Today and Next Week…

Today

- Lecture - Big Data Analysis, Technology Concepts, and Techniques
- Go over Assignment 1 - http://bsharve.github.io
- Python/Portfolio/Github – Installation and Setup

Next Week

- Lecture - Exploratory Data Analysis (EDA) and Statistical Inference

Upcoming Student Research Proposal

- Proposal is Due: 9/28
  - email to: bsharve@email.gwu.edu
- 10/5 we will:
  - Assignment 1 Due and give out Assignment 2 Instructions

Lecture: 4

Big Data Analysis, Technology Concepts, and Techniques

Benjamin Harvey
Department of EMSE and CS
Data Analytics Program
The George Washington University
E-mail: bsharve@email.gw.edu
Web: https://bsharvey.github.io

Agenda 9/21 - Big Data Analysis, Technology Concepts, and Techniques

- Analysis Concepts: Models, EDA, CDA, Data Product, Statistics, Machine Learning, Data Munging,
- A/B Testing
- Correlation and Regression
- Heat Maps
- Time Series Analysis
- Network Analysis
- Spatial Data Analysis
- Classification and Clustering
- Outlier Detection Filtering (including collaborative filtering & content-based filtering)
- Natural Language Processing
- Sentiment Analysis
- Text Analytics

Agenda – Enterprise Technologies and Big Data Business Intelligence

- Research and Development Processing
  - Batch Processing
  - Stream Processing
- Business Intelligence Processing
  - Online Analytical Processing (OLAP)
  - Online Transaction processing (OLTP)
- Extract Transform Load (ETL) Data
- Warehouses
- Data Marts
- Traditional Business Intelligence (BI)
- Big Data Business Intelligence (BI)
  - Understand the business side is key to success and this is what bringing in an Industry/Gov't prof. allows
  - **Analytic Lifecycle**: Business Case Evaluation (BCE) - EACOE
    - Organization and Analysis

Big Data Analysis Techniques and Technology

- Statistical analysis techniques are commonly preferred for exploratory data analysis (EDA), after which computational techniques that leverage the insight gleaned from the statistical study of a dataset can be applied.
- In the long run, an organization will operate its Big Data analysis engine at two speeds:
  - processing **streaming data** as it arrives and
  - performing **batch analysis** of this data as it accumulates to look for patterns and trends.
- The shift from batch to real-time presents other challenges as real-time techniques need to leverage computationally-efficient algorithms.

Big Data Analysis Techniques

Quantitative Analysis

- Quantitative analysis is a data analysis technique that focuses on quantifying the patterns and correlations found in the data

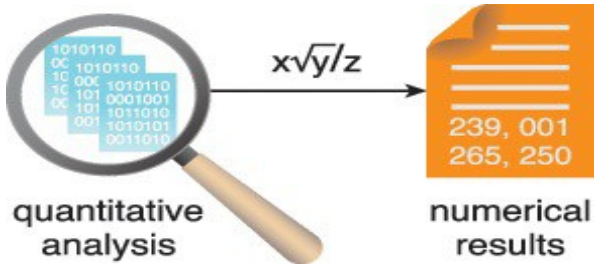


Figure 8.2 The output of quantitative analysis is numerical in nature.

Qualitative Analysis

- Qualitative analysis is a data analysis technique that focuses on describing various data qualities using words

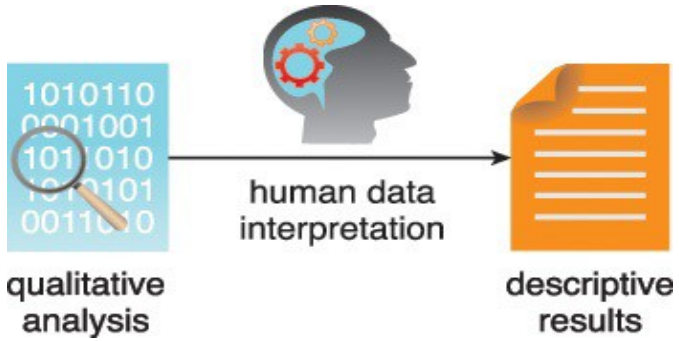
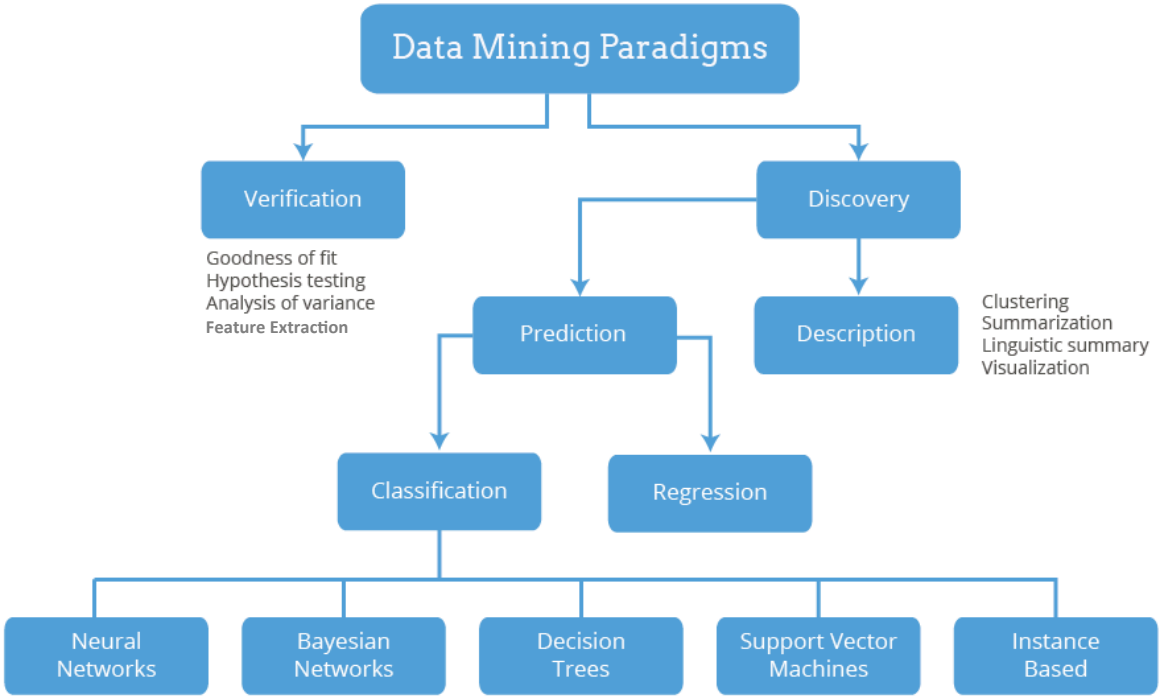


Figure 8.3 Qualitative results are descriptive in nature and not generalizable to the entire dataset

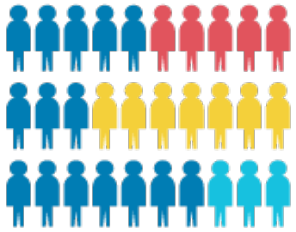
Big Data Analysis Techniques

- Data Mining
  - Or data discovery, is a specialized form of data analysis that targets large datasets.
  - Techniques include automated, software-based technique that sift through massive datasets to identify patterns and trends.

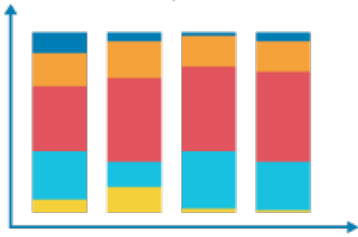


Data Mining Process

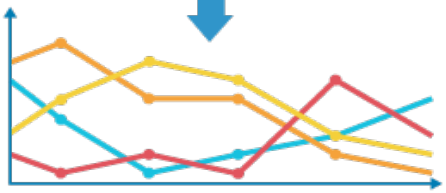
Data Mining Service is an automated process to disclose hidden information in big quantities



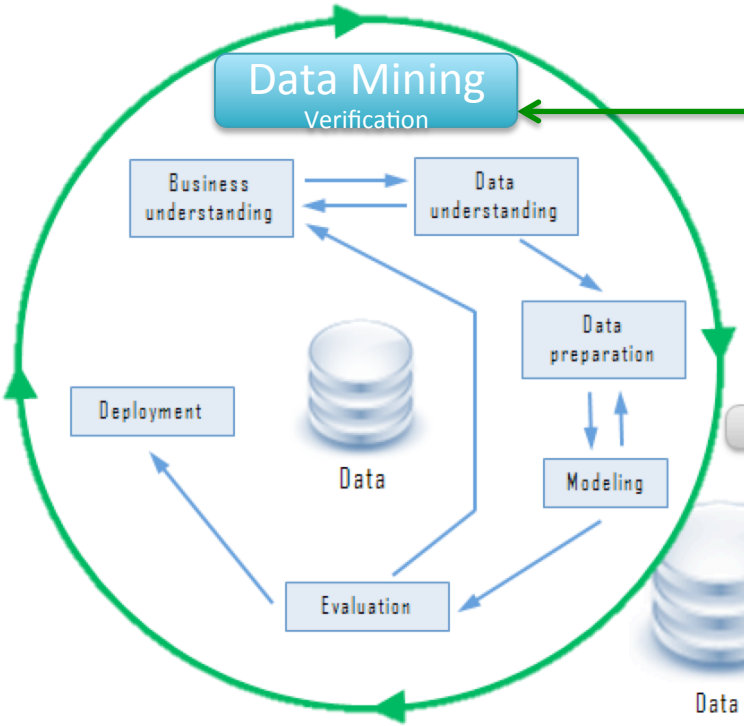
Offer data mining services to our clients all over the world

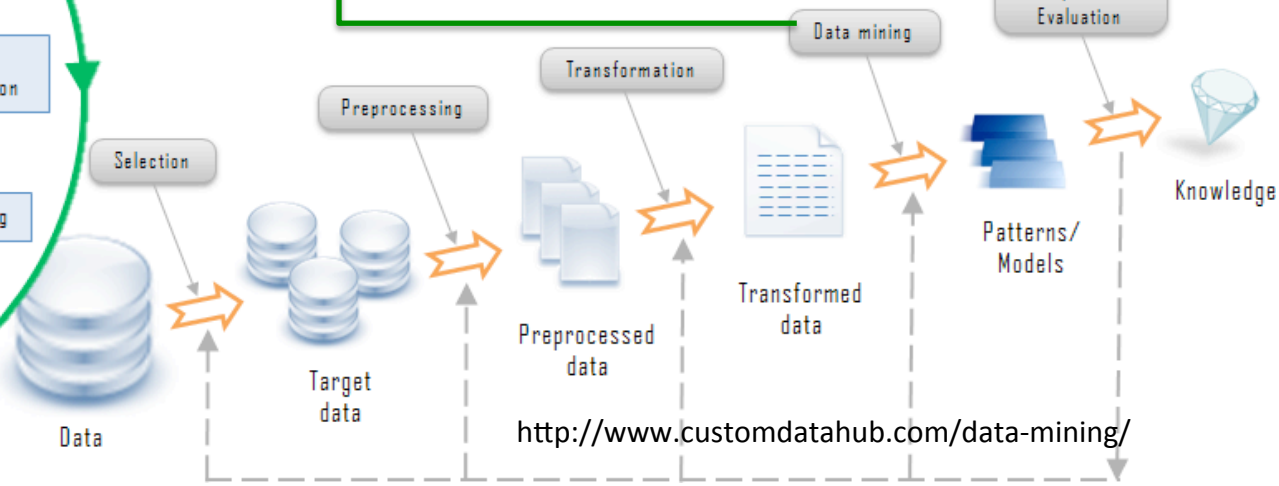


Data mining isn't about simple statistics or moot number



Establishes connections that are crucial in the form of advanced mathematical algorithms





http://www.customdatahub.com/data-mining/

Big Data Analysis Techniques

- Statistical Analysis
  - A/B Testing
    - A/B testing, also known as split or bucket testing, compares two versions of an element to determine which version is superior based on a pre-defined metric.
  - Covariance Correlation
    - Covariance and Correlation is an analysis technique used to determine whether two variables are related to each other.
      - Correlation assumes that both variables are independent.
  - Regression
    - The analysis technique of regression explores how a dependent variable is related to an independent variable within a dataset.
      - applicable to variables that have previously been identified as dependent and independent variables and implies that there is a degree of causation between the variables.

A/B Testing

Figure 8.5 Two different email versions are sent out simultaneously as part of a marketing campaign to see which version brings in more prospective customers.

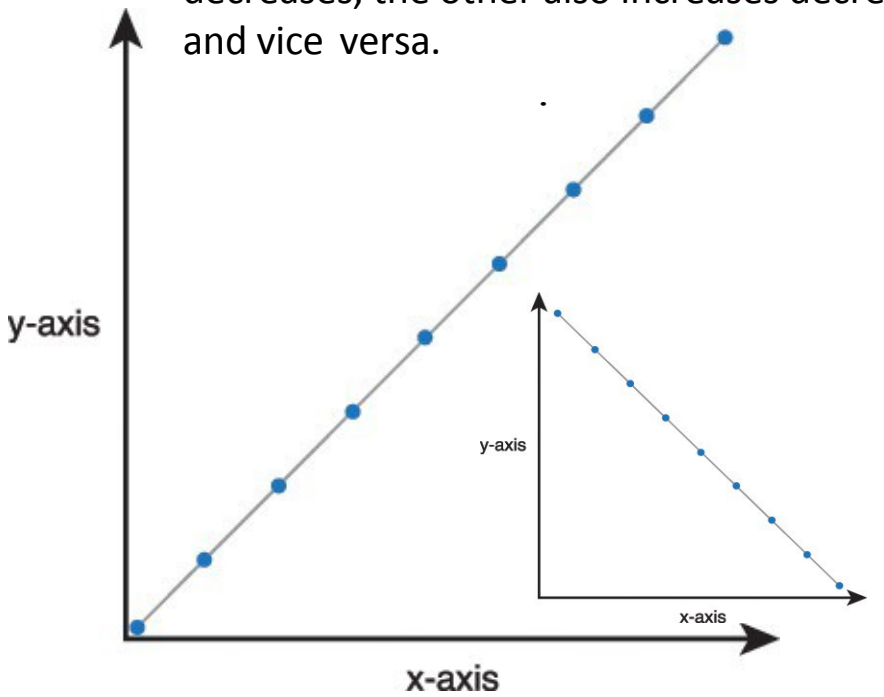
Sample questions can include:

- *Is the new version of a drug better than the old one?*
- *Do customers respond better to advertisements delivered by email or postal mail?*
- *Is the newly designed homepage of the Web site generating more user traffic?*

Correlation

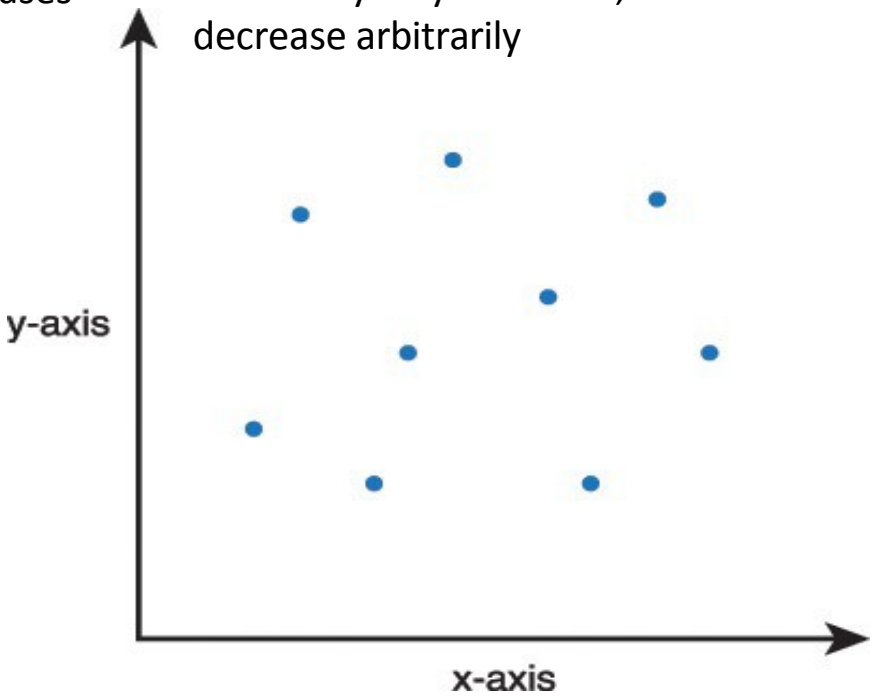
- Correlation is expressed as a decimal number between –1 to +1, which is known as the correlation coefficient. The degree of relationship changes from being strong to weak when moving from –1 to 0 or +1 to 0.
- Correlation assumes that both variables are independent.

Figure 8.6 When one variable increases or decreases, the other also increases decreases and vice versa.



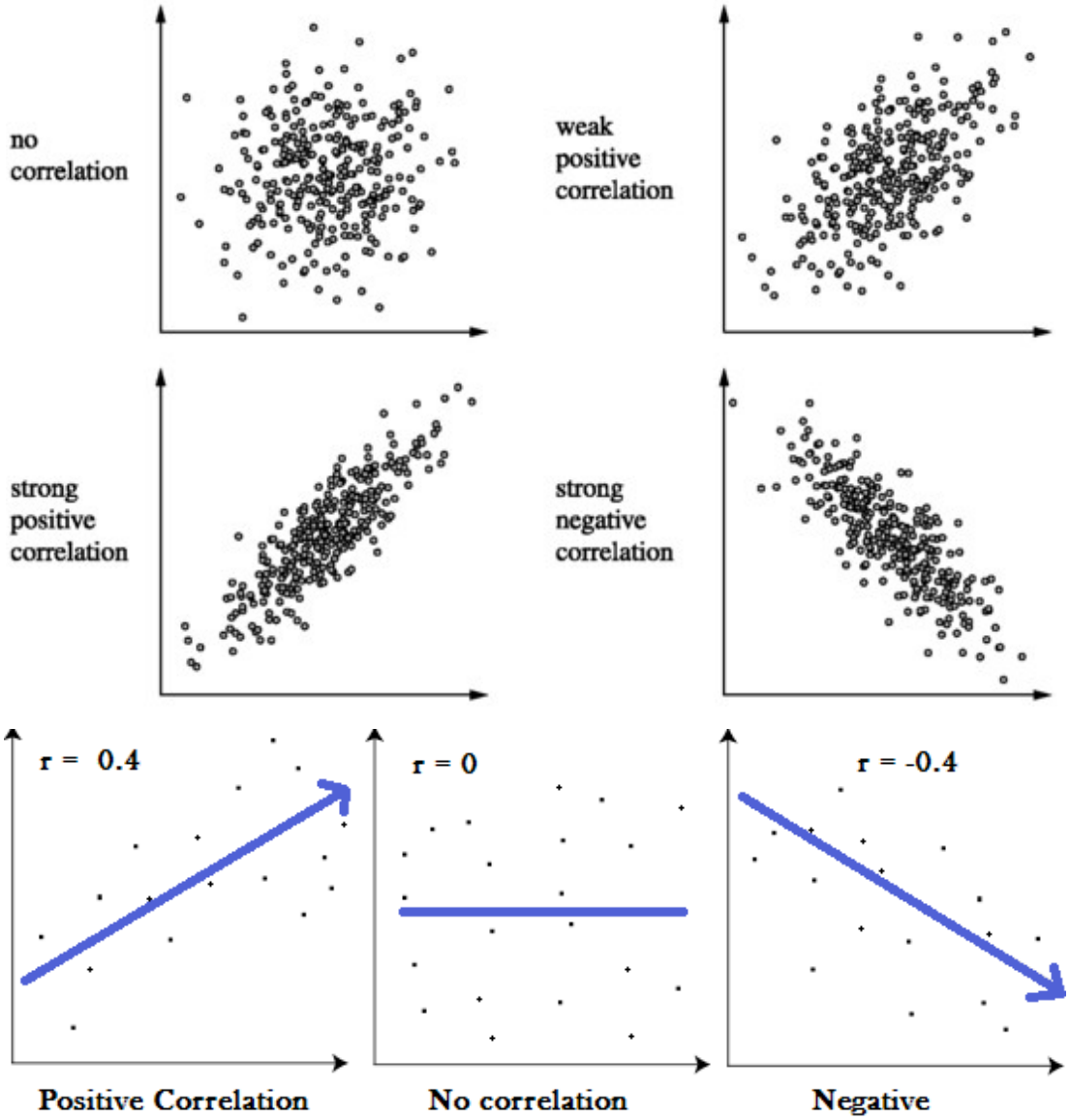
Shows a correlation of +1, which suggests that there is a strong positive relationship between the two variables.

Figure 8.7 When one variable increases, the other may stay the same, or increase or decrease arbitrarily



A slope of –1 suggests that there is a strong negative relationship between the two variables.

Correlation



strong to weak when moving from –1 to 0 or +1 to 0

Regression

Sample questions can include:

- *What will be the temperature of a city that is 250 miles away from the sea?*
- *What will be the grades of a student studying at a high school based on their primary school grades?*
- *What are the chances that a person will be obese based on the amount of their food intake?*

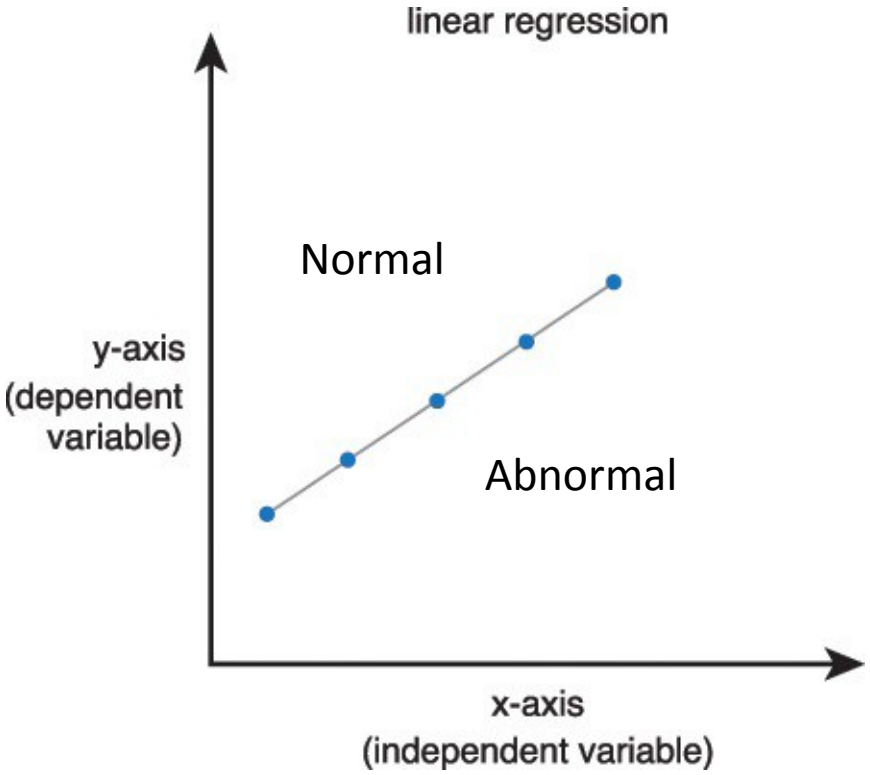


Figure 8.9 Linear regression

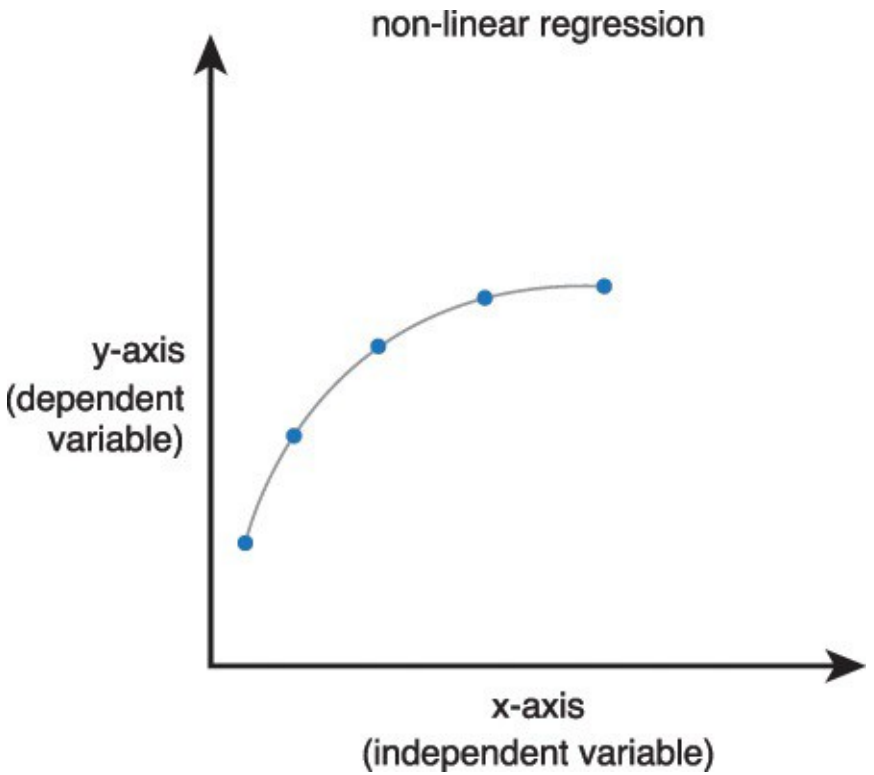
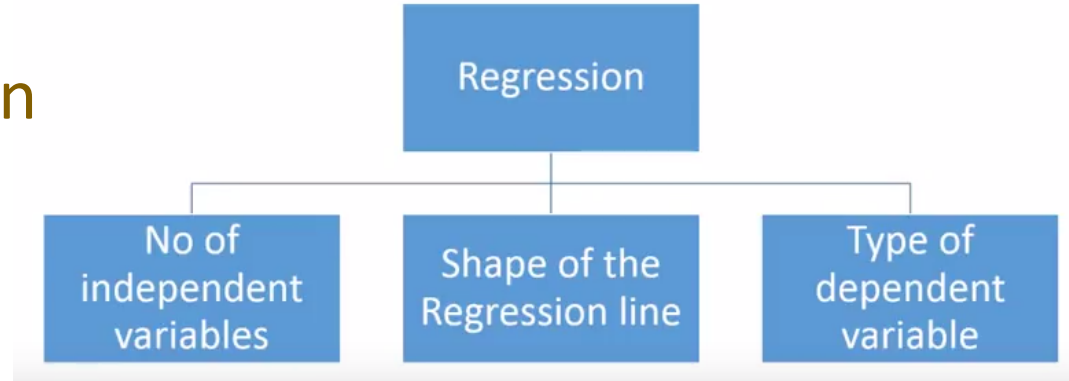


Figure 8.10 Non-linear regression

Regression

- 7 types of Regression that you should know
  - Linear Regression
  - Logistic Regression
  - Polynomial Regression
  - Stepwise Regression
  - Ridge Regression
  - Lasso Regression
  - ElasticNet Regression

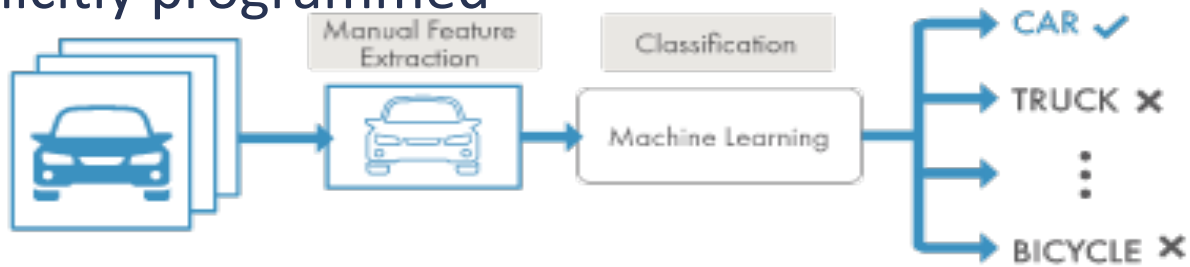


Select regression type based upon

Big Data Analysis Techniques

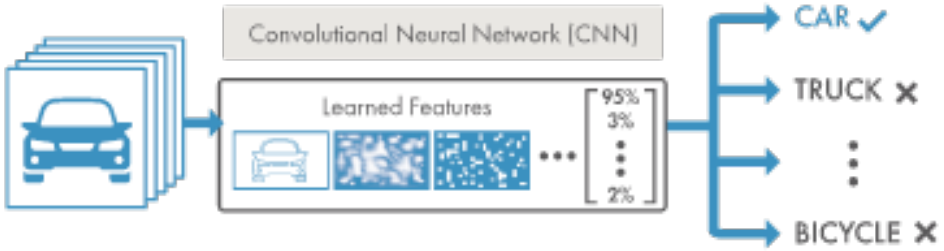
Machine Learning

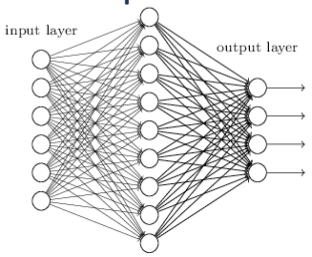
- application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed

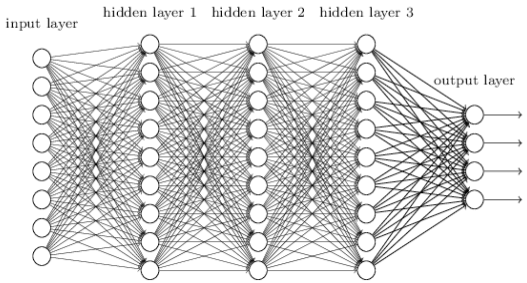


Deep Learning

- Machine Learning technique that use a cascade of many layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input.
- The algorithms may be supervised or unsupervised and applications include pattern analysis (unsupervised) and classification (supervised). are based on the (unsupervised) learning of multiple levels of features or representations of the data. Higher level features are derived from lower level features to form a hierarchical representation.







Machine Learning Techniques

Classification

- Classification is a supervised learning technique by which data is classified into relevant, previously learned categories. It consists of two steps:
  - The system is fed **training data** that is already categorized or labeled, so that it can develop an understanding of the different categories.
  - The system is fed unknown but similar data for classification and based on the understanding it developed from the training data, **build a model**, **then algorithm will classify the unlabeled data**

Clustering

- Clustering is an unsupervised learning technique by which data is divided into different groups so that the data in each group has similar properties.

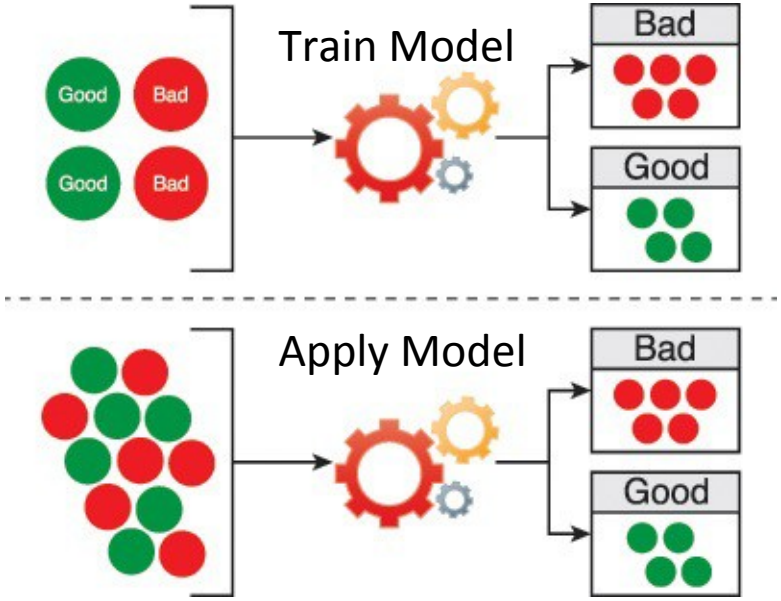
Outlier Detection

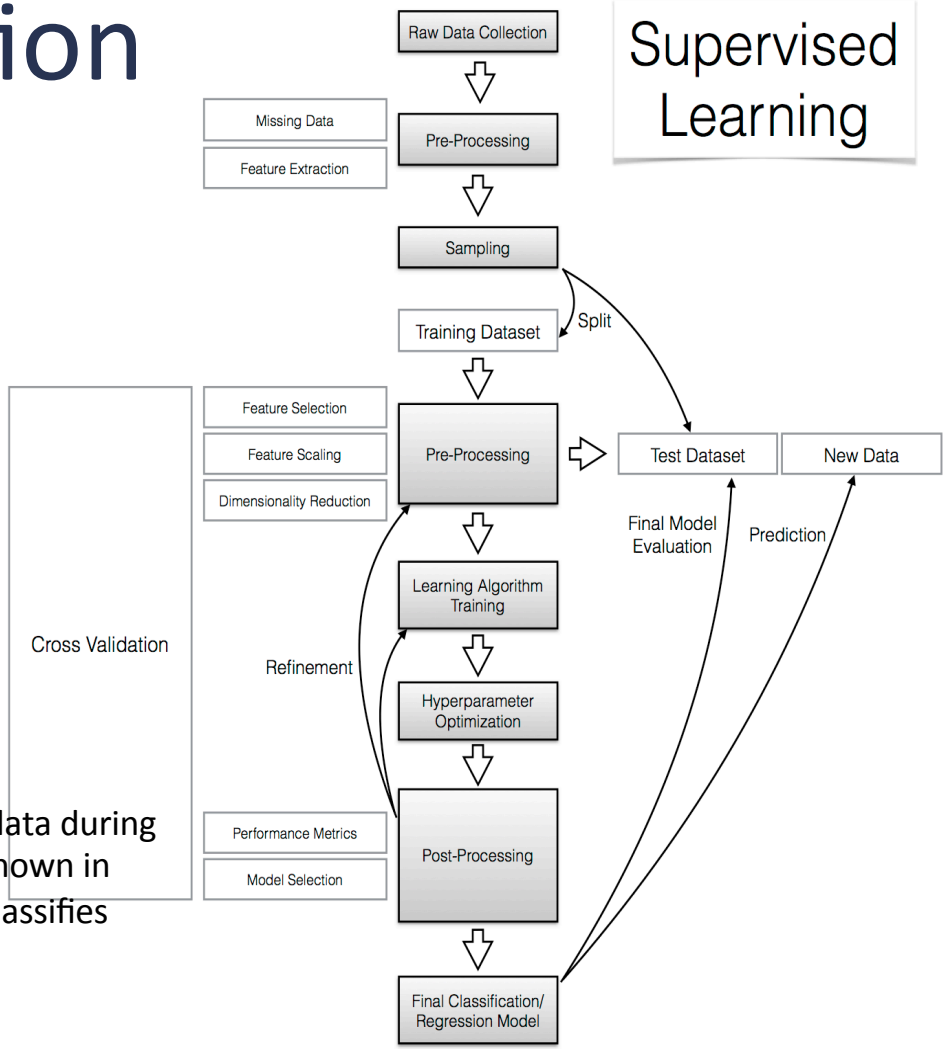
- Outlier detection is the process of finding data that is significantly different from or inconsistent with the rest of the data within a given dataset

Filtering

- Filtering is the automated process of finding relevant items from a pool of items. Filtering is generally applied via the following two approaches:
  - collaborative filtering
  - content-based filtering

Classification



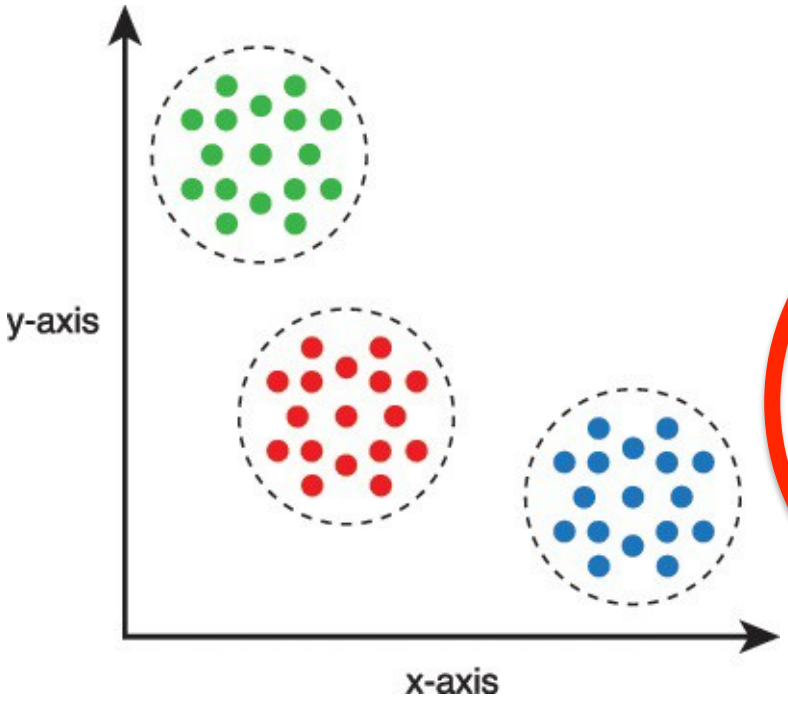


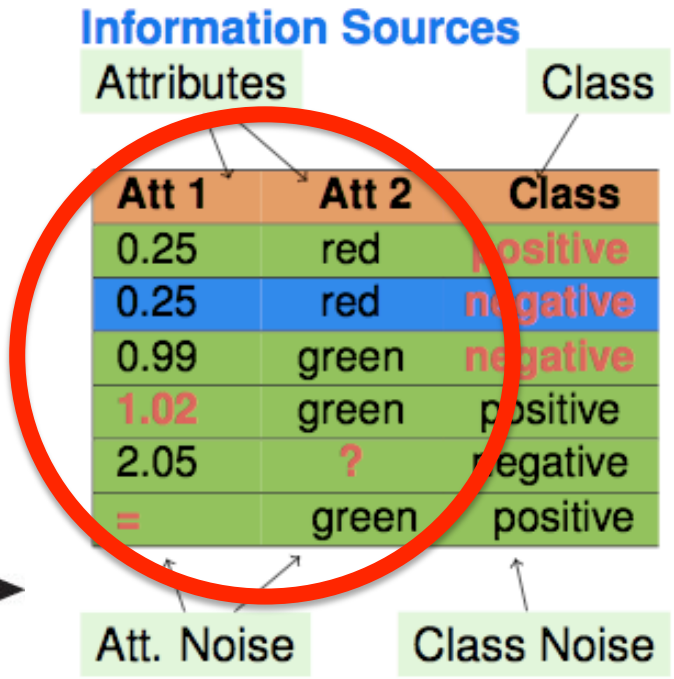
A simplified classification process, the machine is fed labeled data during training that builds its understanding of the classification, as shown in Figure 8.11. The machine is then fed unlabeled data, which it classifies itself.

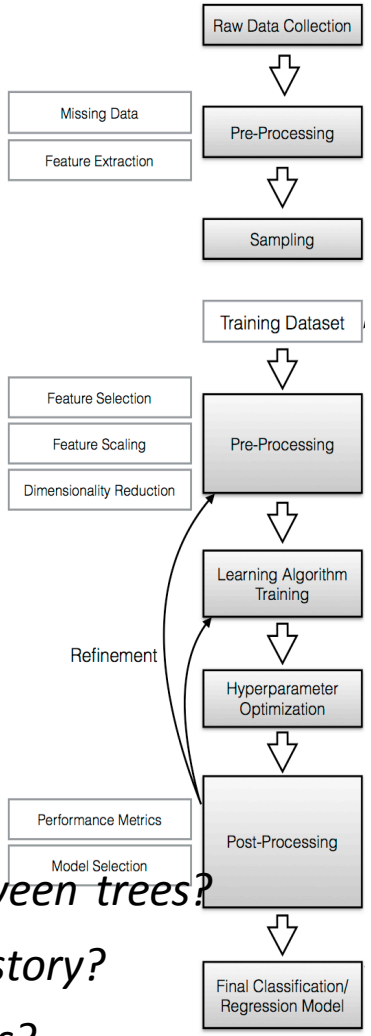
Sample questions can include:

- *Should an applicant's credit card application be accepted or rejected based on other accepted or rejected applications?*
- *Is a tomato a fruit or a vegetable based on the known examples of fruit and vegetables?*
- *Do the medical test results for the patient indicate a risk for a heart attack?*

Clustering







Sample questions can include:

- *How many different species of trees exist based on the similarity between trees?*
- *How many groups of customers exist based upon similar purchase history?*
- *What are the different groups of viruses based on their characteristics?*

Outlier Detection

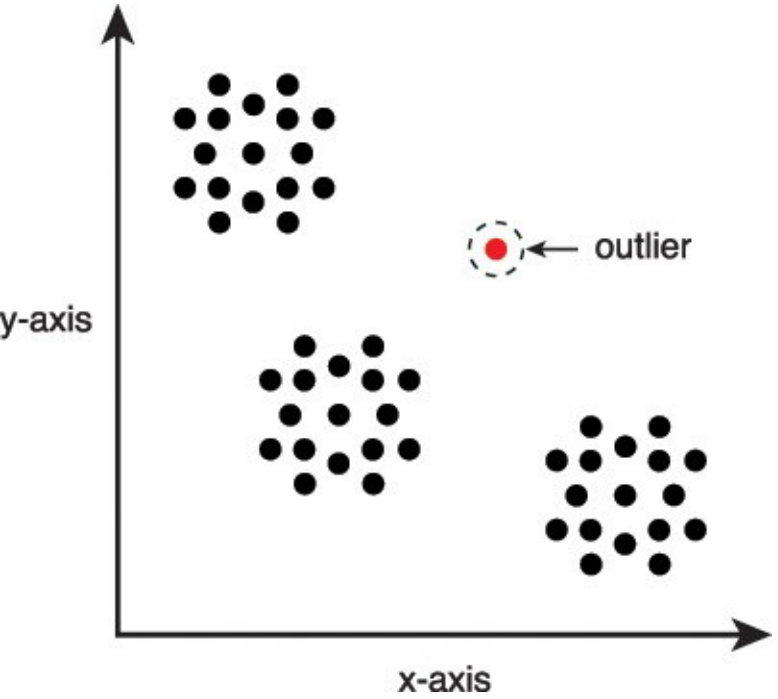


Figure 8.13 A scatter graph highlights an outlier.

Sample questions can include:

- *Is an athlete using performance enhancing drugs?*
- *Are there any wrongly identified fruits and vegetables in the training dataset used for a classification task?*
- *Is there a particular strain of virus that does not respond to medication?*

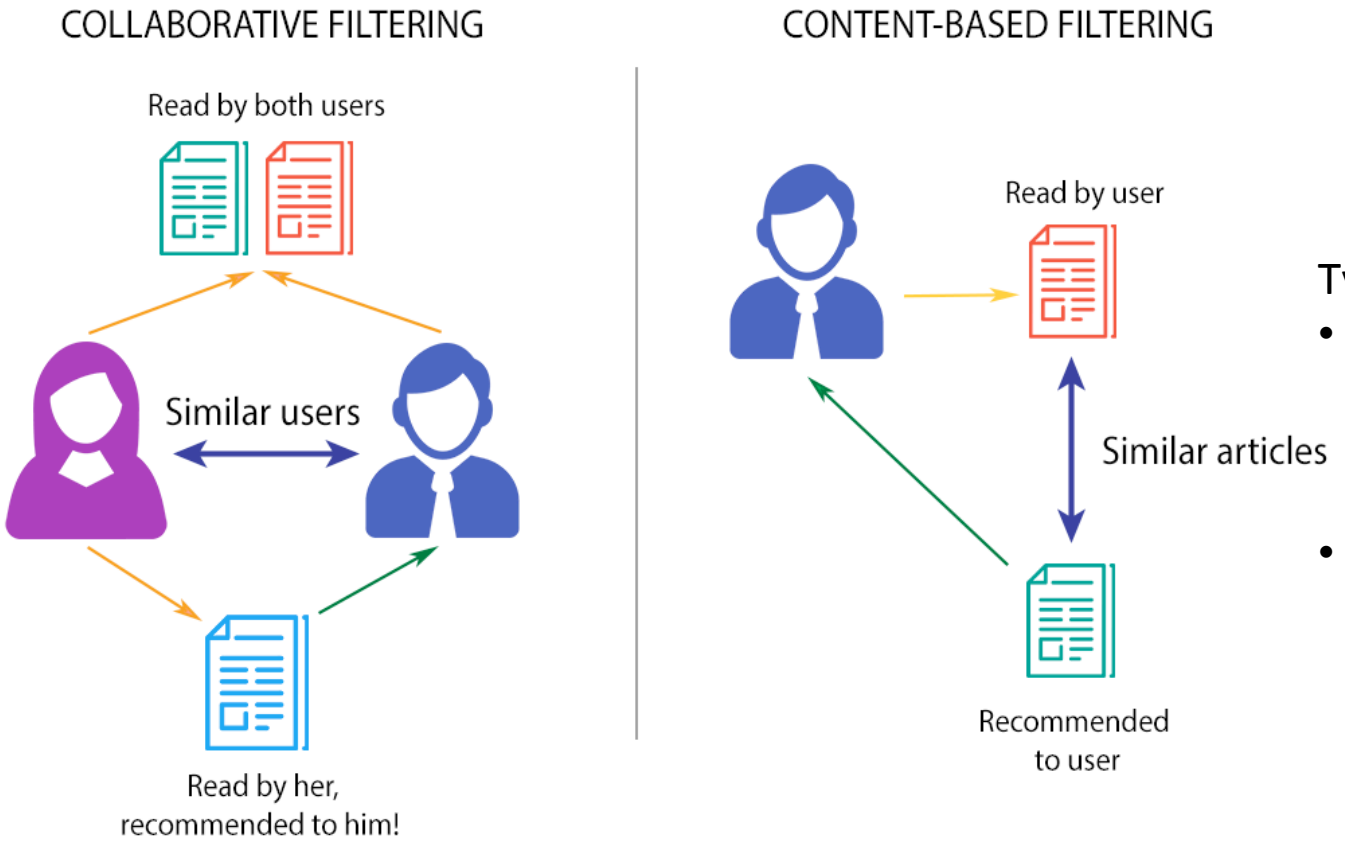
Model-based Approaches:

- Rational
  - Apply a model to represent normal data points
  - Outliers are points that do not fit to that model
- Sample approaches
  - Probabilistic tests based on statistical models − distribution models
  - Depth-based approaches
  - Deviation-based approaches
  - Some subspace outlier detection approaches

Proximity-based Approaches

- Rational
  - Examine the spatial proximity of each object in the data space
  - If the proximity of an object considerably deviates from the proximity of other objects it is considered an outlier
- Sample approaches
  - Distance-based approaches
  - Density-based approaches
  - Some subspace outlier detection approaches

Filtering



Types of filtering:

- Supervised
  - Instance Based
  - Attribute Based
- Unsupervised
  - Instance Based
  - Attribute Based

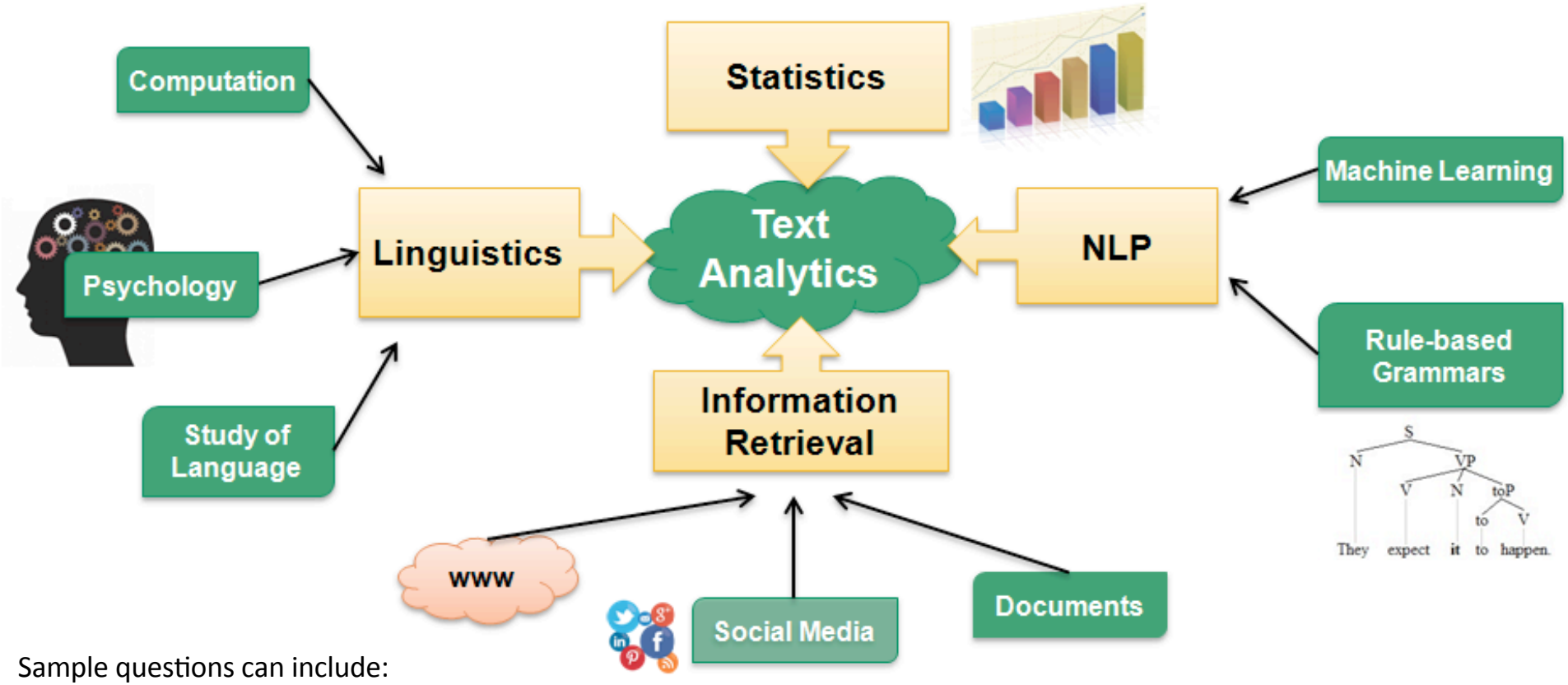
Sample questions can include:

- *How can only the news articles that a user is interested in be displayed?*
- *Which holiday destinations can be recommended based on the travel history of a vacationer?*
- *Which other new users can be suggested as friends based on the current profile of a person?*

Big Data Analysis Techniques

- Semantic Analysis
  - Text Analytics
    - Parsing text within documents to extract:
      - Named Entities – person, group, place, company
      - Pattern-Based Entities – social security number, zip code
      - Concepts – an abstract representation of an entity
      - Facts – relationship between entities
    - Categorization of documents using these extracted entities and facts.
    - Natural Language Processing
      - Natural language processing is a computer's ability to comprehend human speech and text as naturally understood by humans.
    - Sentiment Analysis
      - Sentiment analysis is a specialized form of text analysis that focuses on determining the bias or emotions of individuals.

Natural Language Processing (NLP)



Sample questions can include:

- *How can an automated phone exchange system that can recognize the correct department extension as dictated verbally by the caller be developed?*
- *How can grammatical mistakes be automatically identified?*
- *How can a system that can correctly understand different accents of English language be designed?*

Text Analytics and Sentiment Analysis

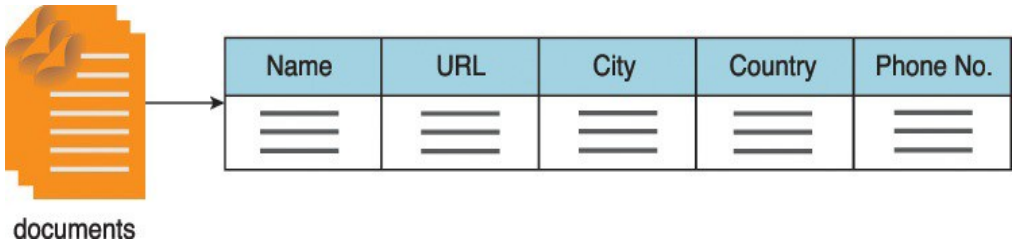


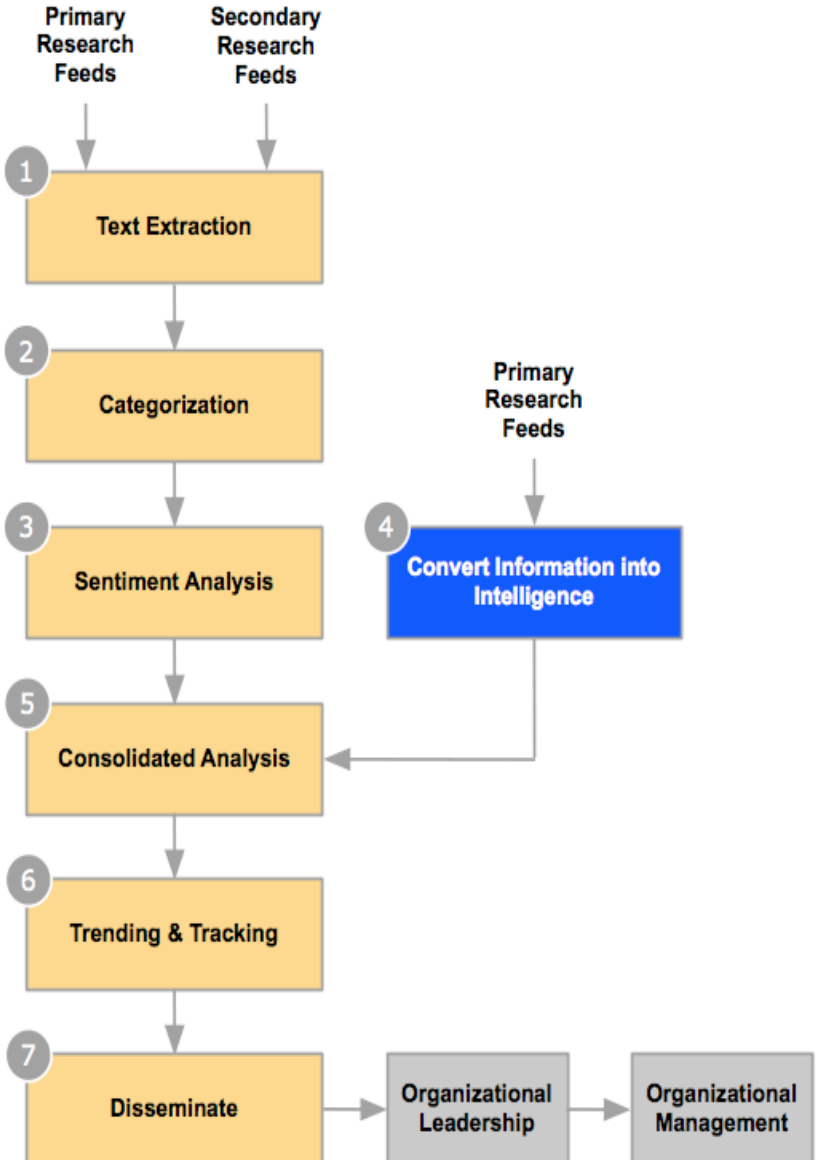
Figure 8.14 Entities are extracted from text files using semantic rules and structured so that they can be searched.

Sample questions can include (Sentiment Analysis):

- *How can customer reactions to the new packaging of the product be gauged?*
- *Which contestant is a likely winner of a singing contest?*
- *Can customer churn be measured by social media comments?*

Sample questions can include (Text Analytics):

- *How can I categorize Web sites based on the content of their Web pages?*
- *How can I find the books that contain content that is relevant to the topic that I am studying?*
- *How can I identify contracts that contain confidential company information?*



Big Data Analysis Techniques

Visual Analysis

- Heat Maps
  - Heat maps are an effective visual analysis **technique for expressing patterns**, data compositions via part-whole relations and geographic distributions of data.
- Time Series Plots
  - Time series plots allow **the analysis of data that is recorded over periodic intervals** of time. This type of analysis makes use of time series, which is a time-ordered collection of values recorded over regular time intervals.
- Network Graphs
  - Within the context of visual analysis, a network graph **depicts an interconnected collection of entities**.
  - It involves plotting entities as nodes and connections as edges between nodes. There are specialized variations of network analysis, including:
    - route optimization
    - social network analysis
    - spread prediction, such as the spread of a contagious disease
- Spatial Data Mapping
  - Spatial or geospatial data is commonly used **to identify the geographic location of individual entities that can then be mapped**. Spatial data analysis is focused on **analyzing location-based data in order to find different geographic relationships** and patterns between entities.

Heat Maps

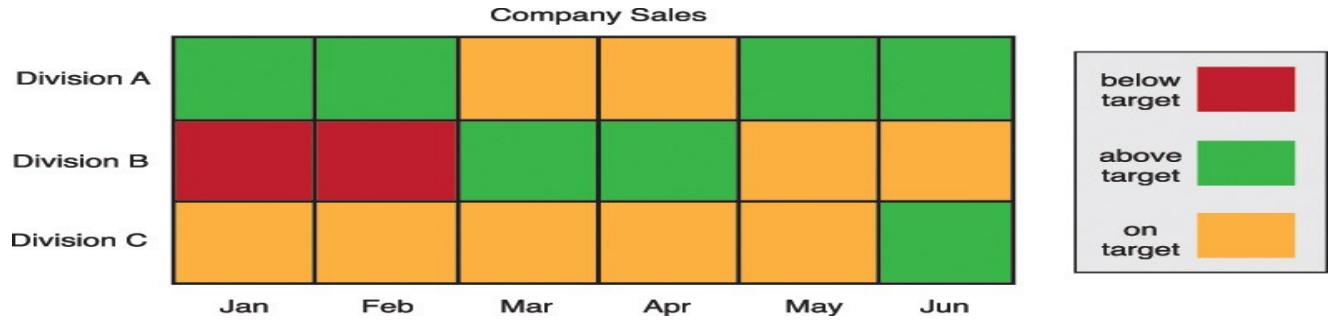


Figure 8.15 This chart heat map depicts the sales of three divisions within a company over a period of six months.

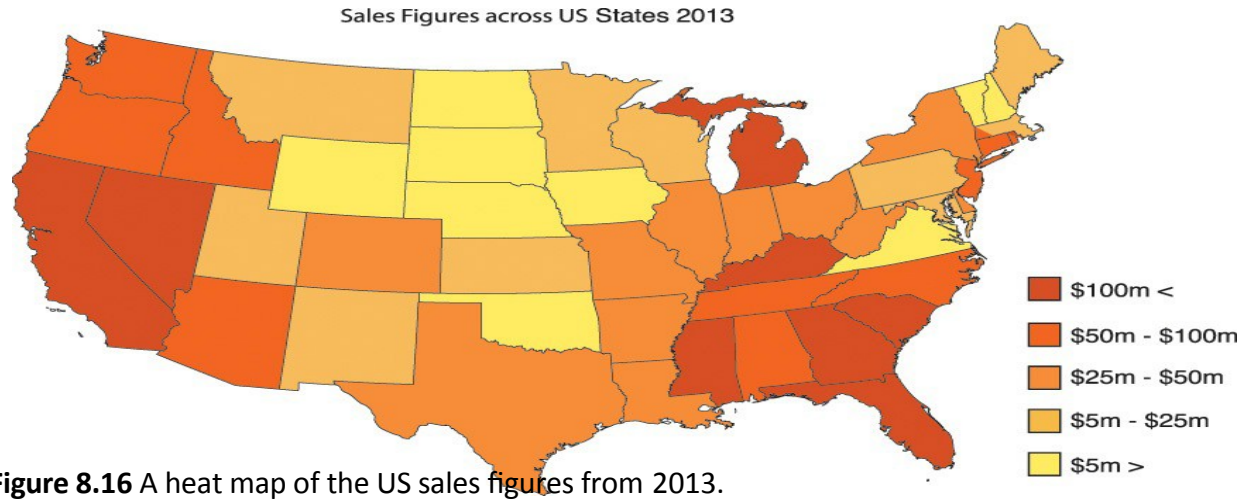


Figure 8.16 A heat map of the US sales figures from 2013.

Sample questions can include:

- *How can I visually identify any patterns related to carbon emissions across a large number of cities around the world?*
- *How can I see if there are any patterns of different types of cancers in relation to different ethnicities?*

Time Series Plots

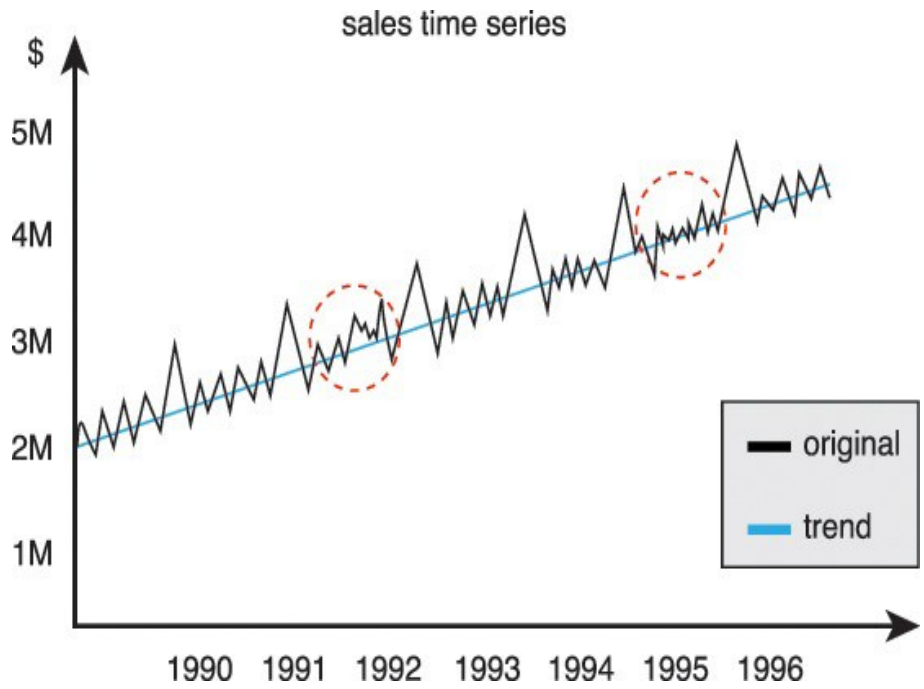
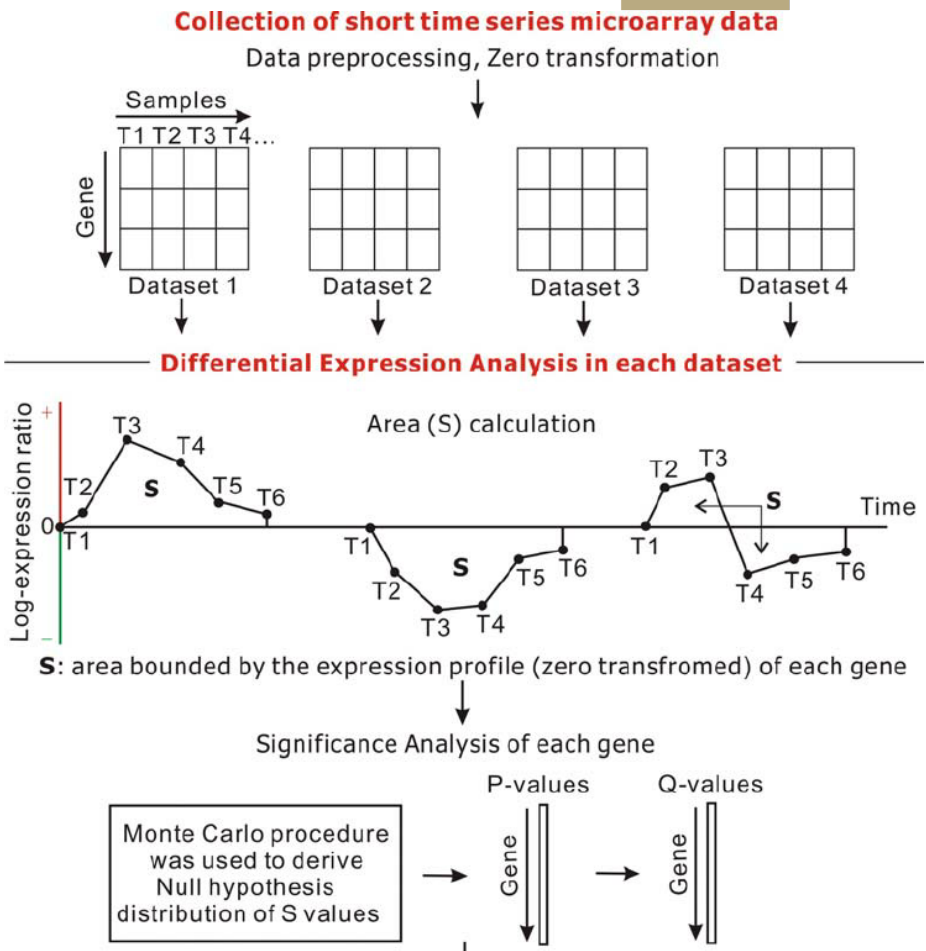


Figure 8.17 A line chart depicts a sales time series from 1990 to 1996.



Sample questions can include:

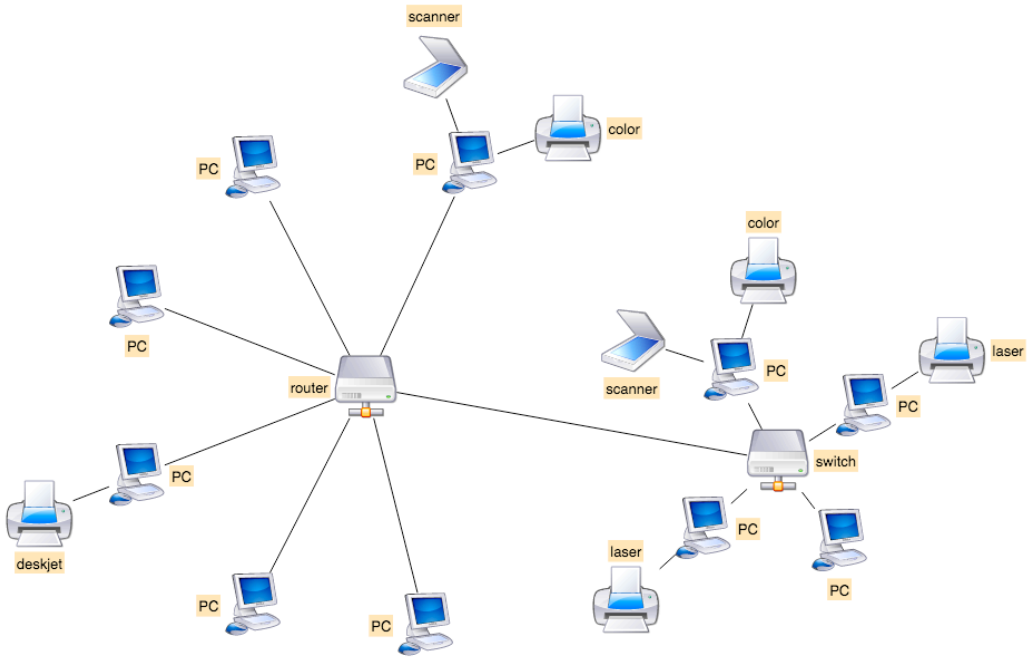
- *How **much** yield should the farmer expect based on historical yield data?*
- *What is the expected **increase** in population in the next 5 years?*
- *Is the current **decrease** in sales a one-off occurrence or does it occur regularly?*

Sun, Ruping & Fu, Xuping & Guo, Fenghua & Ma, Zhaorong & Goulbourne, Chris & Jiang, Mei & Li, Yao & Xie, yi & Mao, Yumin. (2009). A strategy for meta-analysis of short time series microarray datasets. Frontiers in bioscience : a journal and virtual library

Time Series Analysis

- Time Series Decomposition
  - Trend component: long term trend
  - Seasonal component: seasonal variation
  - Cyclical component: repeated but non-periodic fluctuations
  - Irregular component: the residuals
- Time Series Forecasting - to forecast future events based on known past data
  - Autoregressive moving average (ARMA)
  - Autoregressive integrated moving average (ARIMA)
- Time Series Clustering - to partition time series data into groups based on similarity or distance, so that time series in the same cluster are similar
  - Measure of distance/dissimilarity
  - Euclidean distance
  - Manhattan distance
  - Maximum norm
  - Hamming distance
  - The angle between two vectors (inner product)
  - Dynamic Time Warping (DTW) distance
- Time Series Classification- to build a classification model based on labelled time series and then use the model to predict the label of unlabeled time series
- Feature Extraction
  - Singular Value Decomposition (SVD)
  - Discrete Fourier Transform (DFT)
  - Discrete Wavelet Transform (DWT)
  - Piecewise Aggregate Approximation (PAA)
  - Perpetually Important Points (PIP)
  - Piecewise Linear Representation
  - Symbolic Representation

Network Graphs



An example of a using a graph to calculate platform reliability.

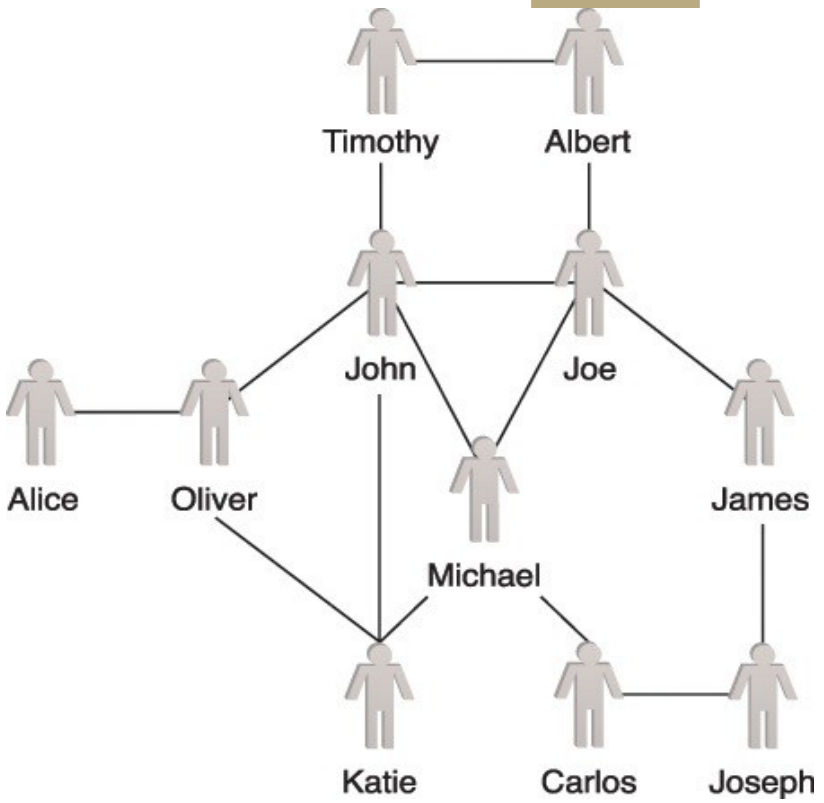


Figure 8.18 An example of a social network graph.

Sample questions may include:

- *How can I identify influencers within a large group of users?*
- *Are two individuals related to each other via a long chain of ancestry?*
- *How can I identify interaction patterns among a very large number of protein-to- protein interactions?*

Spatial Data Mapping

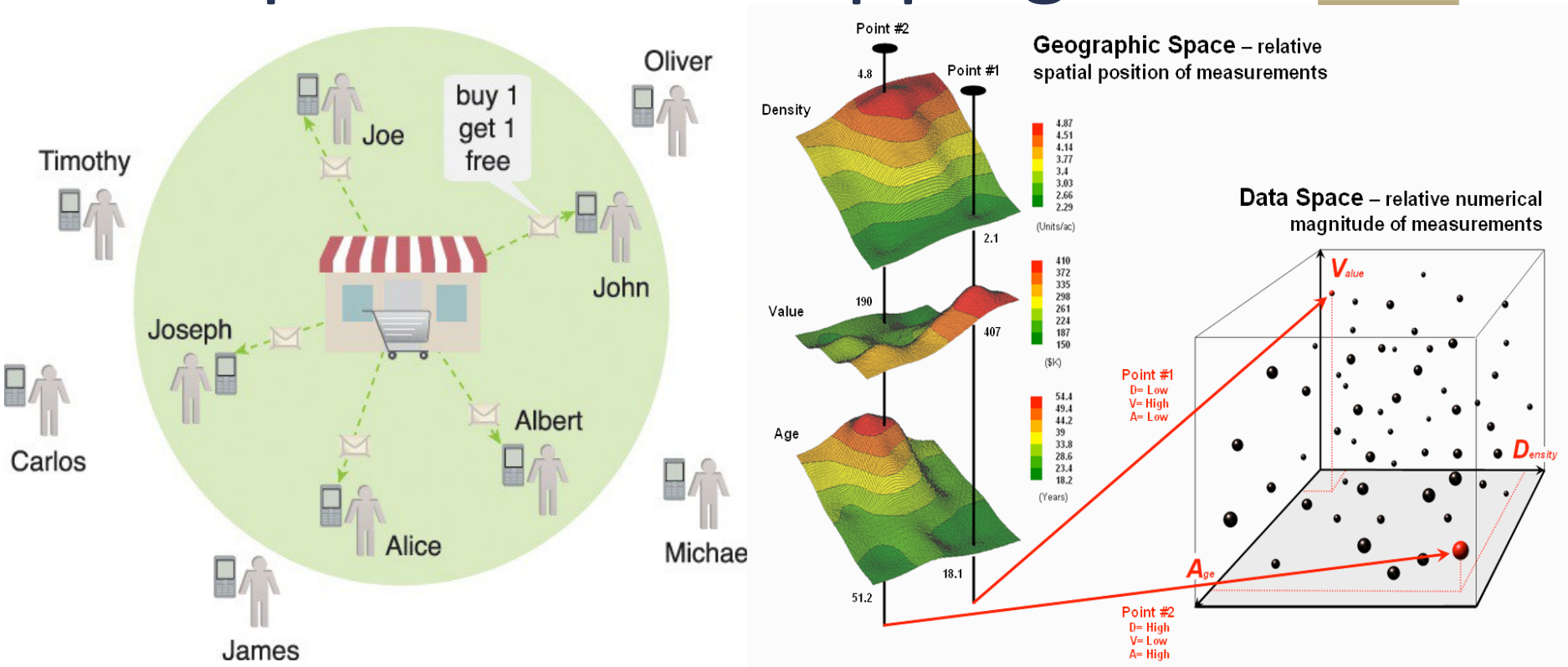


Figure 8.19 Spatial data analysis can be used for targeted marketing by a supermarket is using spatial analysis for targeted marketing

Sample questions can include:

- *How many houses will be affected due to a road widening project?*
- *How far do customers have to commute in order to get to a supermarket?*
- *Where are the high and low concentrations of a particular mineral based on readings taken from a number of sample locations within an area?*

Enterprise Technologies and Big Data Business Intelligence

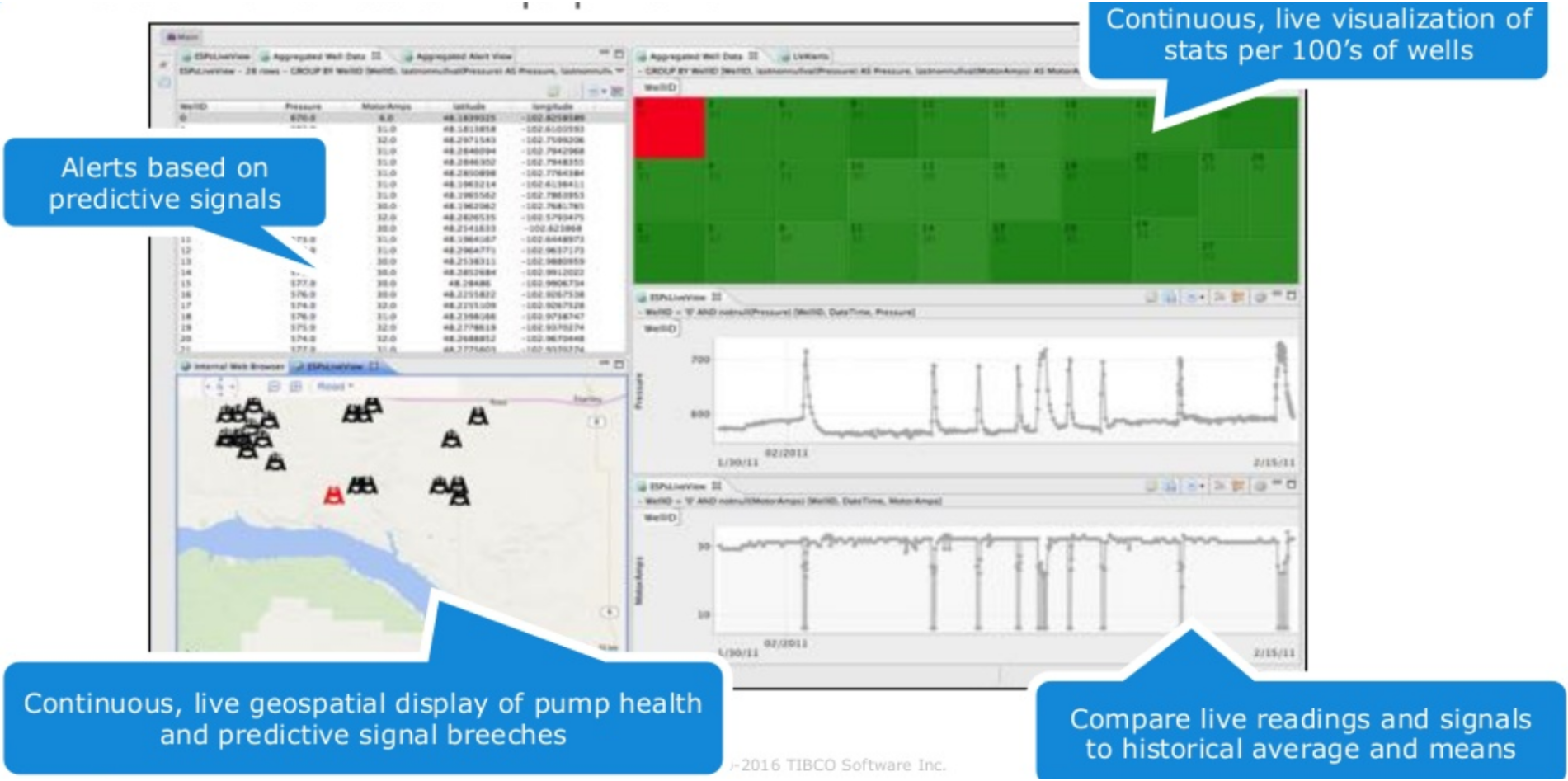
BI Data is collected from throughout the enterprise and warehoused in a **data warehouse**, analytical processing systems answer more complex queries and can provide deeper insight into business operations, but usually used for operations.

It is from these data stores and processing that management gains insight into broader corporate performance, KPI, and for operational insight.

These systems include:

- Operations/Research Storage and Processing Technologies
  - Stream Processing
  - Batch Processing
- Business Storage and Processing Technologies
  - OLAP (Vertica, Netezza, Hadoop)
  - OLTP-oriented databases: (Postgres, MySQL, VoltDB, Oracle),
- Hybrid OLTP and OLAP: (HBase, Cassandra, MongoDB)
- Extract Transform Load (ETL) Processes
- Data Warehouses
- Data Marts

Why Streaming?



Source: Tibco.com

Streaming and OLTP

Streaming Processing

- Under the streaming model, data is fed into analytics tools piece-by-piece. The processing is usually done in real time.
- Online Transaction Processing (OLTP)
  - OLTP is a software system that **processes transaction-oriented data**. The term "online transaction" refers to the completion of an activity in real-time and is not batch-processed. OLTP systems store operational data that is normalized. This data is a common source of structured data and serves as input to many analytic processes

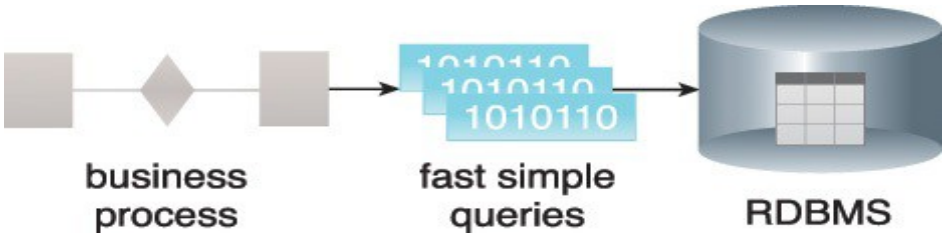


Figure 4.1 OLTP systems perform simple database operations to provide sub-second response times.

Batch and OLAP

Batch Processing

- Under the batch processing model, a set of data is collected over time, then fed into an analytics system. In other words, you collect a batch of information, then send it in for processing.

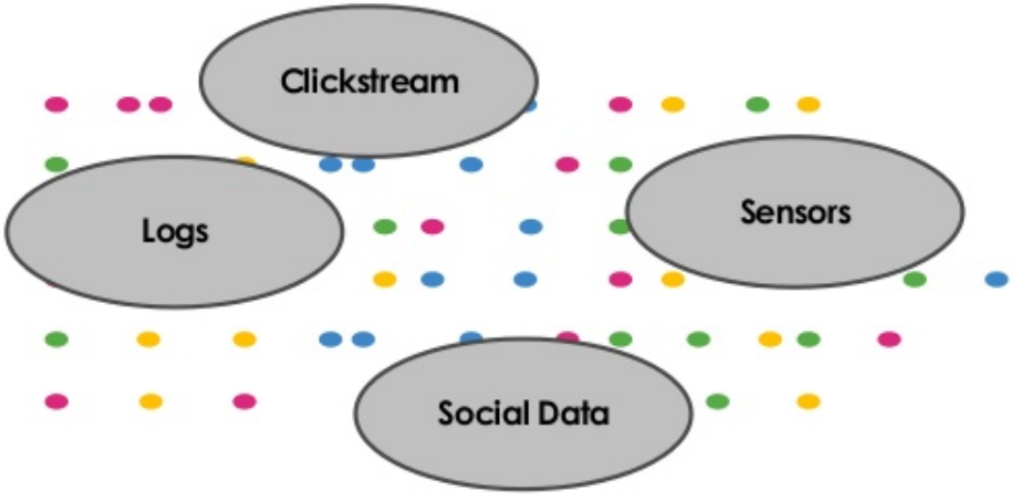
Online Analytic Processing (OLAP)

- Online analytical processing (OLAP) **systems are used for processing data analysis queries**. OLAPs form an integral part of business intelligence, data mining and machine learning processes. They are relevant to Big Data in that they can serve as both a data source as well as a data sink that is capable of receiving data. They are used in diagnostic, predictive and prescriptive analytics

Figure 4.2 OLAP systems use multidimensional databases.

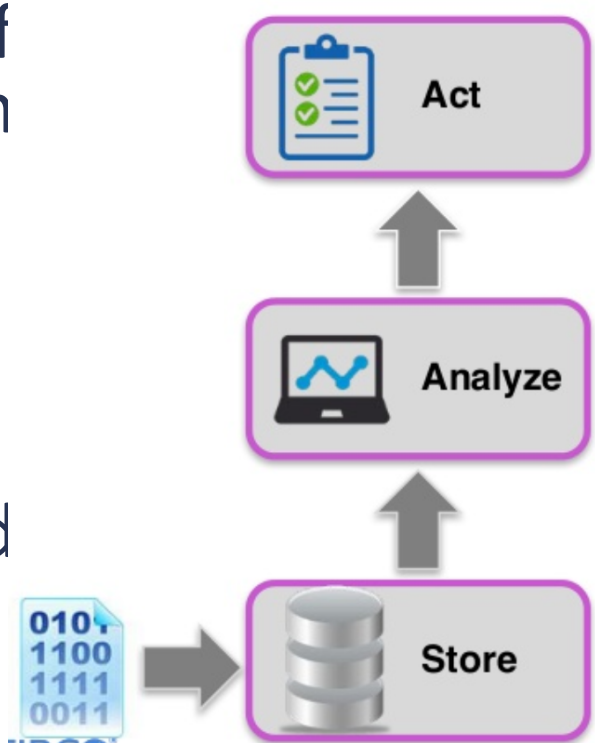
Streaming Analytics: "What Is A Stream"

- Consists of pieces of data typically generated due to a change in state:
  - One or more identifiers
  - Timestamp & payload
  - Immutable: copy then change or delete
- Typically unbounded; there is no end to the data
  - Batch dataset: "bounded"
- Can be raw or derived



Traditional Data Processing – Request & Response

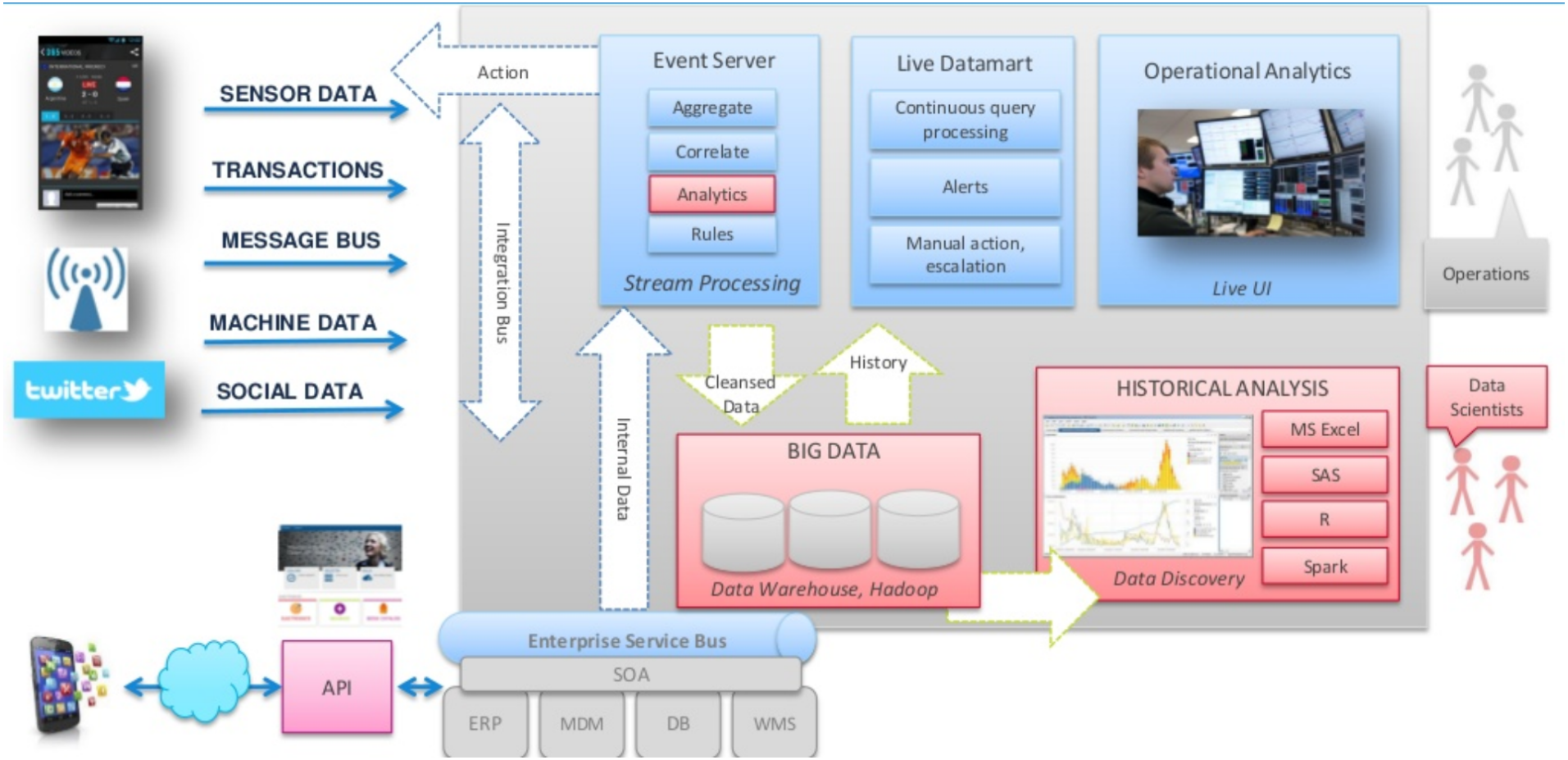
- Data is collected from a variety of sources, and placed in a persisten store
  - Relational database
  - NoSQL store
  - Hadoop environment
- Analytical processes are executed against the stored data to detect opportunities or threats
- Actions are identified, delivered, and executed across various business channels



The New Era: Streaming Analytics

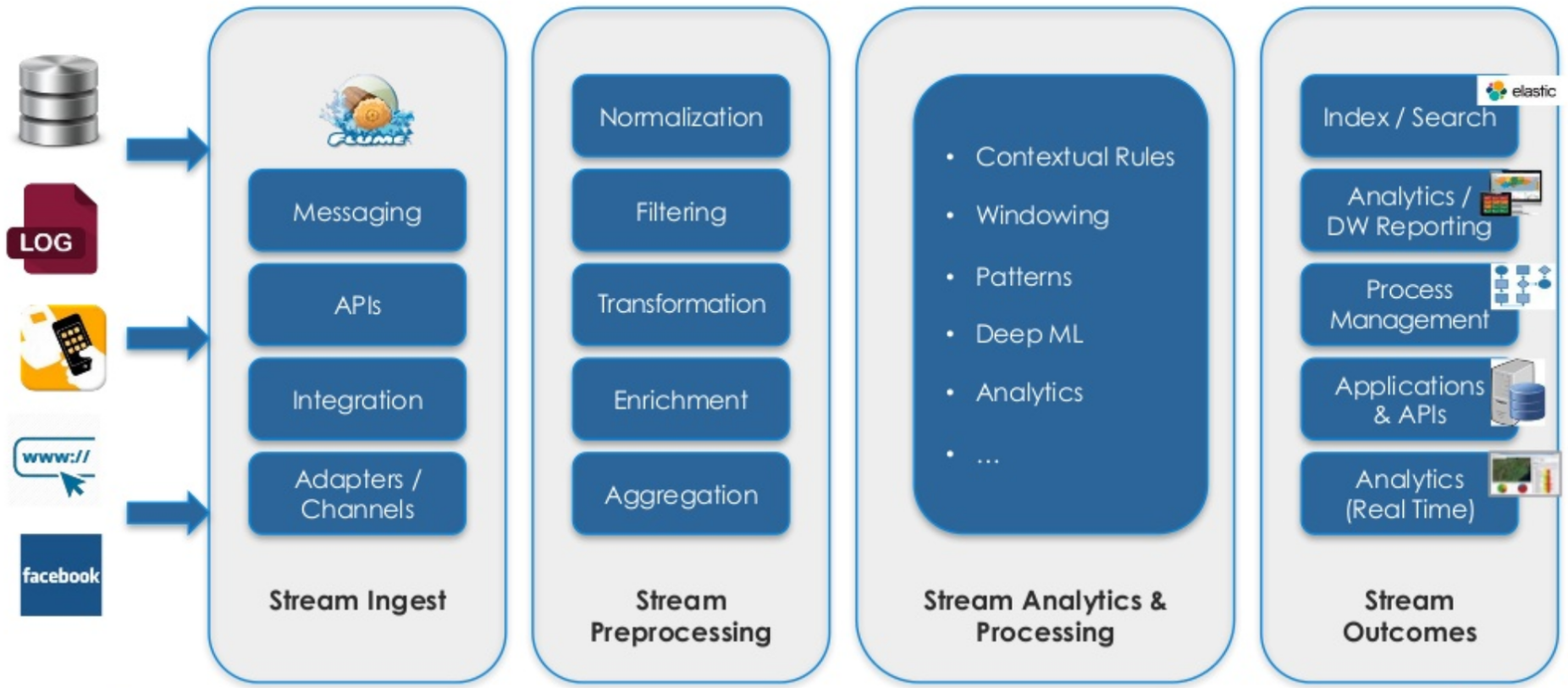
- Events are analyzed and processed in real-time as they arrive.
- Decisions are timely, contextual, and based on fresh data.
- Decision latency is eliminated, resulting in:
  - Superior Customer Experience,
  - Operational Excellence,
  - Instant Awareness and Timely Decisions

Streaming Analytics Reference Architecture



Source: tibco.com

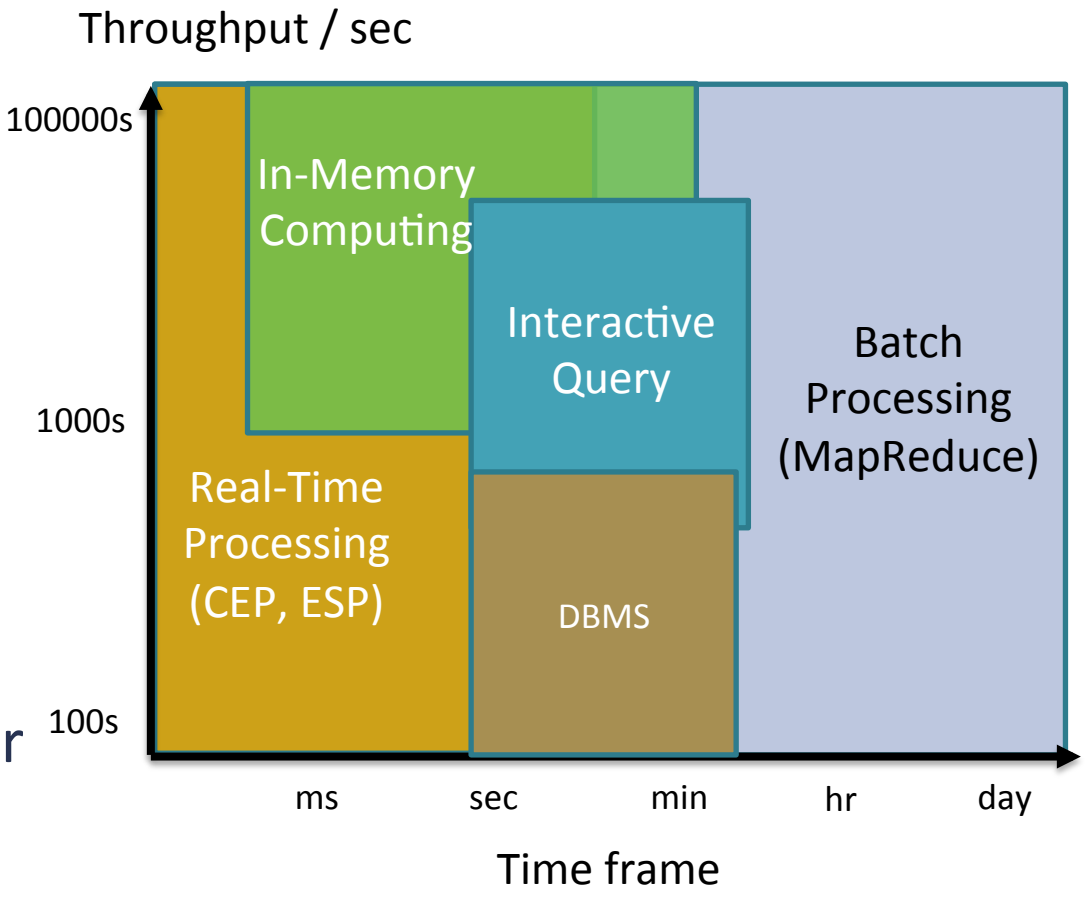
Streaming Analytics Processing Pipeline



Source: tibco.com

Real-time Batch Comparison

- Apache Spark processes data in-memory while Hadoop MapReduce persists back to the disk after a map or reduce action, so Spark should outperform Hadoop MapReduce.
- Spark needs a lot of memory. Much like standard DBs, it loads a process into memory and keeps it there until further notice, for the sake of caching.



Extract Transform Load (ETL)

- Extract Transform Load (ETL) is a process of **loading data from a source system into a target system**. The source system can be a database, a flat file, or an application. Similarly, the target system can be a database or some other storage system.

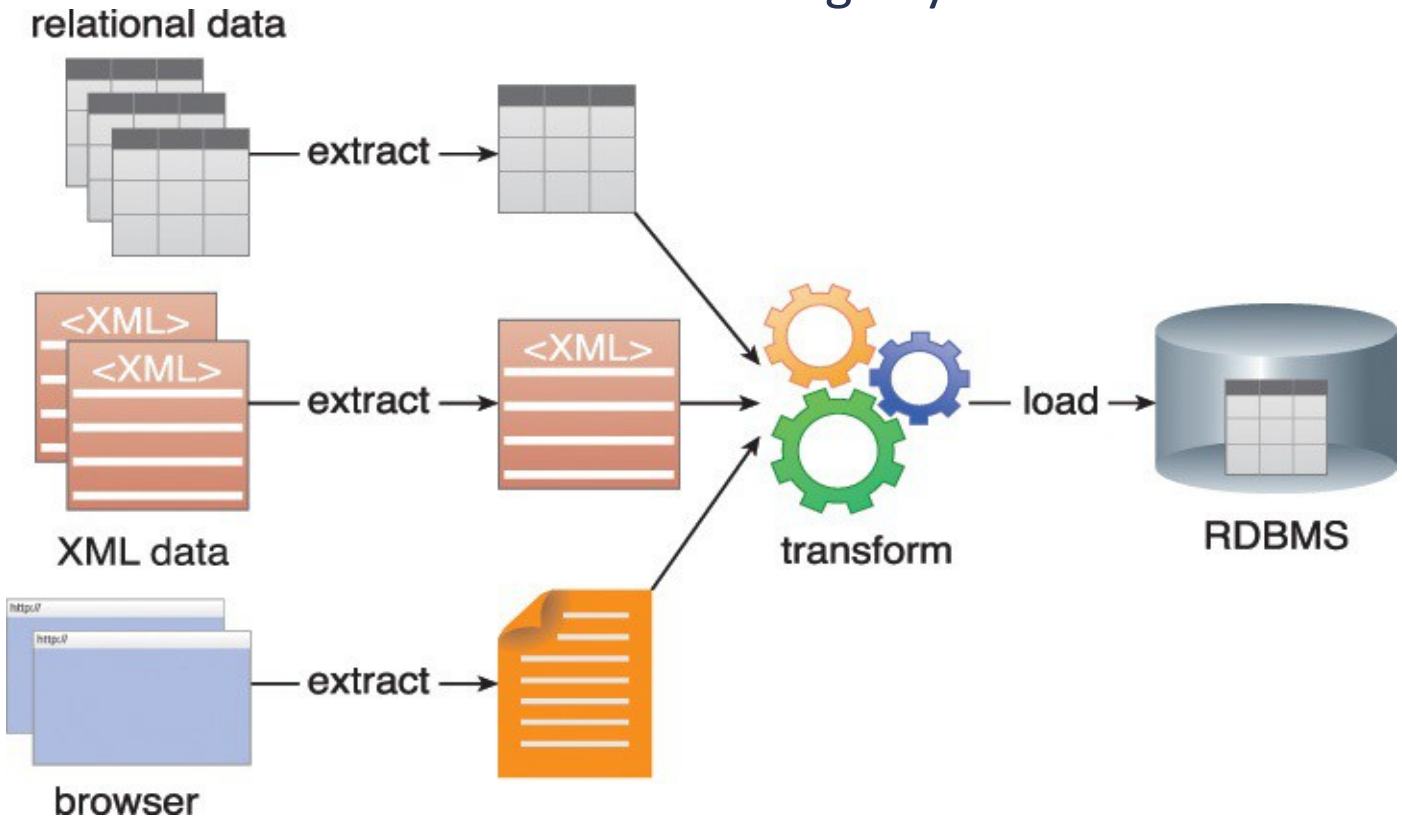


Figure 4.3 An ETL process can extract data from multiple sources and transform it for loading into a single target system.

Data Warehouse

A data warehouse is **a central, enterprise-wide repository consisting of historical and current data**. Data warehouses are **heavily used by BI** to run various analytical queries, and they usually interface with an OLAP system to support multi-dimensional analytical queries.

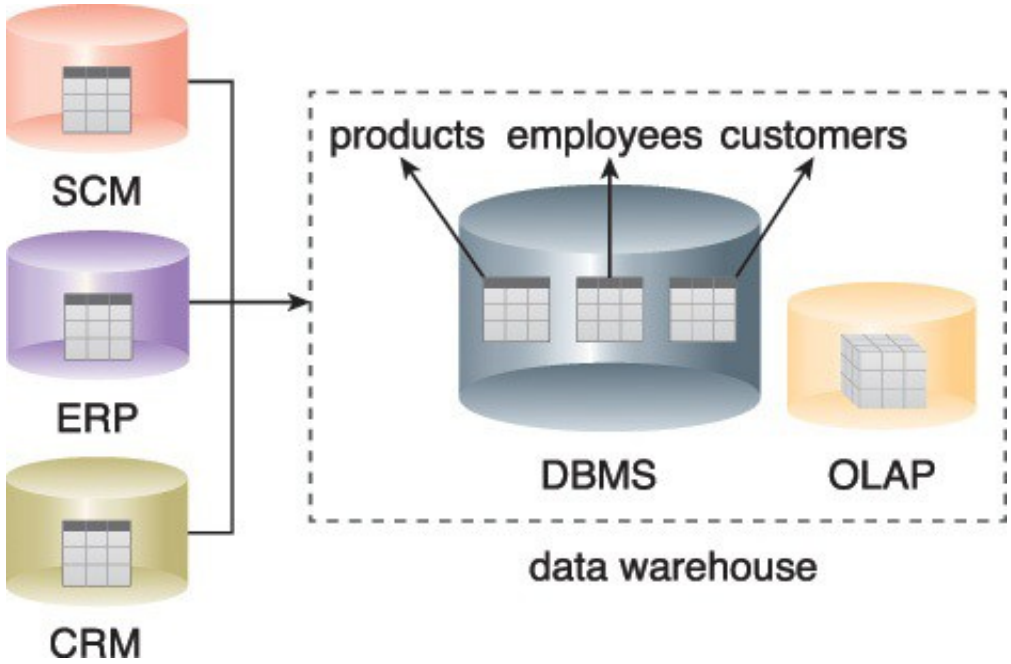


Figure 4.4 Batch jobs periodically load data into a data warehouse from operational systems like ERP, CRM and SCM.

Data Marts

A data mart is **a particular subset of the data** stored in a data warehouse that typically belongs to a department, division, or specific line of business.

Enterprise-wide data is collected and business entities are then extracted. Domain specific entities are persisted into the data warehouse via an ETL process.

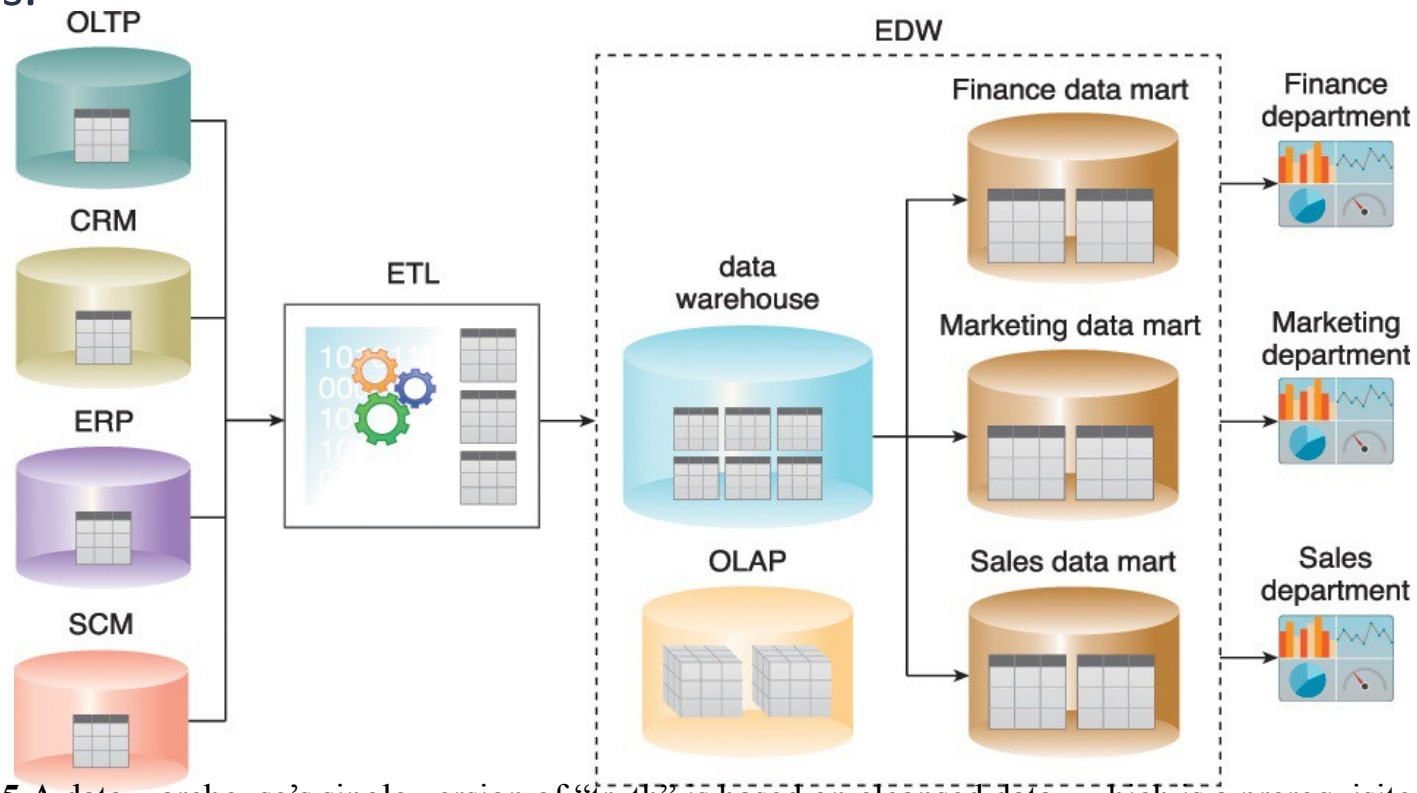


Figure 4.5 A data warehouse's single version of "truth" is based on cleansed data, which is a prerequisite for accurate and error-free reports, as per the output shown on the right.

Ad-hoc Reports and Dashboards

Traditional BI primarily utilizes **descriptive** and **diagnostic** analytics to provide information on historical and current events. It is not "intelligent" because it only provides answers to correctly formulated questions through:

- **Ad-hoc reporting** is a process that involves manually processing data to produce custom-made reports
- **Dashboards** provide a holistic view of key business areas. The information displayed on dashboards is generated at periodic intervals in real-time or near-realtime.

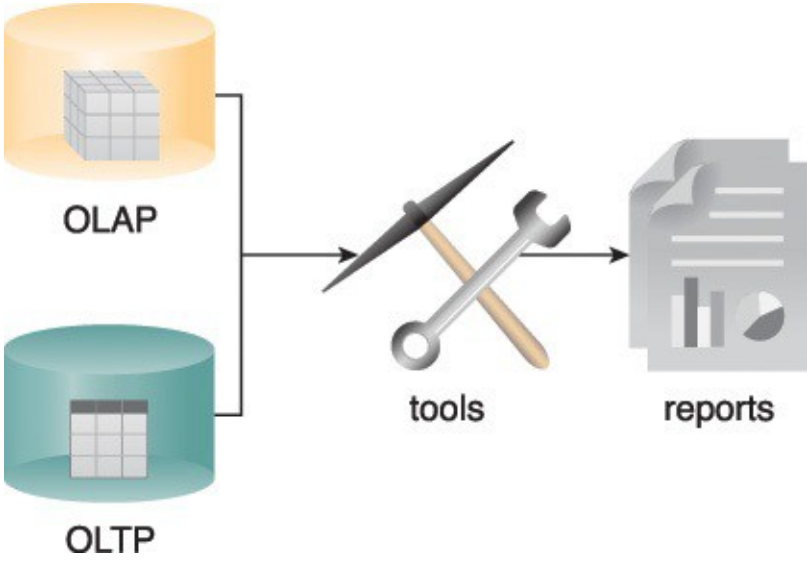


Figure 4.6 OLAP and OLTP data sources can be used by BI tools for both ad-hoc reporting and dashboards.

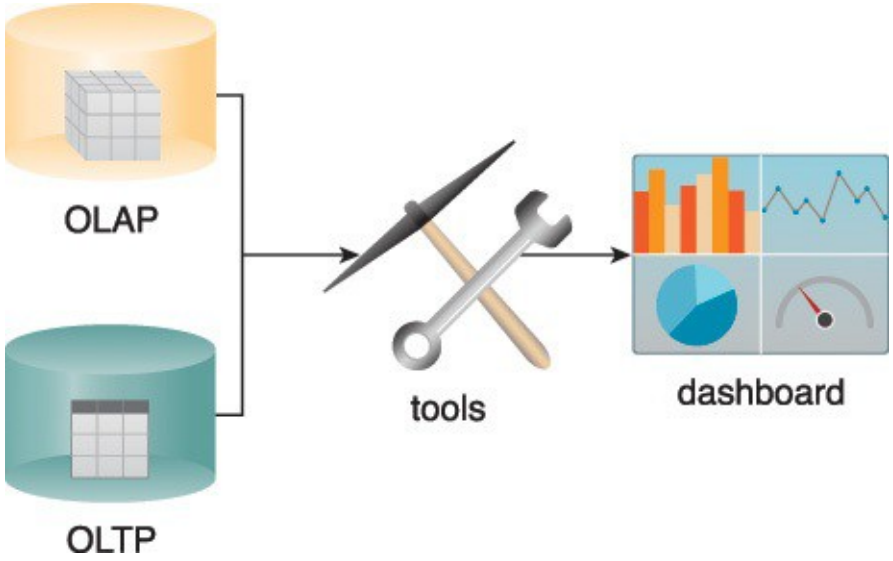


Figure 4.7 BI tools use both OLAP and OLTP to display the information on dashboards.

Traditional Business Intelligence

Traditional BI uses **data warehouses and data marts for reporting and data analysis** because they allow complex data analysis queries with multiple joins and aggregations to be issued

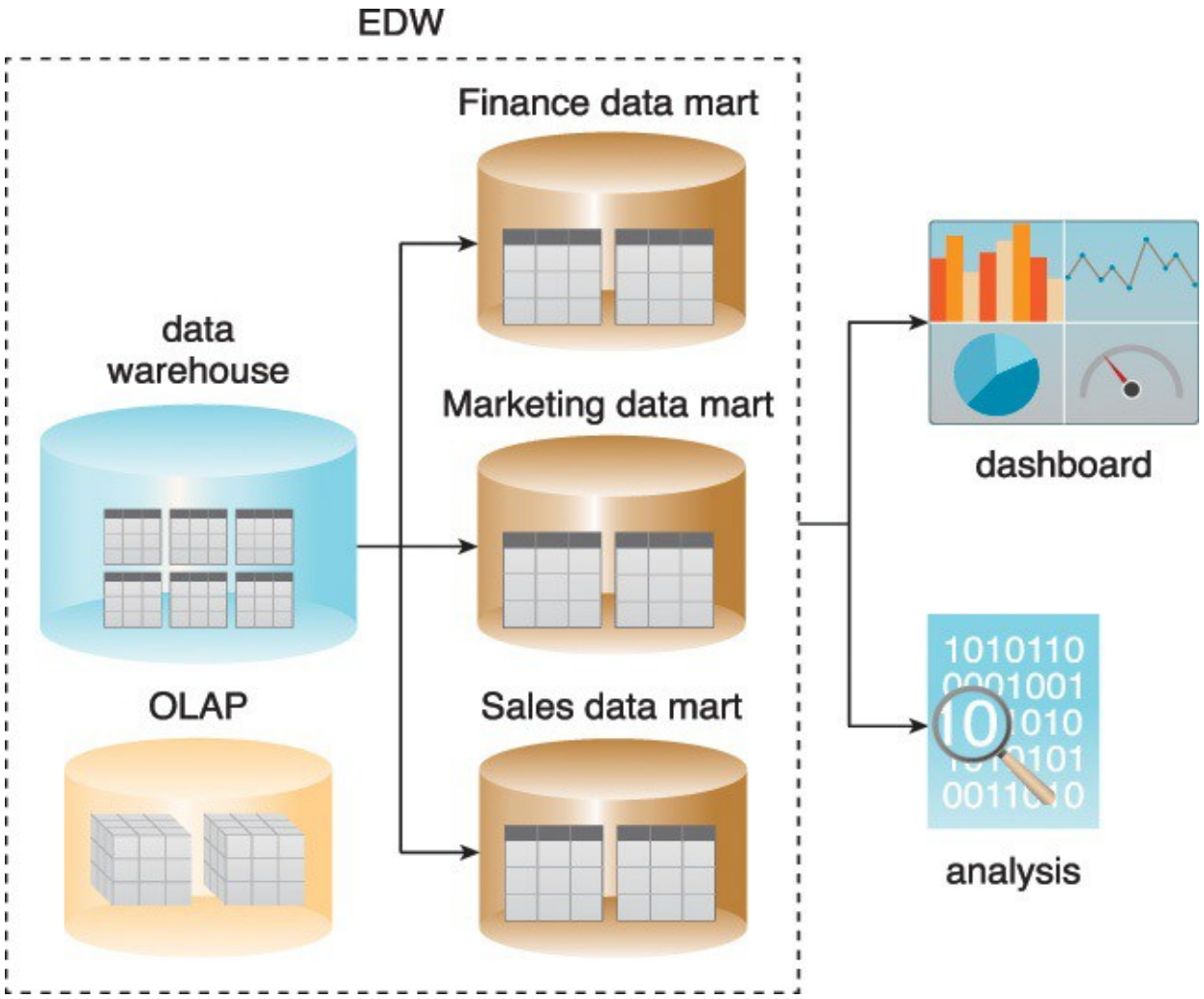


Figure 4.8 An example of traditional BI.

Big Data Business Intelligence

- Big Data BI builds upon traditional BI by **acting on the cleansed, consolidated enterprise- wide data in the data warehouse and combining it with semi-structured and unstructured data sources.**
- It comprises both predictive and prescriptive analytics to facilitate the development of an enterprise-wide understanding of business performance.

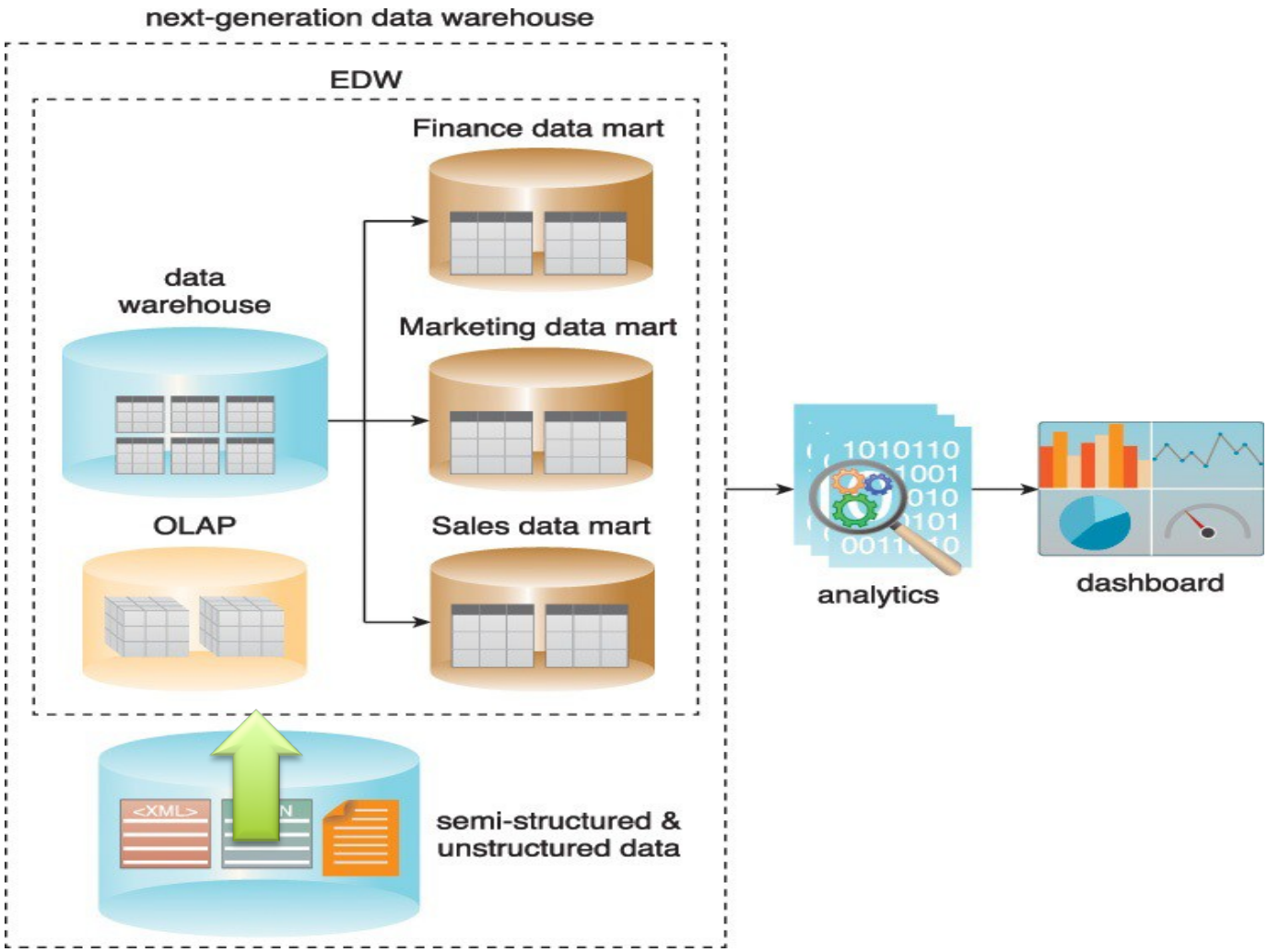
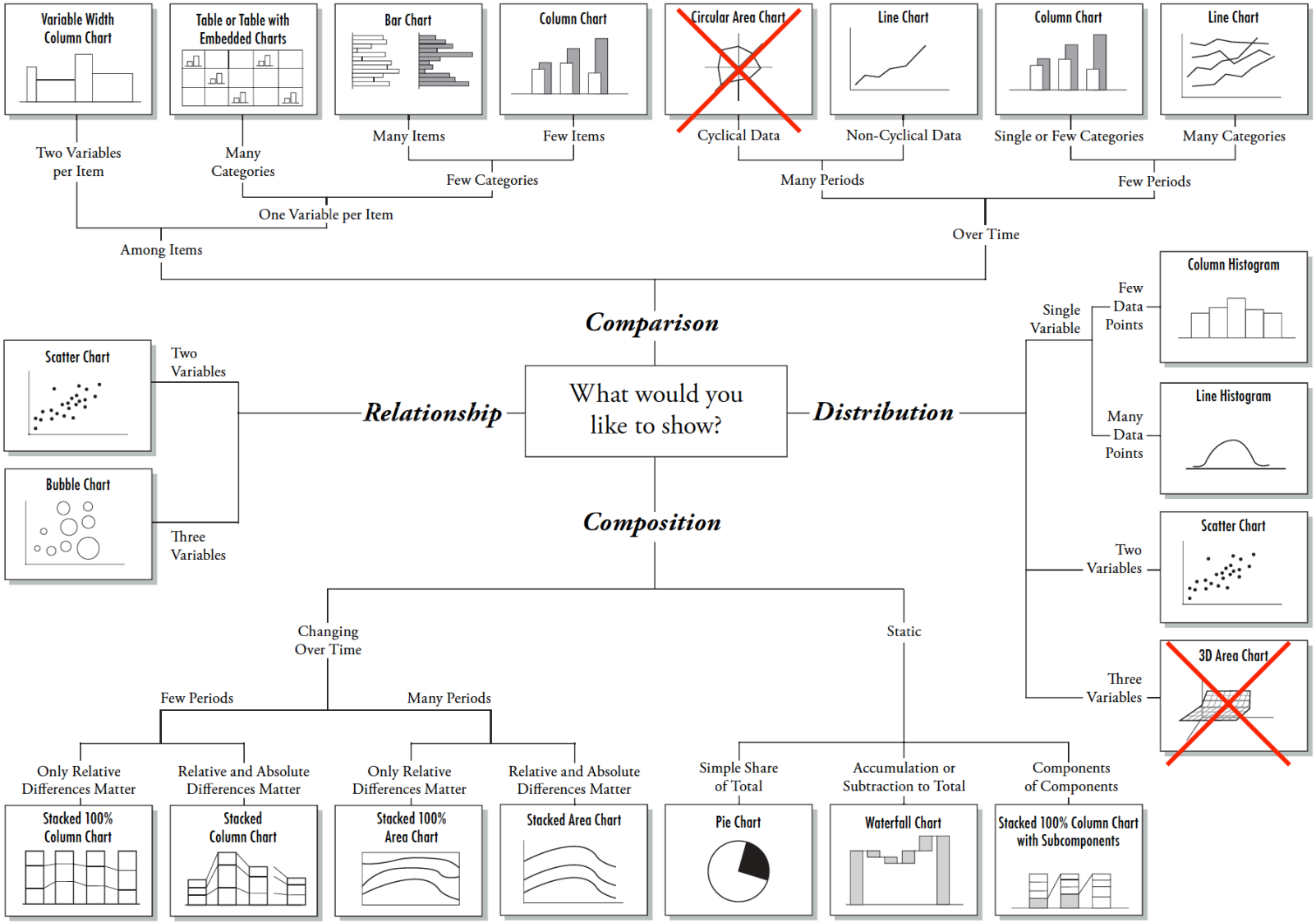


Figure 4.9 A next-generation data warehouse.

Traditional & Big Data Visualizations

- Data visualization is a technique whereby analytical results are graphically communicated, interpreted, and portrayed using elements like **charts, maps, data grids, infographics and alerts.**
- Graphically representing data can make it easier to understand reports, view trends and identify patterns.
- Big Data solutions require data visualization tools that can **seamlessly connect to structured, semi-structured and unstructured data sources** and are further capable of handling millions of data records.
- Data visualization tools for Big Data solutions generally use **in-memory analytical technologies** that reduce the latency normally attributed to traditional, disk-based data visualization tools.

Visualization Types



Python

- Python is a general-purposed high-level programming language
  - Web development
  - Networking
  - Scientific computing
  - Data analytics
  - …

Python for Data Analytics

- The nature of Python makes it perfect-fit for data analytics
  - Easy to learn
  - Readable
  - Scalable
  - Extensive set of libraries
  - Easy integration with other apps
  - Active community & ecosystem

Portfolio's

- iPython is a Python command shell for interactive computing
- Jupyter Notebook (the former iPython Notebook) is a web-based interactive data analysis environment that supports iPython

Commonly Used Data Sciences Tools

Library	Usage
numpy, scipy	Scientific & technical computing
pandas	Data manipulation & aggregation
mlpy, scikit-learn	Machine learning
theano, tensorflow, keras	Deep learning
statsmodels	Statistical analysis
nltk, gensim	Text processing
networkx	Network analysis & visualization
bokeh, matplotlib, seaborn, plotly	Visualization
beautifulsoup, scrapy	Web scraping