BIG DATA

Big Data: refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered (known as the "three v's" of big data). Big data often comes from data mining and arrives in multiple formats.

Big data is a great quantity of diverse information that arrives in increasing volumes and with ever-higher velocity.

Big data can be structured (often numeric, easily formatted and stored) or unstructured (more free-form, less quantifiable).

Nearly every dept in a company can utilize findings from big data analysis, but handling its clutter and noise can pose problems.

Big data can be collected from publicly shared comments on social networks and websites, voluntarily gathered from personal electr-onics and apps, through questionnaires, product purchases, and electronic check-ins.

Big data is most often stored in computer databases and is analyzed using software specifically designed to handle large, complex data sets.

Data analysts look at the relationship between different types of data, such as demographic data and purchase history, to determine whether a correlation exists. Such assessments may be done in-house or externally by a third-party that focuses on processing big data into digestible formats. Businesses often use the assessment of big data by such experts to turn it into actionable information.

Volume: Quantity of data; Size determines the value & potential insight, and if considered big data or not.
Variety: Type & Nature of data. Change from structured to semi- or unstructured challenges the technologies.
Velocity: Speed the data.Big data is often avail. in real-time.
Veracity: Completeness & Accuracy of data. Quality can vary, affecting accurate analysis.
Value: Derived from results of big data analysis.

DATA

Economic Data: Data regarding Interest rates, Asset prices, Exchange rates, and the Consumer Price Index; and other info about the global, national, or regional economy.

Structured Data: Data organized into databases with defined fields, including links between databases.

Unstructured Data: Data that is not organized into predetermined formats, such as databases, and often consists of text, images, or other nontraditional media.

Internal Data: Is owned, captured, and stored by an organization. Includes: Master data identifying customers, vendors, prospects; HR records; Employee/Customer correspondence; and Files specific to the type of business, such as Mfr's inventory records; banks' customer financial records; and insurer's premium records & rating factors.

External Data: Facts and figures available in locations outside a company. Refers to published data from outside the business.

Exploratory Data Analysis (EDA): an approach to analyzing data sets to summarize their main characteristics, often with visual methods.
A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.
Promoted to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection & experi-ments.

EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transf-ormations of variables as needed. EDA encompasses IDA.
EDA Techniques incl: Scatter Plot & Bubble Plot

Initial Data Analysis (IDA): most important distinction between the initial data analysis phase and the main analysis phase, is that during initial data analysis one refrains from any analysis that is aimed at answering the original research question.
IDA phase is guided by the following (4) questions: Quality of Data, Quality of Measurements, Initial transformation, and did the implem-entation of the study fulfill the intentions of the research design.

DATA (cont)

Text Mining: Obtains info through language recognition; more difficult than w/ other models b/c there are no organized fields & no numerical values.

Steps of the Text Mining Process:

1. Rectrieve & prepare the text.
2. Convert unstructured data into structured data.
3. Create a data mining model to help the Org. achieve its objectives.
4. Evaluate the model's effectiveness in multiple areas.

Examples of:

External & Unstructured Data: Social Media, News Reports, Internet Videos.

Internal & Structured Data: Policy Information, Claims History, Customer Data.

External & Structured Data: Telematics, Financial Data, Labor Statistics.

nternal & Unstructured Data: Adjustor's notes, Customer voice records, Surveillance videos.

CyberSecurity

Data loss occurs when valuable or sensitive information on a computer is compromised due to theft, human error, viruses, malware, or power failure. It may also occur due to physical damage or mechanical failure or equipment of an edifice.

Data loss can be caused by external factors, such as a power outage, theft, or a broad-based phishing attack. {{nl} Companies can protect themselves by using data loss prevention procedures in software and by having protocols in place for employees that enable them to safely work with and share business documents.

ESSENTIALS

Data Mining: is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining depends on effective data collection, warehousing, and computer processing.
Data mining programs break down patterns and connections in data based on what information users request or provide.

ESSENTIALS (cont)

Data Science: provides meaningful information based on large amounts of complex data or big data. Data science, or data-driven science, combines different fields of work in statistics and computation to interpret data for decision-making purposes.
Data science uses techniques such as machine learning and artificial intelligence to extract meaningful information and to predict future patterns and behaviors.

Disruptive Innovation: Disruptive Innovation refers to a technology whose application significantly affects the way a market or industry functions. An example of modern disruptive innovation is the Internet, which significantly altered the way companies did business and which negatively impacted companies that were unwilling to adapt to it.
Disruptive innovation refers to a new development that dramatically changes the way a structure or industry functions.

Sequential Pattern Mining: a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence; presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity. *Sequential pattern mining is a special case of structured data mining.*

Leaders

Electronic Commerce (e-commerce): Electronic commerce or e-commerce (sometimes written as eCommerce) is a business model that lets firms and individuals buy and sell things over the internet. E-commerce operates in all four of the following major market segments: *Business to business; * Business to consumer; *Consumer to consumer; and *Consumer to business

Statistics

Scatter Plot: A graphed cluster of dots, each of which represents the values of two variables. The slope of the points suggests the direction of the relationship between the two variables. The amount of scatter suggests the strength of the correlation.
two dimensional plot of point values

Statistics (cont)

Bubble Plot: A Scatter Plot in which the size of the bubble represents a 3rd attribute, such as average accident severity.
Best option for conveying the numerical relationship between three or four sets of values.

Correlation Matrix: A table that summarizes a series of correlations among several variables.
rectangular display of all the correlations between all pairs of data sets with a key (such as color coding) that indicates the strength of the correlation

Regression Model: Estimates relationships between or among variables.
Model uses mathematical functions of statistical regression to predict the numerical value of a target variable based on the values of the explanatory variables

Regression Analysis: A set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome/target variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features').
Primarily used for (2) conceptually distinct purposes.
First, widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning; Second, in some situations, can be used to infer causal relationships between the independent and dependent variables.

Linear Regression: Statistical method that predicts the numerical value of a target variable based on the value of one or more attributes or explanatory variables.
A linear approach to modelling the relationship between a scalar response and 1 or more explanatory variables (also known as dependent & independent variables).

Statistics (cont)

Linear Regression: Falls into 1 of 2 categories:
If the goal is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

Generalized Linear Model (GLM): Removes the normality and constant variance assumption in a linear model and it names a link funcation which defines the relationship between the expected response variable and linear combination of the predictor variables. A flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

GLM consists of (3) elements:
1. An exponential family of probability distributions.
2. A linear predictor - the quantity which incorporates the information about the independent variables into the model.
3. A link function - provides the relationship between the linear predictor and the mean of the distribution function.

Data-Driven Decision Making

Data-Driven Decision Making: gives reference to the collection and analysis of data to guide decisions that improve success.
Data-Informed Decision Making (DIDM): (2) basic approaches:Descriptive & Predictive approach.

Data-Driven Decision Making (cont)

Process for Data-driven Decision Making:

1. **Define the Problem** - provide a business context for using the data *this step is crucial because modeling and analyzing data is not effective without a business context*
2. **Prepare the Data** - Identify the necessary data; Gather quality data; Verify its quality
3. **Analyze & Model** - model the data using big data techniques. *use the appropriate descriptive or predictive approach*
4. **Develop Insights** - identify trends, relationships, behaviors, and events
5. **Make an Actionable Decision** - develop and implement a solution to the problem

FINANCIAL ANALYSIS

Data Warehousing: the electronic storage of a large amount of information by a business or organization. Data warehousing is a vital component of business intelligence that employs analytical techniques on business data.
A data warehouse is designed to run query and analysis on historical data derived from transactional sources for business intelligence and data mining purposes.

Data Analytics: Data analytics is the science of analyzing raw data in order to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption. Data analytics techniques can reveal trends and metrics that would otherwise be lost in the mass of information. This information can then be used to optimize processes to increase the overall efficiency of a business or system.

FINANCIAL ANALYSIS (cont)

Data Analytics Process: involves several different steps:
1. The first step is to determine the data requirements or how the data is grouped. Data may be separated by age, demographic, income, or gender. Data values may be numerical or be divided by category.
2. The second step in data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.
3. Once the data is collected, it must be organized so it can be analyzed. Organization may take place on a spreadsheet or other form of software that can take statistical data.
4. The data is then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete. This step helps correct any errors before it goes on to a data analyst to be analyzed.

Neural Network: A data analysis technique that operates similar to the human brain in its ability to infer rules from data patterns and construct logic to use for data analytics.
A network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes.
Form of AI that enables a computer to learn as it accumulates more data (deep learning).

Neural Network: Disadvantages: The processes for developing the rules and logic may not be transparent.
a neural network can be overtrained if it reviews data in such detail that it can not then operate in a larger framework with other types of data

3 layers of Neural Network:

1. **Input layer** - provides data for the network to analyze
2. **Hidden layer** - uses mathematical functions to learn and recode input data
3. **Output layer** - provides results of the anaysis

SOCIAL NETWORK

Social Network Analysis (Network analysis): Studies the connections and relationships among people in a social network.
Useful tool for making predictions based on trends
Social Network - group of individuals or entities who share relationships and the flow of communication

Node: Each individual or entity is know as this
a basic unit used to build data structures

SOCIAL NETWORK (cont)

Centrality Measures: In a social network context, the quantification of a node's relationship to other nodes in the same network. Determines the efficiency of the flow btwn Social Network commections.
indicators of centrality identify the most important vertices within a graph.

(3) Centrality measures:
1. **Degree** - the number of connections each node has
2. **Closeness** - the average distance or path length btwn a given node and other nodes in the network
3. **Betweenness** - how many times a given node is part of the shortest path btwn 2 other nodes in a network

Financial Technology & Automated Investing

Artificial Intelligence (AI): Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

Deep Learning: is an artificial intelligence (AI) function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabeled. Also known as deep neural learning or deep neural network.
*Deep learning AI is able to learn without human supervision, drawing from data that is both unstructured and unlabeled.; Also a form of machine learning, can be used to help detect fraud or money laundering, among other functions.

Machine Learning: Machine learning is the concept that a computer program can learn and adapt to new data without human intervention. Machine learning is a field of artificial intelligence (AI) that keeps a computer's built-in algorithms current regardless of changes in the worldwide economy.
Machine learning is useful in parsing the immense amount of information that is consistently and readily available in the world to assist in decision making.

Financial Technology & Automated Investing (cont)

Algorithm: An algorithm is set of instructions for solving a problem or accomplishing a task. One common example of an algorithm is a recipe, which consists of specific instructions for preparing a dish/meal. Every computerized device uses algorithms to perform its functions.

Disruptive Technology: Disruptive technology is an innovation that significantly alters the way that consumers, industries, or businesses operate. A disruptive technology sweeps away the systems or habits it replaces because it has attributes that are recognizably superior. Recent disruptive technology examples include e-commerce, online news sites, ride-sharing apps, and GPS systems.
A disruptive technology supersedes an older process, product, or habit.

Association Rule Learning: a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.
association rules are employed today in many application areas including Web usage mining, intrusion detection, continuous production, and bioinformatics.

Process of **Association Rule Generation** is usually split up into two separate steps:
1. A minimum support threshold is applied to find all frequent itemsets in a database.
2. A minimum confidence constraint is applied to these frequent itemsets in order to form rules.

Decision Tree Analysis: (5) Steps

Decision Tree Analysis: (5) Steps:
1. Define the problem with a statement of the decision being considered
2. Create pathways (sequence of events) for each alternative, with each pathway leading to an outcome
3. Assign a probability to each event on a pathway and estimate the value (cost or gain) of the outcome of each pathway
4. Multiply the probability of each event by the value of its outcome to determine the expected value of each pathway
5. Compare expected values to determine the pathway with the highest expected value

DECISION TREE:Analysis, Use, Features, and Inputs

Analysis: Analyzes the consequences, costs, and gains of decisions to compare alternative decisions.

Use: Decision tree analysis helps risk managers choose the best strategy to meet a goal.

Features: The process can be used to analyze both negative and positive consequences.

Inputs: The risk manager inputs the project plan with decision points and information on possible outcomes.

DECISION TREE: Outputs, Advantages, Disadvantages

Outputs: Decision tree analysis produces an analysis of risk for each pathway with options and an expected value for each pathway.

Advantages: Presents a visual portrayal, provides both quantitative and qualitative information, and offers a way to calculate the best pathway through a problem.

Disadvantages: Can be complicated and difficult to explain
also susceptible to oversimplification, which can result in less accurate decision making

EVENT TREE Analysis: (6) Steps:

1. Identify the initiating event (first accidental event that could result in unwanted consequences)
2. Determine consequences of events that could follow the accidental event
3. Construct an event tree diagram that lists barriers in the sequence that would be activated if the designated event occurred
4. Design each pathway to fork at each barrier depending on whether the barrier succeeds or fails
5. Assign an estimated probability to the likelihood of success of failure of each barrier 6. Calculate the frequency of outcomes for each pathway

EVENT TREE: Analysis, Use, and Features

Event Tree Analysis: Analyzes the consequences of accidental events rather than decisions.

Use - Risk managers use event tree analysis to evaluate risk treatment measures and identify, recommend, and justify improvements.

Features - Process typically analyzes only negative consequences.

EVENT TREE: Outputs, Process, Procedures

- List of potential problems, with estimated values for outcomes and frequencies
- Recommendations regarding the effectiveness of barriers

EVENT TREE: Adv & Disadvantages

Advantages:
- offers a visual portrayal of sequences of events following an accident
- shows the effectiveness of control systems
- provides both quantitative and qualitative information

Disadvantages:
- effective only if all potential events are identified
- analysis considers only two options (success or failure of barriers)
- analysis may ignore dependencies that arise within a sequence

CH.10 VOCAB

Economic Data: Data regarding interest rates, asset prices, exchange rates, the Consumer Price Index, and other information about the global, the national, or a regional economy.

Classification Trees: A supervised learning technique that uses a structure similar to a tree to segment data according to known attributes to determine the value of a categorical target variable.

Cluster Analysis: A model that determines previously unknow groupings of data.

Data Mining: The process of extracting hidden patterns from data that is used in a wide variety of applications for research and fraud detection.

Centrality Measures: In a social network context, the quantification of a node's relationship to other nodes in the same network.

Big Data: Sets of data that are too large to be gathered and analyzed by traditional methods.

Data Science: An interdisciplinary field involving the design and the use of techniques to process very large amounts of data from a variety of courses to provide knowledge based on data.