Final Exam ADM1370

Part 1: M/C and T/F

- 75 Questions (75 Marks)
- 60 minutes
- 25 questions from each section
- Similar to quizzes

Part 2: Short Answer

- 6 short answer questions
- 2 questions per section
- 30 marks
- lecture notes and class discussions

Wiki:

- course pack units 1, 2 and 3
- not on technical part of wikis

Excel:

- course pack all units
- short answer will test:
    - theoretical concepts
    - types of business intelligence dashboards

You may be tested on your familiarity with concepts such as:

- Purpose of creating Excel tables
- Difference between conditional and logical functions
- Difference between structured, absolute and relative references
- Purpose of decision analysis / what-if analysis tools such as Data Tables, Goal Seek, and Solver

While you are required to know the purpose of various Excel features and functions that were discussed in class, you will not be asked to recall parameters for functions or their order:

- For example, you should know that VLOOKUP allows you to search for values in a table that correspond to a lookup value, you will not be asked to write the function or specify what parameters are required.

Access:

- course pack: the first chapter only “Data, Text, and Document Management”
- all slides
- Students won’t be tested on technical aspects of MS Access such as those from the lab assignment.

Wiki

- Information technology tools to support **operational, tactical and strategic** level processes and decision making in an organization

- What is an Information System?
  - Provides accurate, timely and useful information
  - An Information System consists of FIVE PARTS, including:
    - ⇨ People, Procedures, Software, Hardware, and Data
- IT = hardware + software + data

- **Data**: symbols , raw facts, must be “put together” in order to form information
- **Information**: data that are processed to be useful; provides answers to "who", "what", "where", and "when" questions
- **Knowledge**: application of data and information; answers "how" questions
- **Understanding**: appreciation of "why"
- **Wisdom**: evaluated understanding.

- Three core activities of information systems:
  - **Input**: Captures raw data from organization or external environment
  - **Processing**: Converts raw data into meaningful form
  - **Output**: Transfers processed information to people or activities that use it

- E-business involves:
  - Consumers
  - Producers
  - Information
  - Connectivity

- **Social Media**: is a group of Internet-based applications that build on the *ideological and technological foundations of Web 2.0*, and that allow the creation and exchange of User Generated Content.

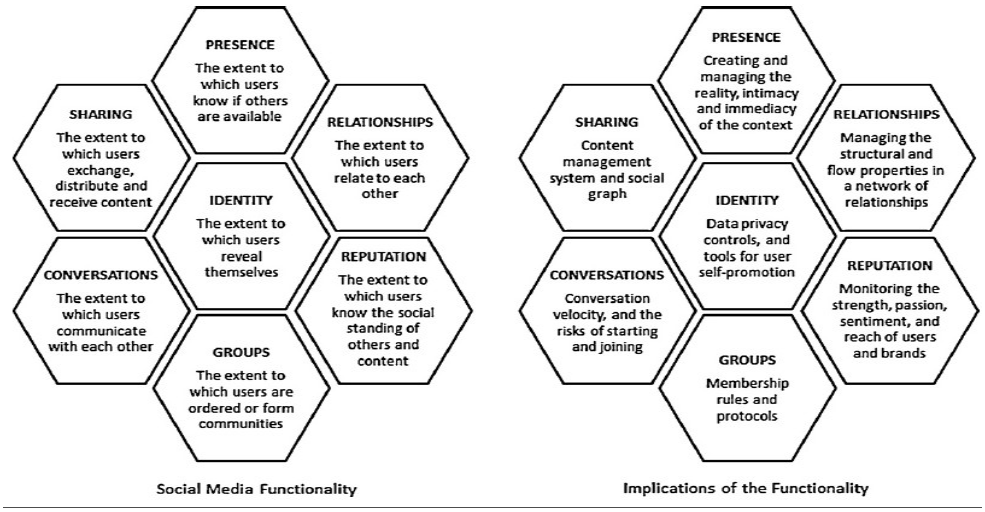
- Business process involves MANNER, FLOW and COORDINATION
- Innovative Connectivity
  - Social media tools **connect** people to processes and information
  - Apps are simply the means of doing that to help prevent information indigestion
  - It **accelerates** the data to information cycle to make information acquisition more efficient
  - Business processes gain **new perspectives**

Exam Question: Gives a company and pitches situation of invention. How did it change nature of business to become more sustainable? Why did it give them an advantage? What was the game changer?

- **Digitalization**: the replacement of physical tasks by digital means
- **Digital Goods**: Goods that can be delivered over a digital network

- ***Competitive advantage*** – a product or service that an organization’s customers place a **greater value** on than similar offerings from a competitor
- ***First-mover advantage*** – occurs when an organization can significantly impact its market share by being **first to market** with a competitive advantage

- **Porter’s Five Forces** = suppliers, buyers, threat of new entrants, rivalry of existing competitors, and substitutes



- Three strategies for generating revenue from Web 2.0 applications:
  - Advertising
  - Subscriptions
  - Commissions
- **Wiki**: a website/online resource which allows users to **add** and **edit** content **collectively.** (A collaboration tool)

- Flattener #4: Uploading / Open-Sourcing:
  - Self-Organizing Collaborative Communities.
  - Individuals as consumers and producers of content.
  - Fast dissemination of information, news and events.
  - Collaboration on open-source & free software.
  - EX: social media, user based content, collective intelligence, forums, communities

- Flattener #9: Informing
  - Ability to build and deploy your own supply chain of information, knowledge and entertainment
  - Self-empowering; empowering individuals to do what they think best with the information they want.
  - The opportunity for people to have private, semiprivate, or public gatherings on the Internet regardless of geography and time
  - EX: Search Engines

- Flattener #10: Steroids
  - The new technologies that are amplifying and turbo-charging all other flatteners.
  - EX: Computing speed and capacity, Instant messaging, Videoconferencing, Computer graphics, Wireless technologies and devices, mobile phones, personal devices
  - Keywords: digital, mobile, personal, virtual.

What were the three specific examples of Digital Steroids cited by Friedman?

1: Computing Speed / Storage / Portability;
2: Peer-to-Peer Features (IM, File Sharing);
3: Voice over IP (e.g. Skype)

- **Business Information Strategies:**

Strategy	Description	Strengths	Weaknesses	When to Use
Multidomestic	Federation of associated business units; decentralized	Ability to quickly react to local conditions	Differing product offerings limit economies of scale, and limited interunit communication limits knowledge sharing	Very heterogeneous markets
Global	Centralized organization with standardized offerings across markets	Standardized product offerings allow achieving economies of scale	Inability to react to local market conditions	Homogeneous markets
Transnational	Some aspects centralized, others decentralized; integrated network	Can achieve benefits of multidomestic and global strategies	Difficult to manage; very complex	Integrated global markets

- **RSS** (Really Simple Syndication): used to publish frequently updated content such as blog entries and news headlines
- **Mashup:** a web application that combines data or functionality from more than one source into a single integrated interface or tool:
  - applications generated by combining content, presentation, or application functionality from disparate sources.
- **Disruptors:** Companies that introduce a significant change in their industries thus cause disruption in the way business is done

Web 1.0 vs. Web 2.0

Web 1.0	Web 2.0—the Social Web
Static pages	Dynamic pages
Author controlled content	User controlled content
Computers	Computers, cell phones, televisions, PDAs, game systems, car dashboards
Users view content	Users create content
Individual users	User communities
Marketing goal: *influence*	Marketing goal: *relationships*
Data: single source	Data: multiple sources, e.g., mashups

- **Enterprise 2.0** refers to Web 2.0 technologies used for some business purpose:
  - ⇨ Promote collaboration and knowledge exchange among employees, consultants and company partners

- Tool-based metrics: Metrics a company uses are determined by what the company is trying to achieve.

- Tactical metrics:
  - ⇨ Example: Increase web traffic
  - ⇨ Example: Increase number of user accounts

- Strategic metrics: aim to fully capture the uses of web 2.0
  - ⇨ Attract listening, talking, support
  - ⇨ Example = united air lines breaks guitars

- Web 2.0 is about commerce, connecting people (THE SOCIAL WEB)
  - ⇨ Consumers become more powerful, and companies try to leverage that (reviews on products, and companies developing from that)

- Web 3.0 is all about connecting KNOWLEDGE (THE SEMANTIC WEB)
  - ⇨ Content, Commerce, Community & Context, Personalization

- Web 4.0 will connect intelligence (senses your needs, futuristic technology) (THE UBIQUITOUS WEB)
  - ⇨ We are somewhere between web 2.0 and web 3.0
  - ⇨ See slide 20, of slideshow - 06
  - ⇨ Things like tagging and hashtags move us toward web 3.0

- **Crowd sourcing:** asking users what they want, and what they are demanding. One challenge can be getting the right answer and valid information. A lot of people that answer won't even know what you're asking

- ⇨ You need a strategy and clear boundaries
- ⇨ Opens the potential for unethical labour (crowdsource your wiki and pay them to write your wiki)
- ⇨ Privacy becomes an issue too

- **Main benefit of web 3.0**: allows us to track information because of tags on the information
- Some of the **barriers** are coming around data. As more and more data is collected, how do we decide to bundle the data and tag it. It becomes hard to organize.
- **Social capital:** the advantage created by a person's location in a structure of relationships. It explains how some people gain more success in a particular setting through their superior connections to other people.
- **Social Network Analysis (SNA)** is the mapping and measuring of relationships and flows between people, groups, organizations, computers, or other information or knowledge processing entities
  - ⇨ **Aggregator, geographical, industry, topical**

- Web 2.0 is often called "the social web"
- Many consider it an extension of the old web, not necessarily something new.
- Web 2.0 Applications:
  - ⇨ **Blog**: Content marketing, where info is shared to users
  - ⇨ **Wikis**: many people can edit and update a site, contributions run it.
  - ⇨ **Social networking services**: defined by a *profile* (Facebook, Twitter, LinkedIn)
  - ⇨ **Sharing Sites**: upload and share various media (YouTube)
  - ⇨ **Widgets and Mash-ups**: stand-alone add-ons
  - ⇨ **RSS (really simple syndication)**: feed to allow users to view updates in real time.
  - ⇨ **Social Bookmarking**: Tracking sites that show favourite and most visited
  - ⇨ **AJAX Technologies**: respond to users' actions without the whole page reloading (HTML, XML, CSS, Java)
  - ⇨ **Social Media**: Collectively all these apps are social media, because control has shifted from organizations to the users individually.
- Businesses have been integrating technologies in order to raise revenues; they risk irrelevancy if they don't participate.
- Online communities have formed: interacting, collaborating and trading
  - ⇨ Associations, Ethnic communities, Gender communities, Affinity portals (organized by interest), Young People, B2B online (technology blogs and meetings and webcasts), Social networking sites (the mega communities)
- Social networking analysis has begun to build relationships with consumers as they use the tech.
- Facebook remains the most used SNS (social networking service)
- <u>Enterprise 2.0 Tools:</u>
  - ⇨ Social media use in businesses is exponentially increasing
  - ⇨ Used for:
    - recruiting and professional networking;
    - Marketing, promotion and sales;
    - Internal collaboration and communication and
    - Supply chain management
- <u>Social Media Objectives and Metrics</u>:
  - ⇨ Metrics: evaluating the use and patterns of social media.
  - ⇨ Response hierarchies are used to enhance: awareness, knowledge, liking, preference and purchase

Excel

Why Learn Excel?

- most **popular** spreadsheet tool today.
- true “killer app” by many experts.
- using Excel at work, and possibly at home.
- used for simple data management and problem-solving, as well as complex decision making.
- used to **cultivate**
- Critical Thinking and Decision Making Skills
- used to enter, analyze, and present (quantitative) data
- automatic calculation functions
- presentation tools
- enhance the presentation and facilitate **sorting** and **filtering** of data in the ranges.

Four skills for a job:

- abstract **reasoning** skills
- system **thinking** skills
- **collaboration** skills
- **experimentation** skills

Worksheet: a spreadsheet that may contain data including text, numbers, formulas, charts etc. Sometimes a charts based worksheet is referred to as a “Chartsheet”

Workbook: a collection of related worksheets within one file

File Tab: Consists of file operations commands such as opening, closing, saving, printing, and sharing files.

Tabs & Ribbon: Each tab corresponds to sets of features displayed horizontally as a ribbon. A ribbon consists of groupings, and controls. Tabs are designed to be task-oriented and consist of several logical groupings of controls that perform similar functions.

Quick Access Toolbar: Contains controls / commands that are most commonly used. Additional controls can be added through Excel Options settings through the Office Button.

Select All button: used to select all elements of the worksheet

Status Bar: Displays information about a selected command or operation in progress. Also displays basic summary information about a selected range of values.

Range: Can consist of contiguous (together) or non-contiguous (not together) cells

Order of precedence rules

Formula (A1=50, B1=10, C1=5)	Order of Precedence Rule	Result
=A1+B1*C1	Multiplication before addition	100
=(A1+B1)*C1	Expression inside parentheses executed before expression outside	300
=A1/B1–C1	Division before subtraction	0
=A1/(B1–C1)	Expression inside parentheses executed before expression outside	10
=A1/B1*C1	Two operators at same precedence level, leftmost operator evaluated first	25
=A1/(B1*C1)	Expression inside parentheses executed before expression outside	1

- Using VLOOKUP:
  - For Closest Matches: e.g. mapping numerical marks to letter grades
  - For Exact Matches: e.g. mapping salary levels to employee categories

- =AVERAGE(number1,number2,..) Excel sums the values in the range and then divides by the number of non-blank cells in the range
- =STDEV(number1,number2,..)
- =MAX(number1,number2,..)

- =MEDIAN(number1,number2,..)
- =PEARSON(number1,number2,..) Pearson correlation coefficient
- =COUNT(range) Counts the number of cells that contains numbers
- =COUNTA(range) Counts the number of cells that are not empty. Both numeric and text entries are included
- =COUNTBLANK(range) Counts the number of empty cells
- =COUNTIF(range,criteria) Counts the number of cells within a range that meets the condition. Will try this one in detail in upcoming lab sessions.

- **Grouping: o**rganizing data so that it can be viewed as a collapsible and expandable outline.
- **Formulation:** Transformation of a real problem scenario into a mathematical model
- **Solution:** Solving the model to obtain the optimal solution
- **Interpretation:** Analyzing results and implementing solution

FV(rate,nper,pmt,pv,type)
- **Rate:** Interest rate per period – e.g. 8 percent annual interest rate
- **Nper:** Total # of payment periods in an annuity
- **Pmt :** Payment made each period. Cannot change over the life of the annuity
- **Pv: (**optional) present value or the lump-sum amount that a series of future payments is worth right now
- **Type:** (optional) the number 0 or 1 indicates when payments are due. If type is omitted, it is assumed to be 0 (end of period

- **Heuristics:**
  - **Heuristic methods are used to speed up the process of finding a good enough solution, where an exhaustive search or advanced problem-solving techniques are impractical.**
    - a "rule of thumb", an educated guess, an intuitive judgment, or common sense.
  - **The most fundamental heuristic is trial and error.**
  - **Other examples of heuristics:**
    - Maximax method, Maximin method , Averaging method, Expected Monetary Value (EMV)

- **Logical Tests: Result of these functions is a logical value (TRUE or FALSE).**
  - **Other Logical Functions:**
    - AND
    - OR
    - NOT

- **Conditional Functions**: Result may be a specified value or a calculated value**.**
  - **Examples:**
    - IF
    - COUNTIF
    - SUMIF
    - AVERAGEIF
    - COUNTIFS, SUMIFS, AVERAGEIFS (to be used when multiple criteria are to be tested)

- **AND Function**
  - Allows you to test the condition of more than just one criterion (condition).
  - Returns either TRUE or FALSE.
  - Only returns TRUE if all tested values are TRUE
  - AND (logical1 [,logical2]…)
    - Example: **=AND(G2="FT",M2>=1)**

- **OR Function**
  - returns a TRUE value if any of the logical conditions are true and a FALSE value if all the logical conditions are false
    OR (logical1 [,logical2]…)
    - Example: **=OR(G2="FT",M2>=1)**

- **NOT Function:**
  - Reverses the value of its argument.

- Used when you want to make sure a value is not equal to one particular value.
  - Example:**=NOT(G2="Blah")**

- **IF Function**
  - Evaluates whether a condition or a logical test is true or false and returns one value if the condition is true, and another value if the condition is false.
  - If true then value "a" else value "b"
  - IF(logical_test, value_if_true, [value_if_false])
  - Example: IF(A2="Yes",B2*C2,0)
- **Nested IF Functions**
  - **A nested IF function is when one IF function is placed inside another IF function to test an additional condition**
  - **The Number of IF functions in total would be one less than the number of possible outputs.**
  - **=IF(D2=1,2%,IF(D2=2,3%, IF(D2=3,4%,"Invalid Pay Grade")))**
  - **Read this as follows:**
    - If the value in D2 is equal to 1, then result is 2%
    - Else… If the value in D2 is equal to 2, then result is 3%
    - Else… If the value in D2 is equal to 3, then result is 4%
    - Else… result is "Invalid Pay Grade"

- **Conditional Functions:**
  - **Summary Functions: COUNTIF, SUMIF, AVERAGEIF**
    - You can calculate the number of cells in a range that match criteria you specify using the **COUNTIF** function, i.e. a conditional count: **=COUNTIF(*range,criteria)***
    - You can add the values in a range that meet criteria you specify using the **SUMIF** function, i.e. a conditional sum: **=SUMIF(*range,criteria[,sum_range])***
    - You use the **AVERAGEIF** function to calculate the average of values in a range that meet criteria you specify, i.e. a conditional average: **=AVERAGEIF(*range,criteria[,average_range])***
- **The COUNTIF, SUMIF, and AVERAGEIF functions specify only one condition to summarize the data.**

- **To summarize the data using several conditions, use:**
  - **COUNTIFS (criteria_range1,criteria1[,criteria_range2, criteria2…])**
  - **SUMIFS (sum_range,criteria_range1,criteria1[,criteria_range2, criteria2…])**
  - **AVERAGEIFS(average_range,criteria_range1,criteria1[,criteria_range2, criteria2…])**

Reporting : Regular Mass Information Dissemination in a standard format

Analytics : Slicing and Dicing with visual feedback and interactivity. Mainly for the Middle Management – for Tactical Decisions and to guide Strategic Decisions

Dashboards : Quick Distilled Snapshots, highlighting the key Indices, for Instant Decision Making

Strategic Dashboards:
- provide quick overview that decision makers need for monitoring the health and opportunities of the business
- typically derived from static snapshots of data, and focus on high-level measures of performance:
- comparisons with targets or levels
- performance indicated categorically or on scales (good/bad ; high/medium/low)

Analytical Dashboards:
- support more sophisticated data analysis by facilitating rich comparisons, and fine grained performance evaluation (drill-down ; slice-and-dice)
- used to not only see what is going on, but to examine the causes

Operational Dashboards:
- for monitoring operations in real-time
- uses simple display media to quickly identify and understand events and to ensure timely responses

What if analysis = goal seek function basically. Solver add-in used.

Access

Importance of Data and Information

- ***Data:*** raw facts that describe the characteristics of an event.
- ***Information:*** data converted into a meaningful and useful context.
- ***Database*** – maintains information about various types of objects (inventory), events (transactions), people (employees), and places (warehouses)
    - ⇨ *Example*:

Raw Data = shares exchanged in the stock market
Information = How did a particular stock perform during the last 6 month? How did it perform compare to other stocks?

QUESTION: What is an important asset of an organization?

a) Its capital
b) Its People
c) Its information
d) **All of the above**

Data Management

- Data management is a structured approach for capturing, storing, processing, integrating, distributing, securing, and archiving data effectively throughout their *life cycle*.

QUESTION: Which one of the following statements is NOT correct?

a) Transactional data captures close interactions with customers
b) Executives use transactional data to make strategic decisions
c) Transactional data are usually large, and with so many detail

- **What is the goal of data management?** To provide the infrastructure and tools to transform raw data into usable information of the highest quality.
- **Data Management Challenges:**
    - Volume of data is increasing exponentially.
    - Data is scattered throughout the organization.
    - Data is created and used offline without going through quality control checks.
    - Data may be redundant and out-of-date, creating a huge maintenance problem.

Traditional File Organization

- **File**: Group of records of same type
- **Record**: Group of related fields
- **Field**: Group of words or a complete number
- **Byte**: Group of bits that represents a single character
- **Bit**: Smallest unit of data; binary digit (0,1)
- **Problems with Traditional:**
    - **Data redundancy**: The presence of duplicate data in multiple data files so that the same data are stored in more than one place or location
    - **Data inconsistency**: The same attribute may have different values.
    - **Data Isolation, Lack of data sharing and availability:** Information cannot flow freely across different functional areas or different parts of the organization.
    - **Poor security:** Management may have no knowledge of who is accessing or making changes to the organization's data

Database Management Systems (DBMS)

- **DBMS =**
    - A software package to create and maintain databases
    - Acts as interface between application programs and physical data files
    - Separates logical and physical views of data

QUESTION: Which one of the following is NOT a problem with traditional file organization?

a) Data redundancy
b) Data inconsistency
c) Program-data dependence
d) Files are organized as records and fields
e) Lack of flexibility in reporting

- ***Entity*** – a person, place, transaction, or event about which information is stored

- ***Entity class (table)*** – a collection of similar entities
- ***Attributes (fields, columns)*** – characteristics or properties of an entity class. Columns in each table contain attributes (ex. Customer ID, Customer Name, Contact Name, Phone)

Relational DBMS

- Represents data as two-dimensional tables. Relates data across tables based on common data element
    - Examples of Relational DBMS: MS Access, DB2, Oracle, MS SQL Server
- ***Primary key*** – a field (or group of fields) that uniquely identifies a given entity in a table
- ***Foreign key*** – a primary key of one table that appears as an attribute in another table and acts to provide a logical relationship between the two tables
- **Select:** Creates subset of rows that meet specific criteria
- **Join**: Combines relational tables to provide users with information
- **Project**: Enables users to create new tables containing only relevant information
- Functions of A DBMS:
    - **Data filtering and profiling:** Inspecting the data for errors, inconsistencies, redundancies, and incomplete information.
    - **Data quality:** Correcting, standardizing, and verifying the integrity of the data.
    - **Data synchronization:** Integrating, matching, or linking data from disparate sources.
    - **Data enrichment:** Enhancing data using information from internal and external data sources.
    - **Data maintenance:** Checking and controlling data integrity over time.

QUESTION: Which of the following is NOT correct about relational database management systems?

a) Relational DBMS consists of two-dimensional tables
b) Access is an example of relational DBMS
c) Relational DBMS increases the chance of data redundancy
d) Relational DBMS generates reports combining multiple tables

Part 2

Data Quality

- Low Quality Info:
    - Data/info from different systems have different entry standards and formats
    - Call centre operators enter abbreviated/erroneous info by accident /save time
    - Third party & external info contains inconsistencies, inaccuracies, and errors
- Potential Problems from low quality info:
    - Inability to accurately track customers
    - Difficulty identifying valuable customers
    - Inability to identify selling opportunities
    - Marketing to nonexistent customers
    - Difficulty tracking revenue due to inaccurate invoices
    - Inability to build strong customer relationships
- **Info Cleansing or Scrubbing**: A process that weeds out and fixes or discards inconsistent, incorrect, or incomplete information. Standardizing Customer name from Operational Systems

Data Warehouse

- For organizational learning purposes, data from many sources and over many time periods must be gathered together & organized in a consistent and useful way
- **Data warehouse**: a copy of transaction data specifically structured for querying, analysis, reporting, and more rigorous data mining
    - Note that the data warehouse contains a copy of the transactions which are not updated or changed later by the transaction system
- **Extraction, transformation, and loading (ETL)** – a process that extracts information from internal and external databases, transforms the information using a common set of enterprise definitions, and loads the information into a data warehouse
- **Data mart** – contains a subset of data warehouse information, contains summarized or highly focused portion of data for a specified business unit or group of users

Data center

- Host and integrate networks, computer systems, and storage devices
- High reliability and availability (redundant power supplies, redundant hardware, etc.)

- High security (physical and data/software)
- Distributed Database:
    - A database that is stored in more than one physical location
    - Reduce the vulnerability of a single, massive central site
    - Increase service and responsiveness to local users
    - Can often run on smaller, less expensive computers
    - Depend on high-quality telecommunications lines
- Two **types** of Databases:
    - **Centralized database:** stores all related files in one physical location
    - **Distributed database:** copies of data at all locations
        - *Partitioned database*: Parts of database are stored in different physical locations
        - *Replicated database*: Duplicate the entire database at all remote locations

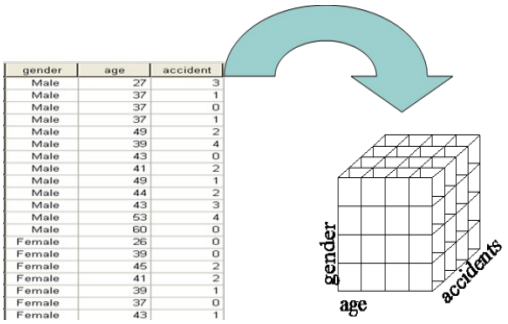
Document Management
Automated control of imaged and electronic documents, spreadsheets, emails, word processing docs, voice and other documents through their life cycle from initial creation to final archiving or destruction.

- **Document Management Systems**: Hardware and software to archive e-documents, convert paper documents to e-documents, then index and store them according to company policy.
    - **DMS** associated with printing and storing documents
        - Improving the security of the contents

Enterprise Content Management

- Includes:
- electronic document management
- Web content management
- digital asset management
- electronic records management (ERM)
- **E-discovery** is the process of gathering electronically stored information in preparation for trial, legal or regulatory investigation, or administrative action as required by law.

Part 3



Multi-dimensional Database and OLAP

- Interactive, exploratory analysis of multidimensional data from multiple dimensions/perspectives

"cube" of data, that can be seen from different prespectives

- *Functionalities*: you could drill down, slice and dice the data, aggregate
- Keeps the data in a relational data base form.

Question:
Which of the following statements is NOT true about OLAP?

a) OLAP is a one dimensional data analysis
b) OLAP database is in a form of a cube
c) IBM is one of vendors of OLAP Products
d) Slice and Dice is one of the functionalities supported by OLAP

Data Mining

- Tools for deep down analysis of large pools of data
    - To find hidden patterns
    - To predict future behavior
    - To infer rules to guide decision-making
    - When looking at customer target marketing :
        - Identifying good prospects
        - Choosing appropriate communication channels
        - Picking suitable messages
- Uses a variety of techniques to find **patterns** and **relationships** in large volumes of information and infers rules that predict future behaviour and guide decision making.

- Common forms of data-mining analysis capabilities include:
    - Classification
    - Cluster analysis
    - Association detection
- **Classification:**
    - Classes are pre-defined
    - Assign each data point to one class
        - Example Applications:
            - Grades {A, B, C, D}
            - Customer Credit Assessment {Low, Medium, High}
            - Fraud detection: {Legitimate, Fraudulent}
            - Medical treatments {Critical, NonCritical}
            - Pattern Recognition {A, B, C, …}
- **Clustering**: a technique to divide an information set into mutually exclusive groups such that the members of each group are as close together as possible to one another and the different groups are as far apart as possible.
- **Clusters** are not predefined
- **CRM** systems depend on cluster analysis to segment customer information and identify behavioral traits.
- Customer relationship management:
    - Matching campaigns to customers
    - Customer segmentation
    - Reduce exposure to credit risk
- **Association detection –** reveals the degree to which variables are related and the nature and frequency of these relationships in the information
- **Market basket analysis –** analyzes such items as Web sites and checkout scanner information to detect customers' buying behaviour & predict future behaviour by identifying affinities among customers' choices of products and services
- **Amazon.**com was one of founders of market basket analysis, with their recommendations program
- **Text Mining**:
    - Interpreting words and concepts in context
    - Discovery of trends and patterns from textual information.

1) Exploration: Word counts, creating topics
2) Pre-processing: Misspelled words, abbreviations (btw, lvm, …)
3) Modeling: Building a decision tree, neural network, etc.

Examples: Medical records, Twitter, Social networking, Negotiations, Contracts (qualitative evaluations collected from written descriptions of the projects and business plans), Classification of documents

Question 2: Floor planning in stores is an application of what?
a) clustering
b) Statistical analysis
c) Association rules
d) classification

Notes from Textbook

- Companies invest in data management to
    - Earn revenues (customer relationship management, CRM)
    - Cut expenses (inventory management)
- Databases store enterprise data
- Data warehouses aggregate this data
- Data management is structured approach to storing, analyzing, processing, securing and archiving data
- **Data lifecycle principles:**
    - **Diminishing value: *more recent, more value***
    - **90/90 data use: *90% of data is not accessed after 90 days***
    - **Data in context: *meaningful context and format for end data is necessary***
- Data visualization is presenting data in ways that are faster & easier to understand
- Master data management is integrating various data sources in a unified view
- ETL = Extract, Transform and Load data mart information
- Data quality has 5 dimensions:

    - Accuracy
    - Accessibility
    - Relevance
    - Timeliness
    - Completeness
- Data problems include: errors, duplicates, compromised, missing
- Text mining:
    - 1) exploration
    - 2) preprocessing
    - 3) categorizing and modelling