

A photograph of a modern university campus. In the background is a tall, curved glass building. In the foreground, there is a large, circular fountain with multiple water jets. The fountain is surrounded by a paved walkway and a modern, curved metal structure that looks like a pergola or a walkway cover. The entire image has a teal/green color overlay.

# **DSBA 6100/MBA 7090**

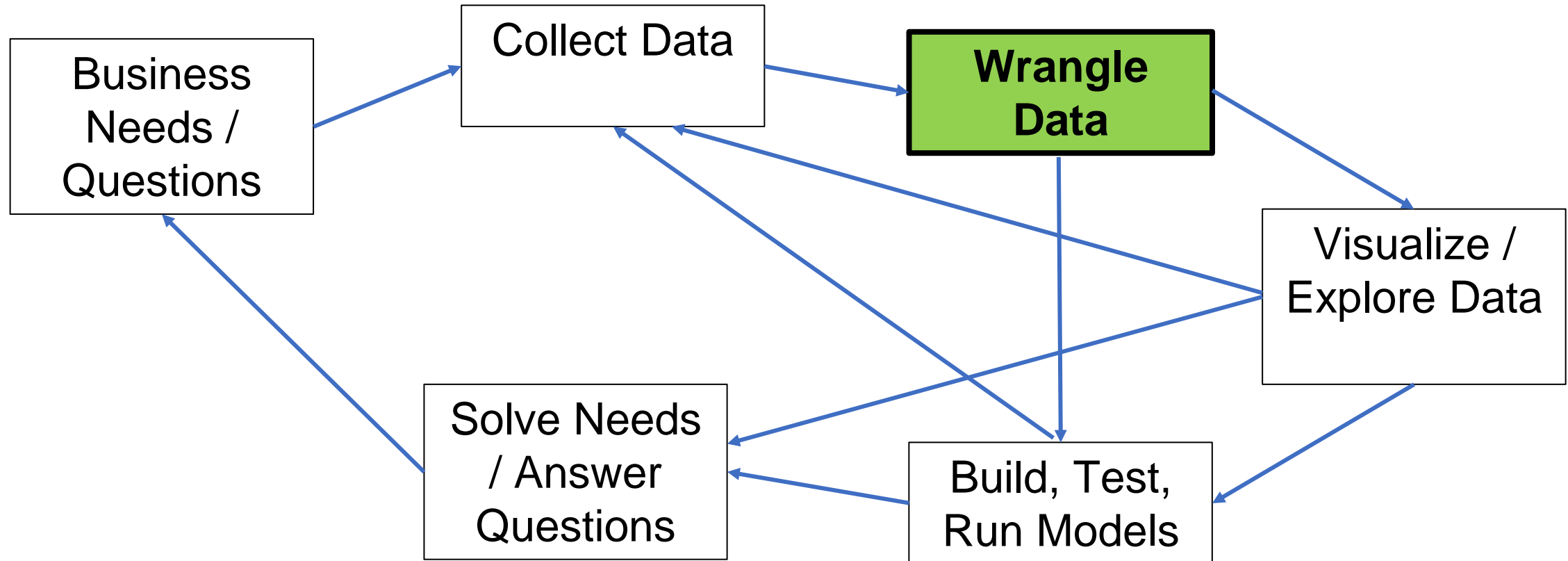
## **Week 4 – Data Sourcing & Management for Analytics**

Dr. D. Blaine Nashold, Jr.

# Outline

- What is data wrangling?
- Why is data wrangling important?
- Steps in data wrangling
- Data wrangling tools

# Big Data Analytics as a Process



# What is Data Wrangling?

- It is another name for “data preparation” for analytics modeling
- Also called “data munging” and “data janitor work”
- It is the process of converting or mapping data from one format into another, cleaning the data, merging, and filtering so it is ready for modeling and analysis
- Goal of data wrangling is to provide a dataset to facilitate analysis – reporting, visualization, prediction, etc.
- Involves understanding the data sources, the formats, pitfalls of extracting from one file format to another
- Most labor-intensive step in the analytics life-cycle
- Many steps in this process are iterative

# Why is data wrangling an important issue?

- Data preparation is an important step before any useful analytics can be done on the data
- Data from multiple sources can provide important clues about a company's business and customers
  - For example, in the food industry, data about production volumes, location data on shipments, weather reports, daily retail sales and social network comments can be combined to have better insights into customer sentiment and demand
- Data scientists spend anywhere from 50% to 80% of their time in collecting and preparing the data before it can be used in analytics models
- In many firms, this involves quite a bit of manual effort, creating bottlenecks for the more useful descriptive or predictive models

# Basic Data Terminology

- For analytics and visualization, we collect data about various attributes of multiple instances of an entity
- A variable stores data about an attribute (e.g., age)
- An observation stores data about different attributes (e.g., name, age, past purchase amount) for a particular instance of the entity (e.g., a customer named “Chandra”)
- A dataset is a collection of attribute values of many many instances of the entity (e.g., all customers)
- For most analytics modeling, each variable is expected to store a fixed type of data (e.g., numeric or text). Hence, when a variable is specified, its type is also specified.

# Basic Data Terminology

- Variables are also classified according to the measurement scale used
- Nominal variables store values which have no logical ordering or sequence
  - For example, the variable “marital status” may store values “married”, “single”, “divorced”
  - There is no logical ordering of the above values
- Ordinal variables store values which have a logical ordering
  - For example, the variable “drink size” may store values “small”, “medium” and “large”
  - There is an order where we know “medium” is more than “small”, and “large” is more than “medium”
  - However, it may not be the case that the difference between “medium” and “large” is the same as the difference between “small” and “medium”



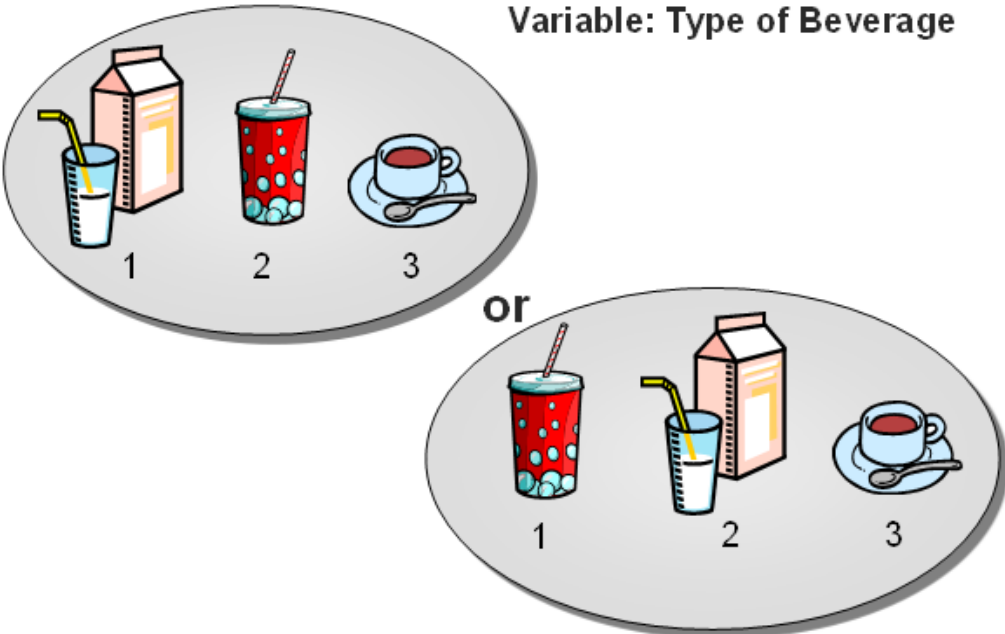


# Basic Data Terminology

sas THE POWER TO KNOW

## Nominal Variables

Variable: Type of Beverage




6

sas THE POWER TO KNOW

## Ordinal Variables

Variable: Size of Beverage



Small Medium Large



# Basic Data Terminology

- Continuous variables store numerical values, which have a logical order and the relative distances between the values are meaningful
  - For example, the volume of a beverage (in oz) can be stored as continuous variable
  - The value 20 (in oz) comes after 18 (in oz)
  - The difference between 20oz and 18oz is the same as the difference between 18oz and 16oz
- Continuous variable can have “interval” or “ratio” level
  - A ratio level has a true zero point (e.g., volume, since there is zero volume which is absence of volume)
  - An interval level has no true zero point (e.g., Celsius temp, where zero Celsius does not imply absence of heat)

# Data Wrangling Challenges

- Multiple formats of data
  - Data is being sourced from multiple and disparate source
    - Web, sensors, smartphones, data warehouses, databases
  - Data needs to be converted into some unified form appropriate for the analytics algorithms or platforms
- Ambiguity in recorded data
  - Data in most free-format storages, such as text descriptions or survey responses are prone to ambiguities and misspellings
    - For example, the FDA, NIH and pharma companies use different terms to describe the same side effects leading to problems when documents from these entities are merged to create a dataset

# Data Wrangling Challenges

- ❑ Some causes of messiness in dataset
  - ❑ Column headers are values, not variable names
  - ❑ Multiple variables stored in one column
  - ❑ Variables stored in both rows and columns
  - ❑ Multiple types of experimental units stored in same table
  - ❑ One type of experimental unit stored in multiple tables
- ❑ What is a tidy dataset?
  - ❑ Observations are in rows
  - ❑ Variables are in columns
  - ❑ Data contained in a single dataset

# Steps in Data Wrangling\*

- Prepare the analytics sandbox
  - Sandbox is the workspace created for exploring data without interfering with live production databases
  - Sandbox can hold all formats of data, including raw data, aggregated data, transformed data
  - Sandbox can grow to be very large, since copies of the data are created to serve multiple analysis models
- Perform ETLT
  - Data is extracted in its raw form and loaded into the data store in the sandbox, where it can then be transformed, if necessary

# The Analytics Sandbox

- Many analytics projects are performed in an analytics sandbox
- Data for modeling is loaded into the sandbox (or workspace) so analysis can be done without interfering with production environment
- Example: A fraud analytics workspace will get a copy of the customer and financial data than directly connecting to the production databases
- It is important for the analytics team to collaborate with IT to balance its need for more data with IT's need for proper data control

# Steps in Data Wrangling\*

- Perform ETLT (continued)
  - Big data platforms like Hadoop & MapReduce are used to move large datasets to the sandbox and do some analysis
  - Advisable to make an inventory of the data available and compare it to the data needed for analysis
  - If external data is needed, APIs are a popular way to access data sources (e.g., Twitter API)
- Learn about the data
  - Meta data or data dictionaries have info about the data fields
  - Provides context to the data and what to expect from analysis

# Steps in Data Wrangling\*

- Condition the data
  - Process of cleaning, normalizing, and performing transformations specific to individual analysis (e.g., creating the outcome variable for prediction)
  - Though book says that this step is done only by IT, data owners, DBA, or data engineer (the data scientist is involved), I think most data scientists have (or expected to have) data skills to do many of the conditioning steps
- Survey and visualize the data
  - Leverage visualization tools to get an overview of the data's characteristics (e.g., unexpected values or ranges)



# Steps in Data Wrangling\*

- Considerations in data conditioning
  - What are data sources and target fields?
  - How clean is the data?
  - How consistent are the files and content? Missing values or deviations from normal?
  - Consistency of data types?
  - Do values in the columns make sense? E.g., negative age or income
  - Evidence of systematic error? Data captured with shifted columns, repurposed column, data capture stopped midway

# Steps in Data Wrangling\*

- What to look for when surveying and visualizing data?
  - Calculations remained consistent within columns and across table
  - Data distributions are acceptable for analysis
  - Granularity or aggregation of data as needed for analysis
  - Population of interest is represented in the data
  - Appropriate time-related measurements as needed for analysis
  - Scales and units are consistent

# Data wrangling tools

- Free tools (source: <http://blog.varonis.com/free-data-wrangling-tools/>)
  - Tabula (to extract data from pdf into csv or excel format)
  - OpenRefine (formerly Google Refine)
  - DataWrangler (a Stanford project that became the commercial venture Trifacta)
  - R packages – e.g., from tidyverse.org (dplyr, tidyr etc.)
  - Python with Pandas library
  - Mr. Data Converter (from csv/tab separated data to other formats)
  - Many companies build their own ad hoc data wrangling programs/scripts using languages such as Python, Java & R

# Data wrangling software companies

- Trifacta
  - A startup based on San Francisco, CA
  - Platform and tools for data wrangling and viewing
- ClearStory Data
  - A startup in Palo Alto, CA
  - Tools for combining data from variety of sources
- Paxata
  - HQ in Redwood City, CA
  - Part of DataRobot Inc.
  - Runs on Apache Spark
  - Uses semantic algorithms to infer meaning of columns and pattern recognition to identify potential duplicates
- Tableau Prep
  - Data cleaning and preparation tool for Tableau visualization

# Sources

- Chapter 2 (Data Analytics Lifecycle) from the book "Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing, and Presenting Data" by EMC Education Services, John Wiley & Sons, 2015.
- Article on data wrangling in New York Times  
([http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?\\_r=0](http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0))
- “Data wrangling” blog by Lucas Biewald  
<http://www.computerworld.com/article/2902920/the-data-science-ecosystem-part-2-data-wrangling.html>
- “ETL Tools and Data Wrangling” blog by Will Davis @ Trifacta  
(<https://www.trifacta.com/blog/etl-tools/>)