Week 1: Introduction to data analytics

Data are everywhere and are produced at an astonishing rate. There are endless potentials of this data wealth. Nevertheless, the challenge is how to dive into the ocean of data and make the best of it. Data analysis processes and tools are the essential navigation steps to learn and understand about the data. Therefore, there are many skills needed in order to comprehend and employ the data effectively.

Analytics is '*the scientific process of transforming data into insight for making better decisions*' [INFORMS]

Data analytics are not new concepts. It is a multi-disciplinary field. It extensively utilizes techniques derived from statistics, operations research, machine learning, computer programming, and data visualization. It is not a single step but it is the whole methodology. Analytics are the integration of procedures, technology, and a team of people to create a valuable knowledge from the data. The insights are used to drive decision-making, raise productivity and gain competitive advantage.

A few examples of successful analytics (https://www.informs.org/Sites/Getting-Started-With-Analytics/Analytics-Success-Stories)

- McKesson's Supply Chain Scenario Modeler Cures Pharmaceutical Distribution Network
  McKesson's, the world's largest healthcare services company, leverages analytics in its supply chain model with the help of IBM. The new model changes its distribution network, supply flows, inventory policies, and other processes. The impact of this new model development has reduced McKesson's working capital by over $1 billion.

- Optimizing Chevron's Refineries
  Chevron developed an in-house software tool called Petro. The analysts run Petro for crude oil selection, product optimization, refinery processing, and optimization. Not only internal applications, but Petro also has been used to identify formulations that have less impact on supply and cost but still meet with the environmental requirements. Over the past 30 years, it is estimated that Petro has generate approximately $10 billion in value.

- Kroger Prescribes Analytics to Optimize Pharmacy Inventory Management
  Kroger with the collaboration with faculty from Wright State University has implemented a new drug inventory optimization system for Kroger pharmacies. In summary, the goals are to reduce costs and increase profit in different areas such as supplying enough drugs to meet with the demands (patients have access to medications whenever they need them), reducing out-of-stock prescriptions, and decreasing inventory cost. As a result of this development, hundreds of millions dollars have been saved.

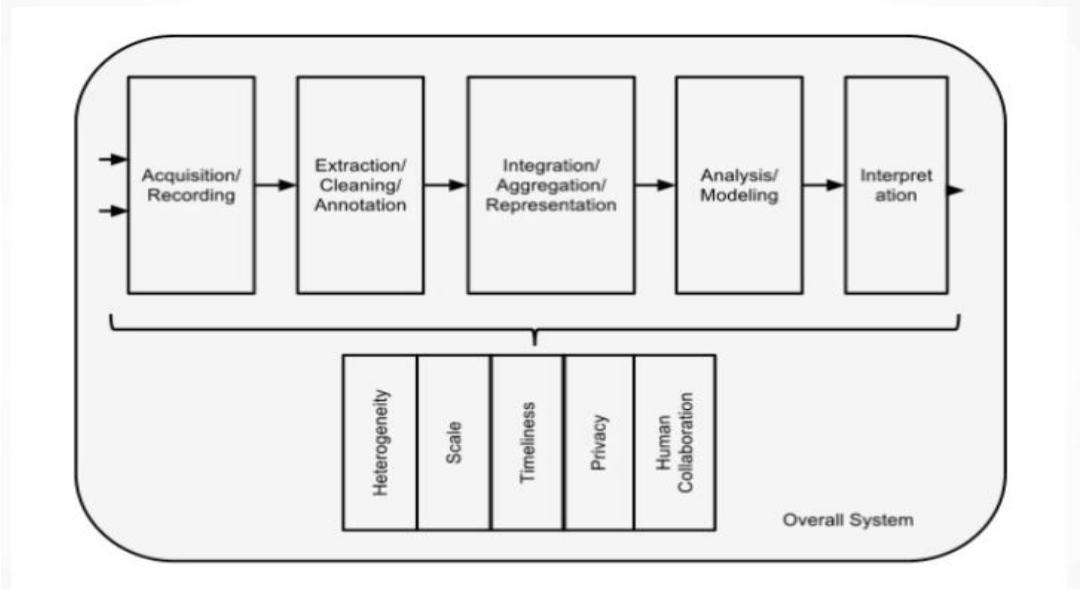
Challenges

Previously, massive sets of data have been limited to scientific communities.
The intensive growth of Internet and social media generate enormous volumes of data. This tremendous size of data is known as 'Big Data'. Enterprises become interested in leveraging the data they have obtained. The aims, generally, are to identify potential customers, predict trends, product/service recommendations, and fraud detection, including the integration of various data analytics to solve the overall organization problems. The big corporations such as Google, Yahoo, Amazon, Netflix can gain insights and make greater revenues from the data they collect from users.

However, the central focus of this challenge is finding the methodology that can handle not only the size of data but also it can process the data with speed.

Another challenge is the structure of data. Enterprises produce data in diverse forms, both structured and unstructured formats. Unstructured data such as emails, documents (text, PDF), data streams, and network graphs, geospatial locations are complicated to work with. However, the data is so valuable with so much information in it. To make the most out of the data, various data sources should be integrated. Not all businesses, however, have the capability to tackle the unstructured data. Inevitably, Industries are forced to come up with cutting-edge technology, raising analytics to novel levels. Other challenges still remain such as cloud computing, security, etc.

The following figure depicts the major steps in big data analysis.



(Big) Data analysis Pipeline
Source: http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf

Risks in Analytics include privacy (e.g. online social media networking such as Facebook accommodates a lot of information, which are revealed to many others), security, making

decisions on incomplete or inaccurate data, using only the data that supports intuitive decisions, drawing the wrong conclusions from the data, etc.

Analytics Tools are comprised of:

- Data mining
- Statistical Analysis
- Predictive Analysis
- Correlation
- Regression
- Forecasting
- Process Modeling
- Operations Research
- Optimizations
- Simulations
- Machine learning
- Visualizations

Techniques in data analysis such as statistics, machine learning, visualizations, and operation research will further discussed in later chapters.

Related Topics:

Business Analytics (BA)

BA applies the statistical and quantitative tools to gain insights of business performance. Many people believe that analytics date back to the early 19th century when Henry Ford attempted to measure the time of each component in the assembly line. Nevertheless, the explosion of business analytics was in the late 1960s with the introduction of computers for decision support systems (DSS) and planning. Business intelligence (BI) is believed to progress from the DSS and became the center of attention in the late 1980s. According to Thomas Davenport (a professor of information system and a research director at International Institute for Analytic), business analytics is the subset of Business Intelligence, while BA emphasizes on statistics, prediction, and optimization, the BI covers broader spectrum including querying, reporting, OLAP, alerts tools, and BA.

Business Analytics consist of three phases: descriptive, predictive and prescriptive analytics.

- *Descriptive analytics* is the first stage of BA, which concerns on answering the questions regarding to what happened and why did it happen. The past or historical data such as scorecards, reports (e.g. sales, marketing, finance), are analyzed to understand past performance (either success or failure).

- *Predictive analytics* applies various algorithms and techniques to the data for making predictions. This stage tries to conclude answers for questions of what or when will it happen.

- *Prescriptive analytics* extends to cover the question such as why it will happen. In this stage, the alternative decisions and their implications are suggested. It is a

continuous stage, where new data can re-predict and re-prescribe to improve the prediction accuracy and better decision options.

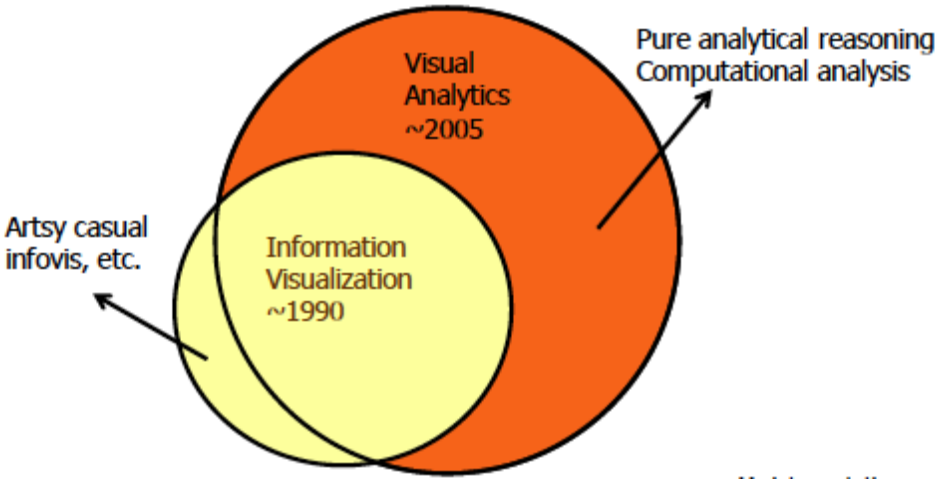
Examples of Applications include differentiating banking customers based on credit risk, offering products to match customer characteristics, customer loyalty programs for gambling businesses, inventory optimizations, etc.

Business analytics have been applied in various domains including financial, marketing, pricing, retail sales, risk& credit, supply chain, and transportation.

Visual Analytics (VA) is defined as "the science of analytical reasoning facilitated by interactive visual interfaces"[Stasko, J]. It is not the area but the umbrella idea.
Visual analytics "combined automated analysis techniques with interactive visualizations for an effective understanding, reasoning, and decision making on the basis of very large and complex data sets" (Keim, et al. 2008)

In fact, information visualization tools do not always include data analysis algorithms.
A paper authored by Shneiderman proposed the integration of computational analysis approaches (e.g. data mining) with information visualization (e.g. Infovis). Therefore, the contribution of each approach results in a more powerful approach for effective understanding, user learning, and decision making especially for very large and complex data. Additionally, the numbers of situations where better analysis for large data sets are rising such as law enforcement, security, and business intelligence also help to promote the idea of visual analytics. The main components for visual analytic consist of interactive visualization, analytical reasoning, and computational analysis.

The following figure shows the overlapped between visual analytics and information visualization. The information visualization principles will be further discussed in a later chapter.



Source: Visual Analytics by Stasko, Georgia Tech.

Visual analytics have been employed in:
http://www.cc.gatech.edu/~stasko/7450/Notes/visualanalytics1.pdf

- Scientific Research
- Regulatory and Legal Communities
- Intelligence Analysis
- DOE and DOD
- Market Assessments
- Capability Analysis – Resumes
- Medical and Pharmaceutical Communities
- National Security and Law Enforcement
- Information Assurance, Web Analytics
- Technology Scanning, Asset and Intellectual Property Management

Reference

Analytics (Dec, 2014). Wikipedia. Retrieved from: http://en.wikipedia.org/wiki/Analytics

Business Analytics (Dec, 2014). Wikipedia. Retrieved from: http://en.wikipedia.org/wiki/Business_analytics

Defining the big data architecture framework (2013). University of Amsterdam. Retrieved from: http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf

D. Keim, G. Andrienko, J.-D. Fekete, C. Gorg, J. Kohlhammer, and G. Melancon (2008), *Visual Analytics: Definition, Process, and Challenges*, in Information Visualization: Human-Centered Issues and Perspectives, (Editors: A. Kerren, J. Stasko, J.D. Fekete, C. North), Springer, pp. 1-18.

Ferire, J. (2013). *Massive data analysis: course overview*. Retrieved from: http://vgc.poly.edu/~juliana/courses/cs9223/Lectures/intro.pdf

Leek, J. (2013) *Data Analysis*. John Hopkins. Coursera. Retrieved from: https://www.coursera.org/course/dataanalysis

Naone, E. (2011) "*The New Big Data*". Technology Review, MIT. Retrieved from: http://www.technologyreview.com/news/425090/the-new-big-data/page/2/

Stasko, J. *Visual Analytics*. Georgia Technological University. Retrieved from: http://www.cc.gatech.edu/~stasko/7450/Notes/visualanalytics1.pdf http://www.cc.gatech.edu/~stasko/7450/syllabus.html

What is Analytics? (Dec, 2014) INFORMS- Institute of operations research and the management sciences. Retrieved from: https://www.informs.org/About-INFORMS/What-is-Analytics