



IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

529 Advanced Data Analytics

August 23/25, 2016

Week 1 Presentation

Data Analytics 529

Course Outline

This course assumes that the student has taken the 527 Introduction to Data Analytics course, has the ability to analyze data sets in Excel, SAS, and R. Also, the student is able to summarize analysis and information in a presentation. The course assumes intermediate knowledge of SQL and data modeling to run queries and process data sets. The course will introduce advanced topics in data analysis and implementation methodologies built upon the 527 course. The course will additionally cover financial services analytic topics. Lastly, the course will further introduce analytic topics in Big Data and its applications.

Upon completion of the course, the student will be able to:

- ◆ Understand advanced data analysis and data management concepts, theories, and implementation methodologies e.g., regression analysis and forecasting
- ◆ Source, process, and model data sets for predictive analysis
- ◆ Use tools and technologies available for data analysis e.g., Excel, SAS, R
- ◆ Perform analysis and summarize findings in a presentation

Data Analytics 529

Class Exercises & Readings

Class exercises, assigned as needed, are due ***one week*** from assigned date. The assignment will be posted in Blackboard and submissions will be collected also via Blackboard.

You will not be able to submit late. No exceptions.

Class readings will be assigned on a weekly basis, as needed. This is an important part of class preparation and augments required and optional book assignments. The expectation is that all required readings and review of weekly presentations are done prior to class.

Data Analytics 529

Midterm and Finals

Midterm grades will be assigned after completion of Week 8 class assignments.

Final grades will be posted during Finals week.

Midterm and Final projects will consist of a presentation and supporting analysis work. Details and deadline for project submissions will be discussed in class.

ITM - 529

- ◆ **A** *Outstanding work reflecting substantial effort: 90-100%*
- ◆ **B** *Adequate work fully meeting that expected of a graduate student: 80-89.99%*
- ◆ **C** *Weak but marginally satisfactory work not fully meeting expectations: 65-79.99%*
- ◆ **E** *Unsatisfactory work: 0-64.99%*
- ◆ *No Exceptions!*

◆	Midterm Project	60%
◆	Class Exercises & Participation	40%

◆	Midterm Project	30%
◆	Final Project	40%
◆	Class Exercises & Participation	30%

Data Analytics 529

Weekly Schedule

Session	Date	Topic	Reading
1	August 23/25		Week 1 topics: Course Overview & Basics
2	Aug/Sept 30/1		Week 2 topics: SAS and R Statistics
3	September 6/8		Week 3 topics: Regression Analysis and Forecasting I
4	September 13/15		Week 4 topics: Regression Analysis and Forecasting II
5	September 20/22		Week 5 topics: Regression Analysis and Forecasting III
6	September 27/29		Week 6 topics: Midterm project workshops
7	October 4/6		No classes – Prepare for Midterm submission
8	October 11/13		Week 8 topics: Midterm - Submission / Optional Presentations
9	October 20		Week 9 topics: (No classes on Oct 18) FS Analytics I
10	October 25/27		Week 10 topics: Financial Services Analytics II
11	November 1/3		Week 11 topics: Big Data Analytics I
12	November 8/10		Week 12 topics: Big Data Analytics II
13	November 15/17		Week 13 topics: Big Data Analytics III
14	November 22		Week 14 topics: Final Project Workshops
15	November 29/1		Week 15 topics: Final Project - Optional Presentations
Final	December 5		Final Project Submission

Data Analytics 529

Project Logistics

- ◆ Class meets in Hermann Hall Room 003 on Tuesdays and Thursdays.
- ◆ Class time is 10:00 PM ~ 11:15 PM. There will be limited time for questions after class.
- ◆ Please use office hours or reach me via contact information below. Office hours are by appointment only. I will confirm a room in Perlstein Hall or Stuart Building for office hours.
- ◆ Course material communication will be done through Blackboard e.g., class materials, assignment submission, etc.
- ◆ Other means of contact:
 - Email: Best means of communication is email. I will respond within 24 hours.
Email: sshin17@iit.edu
- ◆ There will be weekly presentations for the class.
- ◆ Notifications will be sent for postings in Blackboard but please check regularly for updates.

Data Analytics 529

Course Books

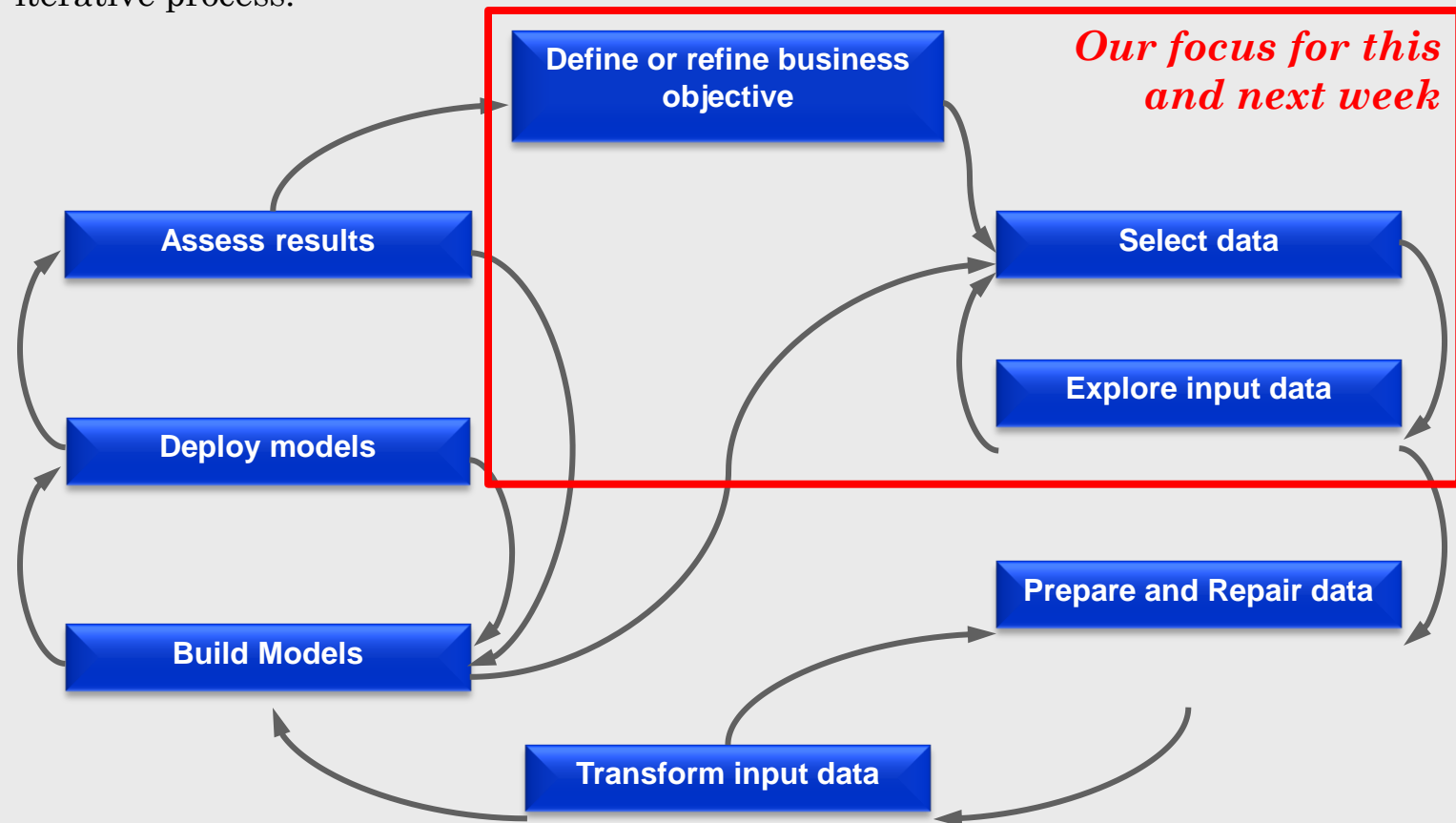
◆ There are no required books for the class.

◆ However, I list a few *optional* references:

1. SAS and R by Ken Kleinman, Nicholas J Horton. ISBN-13: 978-1420070576 ISBN-10: 1420070576
2. (Optional Reference) Data Smart: Using Data Science to Transform Information into Insight 1st Edition by John W. Foreman. ISBN-13: 978-1118661468. ISBN-10: 111866146X
3. (Optional Reference) Data Mining: The Textbook by Charu Aggarwal. ISBN-13: 978-3319141411. ISBN-10: 3319141414
4. (Optional Reference) Microsoft Excel 2013 Data Analysis and Business Modeling 1st Edition by Wayne Winston. ISBN-13: 978-0735669130. ISBN-10: 0735669139
5. (Optional Reference) Data Science for Business: What you need to know about data mining and data-analytic thinking 1st Edition by Foster Provost and Tom Tawcett. ISBN-13: 978-1449361327. ISBN-10: 1449361323

ITM - 529

These steps clarify the purpose and implementation of analytics. Note that this is an iterative process.



Week 1 Topic:

Step 1: Develop a business scenario

Example Business Scenario:

At a large university the administration wants to increase the retention rate of new freshmen who enroll. The Office of Institutional Research is given the task of making recommendations for addressing this issue. They made the decision to build models that would predict those students who were most likely not going to return the following semester. They decided to first build a model to predict those new freshmen students who enrolled in the Fall of 2012 and continued to the Spring of 2013. The objective of this analysis is to build a classification model quickly and easily to measure the propensity for these students to not enroll in the Fall of 2014. This will enable intervention with these identified students.

Week 1 Topic:

Step 2: Develop business objectives

Example Business Objective:

Improve the retention rate of new freshmen who enroll in the university. This will be addressed by building retention models to identify those students who are most likely not going to return the following semester. The first model will be for scoring those new freshmen students who entered in the Fall of 2012 and are currently enrolled in the Spring of 2013. Students in this group who are already receiving special services, such as athletes, will be removed from this study.

Week 1 Topic:

Step 3: Identify and Select Data

Example Data Selection:

- *The data to build the model will be information about those new freshmen students who enrolled in the Fall of 2011 and continued to the Spring of 2012.*
- *The target variable will be defined by examining students and determining who enrolled and those who did not enroll in the Fall of 2012. The target variable will take on a value of 1 for those that did not enroll and a value of 0 for those that did enroll.*
- *The project team will then identify inputs that they feel will be good predictors. These inputs include demographic, financial aid, student life, fall semester statistics, admissions, and other data.*

Week 1 Topic:

Step 4: Example Data Sets

Example Data Sets:

- *Course Data*
- *Demographic Data*
- *Financial Aid Data*
- *Alumni Data*
- *Student Admissions Data*
- *HR Data*
- *Some Calculated Inputs*

Week 1 Topic:

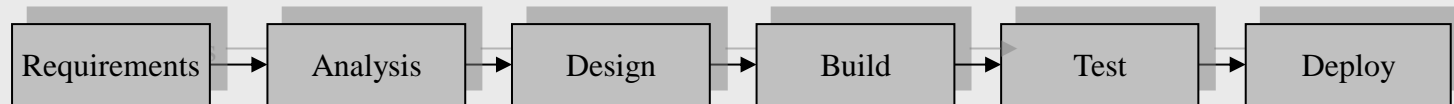
Planning: Iterative Approach

- Iterative Delivery Process is based on the iterative yet very rigorous and disciplined approach to the entire delivery lifecycle
- Its main objective is to accelerate and maximize achieving business value, minimize delivery risk
- It provides a very flexible implementation roadmap that could be adapted to changing business goals and objectives
- It offers an advantage of a very uniform distribution of resource needs throughout the entire project lifecycle.
- It virtually eliminates the needs for extensive user training, post-delivery transition and additional effort to create maintenance and support organization
- All phases of the delivery lifecycle from requirements to deployment are synchronized and based on more frequent and relatively smaller releases.
- They are supported by also iterative approach to prototyping, considered here to be a critical component of an effective delivery strategy

Week 1 Topic:

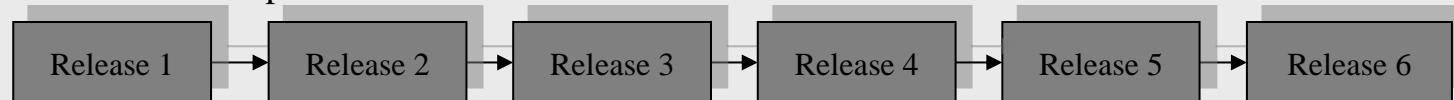
Planning: Waterfall vs Iterative

Waterfall Development Method



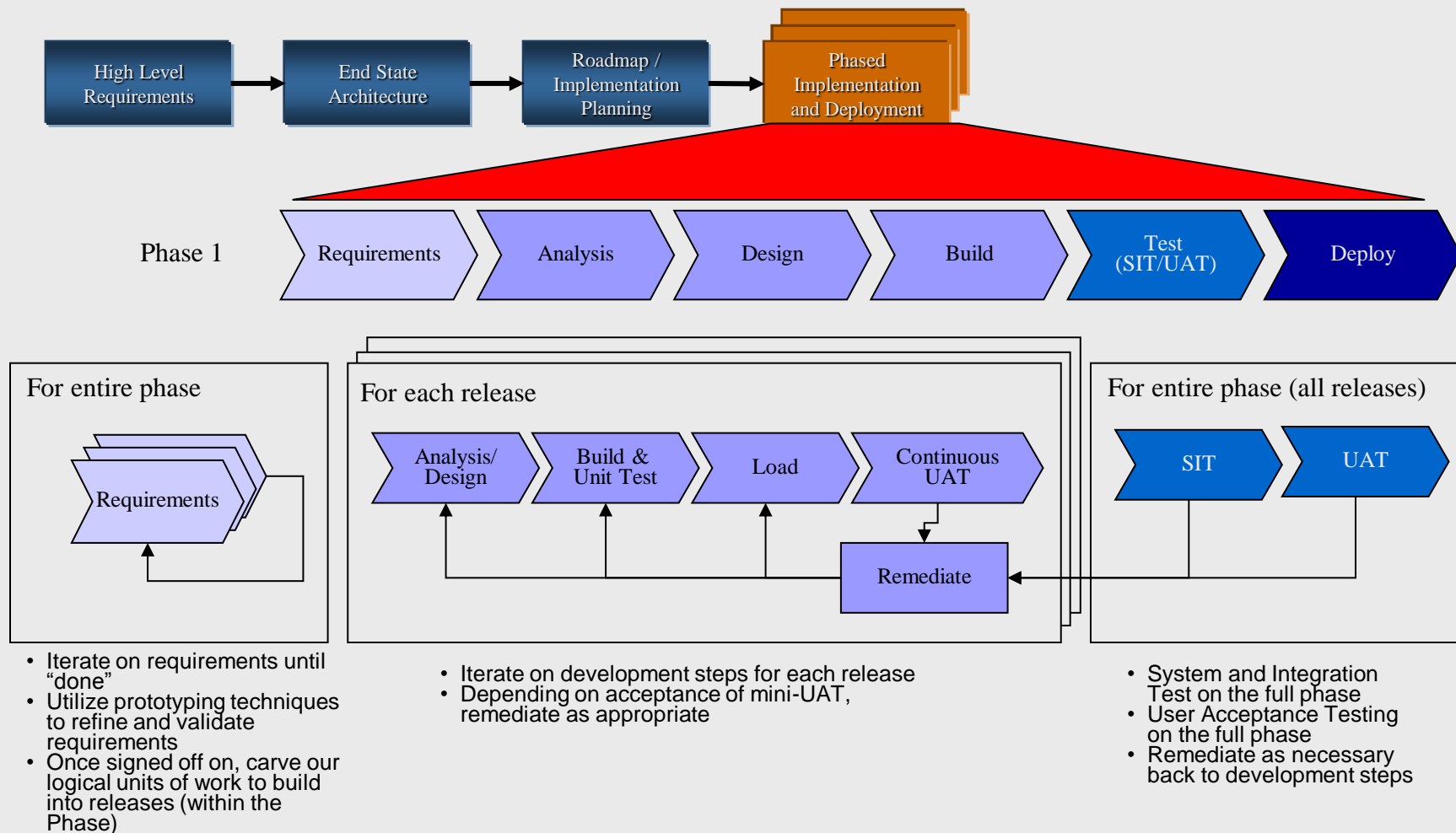
- Long delivery cycle
- Business value achieved at the very end
- Costly and complicated design, development and testing
- Higher overall risk

Iterative Development Process

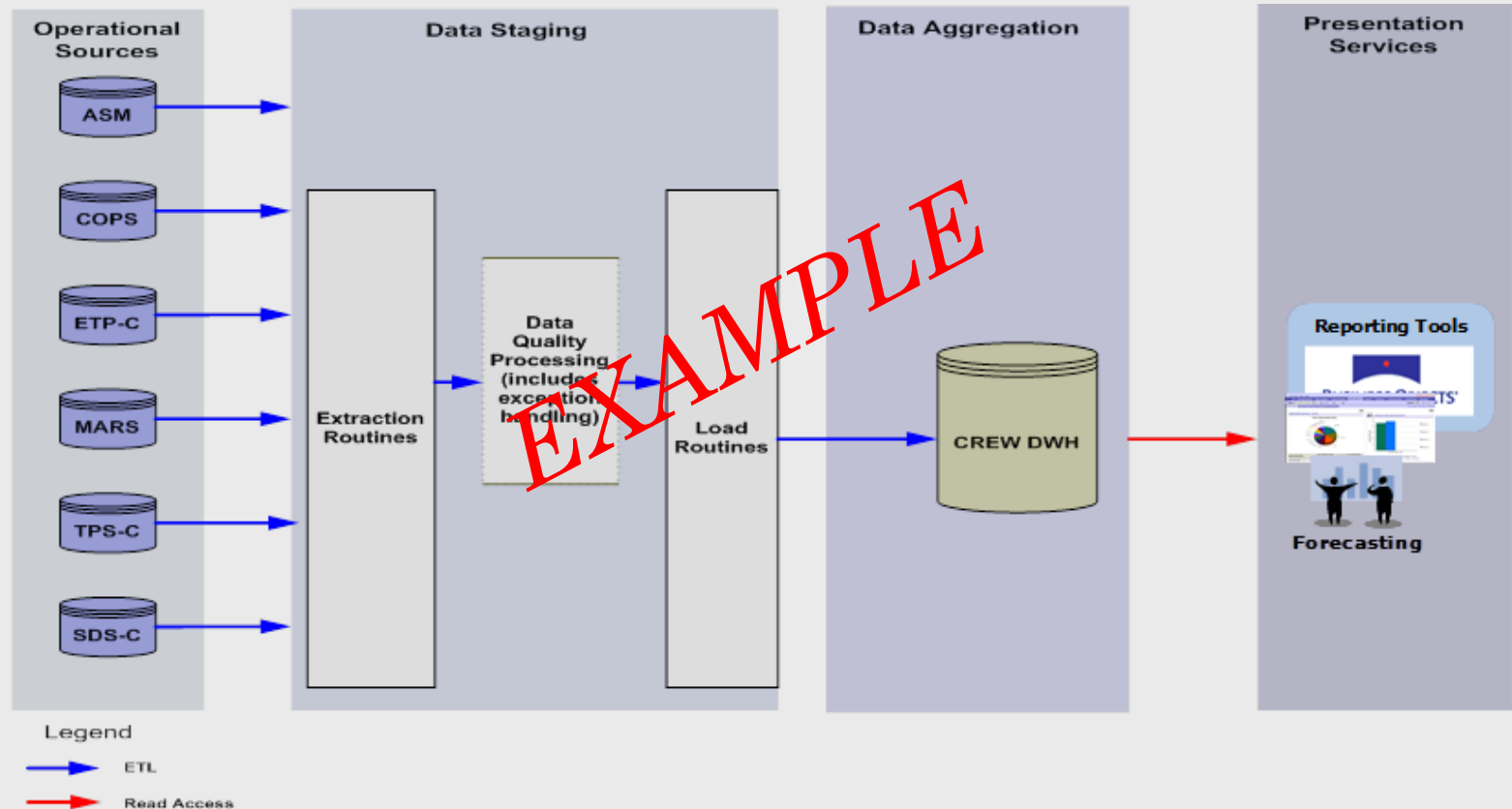


- Smaller segments of functionality are achieved faster
- Requirements that are aging are revalidated quicker
- Costs are incurred more evenly
- Reuse is increased and redundancy decreased
- Strategic changes have smaller overall project impact
- IT builds a track record of wins quicker

Week 1 Topic: Delivery Approach

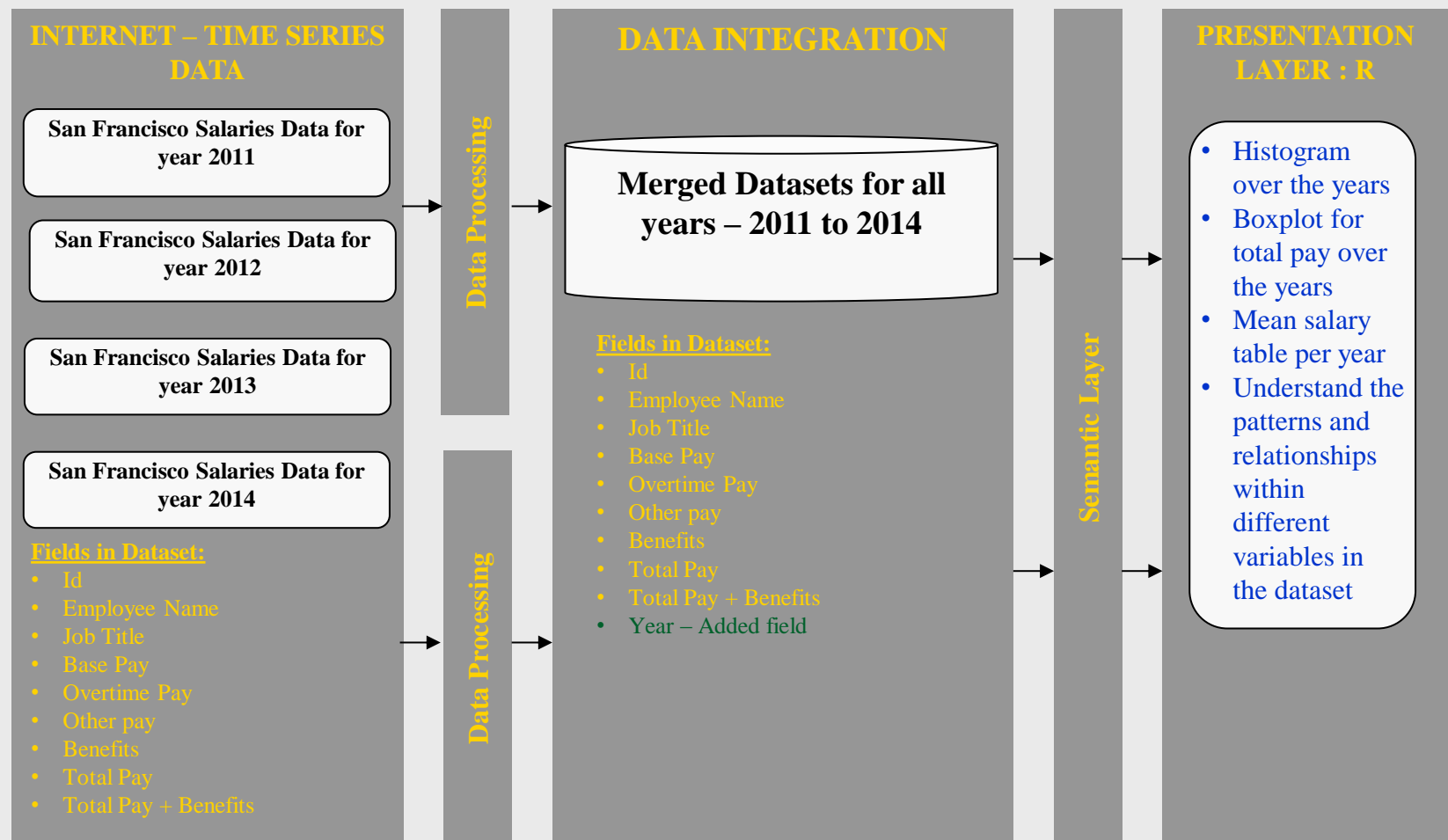


Week 1 Topic: High Level DFD Example



Week 1 Topic:

Sample DFD from a student

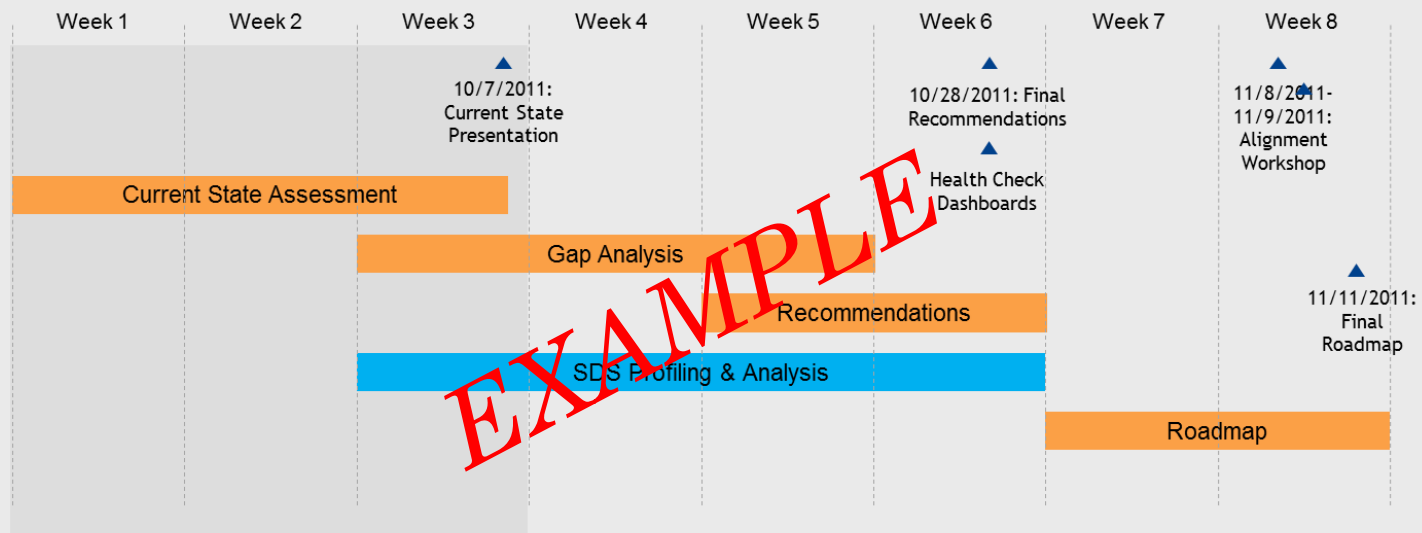


Dataset Ref: <https://www.kaggle.com/kaggle/sf-salaries>

Size: 11.7 MB

Week 1 Topic:

High Level Project Plan Example



- Completed Current State Deliverable
- Crafted the Operating Model and began socialization process for approval
- Delivered Recommendations
- Performed Alignment Workshop
- Delivered Roadmap and Initiated discussions of how to implement

Week 1 Topic:

Sources of data/information

- ◆ Weather Underground: <http://www.wunderground.com/history/>
- ◆ Various government provided data: <http://www.data.gov/>
- ◆ AMERICAN COMMUNITY SURVEY 2013 ACS 1-YEAR PUMS FILES. PUMS Data Link: <http://www.census.gov/programs-surveys/acs/data/pums.html>
- ◆ Kaggle: <https://www.kaggle.com/competitions>
- ◆ A portal for statistical science, the discipline of statistics: <http://www.statsci.org/datasets.html>
- ◆ Statistical forecasting-notes on regression and time series analysis: <http://people.duke.edu/~rnau/411home.htm>
- ◆ Regression and Multivariate Data Analysis: <http://people.stern.nyu.edu/jsimonof/classes/2301/>
- ◆ German Credit: <https://onlinecourses.science.psu.edu/stat857/node/215>
- ◆ Wine Quality: <https://onlinecourses.science.psu.edu/stat857/node/223>
- ◆ Journal of Statistical Education Data Archive: <http://www.stat.ufl.edu/~winner/datasets.html>
- ◆ Princeton Course materials and data: <http://data.princeton.edu/wws509/datasets/#effort>
- ◆ IBM Watson Analytics Datasets: <https://community.watsonanalytics.com/guide-to-sample-datasets/>

Week 1 Topic:

Regression Analysis – a preview

Regression Analysis and its Uses: Regression analysis is a technique for quantifying the relationship between a criterion (or dependent) variable and one or more predictor (or independent) variables.

It is commonly used for:

1. Predicting/Forecasting the dependent variable based on specified values of the predictor variables.
2. Understanding how the predictor variables influence the dependent variable.

READINGS: (In Blackboard)

- 1) An Introduction to Regression Analysis by Alan Sykes:
http://www.law.uchicago.edu/files/files/20.Sykes_.Regression.pdf
- 2) 7 Types of Regression Modeling:
<http://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>
- 3) Regression Basics: <http://people.stern.nyu.edu/jsimonof/classes/2301/pdf/regmback.pdf>

527 Course Review: Parameters and Statistics

- ◆ Statistics are used to approximate population parameters.
- ◆ *Parameters* are characteristics of populations. Because populations usually cannot be measured in their entirety, parameter values are generally unknown. *Statistics* are quantities calculated from the values in the sample.

	Population Parameters	Sample Statistics
Mean	μ	\bar{x}
Variance	σ^2	s^2
Standard Deviation	σ	s

527 Course Review: Statistics

Descriptive Statistics:

- ◆ The goals when you are describing data are to
 - screen for unusual sample data values
 - inspect the spread and shape of continuous variables
 - characterize the central tendency of the sample.

Inferential Statistics:

- ◆ The goals for statistical inference are to
 - estimate or predict unknown parameter values from a population, using a sample
 - make probabilistic statements about population attributes.
- ◆ After you select a random sample of the data, you can start describing the data. Although you want to draw conclusions about your population, you first want to explore and describe your data before you use inferential statistics. Why?
 - Data must be as error free as possible.
 - Unique aspects, such as data values that cluster or show some unusual shape, must be identified.
 - An extreme value of a variable, if not detected, could cause gross errors in the interpretation of the statistics.

527 Course Review: Parameters and Statistics

- ◆ *Statistics* are quantities calculated from the values in the population.
- ◆ Suppose you have x_1, x_2, \dots, x_n , a sample from some population

$$\bar{x} = \frac{1}{n} \sum x_i$$

The mean is an average, a typical value in the distribution.

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

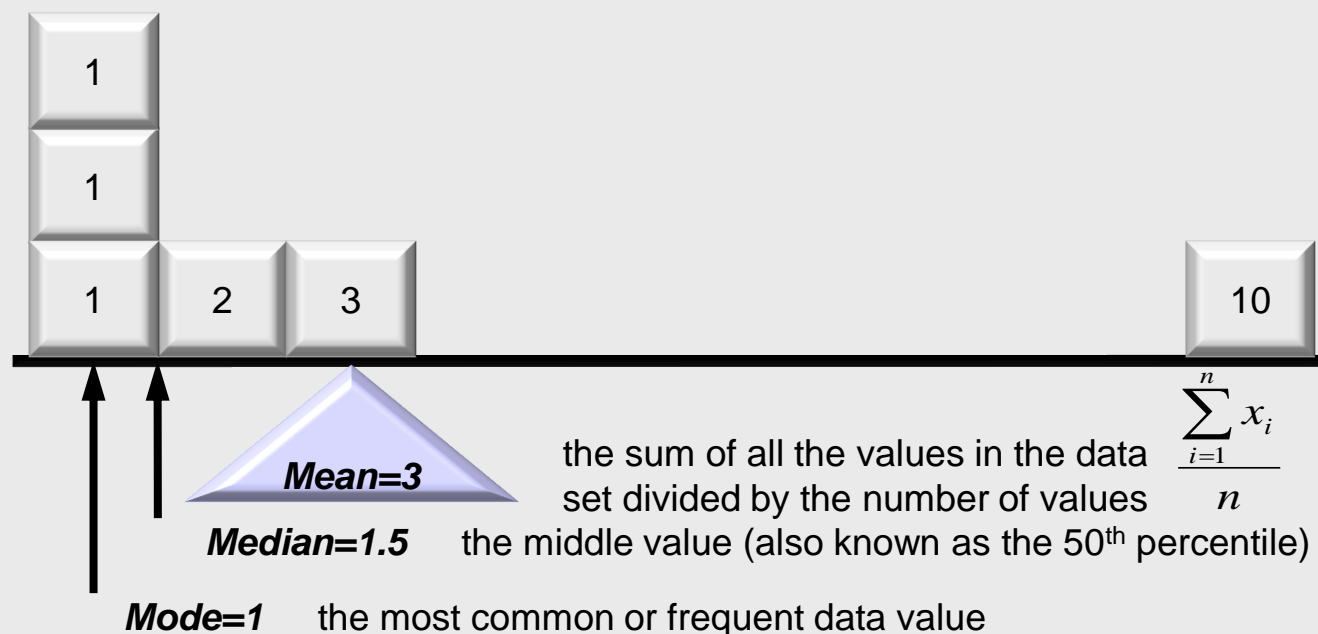
The variance measures the sample variability.

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

The standard deviation is square root of the variance and also measures variability in data. Usually reported in the same units as the mean.

527 Course Review: Mean, Median, Mode

- ◆ A property of the sample mean is that the sum of the differences of each data value from the mean is always 0, that is, $\sum (x_i - \bar{x}) = 0$.
- ◆ The *mean* is the arithmetic balancing point of your data.
- ◆ The *median* is the data point in the middle of a sorted sequence. It is appropriate for either rank scores (variables measured on an ordinal scale) or variables measured on an interval or ratio scale with a skewed distribution.
- ◆ The *mode* is the data point that occurs most frequently. It is most appropriate for variables measured on a nominal scale. There might be several modes in a distribution.

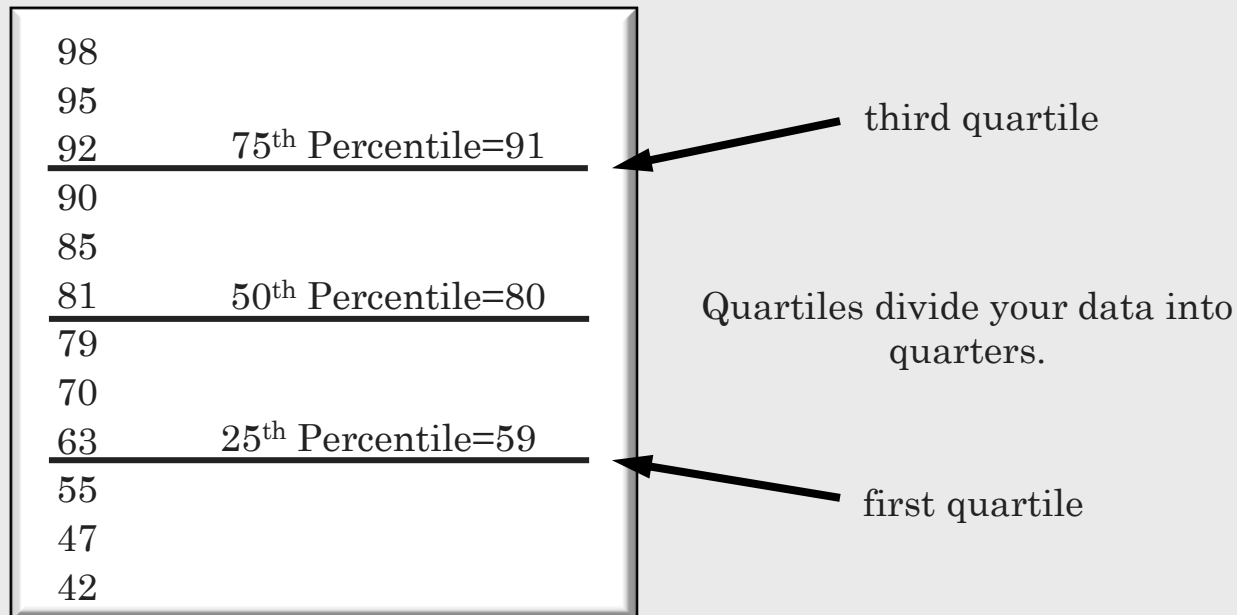


527 Course Review: Distributions

- ◆ A *distribution* is a collection of data values that are arranged in order, along with the relative frequency. For any type of data, it is important that you describe the location, spread, and shape of your distribution using graphical techniques and descriptive statistics.
- ◆ When you examine the distribution of values in a variable, you can determine the following:
 - the range of possible data values
 - the frequency of data values
 - whether the data values accumulate in the middle of the distribution or at one end
 - Are the values symmetrically distributed?
 - Are any values unusual?
 - What is the best estimate of the average of the values for the population?
 - What is the best estimate of the average spread or dispersion of the values for the population?

527 Course Review: Percentiles

- ◆ *Percentiles* locate a position in your data larger than a given proportion of data values.
- ◆ These are commonly reported percentile values:
 - the 25th percentile, also called the first quartile
 - the 50th percentile, also called the median
 - the 75th percentile, also called the third quartile



527 Course Review:

Spread of Distribution - Dispersion

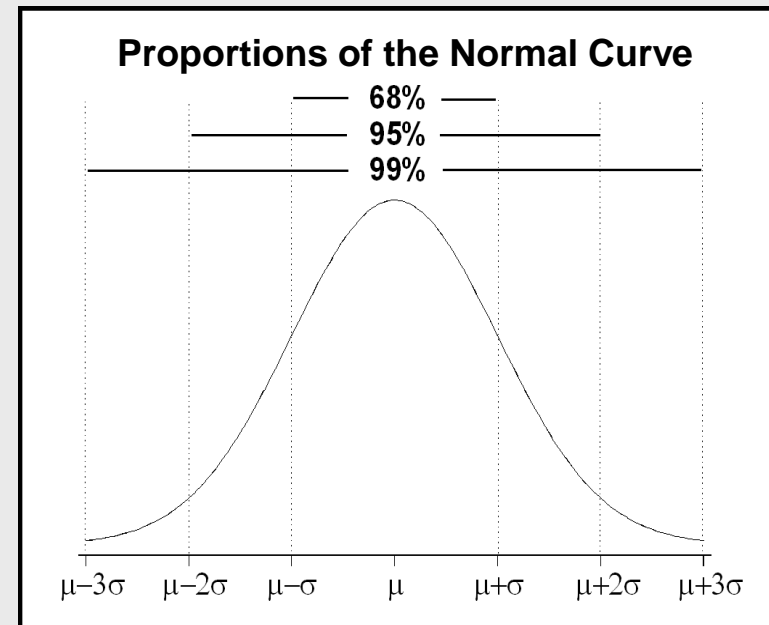
- ◆ Measures of dispersion enable you to characterize the dispersion, or spread, of the data.
- ◆ A value better suited to reflect dispersion is the *interquartile range*. The interquartile range shows the range of the middle 50% of data values.

Measure	Definition
Range	the difference between the maximum and minimum data values
Interquartile Range	the difference between the 25th and 75th percentiles
Variance	a measure of dispersion of the data around the mean
Standard Deviation	a measure of dispersion expressed in the same units of measurement as your data (the square root of the variance)

527 Course Review: Normal Distribution

A normal distribution

- ◆ is symmetric. If you draw a line down the center, you get the same shape on either side.
- ◆ defined by two parameters, μ (the population mean) and σ (the population standard deviation).
- ◆ is bell shaped and has mean = median = mode which describes the midpoint of the distribution
- ◆ Another name for the normal distribution is the Gaussian distribution.
- ◆ The standard normal curve has $\mu=0$ and $\sigma=1$. The area under the curve between any two values can be calculated.
- ◆ Approximately 68% of the total area lies within 1 standard deviation of the mean.
- ◆ Approximately 95% of the total area lies within 1.96 standard deviations of the mean.
- ◆ Approximately 99.7% of the area lies within 3 standard deviations of the mean.



527 Course Review:

Normal Distribution (cont.)

- ◆ Often in analysis, although not always, a normal distribution is assumed.
- ◆ The normal distribution is a mathematical function. The height of the function at any point on the horizontal axis is the “probability density” at that point. Normal distribution probabilities (which can be thought of as the proportion of the area under the curve) tend to be higher near the middle.
- ◆ The center of the distribution is the population mean (μ). The standard deviation (σ) describes how variable the distribution is about μ . A larger standard deviation implies a wider normal distribution. The mean locates the distribution (sets its center point) and the standard deviation scales it.
- ◆ Often, values that are more than two standard deviations from the mean are regarded as unusual. Only about 5% of all values are at least that far away from the mean.
- ◆ You use this information later when you discuss the concepts of confidence intervals.

527 Course Review:

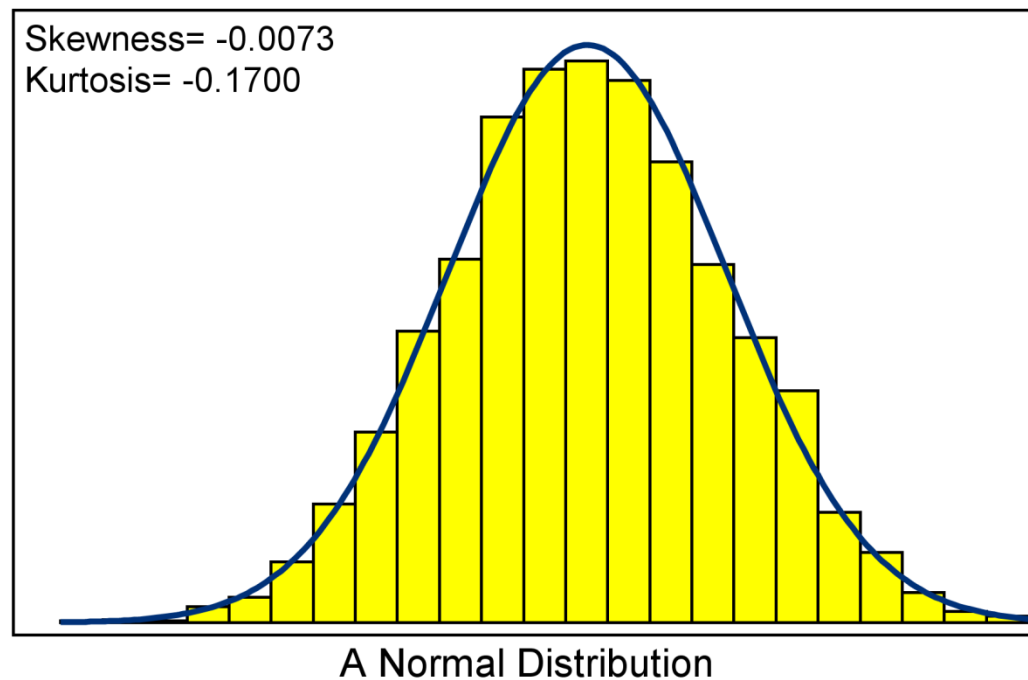
Distributions compared to Normal

- ◆ The distribution of your data might not look normal. There are an infinite number of ways that a population can be distributed. When you look at your data, you might notice the features of the distribution that indicate similarity or difference from the normal distribution.
- ◆ When you evaluate distributions, it is useful to look at statistical measures of the shape of the sample distribution compared to the normal.
- ◆ Two such measures are skewness and kurtosis.



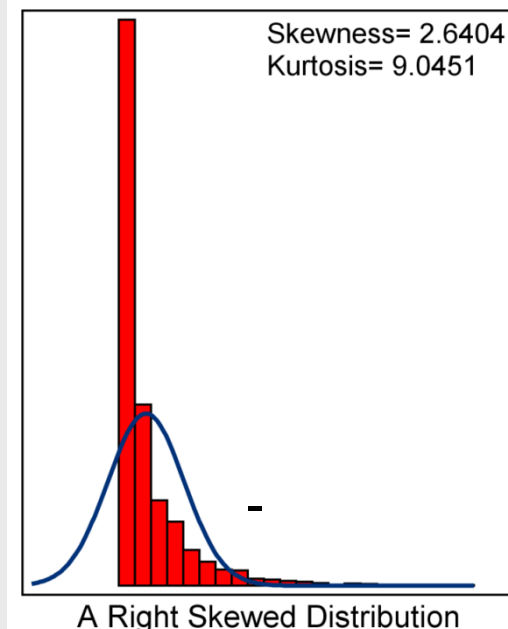
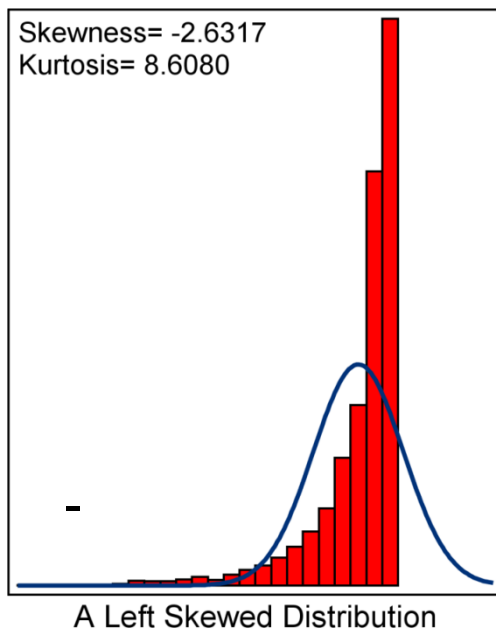
527 Course Review: Normal Distribution

- ◆ A histogram of data from a sample drawn from a normal population generally show values of skewness and kurtosis near zero in SAS output.



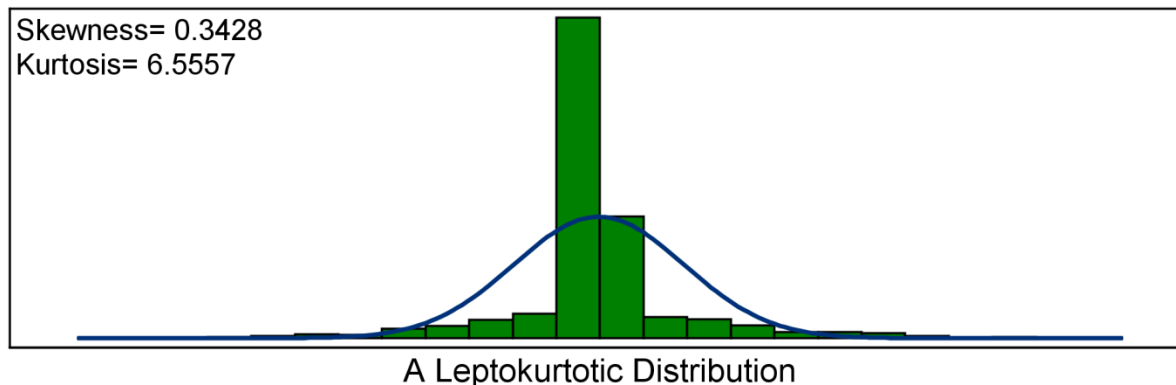
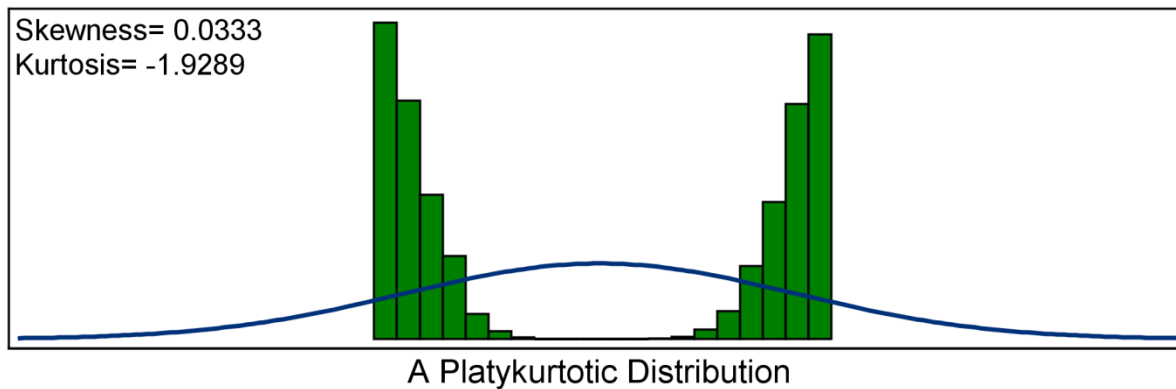
527 Course Review: Skewness

- ◆ One measure of the shape of a distribution is skewness. The *skewness* statistic measures the tendency of your distribution to be more spread out on one side than the other. A distribution that is approximately symmetric has a skewness statistic close to zero.
- ◆ If your distribution is more spread out on the
- ◆ **left** side, then the statistic is negative, and the mean is less than the median. This is sometimes referred to as a *left-skewed* or *negatively skewed* distribution.
- ◆ **right** side, then the statistic is positive, and the mean is greater than the median. This is sometimes referred to as a *right-skewed* or *positively skewed* distribution.



527 Course Review: Kurtosis

- ◆ *Kurtosis* measures the tendency of your data to be distributed toward the center or toward the tails of the distribution. A distribution that is approximately normal has a kurtosis statistic close to zero in SAS. Kurtosis is often very difficult to assess visually.



527 Course Review:

Kurtosis (cont.)

- ◆ If the value of your kurtosis statistic is negative, the distribution is said to be *platykurtic*. If the distribution is both symmetric and platykurtic, then there tends to be a smaller-than-normal proportion of observations in the tails and/or a somewhat flat peak. Rectangular, bimodal, and multimodal distributions tend to have low (negative) values of kurtosis.
- ◆ If the value of the kurtosis statistic is positive, the distribution is said to be *leptokurtic*. If the distribution is both symmetric and leptokurtic, then there tends to be a larger-than-normal proportion of observations in the extreme tails and/or a taller peak than the normal. A leptokurtic distribution is often referred to as *heavy-tailed*. Leptokurtic distributions are also sometimes referred to as *outlier-prone distributions*.
- ◆ Distributions that are asymmetric also tend to have nonzero kurtosis. In these cases, understanding kurtosis is considerably more complex than in situations where the distribution is approximately symmetric.

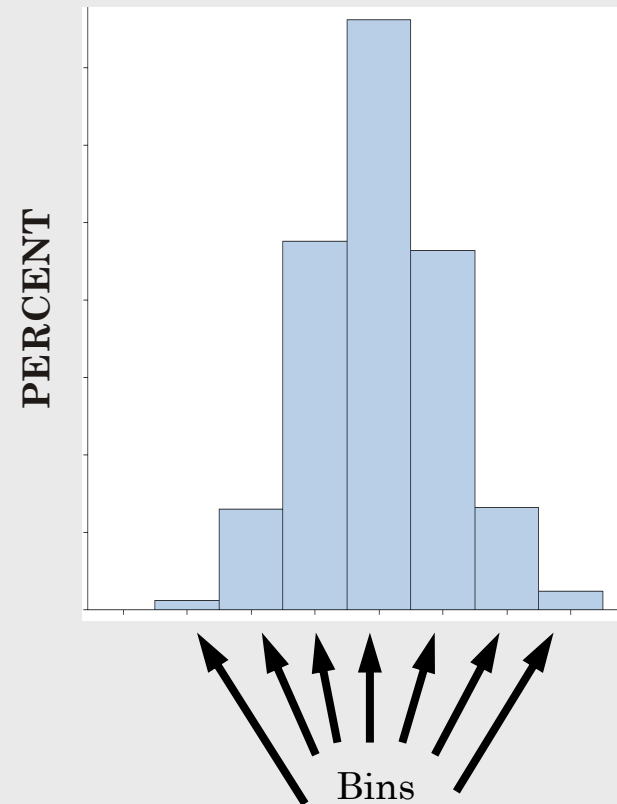
527 Course Review:

Graphical Displays of Distributions

- ◆ You can produce the following three types of plots for examining the distribution of your data values:
 - histograms
 - normal probability plots
 - box plots
 - scatter plots

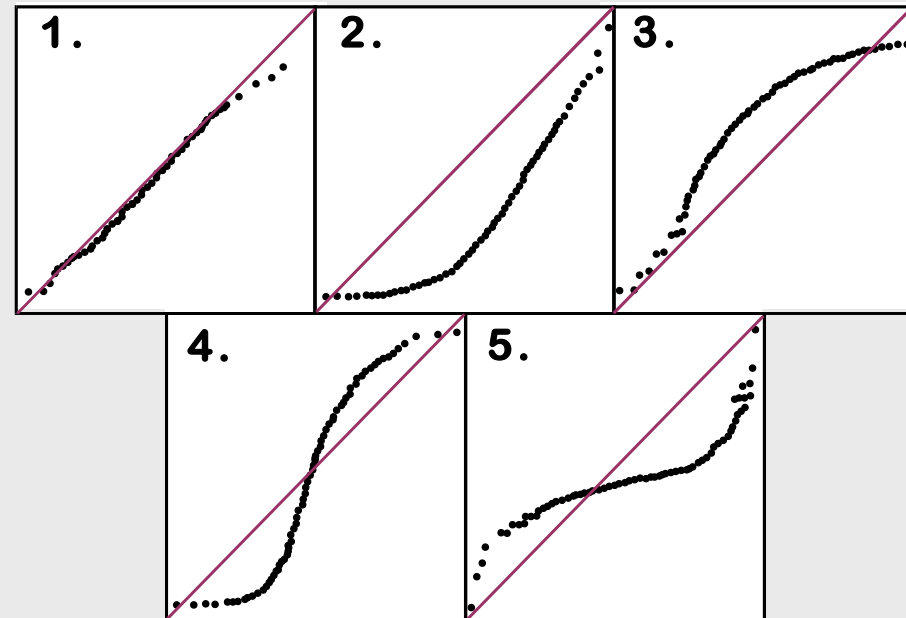
527 Course Review: Histograms

- ◆ Most elementary statistical procedures assume some underlying population probability distribution. It is a good idea to look at your data to see whether the distribution of your sample data can reasonably be assumed to come from a population with the assumed distribution.
- ◆ A histogram is a good way to determine how the probability distribution is shaped.



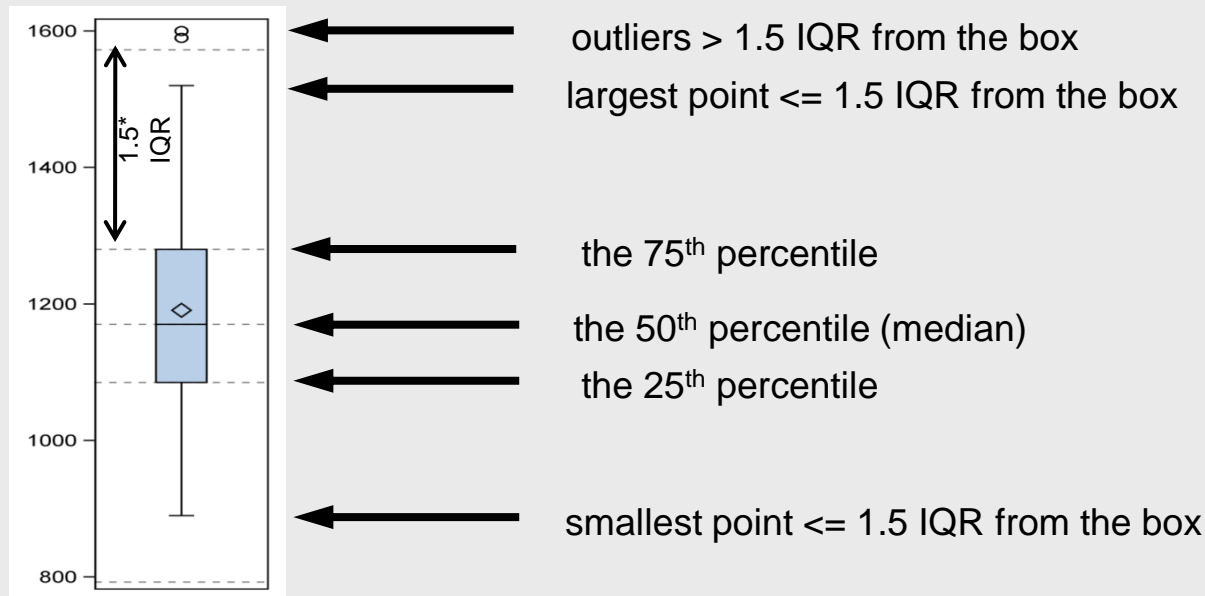
527 Course Review: Normal Probability Plots

- ◆ A *normal probability plot* is a visual method for determining whether your data comes from a distribution that is approximately normal. The vertical axis represents the actual data values, and the horizontal axis displays the expected percentiles from a standard normal distribution.
- ◆ The above diagrams illustrate some possible normal probability plots for data from the following:
- ◆ normal distribution (The observed data follow the reference line.)
- ◆ skewed-to-the-right distribution
- ◆ skewed-to-the-left distribution
- ◆ light-tailed distribution
- ◆ heavy-tailed distribution



527 Course Review: Box Plots

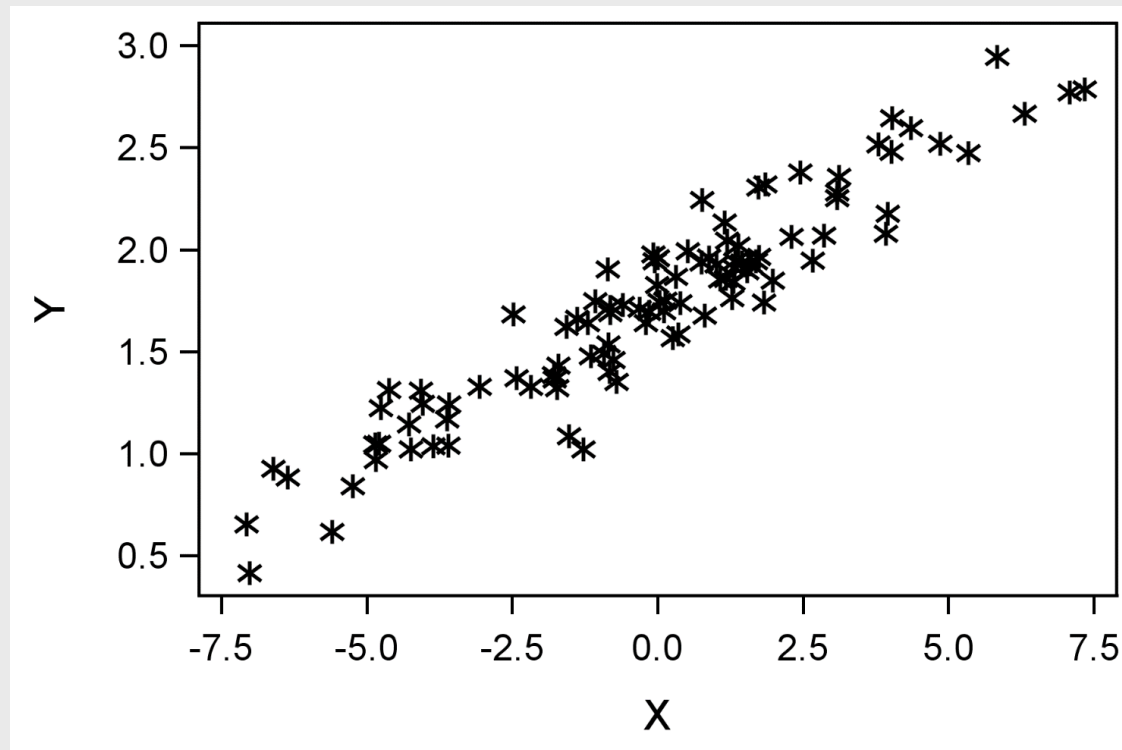
- ◆ *Box plots* (Tukey 1977) (sometimes referred to as *box-and-whisker plots*) provide information about the variability of data and the extreme data values. The box represents the middle 50% of your data (between the 25th and 75th percentile values). You get a rough impression of the symmetry of your distribution by comparing the mean and median, as well as by assessing the symmetry of the box and whiskers around the median line. The whiskers extend from the box as far as the data extends, to a distance of, at most, 1.5 interquartile range (IQR) units. If any values lay more than 1.5 IQR units from either end of the box, they are represented in SAS by individual plot symbols.
- ◆ The plot above shows that the data are approximately symmetric.



The mean is denoted by a \diamond .

527 Course Review: Scatter Plots

- ◆ *Scatter plots* are two-dimensional graphs produced by plotting one variable against another within a set of coordinate axes. The coordinates of each point correspond to the values of the two variables.



527 Course Review: Scatter Plots (cont.)

- ◆ Scatter plots are useful to accomplish the following:
 - explore the relationships between two variables
 - locate outlying or unusual values
 - identify possible trends
 - identify a basic range of Y and X values
 - communicate data analysis results
- ◆ The predicted value can be thought of as the best estimate of the value of the response at a given value of the predictor variable. Scatter plots show graphically the relationship between predictor variables and response variables.
- ◆ Traditionally, predictor variables are plotted on the x axis and response variables are plotted on the y-axis. A preliminary analysis of associations involves discovery of the presence of associations and their nature.

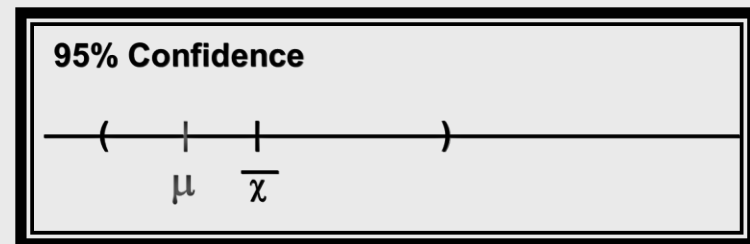
527 Course Review:

Standard Error of a Mean

- ◆ A statistic that measures the variability of your estimate is the *standard error of the mean*. In statistics, assumptions are often made about distributions of parameters. A common one is that the sampling distribution of parameters is normal. This does not necessarily mean that the units of the population are normally distributed. It is often assumed that the parameter itself is normally distributed. Even though most statisticians only take one sample and get one point estimate for the population parameters, it is useful if they can assume normality of the parameter. The variability of a parameter is measured by its standard error.
- ◆ It differs from the sample standard deviation in that:
 - the sample standard deviation is a measure of the variability of data;
 - the standard error of the mean is a measure of the variability of the sample mean.
 - Standard error of the mean = $\frac{s}{\sqrt{n}} = s_{\bar{x}}$
 - where
 - s is the sample standard deviation.
 - n is the sample size.
- ◆ The standard error of the mean is a measure of precision of the parameter estimate. The smaller the standard error, the more precise your estimate.

527 Course Review: Confidence Interval

- ◆ A *confidence interval* is a range of values that you believe is likely to contain the population parameter of interest is defined by an upper and lower bound around a parameter estimate.
- ◆ To construct a confidence interval, a significance level must be chosen.
- ◆ A 95% confidence interval is commonly used to assess the variability of the sample mean. In the Ames housing sales example, you interpret a 95% confidence interval by stating that you are 95% confident that the interval contains the mean sale price for your population of home sales.
- ◆ You want to be as confident as possible, but remember that if you increase the confidence level too much, the width of your interval increases beyond the point where it is informative. For example, a 100% confidence interval would have confidence bounds of negative and positive infinity.
- ◆ A 95% confidence interval represents a range of values within which you are 95% certain that the true population mean exists.
 - One interpretation is that if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.



527 Course Review:

Conditioning our minds to think data

For Week 1, we will cover some history leading up to present day data analytics environment, challenges and technologies available:

- ◆ Defining Data Analytics, Data Mining, and Business Intelligence
- ◆ Data Management Concepts
- ◆ Definition of a Data Warehouse
- ◆ Definition of a Data Marts
- ◆ Next for this course

527 Course Review:

What do we mean by Data Analytics?

For this course, we discuss analysis of data in 3 overlapping disciplines; data analytics, data mining, and business intelligence. You will see in the market that most vendor products choose a discipline to focus their marketing efforts. However, dependent on the solution offering, it can have all or one functional offering.

Theoretically, it's all analysis of data to gather information but using these terms adds *purpose* to the activity. We loosely define the 3 disciplines further but as we progress, we will speak more about *why we use certain techniques for what purpose* rather than noodle over what discipline it belongs to.

- ◆ Data Analytics: Further than just visualization of data, the goal of analytics is to support a decision or prove a hypothesis by using quantitative techniques. This confirmatory approach will validate or summarize information to provide the answer.
- ◆ Data Mining: Describes examining data sets to identify undiscovered patterns and uncover hidden relationships. This exploratory approach may use sophisticated models to determine its patterns.
- ◆ Business Intelligence: Describes data analysis efforts that is focused on answering analytical questions about the business, supports business management processes, or provides views of the business whether it be enterprise or departmental.

527 Course Review: More on Data Analytics

We describe more confirmatory, inference driven data analysis activities to data analytics:

- ◆ **Descriptive Statistics:** Use of descriptive statistics like means, medians, and standard deviations
- ◆ **Graph Distributions:** Using distributions of data to summarize groupings and seek out anomalies in data using visual means e.g., histograms, pie charts, bar charts, etc.
- ◆ **String Operations:** Manipulating, parsing, or translating text values to otherwise interpret information that may be coded or concatenated.
- ◆ **Math Functions:** Using counts, sums, percentages, and other math functions to validate or summarize data. Same techniques used in string operations can also be applied to numeric values.
- ◆ Use of filters and sorts to understand data.
- ◆ Logical operations on data e.g., applying finite ranges, $<$, $=$, $>$.
- ◆ Calculating derived values and applying conditions on data.
- ◆ In general, we answer questions or confirm hypothesis through quantitative means that is not tied to sophisticated models. We tie simple math and view manipulations to this activity.

527 Course Review: More on Data Mining

We describe more model driven exploratory data analysis activities with data mining:

- ◆ **Outlier Analysis:** Identification of unusual data records, that might be interesting or data errors that require further investigation.
- ◆ **Correlations:** Association rule learning or dependency modelling – Searches for relationships between variables. This is sometimes referred to as market basket analysis.
- ◆ **Clustering (Segmentation/Summarization):** The act of grouping similar cases together. Discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- ◆ **Classification:** Predicting a discrete categorical value. The task of generalizing known structure to apply to defined categories.
 - ◆ **Forecasting/Regression:** Discovering patterns in data that can lead to reasonable predictions about the future. Attempts to find a function which models the data with the least error.
 - ◆ **Bayesian:** Use of conditional probabilities to infer an outcome.
 - ◆ **Others include use of Neural Networks and Decision Trees**

527 Course Review:

More on Business Intelligence

We tie Business Intelligence (BI) with a Data Warehouse (DW) solution. Hence, we focus on OLAP functions as BI analysis techniques. We define OLAP as the ability to join discrete sets of data into a dimensional cube structure that is optimized/aggregated for analysis/reporting. With an OLAP cube, you can:

- ◆ **Slice:** Act of taking a subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension.
- ◆ **Dice:** Creating a subset of a cube for analysis by choosing values from dimensions.
- ◆ **Drill Up/Down:** Allowing a user to navigate through the levels of a dimensional hierarchy up (summarized) and down (more detailed).
- ◆ **Roll Up:** Summarization of data along a dimension e.g., totals or derived.
- ◆ However, BI also encompasses reporting functions similar to the way we defined Data Analytics albeit focused on answering business questions and hypothesis.

527 Course Review: Data Analysis Vendor Landscape

One can perform data analytics in spreadsheets as long as one can acquire the data in the right format which is why Excel is so popular amongst the business users. For more sophisticated work, however, we categorize the following top vendors:

Data Mining (*More Business Use*)

SAS Enterprise Miner

SPSS (IBM)

Stata

Tibco Spotfire

Dell StatSoft

Statistical (*More Academic Use*)

MatLab

Mathematica

Minitab

R

SAS JMP

Business Intelligence – OLAP Included

Cognos (IBM)

BusinessObjects (SAP)

MicroStrategy

Hyperion (Oracle)

Informatica

Business Intelligence - Visualization

Tableau

QlikView

Domo

Sisense

first

527 Course Review: Roles in Data Analysis

Business Analyst

- Supports business users
- Serves as subject matter expert for the business domain
- Usually owns a business application, process and/or function
- Performs business analysis mostly in business application or Excel
- Some super users can model and code as necessary

Quantitative Analyst

- Supports business users and analysts
- Serves as subject matter expert for statistical solutions domain
- PhD types that will use mathematical and statistical programming applications to build and execute model based analysis
- Consumes high volumes of raw data for model data feeds

Reporting Analyst

- Supports business users and analysts by building and supporting reports, dashboards, or BI solutions
- Serves as subject matter expert for the reporting domain
- Usually owns the reporting business process or function and in some cases data steward to the data within reports
- Performs data analysis and validations for data processing and management

Data Analyst

- Supports business users and analysts for ad hoc queries or other data related analysis projects
- Serves as subject matter expert for the data domain and typically serves as data stewards
- Performs data analysis, builds reports as necessary, and supports data processing and management tasks

527 Course Review: Database Vendor Landscape

Database Software	Language Support
RDBMS: <ul style="list-style-type: none"> • IBM (Mainframe, DB2, IMS, Cloudant, Informix) • Oracle (Latest 12c) • Microsoft SQL Server (Latest 2014) • SAP (ASE, IQ-columnar) • Teradata (columnar) 	<ul style="list-style-type: none"> • Query: SQL, PL/SQL (Oracle) • API exists for various application build languages • Some recently extended to support JSON, XML
Open Source: PostgreSQL, MySQL, MariaDB Enterprise, Firebird	<ul style="list-style-type: none"> • Query: SQL and other scripting languages
NoSQL (non-relational database systems with no pre-defined structure): Cassandra, MongoDB, Dynamo	<ul style="list-style-type: none"> • Query: Cassandra uses CQL and others use various scripting languages • Open distributed file systems. Cassandra resembles RDBMS table structure while MongoDB uses JSON like file structures
In-Memory: KDB, EXtremeDB, MemSQL, SAP HANA	<ul style="list-style-type: none"> • Query: KDB uses Q and others use direct queries in C/C++, HANA supports Javascripts
Hadoop (a distributed file system) – <i>More on this when we cover Big Data</i>	
Specialty Appliances: <i>Netezza, XtremeData, Greenplum</i> are optimized for analysis using multi processors, lot of memory, faster networks, large disk space, etc.	

527 Course Review: Data Management Components

The following is a view into data management concepts and components for an enterprise. Dependent on the role of the data analyst, one can work in any one of the following areas:

Data Governance			
Organizational Model	Enablement	Standards & Policies	Processes & Procedures
Data Quality			
Profiling / Analysis	Cleansing	Controls	Enrichment / Enhancement
Data Usage			
Reporting	Analytics (OLAP)		Data Mining
Quantitative Analysis	Scorecard / Dashboards		Alerts/ Notifications
Data Management			
System of records	Operational data stores	Data warehouse/data marts	Data movement (ETL/EAI/EII)
Data protection	Metadata management	Reference data management	Master data management
Architecture			
Conceptual	Logical	Physical / technical	
Design patterns	Services	Standards	

527 Course Review:

Data Types – Transaction vs Snapshot

It's important to understand data requirements for analysis as, most likely, you will be defining it for the developers and act as a conduit for the business users' needs. Translating the needs into requirements is not an easy task and in some cases require most time and resources during an implementation. We'll cover data requirements gathering methodologies in the next class. For today, we define what we mean by transaction versus snapshot data:

Transaction Data:

- ◆ Records business events e.g., retail purchases, call detail records, bank deposits/withdrawals, insurance claims, stock trades/quotes, etc.
- ◆ Usually recorded along with date and timestamp for each transaction.
- ◆ Considered raw data or detailed view of data as analysis may be done at a point in time versus looking at each transaction.

Snapshot Data:

- ◆ Records current or past 'state' of a business entity or relationship e.g., customer, account or measures of metric values at a certain point in time.
- ◆ Unlike transaction data, a query will typically want to access only one "time instance" of the snapshot data e.g., balance on the account for month end close.
- ◆ Multiple snapshots can be used for trending or constructing averages over time.

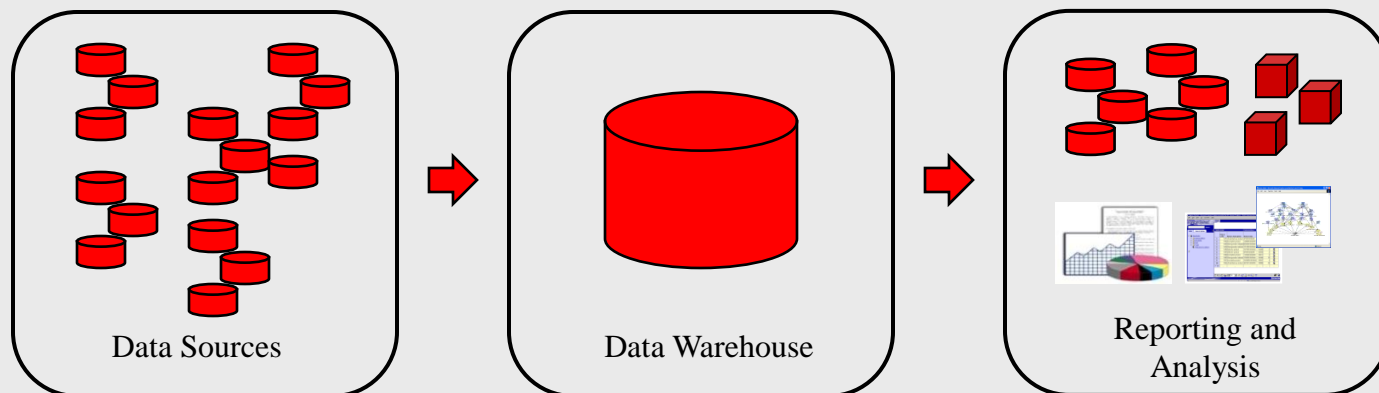
527 Course Review:

Data Warehouse: Definition

Most of all, you need data to do data analysis. Most business requirements will require data from multiple data sources with differing data types and varying degrees of data quality. This is where a data warehouse comes in.

Simply put, *a data warehouse is a data repository that collects data from multiple sources into a uniform structure.* Architecture of a data warehouse and its use varies when you start to consider what that uniform structure looks like.

There were two different philosophies on implementing data warehouses in the beginning. Here, we are talking back in the 90s, many years after the concept was first discussed in the 60s.



527 Course Review:

Data Warehouse: Inmon versus Kimball

When you start to think about building a data warehouse, one of the first things to consider is how you will model this uniform structure. Of course, this is after you have gathered sufficient business requirements and identified applicable data sources to understand the purpose of the data warehouse in the first place. When we say purpose here, we also mean analytics and its data needs as well.

The two philosophies were:

1. Bill Inmon claimed that this uniform structure is in a 3NF*. Analysis is done off of dimensionally structured data marts** that is produced from this 3NF data warehouse. More of a “top down” approach.
2. Ralph Kimball claimed that this uniform structure is a conglomeration of departmental data marts that may share data through an information bus. Hence, a data warehouse itself may also have a dimensional structure. More of a “bottom up” approach.

*3NF was originally defined by E.F. Codd in 1971. This class assumes that you know basics of data modeling. You will need some data modeling skills to prepare your data sets.

**We will discuss data marts and dimensional data modeling in subsequent slides.

527 Course Review:

Data Warehouse: Adoption and Evolution

There was more adoption of Kimball's method as it was considered a "lighter" more practical approach. Investment was easier to justify. This also meant creation of departmental data silos which is another topic all together. He wrote *The Data Warehouse Toolkit* in 1996 which was considered a must read for anyone building a data warehouse at the time.

Following Inmon's approach was a bigger investment with many months spent on design which led to lower adoption by business as it was harder to see benefit in a timely manner. Although, theoretically, it made a lot of sense. He published *Building the Data Warehouse* in 1992. 20+ years after first coining the term.

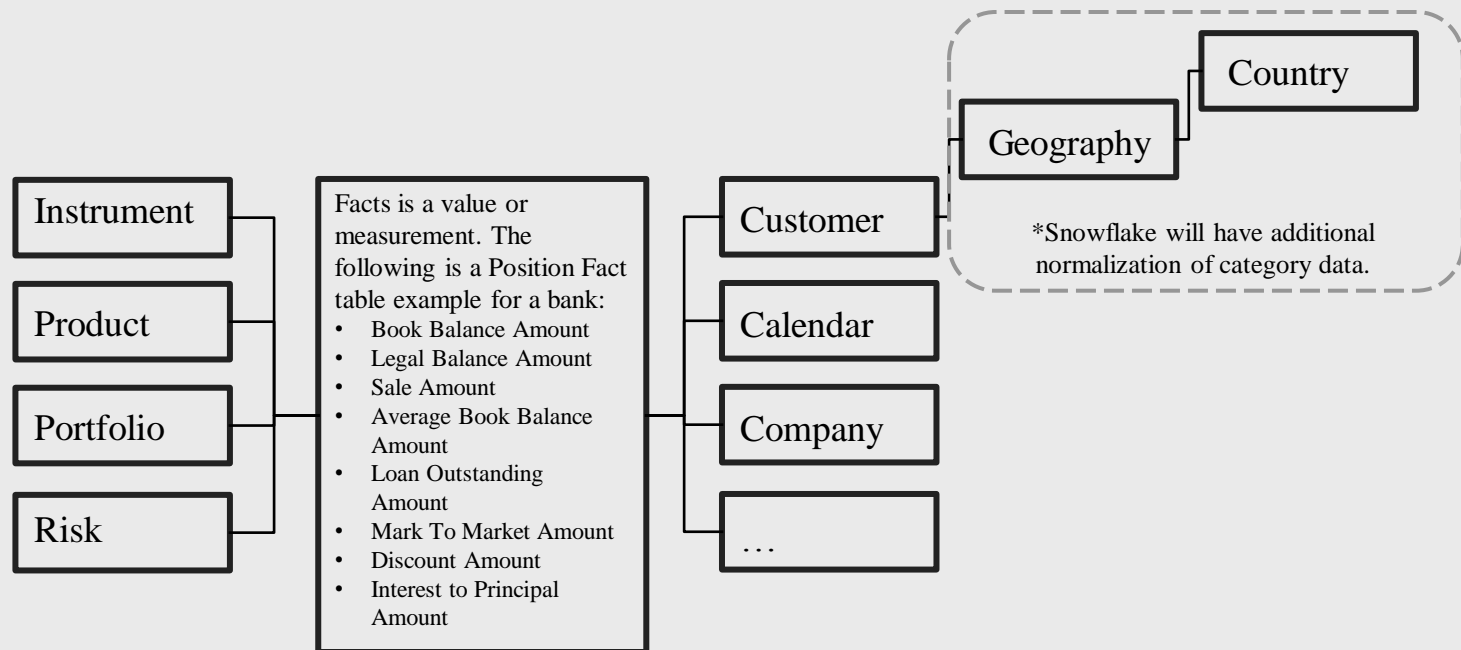
In general, evolution of data warehouses, data analytics, and data management follow technology advancement. Punch cards (used up to about the 70s even though, I still used them in graduate school in the mid 90s) were replaced by magnetic tape (introduced in the 50s) for data storage. Magnetic tapes were replaced with hard disk drives etc. etc.

Today, you can load multiple Terabytes of data into memory for analysis.

527 Course Review:

Data Mart: Definition

A data mart is usually dimensional with facts (measures) and dimensions (categories) hence the term, dimensional modeling. We call this data model a star (and/or snowflake*):



A fact table is not normalized but rather designed to respond to queries fast. Dimensional data is normalized and represents a unique descriptive to measures.

527 Course Review:

Data Solutions Comparisons

- ◆ Data Warehouse
 - Source of consistent, integrated, enterprise-wide data
 - Mechanism providing the analytical and decision support needs of the enterprise
 - Data is at multiple levels of granularity, including transaction-level and summarization
 - Data is typically retained for 3-7 years
 - Data may be sourced from the Operational environment and/or the ODS
- ◆ Data Mart
 - Mechanism providing the analytical and decision support needs of a business function
 - Data is highly summarized and is usually specific to the business function
 - Data is typically retained for 3-7 years
 - Data may be sourced from the ODS and/or the Data Warehouse
- ◆ Operational Data Store (ODS)
 - Mechanism enabling the collection, cleansing, and integration of operational data for population in either a Data Warehouse or a Data Mart
 - Data is typically at the transactional level of detail
 - Data may be retained for short time spans (90-120 days)

527 Course Review:

Data Warehouse: Now and Then

Data Warehouse – *the old days:*

- ◆ Multi-million \$\$, multi-year investment
- ◆ Batch-oriented
- ◆ Limited availability of data – high cost, limited processing power (software and hardware)

Data Warehouse – *new generation:*

- ◆ Speed of data has gotten faster e.g., streaming is not a wish, it's a must
- ◆ Amount of data being generated have increased e.g., in the petabyte range
- ◆ Type of data being processed is more diverse e.g., textural as well as tabular
- ◆ Reclaim of archive data (“dark data”) – more availability
- ◆ More diverse delivery points e.g., handheld devices
- ◆ Cost of technology has plummeted

We will discuss influences of these trends on data management when we discuss Big Data towards the end of the semester.

527 Course Review: People Management

Qualitative items to consider when embarking on a new project is about People & Adoption in the data space. These measures will have an impact on what a developer does and chooses hence should not be ignored especially at initiation phase:

- ◆ Stakeholder Readiness - *understanding opinions*
 - Is the stakeholder sold on the goals, objectives, and value of the project?
 - Does the stakeholder understand or is willing to understand what's involved in implementing the analysis?
- ◆ Organizational Readiness – *understanding situational challenges*
 - Are the needed resources available in the organization? Is the organization receptive to external resources, if not?
 - How long does it take for the organization to adopt new technologies?
- ◆ Financial Readiness
 - How much and how long will the project cost?
 - What are the financial constraints for the project?
- ◆ Data & Technology Readiness
 - What technologies and methods does the organization use currently?
 - Is the data needed for the analysis available? How easily can it be obtained?
 - Is the hardware and software needed for the analysis available?

527 Course Review:

Understanding data - retail

Data:

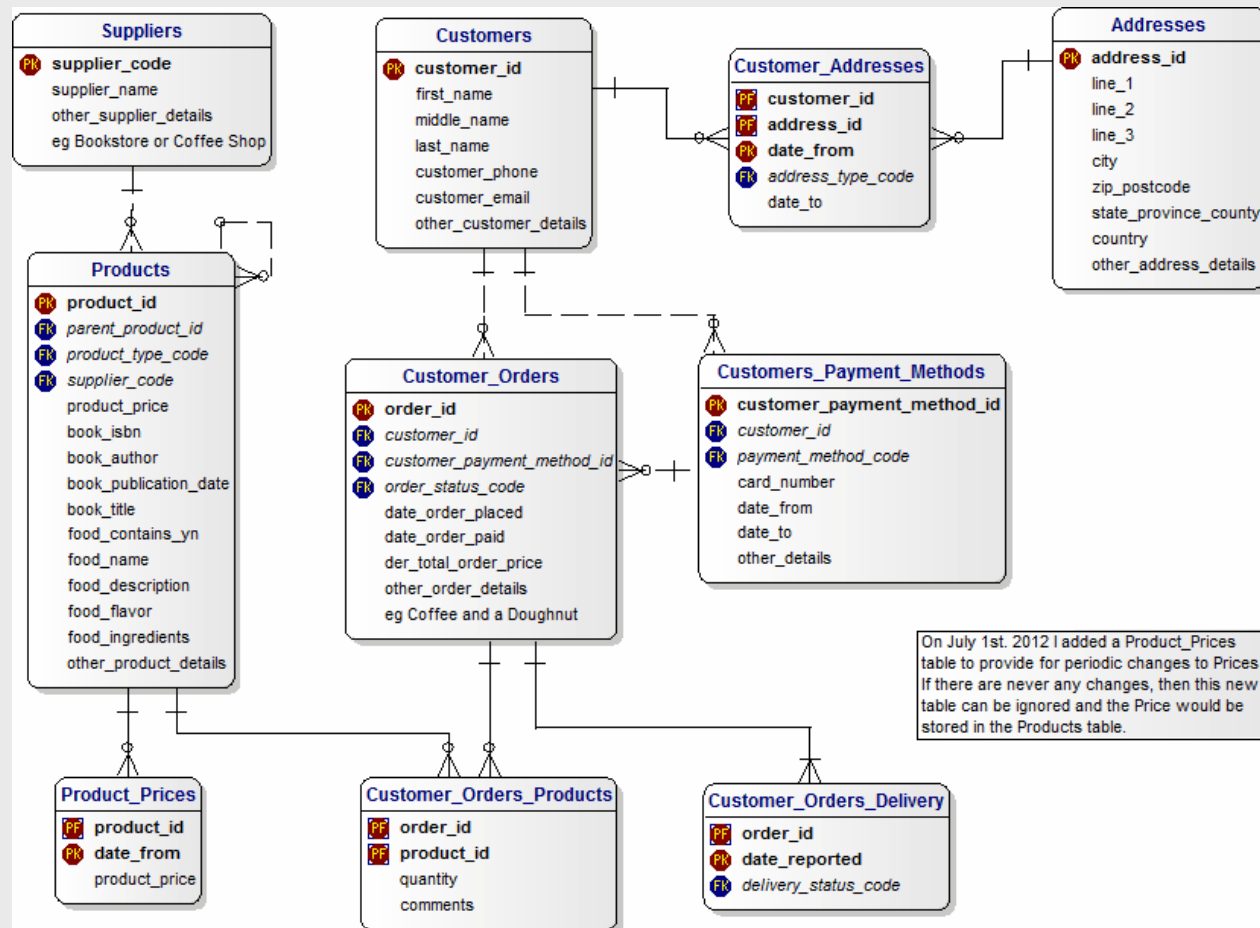
- ◆ (F) Transaction data: purchasing/orders, accounts payable, POS, sales projections, warehouse movements, employee shift records, returns
- ◆ (D) Product data: consumer merchandise, hardware, software, industrial raw materials, any tangible object or service that can be sold or bought along with SKU, EPC, etc.
- ◆ (D) Customer data: name, address, email, phone, demographic, behavioral, financial
- ◆ (D) Store/Branch data: location type, address, manager, size/resources
- ◆ (D) Others include sales reference data (e.g., account type, personnel) and supplier information

Example Analytics:

- ◆ Perform operational analytics to identify economies of scale, inventory management, cash flow analysis, optimal open hours
- ◆ Identify patterns, trends, and anomalies in transactions to mitigate risk and report fraudulent activities e.g., receipt fraud (falsified, stolen or reused receipts are used to return merchandise), price arbitrage (using higher priced product tags to return lower)
- ◆ Cross sell/Up sell using modeling techniques like market basket analysis or by simple product association e.g., diapers and diaper genie, movies with same actor, etc.
- ◆ Launch market campaigns by segmenting like customer groups together according to set criteria e.g., demographic, geographic, income, etc.

527 Course Review: Sample retail data model

http://www.databaseanswers.org/data_models/customers_and_orders/:



527 Course Review:

Understanding data – banking/trading

Data:

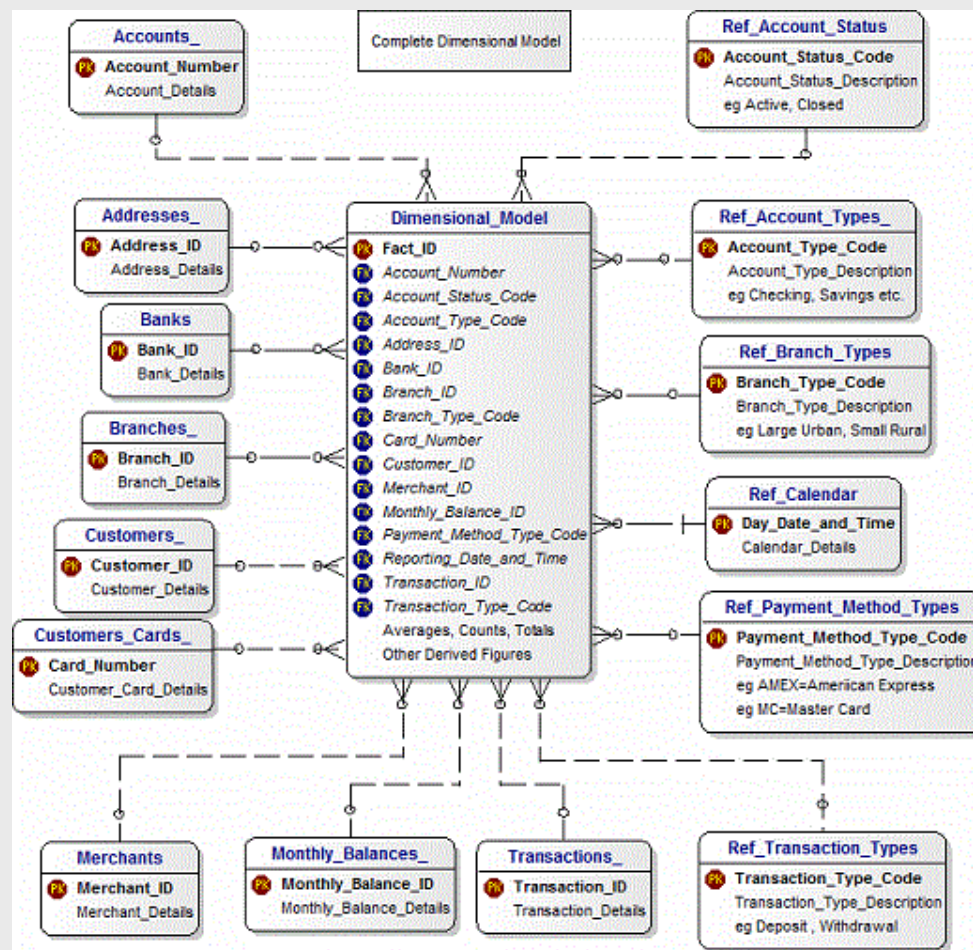
- ◆ (F) Transaction data: trades (price, size), quotes (bid price, ask price, bid size, ask size)
- ◆ (F) Position (financial) data: amount of securities or commodities held e.g., balances of accounts or portfolios
- ◆ (D) Product data (banking): savings, checking, mortgage, credit card, account
- ◆ (D) Instrument data (trading) : tradable assets of any kind e.g., securities, cash
- ◆ (D) Market data: curves, rates, prices, spreads
- ◆ (D) Customer/Obligor/Party data: in addition to the usual CUSIP/SIC/NAIC codes for businesses, SS#, Risk Rating, Obligor (bond issuer, borrower, debtor, contractually/legally obligated entity) Rating

Example Analytics:

- ◆ Banking: Compliance with regulatory measures e.g., AML, KYC, Volcker Rule, Basel, etc.
- ◆ Trading: Generate best execution outlier reports to identify trades that missed best price using transactions
- ◆ Trading: Calculate NBBO (National Best Bid and Offer - this is a regulation that requires brokers to execute customer trades at the best available ask price when buying securities, and the best available bid price when selling securities) matching trades to quotes for a given day

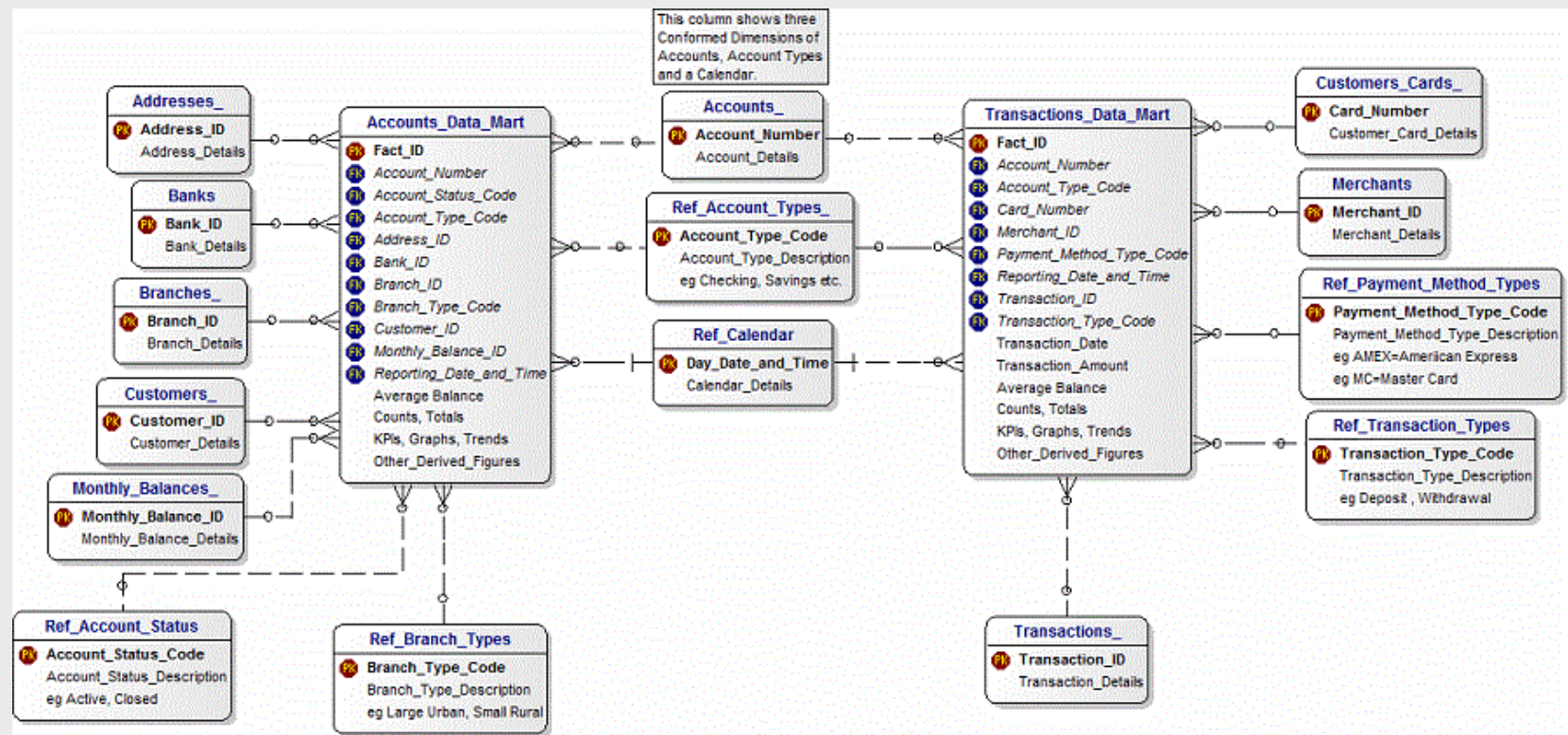
527 Course Review: Sample banking data model

http://www.databaseanswers.org/data_models/retail_banks/:



527 Course Review: Sample banking data model (cont.)

Another example with more facts and dimensions from
http://www.databaseanswers.org/data_models/retail_banks/:



527 Course Review: Data Analysis Methodology

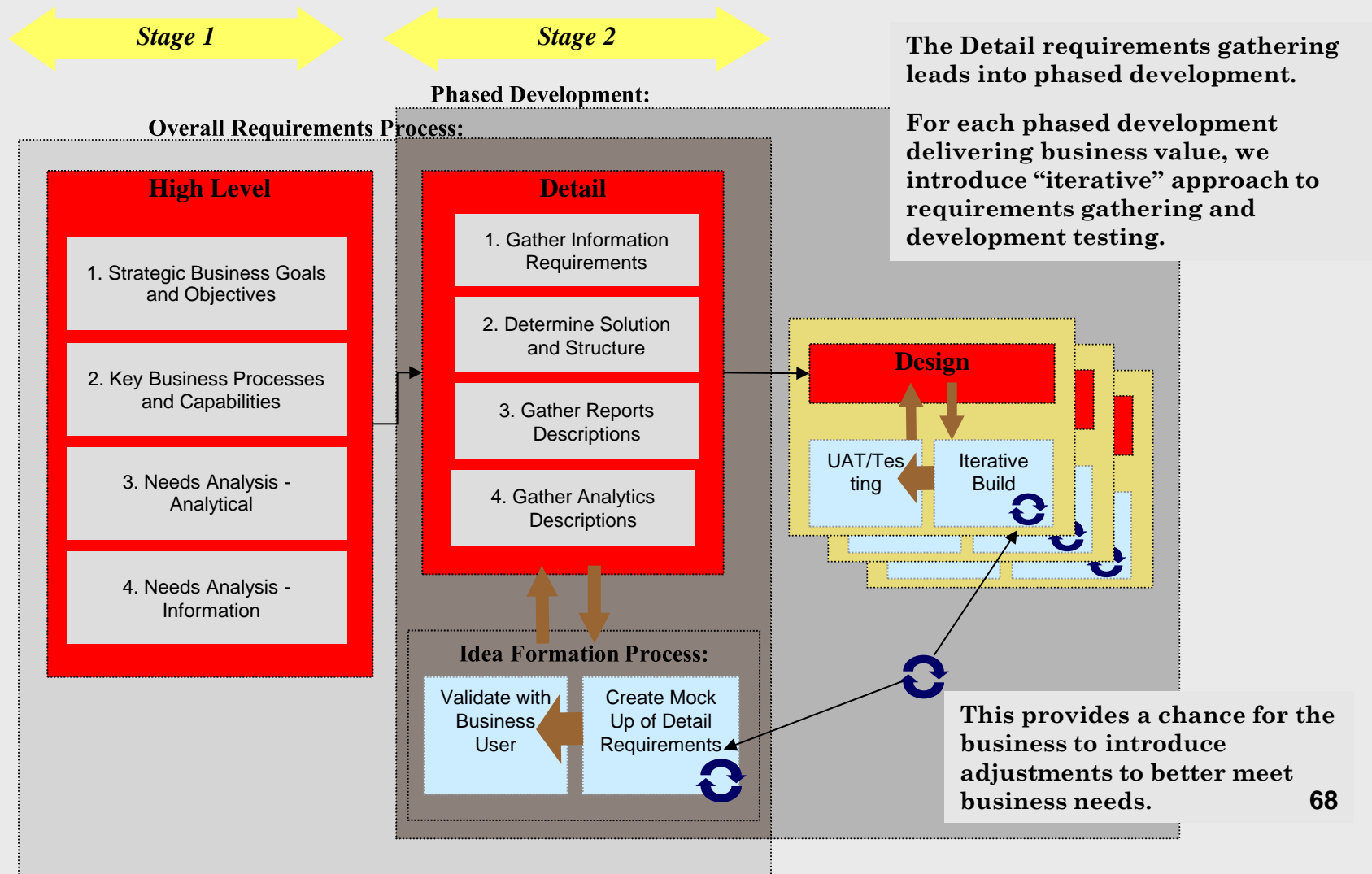
- A. Inspect - All data is inspected and “cleansed”
 - Records are investigated and fixed where appropriate e.g., outliers, type check, range check
 - Consistent default values are assigned to “missing” data
 - Data validation codes are included in the data where appropriate (e.g. invalid zip code) to avoid/compensate/exclude during analysis
- B. Transform - Data is standardized, improved or derived e.g., profitability, scores
 - Promotes consistent analysis using common business rules
 - Reduces analysis “programming” since necessary information is produced prior e.g. calculating months on books, banding score values, computing household product counts
 - Increases understanding of the data
- C. Integrate - All of the required data is in one logical structure e.g., transaction, position, account, customer, household, product, instrument, branch
 - Simplifies data access because all data is located in one location
 - Reduces analysis time since all of the information can be retrieved from a single location through a single query

527 Course Review:

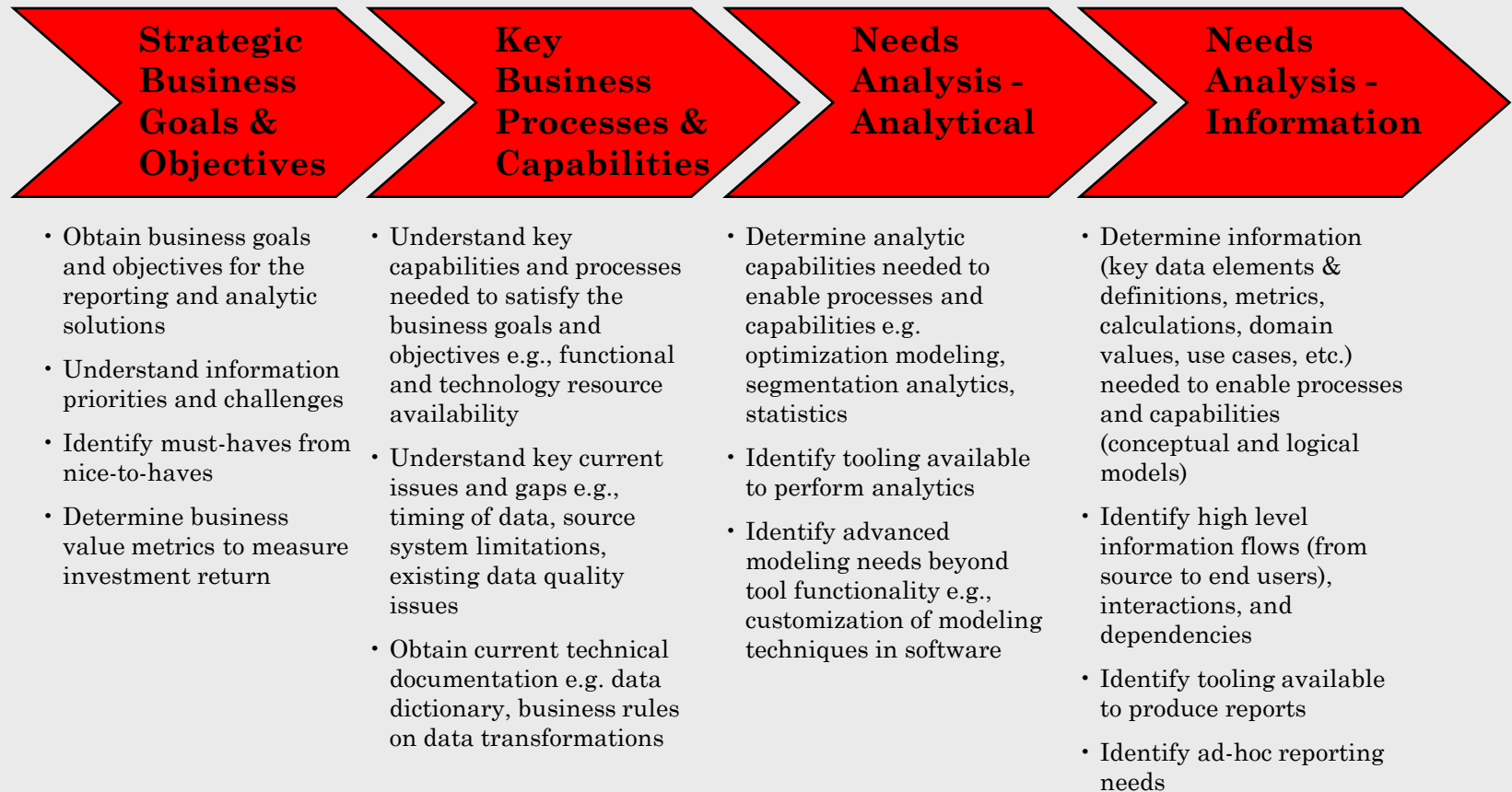
Data Analysis Methodology (cont.)

- D. Organize BI - Data is stored dimensionally
 - Simplifies analysis by providing an intuitive, business oriented data design
 - Enables pre-stored aggregations (cubes w/drill through to detail) to be easily developed and managed
- E. Organize A/DM – Data is stored in modeling specific data structure
 - This could be one de-normalized fact table with select dimensional information included in each row of data
 - Since data mining can only uncover patterns already present in the data, the sample should be large enough to contain significant information, yet small enough to process (dependent on resource capability)
- F. Explore - Search for anticipated relationships, unanticipated trends and anomalies:
 - Clustering discovers groups or structures in the data that are similar, beyond the structures known in the data
 - Classification generalizes a known structure to apply to new data, such as classifying a customer as a good or poor credit risk
- G. Document - Data definitions and transformation rules are documented and accessible
 - Able to understand the data and information gathered from the data

527 Course Review: Requirements Gathering Process

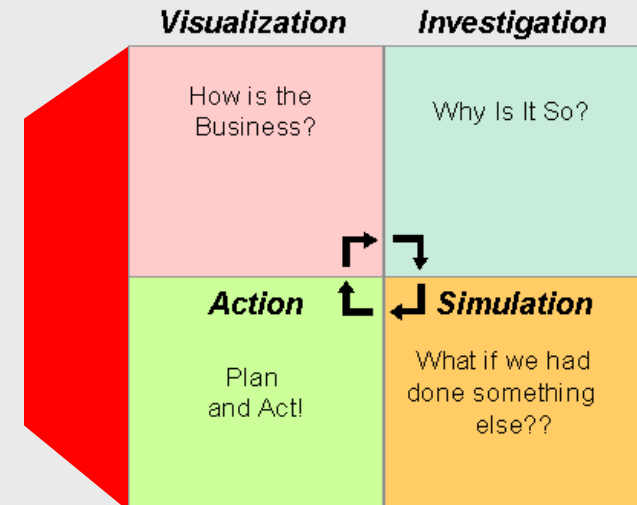


527 Course Review: High Level Requirements Gathering



527 Course Review: Understanding that it's iterative

An analytical environment supports not just reporting, but the full range of information usage listed below:



a) Basic Reporting

Periodic reports with the ability to change the report parameters by the users e.g., time period of reporting, metrics reported

b) Ad-hoc Analysis

Ability to access the data in free-form, create new aggregations and report definitions.

c) Custom Applications

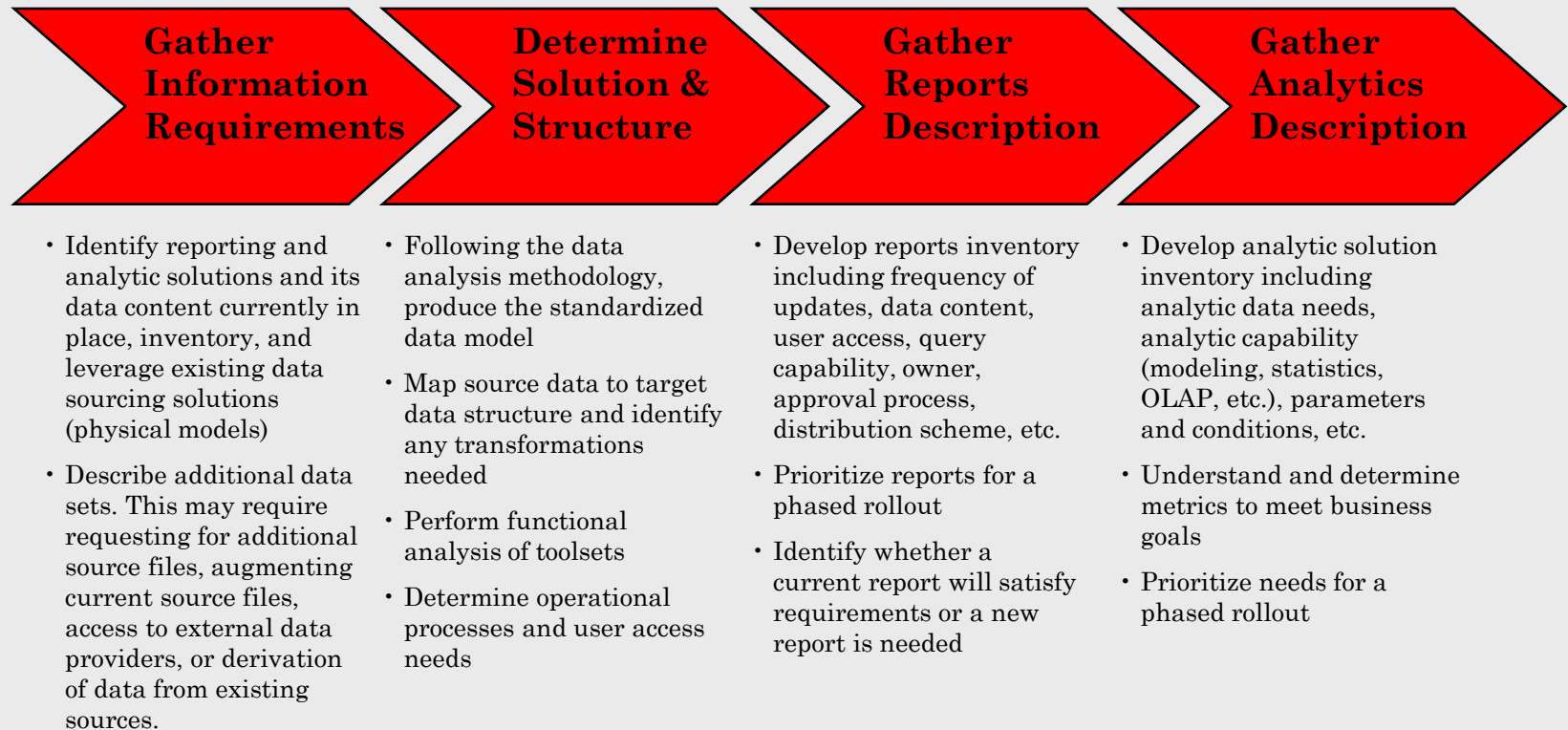
Enterprise application specific functional use e.g., application packages that are used in individual business areas for forecasting, finance, campaign management

d) Intense Analytics

Statistical analysis, modeling and data mining done on an iterative basis to generate segment definitions, offer specifications, credit policies, etc. Requires robust sampling, modeling, scoring and testing capabilities.

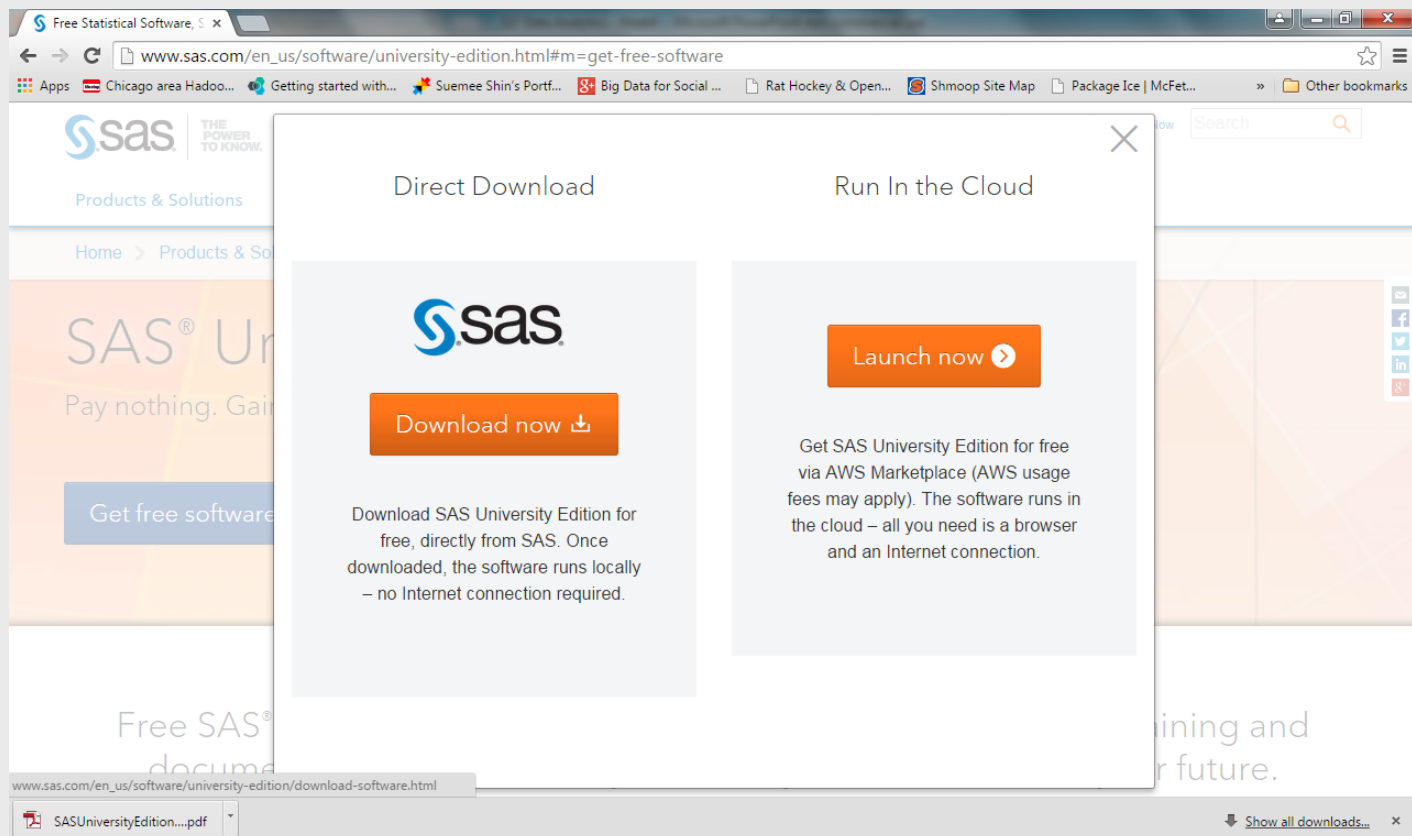
527 Course Review:

Detailed Requirements Gathering



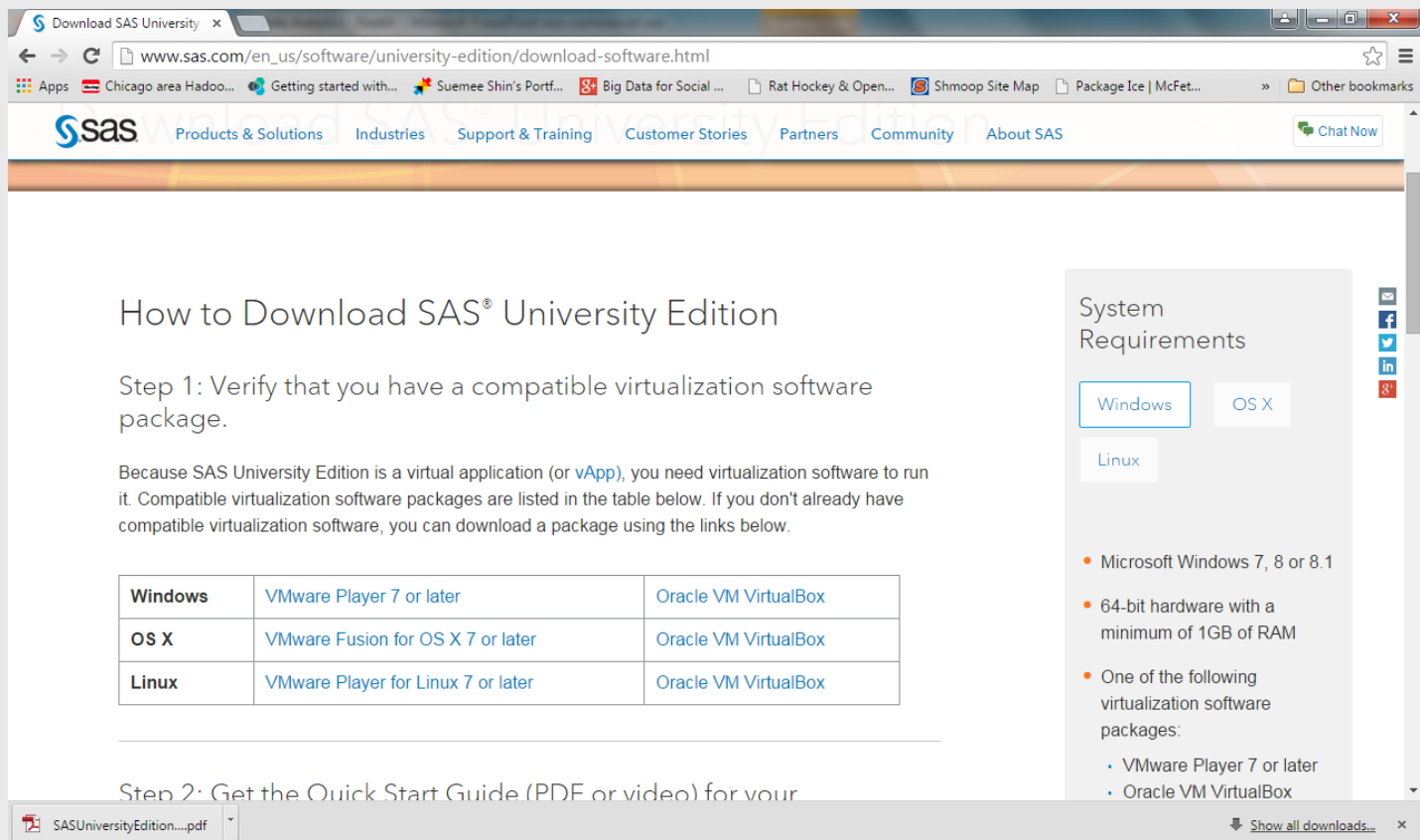
527 Course Review: Installing SAS University Edition

- ◆ Go to: http://www.sas.com/en_us/software/university-edition.html
- ◆ Create a student account. You will get free software and e-learning for one year.



527 Course Review: Getting the right download file

- ◆ Select your system specific download files. Recommend using VirtualBox for Mac users.



The screenshot shows the SAS University Edition download page. The browser address bar displays www.sas.com/en_us/software/university-edition/download-software.html. The page title is "How to Download SAS® University Edition".

Step 1: Verify that you have a compatible virtualization software package.

Because SAS University Edition is a virtual application (or [vApp](#)), you need virtualization software to run it. Compatible virtualization software packages are listed in the table below. If you don't already have compatible virtualization software, you can download a package using the links below.

Windows	VMware Player 7 or later	Oracle VM VirtualBox
OS X	VMware Fusion for OS X 7 or later	Oracle VM VirtualBox
Linux	VMware Player for Linux 7 or later	Oracle VM VirtualBox

Step 2: Get the Quick Start Guide (PDF or video) for your

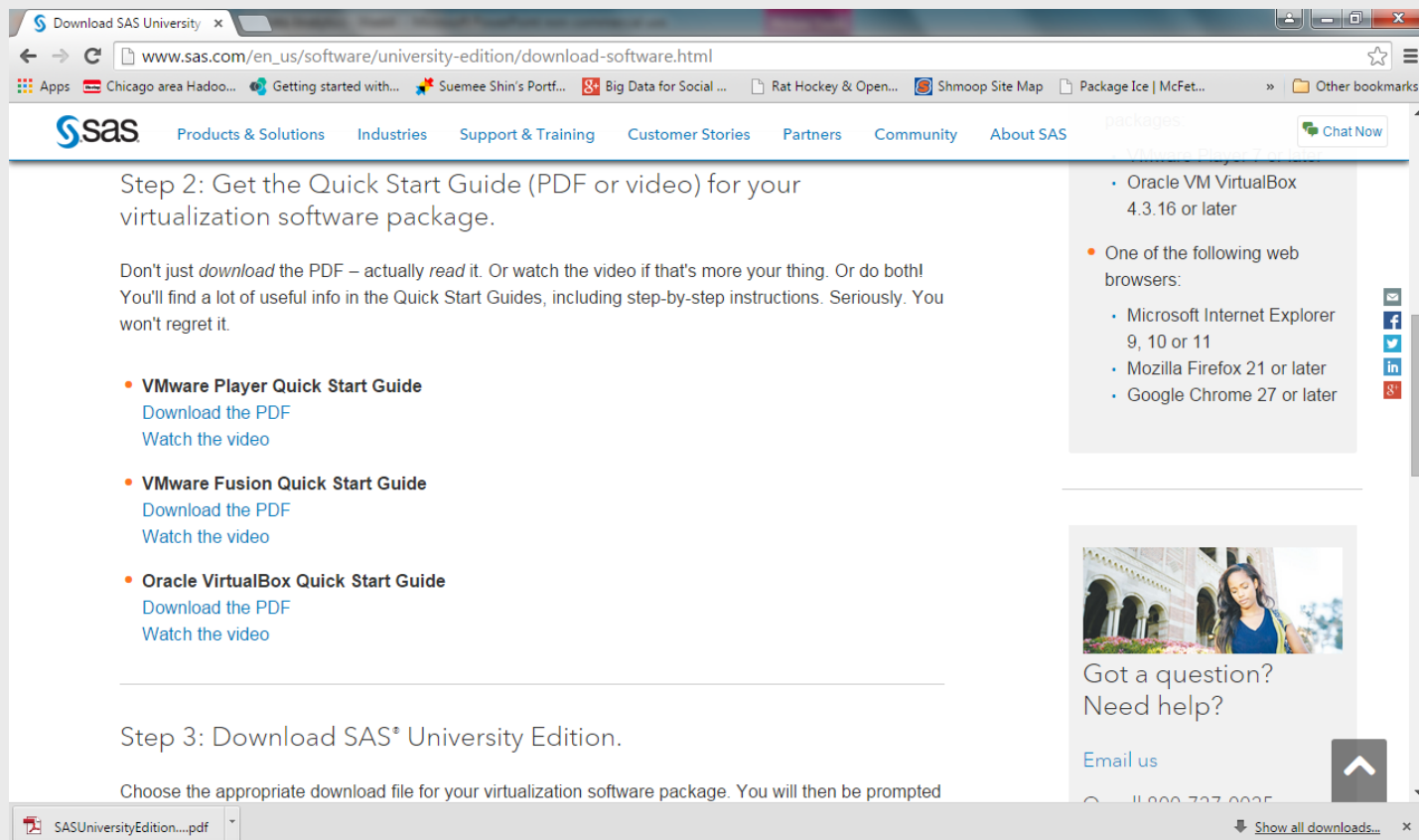
System Requirements

- Microsoft Windows 7, 8 or 8.1
- 64-bit hardware with a minimum of 1GB of RAM
- One of the following virtualization software packages:
 - VMware Player 7 or later
 - Oracle VM VirtualBox

A download bar at the bottom shows a file named "SASUniversityEdition....pdf".

527 Course Review: Follow installation directions

◆ Follow directions in the Quick Start Guide PDF



The screenshot shows a web browser window with the URL www.sas.com/en_us/software/university-edition/download-software.html. The page is titled "Download SAS University" and features the SAS logo and navigation links: Products & Solutions, Industries, Support & Training, Customer Stories, Partners, Community, and About SAS. A "Chat Now" button is also visible.

Step 2: Get the Quick Start Guide (PDF or video) for your virtualization software package.

Don't just *download* the PDF – actually *read* it. Or watch the video if that's more your thing. Or do both! You'll find a lot of useful info in the Quick Start Guides, including step-by-step instructions. Seriously. You won't regret it.

- **VMware Player Quick Start Guide**
[Download the PDF](#)
[Watch the video](#)
- **VMware Fusion Quick Start Guide**
[Download the PDF](#)
[Watch the video](#)
- **Oracle VirtualBox Quick Start Guide**
[Download the PDF](#)
[Watch the video](#)

Step 3: Download SAS® University Edition.

Choose the appropriate download file for your virtualization software package. You will then be prompted

On the right side, there is a "packages" section with the following content:

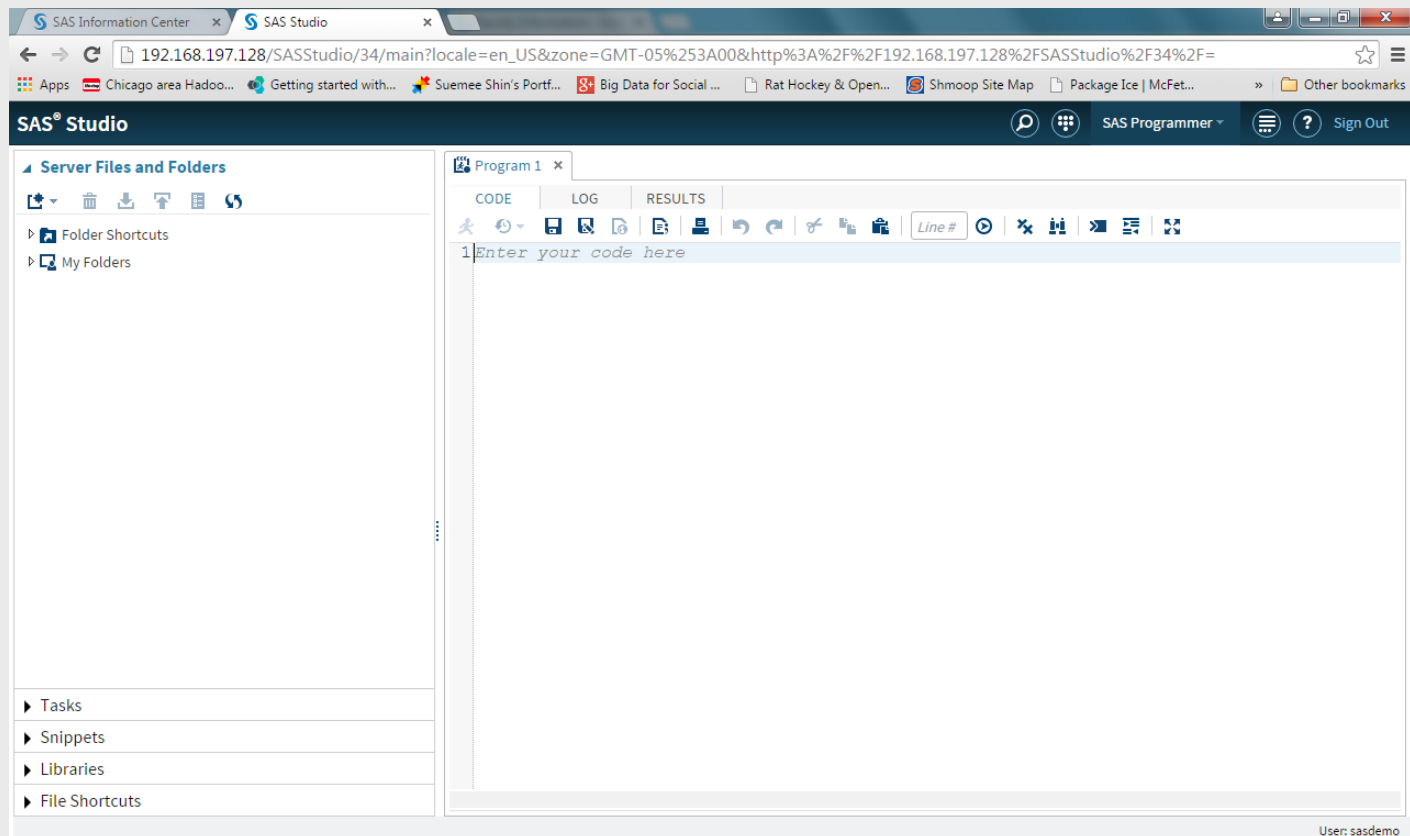
- Oracle VM VirtualBox 4.3.16 or later
- One of the following web browsers:
 - Microsoft Internet Explorer 9, 10 or 11
 - Mozilla Firefox 21 or later
 - Google Chrome 27 or later

Below this, there is a section titled "Got a question? Need help?" with a link to "Email us".

At the bottom, there is a download bar showing a file named "SASUniversityEdition....pdf".

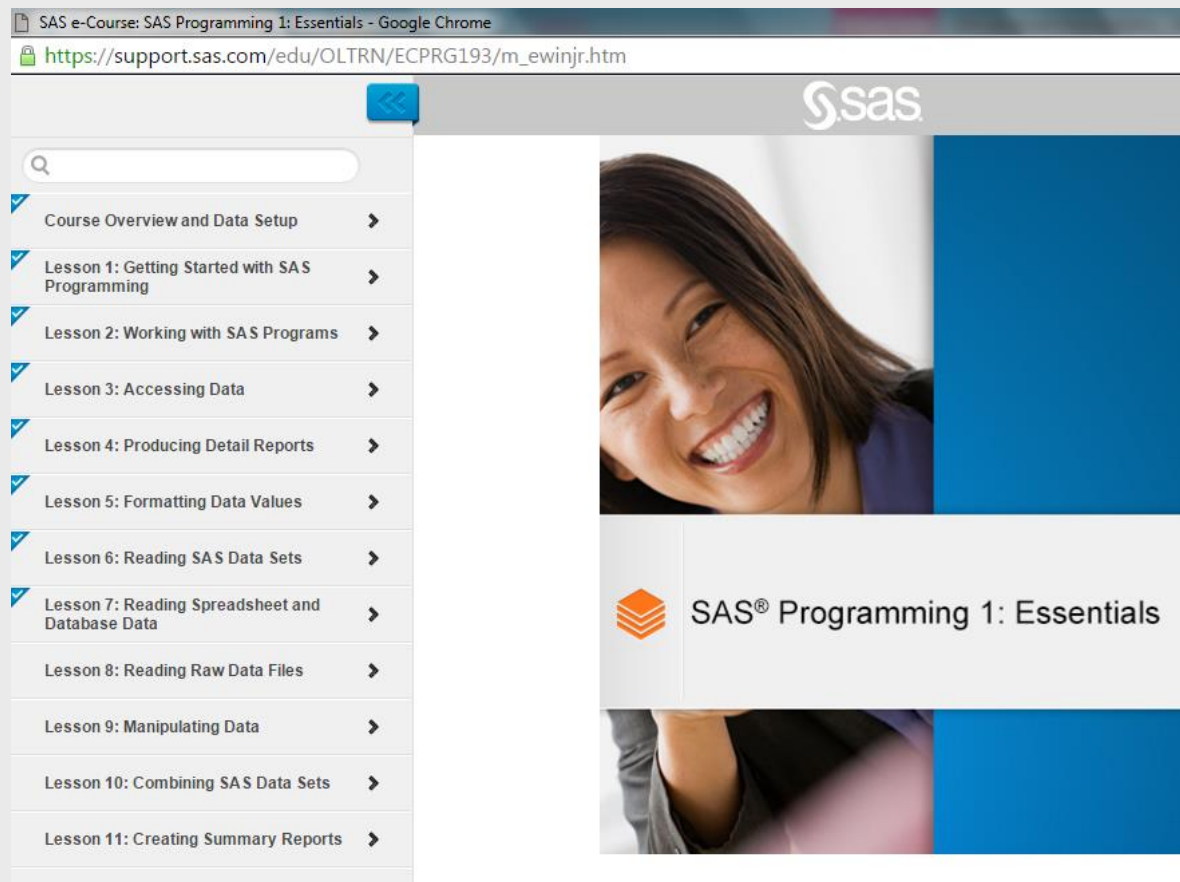
527 Course Review: SAS Studio – start window

- ◆ Get to the point of when you can open a start window:



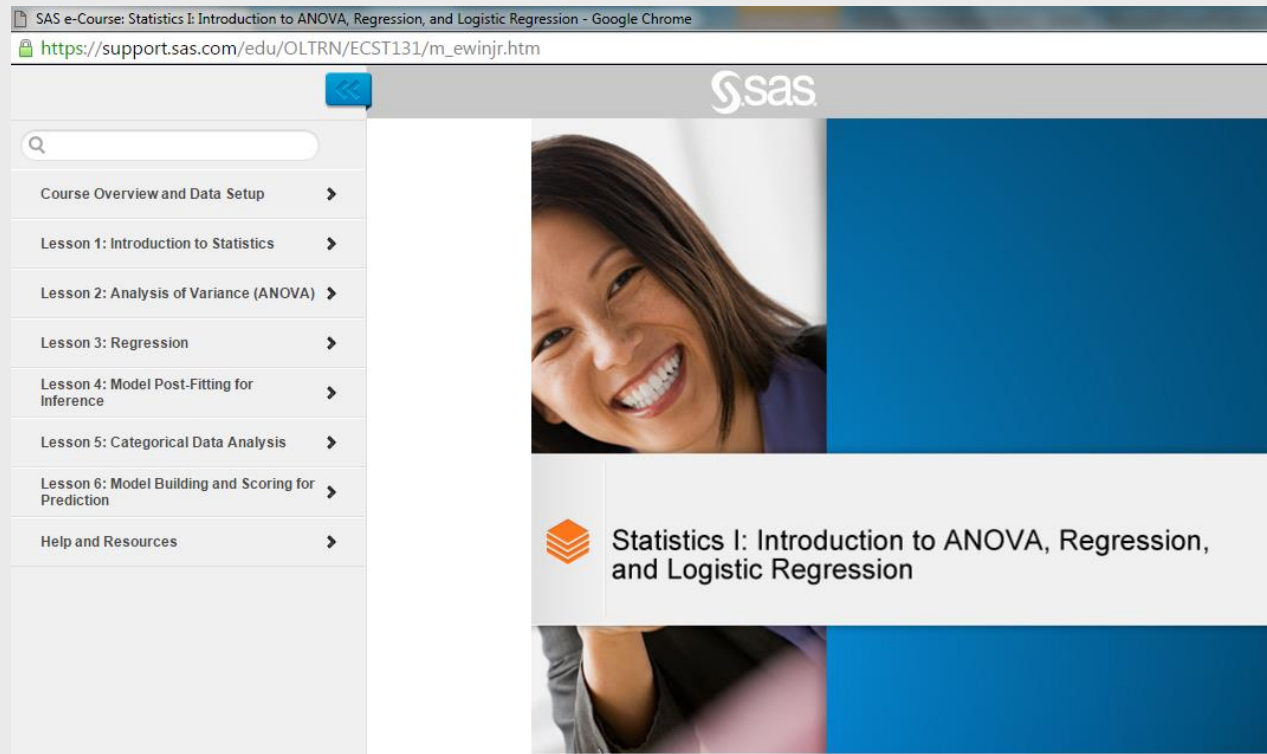
527 Course Review: SAS Programming I Essentials

- ◆ Review SAS Programming I Essentials e-learning course by end of week



527 Course Review: SAS Programming I Essentials

◆ Start Statistics 1 e-learning course!!



527 Course Review: SAS Installation of University Edition

Available Tools and Functions:

- ◆ Site name: 'UNIVERSITY EDITION 2.2 9.4M3 WITH ETS FOR PLAYER'.
- ◆ Expiration: 15JUN2016.
- ◆ Product expiration dates:
 - ◆ ---Base SAS Software 15JUN2016 (CPU A)
 - ◆ ---SAS/STAT 15JUN2016 (CPU A)
 - ◆ ---SAS/ETS 15JUN2016 (CPU A)
 - ◆ ---SAS/IML 15JUN2016 (CPU A)
 - ◆ ---SAS/ACCESS Interface to PC Files 15JUN2016 (CPU A)
 - ◆ ---SAS/IML Studio 15JUN2016 (CPU A)
 - ◆ ---SAS Workspace Server for Local Access 15JUN2016 (CPU A)
 - ◆ ---SAS Workspace Server for Enterprise Access 15JUN2016 (CPU A)
 - ◆ ---High Performance Suite 15JUN2016 (CPU A)

527 Course Review:

SAS Framework and File Types

SAS framework:

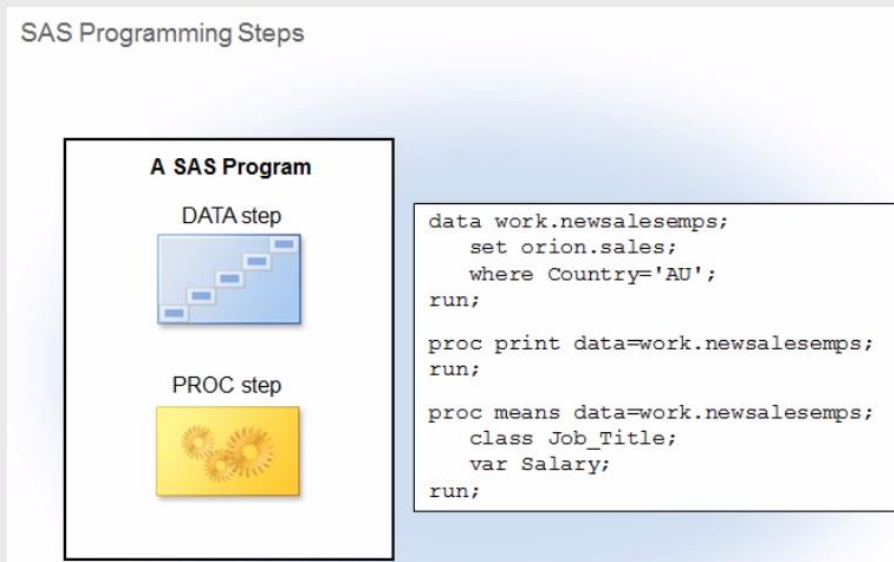
- ◆ **Access data:** Using SAS, you can read any kind of data.
- ◆ **Manage data:** SAS gives you excellent data management capabilities
- ◆ **Analyze data:** For statistical analysis, SAS is the gold standard.
- ◆ **Present data:** You can use SAS to present your data meaningfully.

Three major file types:

- ◆ **Raw data files** contain data that has not been processed by any other computer program. They are text files that contain one record per line, and the record typically contains multiple fields. Raw data files aren't reports; they are unformatted text.
- ◆ **SAS data sets** are specific to SAS. A SAS data set is data in a form that SAS can understand. Like raw data files, SAS data sets contain data. But in SAS data sets, the data is created only by SAS and can be read only by SAS.
- ◆ **SAS program files** contain SAS programming code. These instructions tell SAS how to process your data and what output to create. You can save and reuse SAS program files.

527 Course Review: SAS Steps

- ◆ A SAS program consists of DATA steps and PROC steps. A SAS programming step is comprised of a sequence of statements. Every step has a beginning and ending step boundary. SAS compiles and executes each step independently, based on the step boundaries.
- ◆ A SAS program can also contain global statements, which are outside DATA and PROC steps, and typically affect the SAS session. A TITLE statement is a global statement. After it is defined, a title is displayed on every report, unless the title is cleared or canceled.
- ◆ SAS statements usually begin with an identifying keyword, and always end with a semicolon. SAS statements are free format and can begin and end in any column. A single statement can span multiple lines, and there can be more than one statement per line. Unquoted values can be lowercase, uppercase, or mixed case. This flexibility can result in programs that are difficult to read.



527 Course Review: SAS Comments

- Comments are used to document a program and to mark SAS code as non-executing text. There are two types of comments: *block comments* and *comment statements*.

/ comment */*
** comment statement;*

```
/* create a temporary data
set, newsalesemps, from
the data set orion.sales */

data work.newsalesemps;
  set orion.sales;
  where Country='AU';
run;

proc print data=work.newsalesemps;
run;

proc means data=work.newsalesemps;
  class Job_Title;
  var Salary;
run;
```

/* comment */

- any length
- internal semicolons
- X** nested

```
*create a temporary data set,
newsalesemps, from the data set
orion.sales;

data work.newsalesemps;
  set orion.sales;
  *where Country='AU';
run;

proc print data=work.newsalesemps;
run;

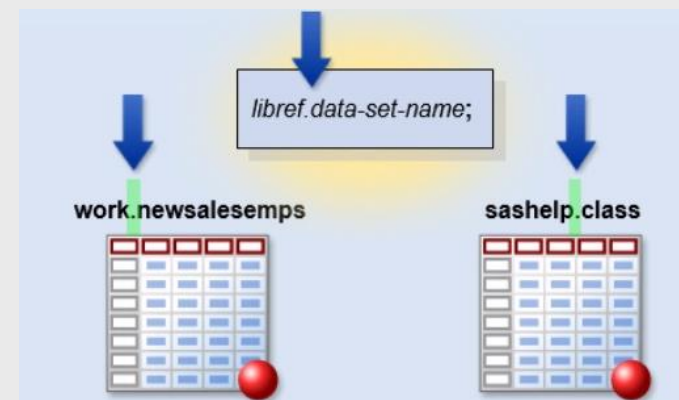
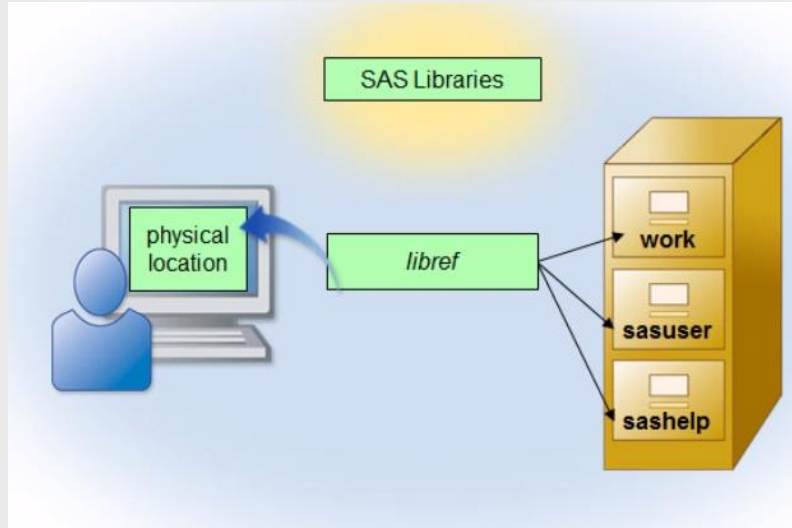
proc means data=work.newsalesemps;
  class Job_Title;
  var Salary;
run;
```

*** comment statement;**

- complete statements
- X** internal semicolons

527 Course Review: SAS Libraries

- ◆ SAS data sets are stored in SAS libraries. A SAS library is a collection of one or more SAS files that are recognized by SAS. SAS automatically provides one temporary and at least one permanent SAS library in every SAS session.
- ◆ **Work** is a temporary library that is used to store and access SAS data sets for the duration of the session. **Sasuser** and **sashelp** are permanent libraries that are available in every SAS session.
- ◆ You refer to a SAS library by a library reference name, or libref. A libref is a shortcut to the physical location of the SAS files.
- ◆ All SAS data sets have a two-level name that consists of the libref and the data set name, separated by a period. Data sets in the **work** library can be referenced with a one-level name, consisting of only the data set name, because **work** is the default library. Data sets in permanent libraries must be referenced with a two-level name.



527 Course Review: SAS Libraries (cont.)

- ◆ You can create and access your own SAS libraries. User-defined libraries are permanent but are not automatically available in a SAS session. You must assign a libref to a user-created library to make it available. You use a LIBNAME statement to associate the libref with the physical location of the library, that is, the physical location of your data. You can submit the LIBNAME statement alone at the start of a SAS session, or you can store it in a SAS program so that the SAS library is defined each time the program runs. If your program needs to reference data sets in multiple locations, you can use multiple LIBNAME statements.
- ◆ In an interactive SAS session, a libref remains in effect until you cancel it, change it, or end your SAS session. To cancel a libref, you submit a LIBNAME statement with the CLEAR option. This clears or disassociates a libref that was previously assigned. To specify a different physical location, you submit a LIBNAME statement with the same libref name but with a different filepath.
- ◆ When a SAS session ends, everything in the **work** library is deleted. The librefs are also deleted. Remember that the contents of permanent libraries still exist in the operating environment, but each time you start a new SAS session, you must resubmit the LIBNAME statement to redefine a libref for each user-created library that you want to access.

```
LIBNAME libref 'SAS-library' <options>;
```

```
LIBNAME libref CLEAR;
```

527 Course Review:

SAS PROC CONTENTS AND PRINT

- ◆ Use PROC CONTENTS with *libref._ALL_* to display the contents of a SAS library. The report will list all the SAS files contained in the library, as well as the descriptor portion of each data set in the library. Use the NODS option in the PROC CONTENTS statement to suppress the descriptor information for each data set.

```
PROC CONTENTS DATA=libref._ALL_ NODS;  
RUN;
```

```
PROC CONTENTS DATA=libref.SAS-data-set;  
RUN;
```

- ◆ After associating a libref with a permanent library, you can write a PROC PRINT step to display a SAS data set within the library.

```
PROC PRINT DATA=libref.SAS-data-set;  
RUN;
```

527 Course Review: SAS Data Sets

- ◆ SAS data sets are specially structured data files that SAS creates and that only SAS can read. A SAS data set is displayed as a table composed of variables and observations. A SAS data set contains a descriptor portion and a data portion.
- ◆ The descriptor portion contains general information about the data set (such as the data set name and the number of observations) and information about the variable attributes (such as name, type, and length). There are two types of variables: **character** and **numeric**. A character variable can store any value and can be up to 32,767 characters long. Numeric variables store numeric values in floating point or binary representation in 8 bytes of storage by default. Other attributes include formats, informats, and labels. You can use PROC CONTENTS to browse the descriptor portion of a data set.
- ◆ The data portion contains the data values. Data values are either **character** or **numeric**. A valid value must exist for every variable in every observation in a SAS data set. A missing value is a valid value in SAS. A missing character value is displayed as a **blank**, and a missing numeric value is displayed as a **period**. You can specify an alternate character to print for missing numeric values using the MISSING= SAS system option. You can use PROC PRINT to display the data portion of a SAS data set.
- ◆ SAS variable and data set names must be 1 to 32 characters in length and start with a **letter** or **underscore**, followed by letters, underscores, and numbers. Variable names are not case sensitive.

/salesemps

Last_Name	Job_Title	Salary
Gentry	Sales Rep. II	26780
Betschkus	Sales Rep. IV	.
Don		26955
Altman	Sales Rep. II	27440

missing values

527 Course Review:

VAR and SUM Statements

- ◆ You can use the VAR statement in a PROC PRINT step to subset the variables in a report. You specify the variables to include and list them in the order in which they are to be displayed.
- ◆ You can use the SUM statement in a PROC PRINT step to calculate and display report totals for the requested numeric variables.

```
PROC PRINT DATA=SAS-data-set;  
    VAR variable(s);  
    SUM variable(s);  
RUN;
```

527 Course Review:

WHERE Statement

- ◆ The WHERE statement in a PROC PRINT step subsets the observations in a report. When you use a WHERE statement, the output contains only the observations that meet the conditions specified in the WHERE expression. This expression is a sequence of operands and operators that form a set of instructions that define the condition. The operands can be constants or variables. Remember that variable operands must be defined in the input data set. Operators include comparison, arithmetic, logical, and special WHERE operators.

Comparison:

Symbol(s)	Mnemonic	Definition
=	EQ	equal to
^= ^= ~=	NE	not equal to
>	GT	greater than
<	LT	less than
>=	GE	greater than or equal to
<=	LE	less than or equal to
	IN	equal to one of a list

Examples

```
where Gender='M';  
where Gender eq 'M';  
where Salary ne .;  
where Salary>50000;  
where Salary lt 50000;  
where Salary<=60000;  
where Country in ('AU','US');
```

527 Course Review: WHERE Statement (Cont.)

Arithmetic:

Symbol	Definition
**	exponentiation
*	multiplication
/	division
+	addition
-	subtraction

Example

```
where Salary+Bonus<=10000;
```

Logical:

WHERE *where-expression-1* AND | OR
where-expression-n;

Symbol(s)	Mnemonic	Definition
&	AND	logical <i>and</i>
	OR	logical <i>or</i>
^ ~	NOT	logical <i>not</i>

Examples

```
where Country ne 'AU' and Salary>=50000;
where Gender eq 'M' or Salary ge 50000;
where Country='AU' | Country='US';
where Country in ('AU' 'US');
where Country not in ('AU', 'US');
```


527 Course Review: WHERE Statement (Cont.)

Contains:

WHERE *where-expression*;

Symbol	Mnemonic	Definition
?	CONTAINS	includes a substring

Examples

```
where Country='AU' and
      Job_Title contains 'Rep';
```

```
where Country='AU' and
      Job_Title ? 'Rep';
```

case sensitive

Mnemonic:

Mnemonic	Definition
BETWEEN-AND	an inclusive range
WHERE SAME AND	augment a where expression
IS NULL	a missing value
IS MISSING	a missing value
LIKE	matches a pattern

Symbol	Replaces
%	any number of characters
-	one character

- 1) Use of ~not~ to exclude
- 2) Use of ~ same and~ to augment
- 3) IS NULL and IS MISSING can be used for both numeric and character variables

527 Course Review: Sorting and Grouping

- ◆ The SORT procedure sorts the observations in a data set. You can sort on one variable or multiple variables, sort on character or numeric variables, and sort in ascending or descending order. By default, SAS replaces the original SAS data set unless you use the OUT= option to specify an output data set. PROC SORT does not generate printed output.
- ◆ Every PROC SORT step must include a BY statement to specify one or more BY variables. These are variables in the input data set whose values are used to sort the data. By default, SAS sorts in ascending order, but you can use the keyword DESCENDING to specify that the values of a variable are to be sorted in descending order. When your SORT step has multiple BY variables, some variables can be in ascending and others in descending order.
- ◆ You can also use a BY statement in PROC PRINT to display observations grouped by a particular variable or variables. The groups are referred to as BY groups. Remember that the input data set must be sorted on the variables specified in the BY statement.

```
PROC SORT DATA=input-SAS-data-set  
             <OUT=ouput-SAS-data-set>;  
             BY <DESCENDING> by-variable(s);  
RUN;
```

527 Course Review:

ID, TITLE, FOOTNOTE Statement

- ◆ You can use the ID statement in a PROC PRINT step to specify a variable to print at the beginning of the row instead of an observation number. The variable that you specify replaces the Obs column.

```
ID variable(s);
```

- ◆ You can enhance a report by adding titles, footnotes, and column labels. Use the global TITLE statement to define up to 10 lines of titles to be displayed at the top of the output from each procedure. Use the global FOOTNOTE statement to define up to 10 lines of footnotes to be displayed at the bottom of the output from each procedure.

```
TITLEn 'text';  
FOOTNOTEn 'text';
```

- ◆ Titles and footnotes remain in effect until you change or cancel them, or until you end your SAS session. Use a null TITLE statement to cancel all titles, and a null FOOTNOTE statement to cancel all footnotes.

527 Course Review:

LABEL Statement

- ◆ Use the LABEL statement in a PROC PRINT step to define temporary labels to display in the report instead of variable names. Labels can be up to 256 characters in length. Most procedures use labels automatically, but PROC PRINT does not. Use the LABEL option in the PROC PRINT statement to tell SAS to display the labels. Alternatively, the SPLIT= option tells PROC PRINT to use the labels and also specifies a split character to control line breaks in column headings.

```
PROC PRINT DATA=SAS-data-set LABEL;  
    LABEL variable='label'  
          variable='label'  
          ... ;  
RUN;
```

```
SPLIT='split-character';
```

527 Course Review:

FORMAT Statement

- ◆ A format is an instruction that tells SAS how to display data values in output reports. You can add a **FORMAT** statement to a PROC PRINT step to specify temporary SAS formats that control how values appear in the report. There are many existing SAS formats that you can use. Character formats begin with a dollar sign, but numeric formats do not.

FORMAT *variable(s) format;*

- ◆ SAS stores date values as the number of days between January 1, 1960, and a specific date. To make the dates in your report recognizable and meaningful, you must apply a SAS date format to the SAS date values.

527 Course Review: FORMAT Statement Examples

Format	Stored Value	Displayed Value
MMDDYY6.	0	010160
MMDDYY8.	0	01/01/60
MMDDYY10.	0	01/01/1960
DDMMYY6.	365	311260
DDMMYY8.	365	31/12/60
DDMMYY10.	365	31/12/1960

Format	Stored Value	Displayed Value
\$4.	Programming	Prog
12.	27134.5864	27135
12.2	27134.5864	27134.59
COMMA12.2	27134.5864	27,134.59
DOLLAR12.2	27134.5864	\$27,134.59
COMMAX12.2	27134.5864	27.134,59
EUROX12.2	27134.5864	€27.134,59

Format	Definition
\$w.	writes standard character data.
w.d	writes standard numeric data.
COMMAw.d	writes numeric values with a comma that separates every three digits and a period that separates the decimal fraction.
DOLLARw.d	writes numeric values with a leading dollar sign, a comma that separates every three digits, and a period that separates the decimal fraction.
COMMAXw.d	writes numeric values with a period that separates every three digits and a comma that separates the decimal fraction.
EUROXw.d	writes numeric values with a leading euro symbol (€), a period that separates every three digits, and a comma that separates the decimal fraction.

527 Course Review:

User Defined FORMAT Statement

- ◆ You can create your own user-defined formats. When you create a user-defined format, you don't associate it with a particular variable or data set. Instead, you create it based on values that you want to display differently. The formats will be available for the remainder of your SAS session. You can apply user-defined formats to a specific variable in a PROC PRINT step.
- ◆ You use the FORMAT procedure to create a format. You assign a format name that can have up to 32 characters. The name of a character format must begin with a dollar sign, followed by a letter or underscore, followed by letters, numbers, and underscores. Names for numeric formats must begin with a letter or underscore, followed by letters, numbers, and underscores. A format name cannot end in a number and cannot be the name of a SAS format.
- ◆ You use a VALUE statement in a PROC FORMAT step to specify the way that you want the data values to appear in your output. You define value-range sets to specify the values to be formatted and the formatted values to display instead of the stored value or values. The value portion of a value-range set can include an individual value, a range of values, a list of values, or a keyword. The keyword OTHER is used to define a value to display if the stored data value does not match any of the defined value-ranges.
- ◆ When you define a numeric format, it is often convenient to use numeric ranges in the value-range sets. Ranges are inclusive by default. To exclude the endpoints, use a less-than symbol after the low end of the range or before the high end.
- ◆ The LOW and HIGH keywords are used to define a continuous range when the lowest and highest values are not known. Remember that for character values, the LOW keyword treats missing values as the lowest possible values. However, for numeric values, LOW does not include missing values.

```
PROC FORMAT;  
  VALUE format-name value-or-range1='formatted-value1'  
                                value-or-range2='formatted-value2'  
                                ...;  
RUN;
```

527 Course Review: Using VALUE Statement

Using the VALUE Statement

```
PROC FORMAT;  
  VALUE format-name value-or-range1= 'formatted-value1 '  
                                value-or-range2= 'formatted-value2 '  
                                ...;  
RUN;
```

value-range sets

	<i>value-or-range</i>	=	<i>formatted-value</i>
value →	'AU' 1	=	'Australia'
range →	'B'-'D' 0-50000	=	'Tier 1'
list →	'U','V' 1,2,3	=	'Below 49.9'

527 Course Review:

Creating new datasets

- ◆ You use a DATA step to create a new SAS data set from an existing SAS data set. The DATA step begins with a DATA statement, which provides the name of the SAS data set to create. Include a SET statement to name the existing SAS data set to be read in as input.
- ◆ You use the WHERE statement to subset the input data set by selecting only the observations that meet a particular condition. To subset based on a SAS date value, you can use a SAS date constant in the WHERE expression. SAS automatically converts a date constant to a SAS date value.

```
DATA output-SAS-data-set;  
    SET input-SAS-data-set;  
    WHERE where-expression;  
RUN;
```

- ◆ By default, the SET statement reads all of the observations and variables from the input data set and writes them to the output data set. You can customize the new data set by selecting only the observations and variables that you want to include. You can use a WHERE statement to select the observations, as long as the variables included in the condition come from the input data set. You can use a DROP statement to list the variables to exclude from the new data set, or use a KEEP statement to list the variables to include. If you use a KEEP statement, you must include every variable to be written, including any new variables.

```
DROP variable-list;  
KEEP variable-list;
```

527 Course Review:

Creating new datasets (cont.)

- ◆ You can subset the original data set with a WHERE statement for variables that are defined in the input data set, and a subsetting IF statement for new variables that are created in the DATA step. Remember that, although IF expressions are similar to WHERE expressions, you cannot use special WHERE operators in IF expressions.

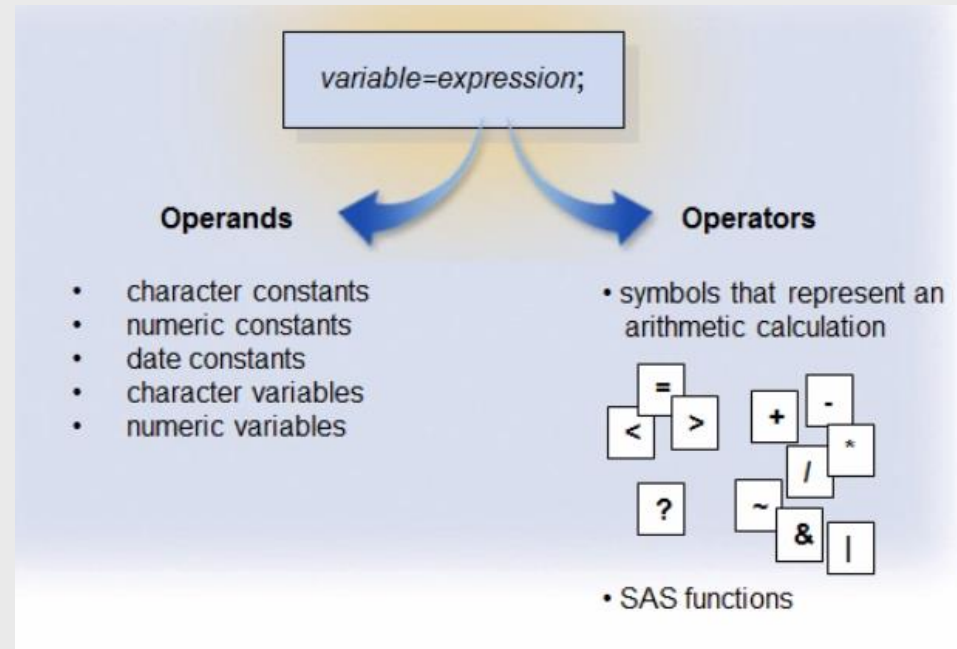
IF expression;

- ◆ To subset observations in a PROC step, you must use a WHERE statement. You cannot use a subsetting IF statement in a PROC step. To subset observations in a DATA step, you can always use a subsetting IF statement. However, a WHERE statement can make your DATA step more efficient because it subsets on input.
- ◆ When you use the LABEL statement in a DATA step, SAS permanently associates the labels to the variables by storing the labels in the descriptor portion of the data set. Using a FORMAT statement in a DATA step permanently associates formats with variables. The format information is also stored in the descriptor portion of the data set. You can use PROC CONTENTS to view the label and format information. PROC PRINT does not display permanent labels unless you use the LABEL or SPLIT= option.

527 Course Review: Assignment Statement

- ◆ You use an assignment statement to create a new variable. The assignment statement evaluates an expression and assigns the resulting value to a new or existing variable. The expression is a sequence of operands and operators. If the expression includes arithmetic operators, SAS performs the numeric operations based on priority, as in math equations. You can use parentheses to clarify or alter the order of operations.

variable=expression;



527 Course Review: Assignment Statement Examples

- ◆ Watch for order of execution:

Symbol	Definition	Priority
**	exponentiation	I
*	multiplication	II
/	division	II
+	addition	III
-	subtraction	III

Example	Type
Salary=26960;	numeric constant
Gender='F';	character constant
Hire_Date='21JAN1995'd;	date constant
Bonus=Salary*.10;	arithmetic expression
BonusMonth=month(Hire_Date);	SAS function