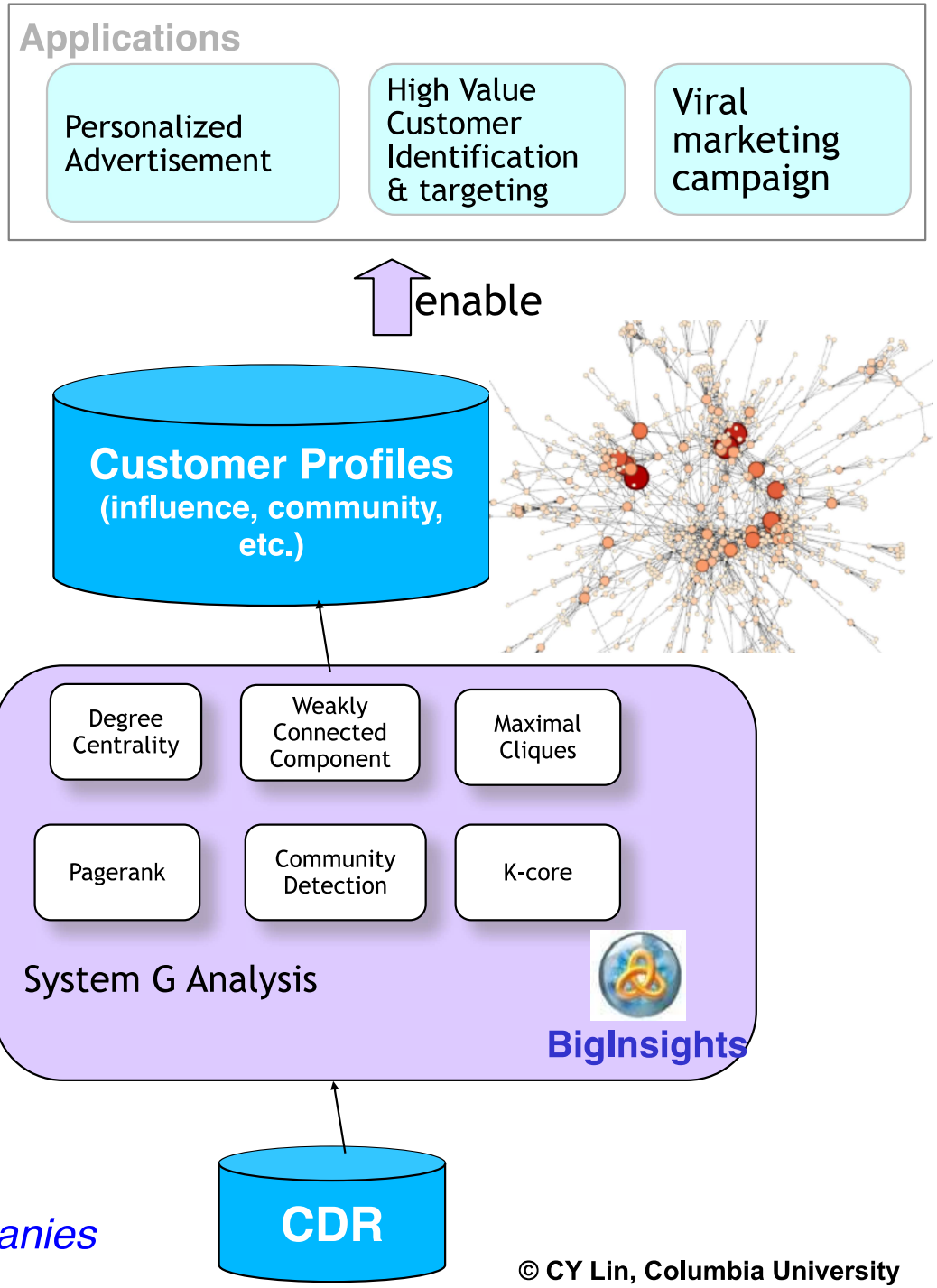
# Use Case 5: Social Media Monitoring



# Use Case 6: Customer Social Analysis for Telco

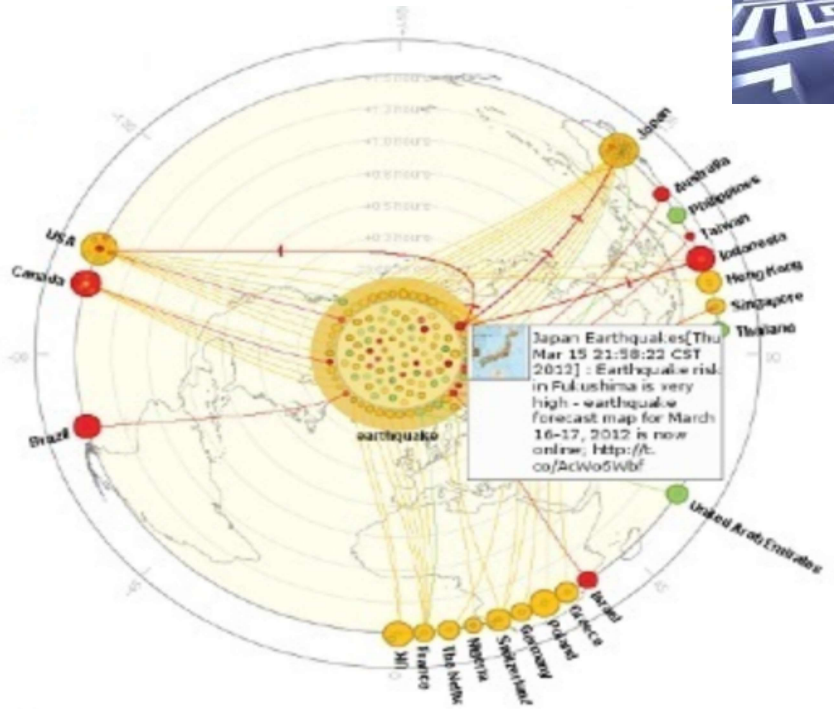
**Goal:** Extract customer social network behaviors to enable Call Detail Records (CDRs) data monetization for Telco.

- Applications based on the extracted social profiles
  - Personalized advertisement (beyond the scope of traditional campaign in Telco)
  - High value customer identification and targeting
  - Viral marketing campaign
- Approach
  - Construct social graphs from CDRs based on {caller, callee, call time, call duration}
  - Extract customer social features (e.g. influence, communities, etc.) from the constructed social graph as customer social profiles
  - Build analytics applications (e.g. personalized advertisement) based on the extracted customer social profiles

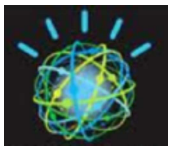



*PoCs with Chinese and Indian Telecomm companies*

# Category 2: Data Exploration


*Enhancing:*




Vivisimo

CÚRAM SOFTWARE

Huge Network Visualization | Network Propagation | I2 3D Network Visualization | Geo Network Visualization | Graphical Model

Communities | Graph Search | Network Info Flow | Bayesian Networks

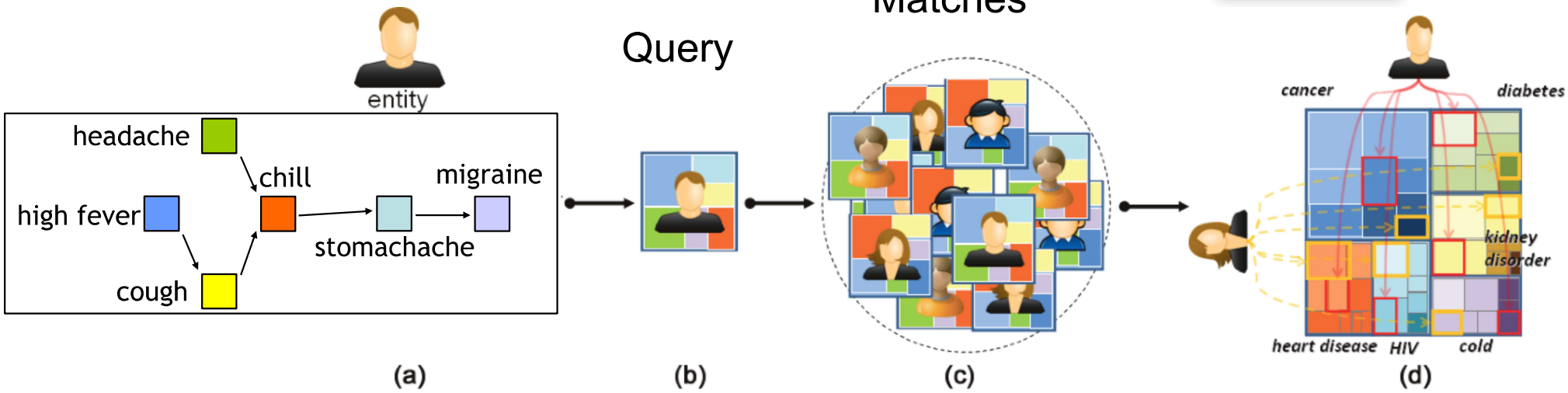
Centralities | Graph Query | Shortest Paths | Latent Net Inference

Ego Net Features | Graph Matching | Graph Sampling | Markov Networks

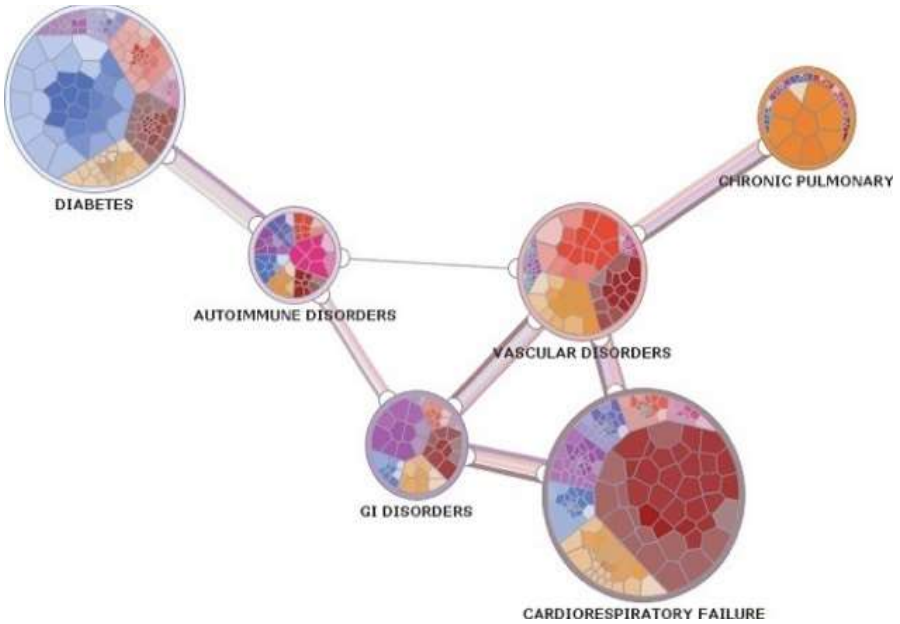
Middleware and Database

# Use Case 7: Graph Analytics and Visualization

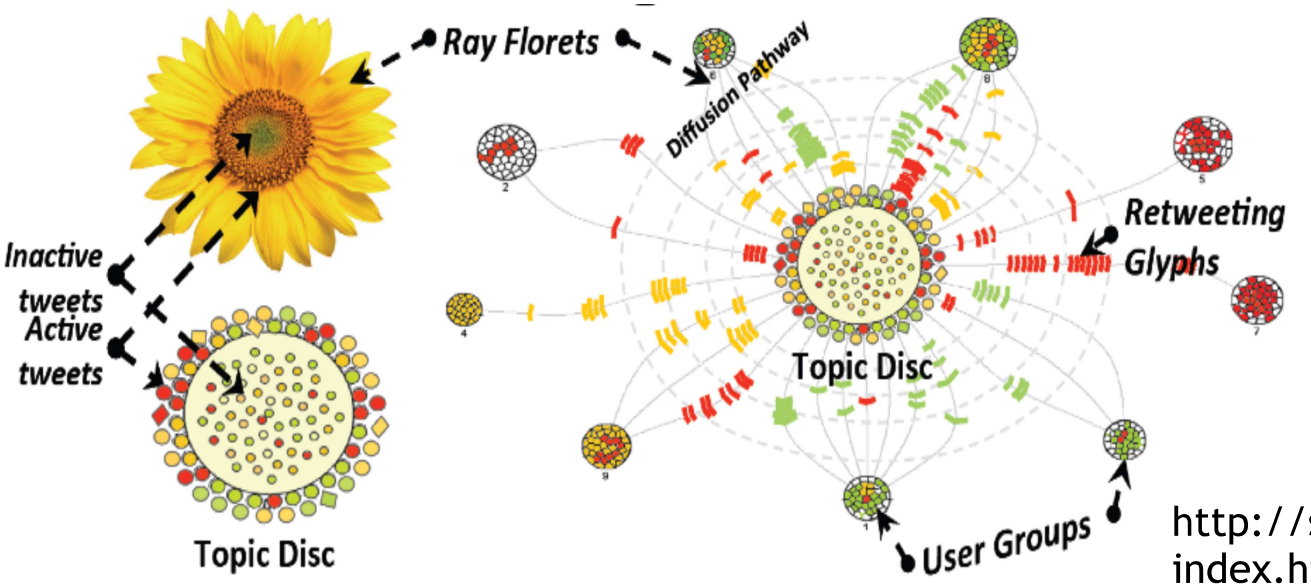
Graph Matching




Graph Communities



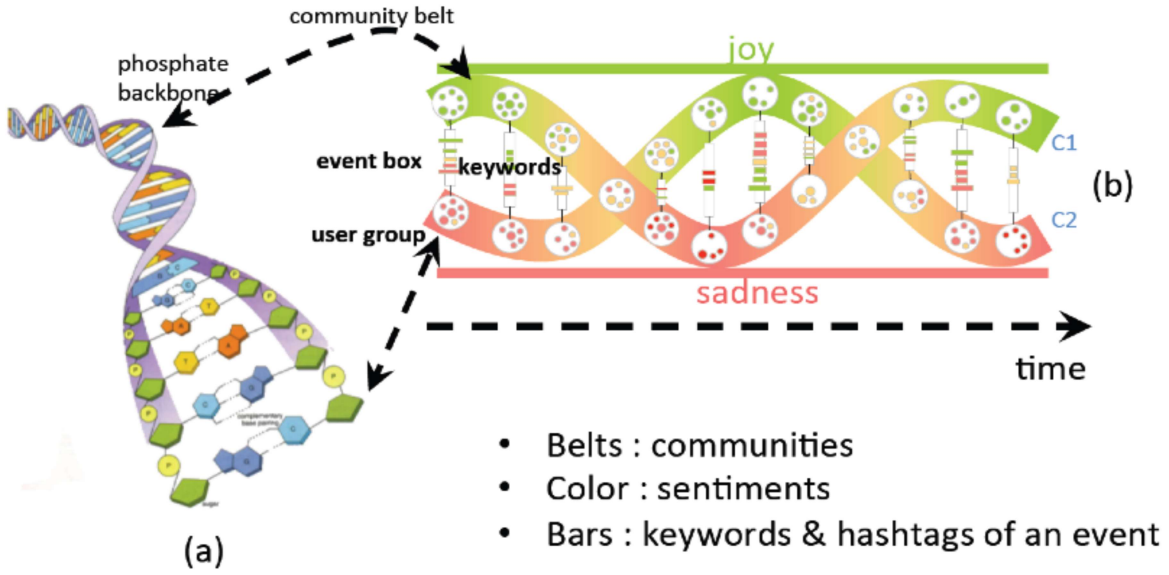
# User Case 8: Visualization for Navigation and Exploration




Whisper : Tracing the information diffusion in Social Media

http://systemg.ibm.com/apps/whisper/index.html

SocialHelix: Visualizaiton of Sentiment Divergence in Social Media

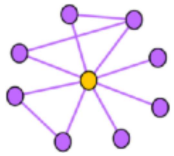


# Use Case 9: Graph Search

Graph Search

existing search engine

query



Info-Socio networks

index

ranking

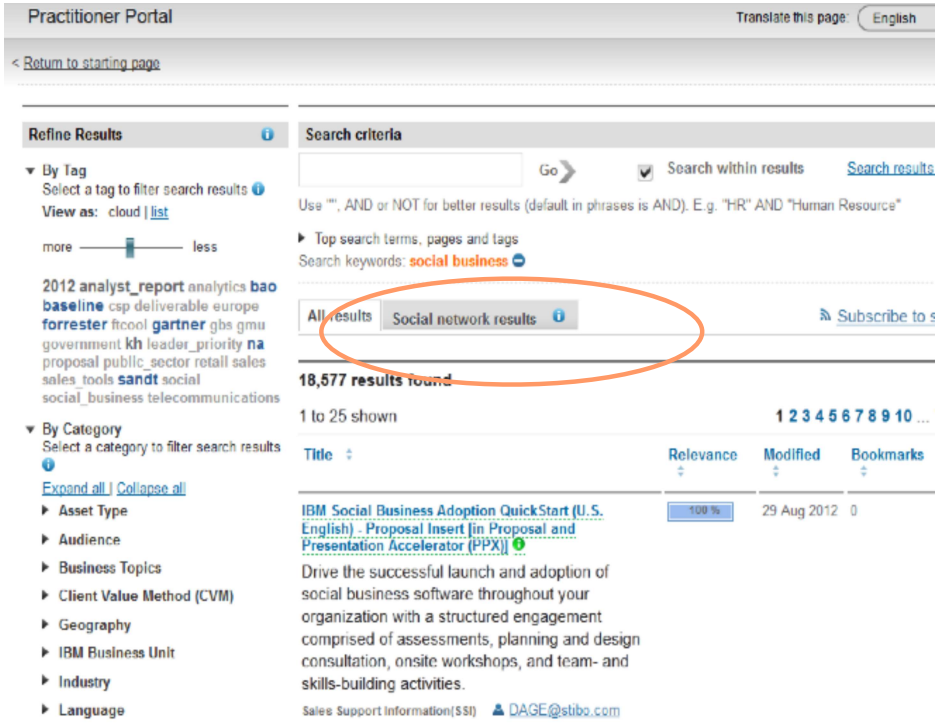
re-ranking

Improved search results

Interest / social network based content recommendations

Graph analysis

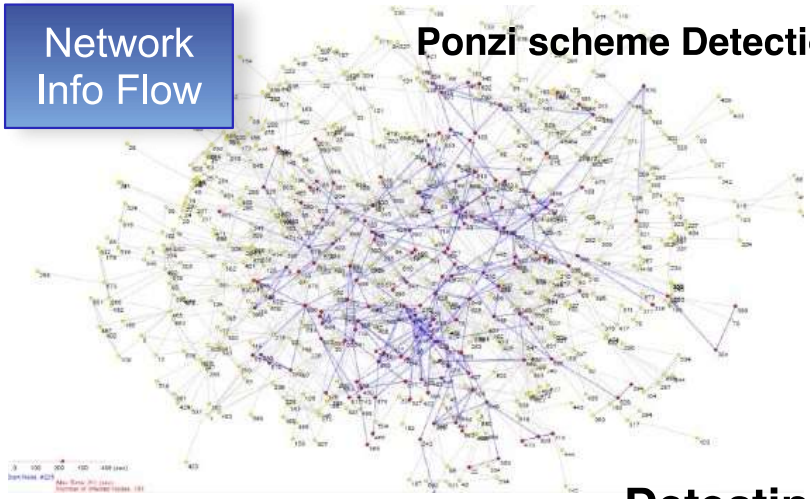
query context



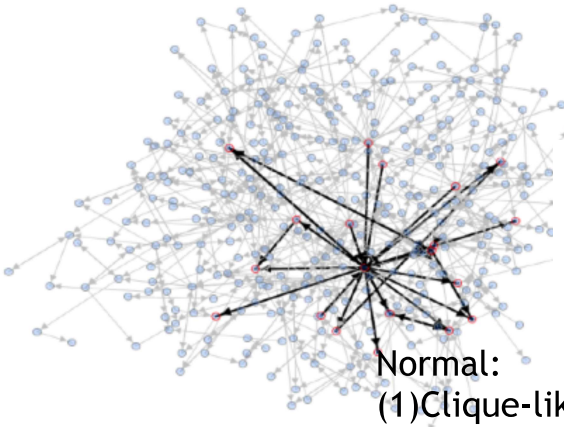
# Category 3: Security

Network Info Flow

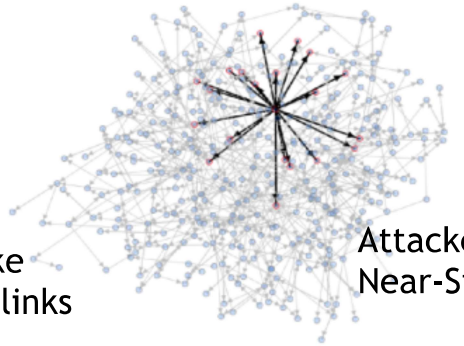
**Ponzi scheme Detection**



Ego Net Features

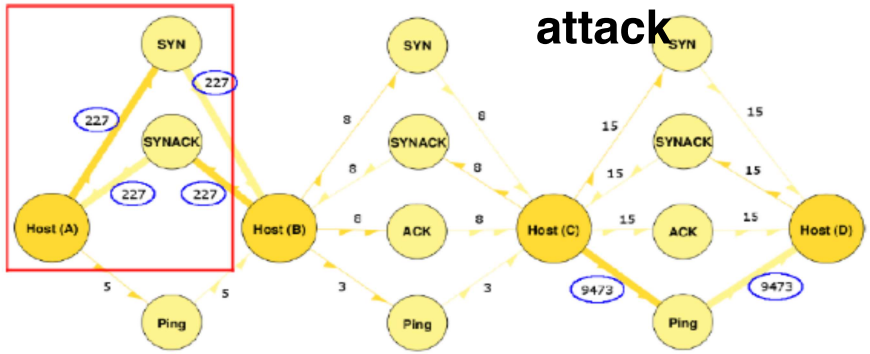


Normal:
(1)Clique-like
(2)Two-way links

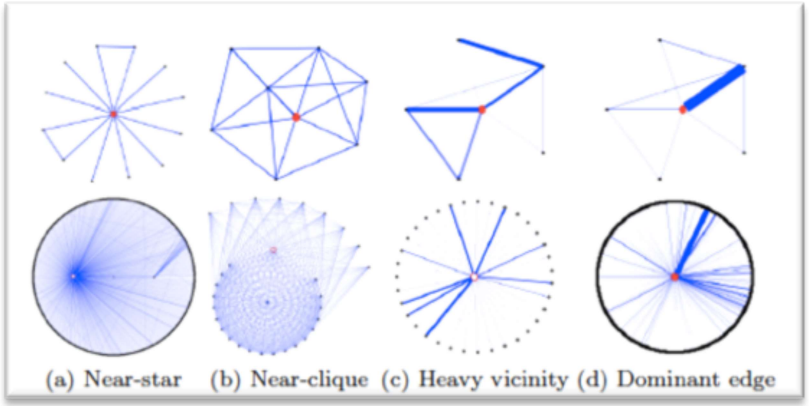


Attacker:
Near-Star

**Detecting DoS attack**




(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.



(a) Near-star (b) Near-clique (c) Heavy vicinity (d) Dominant edge

Graph Visualizations

| | | | |
|---|---|---|---|
| Communities | Graph Search | Network Info Flow | Bayesian Networks |
| Centralities | Graph Query | Shortest Paths | Latent Net Inference |
| Ego Net Features | Graph Matching | Graph Sampling | Markov Networks |

Middleware and Database

# Use Case 10: Anomaly Detection at Multiple Scales

**Based on President Executive Order 13587**

**Goal:** System for Detecting and Predicting Abnormal Behaviors in Organization, through large-scale social network & cognitive analytics and data mining, to decrease insider threats such as espionage, sabotage, colleague-shooting, suicide, etc.


"Enterprise Information Leakage Impacted economy and jobs" Feb 2013

"What's emerged is a multibillion dollar detective industry"
*npr Jan 10, 2013*

**Emails**
**Instant Messaging**
**Web Access**
**Executed Processes**
**Printing**
**Copying**
**Log On/Off**

→ Social sensors
→ Click streams capturer
→ Feed subscription
→ Database access

→ ***Graph analysis***
→ ***Behavior analysis***
→ **Semantics analysis**
→ **Psychological analysis**

→ Multimodality Analysis → Detection, Prediction & Exploration Interface

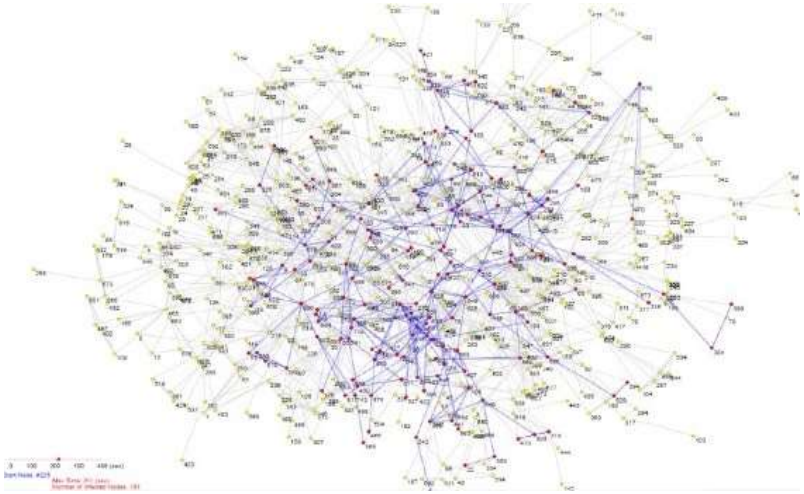
*Infrastructure + ~ 490 Analytics*

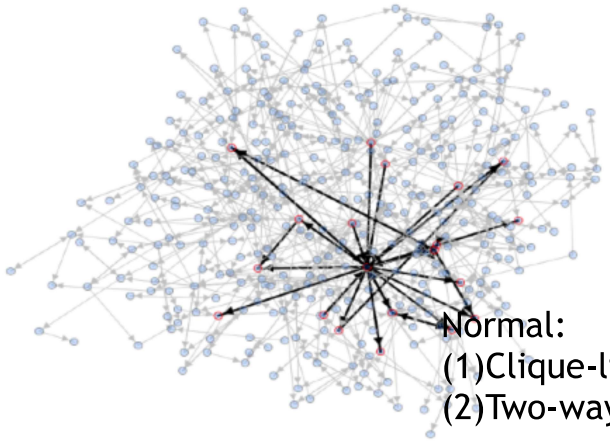
# Use Case 11: Fraud Detection for Bank

Network Info Flow

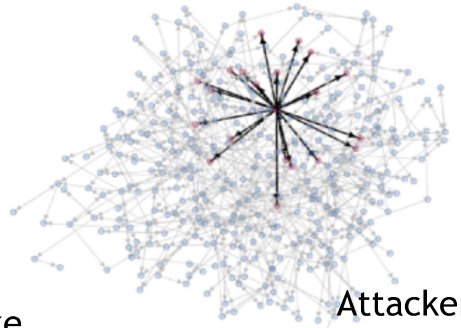
Ego Net Features

**Ponzi scheme Detection**

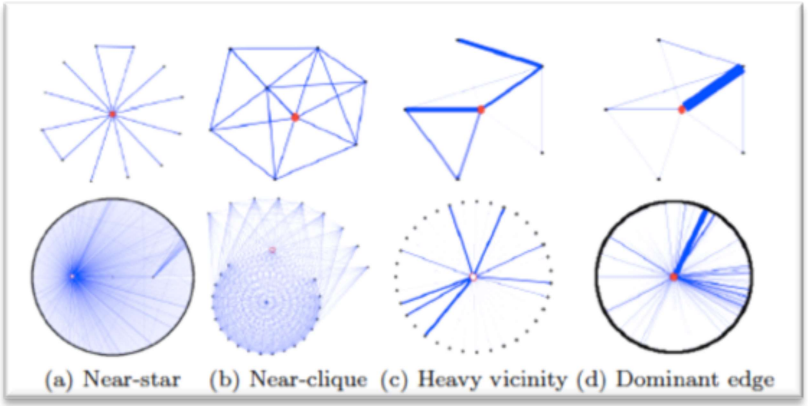




Normal:
(1)Clique-like
(2)Two-way links



Attacker:
Near-Star

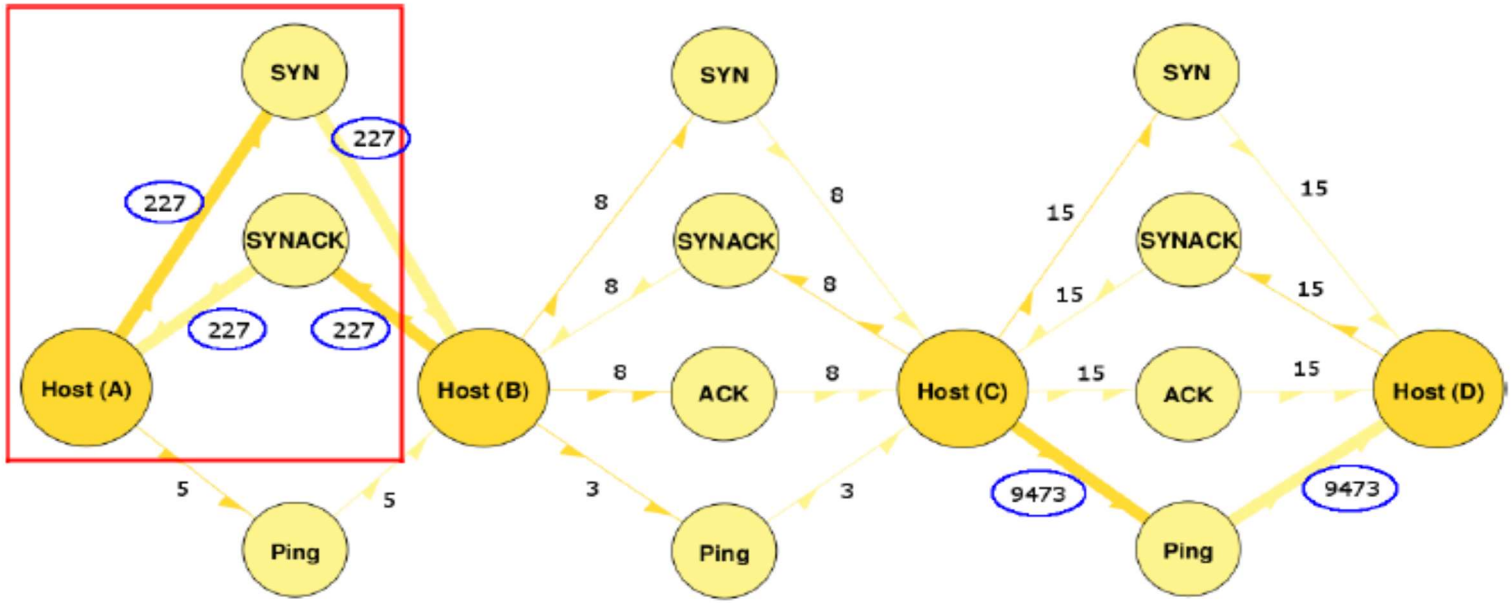


# Use Case 12: Detecting Cyber Attacks

Network Info Flow

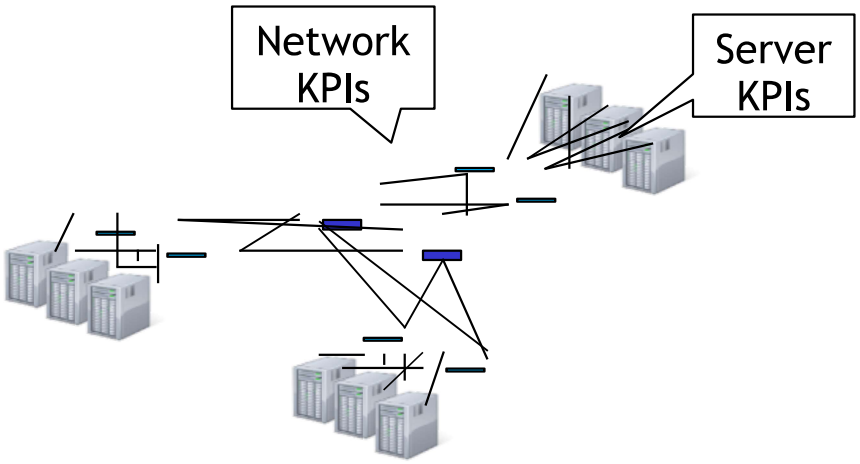
Ego Net Features

**Detecting DoS attack**



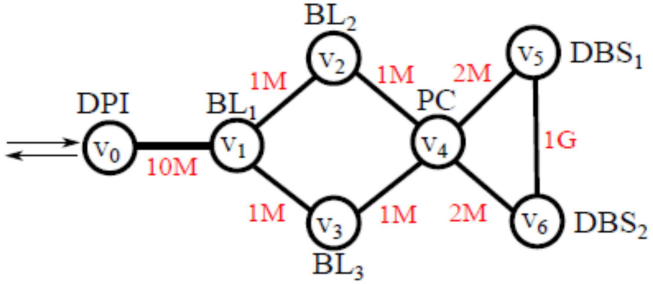

(a) Single large graph representing TCP SYN and ICMP PING network traffic, with two Denial of Service (DoS) attacks taking place.

# Category 4: Operations Analysis




**Cloud Service Placement**

DPI - *Deep Package Inspector* BL - *Business Logic*
PC- *Package classifier* DBS - *DB Server*



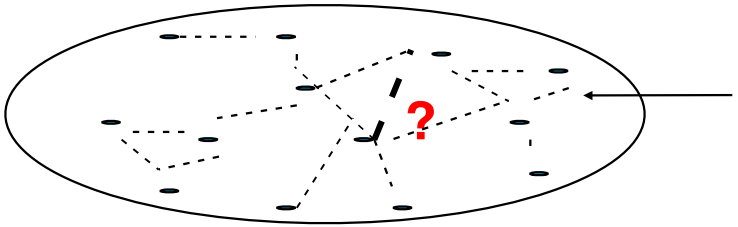

| Memory requirements | |
|---|---|
| $v_0$ | 8G |
| $v_1$ | 2.5G |
| $v_2$ | 2G |
| $v_3$ | 2G |
| $v_4$ | 12G |
| $v_5$ | 20G |
| $v_6$ | 32G |

Graph Matching

Bayesian Network

**KPI time series (e.g., server performance/ load, network performance/load)**

**Causality analyzer**




Varying over time

- KPI (a time series)
- --- (potential) pairwise relationship (e.g., causality)

Graph Visualizations

| | | | |
|---|---|---|---|
| Communities | Graph Search | Network Info Flow | Bayesian Networks |
| Centralities | Graph Query | Shortest Paths | Latent Net Inference |
| Ego Net Features | Graph Matching | Graph Sampling | Markov Networks |

Middleware and Database

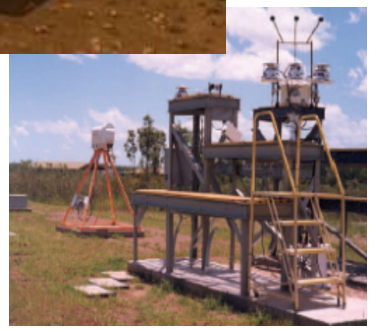
# Use Case 13: Smarter *another* Planet

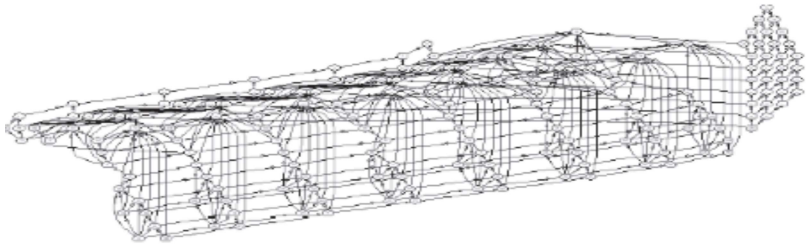
Columbia University

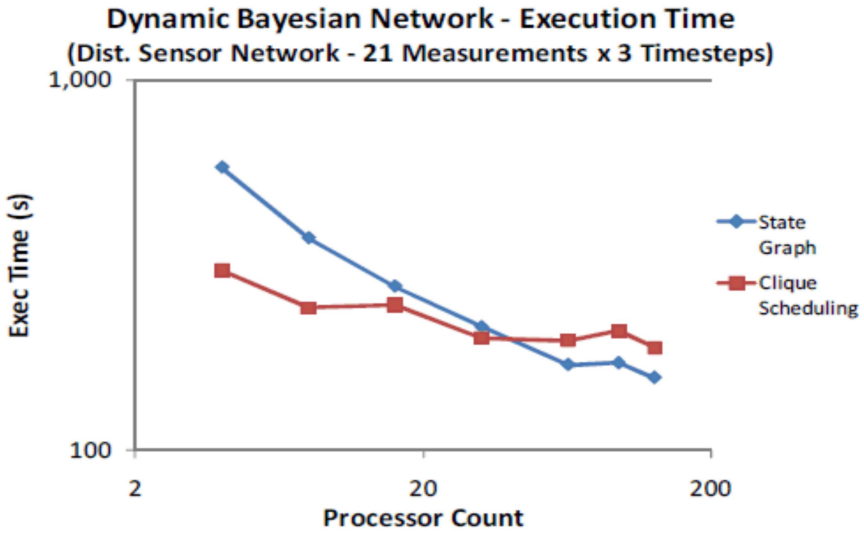
Bayesian Network

**Goal:** Atmospheric Radiation Measurement (ARM) climate research facility provides *24x7 continuous field observations* of cloud, aerosol and radiative processes. **Graphical models** can automate the validation with improvement efficiency and performance.

**Approach:** BN is built to represent the dependence among sensors and replicated across timesteps. BN parameters are learned from over *15 years* of ARM climate data to support distributed climate sensor validation. Inference validates sensors in the connected instruments.








**Bayesian Network**

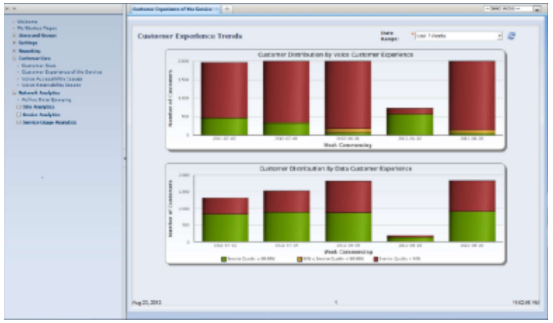
* 3 timesteps * 63 variables
* 3.9 avg states * 4.0 avg indegree
* 16,858 CPT entries

**Junction Tree**

* 67 cliques
* 873,064 PT entries in cliques

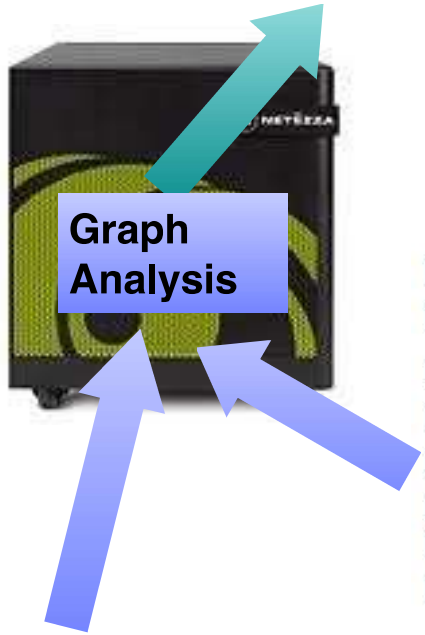
# Use Case 14: Cellular Network Analytics in Telco Operation

**Goal:** Efficiently and uniquely identify *internal* state of Cellular/Telco networks (e.g., performance and load of network elements/links) using probes between monitors placed at selected network elements & endhosts

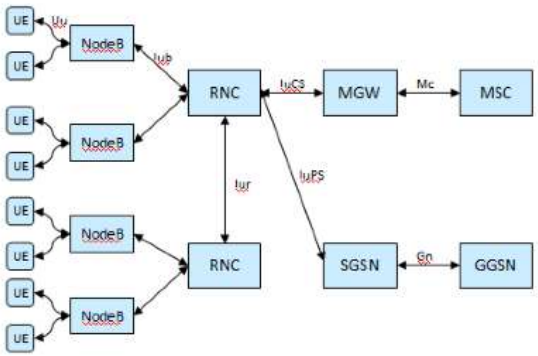


Network load level report

- Applied Graph Analytics to telco network analytics based on CDRs (call detail records): estimate traffic load on CSP network with low monitoring overhead
  - (1)CDRs, already collected for billing purposes, contain information about voice/data calls
  - (2)Traditional NMS* and EMS** typically lack of end-to-end visibility and topology across vendors
  - (3)Employ graph algorithms to analyze network elements which are not reported by the usage data from CDR information
- Approach
  - Cellular network comprises a hierarchy of network elements
  - Map CDR onto network topology and infer load on each network element using graph analysis
  - Estimate network load and localize potential problems



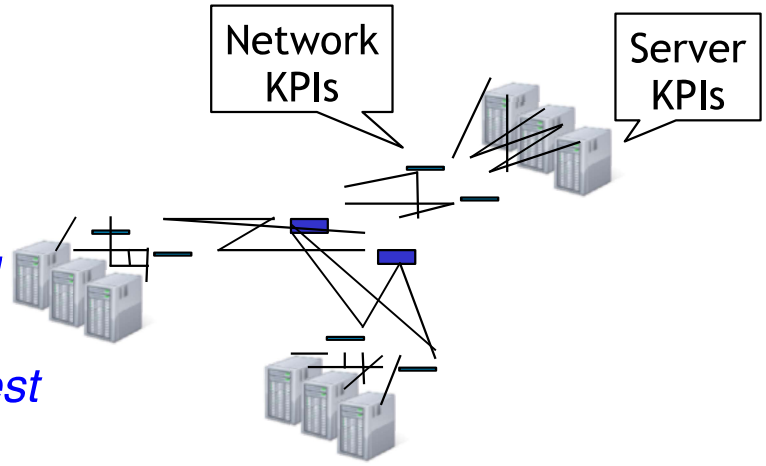

Network topology



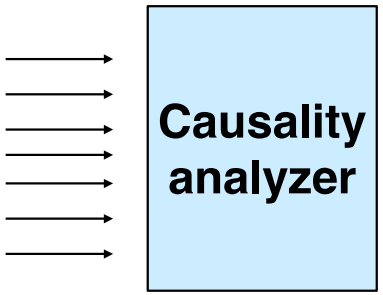
# Use Case 15: Monitoring Large Cloud

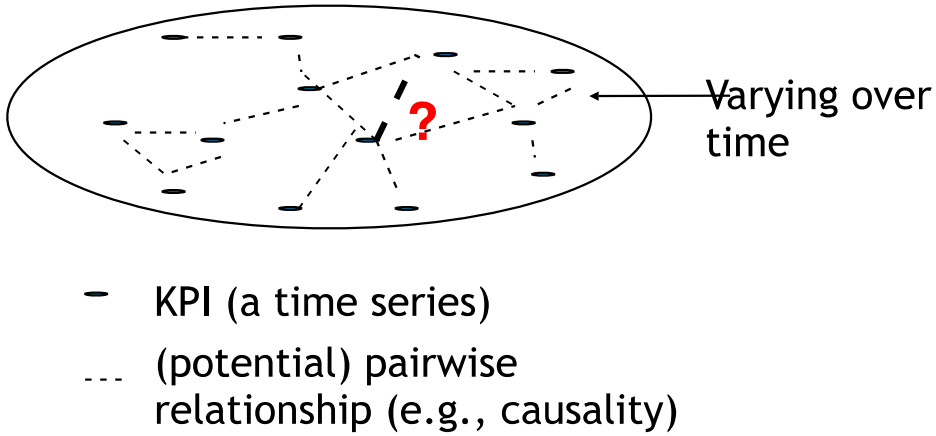
**Goal:** *M*onitoring technology that can track the time-varying state (e.g., causality relationships between KPIs) of a large Cloud when the processing power of monitoring system cannot keep up with the scale of the system & the rate of change

- *Causality relationships (e.g., Granger causality) are crucial performance monitoring & root cause analysis*
- *Challenge: easy to test pairwise relationship, but hard to test multi-variate relationship (e.g., a large number of KPIs)*

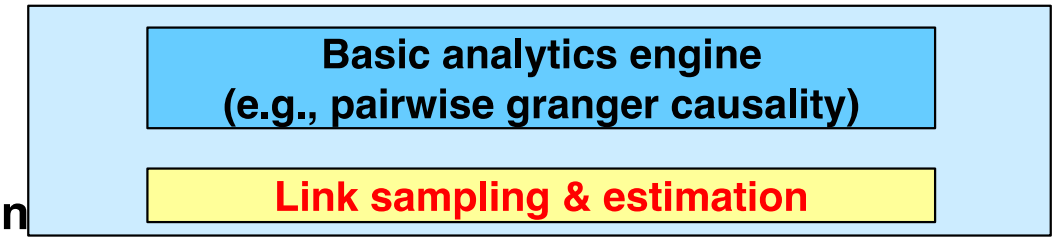



**KPI time series (e.g., server performance/load, network performance/load)**



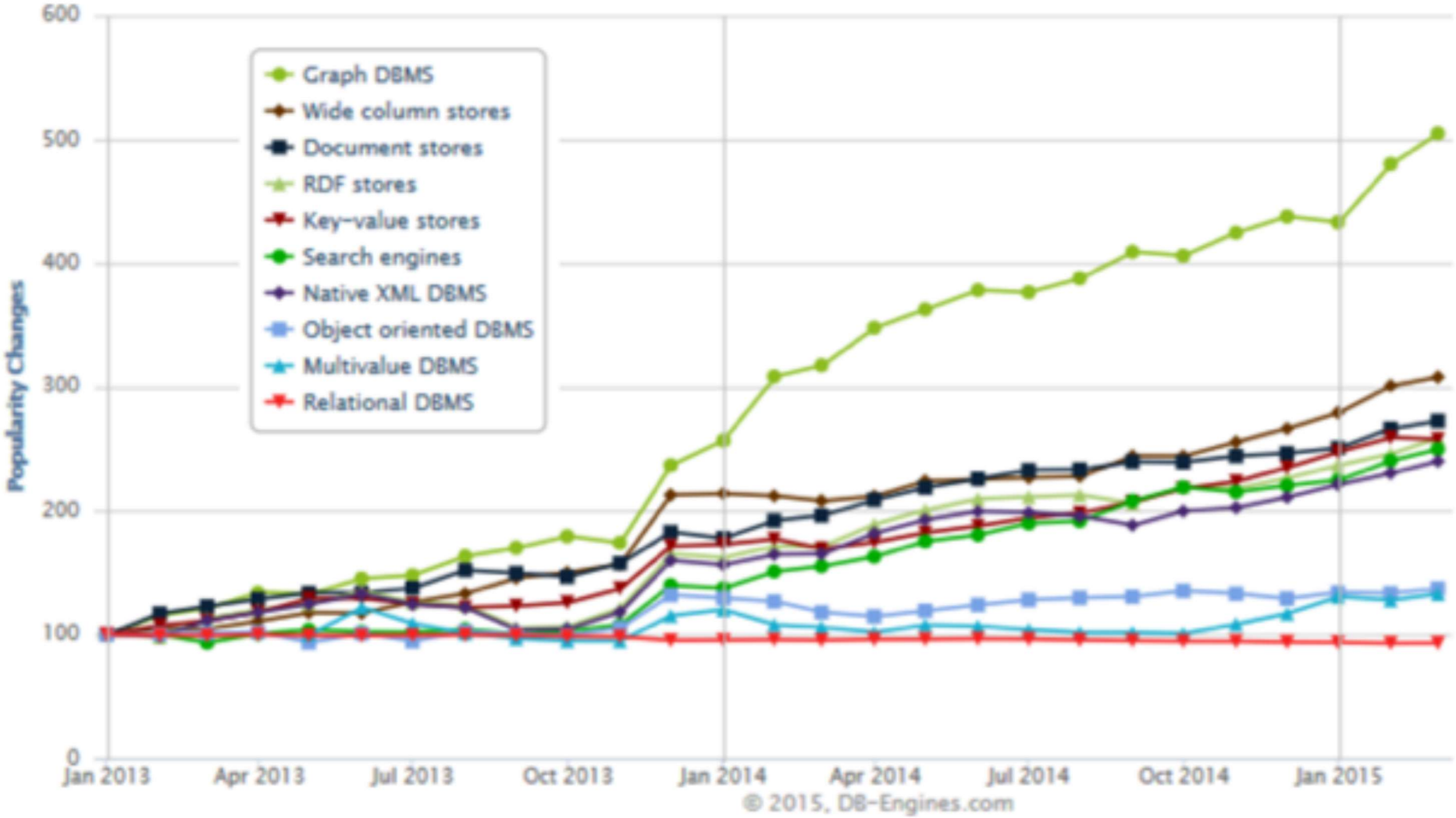




**Our approach: Probabilistic monitoring via sampling & estimation**

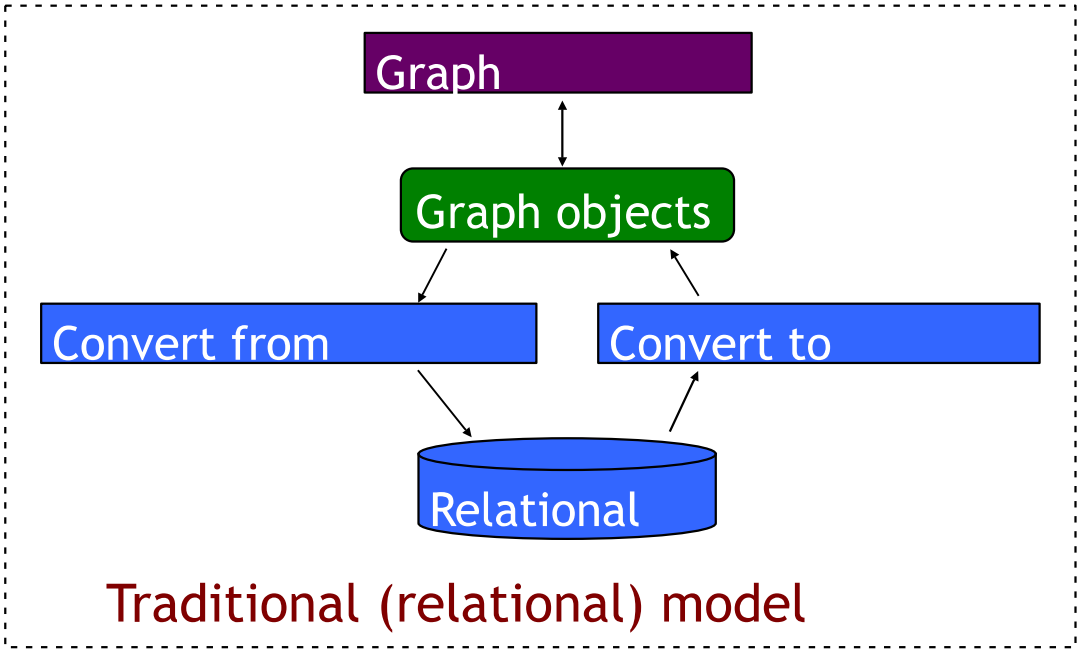


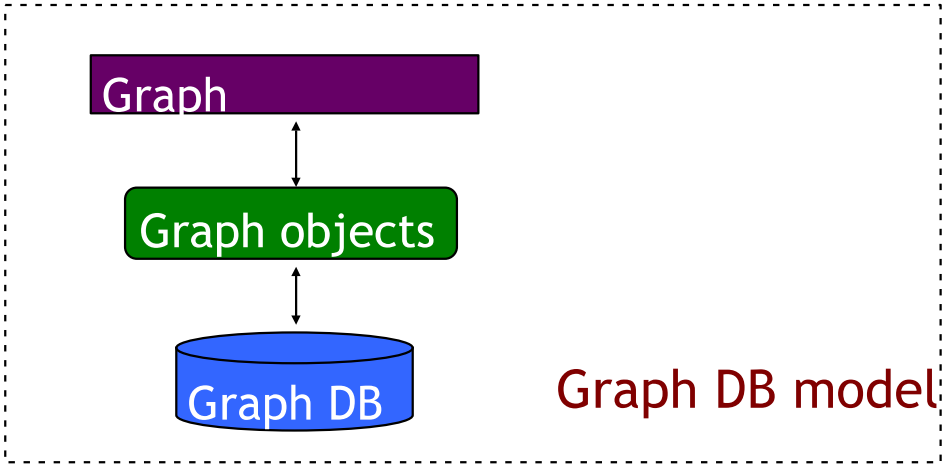

*Select KPI pairs (sampling)→ Test link existence → Estimate unsampled links based on history → Overall graph*

# Category 5: Data Warehouse Augmentation



# Use Case 16: Code Life Cycle Improvement




- Advantages of working directly with graph DB for graph applications
  - (1) Smaller and simpler code
  - (2) Flexible schema → easy schema evolution
  - (3) Code is easier and faster to write, debug and manage
  - (4) Code and Data is easier to transfer and maintain

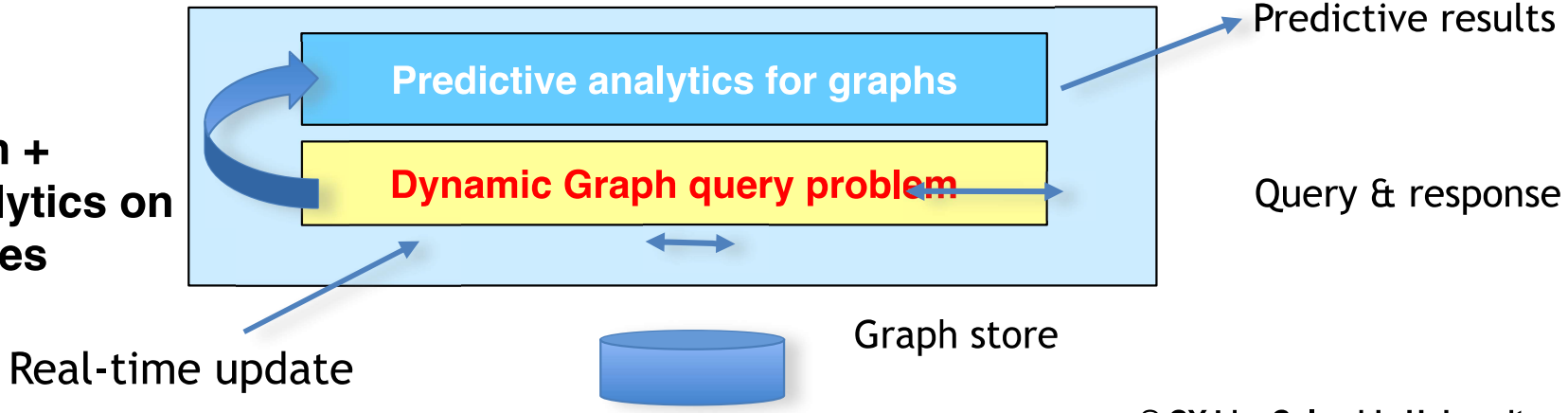
# Use Case 17: Smart Navigation Utilizing Real-time Road Information

**Goal:** *Enable unprecedented level of accuracy in **traffic scheduling** (for a fleet of transportation vehicles) and navigation of individual cars utilizing the **dynamic real-time information** of changing road condition and predictive analysis on the data*

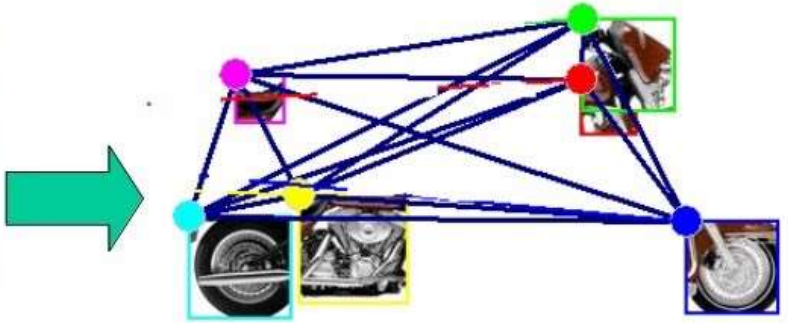
- *Dynamic graph algorithms implemented in System G provide **highly efficient graph query computation** (e.g. shorted path computation) on time-varying graphs (order of magnitudes improvement over existing solutions)*

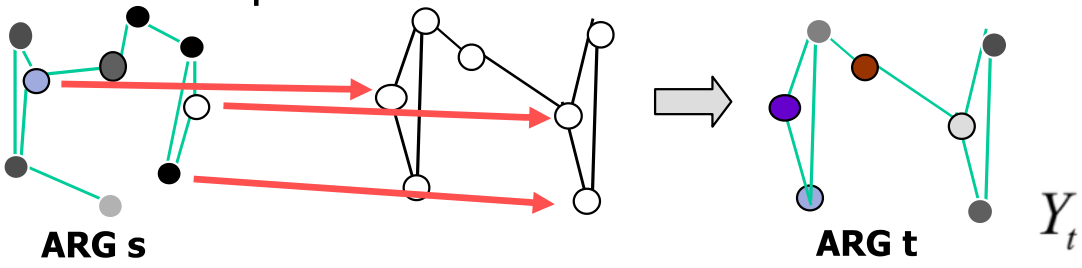
- *High-throughput **real-time predictive analytics** on graph makes it possible to estimate the future traffic condition on the route to make sure that the decision taken now is optimal overall*

**Our approach: Querying over dynamic graph + predictive analytics on graph properties**



# Use Case 18: Graph Analysis for Image and Video Analysis






# Use Case 19: Graph Matching for Genomic Medicine

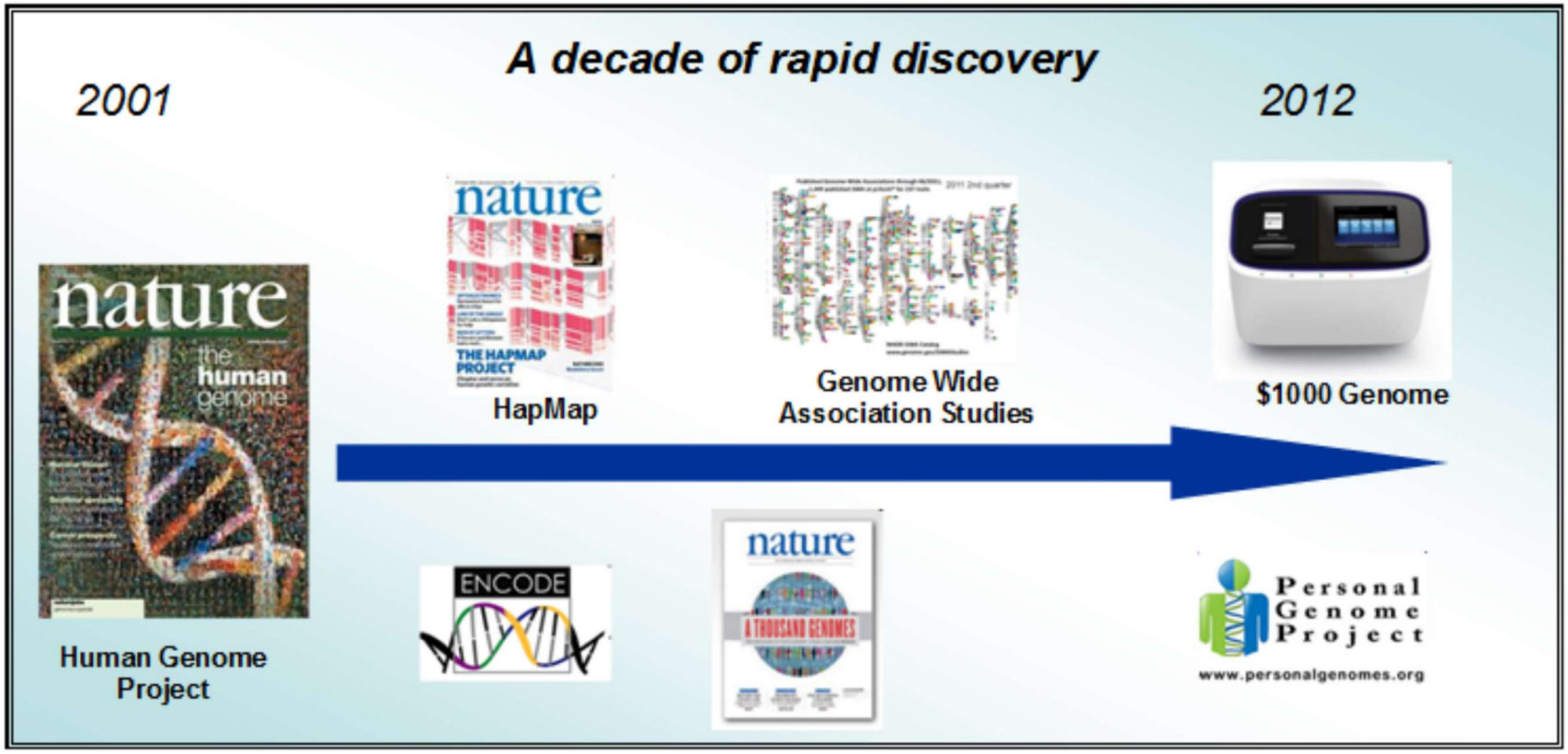
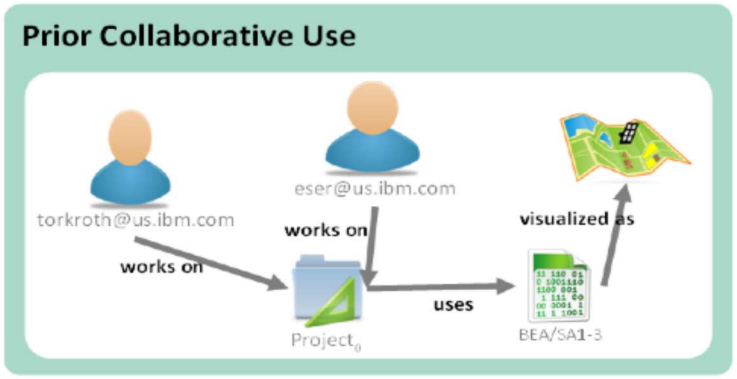
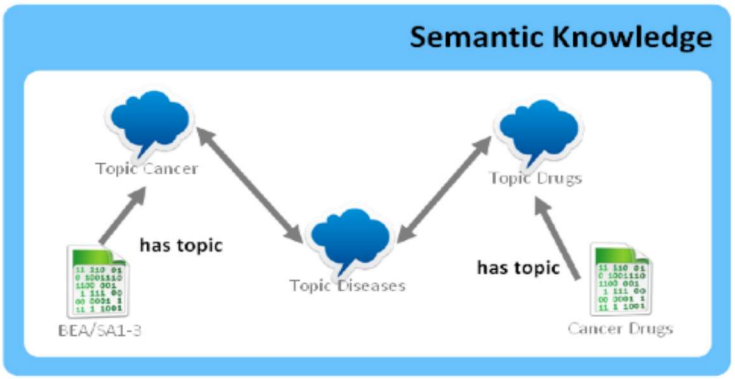
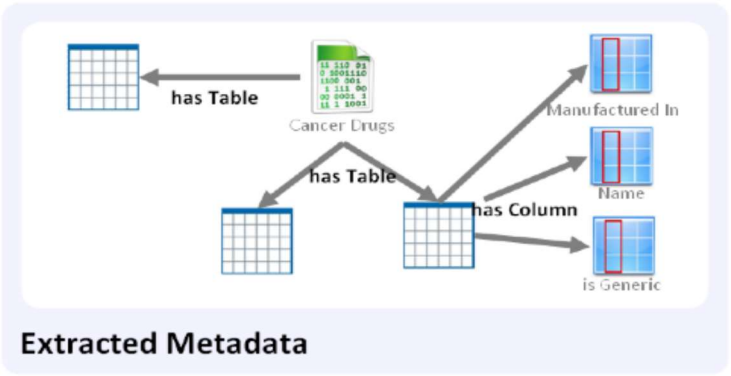


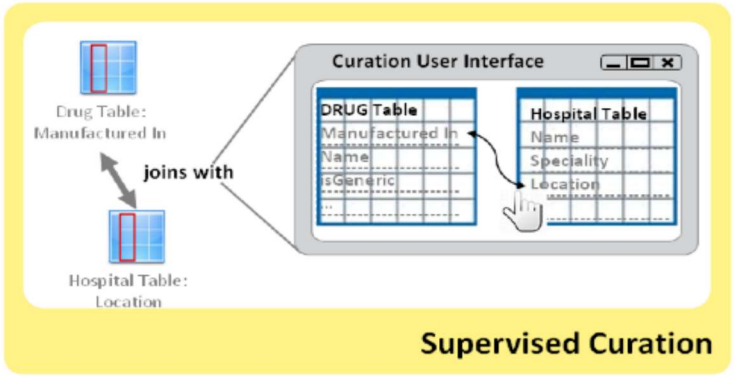

**Figure 1: Since the Human Genome Project, various projects have started to reveal the mysteries of genomes and the $1000 Genome is almost reality.**

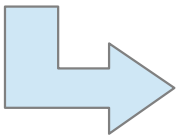
# Use Case 20: Data Curation for Enterprise Data Management

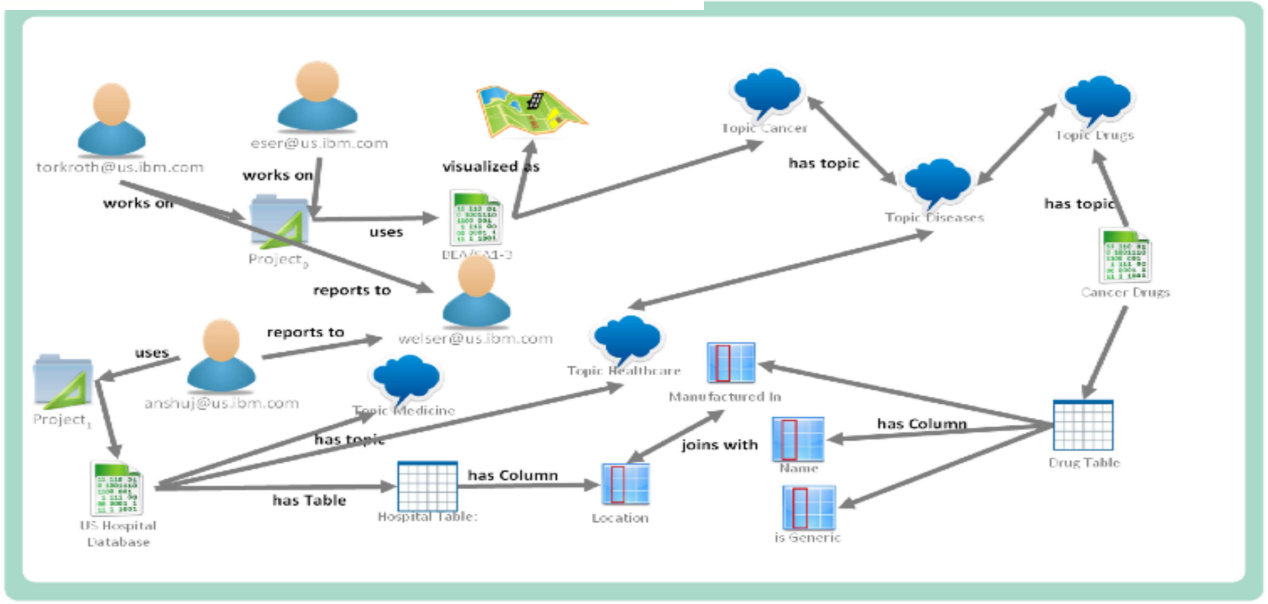


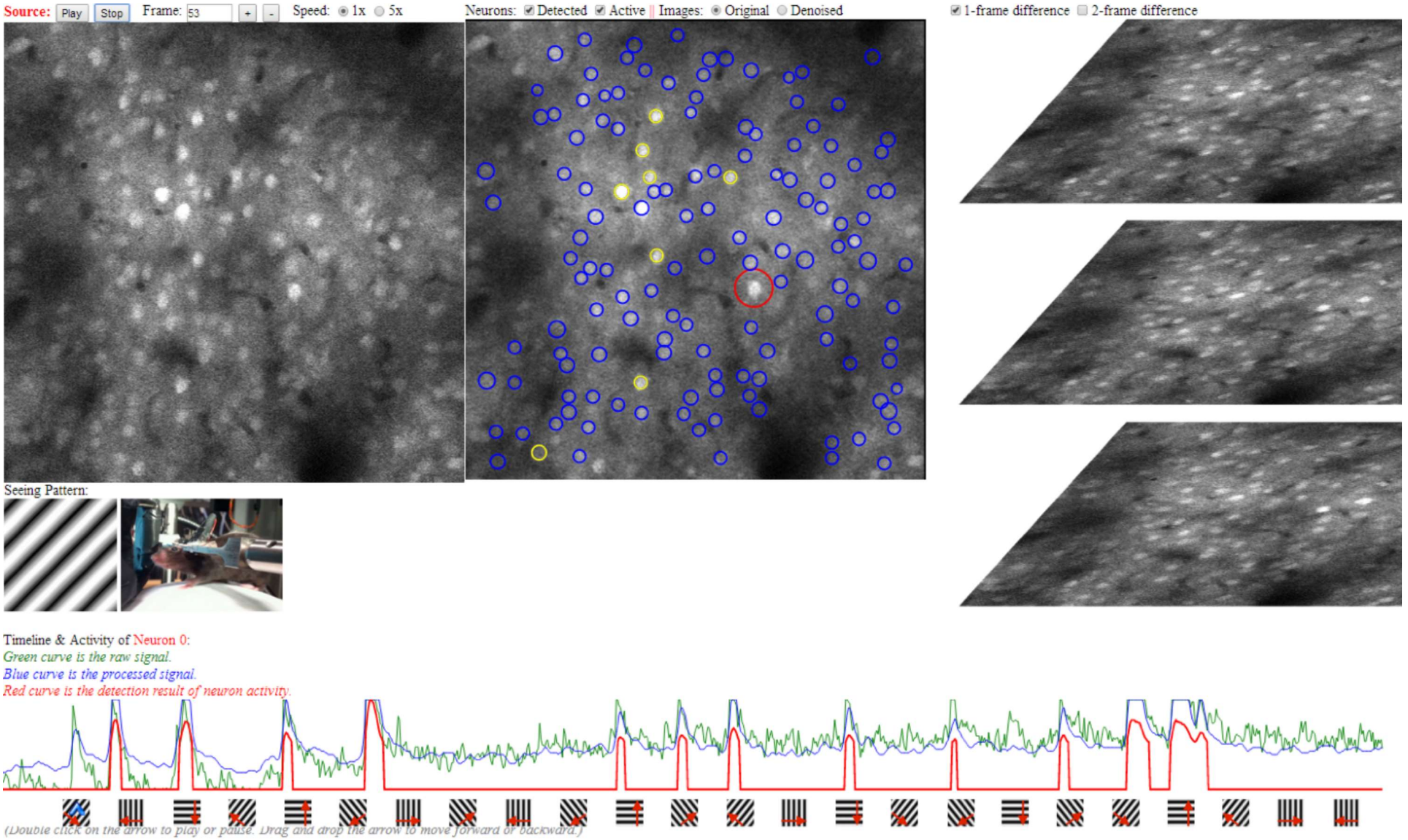





# Use Case 21: Understanding Brain Network



# Use Case 22: Planet Security

- Big Data on Large-Scale Sky Monitoring

# Homework #0: Big Data Environment Setup and Test (due September 24, 5pm)

1. Warm-Up Exercises:
   - Setup Google Cloud account and environment
   - Install Google Cloud SDK
   - Create a Spark cluster
   - Word Count using Google Cloud Storage and Spark
   - Hive and BigQuery

2. Data Analysis — NYC Bike Expert:
   - Load data to a Cloud Storage
   - Simple Analyses through BigQuery

3. Data Analysis — Understanding Shakespeare:
   - Load data to a Cloud Storage
   - Simple Analyses through Word Counts
   - Analyses after running Natural Language Toolkit

# Homework Late Submission Policy

5pm: submission deadline
Next Day 5pm: 10% penalty
Two Days late 5pm: 20% penalty
Three Days late 5pm: 30% penalty
Any late submission more than 3 days (5pm) will not be accepted.