Data Analytics Tools & Technique

DA 6223

The University of Texas at San Antonio

Welcome to DA 6223

Instructor: Min Wang

Email: min.wang3@utsa.edu

Office: BB 4.03.20
Phone: (210) 458-5381

Class Meetings: Fully online in an asynchronous format

Office Hours: 9:00am – 10:30am or By Appointment via the Zoom.


BIG
DATA
INFORMATION
TECHNOLOGIES
MANAGEMENT
PETABYTES
ANALYSIS
RESEARCH
SEARCH
INTERNET
LEARNING
STORAGE
LARGE
SETS
MILLION
SOFTWARE
SAN
NAS
VOLUME
PROCESS
GOVERNMENT
FUTURE
SCIENCE
TECHNOLOGY
DEVELOPMENT
BILLION
NEW
PARALLEL
CRITIQUES
MAPREDUCE
INITIATIVE
TERABYTES
ECONOMIC
DISTRIBUTED
PROJECT
ANNUAL
RELEVANT
USED
TIME
SENSOR
MARKET
VARIETY
RELATED
PEOPLE
TRAFFIC
STATE
CRITIQUE
CURRENTLY
WOULD
NOW
USE
FLOW
NATIONAL
INTELLIGENCE
SECOND
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
EXABYTES
SIZE
DATABASES
PROCESSING
ANALYTICS
SIMULATIONS
FRAMEWORK
BUSINESS
BETWEEN
DEFINITION
STATISTICS
VISUALIZATION
EFFECTIVE
USING
MASSIVE
FACTOR
ONLINE
MANAGE
DAY
WORLDWIDE
SYSTEMS
MAKING
INTEGRATION
RESULTS
SOCIAL
YEARS
DRIVE
REPLICATION
HANDLE
SOURCES
CHALLENGES
LIMITS
COMPLEX
ANNOUNCED
PAST
HAND
CAPACITY
ONE
OUT
INCLUDE
NETWORKS
HYPOTHESIS
END
MOST
NEARLY
APPROACH
MANY
SET
DEPARTMENT
QUERIES
HIGH
USERS
AMOUNT
RATE
REAL
STRUCTURE
WORK
APPLICATIONS
INFRASTRUCTURE
HIGHER
ALGORITHMS
COMPANIES
DETERMINE
WELL
APPROACHES
PROCESSES
PARADIGM
RECORDS
COST
FUNDING
INSIGHT
TOOLS
CENTER
PER
PRIVATE
TERMS
YEAR
CAPABILITIES
SUCCESSFUL
NEEDED
USES
COMPARED
SHARED
LESS
HUGE
EVERY
WORLD
VARIOUS
BASED
TIMES
ARCHITECTURE

Data Analytics Tools & Technique

Chapter 1: Introduction to Big Data

DA 6223

The University of Texas at San Antonio

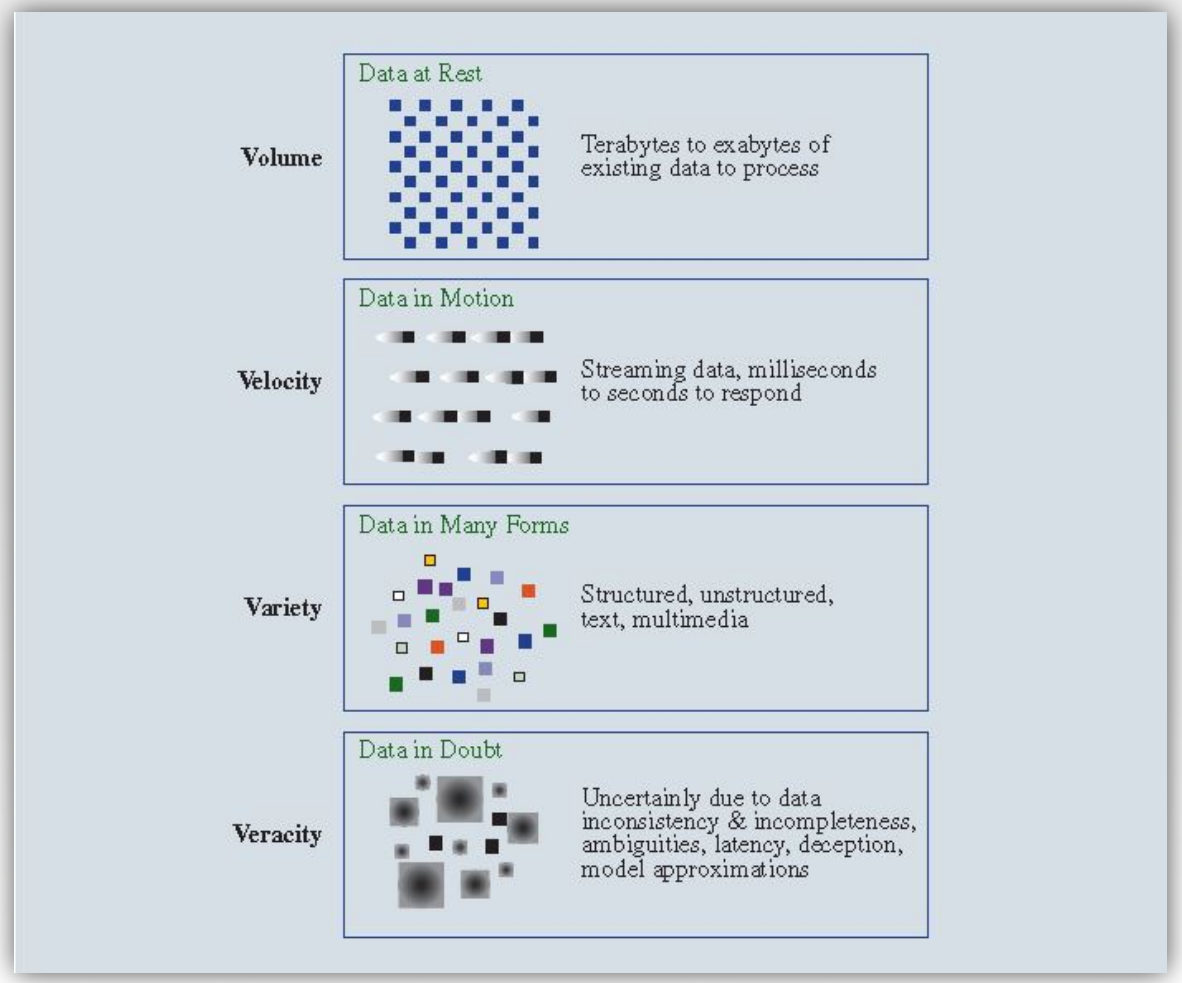
Big Data

Big data: A set of data that cannot be managed, processed, or analyzed with commonly available software in a reasonable amount of time

- Represents opportunities;
- Presents challenges in terms of data storage and processing, security, and available analytical talent;
- More companies are hiring **data scientists** who know how to process and analyze massive amounts of data;

The 4 Vs of Big Data

- Big data are characterized by
  - **V**olume (a large amount of data)
  - **V**elocity (fast collection and processing)
  - **V**ariety (could include nontraditional data such as video, audio, and text)
  - **V**eracity (uncertainty due to various reasons)



New Technologies

- The four Vs have led to new technologies
  - **Hadoop**: An open-source programming environment that supports big data processing through distributed storage and processing over multiple computers.
  - **MapReduce**: A programming model used within Hadoop that performs two major steps: the map step and the reduce step.
- **Data security**: The protection of stored data from destructive forces or unauthorized users

Objectives

- Define business analytics.
- Explain the proliferation of data and how this impacts the need for good analytics.
- Identify scope of business analytics.
- Data for business analytics.
- Problem solving and decision making.
- Name some applications where analytics are ***not*** helpful.
- Explain some of the common pitfalls of analytical practice.

What is Business Analytics?

Analytics is the use of:

data,

information technology,

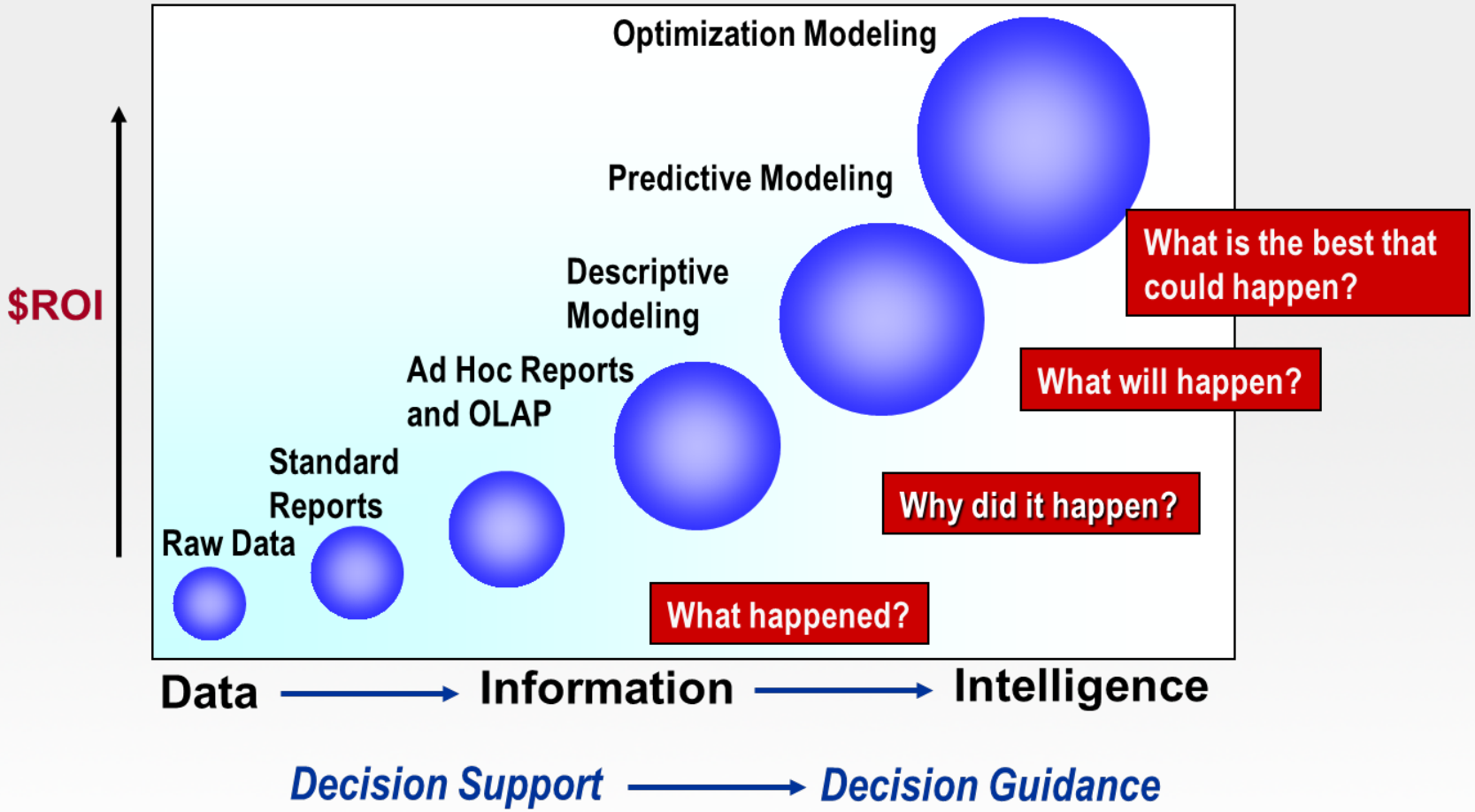
statistical analysis,

quantitative methods, and

mathematical or computer-based models

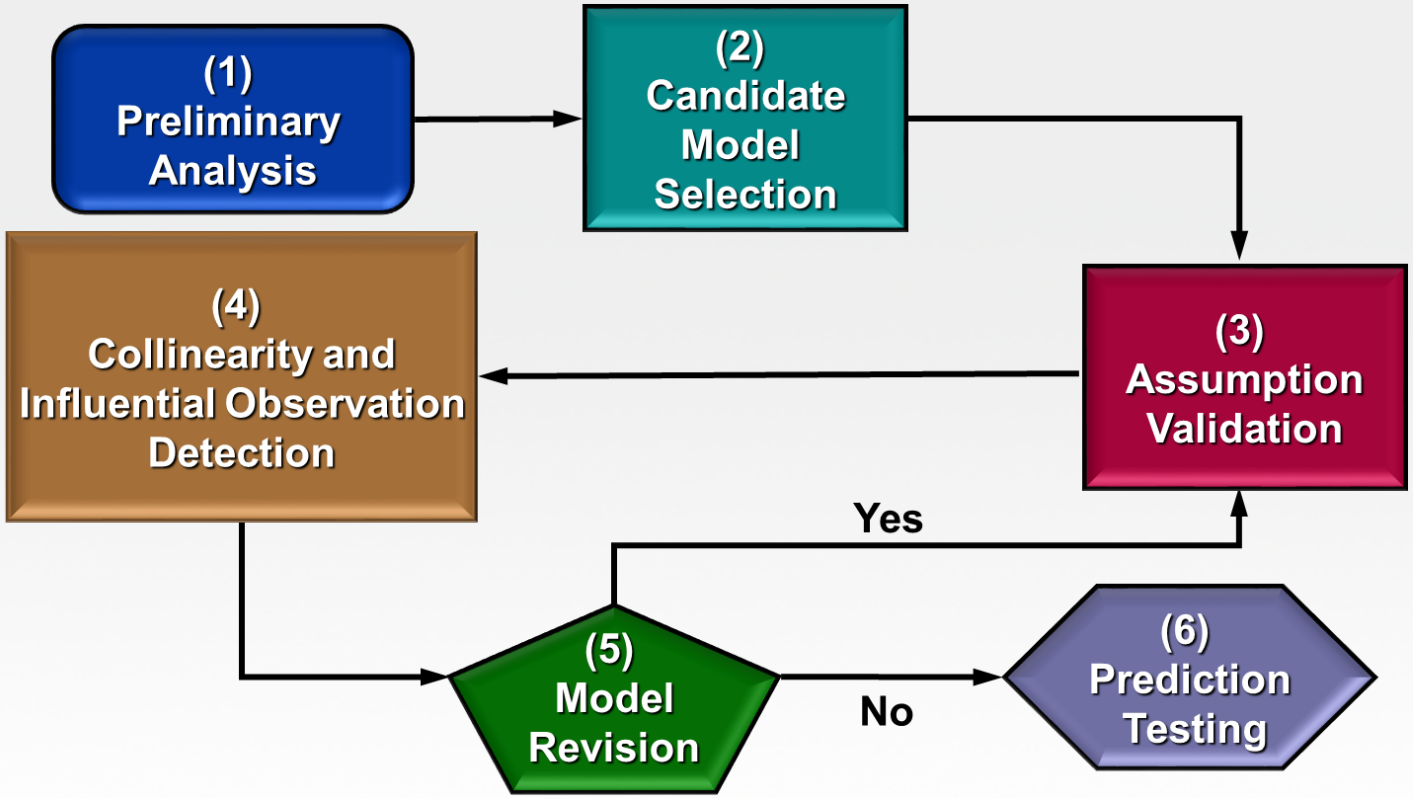
to help managers gain improved insight about their business operations and make better, fact-based decisions.

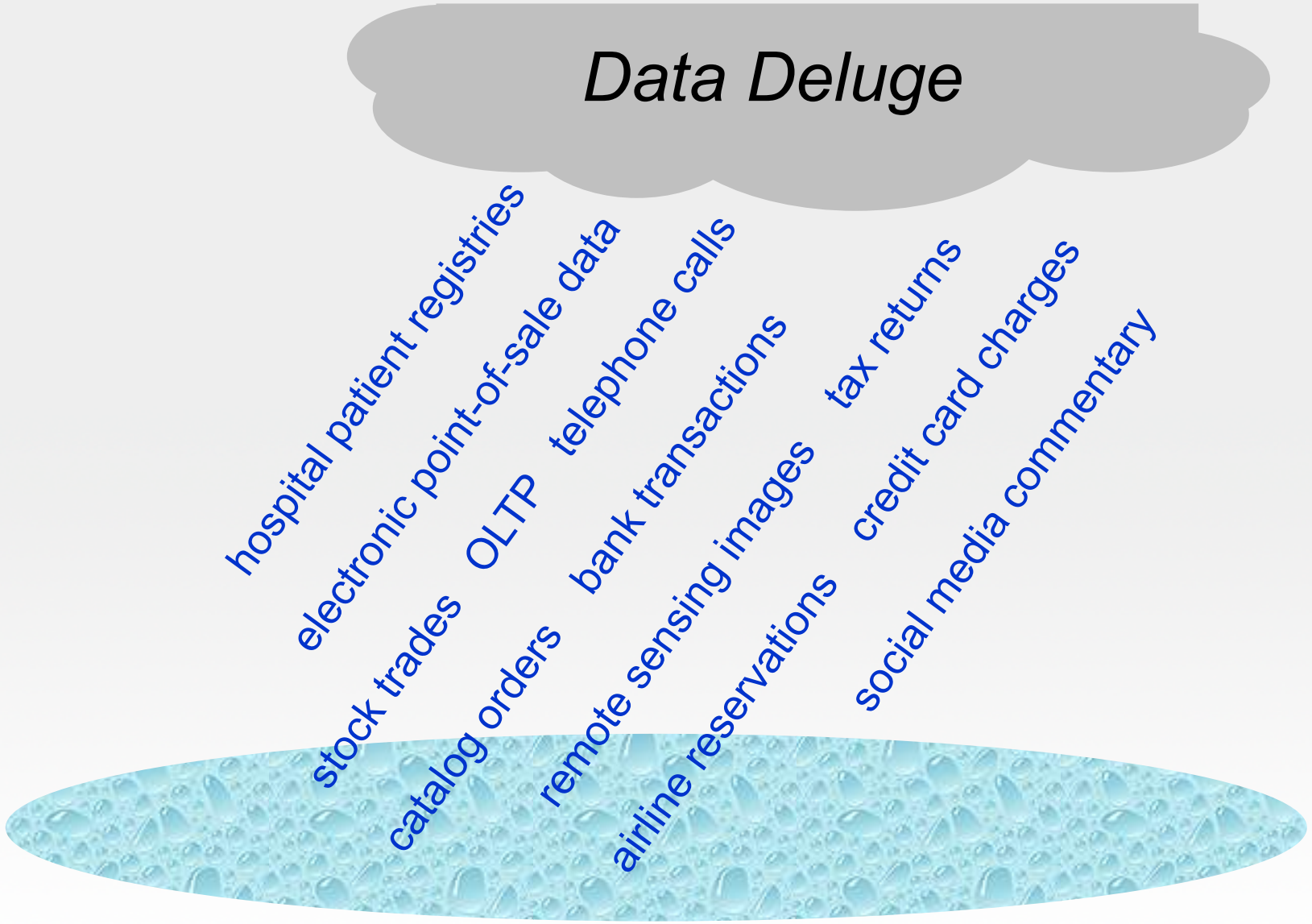
Overview of Statistical Data Analytics



Overview of Statistical Data Analytics

- An Effective Modeling Cycle




Data Deluge
hospital patient registries
electronic point-of-sale data
telephone calls
OLTP
bank transactions
stock trades
tax returns
catalog orders
remote sensing images
credit card charges
airline reservations
social media commentary

The Business Analytics Challenge

- How do we get anything useful out of tons and tons of data?

Scope of Business Analytics

- Descriptive analytics
  - uses data to understand past and present
- Predictive analytics
  - analyzes past performance
- Prescriptive analytics
  - uses optimization techniques

Scope of Business Analytics

Example 1 Retail Markdown Decisions

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

Scope of Business Analytics

Example 2: Harrah's - Caesars Entertainment

- Harrah's owns numerous hotels and casinos
- Uses analytics to:
  - forecast demand for rooms
  - segment customers by gaming activities
- Uses prescriptive models to:
  - set room rates
  - allocate rooms
  - offer perks and rewards to customers

Data for Business Analytics

- Data
  - collected facts and figures
- Database
  - collection of computer files containing data
- Information
  - comes from analyzing data

Examples of Using Data in Business

- Annual reports
- Accounting audits
- Financial profitability analysis
- Economic trends
- Marketing research
- Operations management performance
- Human resource measurements

Data for Business Analytics

- Metrics are used to quantify performance.
- Measures are numerical values of metrics.
- Discrete metrics involve counting
  - on time or not on time
  - number or proportion of on time deliveries
- Continuous metrics are measured on a continuum
- - delivery time
- - package weight
- - purchase price

A Sales Transaction Database File

	A	B	C	D	E	F	G	H
1	**Sales Transactions: July 14**							
2								
3	**Cust ID**	**Region**	**Payment**	**Transaction Code**	**Source**	**Amount**	**Product**	**Time Of Day**
4	10001	East	Paypal	93816545	Web	$20.19	DVD	22:19
5	10002	West	Credit	74083490	Web	$17.85	DVD	13:27
6	10003	North	Credit	64942368	Web	$23.98	DVD	14:27
7	10004	West	Paypal	70560957	Email	$23.51	Book	15:38
8	10005	South	Credit	35208817	Web	$15.33	Book	15:21
9	10006	West	Paypal	20978903	Email	$17.30	DVD	13:11
10	10007	East	Credit	80103311	Web	$177.72	Book	21:59
11	10008	West	Credit	14132683	Web	$21.76	Book	4:04
12	10009	West	Paypal	40128225	Web	$15.92	DVD	19:35
13	10010	South	Paypal	49073721	Web	$23.39	DVD	13:26

Records

Entities

Fields or Attributes

Four Types of Data Based on Measurement Scale

- Nominal: no order, distance or origin; determination of equality *(Male vs. Female)*
- Ordinal: order but no distance or unique origin; determination of greater or lesser values *(Rank)*
- Interval: order and distance but no unique origin; determination of equality of intervals or differences *(Calendar, Time)*
- Ratio: order, distance and unique origin; determination of equality of ratios *(weight, height, distance, money)*

Determine the level of measurement of the variable.

A) nominal B) ratio C) ordinal D) interval

1. A the musical instrument played by a music student;
2. C the medal received (gold, silver, bronze) by an Olympic gymnast;
3. B height of a tree;
4. A the native language of a tourist;
5. D the day of the month;
6. C an officer's rank in the military.

Classifying Data Elements in a Purchasing Database

	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	Spacetime Technologies	A0111	6489	O-Ring	$ 3.00	900	$ 2,700.00	25	10/10/11	10/18/11
5	Steelpin Inc.	A0115	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
6	Steelpin Inc.	A0123	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11
7	Steelpin Inc.	A0204	5319	Shielded Cable/ft.	$ 1.10	16,500	$ 18,150.00	30	09/15/11	10/05/11
8	Steelpin Inc.	A0205	5677	Side Panel	$195.00	120	$ 23,400.00	30	11/02/11	11/13/11
9	Steelpin Inc.	A0207	4312	Bolt-nut package	$ 3.75	4,200	$ 15,750.00	30	09/01/11	09/10/11
10	Alum Sheeting	A0223	4224	Bolt-nut package	$ 3.95	4,500	$ 17,775.00	30	10/15/11	10/20/11
11	Alum Sheeting	A0433	5417	Control Panel	$255.00	500	$ 127,500.00	30	10/20/11	10/27/11
12	Alum Sheeting	A0443	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
13	Alum Sheeting	A0446	5417	Control Panel	$255.00	406	$ 103,530.00	30	09/01/11	09/10/11
14	Spacetime Technologies	A0533	9752	Gasket	$ 4.05	1,500	$ 6,075.00	25	09/20/11	09/25/11
15	Spacetime Technologies	A0555	6489	O-Ring	$ 3.00	1,100	$ 3,300.00	25	10/05/11	10/10/11

Classifying Data Elements in a Purchasing Database

	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	Spacetime Technologies	A0111	6489	O-Ring	$ 3.00	900	$ 2,700.00	25	10/10/11	10/18/11
5	Steelpin Inc.	A0115	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
6	Steelpin Inc.	A0123	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11
7	Steelpin Inc.	A0204	5319	Shielded Cable/ft.	$ 1.10	16,500	$ 18,150.00	30	09/15/11	10/05/11
8	Steelpin Inc.	A0205	5677	Side Panel	$195.00	120	$ 23,400.00	30	11/02/11	11/13/11
9	Steelpin Inc.	A0207	4312	Bolt-nut package	$ 3.75	4,200	$ 15,750.00	30	09/01/11	09/10/11
10	Alum Sheeting	A0223	4224	Bolt-nut package	$ 3.95	4,500	$ 17,775.00	30	10/15/11	10/20/11
11	Alum Sheeting	A0433	5417	Control Panel	$255.00	500	$ 127,500.00	30	10/20/11	10/27/11
12	Alum Sheeting	A0443	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
13	Alum Sheeting	A0446	5417	Control Panel	$255.00	406	$ 103,530.00	30	09/01/11	09/10/11
14	Spacetime Technologies	A0533	9752	Gasket	$ 4.05	1,500	$ 6,075.00	25	09/20/11	09/25/11
15	Spacetime Technologies	A0555	6489	O-Ring	$ 3.00	1,100	$ 3,300.00	25	10/05/11	10/10/11

Categorical
Categorical
Categorical
Categorical
Ratio
Ratio
Ratio
Ratio
Interval
Interval

Problem Solving and Decision Making

- BA represents only a portion of the overall problem solving and decision making process.
- Six steps in the problem solving process
  1. Recognizing the problem
  2. Defining the problem
  3. Structuring the problem
  4. Analyzing the problem
  5. Interpreting results and making a decision
  6. Implementing the solution

Problem Solving and Decision Making

1. Recognizing the 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.

Problem Solving and Decision Making

2. Defining the Problem

- Clearly defining the problem is not a trivial task.
- Complexity increases when the following occur:
  - large number of courses of action
  - several competing objectives
  - external groups are affected
  - problem owner and problem solver are not the same person
  - time constraints exist

Problem Solving and Decision Making

3. Structuring the Problem

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

Problem Solving and Decision Making

4. Analyzing the Problem
   - Identifying and applying appropriate Business Analytics techniques
   - Typically involves experimentation, statistical analysis, or a solution process

Much of this course is devoted to learning BA techniques for use in Step 4.

Problem Solving and Decision Making

5. Interpreting Results and Making a Decision
   - Managers interpret the results from the analysis phase.
   - Incorporate subjective judgment as needed.
   - Understand limitations and model assumptions.
   - Make a decision utilizing the above information.

Problem Solving and Decision Making

6. Implementing the Solution
   - Translate the results of the model back to the real world.
   - Make the solution work in the organization by providing adequate training and resources.

Problem Solving and Decision Making

- Analytics in Practice:
  - Will analytics solve the problem?
  - Can they leverage an existing solution?
  - Is a decision model really needed?
- Guidelines for successful implementation:
  - Use prototyping.
  - Build insight, not black boxes.
  - Remove unneeded complexity.
  - Partner with end users in discovery and design.
  - Develop an analytic champion.

The Methodology: What We Learned Not to Do

- Prediction is more important than inference.
  - Metrics are used "because they work," not based on theory.
  - *p*-values are rough guides rather than firm decision cutoffs.
  - Interpretation of a model might be irrelevant.
  - The preliminary value of a model is determined by its ability to predict a holdout sample.
  - The long-term value of a model is determined by its ability to continue to perform well on new data over time.
  - Models are retired as customer behavior shifts, market trends emerge, and so on.

Using Analytics Intelligently

- Intelligent use of analytics results in the following:
  - better understanding of how technological, economic, and marketplace shifts affect business performance
  - ability to ***consistently and reliably*** distinguish between effective and ineffective interventions
  - efficient use of assets, reduced waste in supplies, and better management of time and resources
  - risk reduction via ***measurable*** outcomes and ***reproducible*** findings
  - early detection of market trends hidden in massive data
  - continuous improvement in decision making over time

Simple Reporting

- **Examples**: OLAP (Online Analytical Processing), QC (Quality Control), descriptive statistics, extrapolation.
- ***Answer questions such as***
  - Where are my key indicators now?
  - Where were my key indicators last week?
  - Is the current process behaving like normal?
  - What is likely to happen tomorrow?

Proactive Analytical Investigation

- **Examples**: inferential statistics, experimentation, empirical validation, forecasting, optimization
- ***Answer questions such as***
  - What does a change in the market mean for my targets?
  - What do other factors tell me about what I can expect from my target?
  - What is the best combination of factors to give me the most efficient use of resources and maximum profitability?
  - What is the highest price the market will tolerate?
  - What will happen in six months if I do nothing?
  - What if I implement an alternative strategy?

- Many companies have data that they do not use or that is used by third parties. These third parties might even resell the data and any derived metrics back to the original company!

- **Example**: retail grocery POS card

Every Little Bit…

- Taking an analytical approach to only a few key business problems with reliable metrics → tangible benefit.
- The benefits and savings derived from early analytical successes → managerial support for further analytical efforts.
- **Everyone has data.**
- **Analytics can connect data to smart decisions.**
- **Proactively analytical companies outpace competition.**

Areas Where Analytics Are Often Used

- New customer acquisition
- Customer loyalty
- Cross-sell / up-sell
- Pricing tolerance
- Supply optimization
- Staffing optimization
- Product placement
- Churn
- Insurance rate setting
- Fraud detection
- …

Which residents in a ZIP code should receive a coupon in the mail for a new store location?

Areas Where Analytics Are Often Used

- Customer loyalty
- …

What advertising strategy best elicits positive sentiment toward the brand?

Areas Where Analytics Are Often Used

- New customer acquisition
- Customer loyalty
- Cross-sell / up-sell
- Pricing tolerance
- Supply optimization
- Staffing optimization
- Price optimization
- Product placement
- Churn
- Insurance rate setting
- Fraud detection
- …

What is the best next product for this customer? What other product is this customer likely to purchase?

Areas Where Analytics Are Often Used

- New customer acquisition
- Customer loyalty
- Cross-sell / up-sell
- Pricing tolerance
- Supply optimization
- Staffing optimization
- Inventory management
- Product placement
- Churn
- Insurance rate setting
- Fraud detection
- …

What is the highest price that the market will bear without substantial loss of demand?

Areas Where Analytics Are Often Used

- New customer acquisition
- Customer loyalty
- Cross-sell / up-sell
- Pricing tolerance
- Supply optimization
- Staffing optimization
- Inventory management
- Product placement
- Churn
- Insurance rate setting
- Fraud detection
- …

How many 60-inch HDTVs should be in stock? (Too many is expensive; too few is lost revenue.)

Areas Where Analytics Are Often Used

- New customer acquisition
- Customer loyalty
- Cross-sell / up-sell
- Pricing tolerance
- Supply optimization
- Staffing optimization
- Inventory optimization
- Product placement
- Churn
- Insurance rate setting
- Fraud detection
- …

What are the best times and best days to have technical experts on the showroom floor?

Areas Where Analytics Are Often Used

- New customer acquisition
- Customer loyalty
- Cross-sell / up-sell
- Pricing tolerance
- Supply optimization
- Staffing optimization
- Financial forecasting
- Product placement
- Churn
- Insurance rate setting
- Fraud detection
- …

What weekly revenue increase can be expected after the Mother's Day sale?

Areas Where Analytics Are Often Used

- New customer acquisition
- Customer loyalty
- Cross-sell / up-sell
- Pricing tolerance
- Supply optimization
- Staffing optimization
- Product placement
- Churn
- Insurance rate setting
- Fraud detection
- …

Will oatmeal sell better near granola bars or near baby food?

Areas Where Analytics Are Often Used

- Churn
- …

Which customers are most likely to switch to a different wireless provider in the next six months?

Areas Where Analytics Are Often Used

- New customer acquisition
- Customer loyalty
- Cross-sell / up-sell
- Pricing tolerance
- Supply optimization
- Staffing optimization
- Product placement
- Churn
- Insurance rate setting
- Fraud detection
- …

How likely is it that this individual will have a claim?

Areas Where Analytics Are Often Used

- New customer acquisition
- Customer loyalty
- Cross-sell / up-sell
- Pricing tolerance
- Supply optimization
- Staffing optimization
- Product placement
- Churn
- Insurance rate setting
- Fraud detection
- …

How can I identify a fraudulent purchase?

When Analytics Are *Not* Helpful

- Snap decisions required

Deciding when to run from danger

- Novel approach (no previous data possible)
- Most salient factors are rare (making decisions to work around unlikely obstacles or miracles)
- Metrics are inappropriate
- Naïve implementation of analytics
- Confirming what you already know

When Analytics Are *Not* Helpful

- Snap decisions required
- Novel approach (no previous data possible)
- Most salient factors are rare (making decisions to work around unlikely obstacles or miracles)
- Metrics are inappropriate
- Naïve implementation of analytics
- Confirming what you already know

Predicting the adoption of a new technology

When Analytics Are *Not* Helpful

- Most salient factors are rare (making decisions to work around unlikely obstacles or miracles)

Planning contingencies for employees winning the lottery

When Analytics Are *Not* Helpful

- Snap decisions required
- Novel approach (no previous data possible)
- Most salient factors are rare (making decisions to work around unlikely obstacles or miracles)
- Expert analysis suggests a particular path
- Metrics are inappropriate
- Naïve implementation of analytics
- Confirming what you already know

The seasoned art critic can recognize a fake

When Analytics Are *Not* Helpful

- Snap decisions required
- Novel approach (no previous data possible)
- Most salient factors are rare (making decisions to work around unlikely obstacles or miracles)

- Metrics are inappropriate
- Naïve implementation of analytics
- Confirming what you already know

Predicting athletes' salaries or quantifying love

When Analytics Are *Not* Helpful

- Snap decisions required
- Novel approach (no previous data possible)
- Most salient factors are rare (making decisions to work around unlikely obstacles or miracles)
- Metrics are inappropriate
- Naïve implementation of analytics
- Confirming what you already know

Only looking at one variable at a time

When Analytics Are *Not* Helpful

- Snap-decisions required
- Novel approach (no previous data possible)
- Most salient factors are rare (making decisions to work around unlikely obstacles or miracles)
- Metrics are inappropriate
- Naïve implementation of analytics

- Confirming what you already know

Ignoring variables that might be important

Expectations Leading the Analysis

- Even sophisticated analytics are not immune to personal bias such as the following:
  - selectively fitting models with variables because they place someone's opinion or agenda in a positive light
  - ignoring information that might disprove a hypothesis.
- Personal bias in model fitting, whether intentional or otherwise, can diminish the usefulness of your analytical efforts.

Trustworthy Analytics

- Let the data guide your conclusions.
- Ask the following questions:
  - Are my assumptions about the causes of my data patterns warranted?
  - Should I try something different?
- Assign a cynic to the analytical team whose purpose is to question the assumptions.
  - What would my critic say is the flaw with my analysis?
  - Investigate the data in such a way that a critic's concerns can be ruled out.