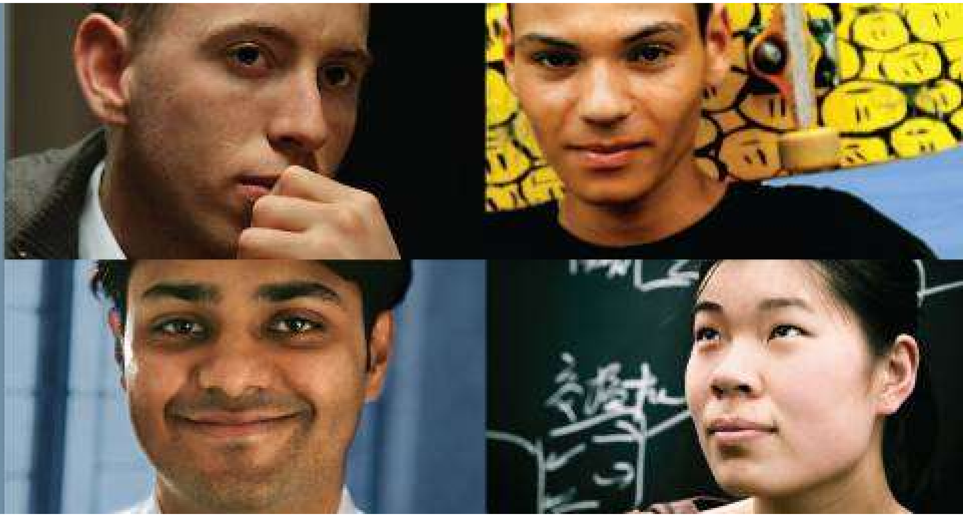
CS433: Internet of Things
NCS463: Internet of Things

Dr. Ahmed Shalaby

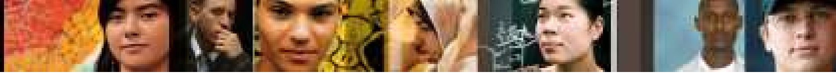
http://bu.edu.eg/staff/ahmedshalaby14

Chapter 1: Data and the Internet of Things



Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™



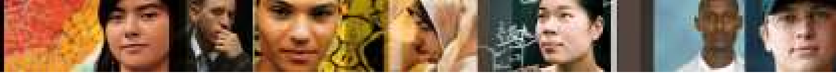
Chapter 1 - Sections & Objectives

- 1.1 Value of Data
  - Demonstrate the value of data.
- 1.2 Data and Big Data
  - Explain the concept of big data.
- 1.3 Managing Big Data
  - Demonstrate knowledge of data management approaches in the IoE.

1.1 The Value of Data

The Data Aspect of a Connected World

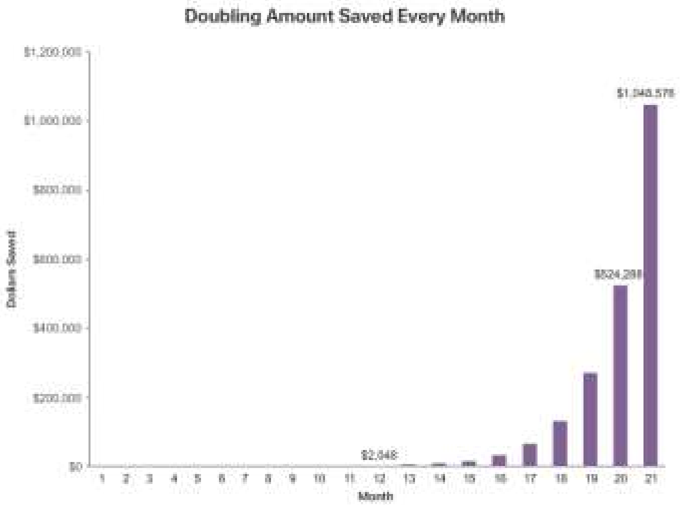
- The Value of Data
  - The amount of data to be stored and analyzed is expanding.
  - The variety of data will reach new areas.
  - The digital transformation will impact three elements of our lives: business, social, and environmental.
- What is Data?
  - Data can be many things.
    - Words in a book, article, or blog
    - Contents of a spreadsheet or database
    - Pictures or video
    - A stream of measurements from a device
  - Useful data is information.
  - Determine the amount of data to be collected.
  - Not all data can be used as-is.
  - Data analysis provides useful information and/or trends.

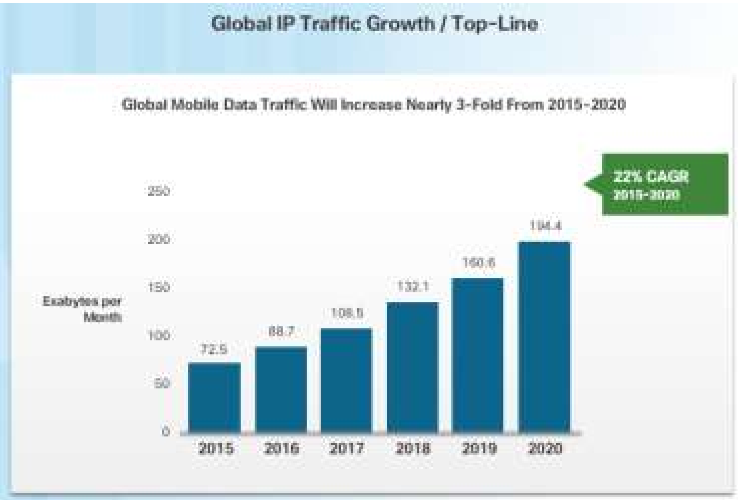


The Value of Data

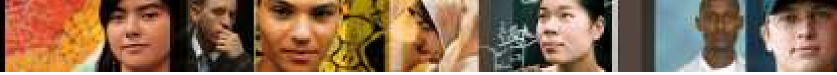
Data is Growing Exponentially

- Estimating Exponential Growth
  - Two types: linear and exponential
  - Exponential growth is more dramatic.
- Growth of Data
  - Today’s data is growing exponentially.
  - Sample data growth forecast for 2015 to 2020 from Cisco’s Visual Networking Index (VNI)
    - Consumer mobile data traffic will reach 26.1 exabytes per moth in 2020.
    - IP traffic will reach 194.4 exabytes per month in 2020.
    - 64% of all global Internet traffic will cross content delivery networks in 2020.





Source: Cisco VNI Global IP Traffic Forecast, 2015-2020



The Value of Data

Data Growth Changes Our Lives

- Data Growth Impact
  - Fueled by the proliferation of IoT devices
  - Including sensors, wireless end devices, and mobile networks
- Business Example: Kaggle
  - Kaggle is a platform that connects businesses and other organizations that have questions about data to the people who know how to find the answers.
  - Kaggle runs online competitions.
- Social Example: DrivenData
  - Brings cutting-edge practices in data science and crowdsourcing to people and organizations that are addressing these challenges
- Environmental Example: Climate Change
  - NASA and Cisco partnership – Planetary Skin
  - Online collaborative global monitoring platform
  - Captures, collects, analyzes and reports data on environmental conditions

1.2 Data and Big Data

Where Does Big Data Come From

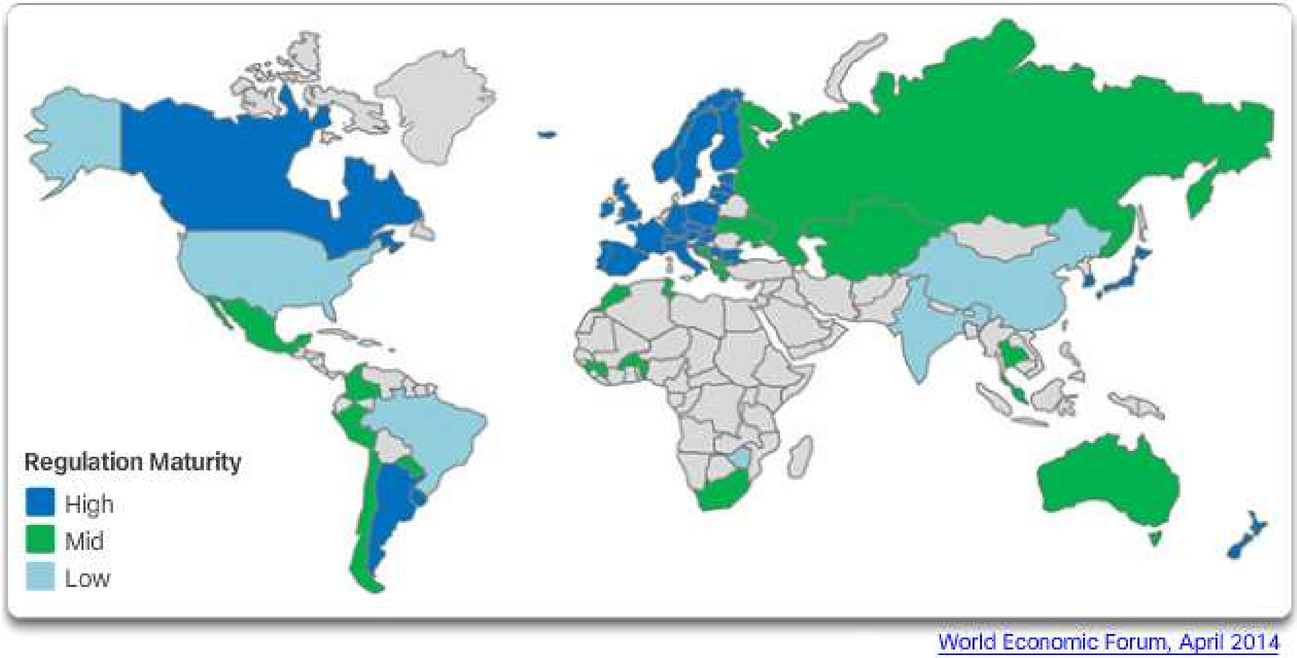
- Defining Big Data
  - Data that is so vast, fast, or complex that it becomes impossible to store, process, and analyze using traditional data storage and analytics applications
- Big Data Characteristics
  - 4 big Vs of Big Data: volume, velocity, variety, and veracity
  - Volume – amount of data
  - Velocity – rate data is generated
  - Variety – type of data
  - Veracity – preventing inaccurate data from spoiling a data set
- How much Data is Big Data
  - IBM's Paul Zikopaulos stated it takes 200 to 600 Terabytes to qualify as Big Data

SCALE OF DATA
VOLUME
FORMS OF DATA
VARIETY
BIG DATA
VELOCITY
ANALYSIS OF DATA FLOW
VERACITY
UNCERTAINTY OF DATA

Data and Big Data

Open Data and Private Data

- Open Data
  - The Open Knowledge Foundation describes Open Data as “any content, information or data that people are free to use, reuse, and redistribute without any legal, technological, or social restriction.”
- Private Data
  - Data related to an expectation of privacy and regulated by a particular country/government

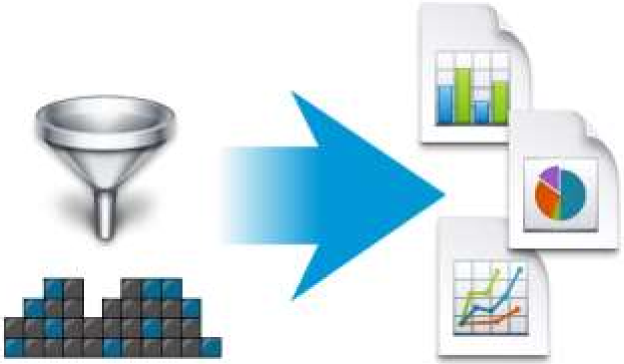


World Economic Forum, April 2014

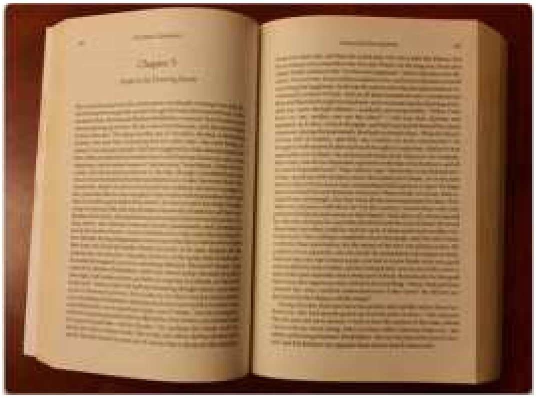
Data and Big Data

Structured and Unstructured Data

- Structured Data
  - Data entered and maintained in fixed fields within a file or record
  - Easily entered, classified, queried, and analyzed
  - Relational databases or spreadsheets



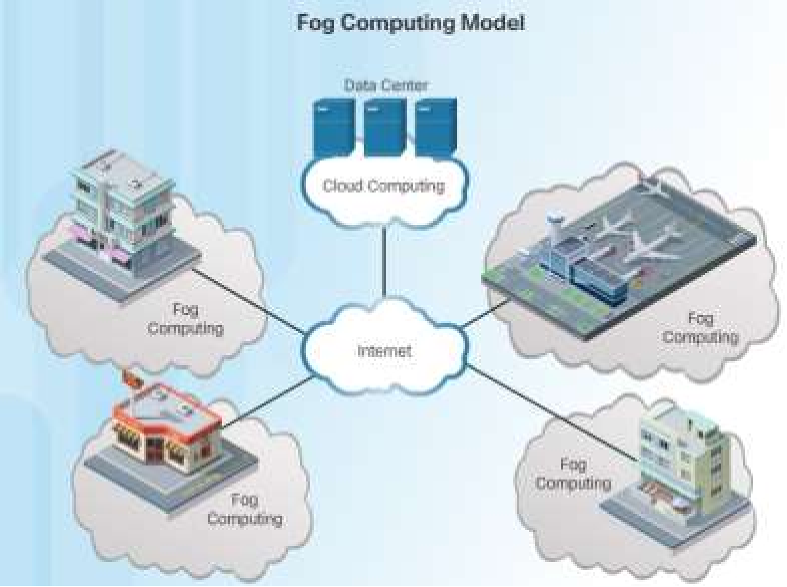
- Unstructured Data
  - Lacks organization
  - Raw data
  - Photo contents, audio, video, web pages, blogs, books, journals, white papers, PowerPoint presentations, articles, email, wikis, word processing documents, and text in general



Data and Big Data

Data at Rest and Data in Motion

- Data at Rest
  - Data stored in a physical location such as a server hard drive or within a data center
  - Follows the traditional data analysis flow of **Store > Analyze > Notify > Act**
- Data in Motion
  - Dynamic data that requires real-time processing before the data becomes irrelevant or obsolete
  - Analysis and action happen sooner rather than later
  - Data analysis flow is **Analyze > Act > Notify > Store**



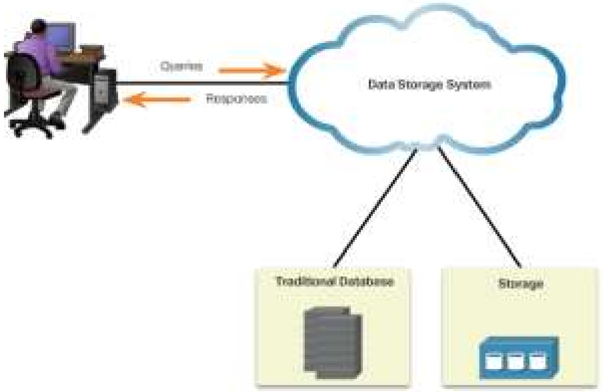
1.3 Managing Big Data

Evolution to Big Data

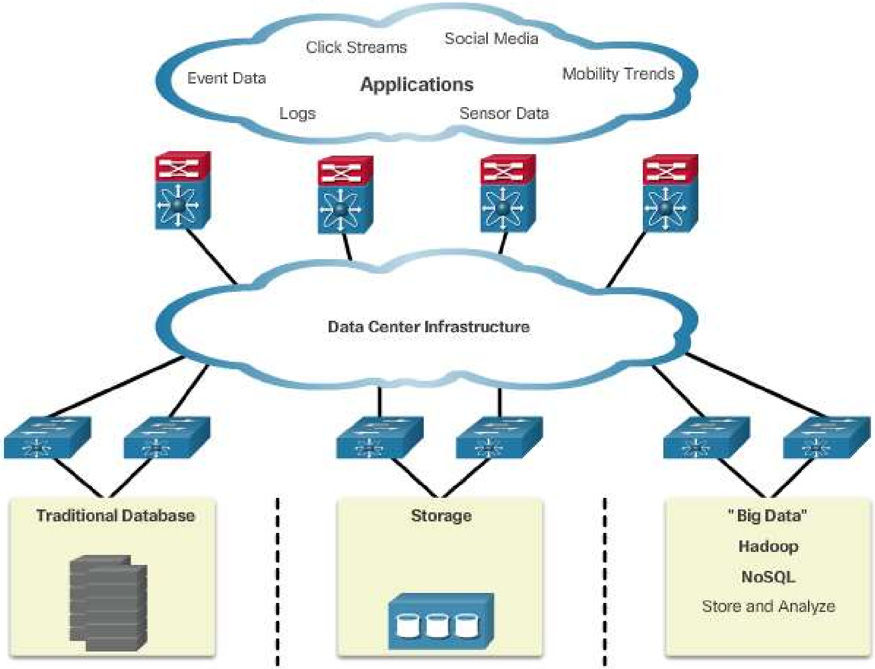
Traditional to Big Data Infrastructure

- Database servers and traditional data processing tools
- Distributed data systems across horizontally coupled, independent resources to achieve the scalability needed for the efficient processing of extensive data sets
- Onsite and cloud computing solutions

Traditional Database Management System

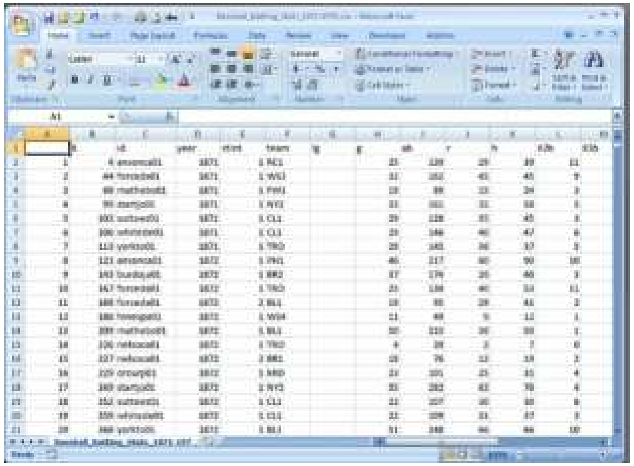


Big Data Infrastructure

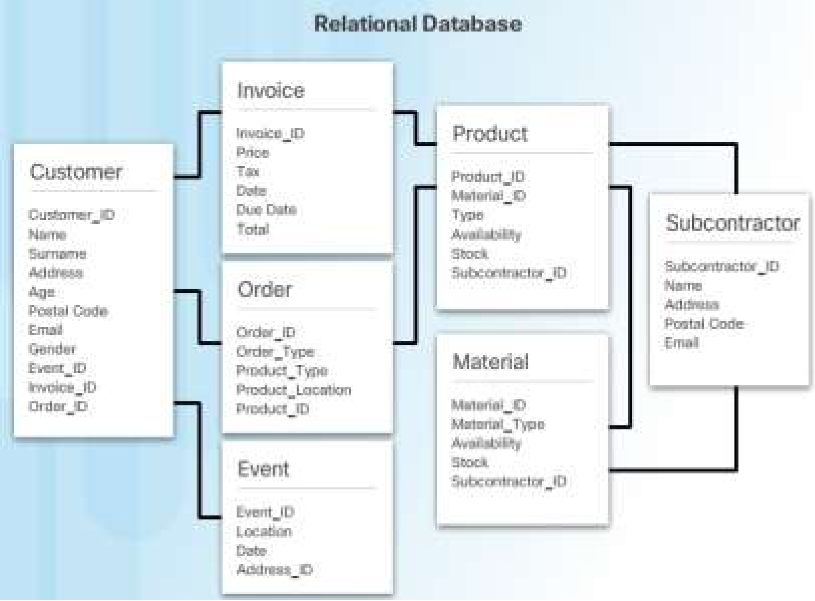


Managing Big Data

Basic Data Management Technologies



- Flat file database – stores records in a single file with no hierarchical structure such as a spreadsheet
- Relational database – capture relationships between different sets of data, creating more useful information



Relational Database

Customer

Customer_ID
Name
Surname
Address
Age
Postal Code
Email
Gender
Event_ID
Invoice_ID
Order_ID

Invoice

Invoice_ID
Price
Tax
Date
Due Date
Total

Order

Order_ID
Order_Type
Product_Type
Product_Location
Product_ID

Event

Event_ID
Location
Date
Address_ID

Product

Product_ID
Material_ID
Type
Availability
Stock
Subcontractor_ID

Material

Material_ID
Material_Type
Availability
Stock
Subcontractor_ID

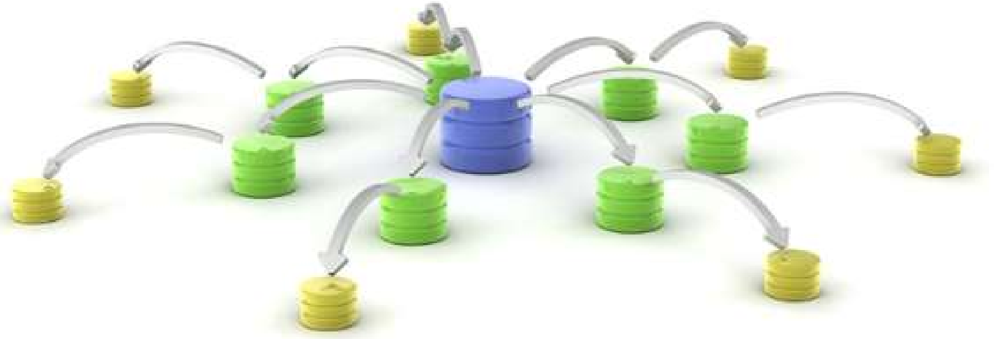
Subcontractor

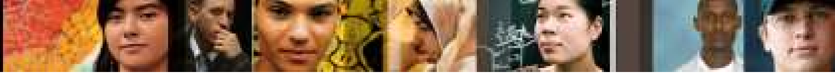
Subcontractor_ID
Name
Address
Postal Code
Email

Managing Big Data

Basic Data Management Technologies

- Relational Database Management System is the dominant database technology with no challenge for over 30 years.
- Big Data analytics becomes increasingly difficult to manage with a relational database management system (RDBMS)
- Hadoop Distributed File System (HDFS) is a distributed, fault tolerant file system created to deal with big data volumes.
- NoSQL database structure created to make database design simpler with faster. Meets the demands of Web applications.
- SQLite – simple and easy to use SQL database engine that is the most widely deployed database in the world.

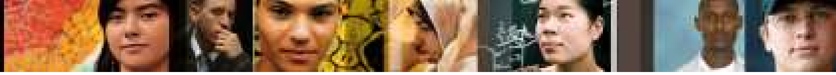




Chapter Summary

Summary

- Data can be words in a book, contents of a spreadsheet, photos, files, or streams of measurements sent by a device.
- Data growth can be linear and exponential. Exponential is a more dramatic increase.
- Four Vs of Big Data are volume, velocity, variety, and veracity.
- Structured data is data entered in fixed fields within a database file or record. Unstructured data does not have a fixed schema that identifies the type of data.
- Data at rest is static data stored in a physical location.
- Data in motion analyzes and extracts value from the data before it is stored.
- A flat file database is like a spreadsheet storing records in a single file with no hierarchical structure.
- A relational database captures the relationships between different data sets and can provide more useful information.

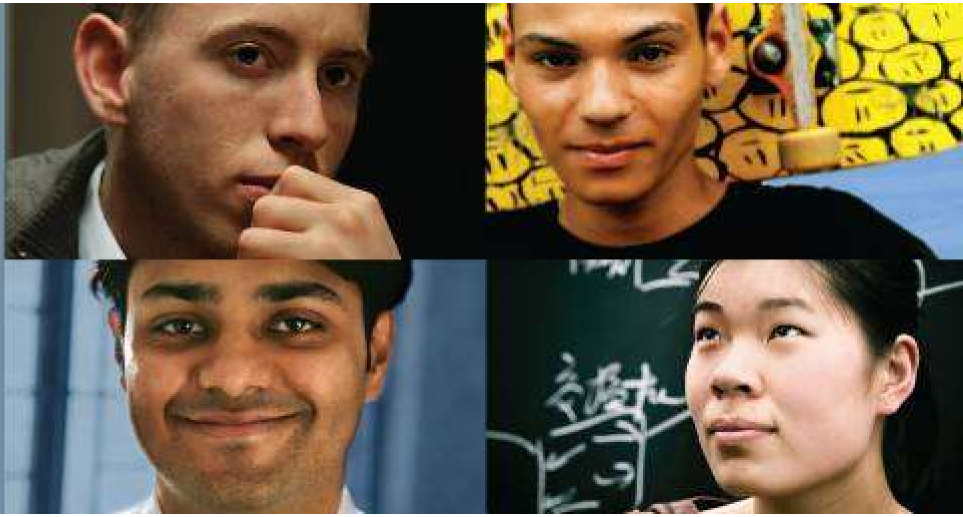


Chapter Summary

Summary

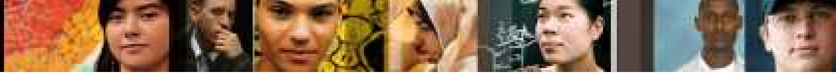
- Hadoop was created to deal with big data volumes.
- A NoSQL database stores and accesses data differently than relational database.
- SQLite is a simple and easy to use SQL database engine that is the most widely deployed database in the world.

Chapter 2: Fundamentals of Data Analysis



Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™



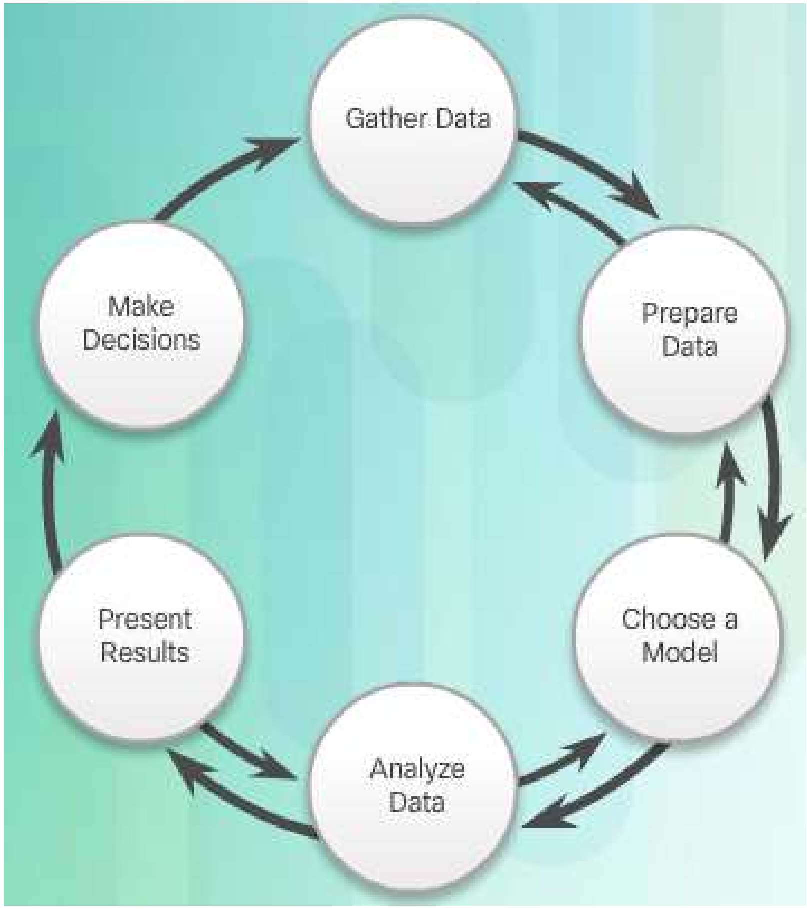
Chapter 2 - Sections & Objectives

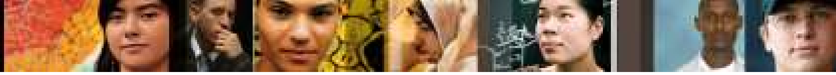
- 2.1 What is Data Analysis
  - Explain how data is used to create knowledge.
- 2.2 Using Big Data
  - Use software tools to visualize a data analysis following the Data Analysis Lifecycle process.
- 2.3 Data Acquisition and Preparation
  - Configure data for analysis.
- 2.4 Big Data Ethics
  - Explain why ethics are important when using Big Data.
- 2.5 Preparation for Chapter 2 Internet Meter Labs
  - Analyze data by using an external application and SQLite.
- 2.6 Summary
  - Summarize the concepts presented in this chapter.

2.1 What is Data Analysis?

Analytics Models

- The six-step Data Analysis Lifecycle
- Data Analytics tools should provide:
  - Ease of use
  - Data manipulation
  - Sharing
  - Interactive exploration





What Are Analytics?

Analytics Models cont…

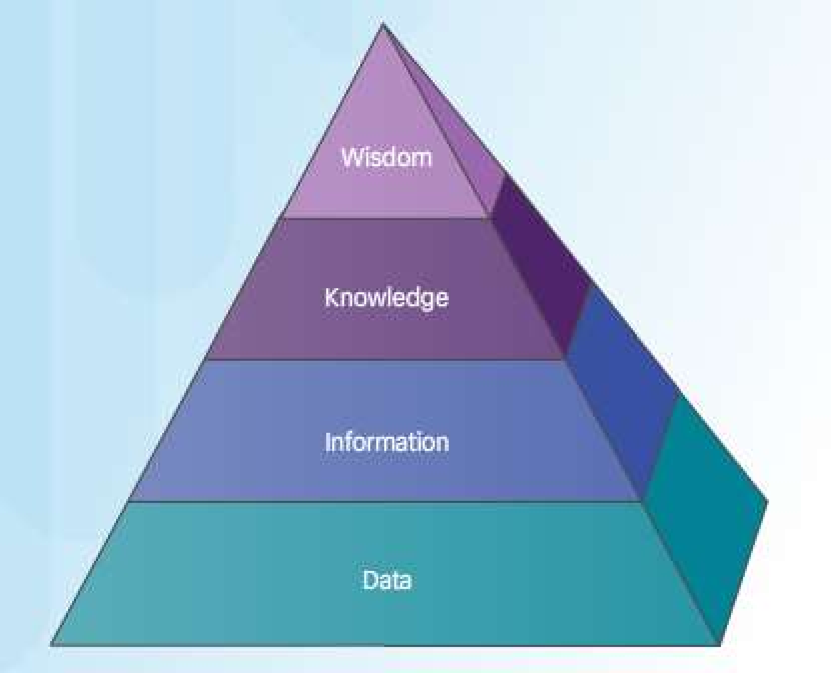
- The Python programming language has become a commonly used tool for handling and manipulating data.
- Python will be used in this course to perform data cleaning, analysis, and manipulation.
- Jupyter Notebooks will be used as both a document for written instructions as well as a Python command interface for running code.

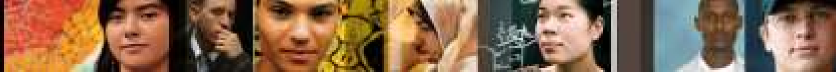
- The libraries that will be used in this course:
  - **NumPy** – This library adds support for arrays and matrices. It also has many built-in mathematical functions for use on data sets.
  - **Pandas** – This library adds support for tables and time series. Pandas is used to manipulate and clean data, among other uses.
  - **Matplotlib** – This library adds support for data visualization. Matplotlib is a plotting library capable of creating simple line plots to complicated 3D and contour plots.

2.2 Using Big Data

Types of Data Analysis

- Scalable technologies are enabling data center administrators to manage the top three aspects of Big Data:
  - **Volume**
  - **Velocity**
  - **Variety**
- The **Data**, **Information**, **Knowledge**, and **Wisdom** (DIKW) model shows the transitions that data undergoes until it gains enough value to inform wise decisions. This is called **Business Intelligence**

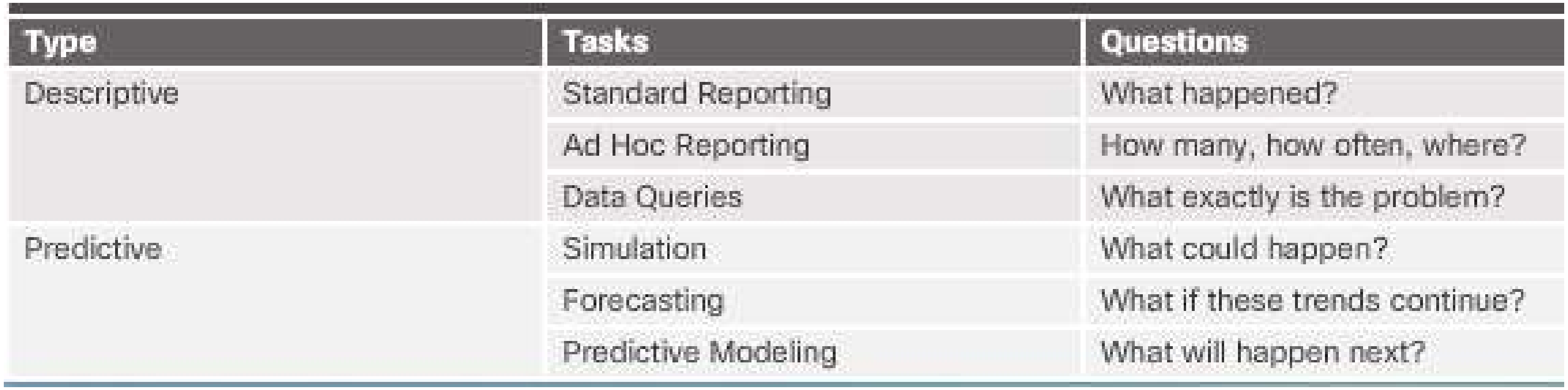




Using Big Data

Why Analyze Big Data?

- Multiple types of analytics provide organizations and people with information that can drive innovation, improve efficiency and mitigate risk.
  - **Descriptive analytics** - Relies solely on historical data to provide regular reports on events that have already happened.
  - **Predictive analytics** - Can infer missing data and establish a future trend line based on past data. It uses simulation models and forecasting to suggest what could happen.
  - **Prescriptive analytics** - Recommends actions or decisions based on a complex set of targets, constraints, and choices.

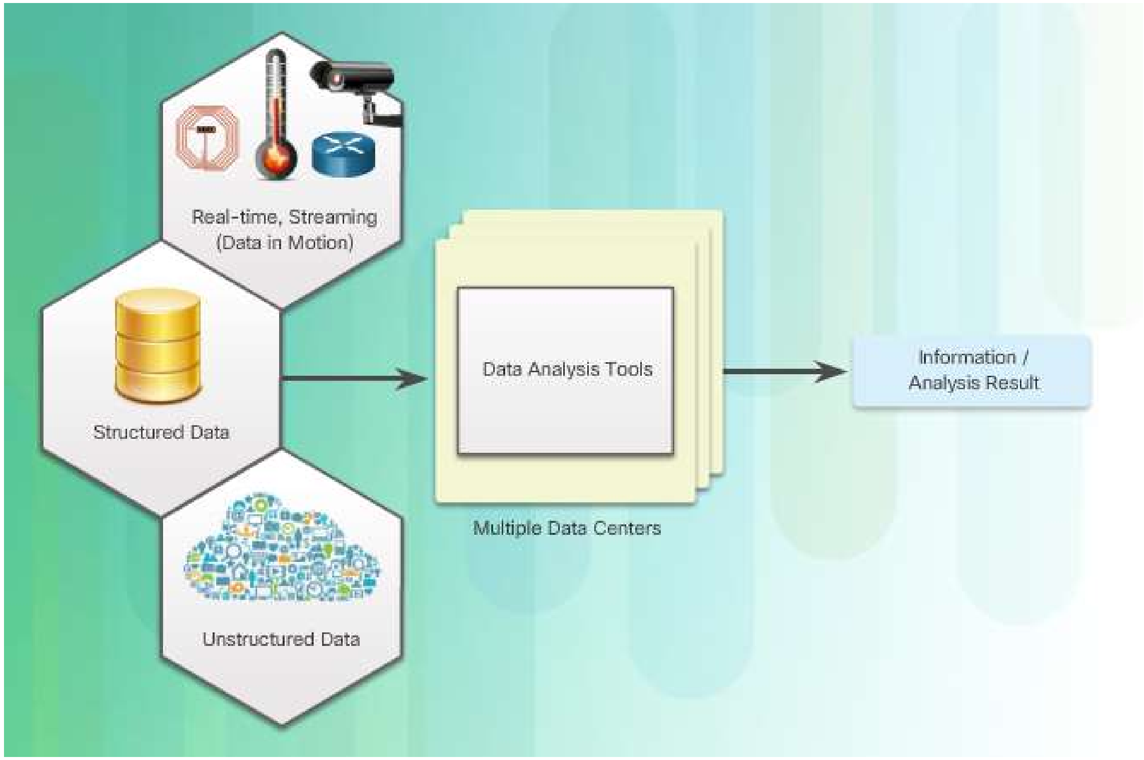


Type	Tasks	Questions
Descriptive	Standard Reporting	What happened?
	Ad Hoc Reporting	How many, how often, where?
	Data Queries	What exactly is the problem?
Predictive	Simulation	What could happen?
	Forecasting	What if these trends continue?
	Predictive Modeling	What will happen next?

Using Big Data

Timely Analysis of Big Data

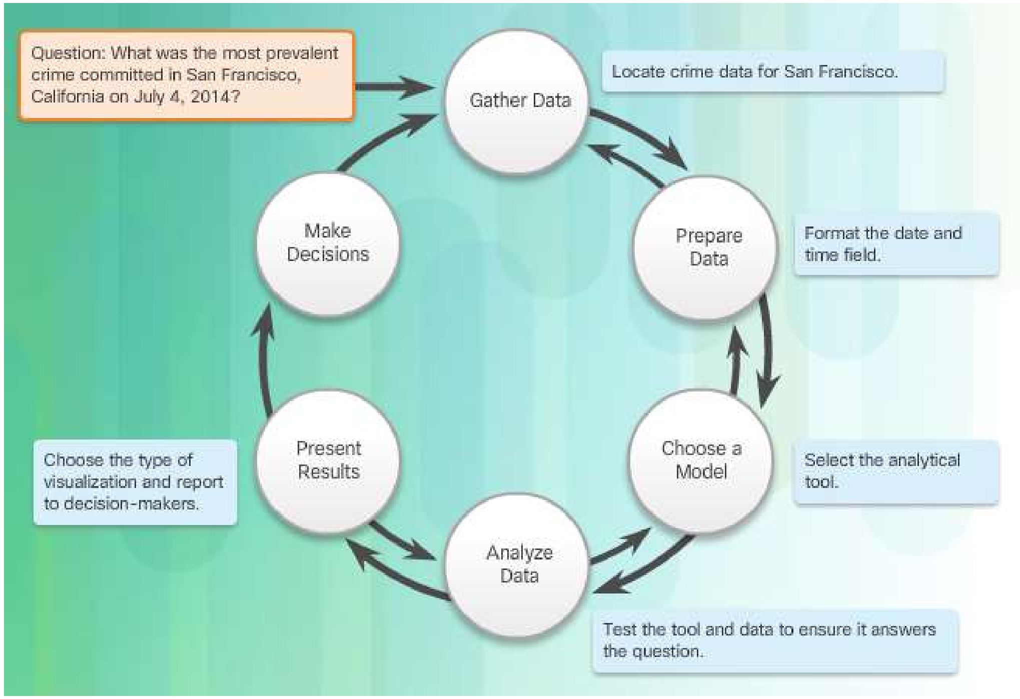
- With Big Data, much of the value of data is derived from creating opportunities to take action immediately.
- Data-driven decisions can have the following benefits:
  - Increased time to research and develop products and services
  - Increased efficiency and faster manufacturing
  - Faster time to market
  - More effective marketing and advertising



Using Big Data

Data Analysis Lifecycle

- **Gathering the data** - The process of locating data and then determining if there is enough data to complete the analysis.
- **Preparing the data** - This step can involve many tasks to transform the data into a format appropriate for the tool that will be used.
- **Choosing a model** - This step includes choosing an analysis technique that will best answer the question with the data available.
- **Analyzing the data** - The process of testing the model against the data and determining if the model and the analyzed data are reliable. Were you able to answer the question with the selected tool?
- **Presenting the results** - The process of communicating the results to decision-makers.
- **Making decisions** - Organizational leaders incorporate the new knowledge as part of the overall strategy. The process begins anew with gathering data.



Question: What was the most prevalent crime committed in San Francisco, California on July 4, 2014?

Gather Data

Locate crime data for San Francisco.

Prepare Data

Format the date and time field.

Choose a Model

Select the analytical tool.

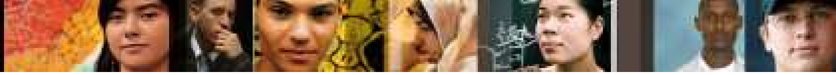
Analyze Data

Test the tool and data to ensure it answers the question.

Present Results

Choose the type of visualization and report to decision-makers.

Make Decisions



2.3 Data Acquisition and Preparation

Sources of Data

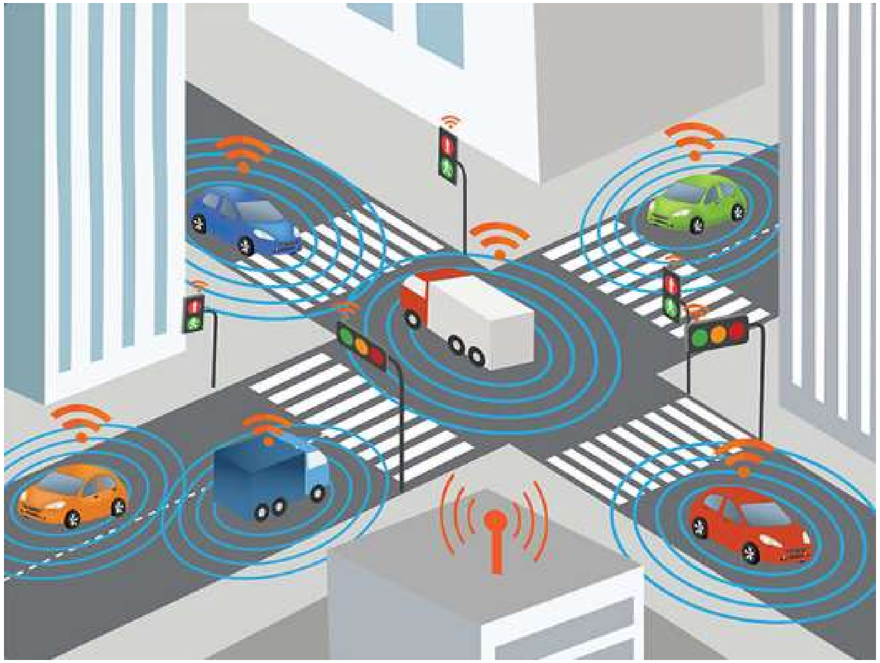
There are many different sources of data.

- A vast amount of historical data can be found in files such as:
  - MS Word documents
  - Emails
  - Spreadsheets
  - MS PowerPoints
  - PDFs
  - HTML
  - and plaintext files
- Public and Private Archives
- CSV, JSON, and XML files use plaintext, a common format, and are compatible with a wide range of applications
- The Web can be mined for data using a web scraping application

Data Acquisition and Preparation

Sources of Data cont…

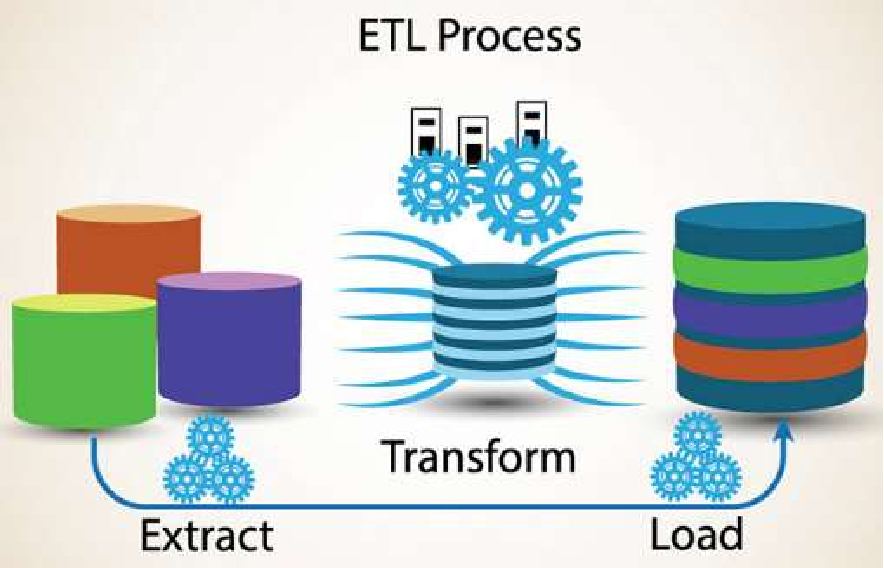
- The IoT uses sensors create data
  - Sensors in smartphones, cars, airplanes, street lamps, and home appliances capture raw data
- The list of things with sensors grows every year
  - The IoT contributes to the growth of Big Data



Data Acquisition and Preparation

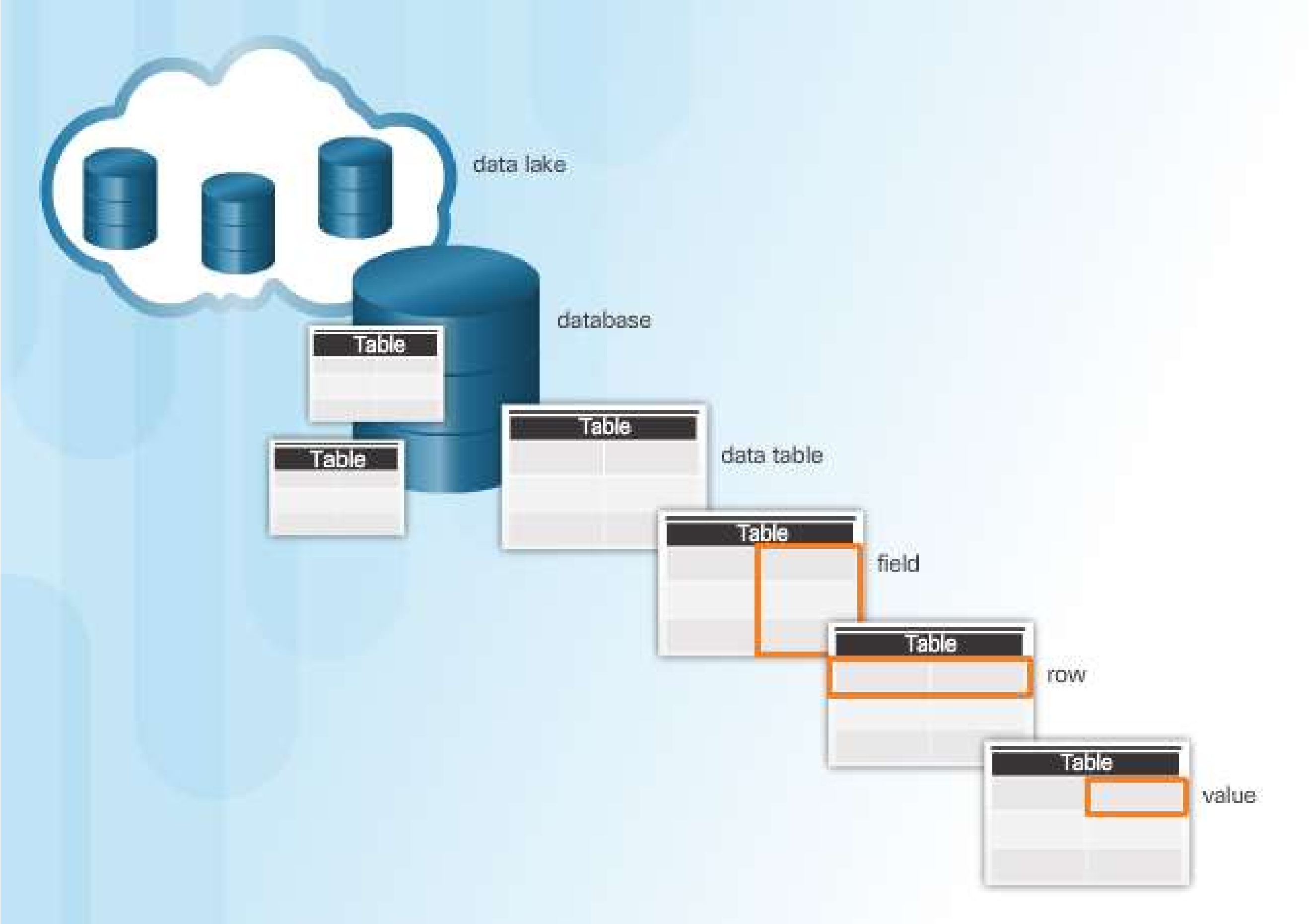
Data Preparation

- Collected data may not be compatible or formatted correctly
  - Data must be prepared before it can be added to a data set
- Extract, Transform and Load (ETL)
  - process for collecting data from a variety of sources, transforming the data, and then loading the data into a database



Data Structures

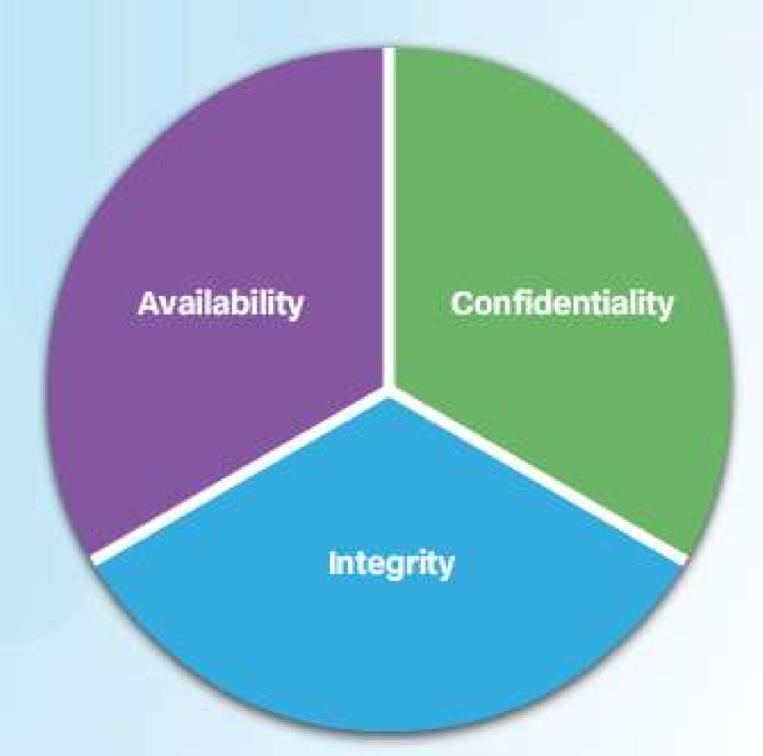
- Relational Database Tables
  - Fields (Column)
  - Rows (Rows)
  - Values (Cells)
- Python
  - Strings
  - Lists
  - Tuples
  - Sets
  - Dictionaries



2.4 Big Data Ethics

What are the Ethical Concerns?

- Data protection regulations varies from country to country
- Confidentiality, integrity and availability, known as the CIA triad is a guideline for data security in an organization
- Four general cloud security controls:
  - Deterrent
  - Preventive
  - Detective
  - Corrective

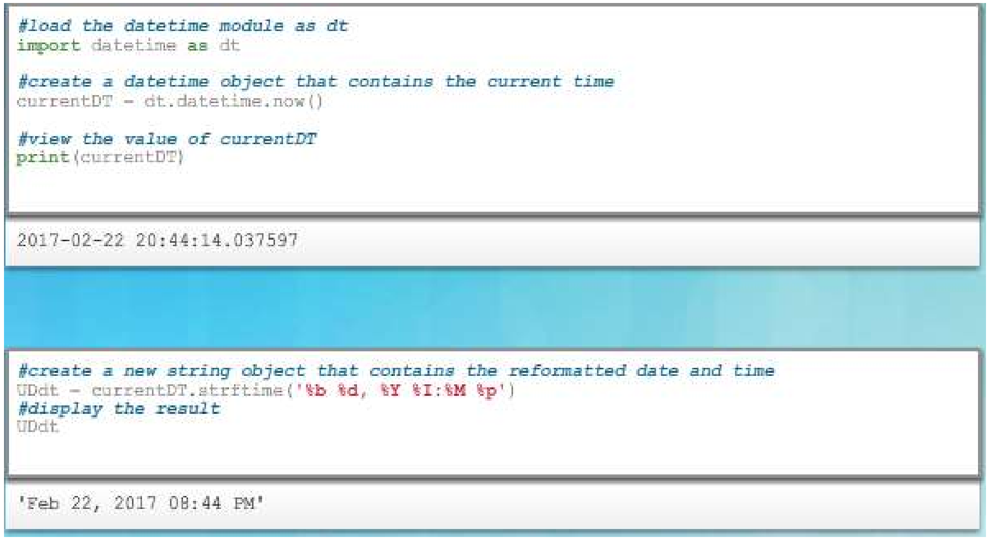


2.5 Preparation for Chapter 2 Internet Meter Labs

Part 1

- The **datetime** module is included in most Python distributions as a standard library; however, it must be imported to be used in your code.

- The csv module allows reading and writing to .csv files.

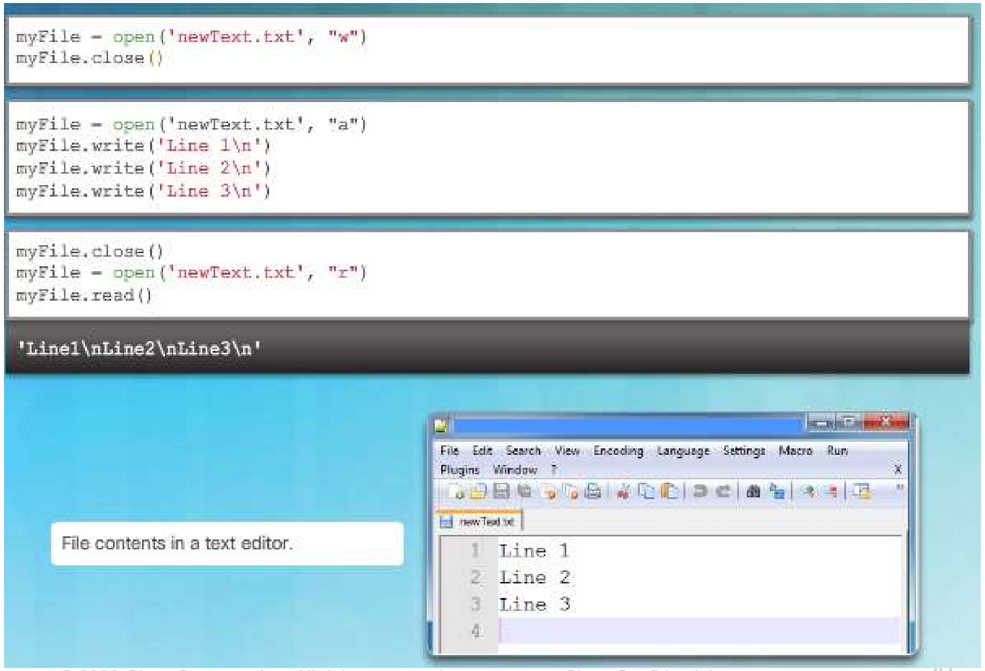


```
#load the datetime module as dt
import datetime as dt

#create a datetime object that contains the current time
currentDT = dt.datetime.now()

#view the value of currentDT
print(currentDT)
```

```
2017-02-22 20:44:14.037597
```

```
#create a new string object that contains the reformatted date and time
UDdt = currentDT.strftime('%b %d, %Y %I:%M %p')
#display the result
UDdt
```

```
'Feb 22, 2017 08:44 PM'
```



```
myFile = open('newText.txt', "w")
myFile.close()
```

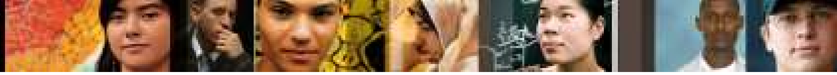
```
myFile = open('newText.txt', "a")
myFile.write('Line 1\n')
myFile.write('Line 2\n')
myFile.write('Line 3\n')
```

```
myFile.close()
myFile = open('newText.txt', "r")
myFile.read()
```

```
'Line1\nLine2\nLine3\n'
```

File contents in a text editor.

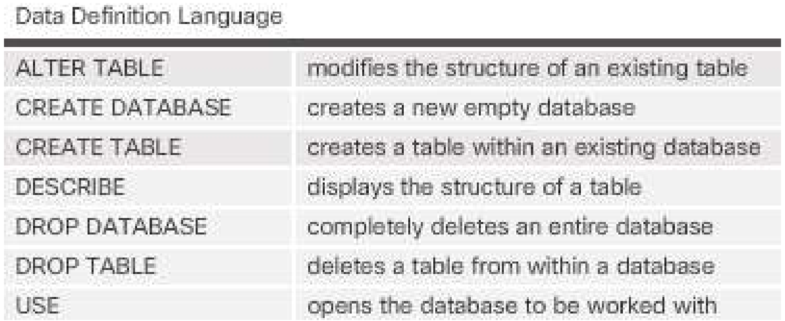
```
Line 1
Line 2
Line 3
```



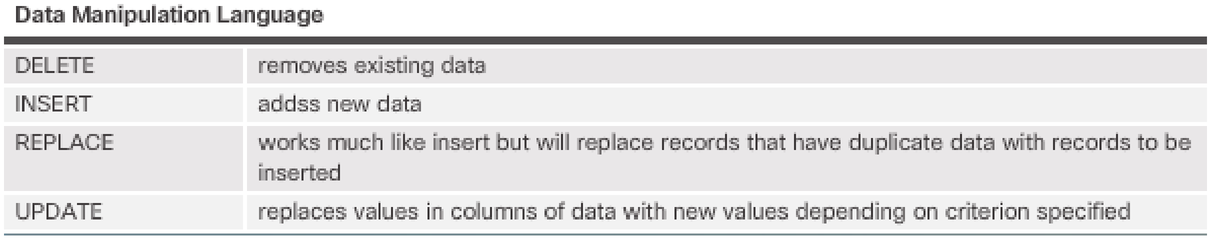
Preparation for Chapter 2 Internet Meter Labs

Part 2

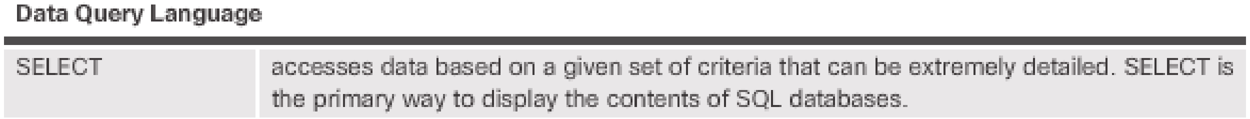
- SQLite is an SQL implementation using a client server method of operation
  - Uses connections established between Python and an SQL database by creating an SQL connection object
- SQL can be said to be a language composed of three special purpose languages
  - Data Definition Language
  - Data Manipulation Language
  - Data Query Language



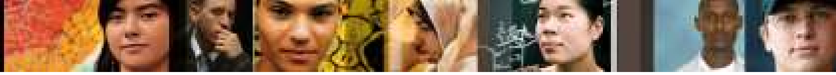
Data Definition Language	
ALTER TABLE	modifies the structure of an existing table
CREATE DATABASE	creates a new empty database
CREATE TABLE	creates a table within an existing database
DESCRIBE	displays the structure of a table
DROP DATABASE	completely deletes an entire database
DROP TABLE	deletes a table from within a database
USE	opens the database to be worked with



Data Manipulation Language	
DELETE	removes existing data
INSERT	addss new data
REPLACE	works much like insert but will replace records that have duplicate data with records to be inserted
UPDATE	replaces values in columns of data with new values depending on criterion specified



Data Query Language	
SELECT	accesses data based on a given set of criteria that can be extremely detailed. SELECT is the primary way to display the contents of SQL databases.

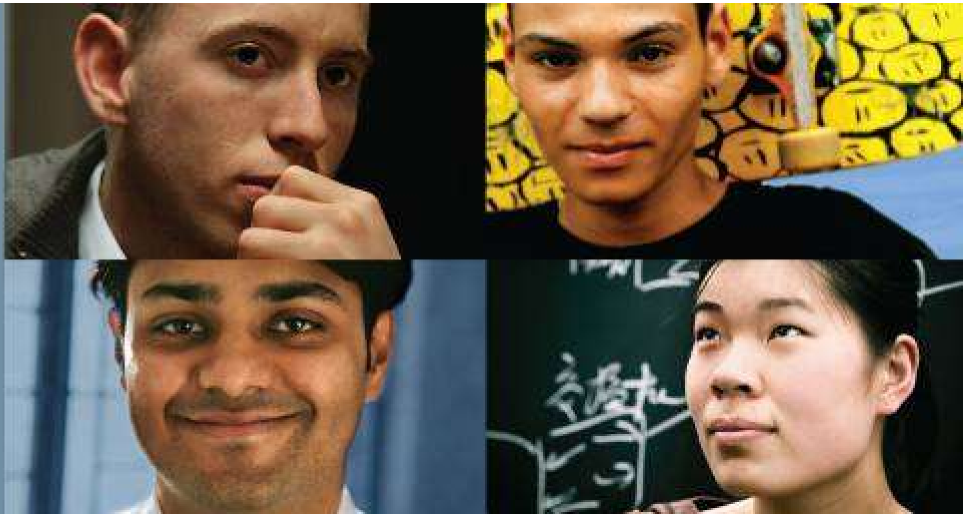


Chapter Summary

Summary

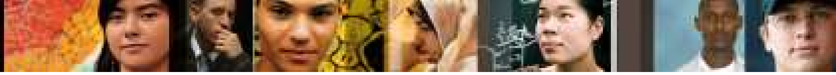
- Data can no longer be stored on a few machines or processed with just one tool
- Decision makers will increasingly rely on data analytics to extract the required information at the right time, in the right place, to make the right decision
- Descriptive analytics relies solely on historical data
- Predictive analytics attempts to predict what may happen
- Prescriptive analytics predicts outcomes and suggests courses of actions that will hold the greatest benefit for an organization
- Files, the Internet, sensors, and databases are all good sources of data.
- Extract, Transform and Load (ETL) is a process for collecting data from a variety of sources, transforming the data, and then loading the data into a database
- The CIA triad is a guideline for data security for an organization

Chapter 3: Data Analysis



Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™



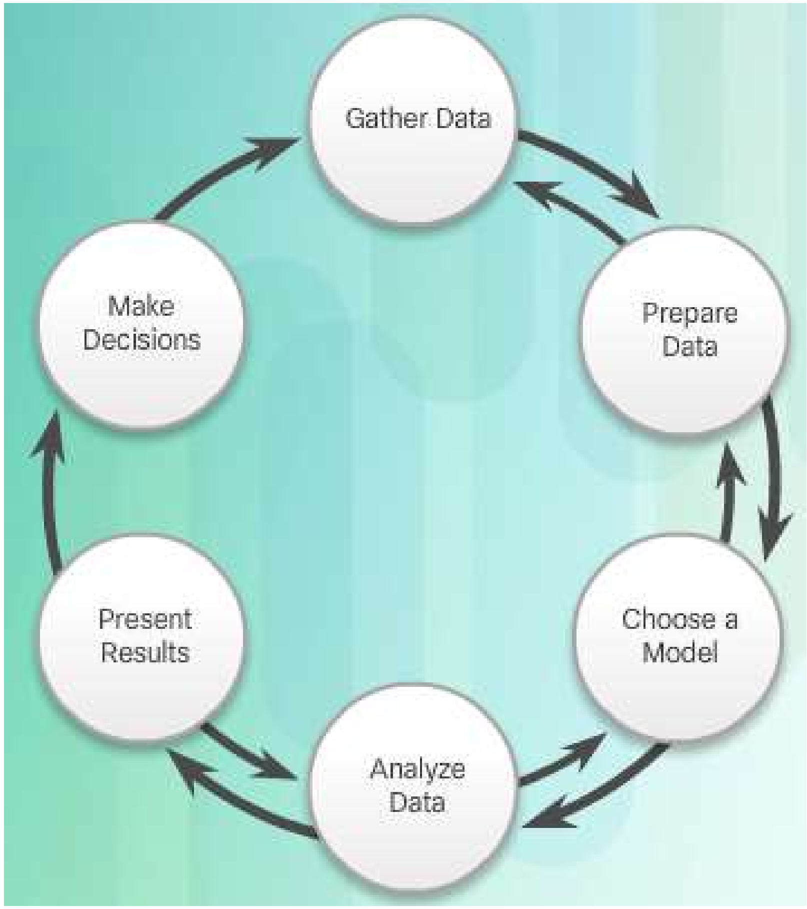
Chapter 3 - Sections & Objectives

- 3.1 Analyzing Data
  - Analyze data using basic statistics.
- 3.2 Preparation for Chapter 3 Internet Meter Lab
  - Configure data for analysis.
- 3.3 Summary
  - Summarize the concepts presented in this chapter.

3.1 Analyzing Data

Preliminaries

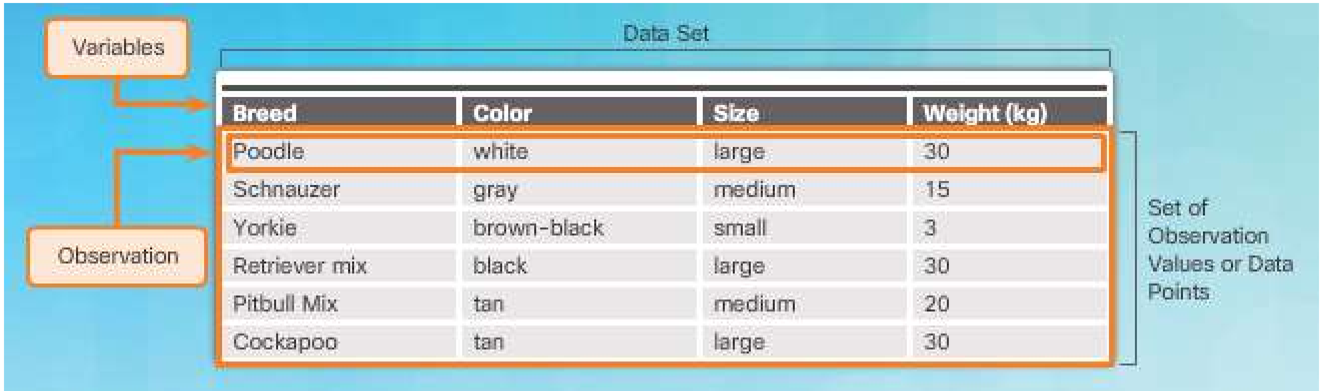
- Data is changed from its raw format into information after it has been gathered, prepared, analyzed, and presented in a usable format.
- Exploratory data analysis is a set of procedures designed to produce descriptive and graphical summaries of data with the notion that the results *may* reveal interesting patterns



Analyzing Data

Preliminaries cont…

- **IoT Concerns**
  - IoT data may come in large volume and in different forms.
  - IoT data may require more advanced analytic tools for structured and unstructured data
  - IoT data is frequently streaming in real time or nearly real time.
- **Observations, Variables, and Values**
  - A variable is anything that varies from one instance to another and is something that can be measured, manipulated or controlled.
  - The recordings of the values, patterns and occurrences for a set of variables is an observation.
  - The set of values for a specific observation is called a data point.

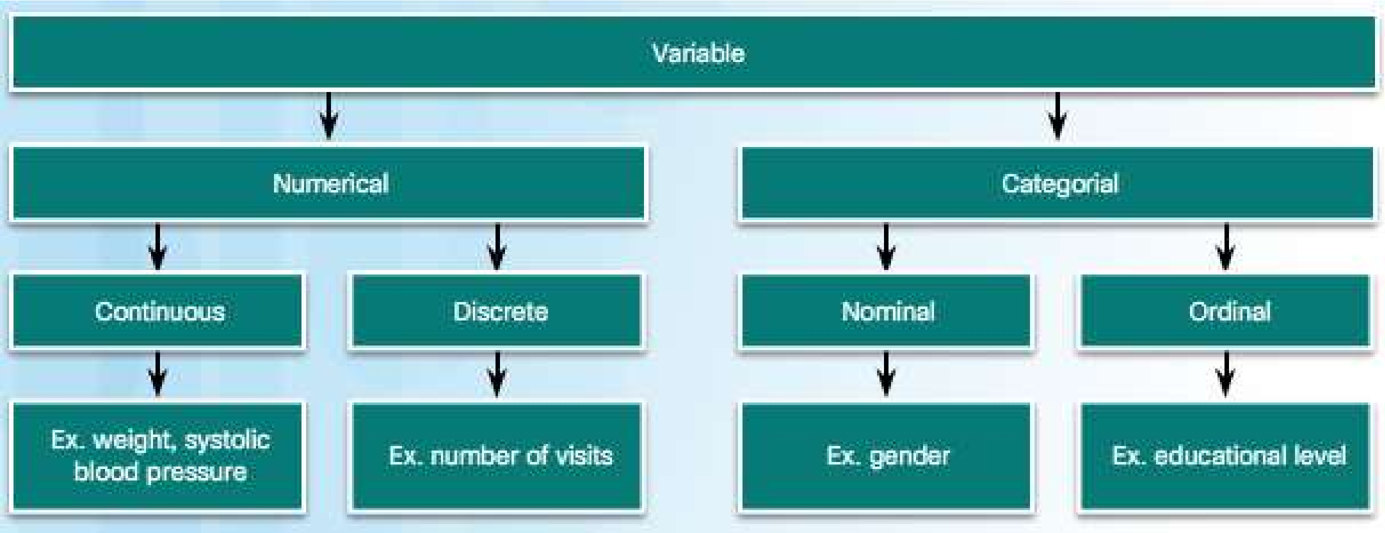


Breed	Color	Size	Weight (kg)
Poodle	white	large	30
Schnauzer	gray	medium	15
Yorkie	brown-black	small	3
Retriever mix	black	large	30
Pitbull Mix	tan	medium	20
Cockapoo	tan	large	30

Analyzing Data

Preliminaries cont…

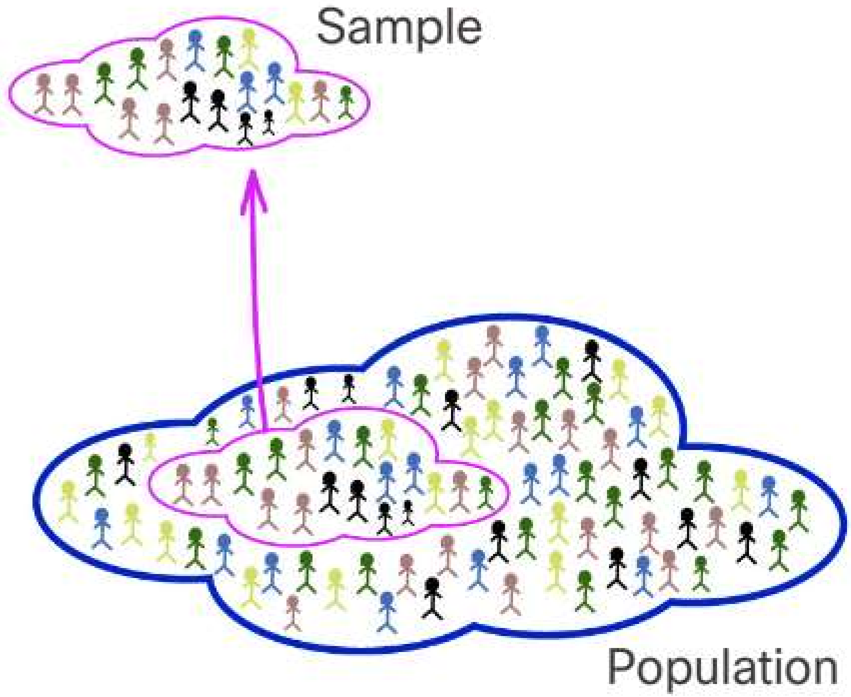
- Categorical variables include:
  - Nominal – Two or more categories or names that identify the object
  - Ordinal – Two or more categories in which order matter in the value
- Numerical variables include:
  - Continuous – quantitative along a continuum or range of values
  - Ratio - Interval variables where zero (0) means none
  - Discrete - Quantitative with a specific value from a finite set of values

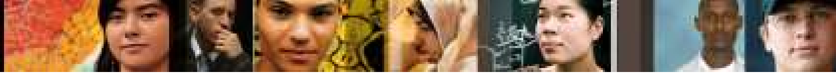


Analyzing Data

Statistical Analysis

- Statistics is the collection and analysis of data using mathematical techniques.
- Sample and Population
    - A population is a group of similar entities such as people, objects, or events that share some common set of characteristics.
    - A sample is a representative group from the population.

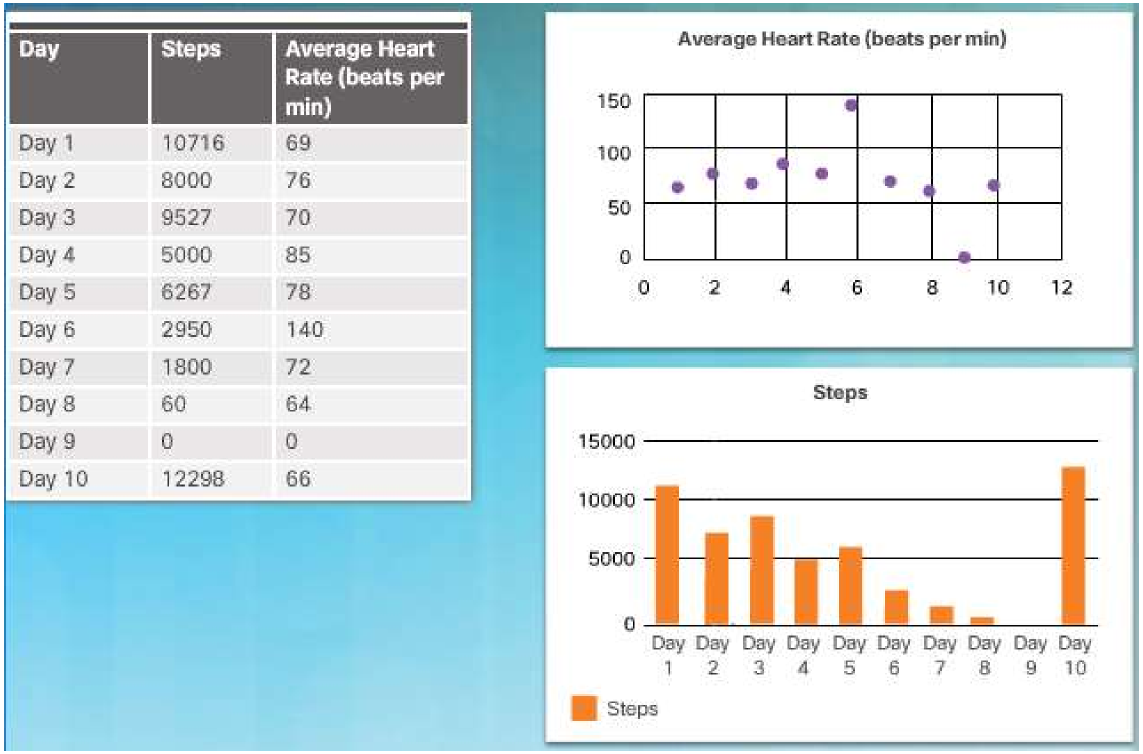




Analyzing Data

Statistical Analysis cont…

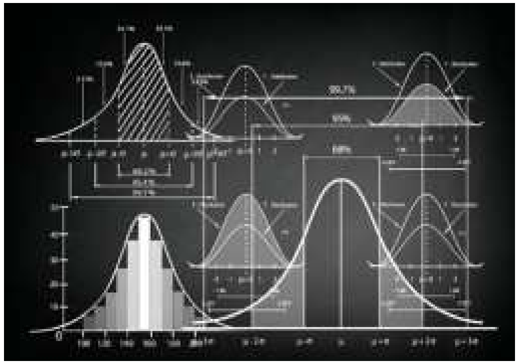
- Descriptive statistics
  - describe or summarize the values and observations of a data set.
- Inferential statistics
  - process of collecting, analyzing and interpreting data gathered from a sample to make generalizations or predictions about a population



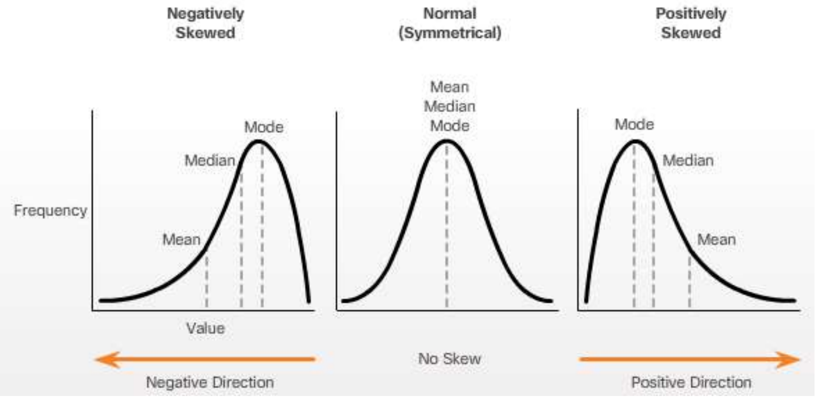
Day	Steps	Average Heart Rate (beats per min)
Day 1	10716	69
Day 2	8000	76
Day 3	9527	70
Day 4	5000	85
Day 5	6267	78
Day 6	2950	140
Day 7	1800	72
Day 8	60	64
Day 9	0	0
Day 10	12298	66

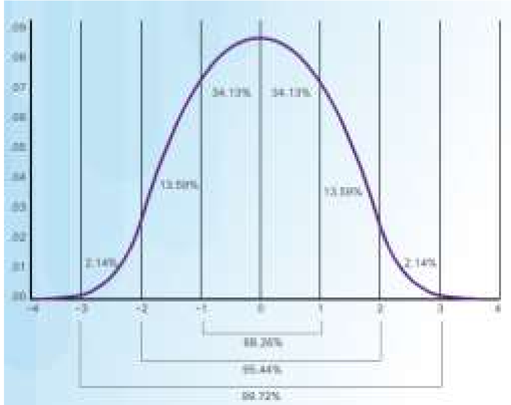
Analyzing Data

Characteristics of Samples



- Distribution
  - a variable and its frequency or probability
- Centrality
  - The mean, median, and mode
- Dispersion
  - the variability in the distribution

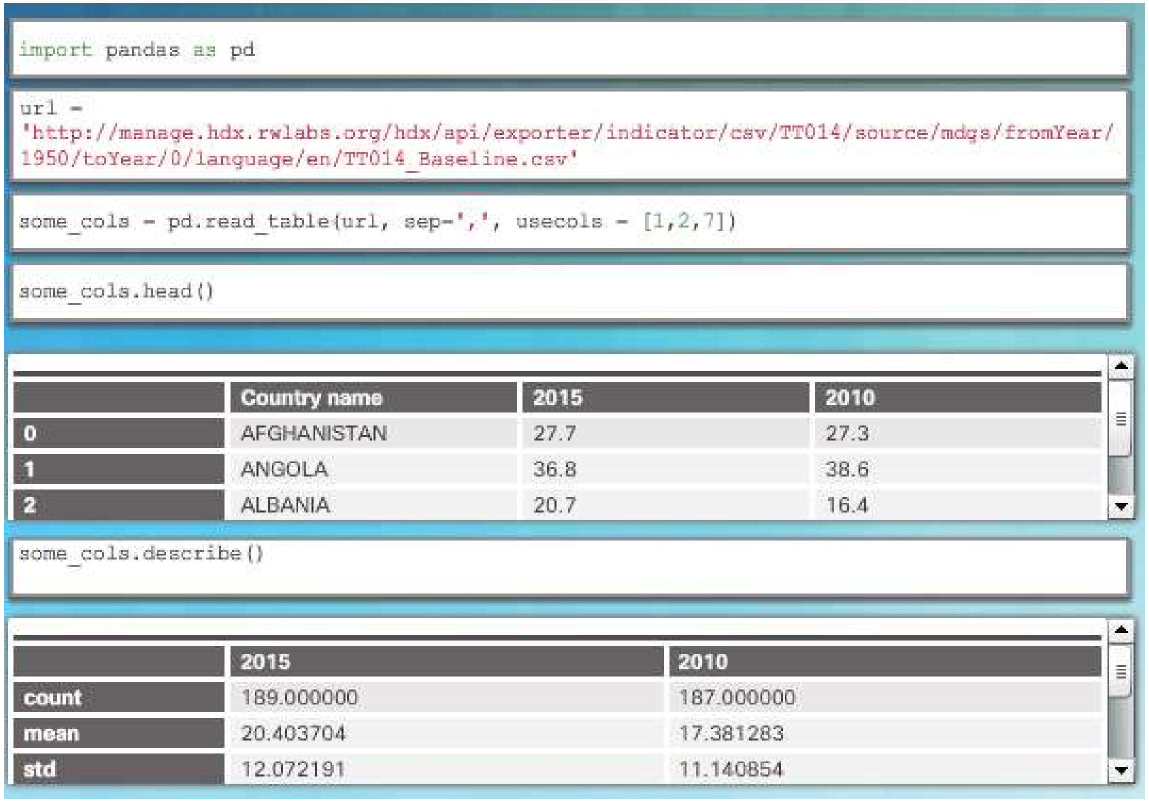




Analyzing Data

Analysis Using Descriptive Statistics

- Pandas
  - open source library for Python that adds high-performance data structures and tools for analysis of large data sets
  - Import data from files
  - Import data from web
  - Descriptive statistics in pandas



```
import pandas as pd
```

```
url =
'http://manage.hdx.rwlabs.org/hdx/api/exporter/indicator/csv/TT014/source/mdgs/fromYear/1950/toYear/0/language/en/TT014_Baseline.csv'
```

```
some_cols = pd.read_table(url, sep=',', usecols = [1,2,7])
```

```
some_cols.head()
```

	Country name	2015	2010
0	AFGHANISTAN	27.7	27.3
1	ANGOLA	36.8	38.6
2	ALBANIA	20.7	16.4

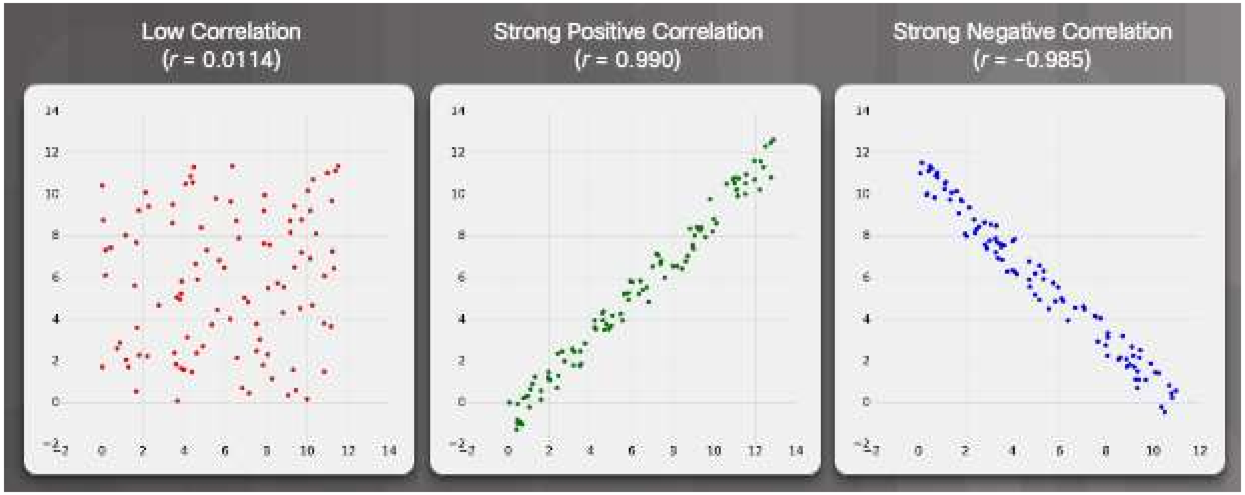
```
some_cols.describe()
```

	2015	2010
count	189.000000	187.000000
mean	20.403704	17.381283
std	12.072191	11.140854

Analyzing Data

Analysis Using Correlation

- “Correlation does not imply causation”
  - Causation is a relationship in which one thing changes, or is created, directly because of something else.
  - Correlation is a relationship between phenomena in which two or more things change at a similar rate.
  - Correlations can be positive or negative.



Low Correlation (r = 0.0114)

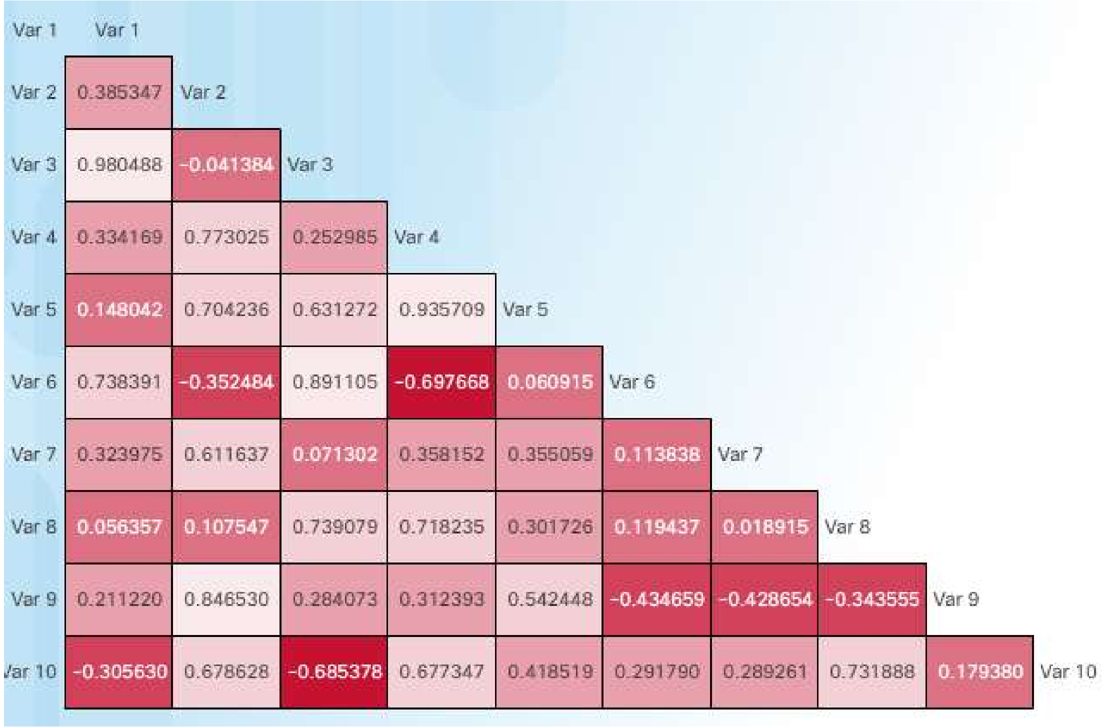
Strong Positive Correlation (r = 0.990)

Strong Negative Correlation (r = −0.985)

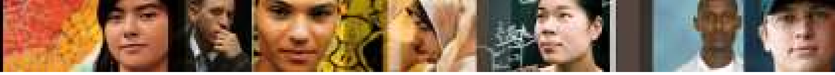
Analyzing Data

Analysis Using Correlation cont…

- Correlations can be calculated for multiple variables simultaneously
- Heat map
  - values for correlation coefficients relate to one another



Var 1	Var 1									
Var 2	0.385347	Var 2								
Var 3	0.980488	-0.041384	Var 3							
Var 4	0.334169	0.773025	0.252985	Var 4						
Var 5	0.148042	0.704236	0.631272	0.935709	Var 5					
Var 6	0.738391	-0.352484	0.891105	-0.697668	0.060915	Var 6				
Var 7	0.323975	0.611637	0.071302	0.358152	0.355059	0.113838	Var 7			
Var 8	0.056357	0.107547	0.739079	0.718235	0.301726	0.119437	0.018915	Var 8		
Var 9	0.211220	0.846530	0.284073	0.312393	0.542448	-0.434659	-0.428654	-0.343555	Var 9	
Var 10	-0.305630	0.678628	-0.685378	0.677347	0.418519	0.291790	0.289261	0.731888	0.179380	Var 10

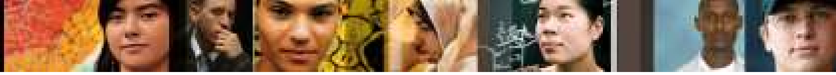


3.2 Preparation for Chapter 3 Internet Meter Lab

Basic Analysis with pandas

- More often than not, the data sets that you work with will have incompatibilities
- Cleaning data can involve removing missing or unwanted values, or altering the format of the values to make them consistent

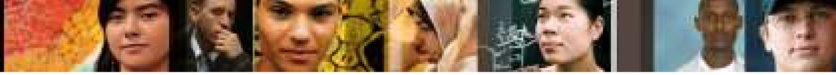
- **NaNs** (Not a Number) values are used to represent data that is undefined or cannot be represented. pandas refers to missing data as NaN values
  - NaTs are used for timestamps
- **Pandas** has many built-in functions for:
  - converting the datatypes
  - manipulating data frames
  - running statistical analysis on data sets.



Chapter Summary

Summary

- Exploratory data analysis produces descriptive and graphical summaries of data with the notion that the results *may* reveal interesting patterns.
- IoT data may be structured or unstructured and data must be organized in real time.
- Observations, variables, and values are critical to an analysis.
- Variables include Numerical (Continuous and Discrete) and Categorical (Nominal and Ordinal)
- Statistics is the collection and analysis of data using mathematical techniques.
    - The interpretation of data and the presentation of findings.
    - The discovery of patterns or relationships between variables.
- Statistics uses samples and populations.
- Statistical analysis includes descriptive and inferential statistics.

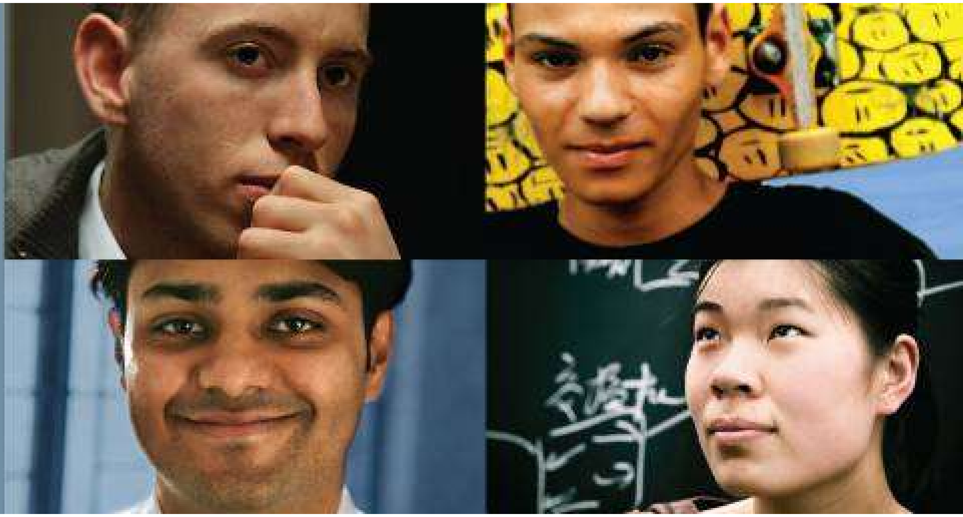


Chapter Summary

Summary cont…

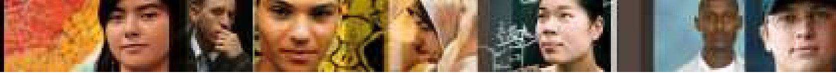
- Distribution is a simple association between a value and the number or percentage of times it appears in a data sample.
- Centrality includes the mean, median, and mode.
  - These values that are closer to the center of the distribution occur with greater frequency.
- Dispersion is the variability in the distribution.
- Pandas is an open source library for Python with tools for analysis of large data sets
  - Importing data from files
  - Importing data from Web
  - Viewing descriptive statistics
- “Correlation does not imply causation”
- Data commonly needs cleaning, converting, and manipulating before data analysis.

Chapter 4: Advanced Data Analytics and Machine Learning



Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™



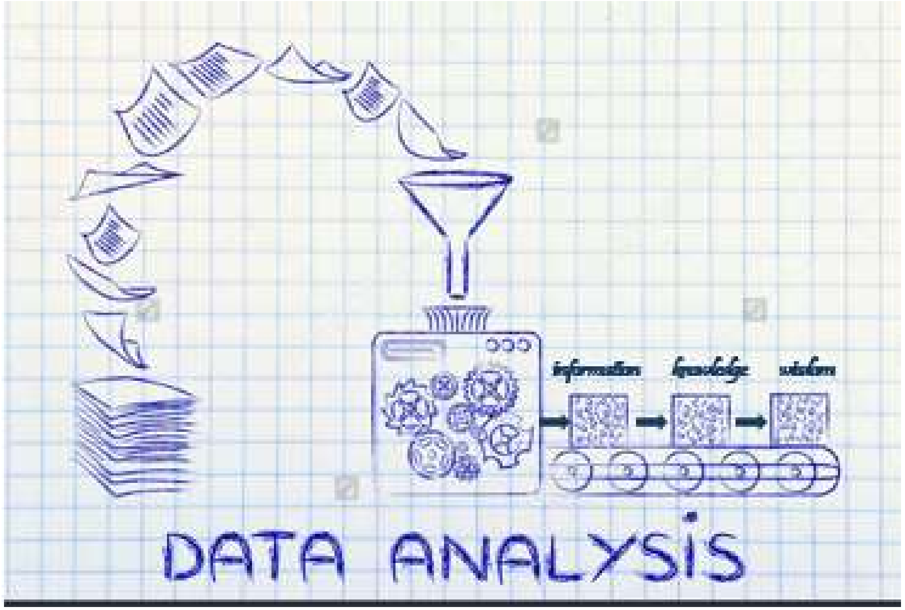
Chapter 4 - Sections & Objectives

- 4.1 Predictive Analytics
  - Identify the likelihood of future outcomes through the use of data, statistical algorithms and machine learning techniques, based on historical data.
- 4.2 Model Evaluation
  - Examine the various evaluation metrics used in predictive analytics.
- 4.3 Preparation for Chapter 4 Labs

4.1 Machine Learning

Looking Ahead

- Characteristics that distinguish Big Data from data:
  - Volume
  - Velocity
  - Variety
  - Veracity
- Big Data is used to create predictive models that answer:
  - What will happen?
  - How should we act?



Machine Learning

What is Machine Learning?

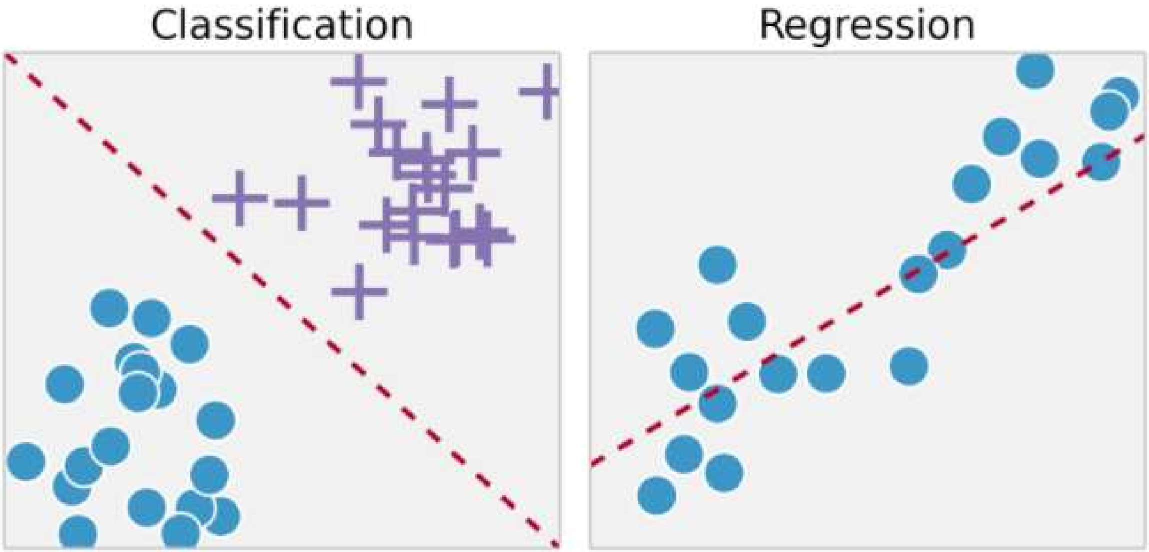
- Kevin Patrick Murphey defines machine learning as “…a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty.”

  - Machine learning algorithms improve their performance on specific tasks based on repeated performance of those tasks. Machine learning methods are applied to a wide range of applications including speech recognition, medical diagnostics, self-driving cars, sales recommendation engine, and many others.

Machine Learning

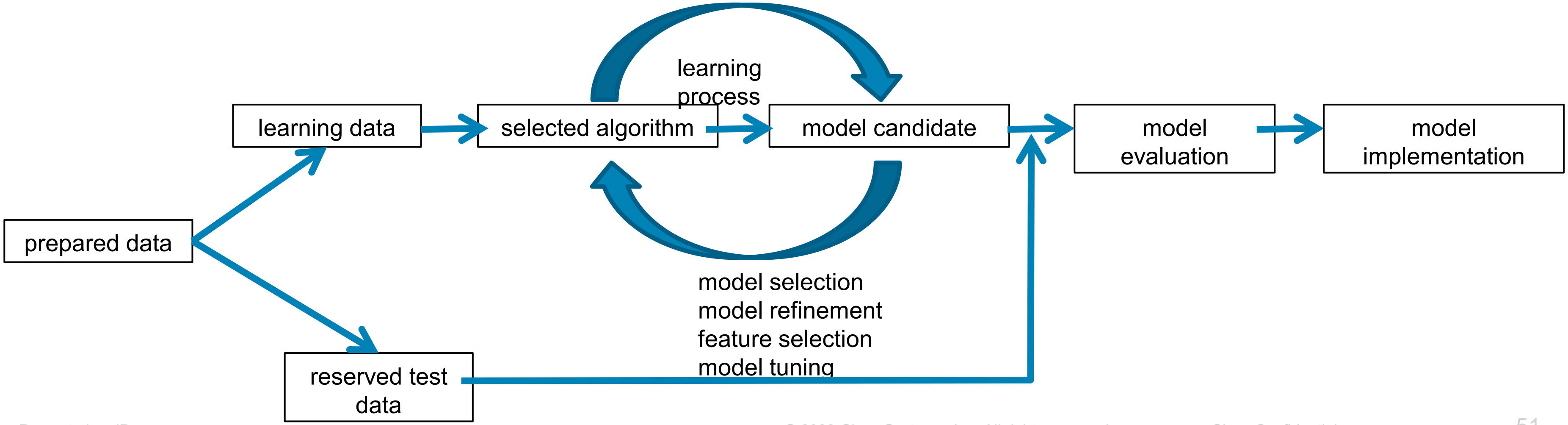
Types of Machine Learning Analysis

- Two main categories of machine learning algorithms:
  - Supervised – commonly used for predictive analytics. The are used to solve regression and classification problems.
  - Unsupervised – they autonomously discover patterns in data. Examples of problems solved with unsupervised methods are clustering and association.



A Machine Learning Process

- Developing machine learning solutions can be simplified into the following steps:
  - Step 1 – Prepare the data
  - Step 2 – Create a learning set
  - Step 3 – Create a test set
  - Step 4 – Create a loop
  - Step 5 – test the solution
  - Step 6 – Implement the solution



Machine Learning

Common Applications of Machine Learning

- Predictive analytics algorithms have a wide range of applications, including the use of analytics technology in the fields of entertainment, agriculture, medicine, and retail sales.

Large retail chain stores use IoT sensors to identify the location of shoppers within their stores. The predictive analytics system then sends targeted sales offers to the shopper's cell phone in real time.

Classification algorithms can help viewers find videos they will like. Based on a customer's video rental behavior and the behavior of other customers, the algorithm predicts which videos a customer is likely to enjoy and makes recommendations to the customer.

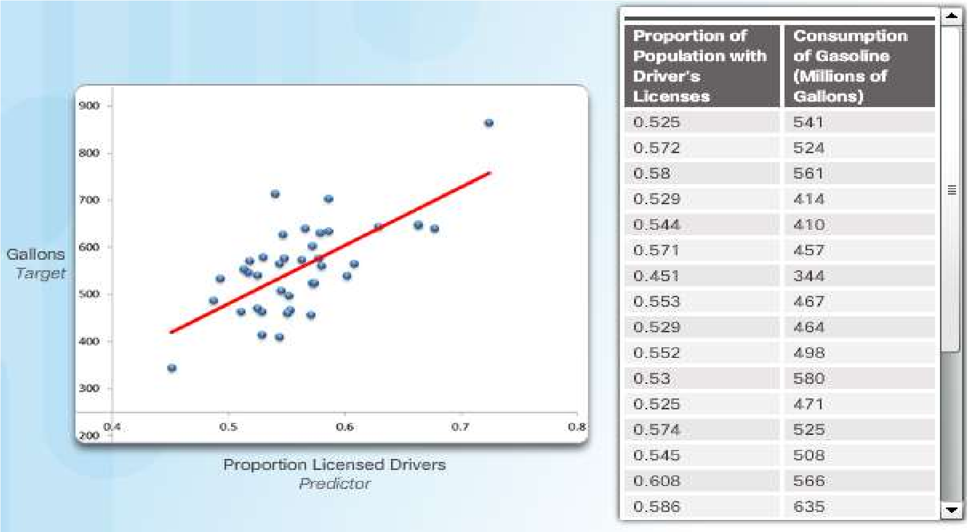
Farmers use cellphones to provide researchers with images of plant diseases. These images are used in image recognition systems to diagnose plant diseases. Combined with environmental data regression algorithms could then predict future outbreaks of disease.

A machine learning classification algorithm uses 20 input variables to predict the possibility of breast cancer. This approach can accurately identify patients who should carefully be watched for early detection of the disease.

Regression

Regression Analysis

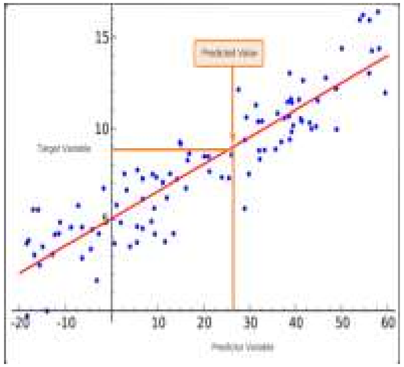
- Regression Analysis is one of the oldest and most commonly used statistical methods for analyzing data.
- The main goal of regression is to qualify the mathematical relationship between one or more independent variables (predictor variable(s)), and a dependent one (target variable).

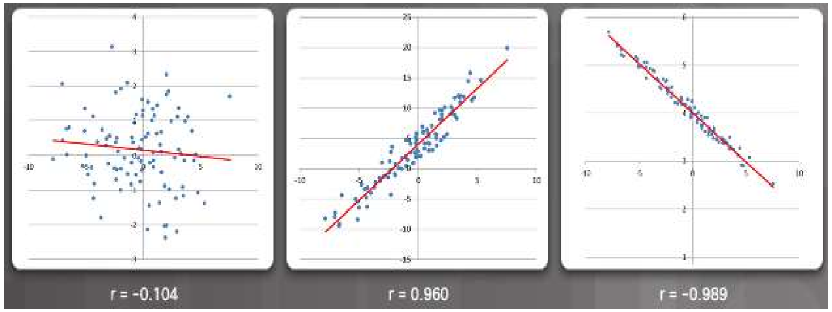


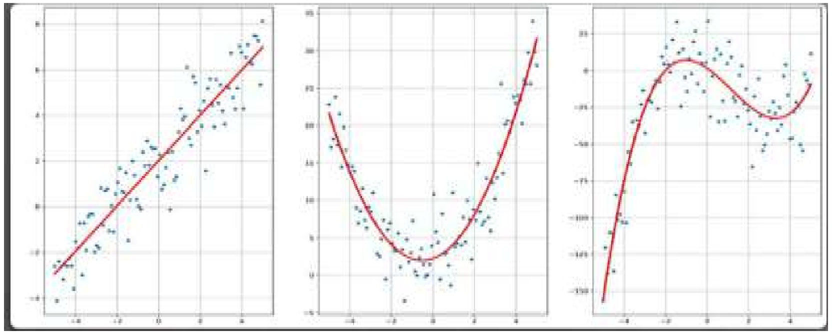
Proportion of Population with Driver's Licenses	Consumption of Gasoline (Millions of Gallons)
0.525	541
0.572	524
0.58	561
0.529	414
0.544	410
0.571	457
0.451	344
0.553	467
0.529	464
0.552	498
0.53	580
0.525	471
0.574	525
0.545	508
0.608	566
0.586	635

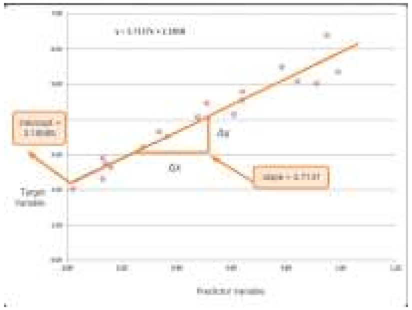
Linear Regression

- Linear regressions are the simplest from both a computational and mathematical point of view.
  - The term linear implies that the regression function will always try to fit the data using a weighted average of other functions, whether those functions are linear or not.









Regression

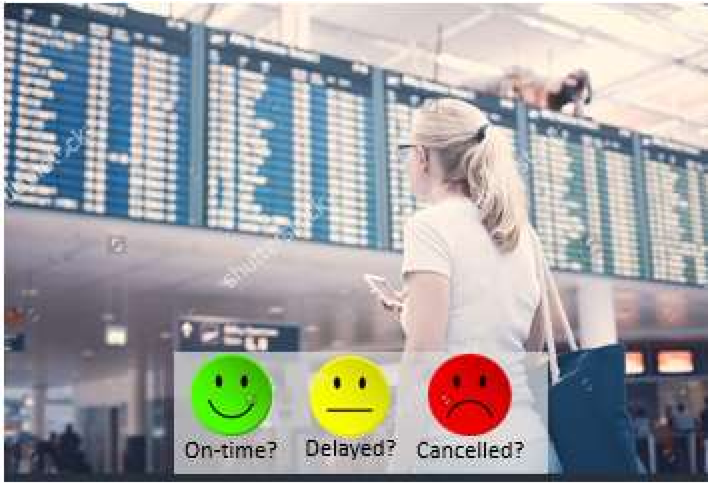
Applications of Regression Analysis

- Regression Analysis has many applications. It is frequently used in business and financial analysis with historical data to inform strategies for future action.
- It can be used to predict trends in economics and can inform political action to guide economic growth.
- Customer behavior can also be predicted to determine normal from possibly fraudulent behavior in fields of insurance and consumer credit.

Statistical Analysis

Classification Problems

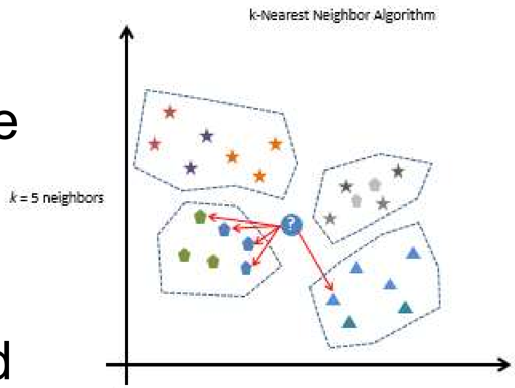
- Classification can be seen as a regression problem where the target variable is **discrete,** and represents a class in which a human expert has classified the data sample.
  - For example, a web-based travel company is interested in providing a reliability rating for the flights that it finds for customers. Via trial error of different models, it has been determined which variables among all the ones in the dataset are the most relevant for the classifications. This is also known as the variables with the highest discriminant power. Only these relevant features are extracted from the data and used to train the classifier.

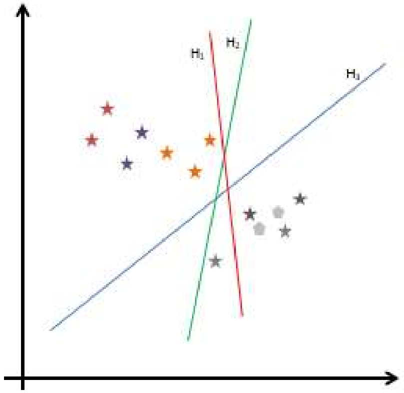


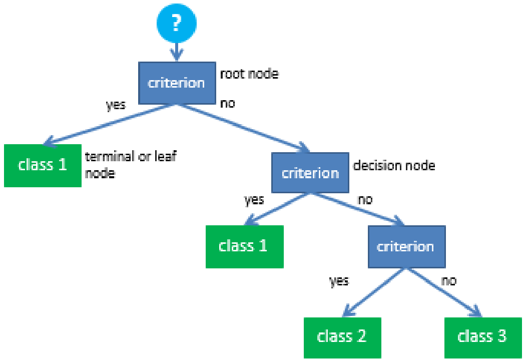
Statistical Analysis

Classification Algorithms

- **k-nearest neighbor (k-NN)** - k-NN is possibly the simplest classifier, which uses the distance between training examples as a measure of similarity. To visualize how a k-NN classifier works, imagine that each sample has two features, for which the values can be represented in a 2D plot.
- **Support vector machines (SVM)** - Support vector machines (SVM) are examples of supervised machine learning classifiers. Rather than basing the assignment of category membership on distances from other points, support vector machines compute the border, or hyperplane, that better separates groups.
- **Decision trees** - Decision trees represent a classification problem as a set of decisions based on the values of the features. Each node of the tree represents a threshold over the value of a feature, and splits the training samples in two smaller sets.







Statistical Analysis

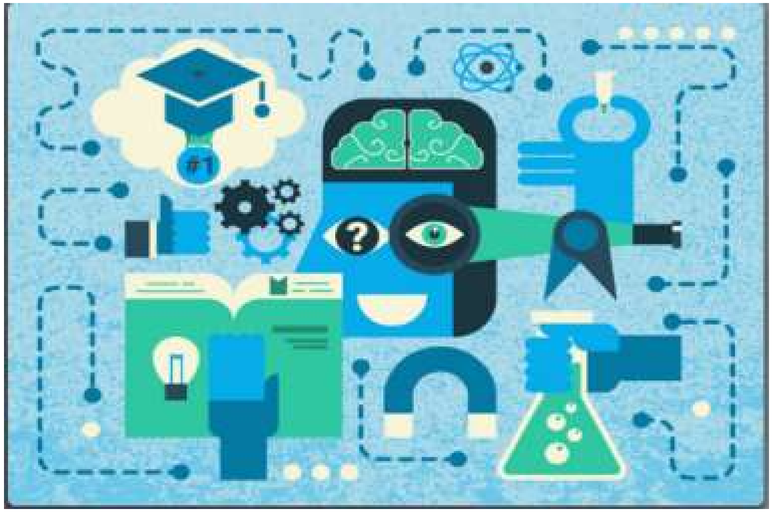
Applications of Classifications

- Classification algorithms have many applications. For example:
  - **Risk Assessment** - Classification systems can be used to determine which of many factors contribute to the likelihood of various risks.
  - **Medical Diagnostics** - Classification systems can use guided questions to build a decision tree that can help diagnose various diseases and risks of disease.
  - **Image Recognition** - In handwriting recognition, a system may be working at the task of identifying handwritten numerals.

4.2 Validity and Reliability

Issues in Using analysis

- The six step process for scientific discovery are:
  - Ask a question about an observation
  - Perform research
  - Form a hypothesis
  - Test the hypothesis
  - Analyze the data from the experiments to draw a conclusion
  - Communicate the results



Validity and Reliability

Validity

- While there are many terms used to describe types of validity, researchers typically distinguish between four types of validity:
  - **Construct validity** - Does the study actually measure what it claims to measure?
  - **Internal validity** - Was the experiment designed correctly? Does it include all the steps of the scientific method?
  - **External validity** - Can the conclusions apply to other situations or other people in other places at other times? Are there any other causal relationships in the study that might account for the results?
  - **Conclusion validity** - Based on the relationships in the data, are the conclusions of the study reasonable?

Validity and Reliability

Reliability

- A Reliable experiment or study means that someone else can repeat it and achieve the same results. Researchers distinguish between four types or reliability:
  - **Inter-rater reliability** - How similarly do different people score on the same test?
  - **Test-Retest Reliability** - How much variation is there between scores for the same person taking a test multiple times?
  - **Parallel-Forms Reliability** - How similar are the results of two different tests that are constructed from the same content?
  - **Internal Consistency Reliability** - What is the variation of results for different items in the same test?

Error in Analysis

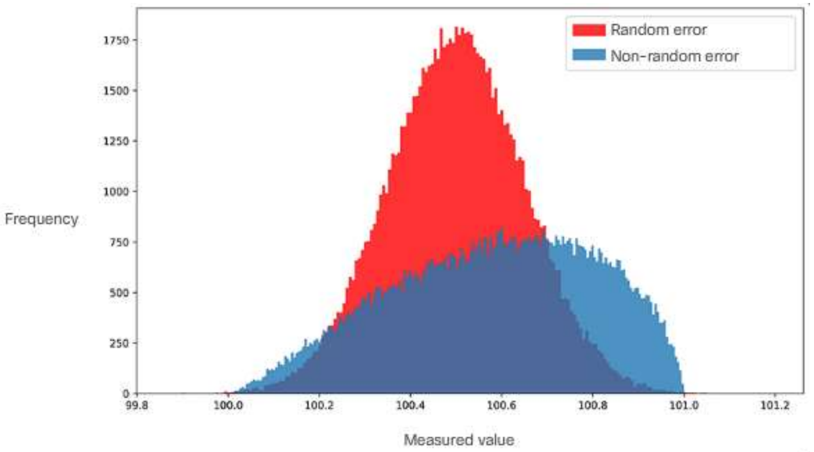
Error in Data Analytics

- Errors, and more in general, uncertainty, affect the data analytics process at different levels:
  - The first type of error is the **measurement error**. Any device for taking measurements is limited in its precision. Therefore, all measurements have a built-in error component.
  - Another type of error is the **prediction error**. In supervised learning, the prediction error is quantified as the difference between the value predicted by the model and the observed value.

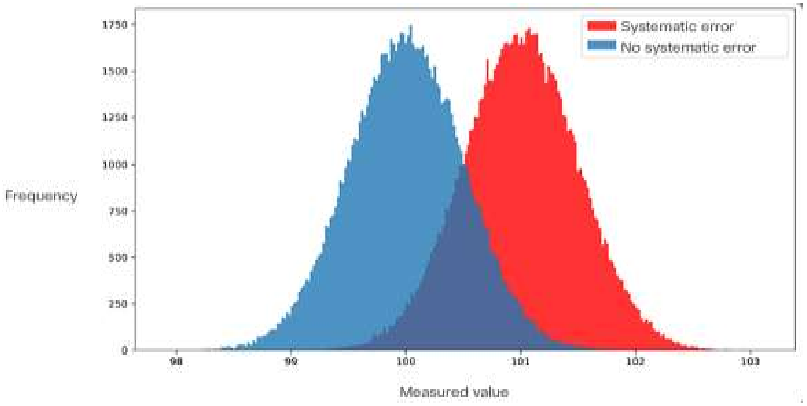
Error in Analysis

Types and Sources of Measurement Error

- Measurement errors can be categorized into these three groups:
  - **Gross errors** - These are caused by a mistake in the instrument being used to take the measurement, or in recording the result of the measurement.
  - **Random errors** – These are caused by factors that randomly impact the measurement over a sample of data.
  - **Systematic errors** – These are caused by instrumental or environmental factors that impact all measurements taken over a given period of time.



Random errors



Systematic errors

Error in Analysis

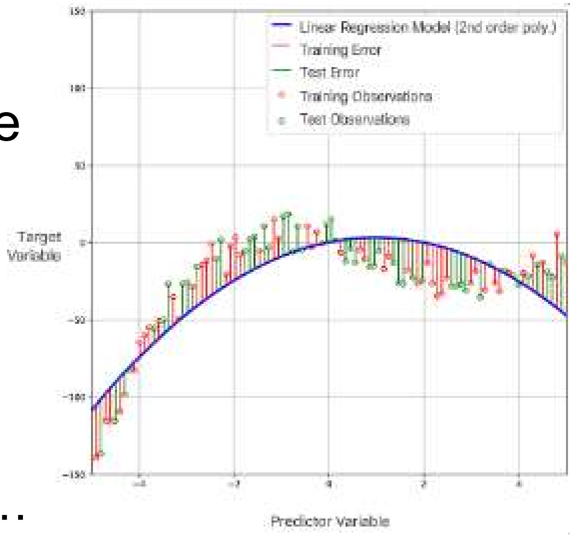
Random Error Distribution

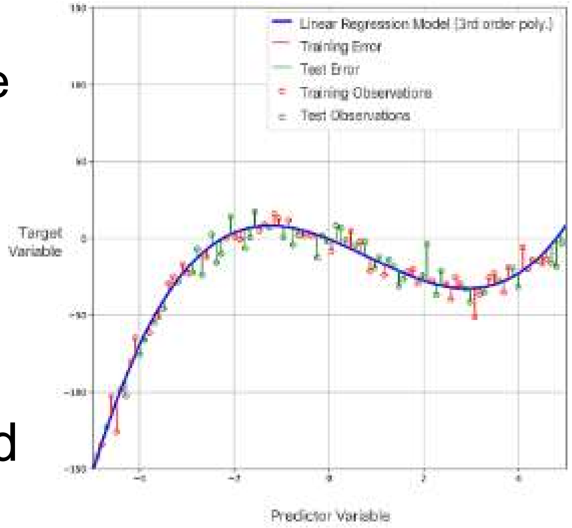
- **Random errors** tend to create a normal distribution around the mean of the observation. It is possible to build a statistical model of the error, in which case regression and classification algorithms can easily take it into account.
- **Systematic errors** tend to shift the distribution of the observations (right side of the figure) in one direction or another. A systematic error is therefore harder to deal with, because the true value is not known, so the only way to detect a systematic error is to use another measurement system that we deem more reliable.

Error in Analysis

Errors in Predictive Analytics

- **Prediction error is a difference** between the value predicted by the regression or classification model, and the measured value.
- **Prediction error is the distance** between the regression function, and the data points. The prediction error has **two components**
  - The first component is caused by the choice of model… we make an assumption on how the data is distributed, which is inevitably an approximation.
  - Even when the chosen model perfectly reflects the true distribution, there will still be differences between predicted and actual values because of the measurement error.
- In machine learning, the first cause of prediction error is often called **bias** of a model, while the second is **variance**. One cannot minimize both, and this situation is often called the **bias-variance tradeoff.**





Model Evaluation

Misleading Research

- Understanding the impact of validity, reliability, and errors in a pattern of data is an important first step to ensuring that your conclusions are based on a solid research design.
- Misleading, bad, or erroneous research is more common than you may think. In fact, John P.A. Ioannidis states that most research findings are false.

Guidelines for Evaluating Results

- There are several guidelines you can following when evaluating the results reported by a research study or a data analysis report:
  - **Statistics** - Does the study have a large enough sample size to support the findings?
  - **Research design** - Did the architects of the study follow generally accepted methods of research design?
  - **Duration** - Does the research appropriately account for the impact on time?
  - **Correlation and causation** - Just because two variables are correlated does not mean that one caused the other.
  - **Alignment to other studies** - Do the results confirm or align with other studies in the field?
  - **Peer review** - Has the study been reviewed by experts in the same field?

4.3 Preparation for Chapter 4 Labs

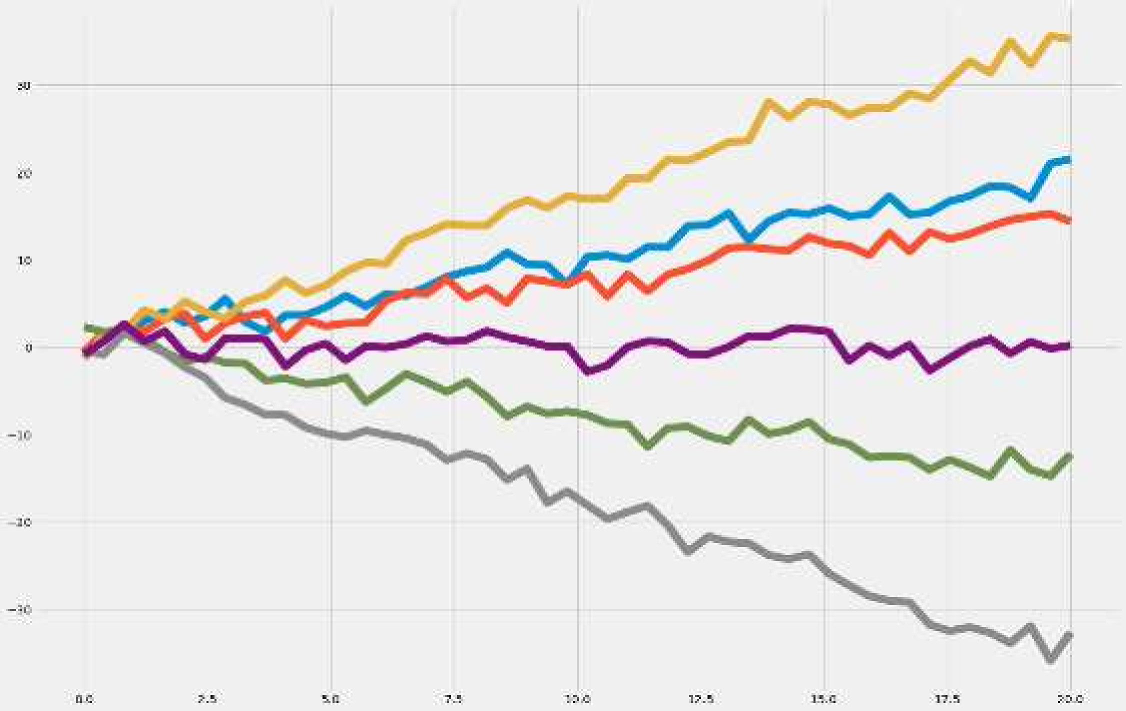
Using scikit-learn for Regression Analysis

- **scikit-learn is a machine learning library for Python** built on NumPy, SciPy, and matplotlib
- In the first lab, you will use regression analysis to view historical data about the growth of Internet traffic. You will quantify the relationship between the year and the measurement of Internet traffic.

Preparation for Chapter 4 Labs

Style Sheets for Plots

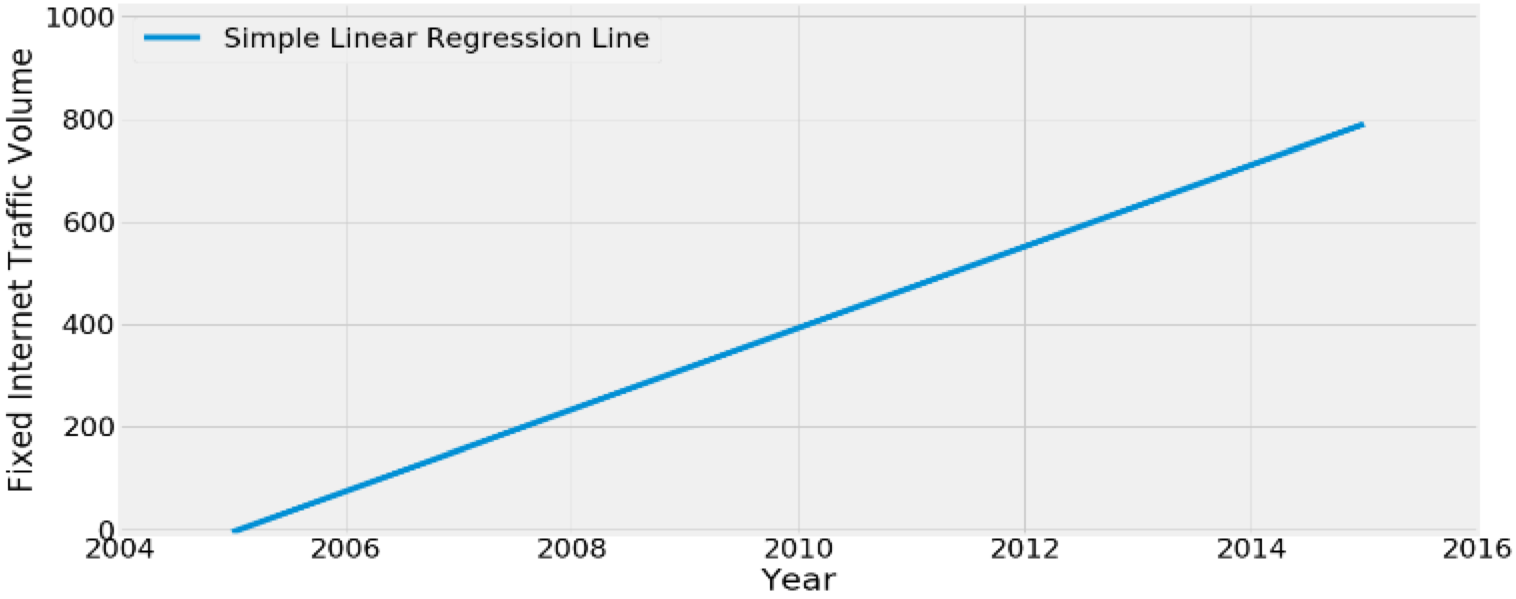
- You will install pandas, numpy, and matplotlib. The matplotlib library includes different styles for showing your plots.



Preparation for Chapter 4 Labs

Fitting the Data

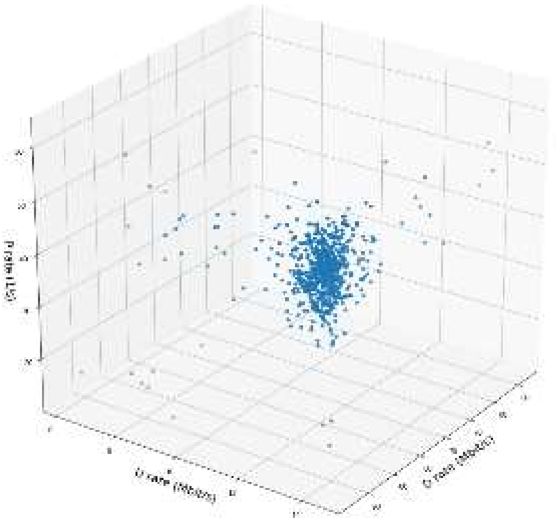
- To do linear regression in Python, you will call on the Numpy class, polyfit. Although polyfit has many arguments, you will only define the values for x, y, and deg. The value for x and y will be used for the x and y axis. Using polyfit will allow you to plot the simple linear regression shown in the figure. The value for deg will define the degree of fit..



Preparation for Chapter 4 Labs

Plotting in 3D

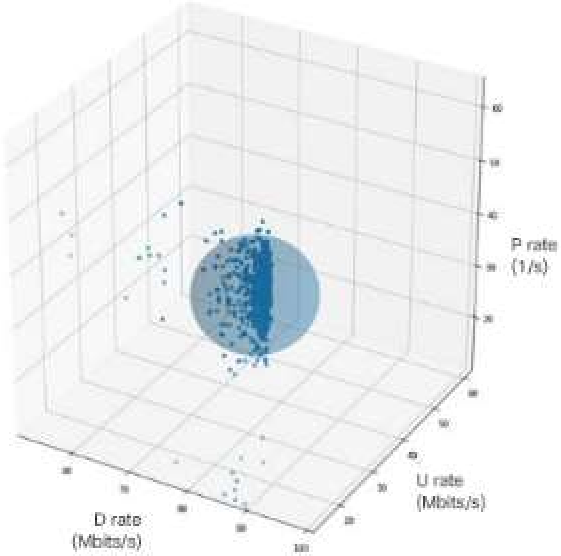
- You will visualize data in three dimensions. To do so, you will extend the matplotlib library by installing the mpl_toolkits class from the mplot3d library. You will then use the Internet meter data to create a 3D plot to display three axis: download rate (x axis); upload rate (y axis); and ping rate (z axis). This visualization will display where the rates for most of the pings cluster

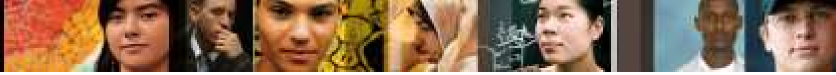


Preparation for Chapter 4 Labs

Visualizing the Boundary for Anomalies

- Data Anomalies can be caused by corruptions or distortions during measurement, transmission, or storage. These values are considered outliers. They deviate so far from expected values that they could distort the results of the analysis.
- Anomalies are frequently removed from the data set after careful consideration.
- The sphere shows the decision boundary between normal data and anomalous data.



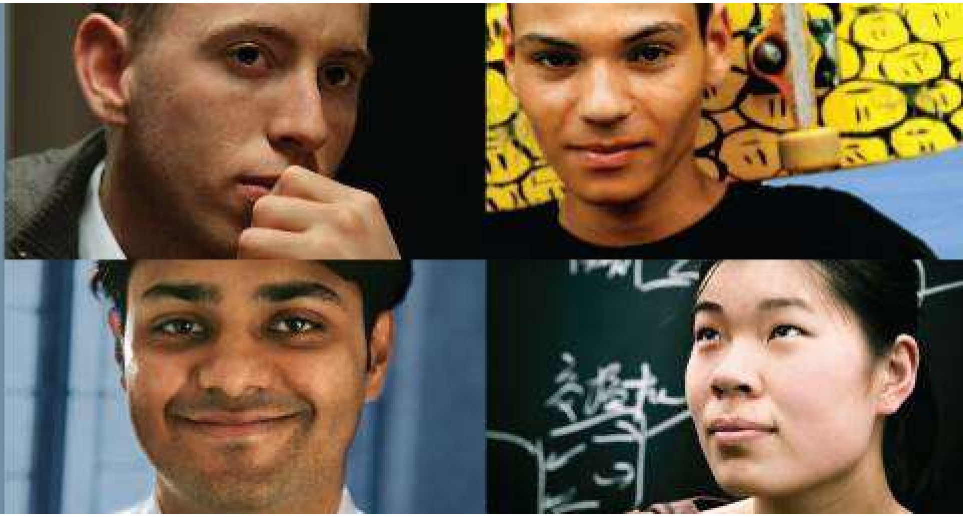


Chapter Summary

Summary

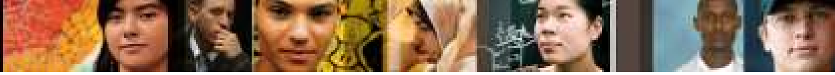
- Big Data is characterized by volume, velocity, variety and veracity.
- Examples of supervised machine learning approaches, ie: Regression and Classification.
  - Regression uses historical relationship between one or more independent variables and a dependent variable to predict future values of dependent variables.
  - Classification models are knows as classifiers. There are numerous classifier algorithms. Example: k-nearest neighbor, Support vector machine and Decision tree.
- The chapter discusses the six step process used by the scientific method for validating the evaluation model.
- The four types of validity are: construct, internal, external, and conclusion.
- The four types of reliability are: inter-rater, test-retest, parallel-forms, and internal consistency.
- Error is the difference between the actual value and the measured value of an observation.

Chapter 5: Storytelling with Data



Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™



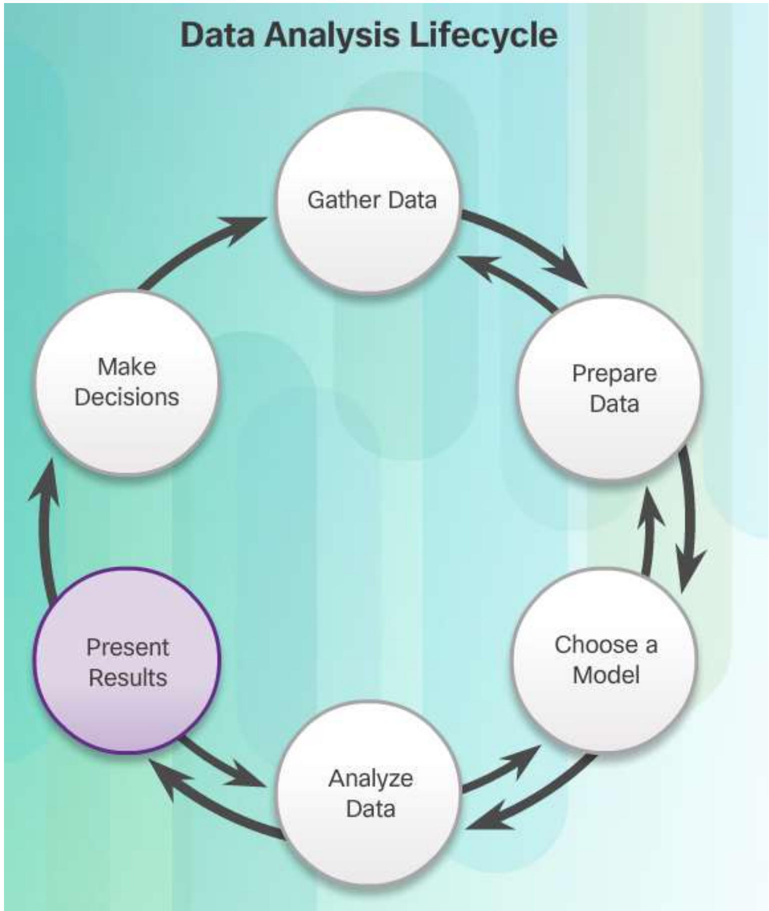
Chapter 5 - Sections & Objectives

- 5.1 Building a Data Story
  - Explain the fundamentals of creating an argument from data.
- 5.2 The Power of Visualization
  - Explain how to use Python libraries to create the appropriate visualizations for a communicative purpose.
- 5.3 Preparation for Chapter 5 Labs
  - Describe the sources of Big Data.

5.1 Building a Data Story

Telling a Story

- Results of data analysis are shown during the *Present Results* part of the data analysis lifecycle.
- Results drive changes made by decision makers.
- Do <u>not</u> spend too much time on the data. Give enough to explain your point.



Building a Data Story

Audience

- Who is your audience?
  - Who will hear the story?
  - What is the listener’s motivation?
  - What is the listener’s level of knowledge and familiarity with the business problem?
  - What are possible reactions?
- Where is your audience?
  - Online
  - Audio only
  - Face-to-face
  - Will the presentation be shared?
- When is your audience available?
  - What to do if someone cannot attend
  - Record the presentation?
  - Confidentiality/Security concerns

Building a Data Story

Business Value and Goal

- Business value means different things to different audiences so be clear on why someone should care about the story being told
- What do you want members of the audience to take away?
- What is the call to action, if any?

Building a Data Story

Using Evidence

- Should be critical to the end goal
- If a piece of evidence does not support concluding remarks or is secondary to the primary focus, consider leaving the evidence out of the presentation.

Deductive Reasoning

- Uses facts or premises to arrive at a conclusion
- Considered “top-down” because it moves from a general premise to specific facts derived from the general premise
- Sound deductive reasoning leads to conclusions that are true.
- Example: syllogism All mammals have eyes. Humans are mammals. Therefore, humans have eyes.

Building a Data Story

Inductive Reasoning

- Moves from specific to general
- Create a conclusion based on observations, patterns, and hypotheses
- We sample a population, study the sample, and then make inferences that we believe will be true for the entire population.
- Be sure the sample represents the population to which the conclusion is being applied

Building a Data Story

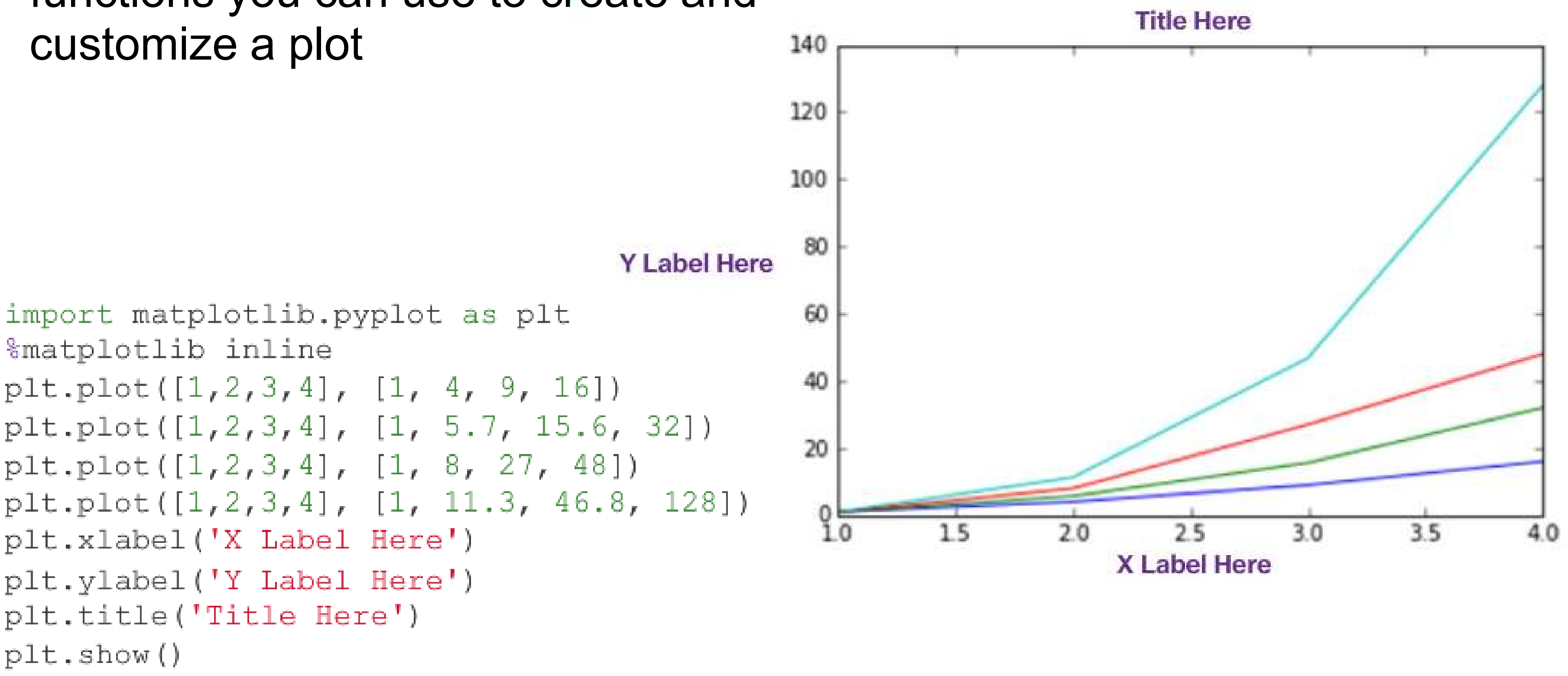
Fallacies

- Argument might not apply a rule of logic
- Argument might leave out or misinterpret a crucial premise
- Conclusion might not follow logically from the premise(s)
- Formal Fallacy
  - One or more premises shown to be false
  - If milk is kept in the refrigerator, it will not spoil. The milk is spoiled. Therefore, the milk was not kept in the refrigerator.
- Informal Fallacy
  - One or more premises do not adequately support the conclusion
  - Some people have psychic powers. Can you prove it? No one has been able to disprove it.

5.2 The Power of Visualization

Pyplot

- Pyplot is a Matplotlib module.
- Pyplot includes a collection of style functions you can use to create and customize a plot

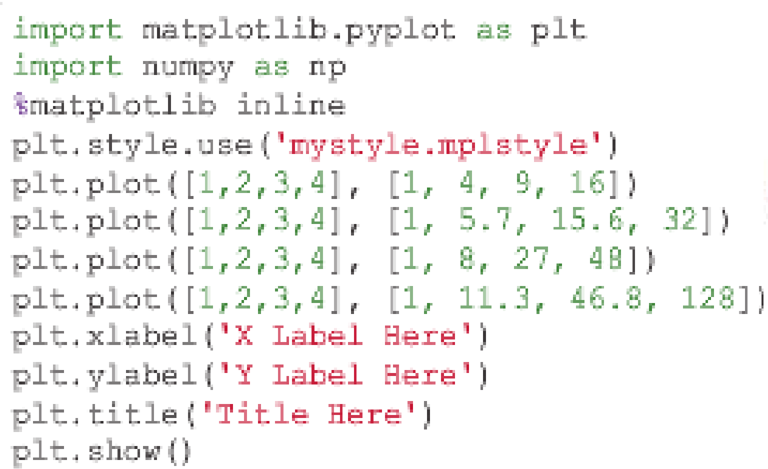


```
import matplotlib.pyplot as plt
%matplotlib inline
plt.plot([1,2,3,4], [1, 4, 9, 16])
plt.plot([1,2,3,4], [1, 5.7, 15.6, 32])
plt.plot([1,2,3,4], [1, 8, 27, 48])
plt.plot([1,2,3,4], [1, 11.3, 46.8, 128])
plt.xlabel('X Label Here')
plt.ylabel('Y Label Here')
plt.title('Title Here')
plt.show()
```

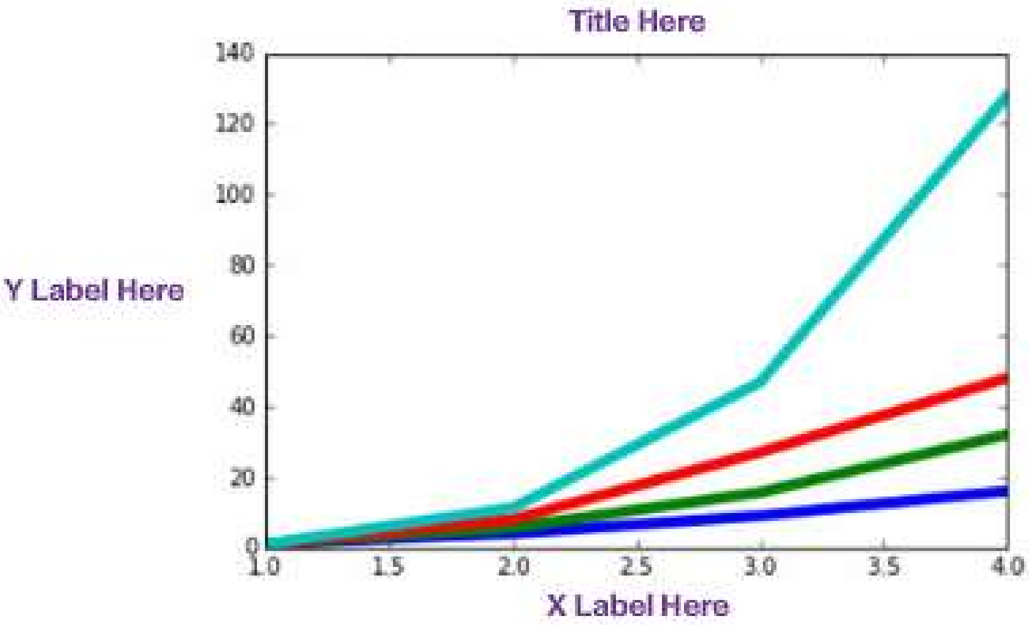
The Power of Visualization

Pyplot Custom Style Sheet

- In pyplot you can create a custom style sheet so all plots have the same style feature and you avoid making minor errors to the inline code.
  - If you store the style sheet in a non-default location, you must provide path information when you reference it.



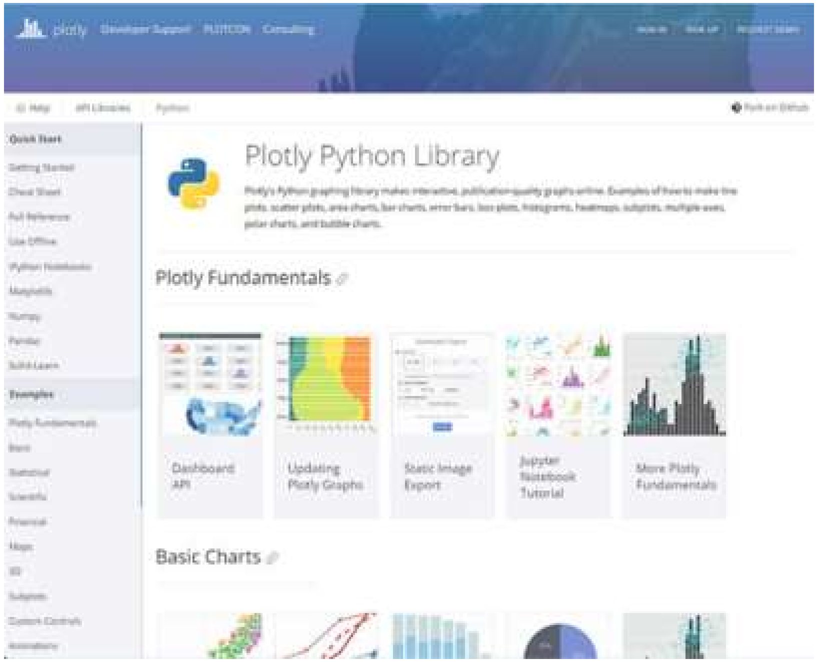
```
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline
plt.style.use('mystyle.mplstyle')
plt.plot([1,2,3,4], [1, 4, 9, 16])
plt.plot([1,2,3,4], [1, 5.7, 15.6, 32])
plt.plot([1,2,3,4], [1, 8, 27, 48])
plt.plot([1,2,3,4], [1, 11.3, 46.8, 128])
plt.xlabel('X Label Here')
plt.ylabel('Y Label Here')
plt.title('Title Here')
plt.show()
```



The Power of Visualization

Plotly

- Plotly is an online tool to generate data visualizations.
- Has resources including free content, API libraries, figure converters, apps for Google Chrome, and an open source JavaScript library
- Code, images, and data can be exported

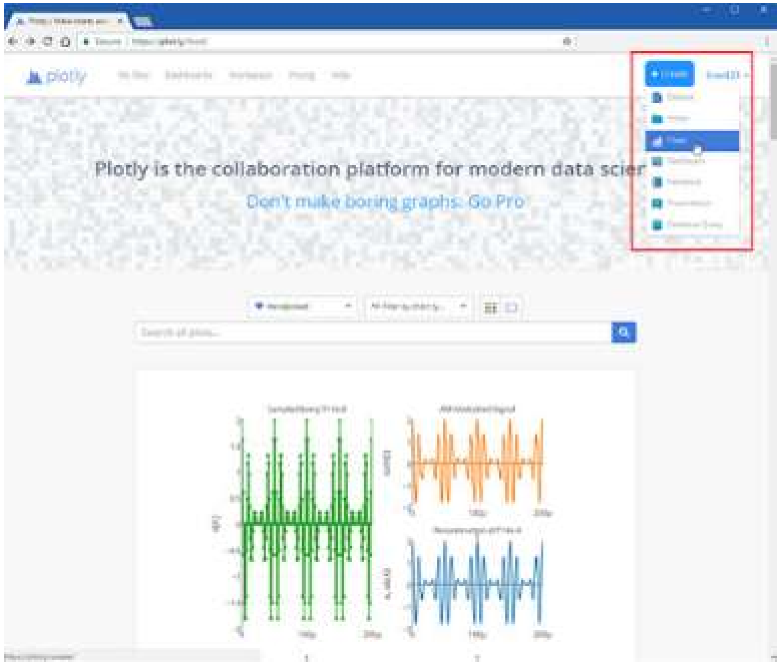


Plotly Python Library

Plotly's Python graphing library makes interactive, publication-quality graphs online. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, and bubble charts.

Plotly Fundamentals

Basic Charts

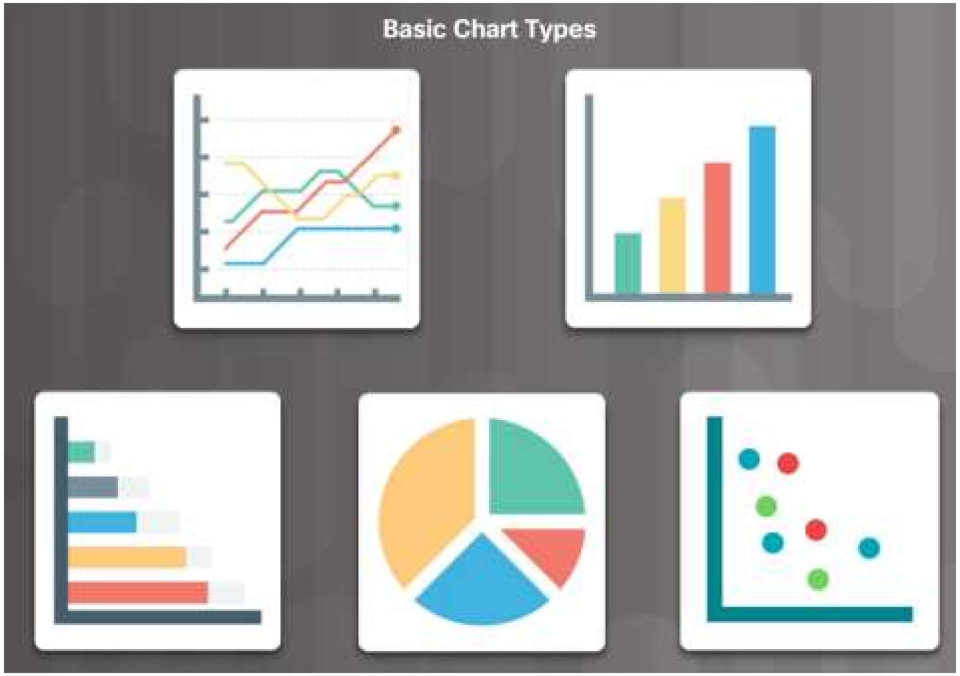


Plotly is the collaboration platform for modern data scier

Don't make boring graphs. Go Pro

Common Types of Data Visualizations

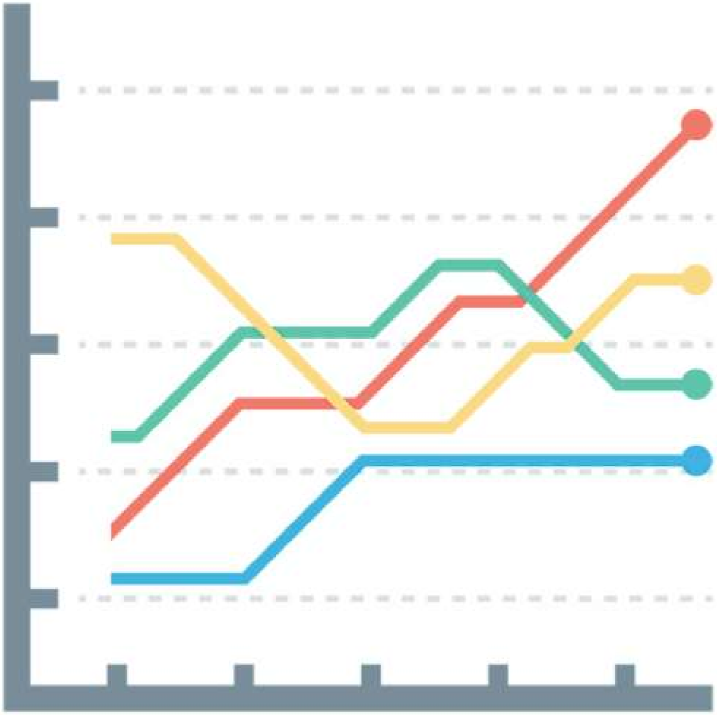
- How many variables?
- How many data points are in each variable?
- Is the data over time or comparing items?
- Most popular charts: line, column, bar, pie, and scatter



The Power of Visualization

Line Charts

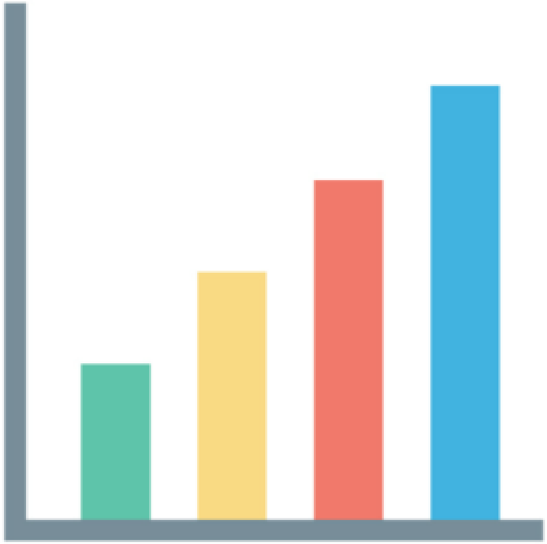
- One of the most commonly used
  - Used when there is a continuous set of data, the number of data points is high, and you would like to show a trend in data over time
- Examples
  - Quarterly sales for past five years
  - Number of customers per week in the year
- Best practices
  - Label axes.
  - Plot time on the x-axis (horizontal).
  - Plot data values on the y-axis (vertical).
  - Keep data sets to a minimum.
  - Minimize gridlines.
  - Modify the axis starting point if necessary.



The Power of Visualization

Column Chart

- Positioned Vertically
  - Most commonly used when you want to display the value of a specific data point compared across similar categories
- Examples
  - Population of five nations
  - Yearly sales for four companies
- Best practices
  - Label axes.
  - If time is used, plot on the x-axis (horizontal).
  - Use solid colors for columns.
  - Avoid using more than 7 categories on the horizontal axis.
  - Start the y-axis value at zero.
  - Spacing between columns should be about half the width of a column.



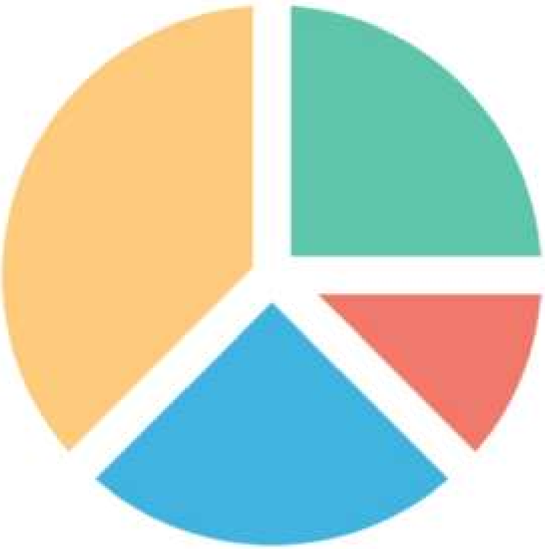
The Power of Visualization

Bar Chart

- Positioned Horizontally
  - Most commonly used when you want to display the value of a specific data point compared across similar categories and the names for each data point is long.
- Examples
  - GDP of 25 nations
  - Car sales by salesman
- Best practices
  - Label axes.
  - Order bars from longest to shortest.
  - Use solid colors for bars.
  - Avoid using more than 7 categories on the horizontal axis.
  - Start the x-axis value at zero.
  - Spacing between rows should be about half the width of a column.

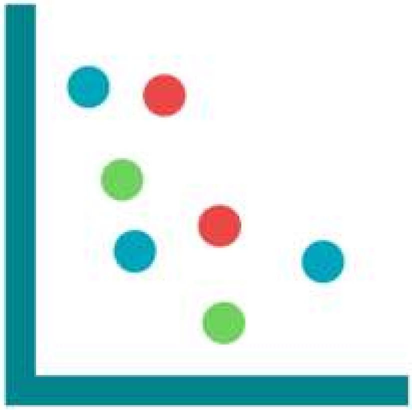
Pie Charts

- Pie chart
  - Used to show the composition of a static number.
  - Segments show a percentage of that number
  - Segments total 100%
- Examples
  - Annual expenses by type
  - Energy sources by type used
- Best practices
  - Keep categories to a minimum. Consolidate when necessary.
  - Use different colors for different segments and order by size.
  - Ensure segment values total 100%.



The Power of Visualization
Scatter Plot

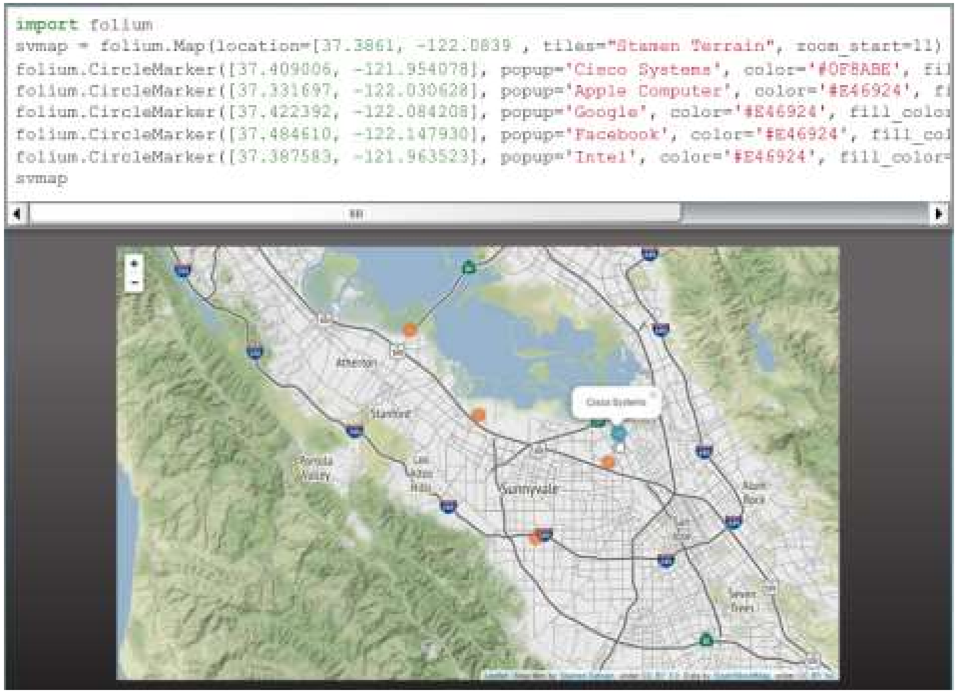
- Clustering and Correlation Visualizations
  - Used to show correlation or distribution of data points
  - Useful in showing clustering or identifying data outliers
- Examples
  - Comparing life expectancy to GDP
  - Comparing daily sales of ice cream to average temperature
- Best practices
  - Label axes.
  - Ensure data set is large enough.
  - Start y-axis at zero. X-axis start value depends on data.
  - Consider adding a trend line, but don’t use more than two.



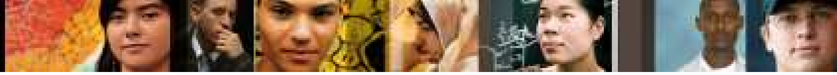
5.3 Chapter 5 Preparation for Chapter 5 Labs

Folium Library

- Combines the strength of Python scripts with the mapping abilities of the Leaflet.js library
- Allows Python data frames to be displayed within an interactive Leaflet map
- Tileset – collection of raster or vector data that can display a map on mobile devices and within a browser



```
import folium
svmap = folium.Map(location=[37.3861, -122.0839 , tiles="Stamen Terrain", zoom_start=11)
folium.CircleMarker([37.409006, -121.954078], popup='Cisco Systems', color='#0F8ABE', fil
folium.CircleMarker([37.331697, -122.030628], popup='Apple Computer', color='#E46924', fi
folium.CircleMarker([37.422392, -122.084208], popup='Google', color='#E46924', fill_colo
folium.CircleMarker([37.484610, -122.147930], popup='Facebook', color='#E46924', fill_co
folium.CircleMarker([37.387583, -121.963523], popup='Intel', color='#E46924', fill_color=
svmap
```

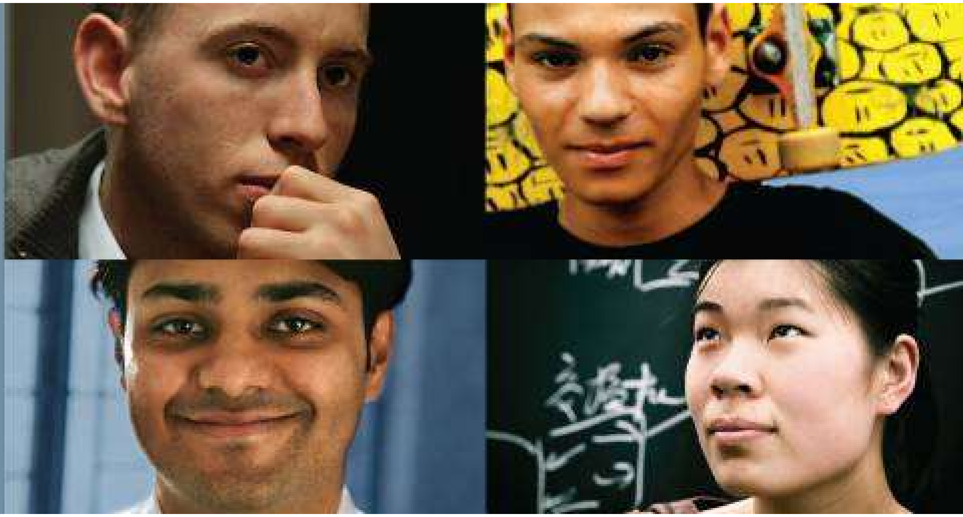


Chapter Summary

Summary

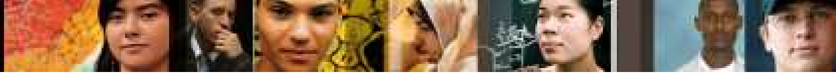
- Data can be summarized using visualizations to help others understand the data.
- Must know who your audience is, where they are, and when the audience is available?
- Evidence presented can be derived from deductive reasoning or inductive reasoning and should not suffer from a logical fallacy (formal or informal).
- Deductive reasoning uses facts, propositions, or other statements of truth to arrive at a conclusion.
- Inductive reasoning creates a conclusion based on observations, patterns, and hypotheses.
- Types of charts used in visualizations are line, column, bar, pie, and scatter.
- Pyplot is a matplotlib extension that includes style functions used to create and customize a plot.
- Plotly is an online tool used to create a visualization.

Chapter 6: Architecture for Big Data and Data Engineering



Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™



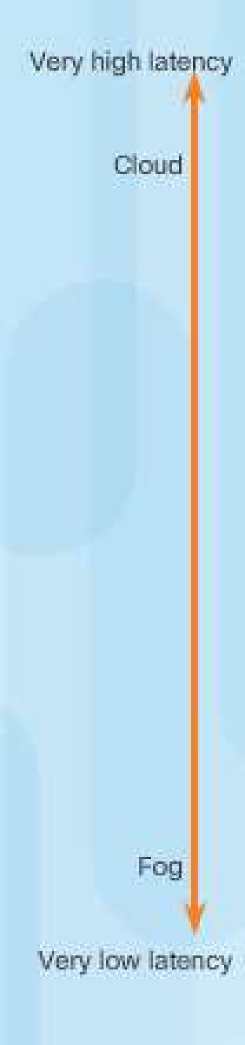
Chapter 6 - Sections & Objectives

- 6.1 Scaling Data Analytics
  - Explain how the virtualized data center supports Big Data and analytics.
- 6.2 Introduction to Data Engineering
  - Explain the history, theory, concept, design, and barriers behind data engineering needs.
- 6.3 The Big Data Pipeline
  - Explain how a big data pipeline supplies streaming IoT data for analysis.
- 6.4 The Image Processing Labs
  - Analyze digital image data.

6.1 Scaling Data Analysis

Edge Analytics and Cloud Analytics

- Transforming data into valuable insights requires computing and storage capacity.
- Device-Network-Cloud - all data points collected by sensors are sent directly to the cloud for storage and processing. This is what happens with most of the wearables used to track fitness activities.
- Device-Gateway-Network-Cloud - when the numbers of sensors increase, or when the processing of the data requires a much shorter response time, data can be processed very near the source of its creation on the gateway or other intermediate places on the network. Known as fog computing.



Scaling Data Analysis

Data Centers and Cloud Computing

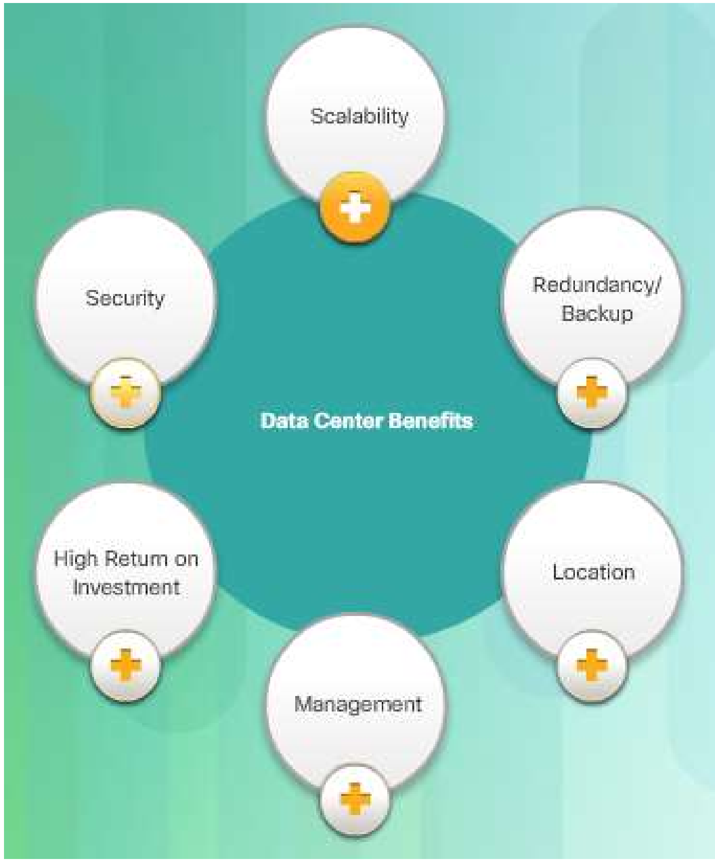
- Cloud Computing supports the four V’s of Big Data: Volume, Variety, Velocity, Veracity
- Enterprise access to data anywhere anytime
- Pay-as-you-go model where you only subscribe to services that are needed
- Reduces costs by not having to purchase costly hardware or physical infrastructure
- Scalable computer storage and processing
- The 3 Main Cloud Services are:
  - SaaS – Software as a service
  - Paas – Platform as a service
  - IaaS – Infrastructure as a service

Software
Platform
Infrastructure

Scaling Data Analysis

Benefits of a Data Center

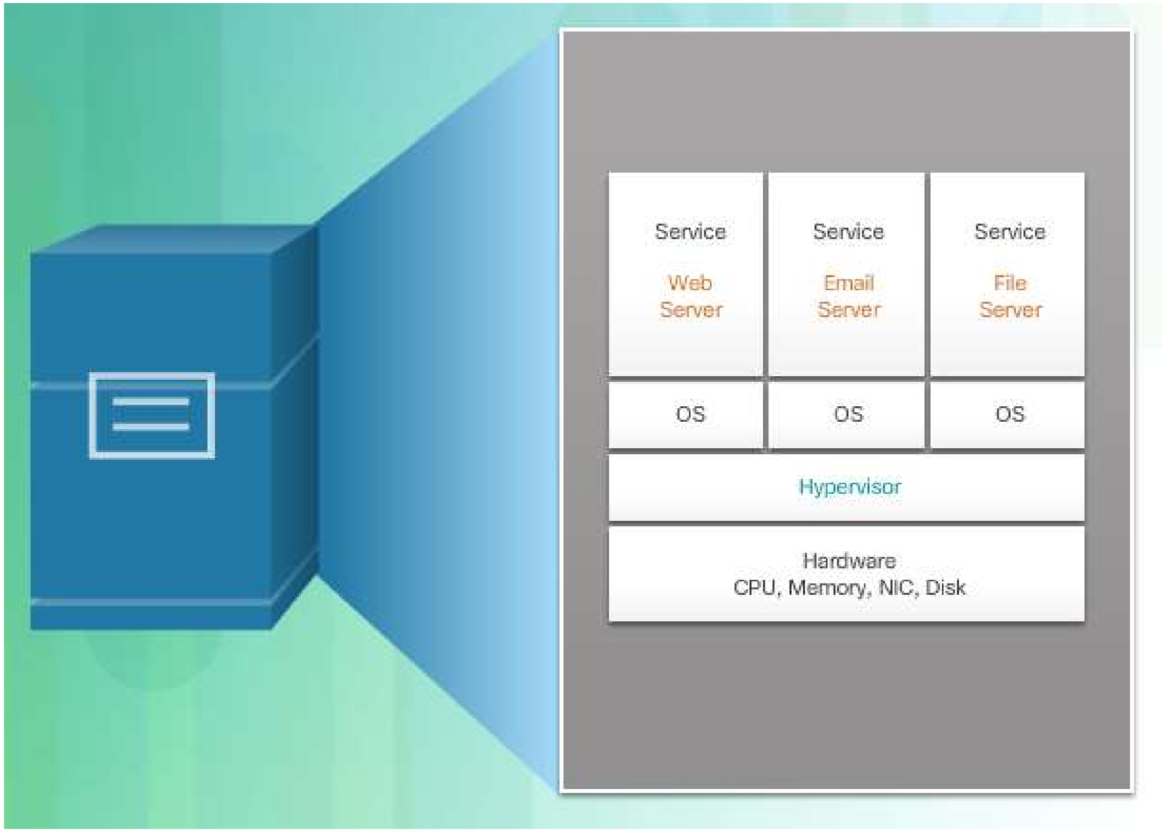
- Some organizations create and maintain their own data centers in-house
- Other organizations rent data center servers at co-location facilities (colos)
- Other organizations use public, cloud-based services like Amazon Web Services, Microsoft Azure, Rackspace, and Google.
- Data centers provide:
  - Scalability,
  - Redundancy/Backup,
  - Location,
  - Management,
  - High return on investment,
  - Security



Scaling Data Analytics

What is Virtualization?

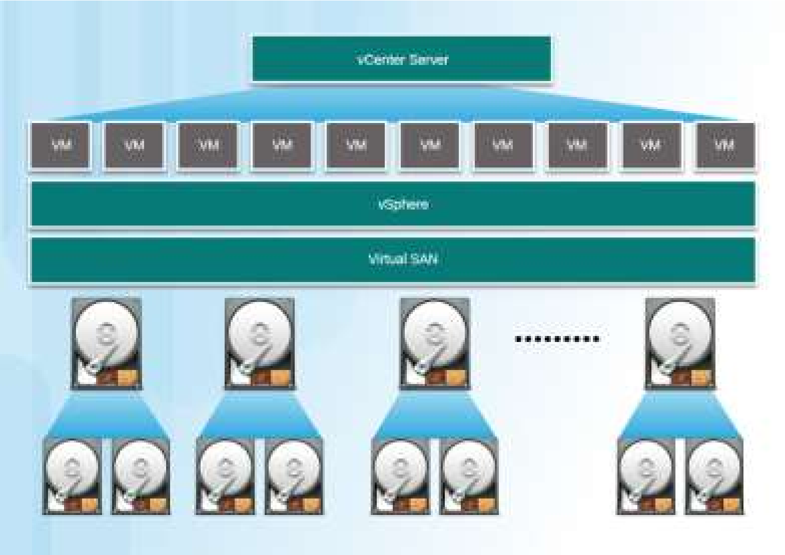
- Virtualization separates the OS from the hardware.
- A hypervisor is software that creates and runs virtual machine (VM) instances.
- Containers are a specialized “virtual area”.

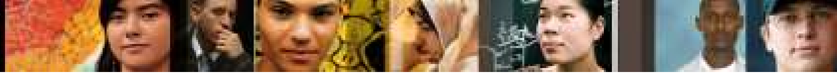


Scaling Data Analytics

The Virtualized Data Center

- Data centers use virtualization to cut costs and expand offerings as cloud providers.
- Storage virtualization combines physical storage from multiple network storage devices into what appears to be a single storage device.
- Network virtualization (NV) is the creation of virtual networks within a virtualized infrastructure.





6.2 Introduction to Data Engineering

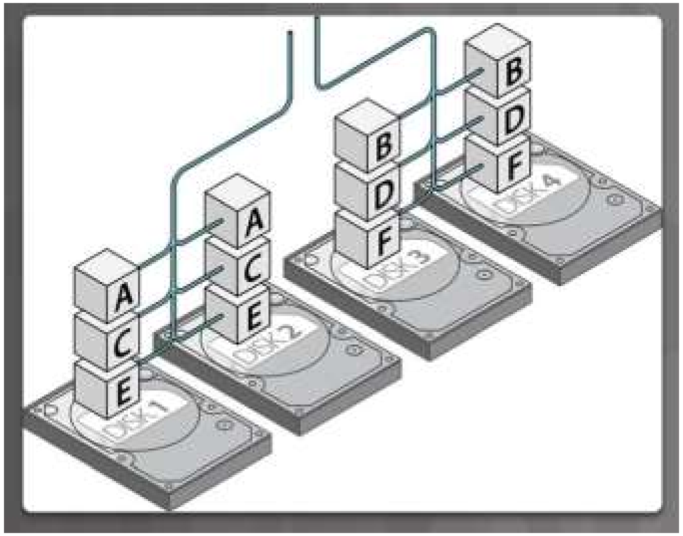
What is Data Engineering?

- Data engineering typically involves a business-related, computer-based information system where information (data) is captured or generated, processed, stored, distributed, and analyzed.
- The ability to capture data and analyze it in a meaningful way is typically done with a database and database management system.
- The relational database emerged around the same time as the personal computer revolution.
  - The relational database and the structured query language (SQL) programming language are the foundation of the relational database management system (RDMS).
- The emergence of the Web 2.0, E-commerce and Google made it obvious that the relational database could not handle the volume and speed of web requests and searches.
- Non-relational databases like NoSQL and Object databases were created to meet the demands of the modern Web.
- Google helped pioneer the emergence of Big Data by openly publishing a paper on MapReduce and distributed processing and storage.

Introduction to Data Engineering

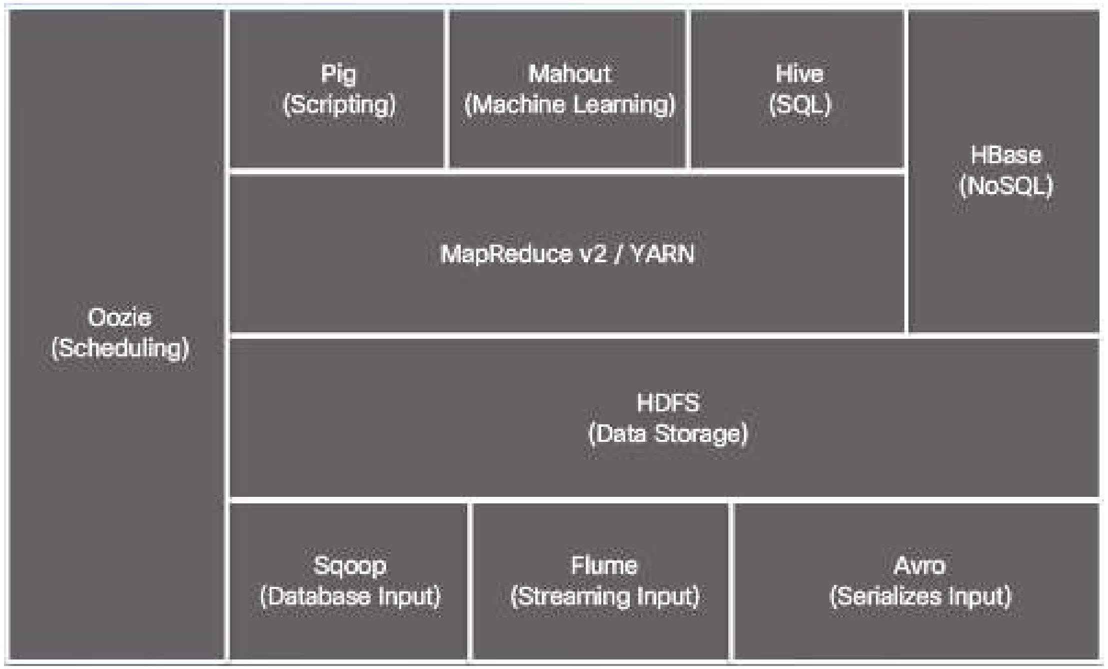
Big Data Systems

- Scalability is the ability to scale both data storage as well as data processing.
- Speed and availability are the primary concern for many companies working with Big Data.
- Fault tolerance is similar to availability in that a company’s business needs to be constantly online and available 24/7.



What is Hadoop?

- The Hadoop Distributed File System (HDFS) is a redundant filesystem that stores data by distributing it across many computers.
- MapReduce is a distributed processing framework for parallelizing algorithms across large numbers of commodity servers.
- Hadoop is not a single application but an ecosystem of applications all working together.



Oozie (Scheduling)

Pig (Scripting)

Mahout (Machine Learning)

Hive (SQL)

HBase (NoSQL)

MapReduce v2 / YARN

HDFS (Data Storage)

Sqoop (Database Input)

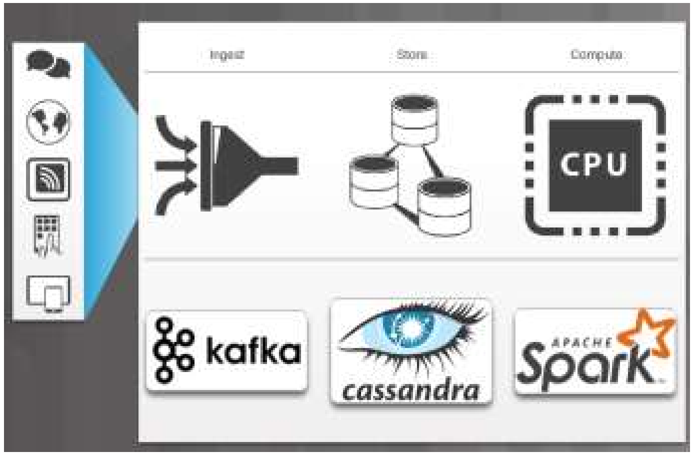
Flume (Streaming Input)

Avro (Serializes Input)

6.3 The Big Data Pipeline

Data Ingestion

- The big data pipeline consists of: data ingestion, data storage, and data processing.
- To ingest data in real-time, a distributed streaming platform such as Kafka must be used.
- What makes Kafka different than traditional message brokers is the use of transaction logs.



The Big Data Pipeline

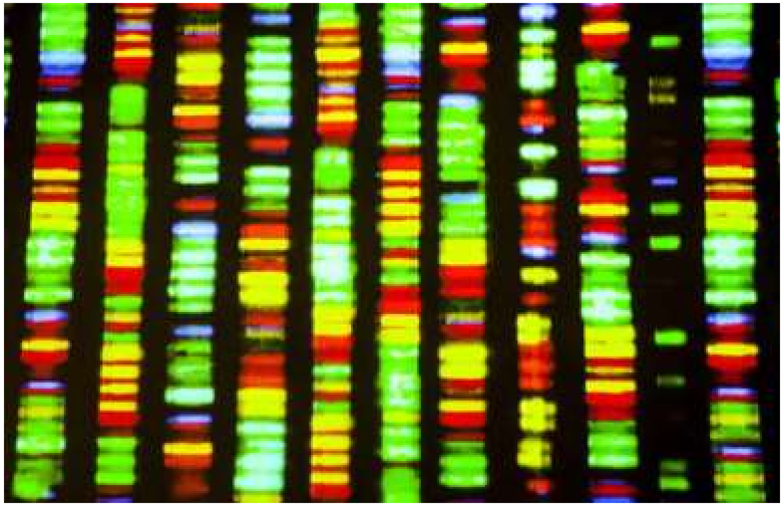
Data Storage

- Big Data generates vast amounts of data that must be stored.
- Cassandra is an open-source NoSQL distributed database management system.
- Cassandra uses the Cassandra File System (CFS).
- With the CFS, analytic metadata is stored in a keyspace.

The Big Data Pipeline

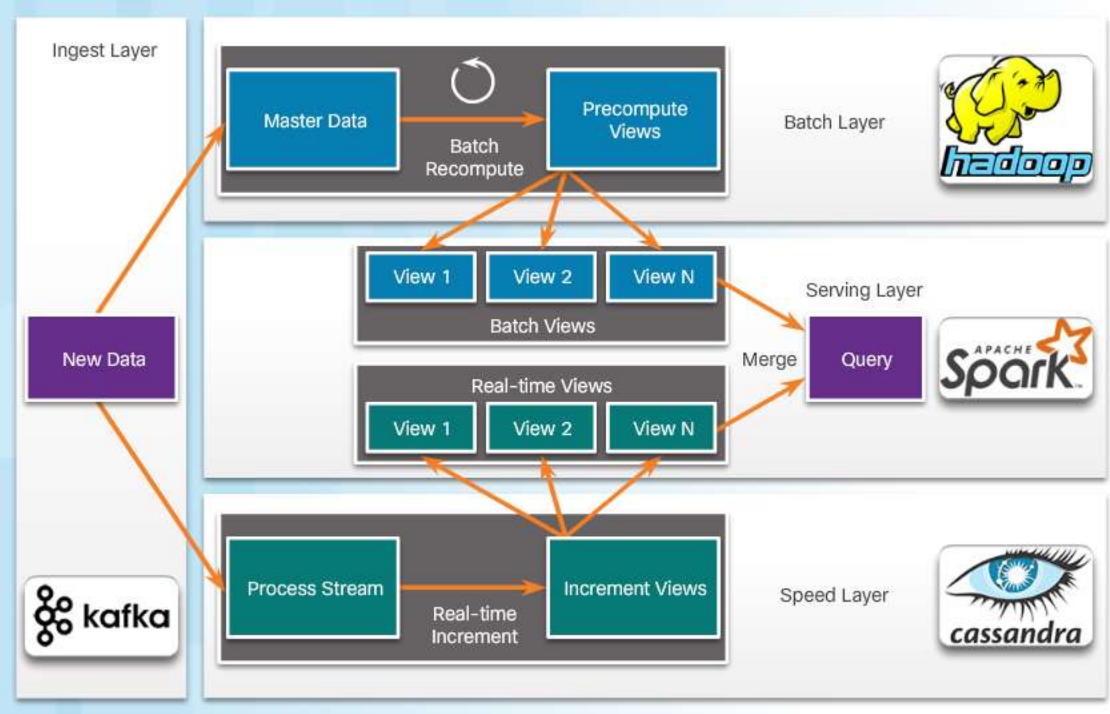
Compute

- The size of the data sets being used in many different fields is a challenge for Big Data.
- Spark is an open-source, distributed data processing engine used for Big Data.
- Spark is able to run right on top of an Hadoop instance, using HDFS for storage and YARN for cluster management.



The Lambda Architecture

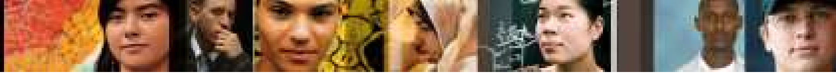
- Lambda is a data processing architecture that uses both stream processing and batch processing to get accurate views of both "live" data and batch data.



6.4 The Image Processing Lab

Digital Images as Data

- Data also includes media, such as images, video, and sound, as data.



Chapter Summary

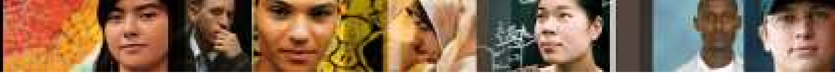
Summary

- Virtualized data center supports Big Data and analytics.
- With fog computing data can be processed almost immediately after it is generated.
- Data centers are centralized locations containing large amounts of computing and networking equipment.
- Virtualization separates the OS from the hardware.
- Storage virtualization combines physical storage from multiple network storage devices into what appears to be a single storage device.
- Network virtualization (NV) is the creation of virtual networks within a virtualized infrastructure.

Chapter Summary

Summary

- Data engineering involves a business-related, computer-based information system where information (data) is captured or generated, processed, stored, distributed, and analyzed.
- Scalability means designing a solution that can meet the exponential growth demands of large companies.
- The Hadoop Distributed File System (HDFS) is the filesystem where Hadoop stores data.
- MapReduce is a distributed processing framework for parallelizing algorithms across large numbers of commodity servers.
- Kafka is used to pipe real-time streaming data between different systems and applications.



Chapter Summary

Summary

- Cassandra uses the Cassandra File System (CFS) which is is not a master-slave architecture like HDFS.
- Cassandra is an open-source NoSQL distributed database management system.
- Spark is an open-source, distributed data processing engine used for Big Data jobs.
- Lambda is a data processing architecture that uses both stream processing and batch processing to get accurate views of both “live” data and batch data.
- In the digital age, media is numeric data also. It is represented by ones and zeros as digital data.