Introduction to Big Data

Day 1
Introduction to Big Data, Analytics and Business Problem Framing

YORK UNIVERSITÉ UNIVERSITY

Week 1 – Major Topics Introduced

The following major topics are introduced and described this week.

- Big Data
- Data Analytics
- Business Questions
- Big Data and Analytics
- Analytical Models
- Statistical Methods
- Python Jupyter Notebooks

Understanding these concepts will provide you with a solid foundation to master the material and develop the necessary skills through the rest of the course.

Lesson 1 - Learning Objectives

Goal of this content:

- Describe the curriculum and the course learning objectives and grading scheme
- Define big data in terms of structure, sources, impact, business opportunities and key characteristics
- Define data analytics based on its recent history, terminology, purpose, structure and capabilities
- Describe key areas data analytics functionality that helps answer different types of business questions
- Describe how combining Big Data with Analytics can drive improved business performance through proper business problem framing
- Describe the structure, purpose and limitations of analytical models
- Describe the roles that statistical methods play in creating useful information from raw data

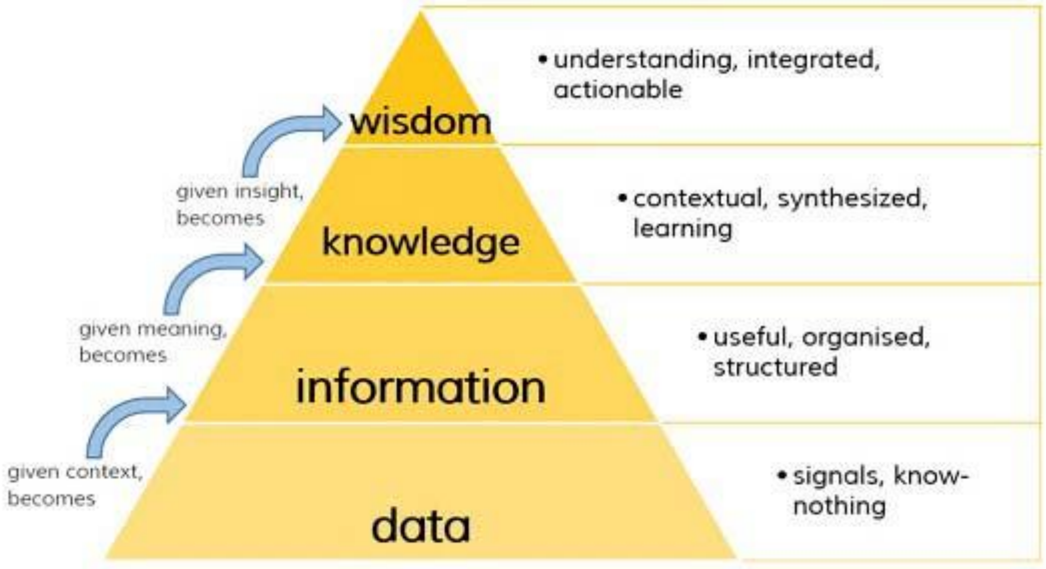
What is Data?

- The first English use of the word "data" is from the 1640s.
- Using the word "data" to mean "transmittable and storable computer information" was first done in 1946.
- The expression "data processing" was first used in 1954.[4]

- The Latin word *data* is the plural of *datum*, **"(thing) given"** or **"to give"**
- **Data = Signals, Symbols…Inherently Meaningless**

DIKW Pyramid

- Data, information, knowledge and wisdom are closely related concepts, but each has its own role in relation to the other, and each term has its own meaning.



- According to a common view, data is collected and analyzed;
- Data only becomes information suitable for making decisions once it has been analyzed in some fashion.

Big Data

Big Data Definitions

- A variety of definitions and perspectives
- Commonly described according to the 3 V`s
  - Volume
  - Velocity
  - Variety
- Additional description related to
  - Veracity
  - Value

Small vs. Big

The ultimate goal for data analysis to get timely insights to support decision making.

Categorizing data into Small and Big help to tackle challenges in analyzing data of each world separately with proper tools.

The line between two categories varies with emerging advanced data processing systems which makes even big data querying much faster and less complex.

Big Data vs Small Data science:

1. populations, not samples;
2. messy, not clean data, and;
3. correlations, not causality

Big Data Concepts

- Consider what makes data ``big``
- Big data isn`t really about size or volume (although it is a factor)
- Big data
  - is challenging to manage
  - can be impactful
  - can be transformational
  - is changing how we live our lives
  - is changing how companies compete
  - should be important to our strategy
  - can be complex
  - by itself does not create value
  - can create value if analyzed and acted on

Big Data Concepts

- Big data
  - is generated by living our personal and corporate lives
  - can be described as ``non-traditional``
  - requires innovations in technology to manage and apply it
  - is largely created by someone other than us
  - can be misleading

Big Data Structure

- Structure of data dictates its `Variety``
  - Structured
    - Data is organized in a tabular format at the ``field`` level
    - Eg. Spreadsheets and data base tables
  - Unstructured
    - The structure is unknown
    - Eg. Images, audio, video,
  - Multi-structured
    - The structure includes a combination of tabular, hierarchical, tagged, other and unknown
    - Eg. Emails and Twitter feeds contain some known fields such as the ``From`` and ``To`` parties combined with unstructured text

Big Data Sources

- Types of Sources
  - Internal
    - Generated by internally managed business activities
    - Generated by staff, contractors and partners known to us
    - Examples
      - Systems for ERP, Operations, CRM, transaction records, production
      - Sensors, text documents, audio, video, images, maps, web logs, etc
  - External
    - Generated by external activities
    - Generated by organizations, people or equipment who are likely unknown
    - Examples
      - Social media, open data, public data, blogs, news feeds, audio, video, images, sensors, IOT devices, smart phone communications, wearable devices, web sites, data markets, location data using GPS, data subscription services

Big Data

- Example data
  - Full motion video
  - Multi/hyper-spectral imagery
  - Cell phone calls
  - Register transactions
  - Lidar/point clouds
  - Email/tweets

- Space/time critical

Big Data

- Query Optimization
  - Traditional data types solved long ago
  - Big problems with extended data types
  - Revert to full table scans

- One solution: massively parallel systems, data partitioning, etc.
  - IBM's Netezza, Oracle's Exadata, Microsoft's SQL Azure, Apache's Hadoop, Teradata,
    among others
- Can a finely tuned query win?

Big Data Impact

- Areas where Big Data makes an impact
  - Enhanced Surveillance and Monitoring
  - Customer Intimacy and Relationships
  - Government and Corporate Transparency
  - Fraud detection
  - Improved risk management
  - Enriched information for predictive models
  - Personal fitness management
  - Consumer purchase choices
  - Micro-market segmentation
  - Smart grid management for energy conservation
  - Improved physical asset reliability
  - Location and route optimization
  - Crowdsourcing of opinions

Big Data

- How to stream for real time event processing
  - Store to disk/post process
  - Analyst with manual inspection
  - Slow

- How to persist/partition and rapidly search

Big Data

- Search criterion controlling storage
  - Based upon predicate filtering
  - Temporal, then spatial, or vice versa
  - Placenames
  - Type of attribute/tags
    - Sensor platform attributes

  - Column-oriented stores

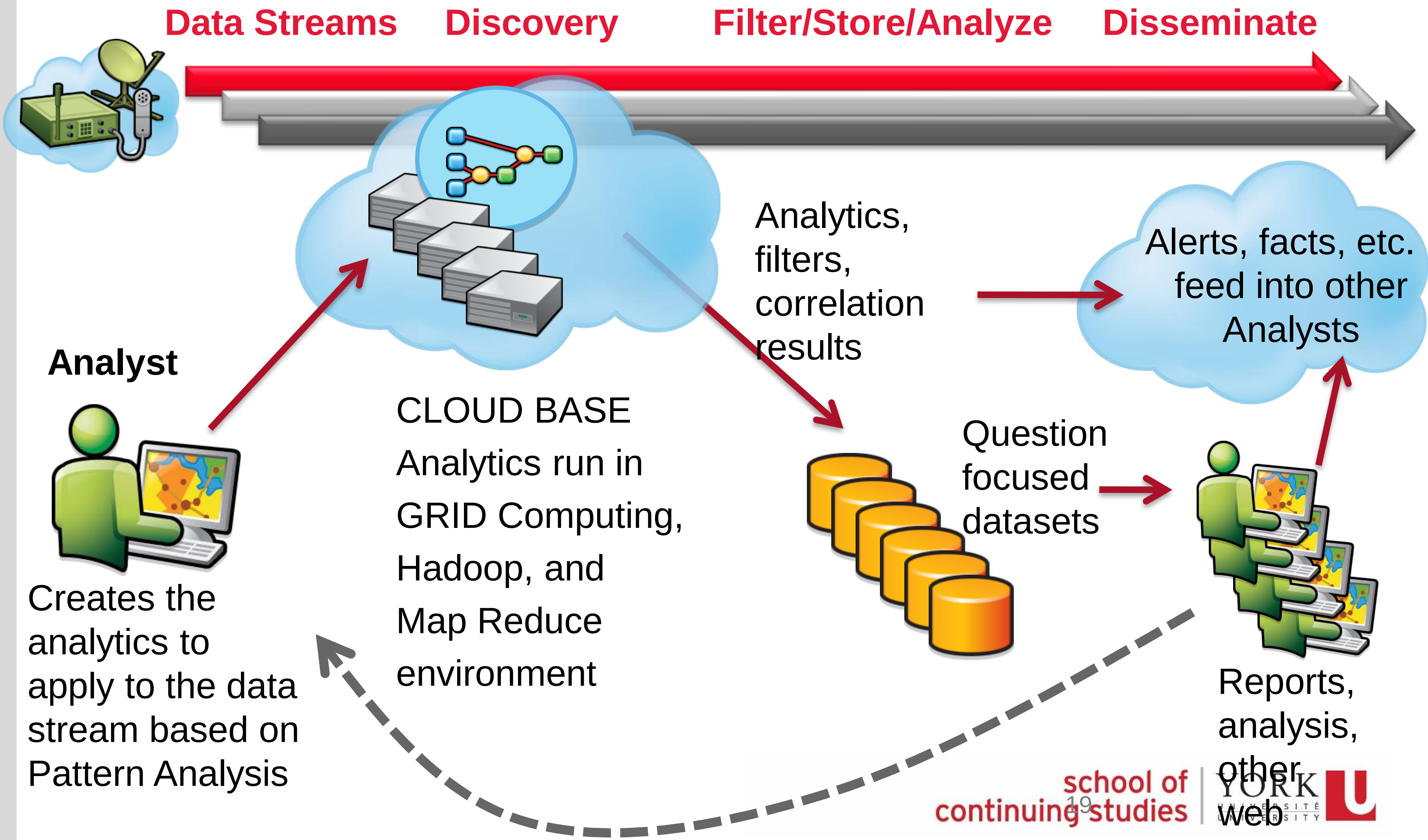
Big Data

- Peeking at data as it flows in

- Identify interesting bits, ignore most
  - When is something near, when does something cross …
  - Query optimization problem

- Existing frameworks
  - Microsoft, Oracle, IBM, etc.

Move is to dynamic data, applying analytics to large volumes, reporting facts as available



Data Streams

Discovery

Filter/Store/Analyze

Disseminate

Analyst

Creates the analytics to apply to the data stream based on Pattern Analysis

CLOUD BASE
Analytics run in
GRID Computing,
Hadoop, and
Map Reduce
environment

Analytics, filters, correlation results

Alerts, facts, etc. feed into other Analysts

Question focused datasets

Reports, analysis, other web pages

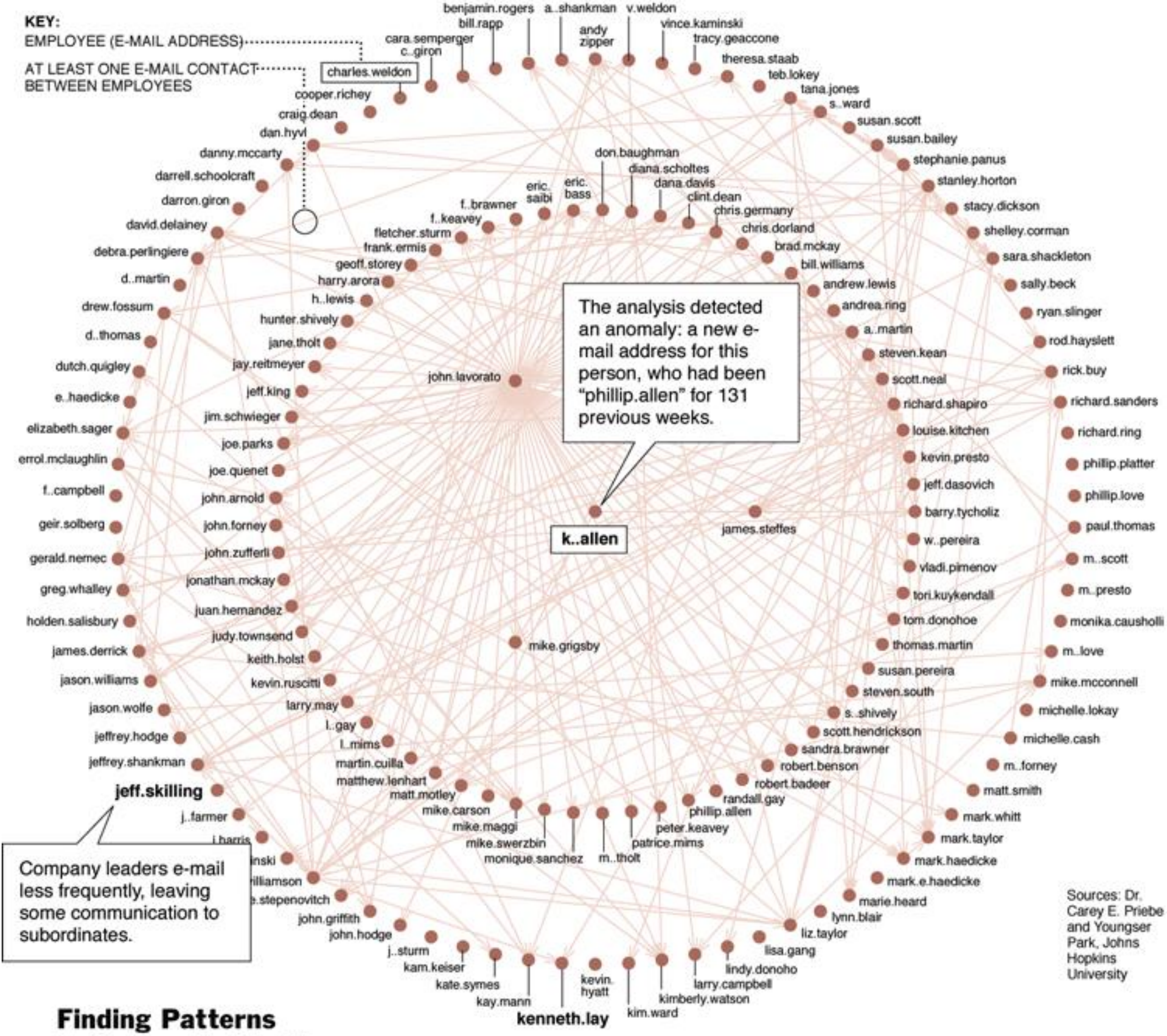
Big Data

- Detecting patterns, connecting thing together
  - Social media type stuff with spatial/temporal
    - Cash register transactions, cell phone calls
    - Pattern of life
  - “Connecting the dots spatially”
- Knowns and unknowns
  - How to assign unknowns to knows
  - How to assign confidences

Temporal Anomalies

- A half-million Enron e-mails from ~150 accounts were sent from 1999 to 2001, a period when Enron executives were manipulating financial data, making false public statements, engaging in insider trading, and the company was coming under scrutiny by regulators
- The graph reveals a map of a week's e-mail patterns in May 2001, when a new name suddenly appeared
    - This week's pattern differed greatly from others, suggesting different conversations were taking place that might interest investigators

Temporal Anomalies



The analysis detected an anomaly: a new e-mail address for this person, who had been "phillip.allen" for 131 previous weeks.

Company leaders e-mail less frequently, leaving some communication to subordinates.

Sources: Dr. Carey E. Priebe and Youngser Park, Johns Hopkins University

Finding Patterns In Corporate Chatter

Computer scientists are analyzing about a half million Enron e-mails. Here is a map of a week's e-mail patterns in May 2001, when a new name suddenly appeared. Scientists found that this week's pattern differed greatly from others, suggesting different conversations were taking place that might interest investigators. Next step: word analysis of these messages.

Big Data

- Spatio-temporal web crawlers
  - Trends and spatial activity
  - Social media
  - Meaningful persistence
    - Fast, geolocate, query

Non-traditional Data

- Lots of non-spatial data
  - CSV/TXT files, Excel spreadsheets, news feeds, social media

- Coarse grained spatial data
  - City level, not down to 10 meters …
  - E.g., Fukushima radiation

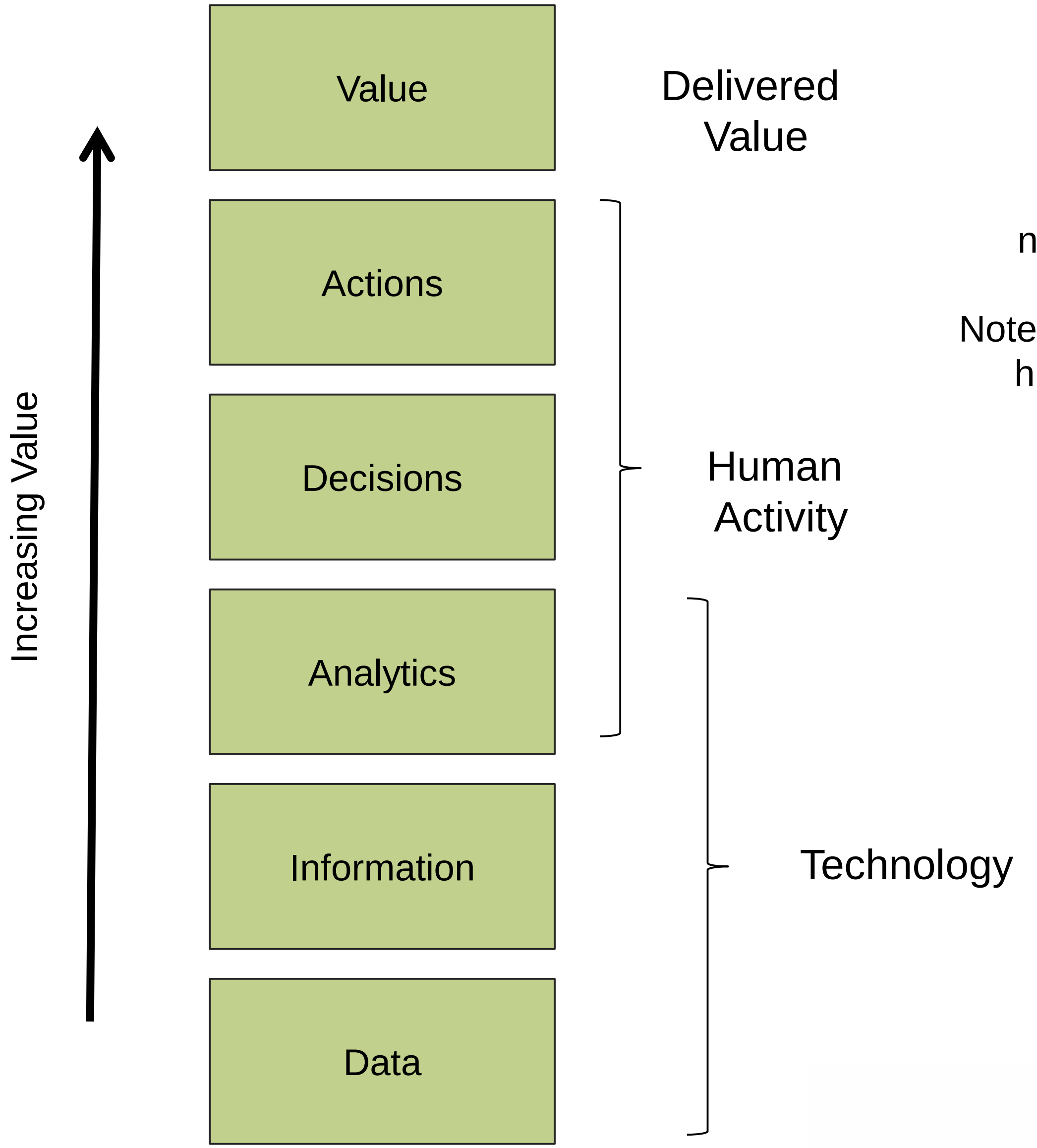
- Geoprocessing and trend analysis/detection

Big Data Opportunities

- It is important to understand all of the components needed to transform data into business value
- A Big Data Opportunity must specify how these levels will contribute to value
- These levels represent a value chain from Data (level 1) to Value (Level 6)
- All of these components must function to deliver value at the top of the chain

  - Level 6- Value
    - Measurable improvement in a business result. This is ``value
  - Level 5 - Action
    - Execution of `the right work or activity to drive the impact
  - Level 4 - Decision
    - Selective choice made to take an action that drives impact
  - Level 3 - Analytics
    - Determining the business questions to ask
    - Generating new information and insights
  - Level 2- Information
    - Answers to questions that will guide decisions
  - Level 1 - Data
    - Building blocks used to create meaningful answers to drive actions

Big Data Value Chain



This model shows the activities needed to transform data into value

Note how technology components enable human components to create value

Examples of Value

- Lower cost
- Higher revenue
- Lower risk
- Cleaner environment
- Safer workplace
- Satisfied customers

Big Data Opportunities

- Discovering Opportunities
  - Determine what business impact is desired and how creates value
  - Understand what actions and decisions will lead to the desired impact
  - Determine what questions need to be answered to guide the decisions
  - Ensure experienced people can frame the questions that can be answered
  - Provide information to answer the questions
  - Information is created by integrating available data and using analytics to generate net new insights
  - Data is gathered and must be in an ``analyzable`` condition
  - Data conditions include security, quality, accessibility, integration and timeliness

Big Data Key Characteristics

- Big Data is characterized by
- Volume
  - Describes the quantity of data in terms of breadth, granularity and history
  - As data volumes grow, traditional technology has difficulty scaling
- Variety
  - Describes multiple formats, structures and time intervals
  - As variety increases, traditional technology has difficulty keeping pace it
- Velocity
  - Describes the trend towards data flowing faster and faster nearing real-time
  - As velocity increases, traditional technology has difficulty keeping pace
- Veracity
  - Describes the level of confidence and trust we place in the data
  - As more data is acquired that was generated externally, this issue increases
- Value
  - Describes the positive impact made by utilizing data
  - Companies have difficulty with this

Data Analytics

Data Analytics Terminology

- Business Intelligence
  - An organization`s capability to intelligently set and accomplish goals
- Analytics
  - The ability to generate meaningful insights to support
- Data Warehousing
  - The ability to organize data into a form that can be analyzed
- Data-Driven
  - Utilize data as evidence to guide decision making and results monitoring
- Big Data Analytics
  - The application of analytics methods to the wide range of available big data sources (internal and external) to generate new and powerful insights.

Data Analytics Recent History

- The 1990`s
  - Early applications of data warehousing to unify tabular, internal data to enable integrated business reporting
  - Analytics was in the form of using integrated facts to drive improved decision making
- The 2000`s
  - Maturing of the internet as a platform for entertainment, commerce and communications
  - Emergence of the smart phone, GPS systems and social media
  - Data started taking on the characteristics of ``big``
  - Analytics was in the form of advanced monitoring and control
- The 2010`s
  - Emergence of the digital economy and society
  - Mass installation of surveillance and sensor systems
  - Maturing of social media and digital communities
  - Analytics broadened into a range of new capabilities including predictions

Data Analytics Value Concepts

- Analytics uses available data to generate different categories of insights.
- Analytics is not a single step in a value chain
- Analytics is a series of activities that build on each other as new insights are generated and applied to subsequent steps
- New value is added to an original data set as broader sets of insights are generated from it.
- If data is a basic asset in a company, then analytics is the engine that drives value from it.

Data Analytics Value Examples

- Value delivered by Analytics is described in a Value Chain model discussed earlier
- The 5 major categories of value that can be enabled by data analytics are
  - Effectiveness
    - ensures that the “right” work is done to accomplish goals
  - Efficiency
    - ensures that work is done “well” and fewer resources are consumed to produce a given level of outputs
  - Agility
    - enables organizations to respond to changing market conditions
  - Quality
    - enables consistent results with fewer defects and less re-work
  - Innovation
    - enables new disruptive models to be tested and implemented

Data Analytics Purpose

- Data Analytics Generates New Information and Insight
  - New information provides:
    - answers questions that could not otherwise be answered
    - facts about time periods such as the past, present of future
    - descriptions of scenarios that never actually took place
  - Insight provides knowledge:
    - that can generally be applied to our domain
    - that helps us understand how our world actually works
    - that can be applied in a competitive sense
    - that can help us improve organizational performance

Data Analytics Structure

- Data Analytics includes
  - Framed business problems including assumptions
  - Defined analytic problems to address the business problem
  - Training and testing data sets
  - Algorithms to create models from the training data
  - Evaluation of models based on test data sets
  - Implemented and deployed models
  - Decision variables, outcome variables, parameters and constraints
  - Visualizations used to interpret model results
  - Human expertise with mental processes to interpret and apply the results
  - Human expertise with communication skills to tell stories and share the results

Data Analytics Capabilities

- Major Categories of Capabilities are Based on Functionality
  - Discovery Analytics
    - Statistical and visualization methods for finding useful patterns and relationships
  - Descriptive Analytics
    - Measurement methods used to quantify and describe a domain using statistical techniques
  - Diagnostic Analytics
    - Abnormal condition detection and root cause analysis
  - Predictive Analytics
    - Models used to estimate future conditions and the probability of events
  - Prescriptive Analytics
    - Models that guide decision makers towards a feasible scenario using simulation or an optimum scenario using optimization

Business Questions

Business Question Concepts

- Business questions
  - provide a foundation to define the requirements of an analytics solution
  - contain two components, a fact and multiple qualifiers
  - analytic questions usually have a quantitative fact
  - operational questions may have qualitative or quantitative facts
  - qualifiers are qualitative and become dimensions
- Example
  - Question
    - How many customers will respond to our marketing programs over the next 3 years by month, by product and by campaign
    - Fact
      - count of customers
    - Qualifiers
      - month, product and campaign

Business Question Types

- Operational Questions
  - Focus on a specific event, condition or transaction
  - Examples:
    - Which customer requires a follow up
    - When the package arrive
    - What is the delivery status of the last shipment
- Analytical Questions
  - Focus on counts, totals and aggregates grouped by criteria
  - Examples:
    - How is our customer base broken down by satisfaction category
    - How many products are in the low quality category by shift
    - How many accidents do we expect at rush hour in the financial district during the upcoming holiday season

Business Questions and Analytics

Categories of Business Questions with Examples

- Who
  - Segment employees or customers based on descriptive analytics
- What
  - Identify an abnormal condition based on diagnostic analytics
- Where
  - Group intersections by accident likelihood based on predictive analytics
- When
  - Forecast future demand by month based on predictive analytics
- Why
  - Determine root cause of quality defects based on diagnostic analytics
- How
  - Prescribe how to respond to a delay based on prescriptive analytics

Big Data and Analytics

Business Performance Concepts

- Performance describes how closely a controlled variable tracks to a target variable established by a goal.
- High performance exists when the controlled variable is within an acceptable tolerance of a target
- Low performance exists when the margin between the controlled variable and a target is significantly large
- The management of performance (performance management, business performance management and corporate performance management) adjusts the decision (input) variables to maintain an acceptable level of performance
- Managers must determine what decisions and actions will result in desired levels of performance
- Decision variables are manipulated to drive output variables to an acceptable level based on a target

Business Performance Measurements

- Measurements quantify the position, level or condition of input and output variables.
- Definitions
  - Measures
    - Quantitative data elements recorded at a point in time
    - Describe properties or attributes of things we need to manage
  - Metrics
    - Information derived from measures to inform decision making
    - Business metrics are outcome variables with targets
    - Metrics are input variables that do not have targets
    - Indicators provide evidence of a condition similar to metrics
    - Performance Indicators provide evidence of performance levels
    - Key Performance Indicators (KPI`s) are a few strategic indicators

Business Problem Framing

- Business problems initially are commonly vague and ambiguous
- Problem Framing provides structure to define:
  - The heart of the problem
  - A short description of the problem in business terms
  - Why the problem should be solved
  - Outcome variable(s) including target levels and deadlines
  - Input decision variable(s) that can be manipulated
  - Input parameters based on assumptions
  - Historical or timing implications of the problem

Analytical Models

Analytical Models Definitions

- Models are representations of reality

- They are based on assumptions and simplifications

- George Box (a famous statistician) quote:

` ``*All models are wrong. Some are more useful than others*``

- Models transform input variables into output variables

Analytical Models Structure

- Models represent a simplified view of reality

- Input variables are processed by an algorithm or a set of rules to produce output variables

- A simple model may be built using an algorithm that models the problem

- A complex model may require many sub-models to work together, each one addressing a piece of the larger problem

Analytical Models Categories

- Empirical
  - model is created based strictly on observed data
- Mechanistic
  - model is created based on theoretical rules and formula
- Deterministic
  - model does not consider randomness or uncertainty
- Stochastic
  - model considers randomness and provides a range of output values
- Continuous
  - model works with fractions and real numbers
- Discrete
  - model works with integers and whole numbers

Analytical Models Purpose

- Models generate new information and insights that enable the class of business questions that were created by framing a business problem

- The focus of information and insights is to ensure that they are useful and relevant relate by answering key business questions that are framed as part of the analytics effort

Analytical Models Limitations

- Models must be calibrated over time to ensure that the error produced by the model is recognized and understood
- Models based on empirical data are limited to the range of data within the training data set.
- Models may not generalize beyond the range of observed data.

Statistical Methods

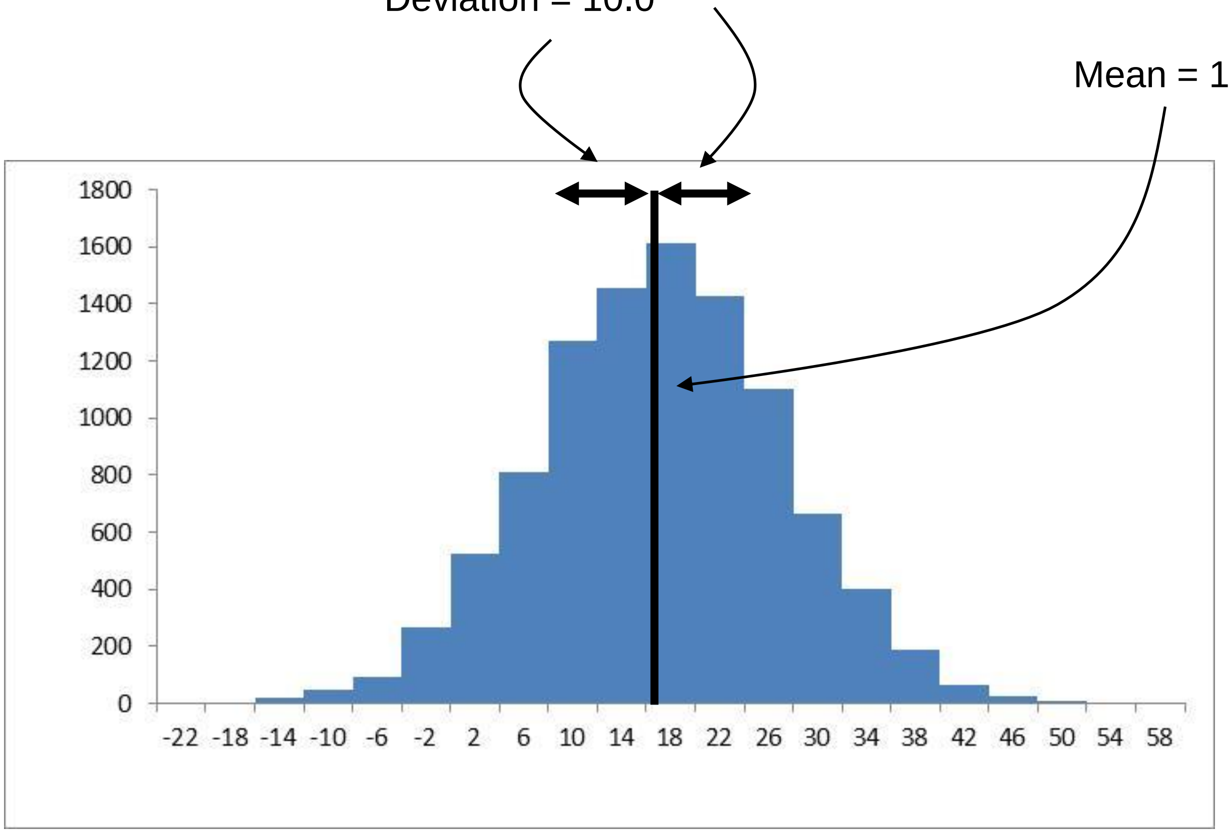
Statistical Methods Overview

- Techniques for understanding and summarizing data values for a single variable
- Descriptive statistics
  - Provides methods for describing data values to understand similarities and differences about attributes within the sample
  - Provides measures of central tendency and dispersion
  - How data clusters around a defined point and how it spreads away from that point provides us with the shape of the data for a given variable.
- Inferential statistics
  - Provides methods to generalize our belief about the greater population represented by the sample that was gathered
  - Allows us to use a generalized model in a wide range of analytics applications
  - Helps us design experiments to gather samples and generalize the results to the broader population
- Techniques for describing the relationships between two variables
  - Correlation

Statistical Methods to Understand Data Shape

- Variables from a data sample possess a set of properties that define its shape
- Values cluster around the central tendency
- Central tendency is measured by the following descriptive statistics
  - Mean is the arithmetic average
  - Median is the middle value
  - Mode is the most common value
- The amount of spread is called dispersion or variation
- Dispersion is measured by the following descriptive statistics
  - Variance
  - Standard Deviation is the square root of variance
- Methods of calculation and application will be discussed later in the course
- The measures of central tendency and spread provide a signature about the broader population described by the variable.

Statistical Methods to Understand Data Shape



Standard Deviation = 10.0

Mean = 15.1

Frequency Histogram for Variable X

Python

Python Introduction

- Python is a popular programming language used in data analytics
- It contains a rich set of libraries to do mathematical, statistical and visualization functions
- It is available free of cost as an open source product.
- An interactive form-base interface called a Jupyter Notebook provides a structured method for executing Python functions and commands

Python Applied to This Course

- The intent of using Python in this course is to provide a software platform for you to view demonstrations and carry out practical exercises in the following areas.
  - Calculate statistics
  - Visualize data
  - Execute SQL queries
- Your exercises will be carried out in a structured format using the Jupyter Notebook interface for Python.

Lesson Review

Lesson Review

- Consider the following questions based material describe in this lesson and be prepared to discuss with the instructor in class next week.
  - Describe the key components to transform data into business value
  - Describe how Big Data is different from traditional Small Data
  - Identify some of the historical events since the year 2000 that have catalysts for the Big Data era
  - Differentiate data from information
  - Explain the major purposes of analytics
  - Describe how business problems need to be translated and framed to take advantage of analytics including the essential components of a framed problem
  - Describe how data analytics relates to business performance management
  - Describe the main role of statistics in describing data
  - Describe the main set of analytics capabilities that are in common use

Lesson Summary

Lesson 1 Summary

During this lesson, you learned to:

- Describe the curriculum and the course learning objectives and grading scheme
- Define big data in terms of structure, sources, impact, business opportunities and key characteristics
- Define data analytics based on its recent history, terminology, purpose, structure and capabilities
- Describe key areas data analytics functionality that helps answer different types of business questions
- Describe how combining Big Data with Analytics can drive improved business performance through proper business problem framing
- Describe the structure, purpose and limitations of analytical models
- Describe the roles that statistical methods play in creating useful information from raw data