

# Transforming Data into Knowledge

Matthew Renze

Iowa Code Camp

Spring 2013

# About Me

- Independent software consultant
- 13 years of Agile software development experience
- Data-driven desktop, server, and web apps
  - Web-based GIS data warehouse
  - Energy data ETL application
  - Global data management system
  - Intelligent lighting control systems

# About Me

- Education
  - BS in Computer Science
  - BA in Philosophy
    - Minor in Economics
    - Focus on Artificial Intelligence and Machine Learning
  - AS in MIS and Business Administration
- Training
  - Kimball Group Training in Data Warehousing
  - ESRI ArcGIS, ArcSDE, ArcGIS Server Training
  - Various data analysis and statistics courses

# Purpose

To provide a high-level overview of the tools the software industry uses to transform data into knowledge, specifically for the purpose of making better decisions

# Topics will include

- Sensor Data
- Transactional Data
- Semi-structured / Unstructured Data
- Data ETL
- Data Warehouses
- OLAP Cubes
- Statistical Analysis
- Data Visualization
- Data Exploration
- Data Mining
- Machine Learning
- Growing industry trends

# Audience

- Anyone who is interested in:
  - Transforming data into knowledge
  - Understanding the data value chain
  - Learning about the tools used in data analysis
- Session is 100-level
  - No previous technical knowledge is required
  - Presentation will be very high-level and very broad in scope

# Motivation

- Question:
  - Why do we want to transform data into knowledge?
- Answers:
  - To make better decisions
  - To understand our world
  - To make predictions about the future
  - Knowledge is power and power is awesome!

# Motivation

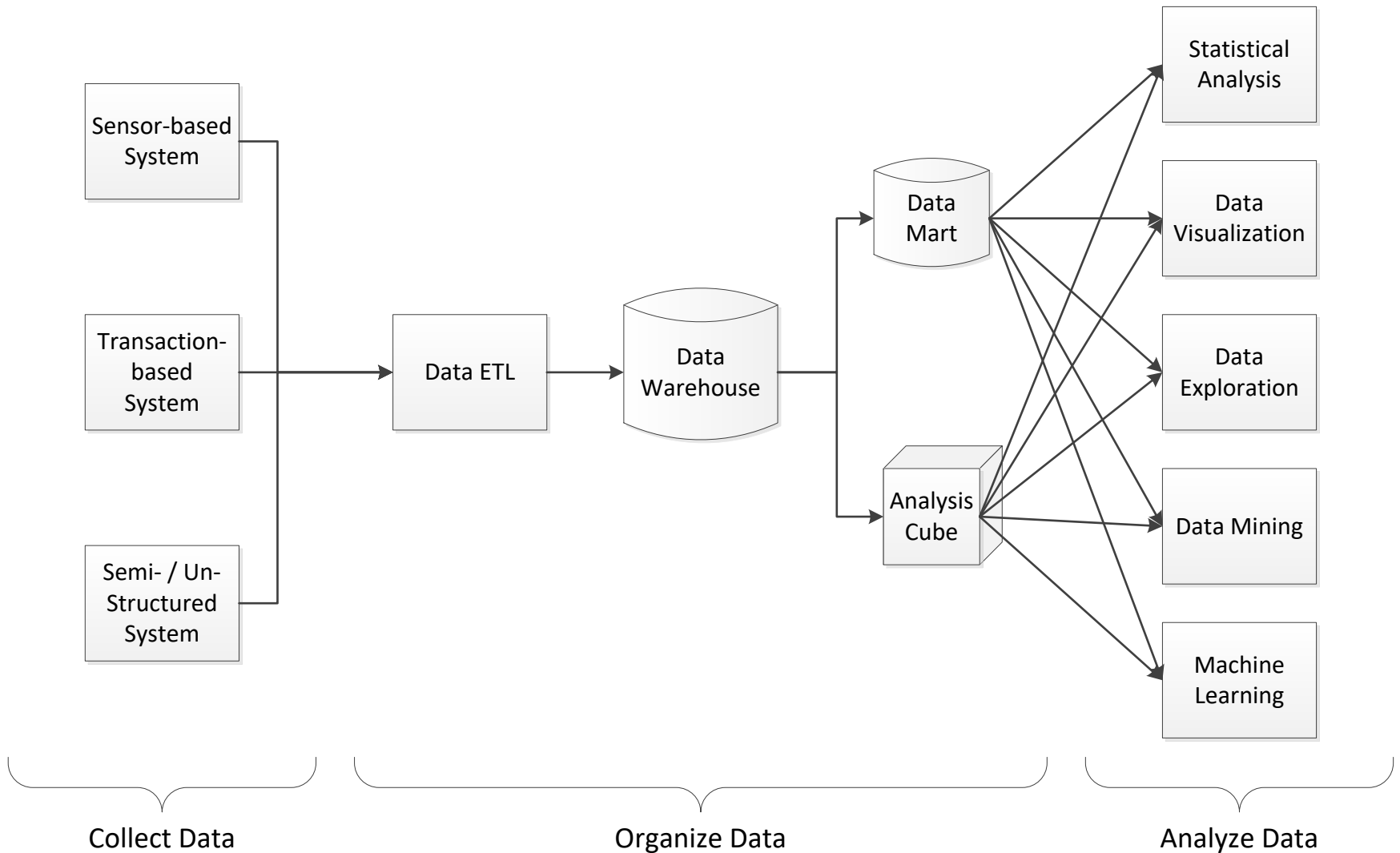
- Question:
  - What are data and knowledge?
- Answers:
  - **Data** are quantitative or qualitative values belonging to objects
  - **Knowledge** (for our purposes) is any model (mental or computational) that helps us make better decisions

# Motivation

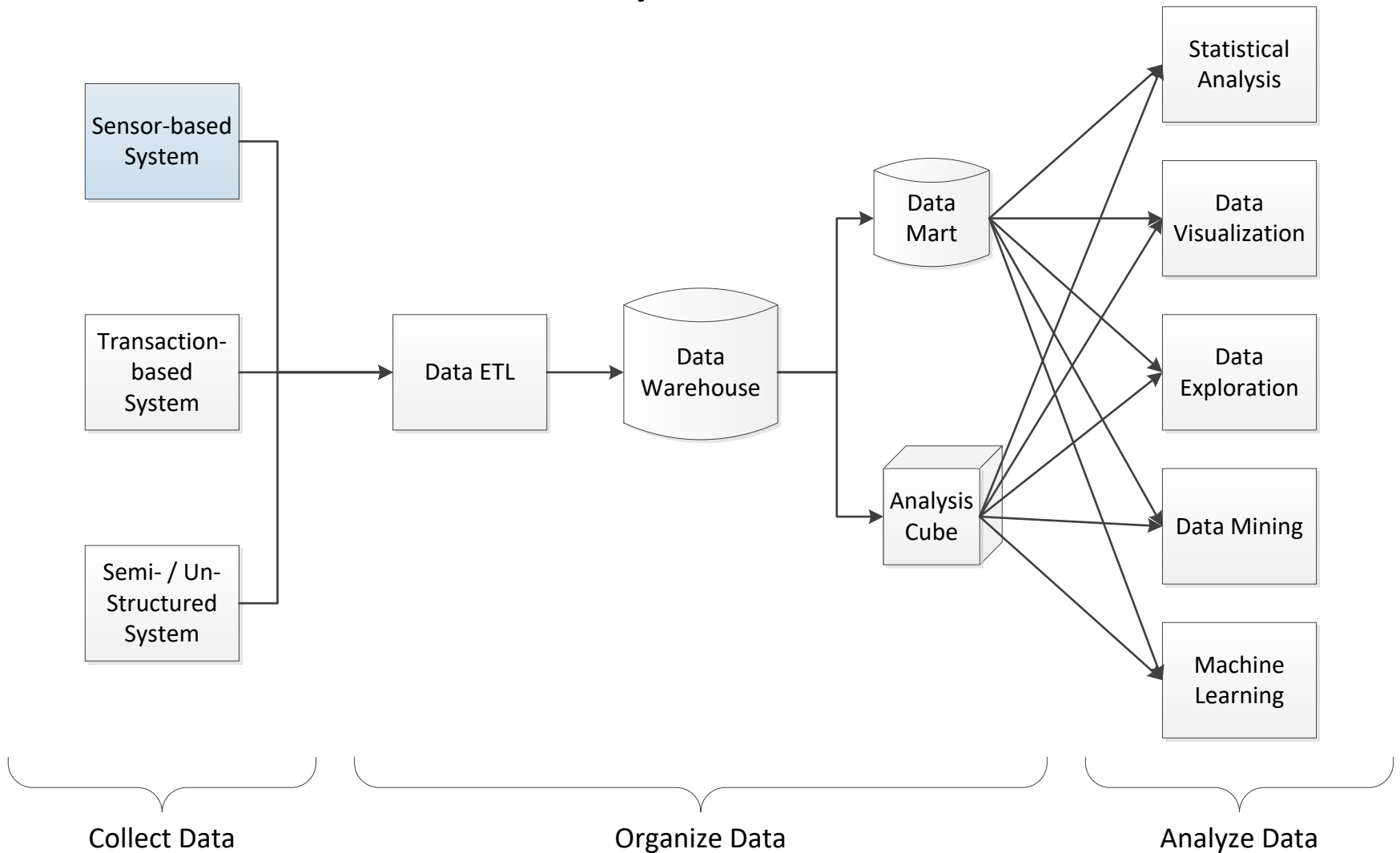
- Question:
  - How do we transform data into knowledge?
- Answer:
  1. Collect Data
  2. Organize Data
  3. Analyze Data



# Overview

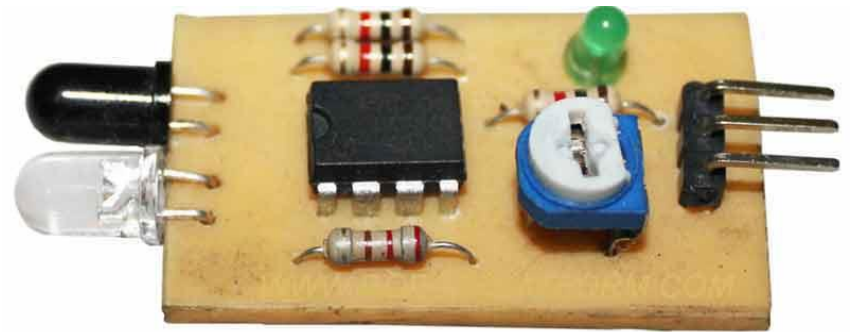


# Sensor-based Systems



# Sensor

- Converts an observable physical quantity into a representation that can be read by an observer (i.e., data)
- For example:  
Temperature => 98.6°F

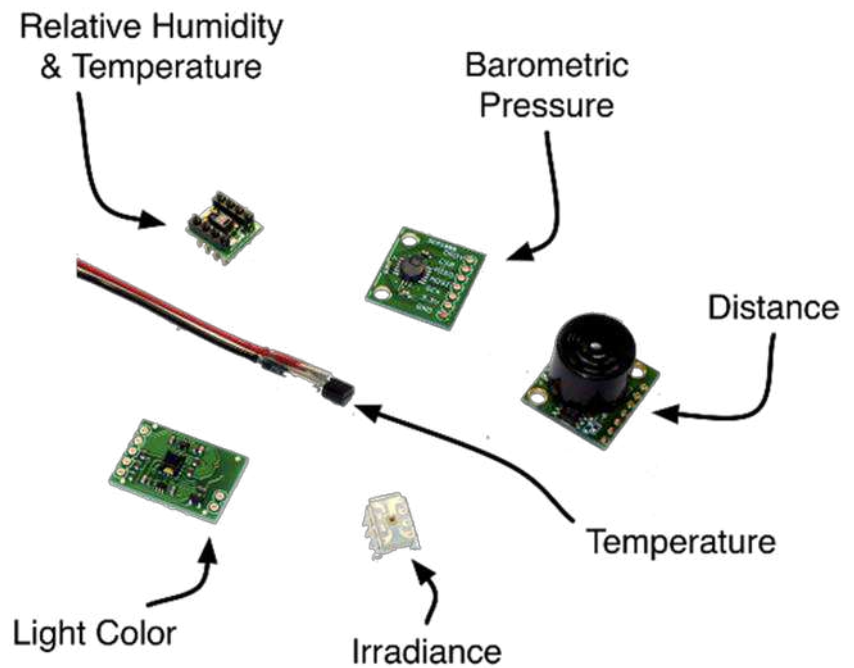


*IR SENSOR (TRANSCIEVER)*

Source: Wikipedia

# Types of Sensors

- Thermometer
- Rain Gauge
- Air Flow
- Smoke Detector
- Breathalyzer
- Fish Counter



Source: Sensorpedia

# Data Logger

- Device that connects to a series of sensors
- Reads the values of those sensors at a regular time interval
- Writes the values to a log file or database
- Typically connected to a network via ethernet
- Wireless data logger networks exist too



Source: Onset Computer Corporation

# Control System

- Device that connects to a series of sensors and actuators
- Reads sensors at regular intervals
- Sends commands to actuators
- Runs a program that maps sensor inputs to expected actuator outputs
- Examples: HVAC, Industrial Control



NAE55



NAE45

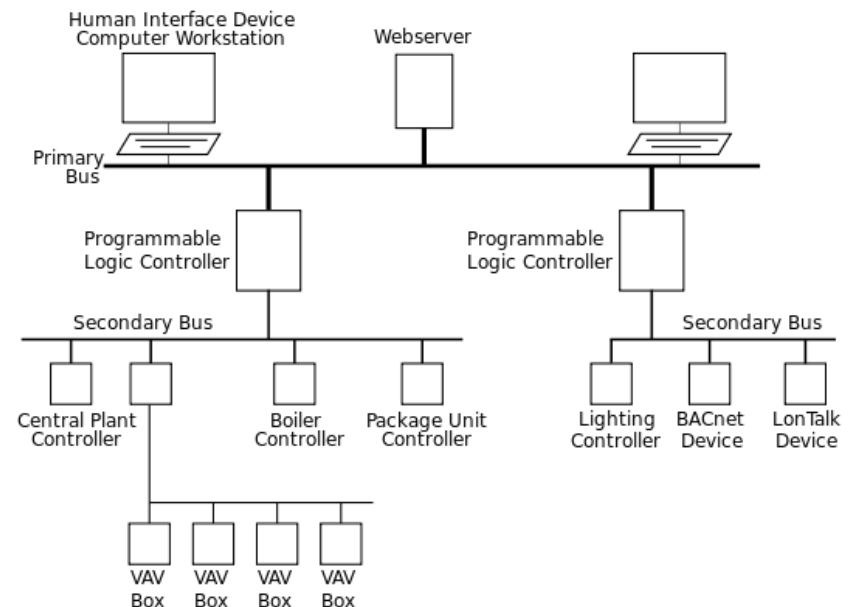


NAE85

Source: Johnson Controls

# Automation System

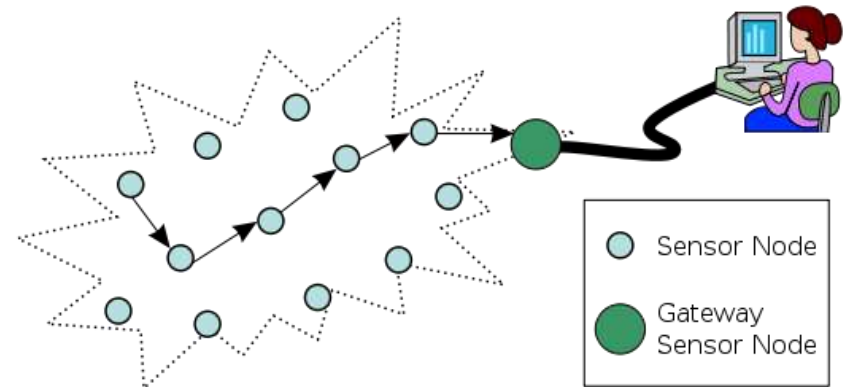
- Computer system that connects to control systems, data loggers, and sensors
- Reads data from sources
- Sends commands to control systems
- Runs programs to automate entire system
- Typically found in large buildings



Source: Wikipedia

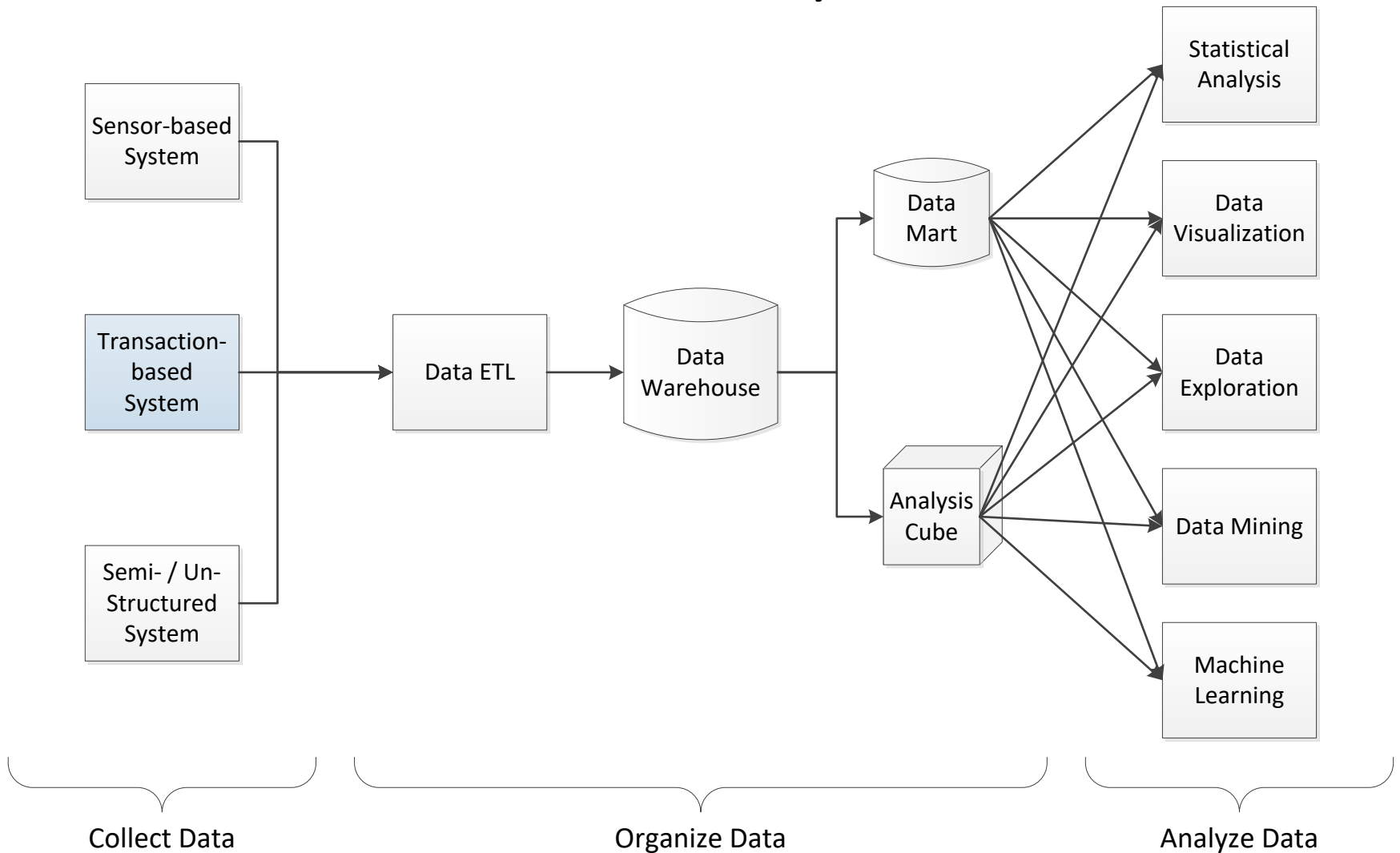
# Wireless Sensor Network

- A set of wireless sensor devices that can communicate with one another
- Data is forwarded through sensor nodes to a gateway node
- Uses peer-to-peer communication to maximize reliability and resilience



Source: Wikipedia

# Transaction-based Systems



# Transaction

- An exchange between two or more entities
- Occurs at a point in time (i.e., an event)
- Types of Transactions
  - Business Transactions
    - Banking Transactions
    - Sales Transactions
  - Communications
    - Tweets on Twitter feed

Sales Transactions

ID	Date	Customer	Product	Quantity
1	2012-10-27	John	Pizza	2
2	2012-10-27	John	Soda	2
3	2012-10-27	Jill	Salad	1
4	2012-10-27	Bob	Milk	1
5	2012-10-28	Sue	Soda	3
6	2012-10-28	Bob	Pizza	2
7	2012-10-28	Jill	Pizza	1
8	2012-10-28	Jill	Soda	3

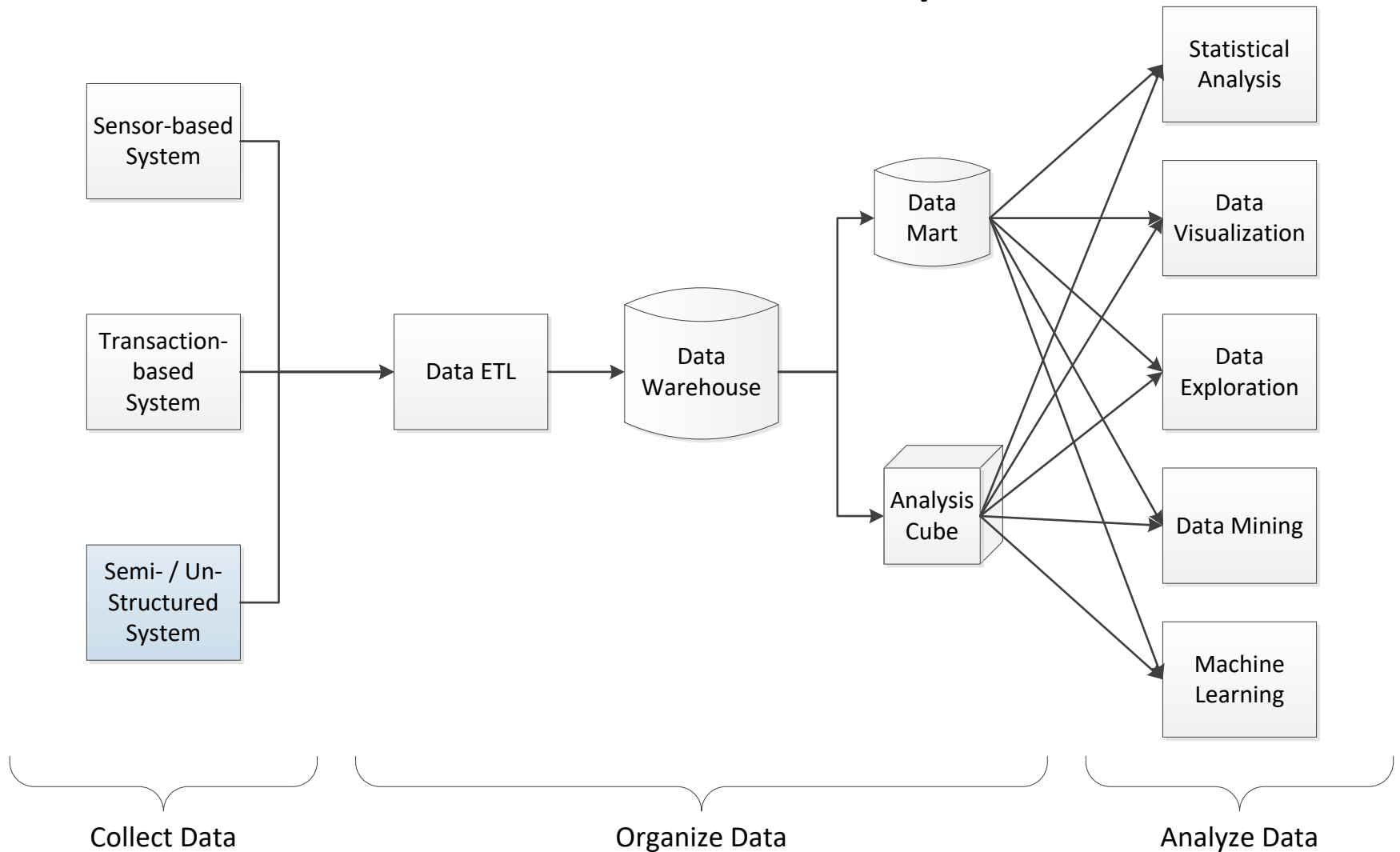
# Operational System

- Generic term used to describe any application or system used to process day-to-day transactions
- Typically focused on data entry of some kind
- Typically uses an OLTP (On-Line Transaction Processing) relational database to store data

# Operational System Examples

- Line-of-Business (LOB) Application
  - Data Entry Applications
  - Retail Point-of-Sale (POS) Terminal
- Enterprise Applications
  - Customer Relations Management (CRM)
  - Enterprise Resource Planning (ERP)
- Social Networking Applications
  - Facebook
  - Twitter

# Semi- / Unstructured Systems



# Semi-Structured Data

- Data that do not conform to the standard relational data model
- Data are self-describing
- Uses tags or markers separate semantic elements
- Typically hierarchical in nature

```
<body>
  <header>
    <h1>Matthew Renze</h1>
  </header>
  <nav>
    <ul>
      <li><a href="/Home.html">Home</a></li>
      <li><a href="/About.html">About</a></li>
      <li><a href="/Software.html">Software</a></li>
      <li><a href="/Events.html">Events</a></li>
      <li><a href="/Contact.html">Contact</a></li>
    </ul>
  </nav>
  <aside>
    <figure>
      Matthew is a software consultant whose
    </figcaption>
    </figure>
  </aside>
  <article>
    <header>
      <h1>Welcome</h1>
    </header>
    <p>I am currently in the process of getting my new webs
    <ul>
      <li><a href="/About.html">About</a> - information a
      <li><a href="/Software.html">Software</a> - a list
      <li><a href="/Events.html">Events</a> - upcoming an
      <li><a href="/Contact.html">Contact</a> - how to ge
    </ul>
  </article>
  <footer>
    <p>&copy; 2012 Matthew Renze</p>
  </footer>
</body>
```

# Semi-Structured Data

- HTML (Hyper-Text Markup Language)
- XML (eXtensible Markup Language)
- JSON (Java Script Object Notation)

```
<?xml version="1.0" encoding="utf-8"?>
<person>
  <firstName>John</firstName>
  <lastName>Smith</lastName>
  <age>25</age>
  <address>
    <streetAddress>21 2nd Street</streetAddress>
    <city>New York</city>
    <state>NY</state>
    <postalCode>10021</postalCode>
  </address>
  <phoneNumbers>
    <phoneNumber type="home">212 555-1234</phoneNumber>
    <phoneNumber type="fax">646 555-4567</phoneNumber>
  </phoneNumbers>
</person>
```

```
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021"
  },
  "phoneNumber": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "fax",
      "number": "646 555-4567"
    }
  ]
}
```

# Unstructured Data

- Data that do not conform to the standard relational data model
- Data are not self-describing (i.e., no tags)
- Technically, there is always structure, just not structure in a relational data model sense



A bunny with a pancake on his head

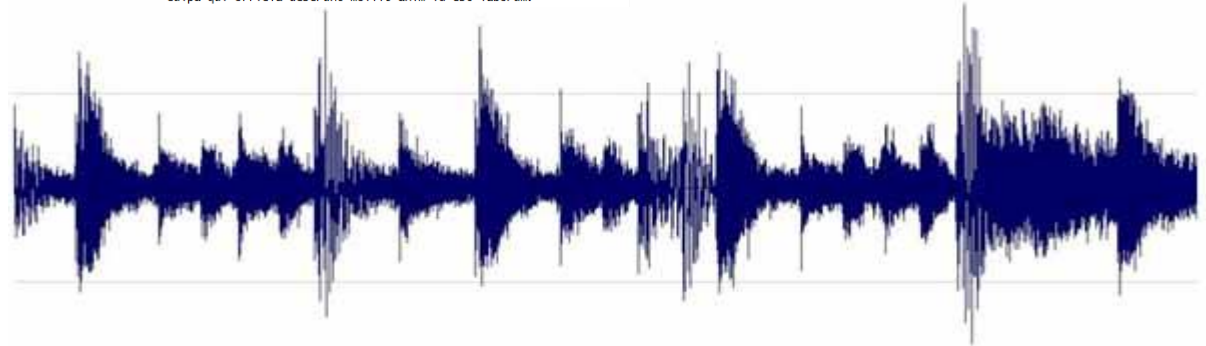
# Unstructured Data

- Text
- Images
- Audio
- Video

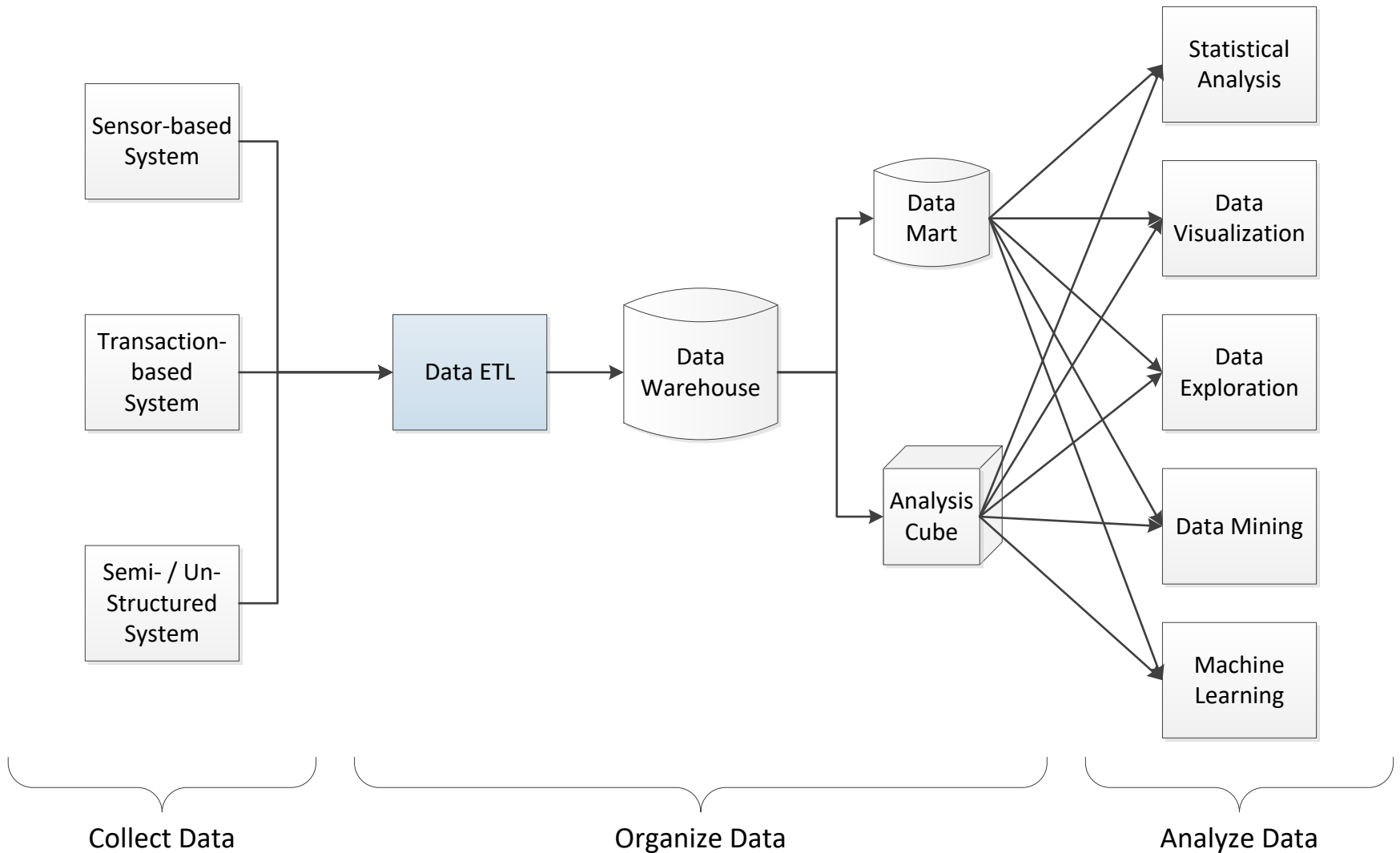
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.



# Data ETL



# Data ETL

- ETL stands for Extract, Transform, and Load
- Process of extracting, transforming, and loading data from a source data system into a data warehouse
- Occurs in a Data Staging Area
- Typically done as a nightly routine

# Data Extraction

- Process of extracting data from the operational system
- Data can be extracted:
  - Directly from operational database (via SQL)
  - Indirectly from operational database export files
  - Indirectly from an Operational Data Store (ODS)
    - An ODS is typically a replica or mirror of the operational database; however, the term can mean many things

# Feature Extraction

- When performing ETL on unstructured data, features must be extracted from the raw data
- Examples:
  - Extracting word counts from a text document
  - Extracting text from a document image (OCR)
  - Extracting faces from images
- Purpose is for dimensionality reduction

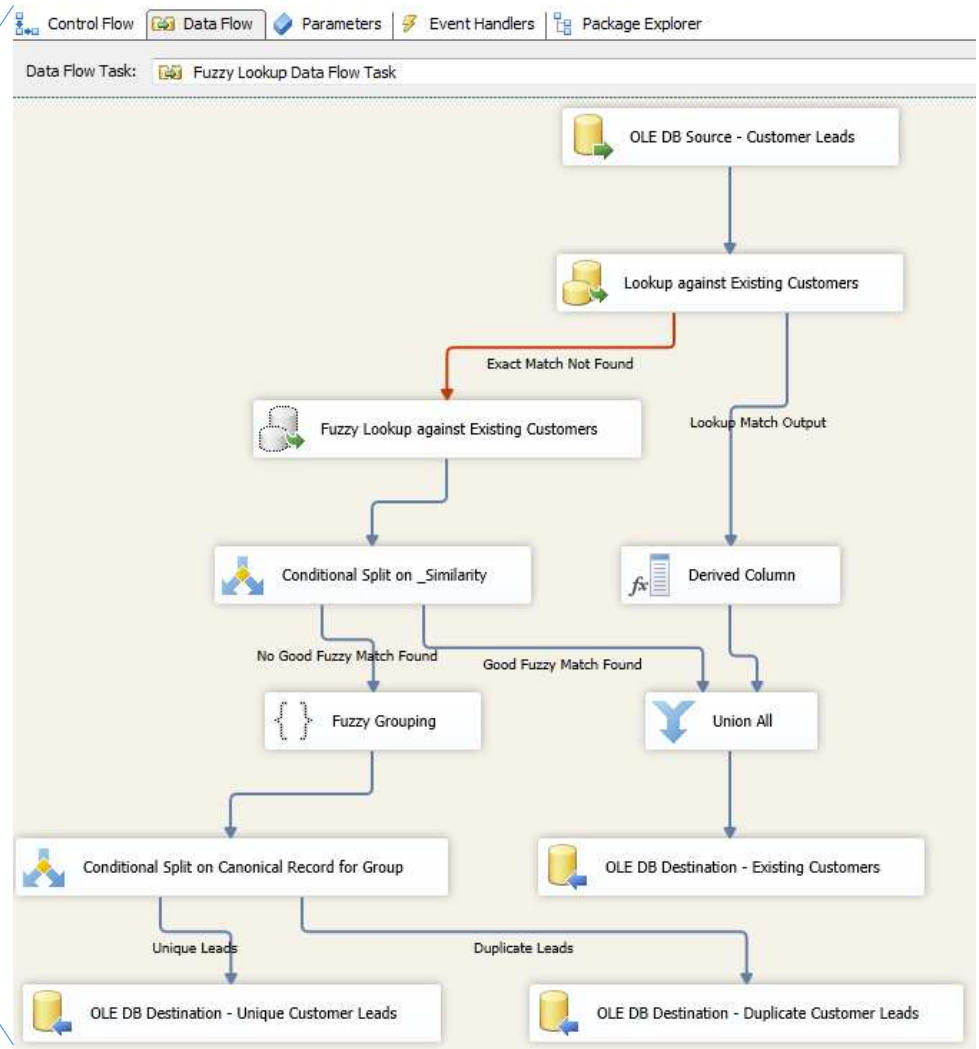
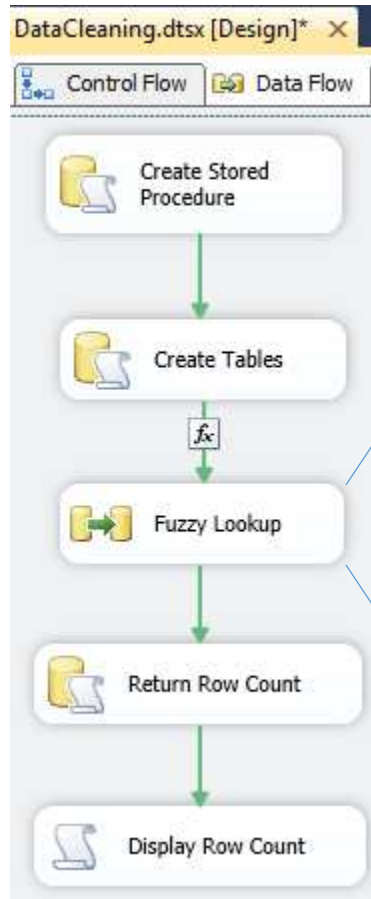
# Data Transformation

- Process of transforming operational data into a form best suited for reporting and analysis
- Typical data transformations include:
  - Projections (i.e., selecting a subset of columns)
  - Decoding (e.g., “M” to “Male”, “F” to “Female”)
  - Joining data from multiple data sources
  - Performing table lookups
  - Calculating (e.g.,  $\text{amount} = \text{price} * \text{quantity}$ )
  - Transposing (e.g., pivoting columns into rows)
  - Cleaning Data (i.e., cleaning up bad data)

# Data Loading

- Process of load data into the data warehouse or analysis cube
- Typically done as a bulk insert

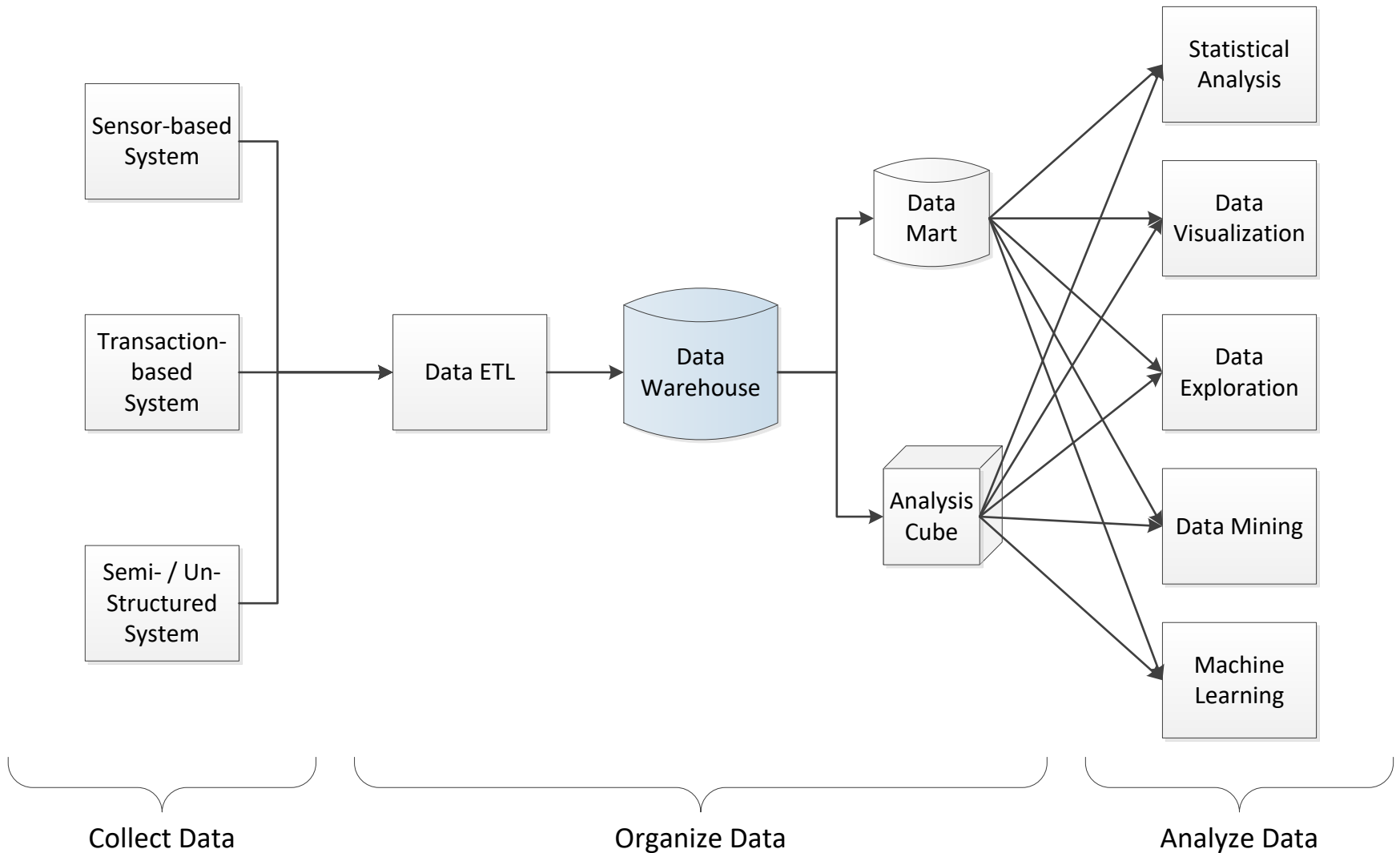
# ETL Package



# Popular ETL Software

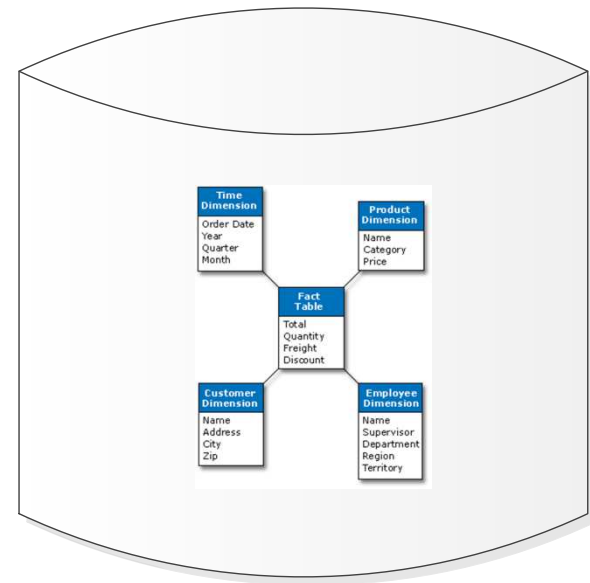
- IBM – Information Server (Datastage)
- Informatica – PowerCenter
- Microsoft – SQL Server Integrations Services (SSIS)
- Oracle – Data Integrator / Warehouse Builder
- SAP Business Objects – Data Integrator
- SAS – Data Integration Studio
- Clover ETL (open source)
- Many companies still hand code their ETL in SQL

# Data Warehouse



# Data Warehouse

- Database optimized for reporting and analysis
- Typically integrates data from several operational data sources



# Two Schools of Thought

## Inmon

- Top-down design
- Entity-relational model
- Normalized (3NF)



## Kimball

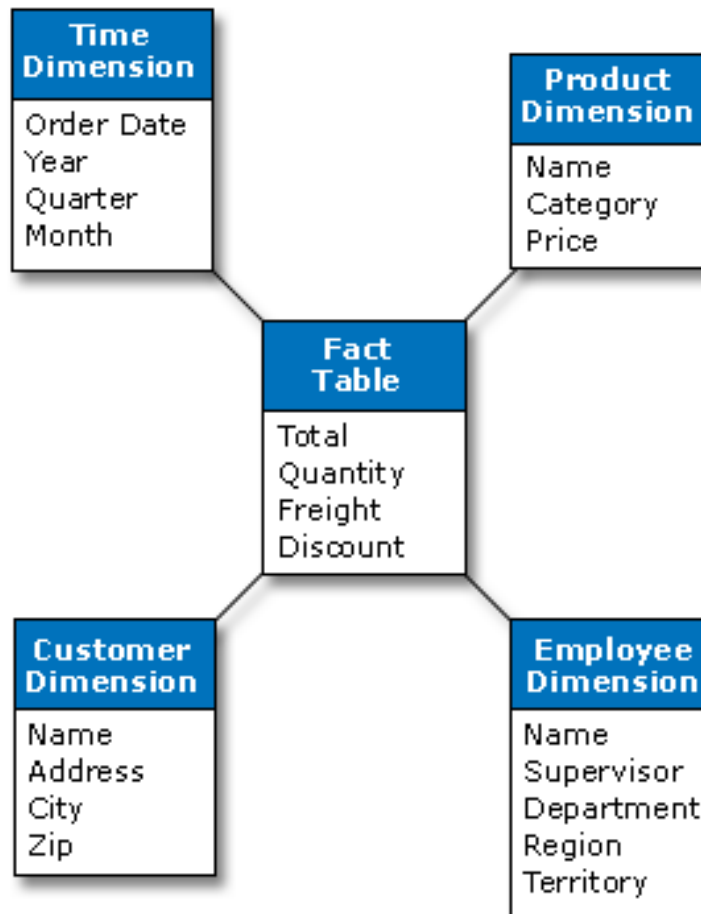
- Bottom-up design
- Dimensional model
- Denormalized (Star Schema)



# Dimensional Model

- Fact
  - aka: Measures
  - A value or measurement
  - Example: Price = \$100, Temperature = 98.6°
- Dimension
  - Give context to facts
  - Categorizes facts into non-overlapping regions
  - Example: Date, Customer, Region

# Star Schema



Source: Microsoft

# Operational System vs. Data Warehouse

## Operational System

- Optimized for writing data quickly and maintaining data integrity
- Normalized to minimize duplication of data via 3NF (3<sup>rd</sup>-Normal Form)
- Typically queried in very narrow and specific ways

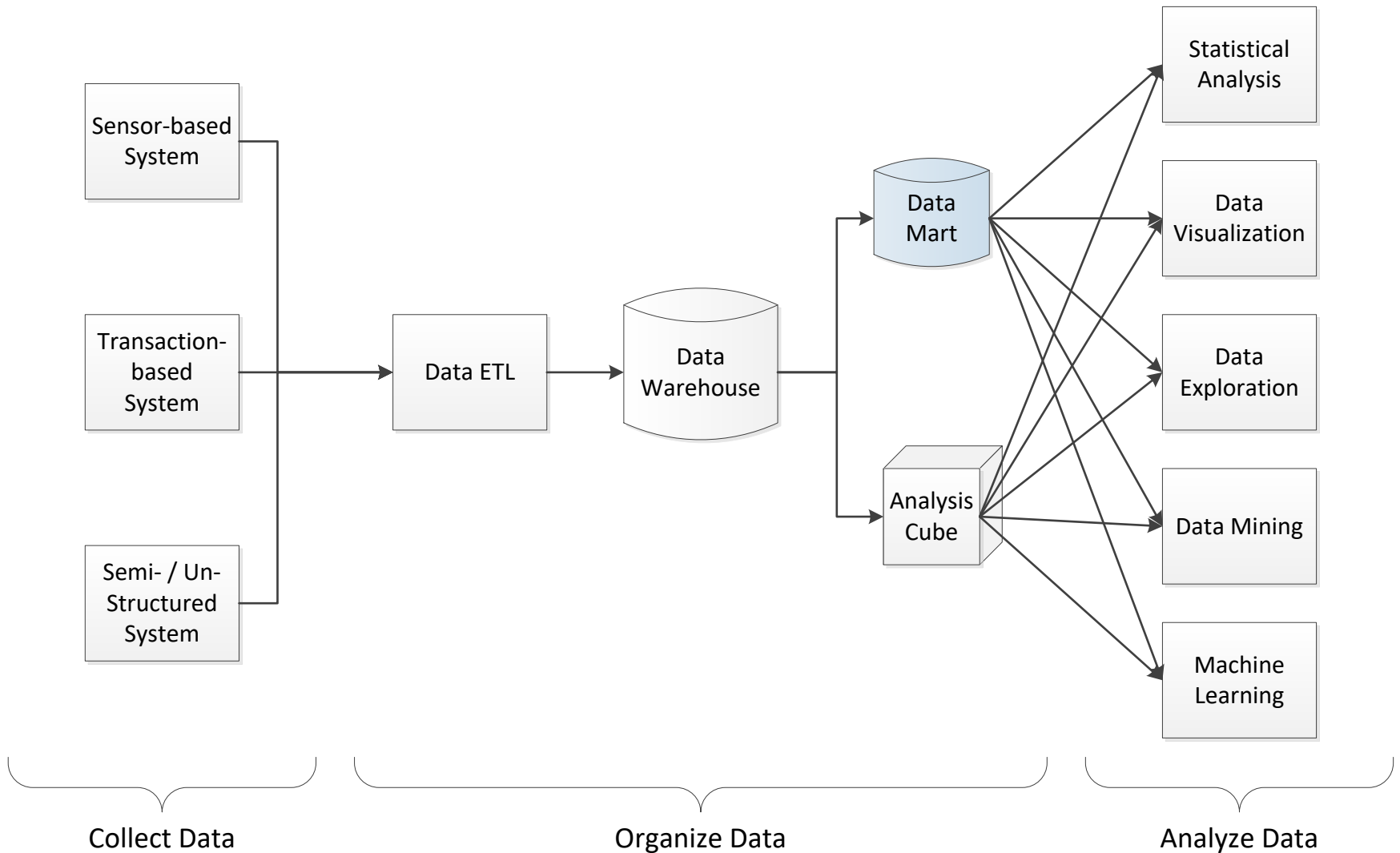
## Data Warehouse

- Optimized for reading data quickly for reporting and analysis
- Denormalized to maximize speed of analysis via Star Schema
- Queried in very broad and unexpected ways

# Data Warehouse Providers

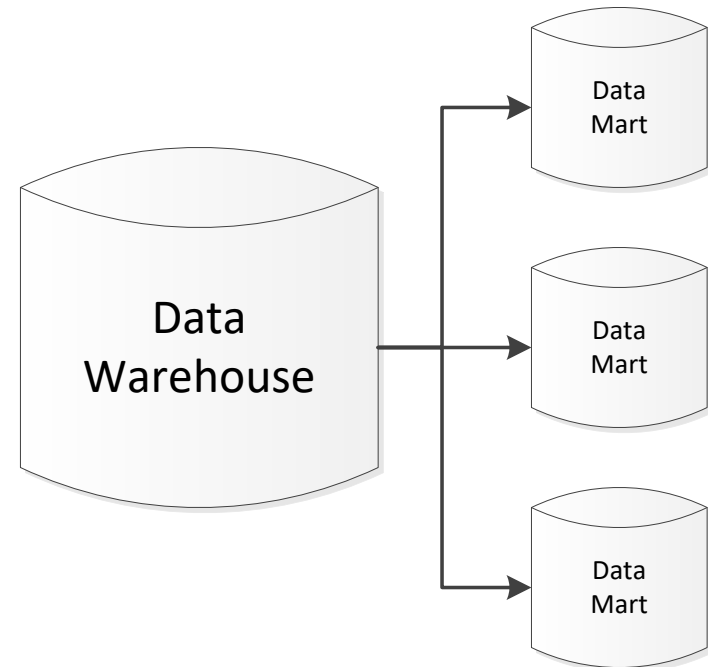
- IBM – Infosphere and Netezza
- Microsoft – SQL Server
- Oracle – Database 11g and Exadata
- SAP – Sybase
- Teradata – Active Enterprise Data Warehouse

# Data Mart

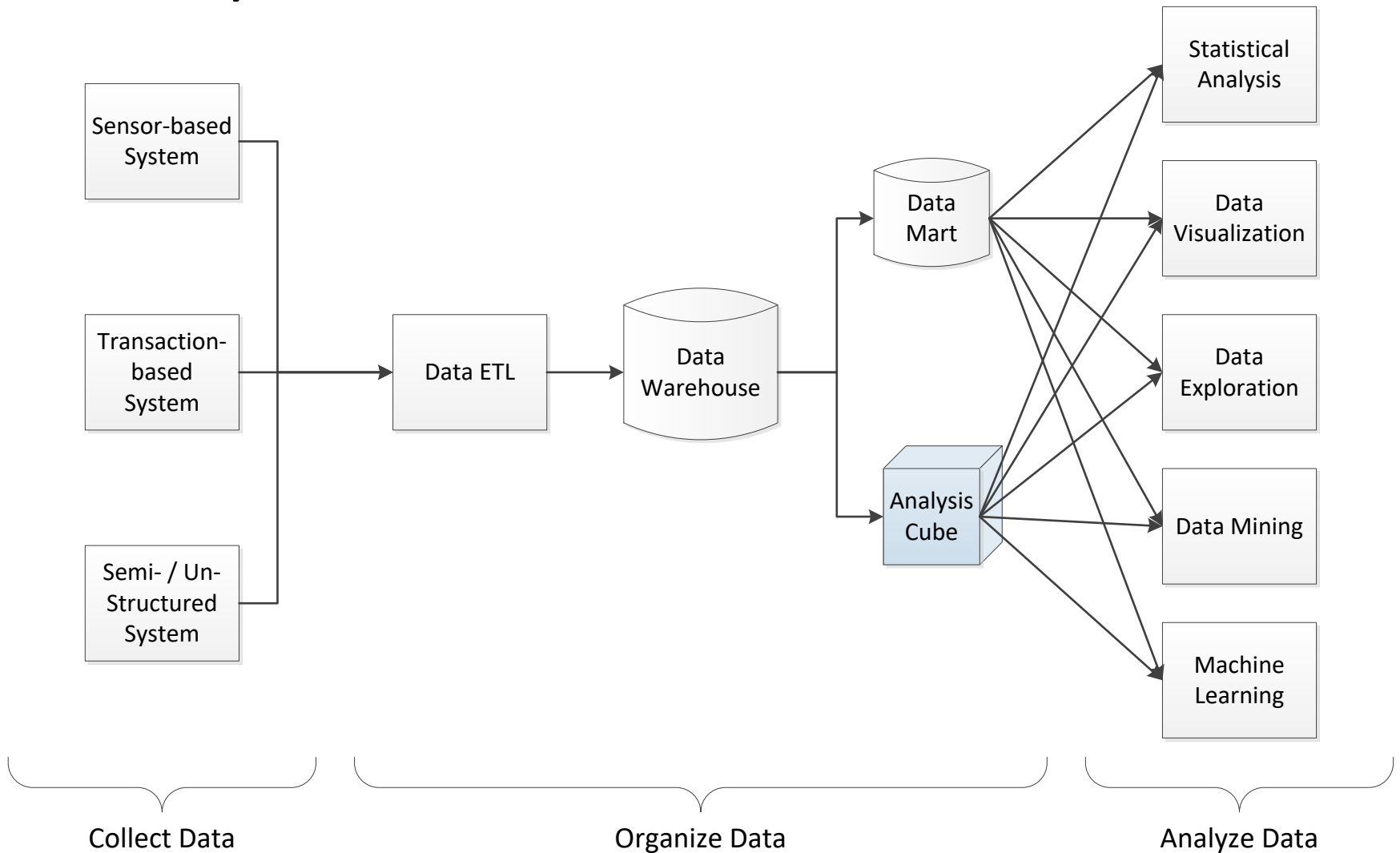


# Data Mart

- Provide users with access to the data in the data warehouse
- Subset of the data warehouse oriented to a specific department or team

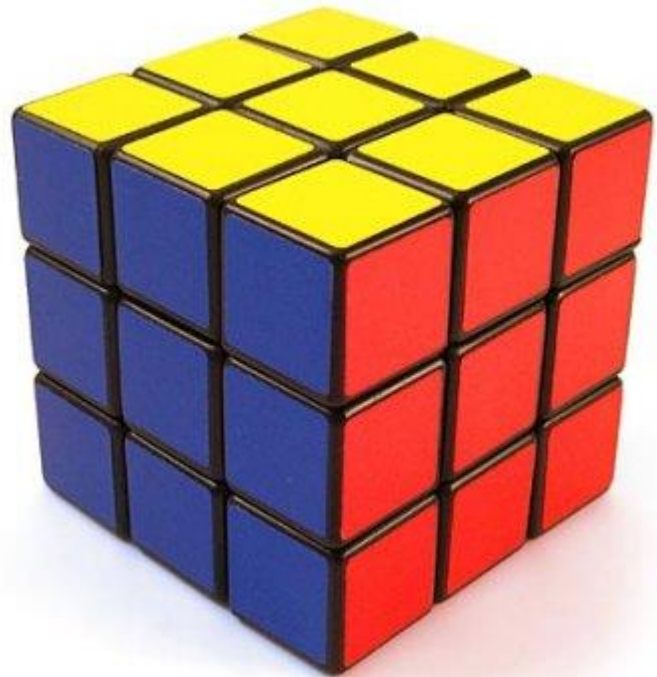


# Analysis Cube



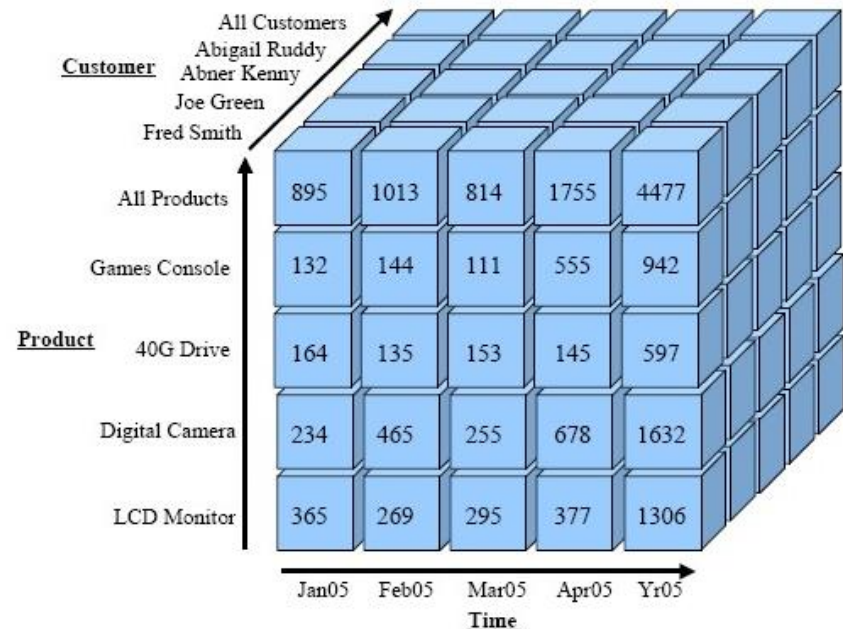
# Analysis Cube

- Multi-dimensional array
- Extremely fast for analysis operations
- Like a spreadsheet but with more than two dimensions
- Both server-based cubes or desktop cubes
- aka: OLAP Cube, Multi-Dimensional Cube



# Analysis Cube

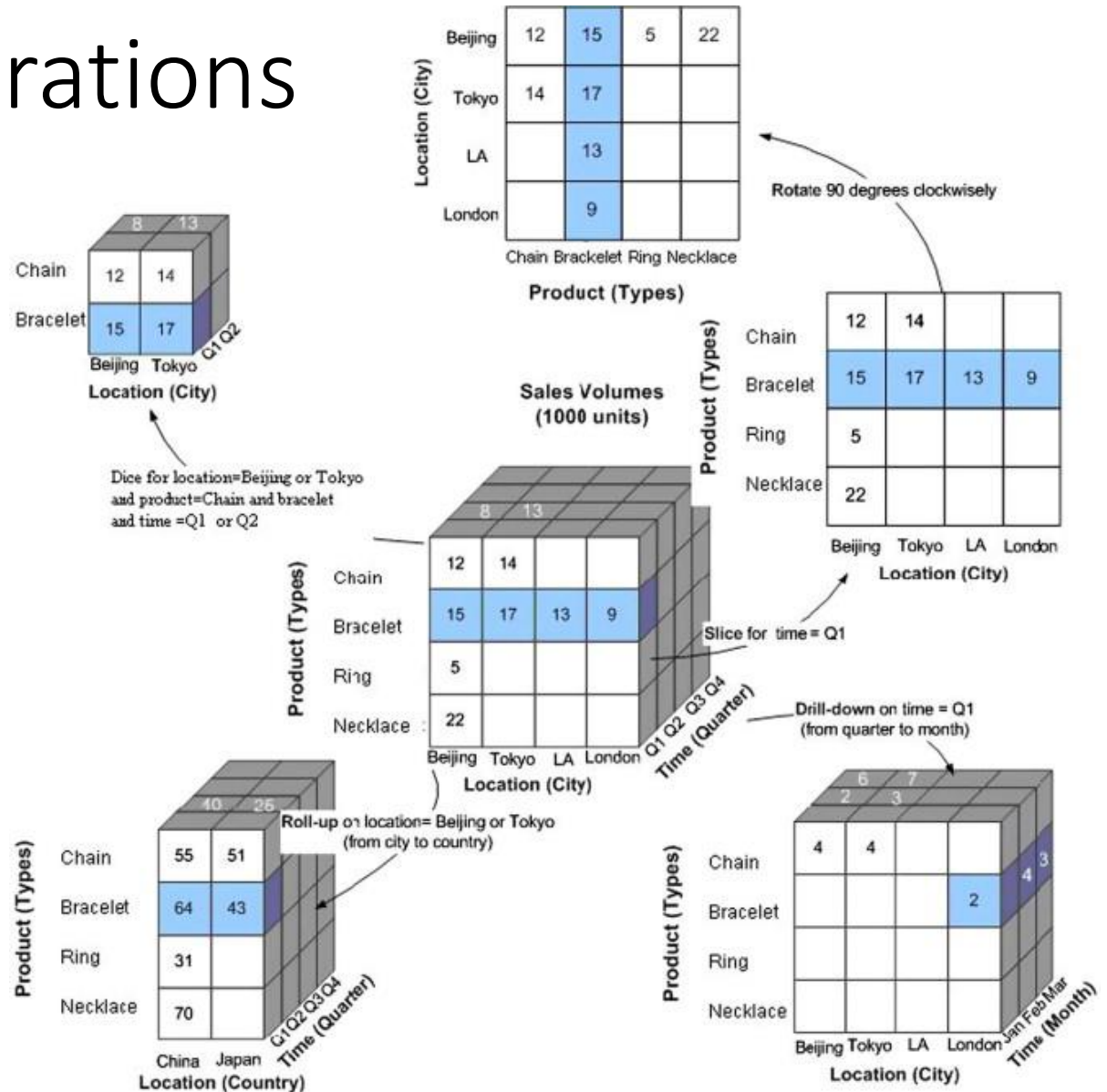
- Facts stored in each cell
- Dimensions of data map to dimensions of cube
- Dimensions can be hierarchically organized
- Typically more than three dimensions (hypercube)



Source: [http://gerardnico.com/wiki/database/oracle/oracle\\_olap](http://gerardnico.com/wiki/database/oracle/oracle_olap)

# Cube Operations

- Slice
- Dice
- Pivot
- Drill Down
- Roll Up



# Analysis Cube Query Languages

- MDX – Multi-Dimensional Expressions
- XMLA – XML for Analysis

Sample MDX Query:

```
select
    { [Measures].[Store Sales] } on columns,
    { [Date].[2011], [Date].[2012] } on rows
from Sales
where ( [Store].[USA].[CA] )
```



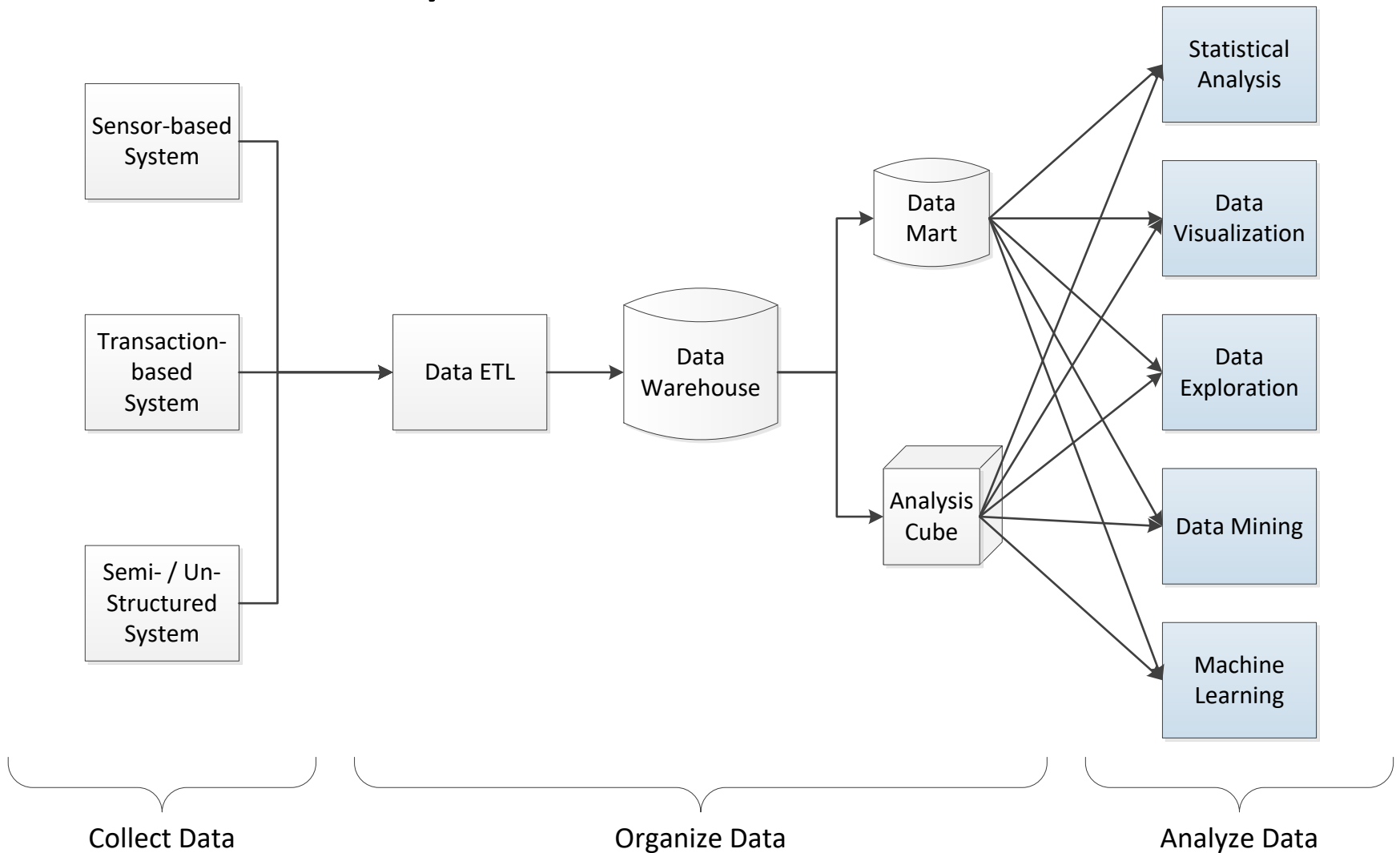
Query Results:

	Store Sales
2011	32663.74
2012	65303.44

# Analysis Cube Providers

- IBM – Cognos TM1
- Microsoft – SQL Server Analysis Services (SSAS)
- MicroStrategy – Intelligence Server
- Oracle – Hyperion Essbase / OLAP Option
- SAP – NetWeaver BW (InfoCubes)
- SAS – OLAP Cube Studio
- Pentaho – Mondrian (open source)

# Data Analysis

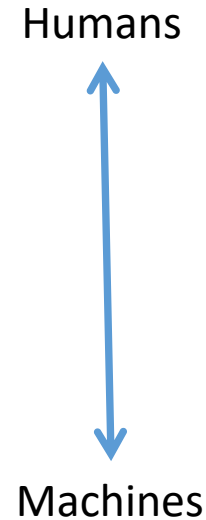


# Data Analysis

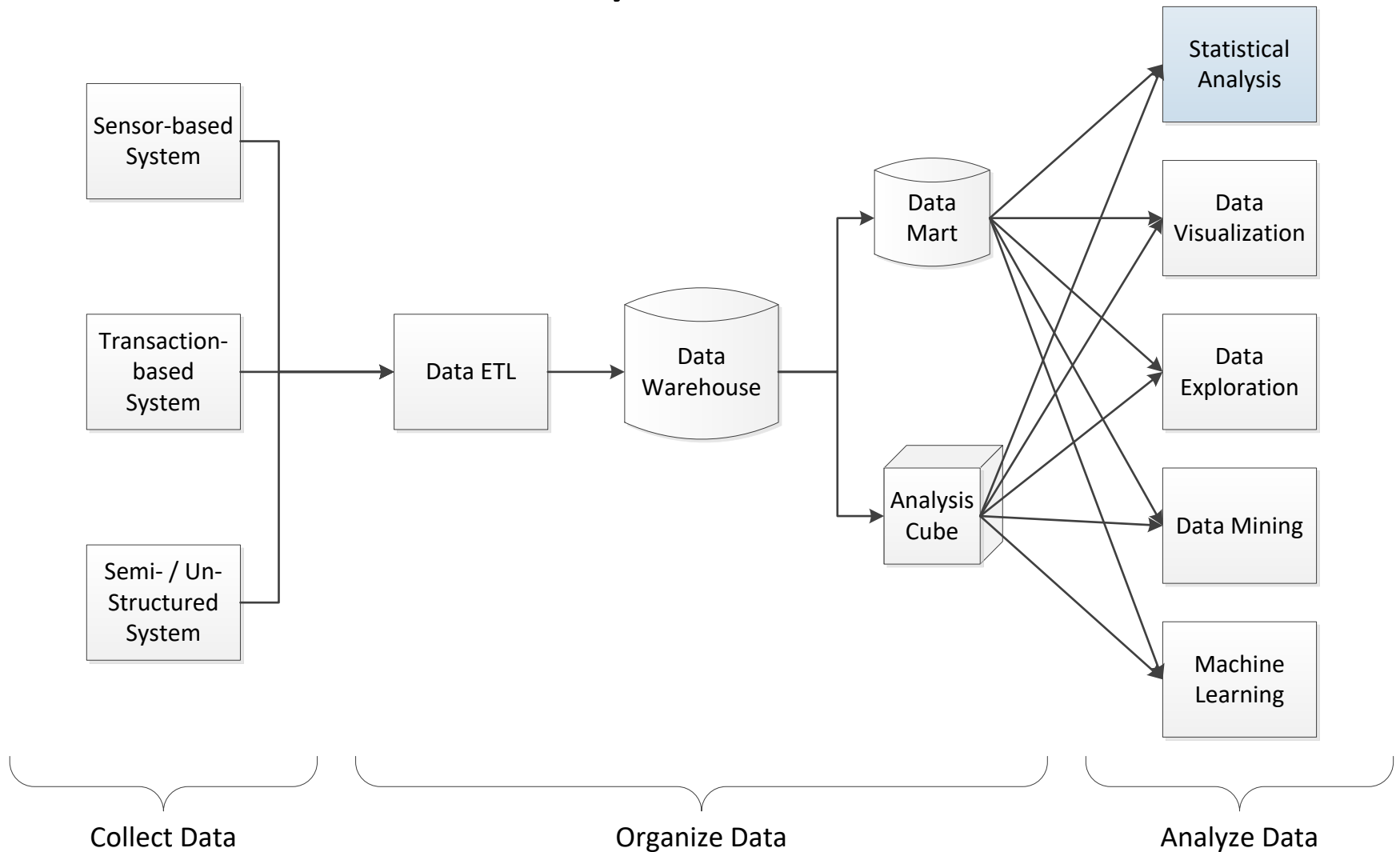
- Process of decomposing data into constituent parts in order to study it and extract new information
- Common buzz words:
  - Decision Support Systems
  - OLAP (On-line Analytical Processing)
  - Business Intelligence
  - Data Analytics

# Methods of Data Analysis

- Statistical Analysis
- Data Visualization
- Data Exploration
- Data Mining
- Machine Learning

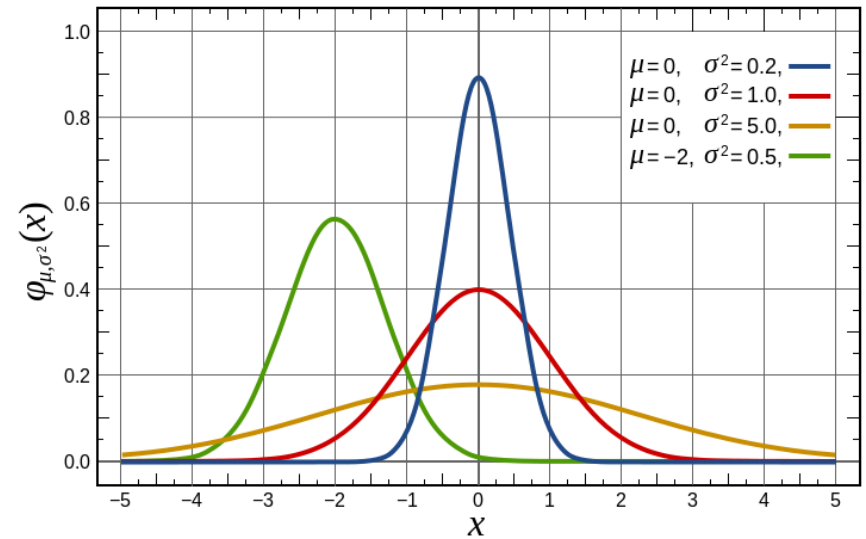


# Statistical Analysis



# Statistical Analysis

- Data analysis using statistical methods



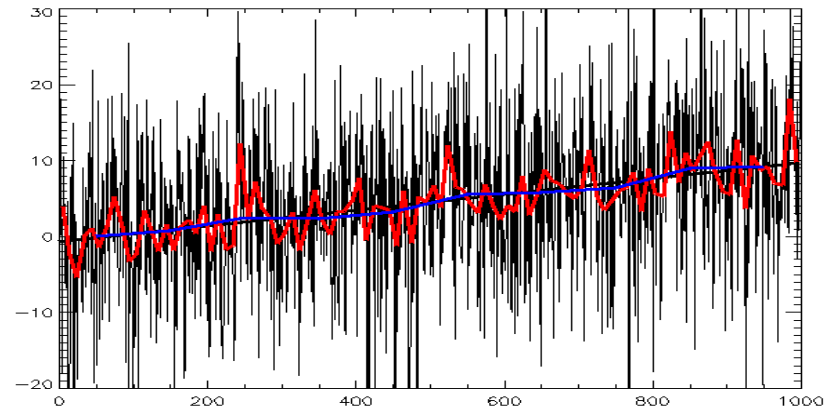
Source: Wikipedia

# Types of Statistical Analysis

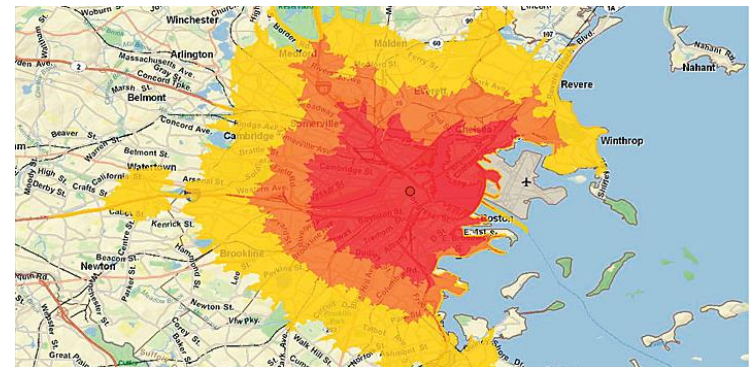
- Descriptive
  - Describes data in quantitative or qualitative ways
- Inferential
  - Draws conclusions about a population from a sample
- Exploratory
  - Discovers knowledge from data by exploring it
- Predictive
  - Makes predictions about new data given existing data

# Types of Statistical Analysis

- Time Series Analysis
  - Analysis of data changing over time
- Geo-Spatial Analysis
  - Analysis of data that has a geographical or spatial properties



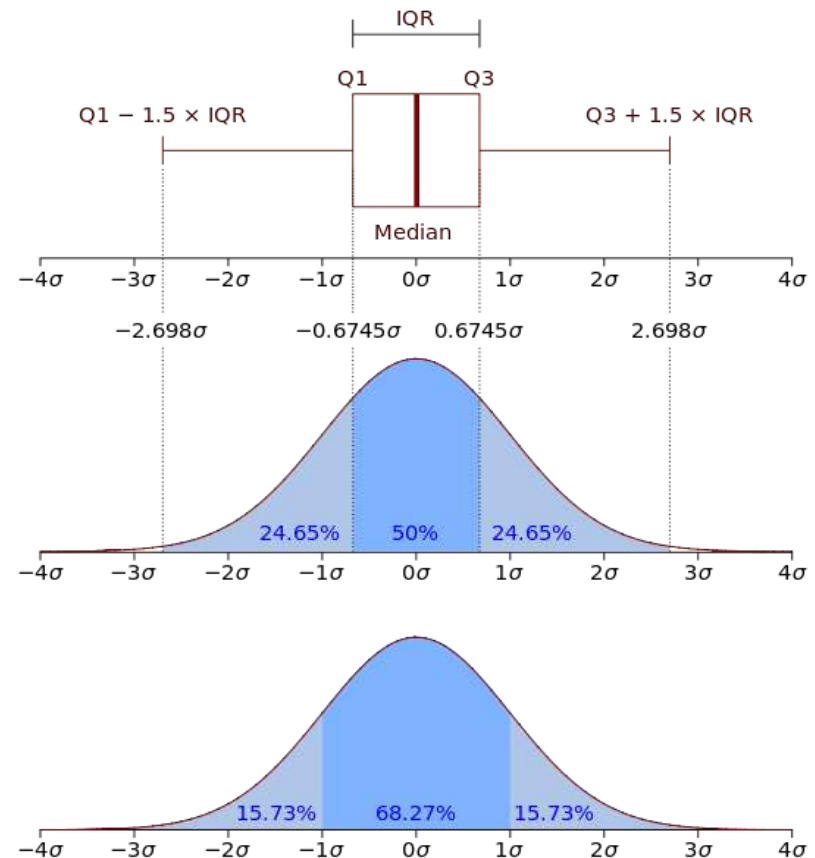
Source: Wikipedia



Source: ESRI

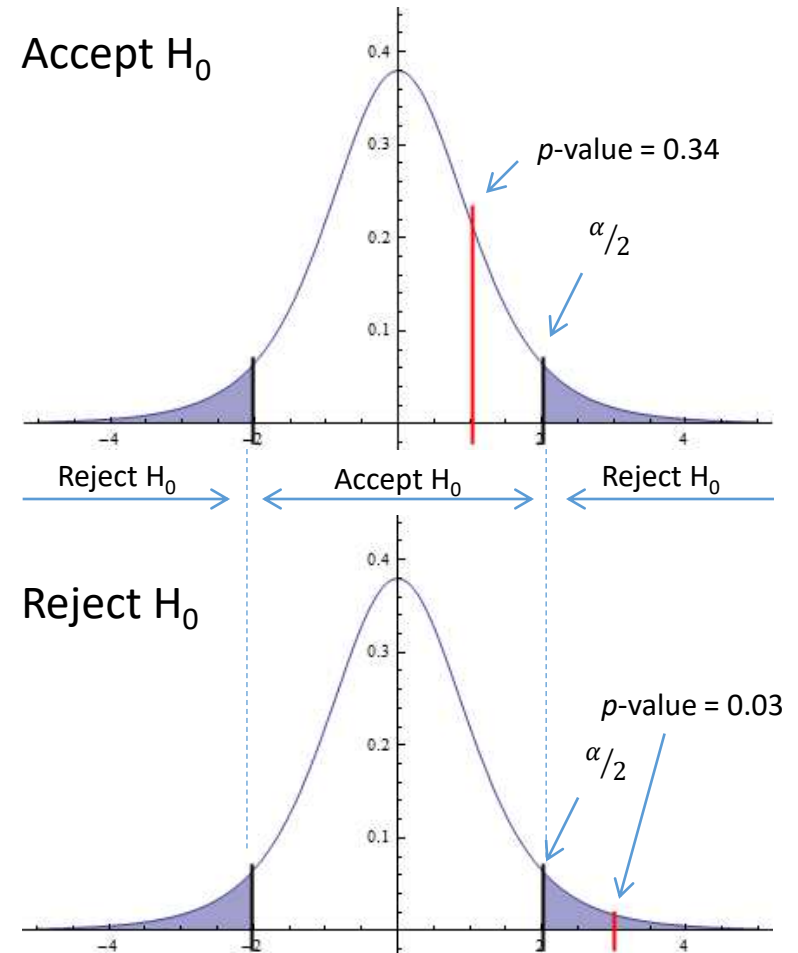
# Descriptive Statistics

- Univariate Analysis
  - Central Tendency: Mean, Median, Mode
  - Dispersion: Min, Max, Range, Quantiles, Variance, Standard Deviation
- Bivariate Analysis
  - Relationship: Covariance, Correlation coefficient
- Multivariate Analysis



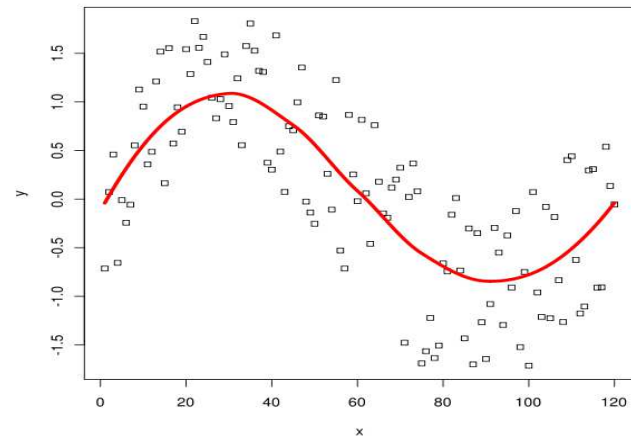
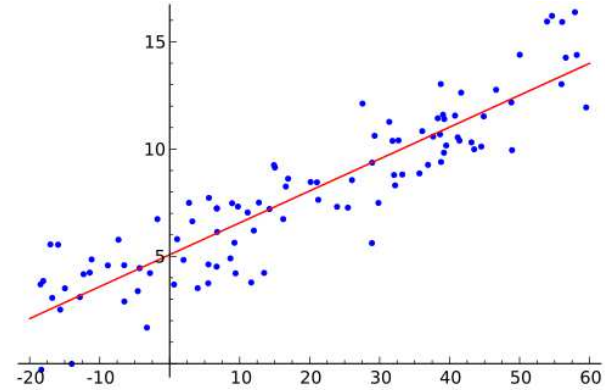
# Hypothesis Testing

- Used to determine statistical significance of an observation
- Start with a question
- State the null hypothesis ( $H_0$ ) and an alternate hypothesis ( $H_1$ )
- Either accept  $H_0$  or reject  $H_0$  based on  $p$ -value
- Need sufficient sample data to make inferences about the population in general



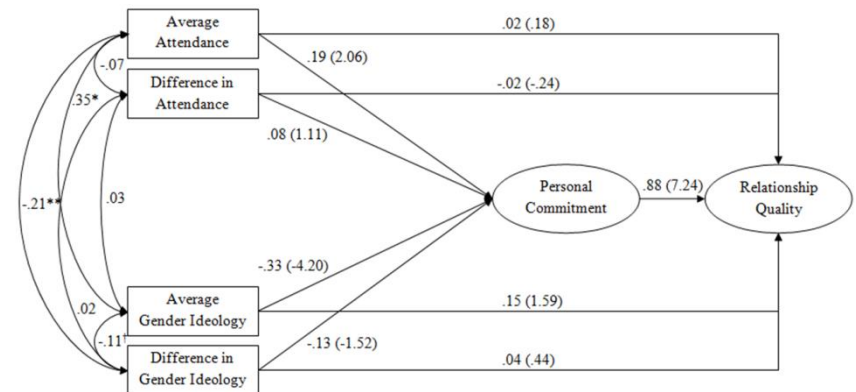
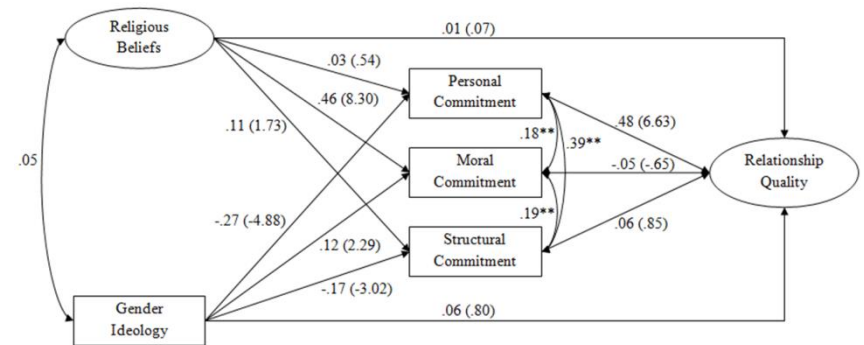
# Regression Analysis

- Technique for estimating the relationship between two or more variables
- Result is a function
- Types of Regression:
  - Linear Regression
  - Non-Linear Regression
  - Logistic Regression
  - Multivariate Regression



# Statistical Modeling

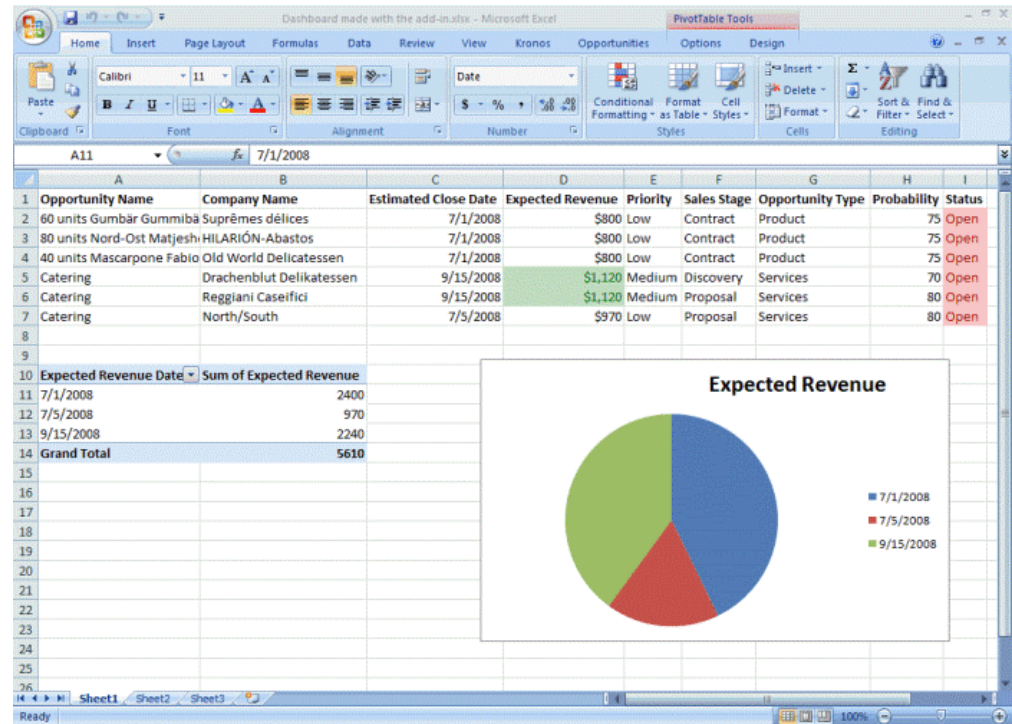
- Statistical model of variables and their relationships
- Multiple uses:
  - Explanatory models
  - Predictive models
- Types of models:
  - Linear Model
  - Bayesian Model
  - Multi-level Model
  - Structural Equation Model



Source: Karen Bittner

# Spreadsheet

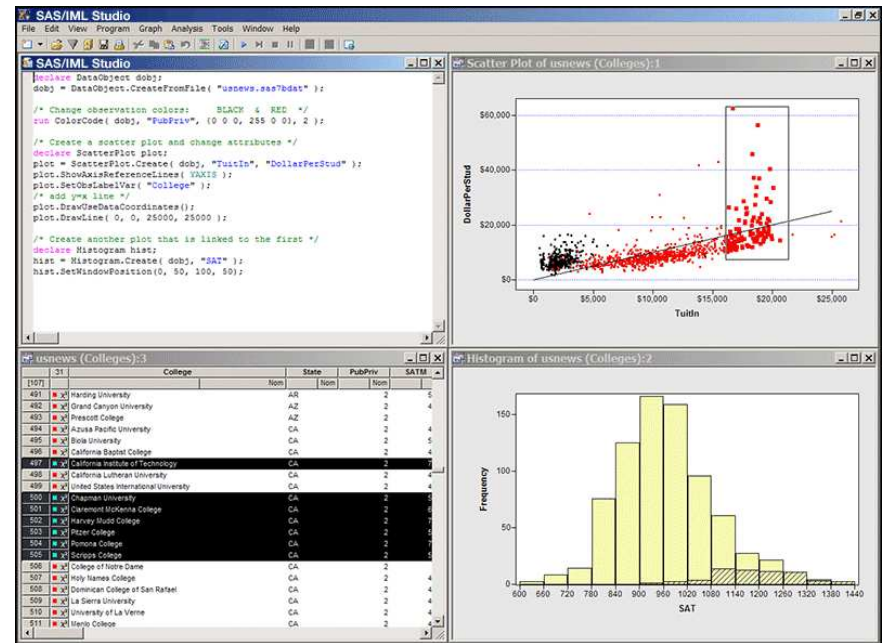
- Most popular tool for basic statistical analysis
- Plug-ins available for more rigorous statistical analysis



Source: Microsoft

# Statistical Analysis Software

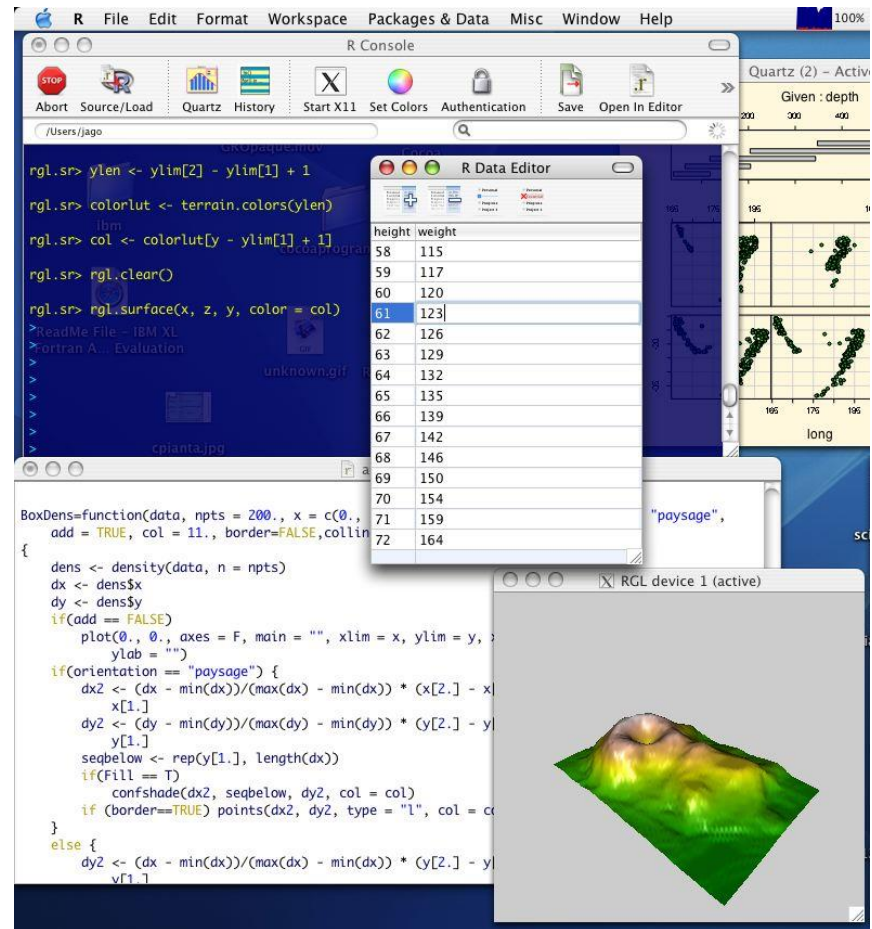
- Very powerful tool for data analysis
- Steep learning curve
- Graphical User Interface
- Command-line Interface
- Provides both analytic and graphical methods
- Popular software:
  - SAS, SPSS, Minitab, Stata



Source: SAS

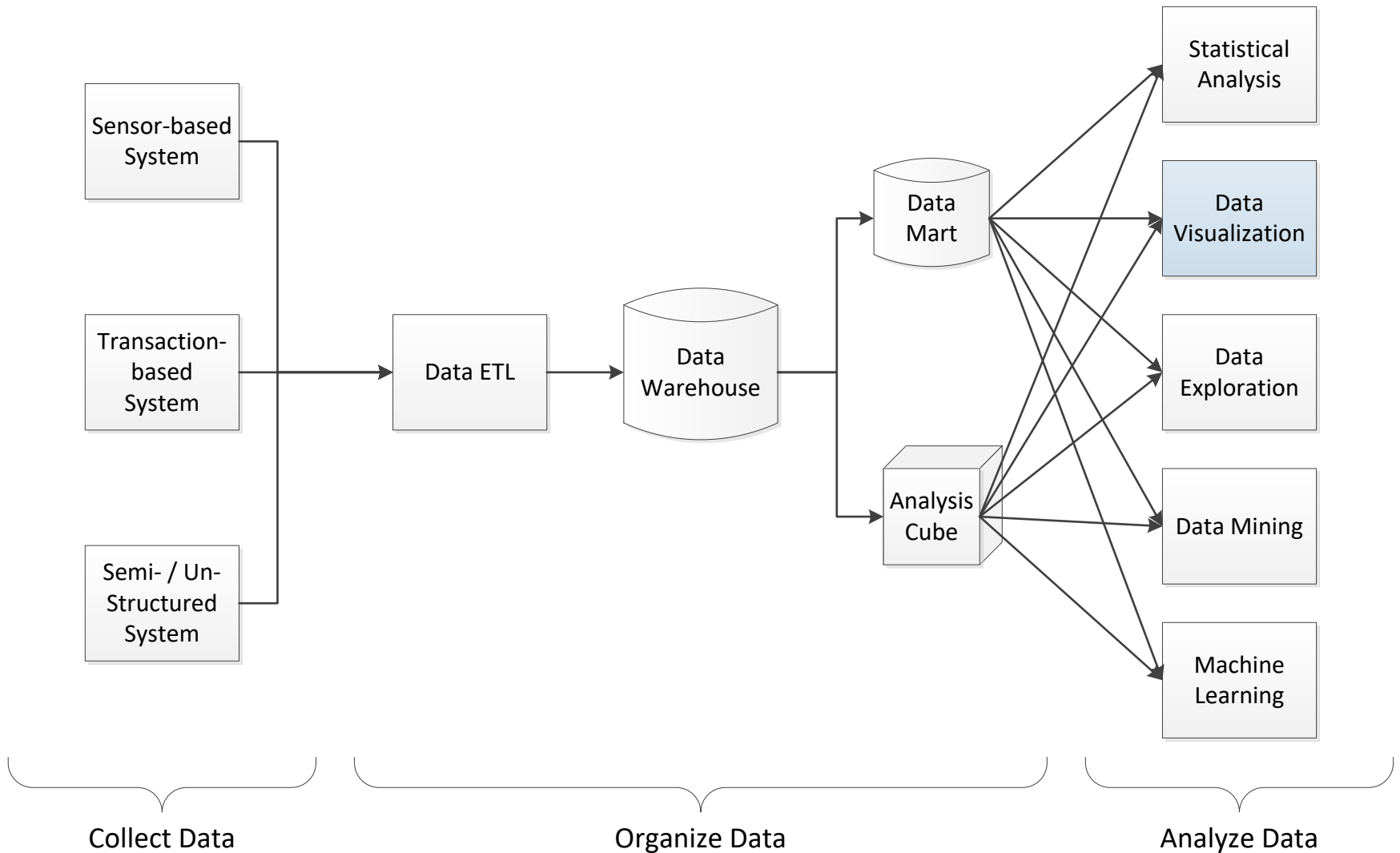
# Statistical Programming Language

- Most powerful type of data analysis tool
- Steepest learning curve
- Uses a command-line interpreter (like Python)
- Popular languages:
  - SAS
  - R



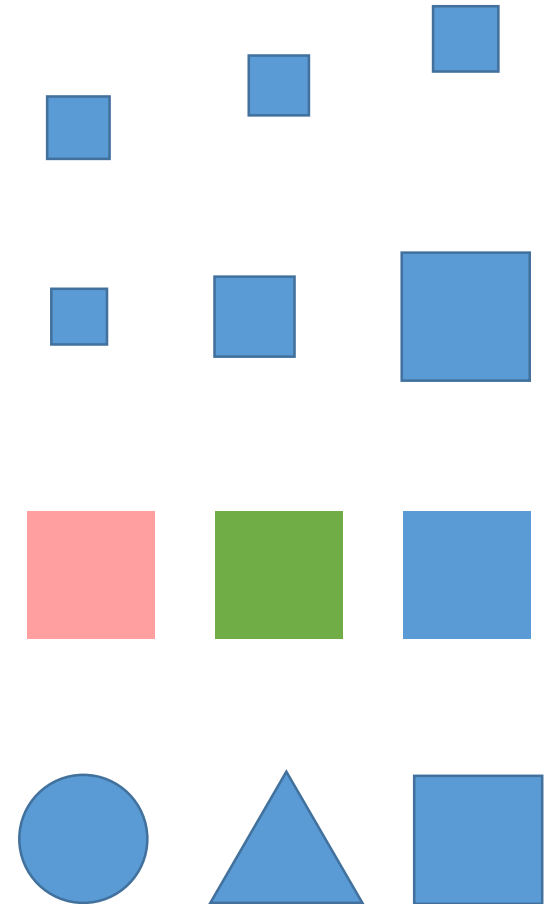
Source: The R Project

# Data Visualization



# Data Visualization

- Representation of data via visual means
- Human brain is exceptionally good at visual pattern recognition
- Map dimensions of data to visual qualities
  - Location
  - Size
  - Color
  - Shape



# Tabular

- Table
  - Data organized into rows and columns
  - Rows represent items
  - Columns represent properties of items
- Cross-tab Matrix
  - Data organized in a 2-dimensional matrix
  - Cells contain aggregate values scoped to intersection of row/column

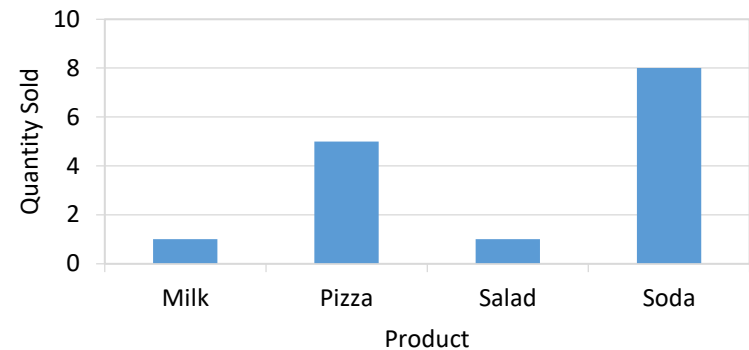
ID	Date	Customer	Product	Quantity
1	2012-10-27	John	Pizza	2
2	2012-10-27	John	Soda	2
3	2012-10-27	Jill	Salad	1
4	2012-10-27	Bob	Milk	1
5	2012-10-28	Sue	Soda	3
6	2012-10-28	Bob	Pizza	2
7	2012-10-28	Jill	Pizza	1
8	2012-10-28	Jill	Soda	3

	Milk	Pizza	Salad	Soda	Total
Bob	1	2	1	0	4
Jill	0	1	0	3	4
John	0	2	0	2	4
Sue	0	0	0	3	3
Total	1	5	1	8	15

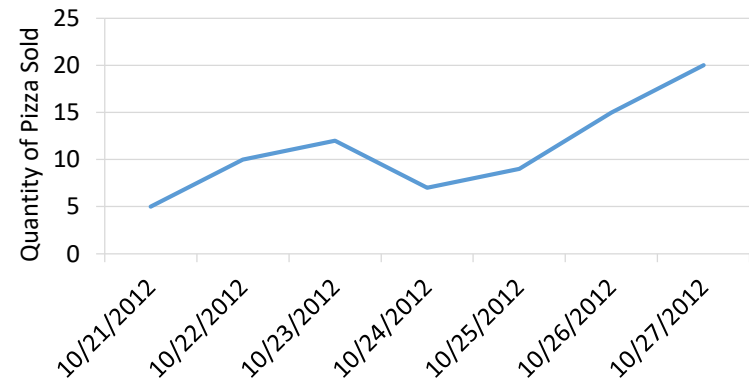
# Charts and Graphs

- Visual representation of multivariate data
- Discrete and continuous data representations
- Common examples:
  - Bar/Column Chart
  - Line Graph
  - Pie Chart
  - Scatter Plot

Sales by Product

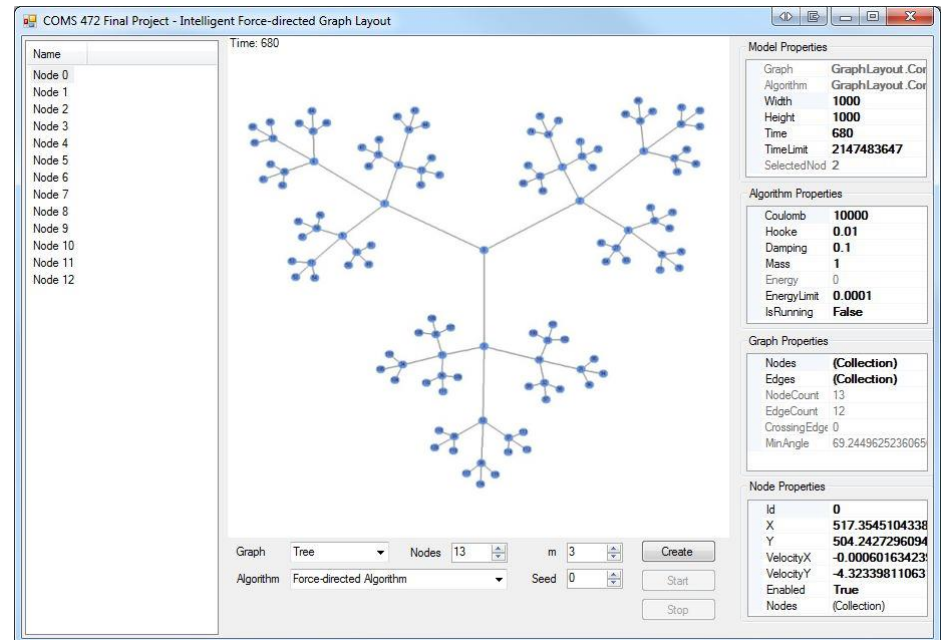


Daily Pizza Sales



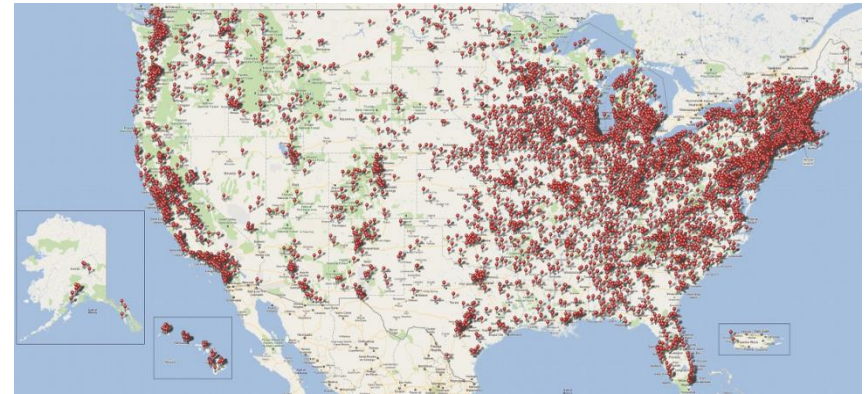
# Tree and Graph Visualization

- Visual representation of tree and graph data structures
- Common examples:
  - Tree map
  - Hierarchy chart
  - Graph diagram
  - Network diagram

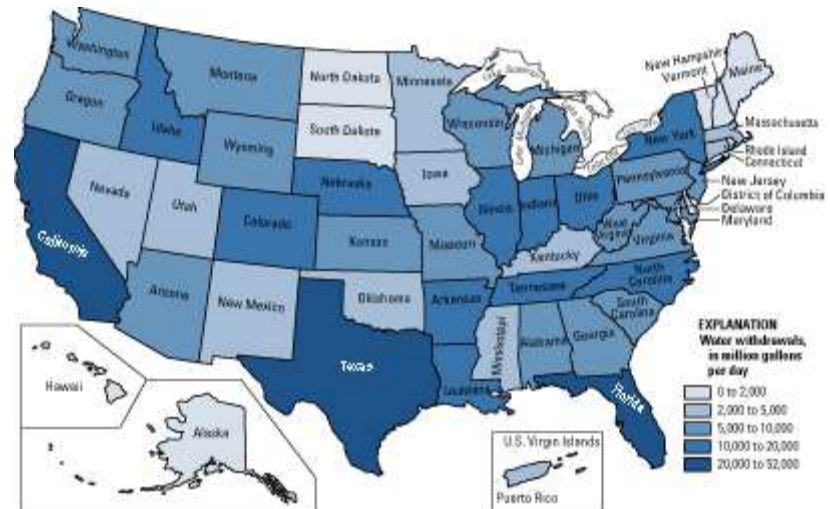


# Data Maps

- Visual representation of geo-spatial data
- Types of data maps:
  - Dot Map
    - Dots represents location and spatial distribution of data
  - Choropleth
    - Colors represent values of data within a boundary




Source: USDA



Source: Wikipedia

# Report

- Provide formatted data to users
- Can contain text, tables, and graphs
- Usually printable and exportable

					
123 Main Street Any City, USA					
Site Type: Radius					
Latitude: 38.8828 Longitude: -77.1175 Radius: 1.0 miles					
	Census 2000	2005	2010	2005-2010 Change	2005-2010 Annual Rate
Population	31,400	33,753	35,478	1,725	1%
Households	14,740	16,184	17,159	975	1.18%
Median Age	34.3	36.3	39.1	2.8	1.5%

Census 2000 Households by Income and Age of Householder							
	< 25	25 - 34	35 - 44	45 - 54	55 - 64	65 - 74	75+
HH Income Base	982	4,343	3,202	2,490	1,406	893	1,404
<\$10,000	108	159	55	101	42	66	116
\$10,000 - \$14,999	46	62	89	48	25	87	153
\$15,000 - \$24,999	105	161	157	89	94	106	205
\$25,000 - \$34,999	84	314	159	122	75	123	213
\$35,000 - \$49,999	147	605	365	145	125	109	187
\$50,000 - \$74,999	252	1,107	787	575	255	161	255
\$75,000 - \$99,999	98	555	515	360	301	123	77
\$100,000 - \$149,999	77	942	661	509	301	86	135
\$150,000 - \$199,999	33	268	209	245	64	26	14
\$200,000+	32	160	185	273	124	26	27
Median HH Income	\$50,121	\$58,137	\$73,920	\$83,870	\$80,436	\$47,744	\$36,356
Average HH Income	\$62,168	\$82,546	\$90,366	\$107,647	\$93,443	\$59,969	\$49,524

Percent Distribution							
	< 25	25 - 34	35 - 44	45 - 54	55 - 64	65 - 74	75+
HH Income Base	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
<\$10,000	11.0%	3.7%	1.7%	4.1%	3.0%	7.4%	8.3%
\$10,000 - \$14,999	4.7%	1.4%	2.8%	1.9%	1.8%	7.5%	10.5%
\$15,000 - \$24,999	10.7%	3.7%	4.9%	3.6%	6.7%	11.9%	14.7%
\$25,000 - \$34,999	8.6%	7.2%	5.0%	4.9%	5.3%	13.8%	15.2%
\$35,000 - \$49,999	15.0%	13.9%	12.0%	5.9%	8.9%	12.2%	13.3%
\$50,000 - \$74,999	25.7%	25.5%	24.6%	23.1%	18.1%	18.0%	18.2%
\$75,000 - \$99,999	10.0%	13.0%	16.1%	16.3%	21.4%	18.8%	8.5%
\$100,000 - \$149,999	7.8%	21.7%	20.6%	20.4%	21.4%	8.6%	11.0%
\$150,000 - \$199,999	3.4%	6.2%	6.5%	9.8%	4.6%	2.9%	1.0%
\$200,000+	3.3%	3.7%	5.8%	11.0%	8.8%	2.9%	1.9%

**Data Note:** Census 2000 income is expressed in current (1999) dollars.

**Source:** U.S. Bureau of the Census, 2000 Census of Population and Housing; ESRI forecasts for 2005 and 2010.

©2005 ESRI      On-demand reports and maps from Business Analyst Online. Order at [www.esri.com](http://www.esri.com) or call 800-735-7489      5/17/2005      Page 1 of 3

Source: ESRI

# Types of Reports

- Canned Report
  - static query, parameters, and layout
- Parameterized Report
  - static query and layout but dynamic parameters
- Drill-down Report
  - user clicks link to get additional details
- Ad Hoc Report
  - dynamic query, dynamic layout

Less interactive



More interactive

# Dashboard

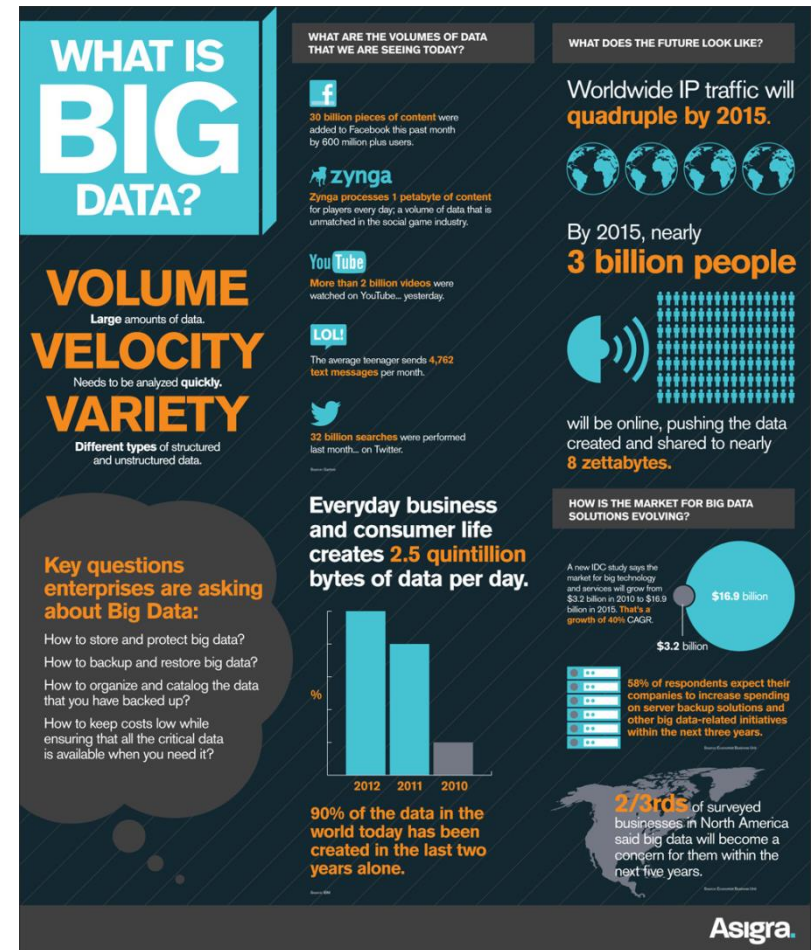
- Prove multiple KPIs (Key Performance Indicators) in a single view
- Like the dashboard of your car
- Typically provide drill-down capabilities



Source: Google Analytics

# Infographic

- Visual representation of data as information
- Tells a story about data
- Intersection of data visualization and design
- Most are static but they are becoming interactive

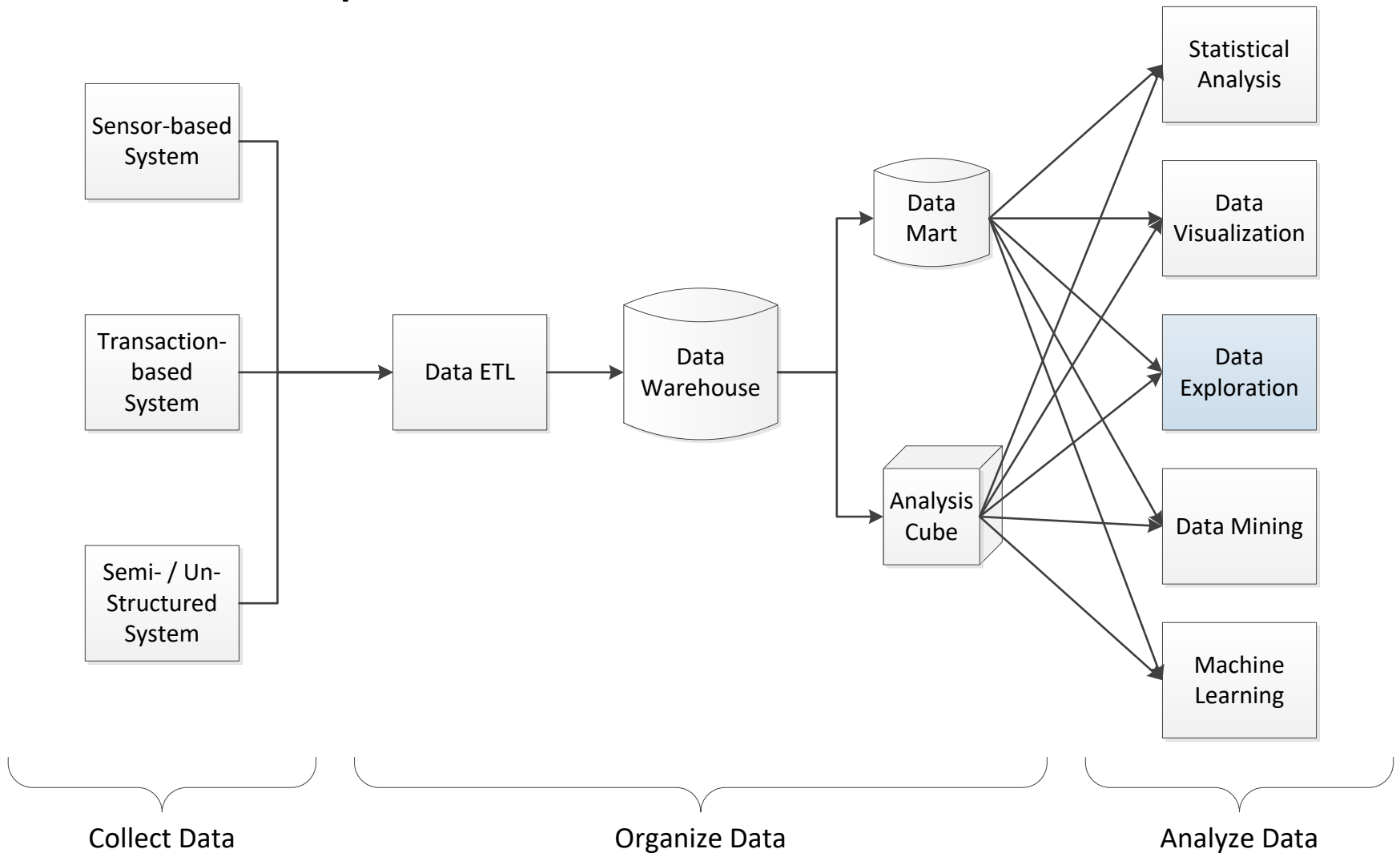


Source: <http://visual.ly/what-big-data>

# Reporting Providers

- IBM – Cognos Business Intelligence
- Microsoft – SQL Server Reporting Services (SSRS)
- Oracle – Oracle Reports
- SAP Business Objects – Crystal Reports

# Data Exploration

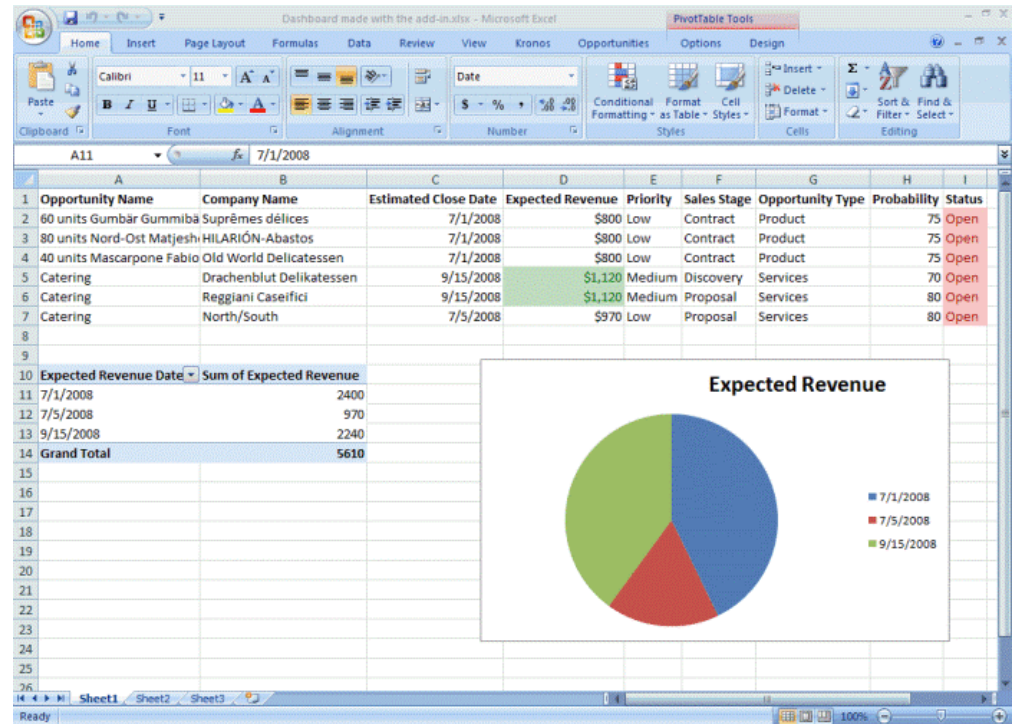


# Data Exploration

- Rather than starting with a specific question, we explore the data to discover knowledge
- Requires interactive tools
- Requires a rapid feedback loop
- Relies heavily on data visualization
- aka: Exploratory data analysis

# Spreadsheet

- Most popular software tool for exploratory data analysis
- Interactive sorting and filtering
- Interactive data visualization

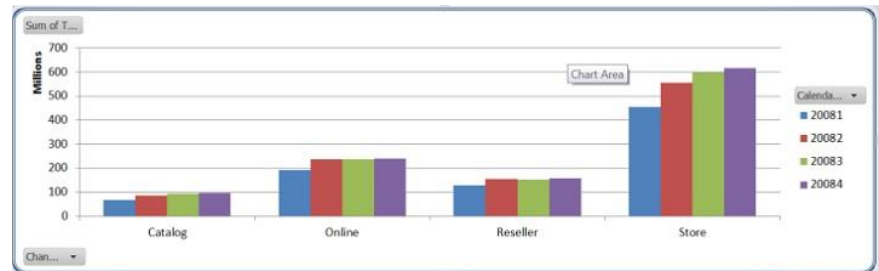


Source: Microsoft

# Pivot Table and Pivot Chart

- Pivot Table
  - Like a cross-tabulation matrix but interactive
- Pivot Chart
  - Provides an interactive graphical representation of a pivot table

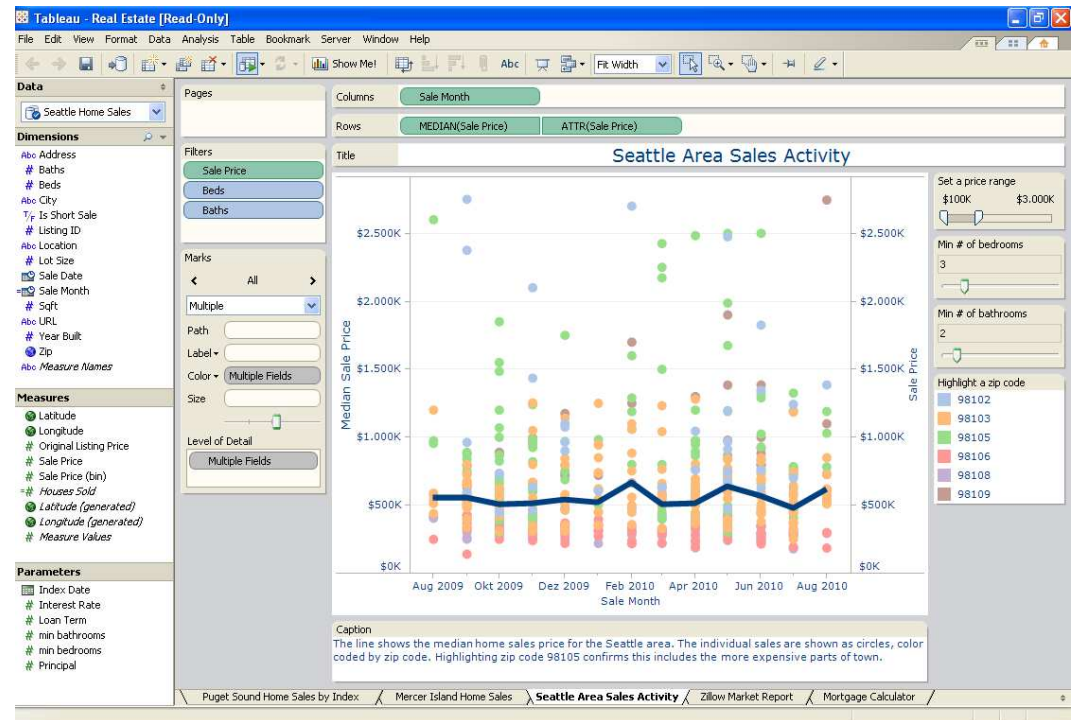
Sum of Units	Ship Date ▼					
Region ▼	1/31/2005	2/28/2005	3/31/2005	4/30/2005	5/31/2005	6/30/2005
East	66	80	102	116	127	125
North	96	117	138	151	154	156
South	123	141	157	178	191	202
West	78	97	117	136	150	157
(blank)						
Grand Total	363	435	514	581	622	640



Source: Microsoft

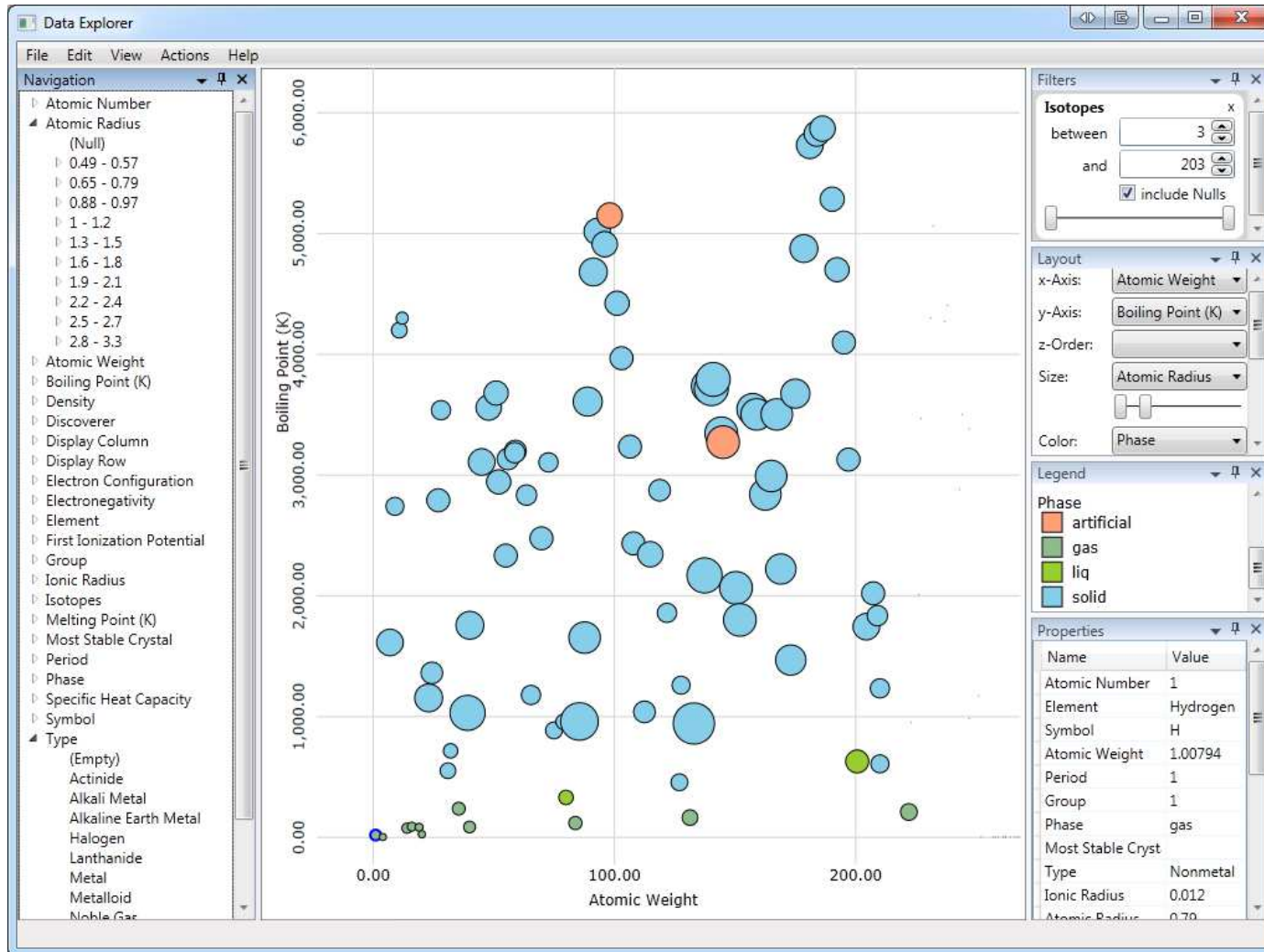
# Data Explorer

- Interactive data visualization tool
- Highly visual
- Highly interactive
- Rapid feedback
- Popular software:
  - Tableau
  - Spotfire

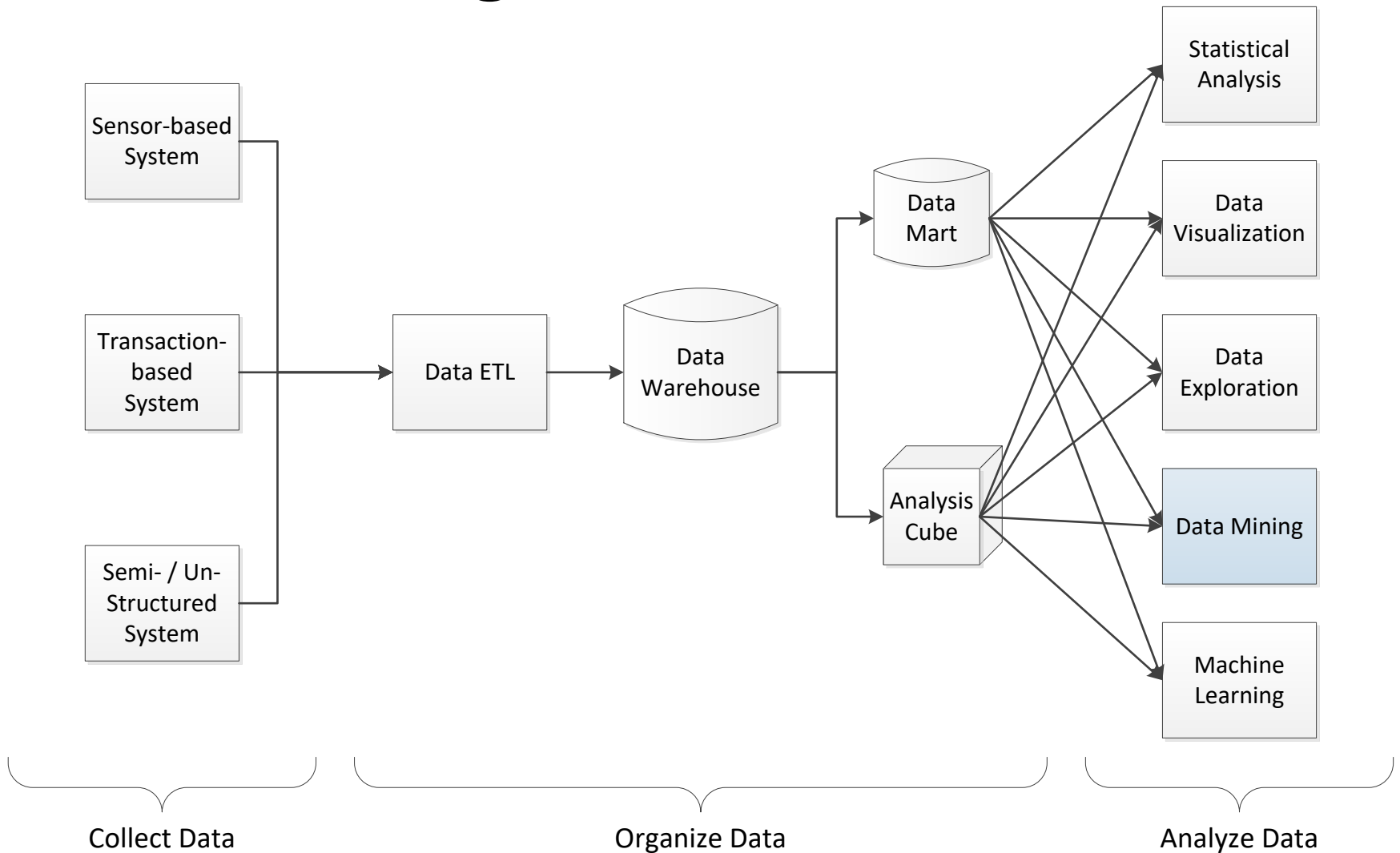


Source: Tableau

# Casual Data Explorer

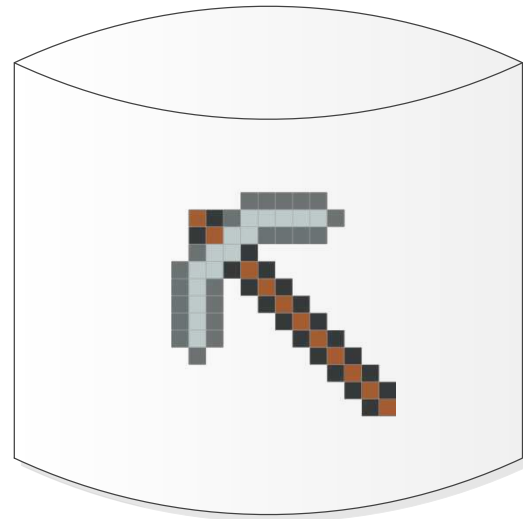


# Data Mining



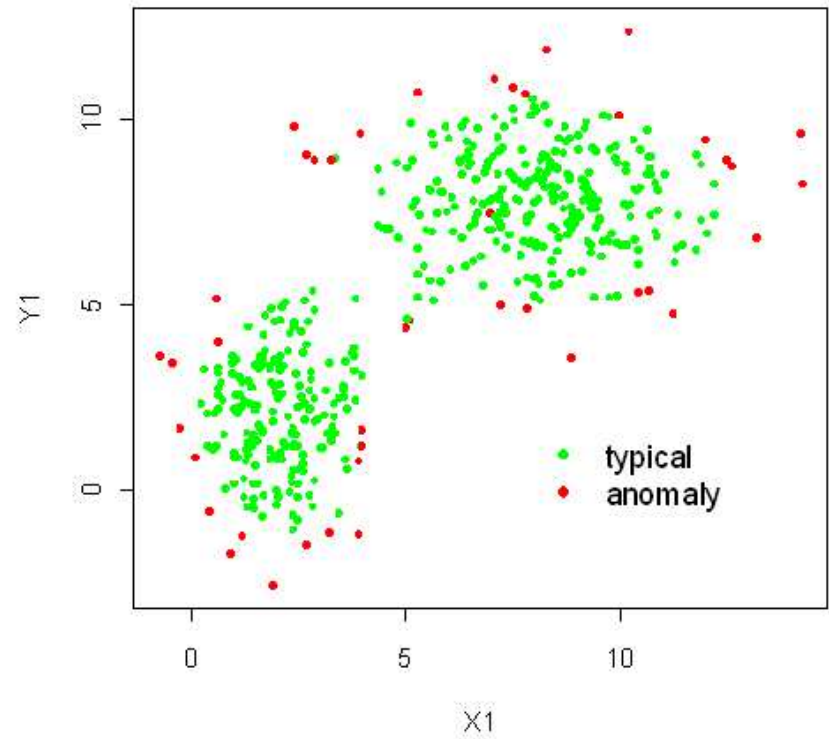
# Data Mining

- Automated or semi-automated exploratory analysis of large sets of data
- Used to discover previously unknown patterns in data
- Sub-field of machine learning (“Applied ML”)



# Anomaly Detection

- Detection of outliers (i.e., patterns of data that do not conform to the rest of the data)
- Applications:
  - Fraud detection
  - Intrusion detection
  - Cleaning data



Source: Oracle

# Association Rule Learning

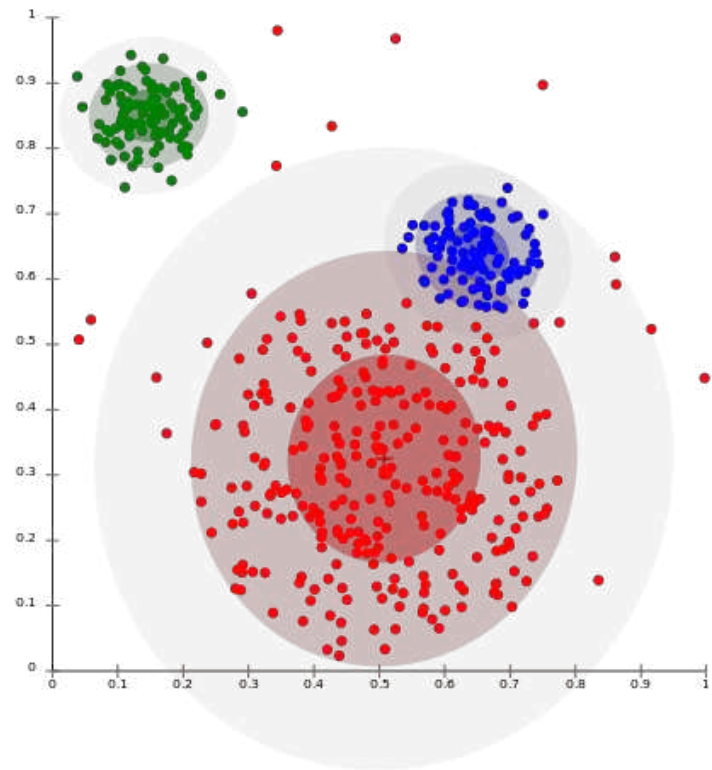
- Discovers relationships between variables in large databases
- Applications:
  - Market Basket Analysis
- Classic Example:
  - Beer and Diapers

Orders\$1_Pro duct Sub- Category	Product 2 - Sub-Category						
	Appliances	Binders and Binder Accessori..	Bookcases	Chairs & Chairmats	Computer Peripheral s	Copiers and Fax	Envelc
Appliances		■	■	■	■	■	■
Binders and ..	■		■	■	■	■	■
Bookcases	■	■		■	■	■	■
Chairs & Ch..	■	■	■		■	■	■
Computer P..	■	■	■	■		■	■
Copiers and ..	■	■	■	■	■		■
Envelopes	■	■	■	■	■	■	■
Labels	■	■	■	■	■	■	■
Office Furnis..	■	■	■	■	■	■	■
Office Machi..	■	■	■	■	■	■	■
Paper	■	■	■	■	■	■	■
Pens & Art S..	■	■	■	■	■	■	■
Rubber Ban..	■	■	■	■	■	■	■
Scissors, Ru..	■	■	■	■	■	■	■
Storage & O..	■	■	■	■	■	■	■
Tables	■	■	■	■	■	■	■
Telephones ..	■	■	■	■	■	■	■

Source: Tableau

# Cluster Analysis

- Assigns a set of objects into groups of similar properties
- Applications:
  - Market Segmentation
  - Image Recognition
  - Crime Analysis

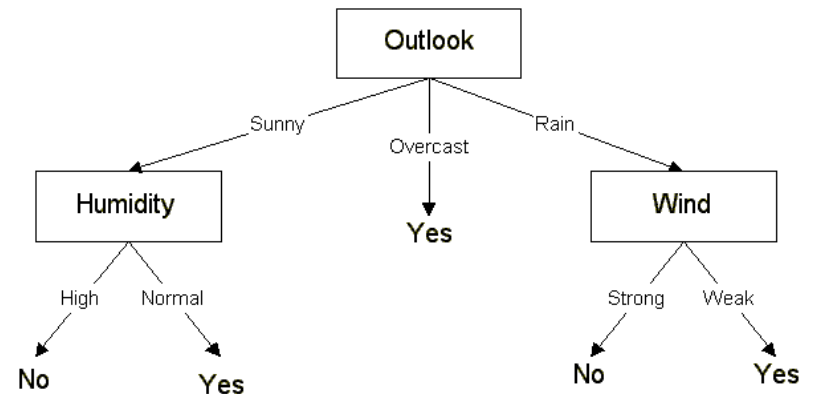


Source: Wikipedia

# Decision Trees

- Builds a decision tree as a model for mapping input variables to an output variable
- Decisions branches ordered to maximize information gain
- Applications:
  - Medical diagnostics
  - Loan approval systems

Decision Tree for Playing Tennis

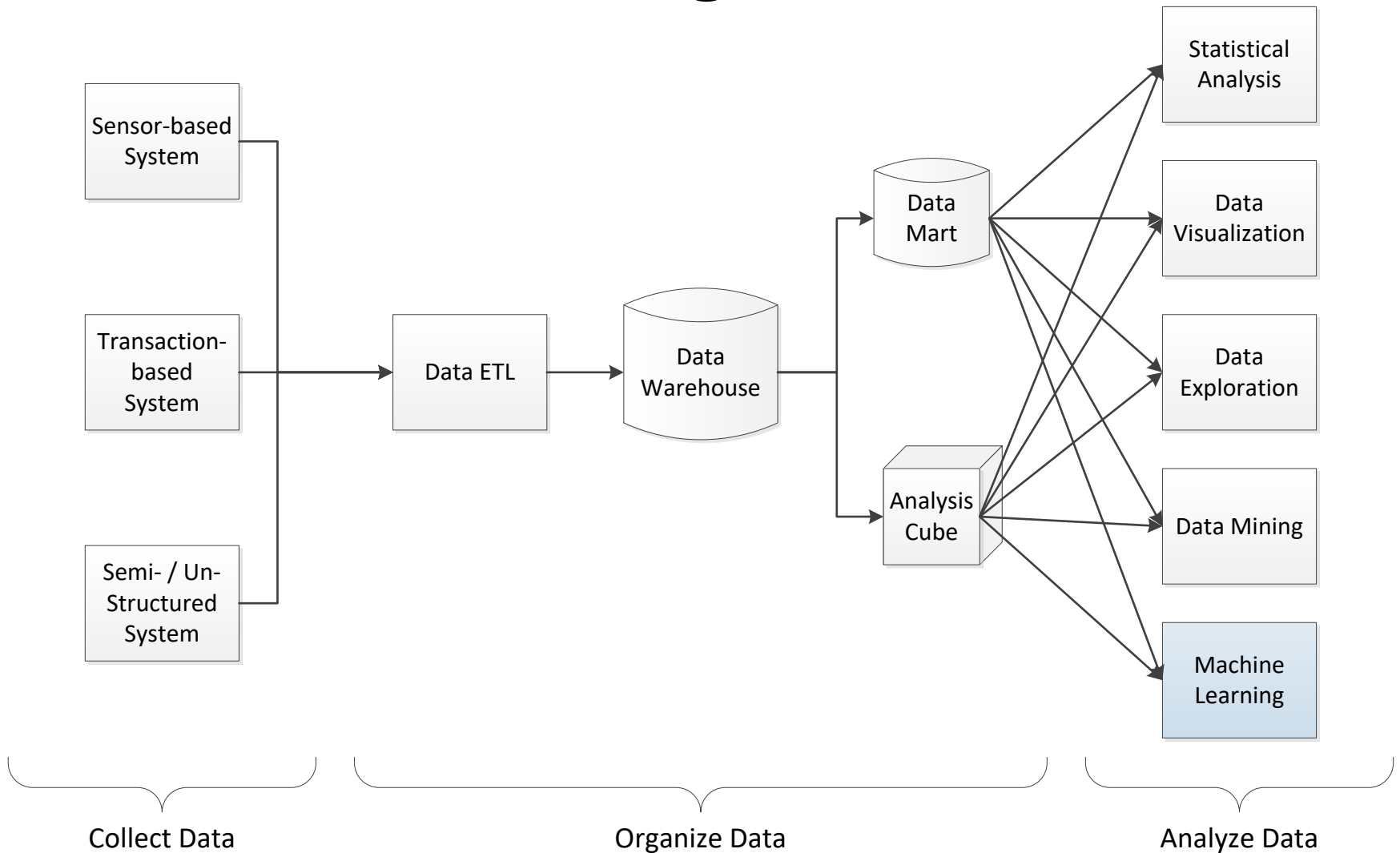


Source: Machine Learning (Tom Mitchell)

# Data Mining Software Providers

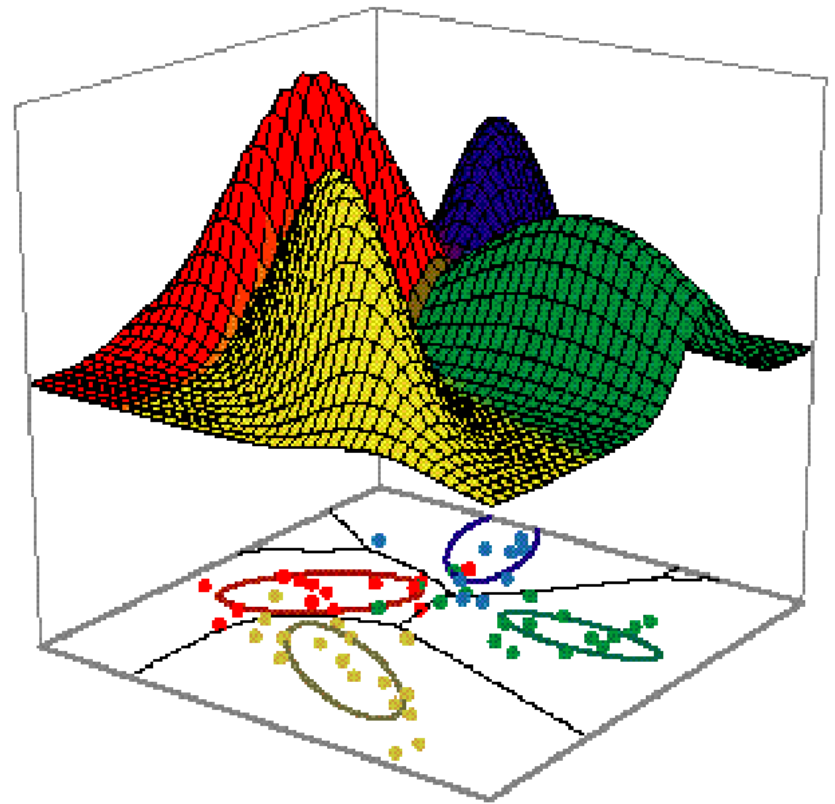
- IBM – Intelligent Miner
- Microsoft – SQL Server Analysis Services
- Oracle – Oracle Data Mining (ODM)
- SAS – Enterprise Miner
- Weka (open source)
- Rapid Miner (open source)

# Machine Learning



# Machine Learning

- Study of algorithms that use existing data to make decisions or predictions about future data
- The algorithm learns the patterns in the data in order to make intelligent decisions
- Sub-field of Artificial Intelligence



Source: Pattern Classification (Duda, Hark, Stork)

# Data Mining vs. Machine Learning

## Data Mining

- Goal is to discover previously unknown knowledge from data
- Uses existing data to discover patterns so that humans can make better decisions
- Uses existing data only

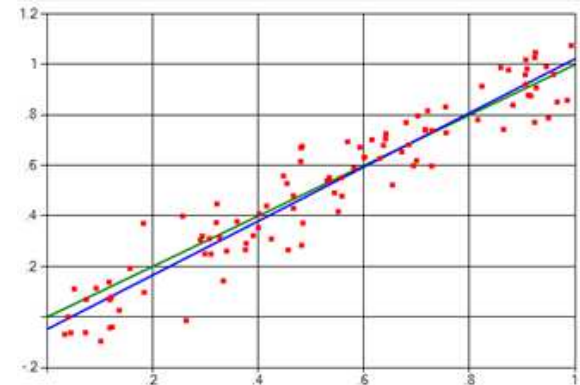
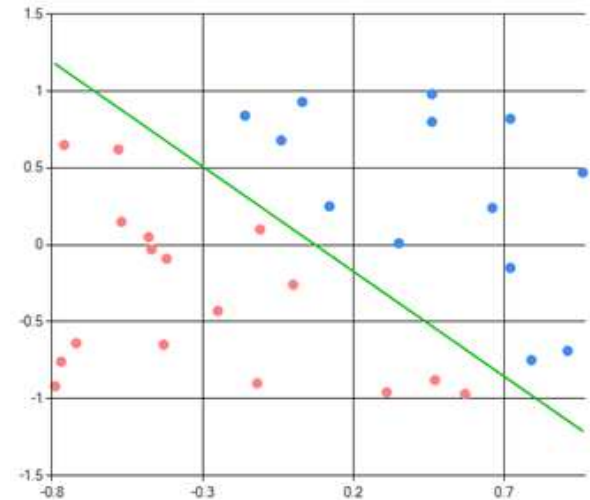
## Machine Learning

- Goal is to reproduce intelligent decision making
- Uses data to create a knowledge model to make decisions autonomously
- Uses existing data to make predictions about new incoming data

Note: All data mining models can be used in machine learning

# Output Types

- Classification
  - Output is a discrete value
  - For example:
    - {true, false};
    - {sunny, cloudy, rainy}
- Regression
  - Output is a continuous value
  - For example:
    - Temp = 98.6°F;
    - Profit = \$100

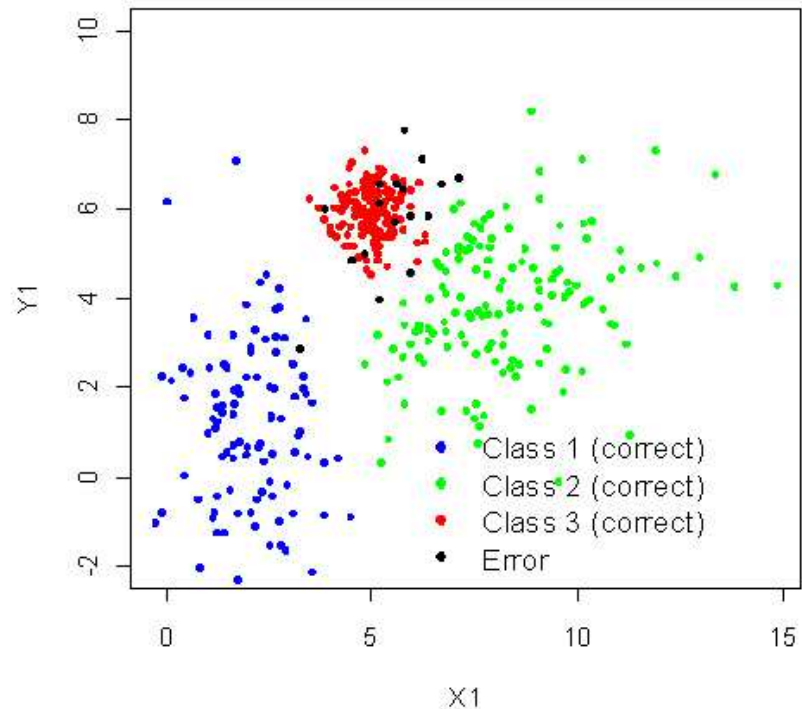


# Training Types

- Supervised
  - Human labels data as input vs. output
  - Machine learns function mapping input to output
- Unsupervised
  - Machine learns structure of unlabeled data
  - Essentially data mining
- Reinforcement Learning
  - Machine learns good decisions from reinforcement
  - Tradeoff between exploration and exploitation

# Classification

- Attempts to classify new data given known classes of existing data
- Supervised version of cluster analysis
- Applications:
  - Spam Detection
  - Credit Scoring
  - Image Recognition
  - Document Classification

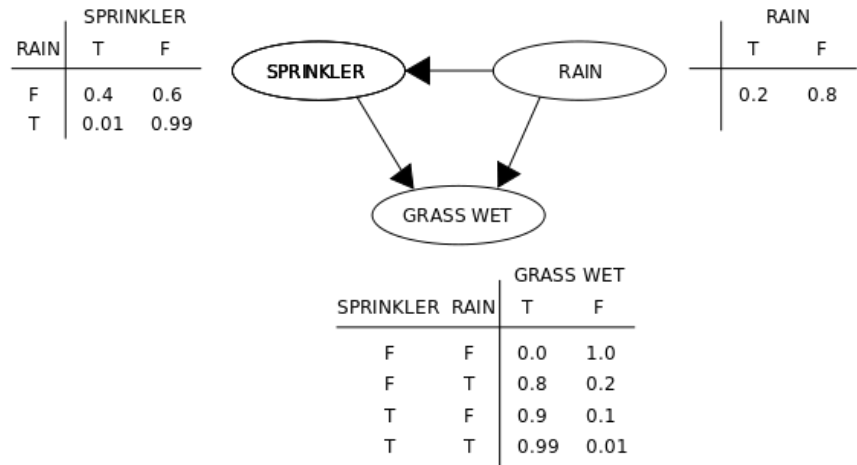


Source: Oracle

# Bayesian Networks

- Graph of variables (nodes) and conditional probabilities (edges)
- Used to calculate probability of
  - Causes given effects (diagnostic)
  - Effects given causes (predictive)
- aka: Belief Network
- Naïve Bayes Network
  - Assumes all variables are conditionally independent

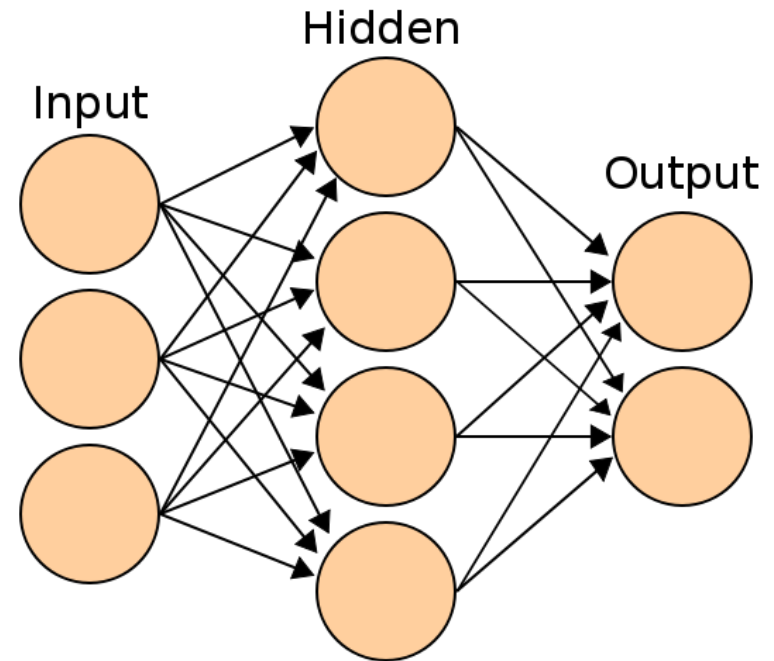
What Causes Wet Grass?



Source: Wikipedia

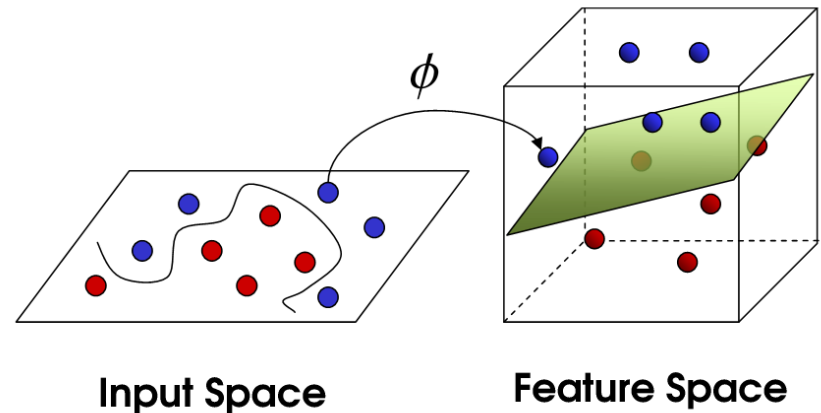
# Neural Networks

- Mathematical model inspired by biological neural networks
- Nodes have summation and activation functions
- May contain one or more hidden layers
- Backpropagation used for credit assignment
- Feedforward vs. feedback



# Support Vector Machines

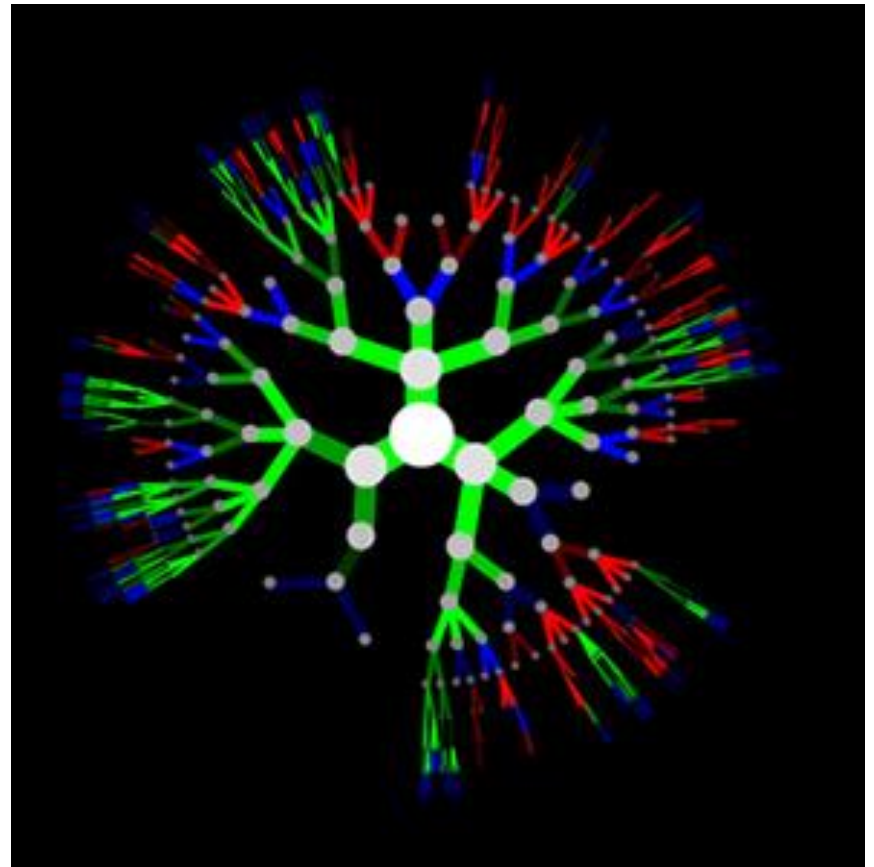
- Like a linear classifier
- Uses a kernel function to map non-linearly separable data to high dimensional space
- Maximum-margin hyperplane can then linearly separate data



Source: Norikazu Takahashi

# Genetic Programming

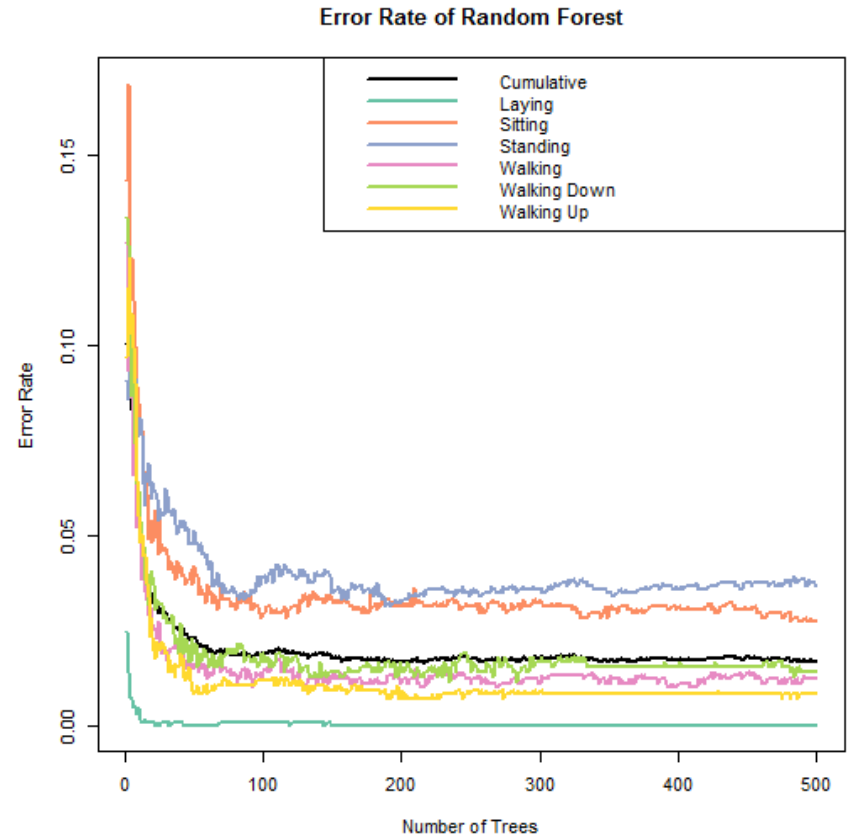
- Uses an evolutionary algorithm to seek optimal decisions given an environment
- Based on biological evolution
  - Genetic crossover
  - Genetic mutation
- Successful agents reproduce; unsuccessful agents die off



Source: Genetic Programming Tree Visualizer

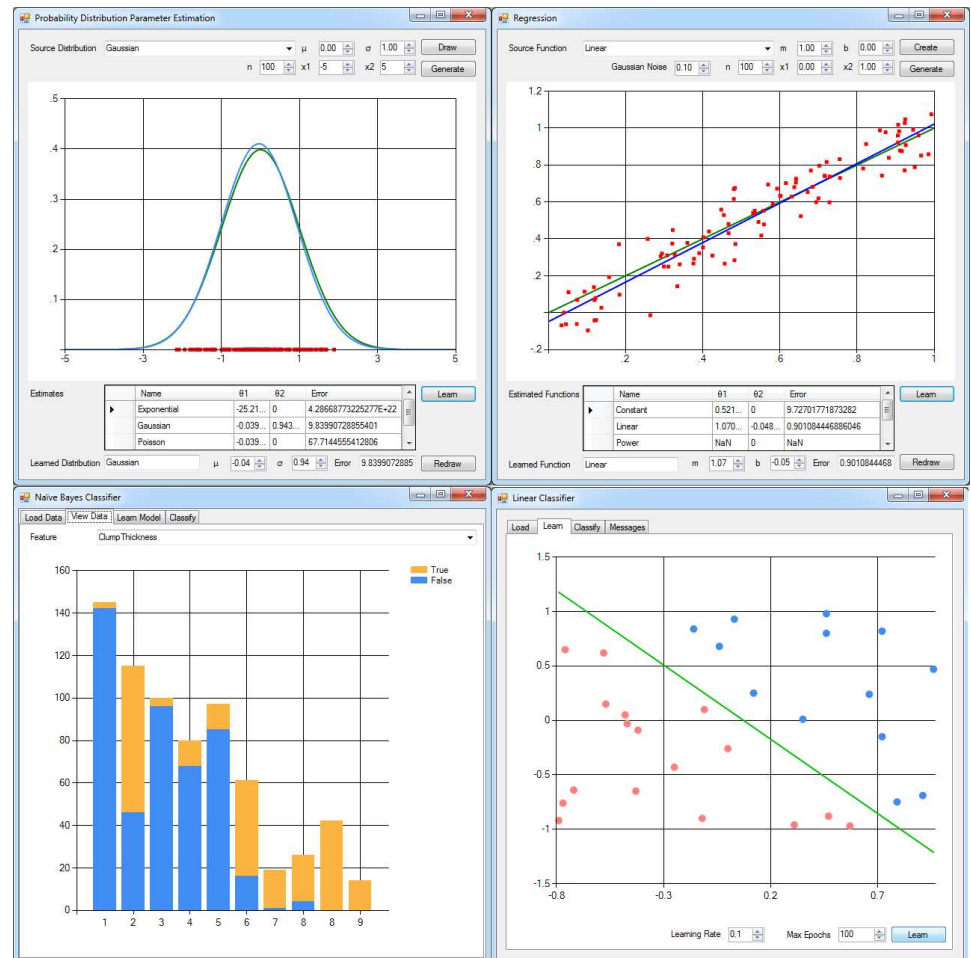
# Ensemble Learning

- Uses multiple machine learning methods to produce better results than any single method
- Multiple *weak* learners vs. one *strong* learner
- If individual answers are better than random then we can aggregate
- Examples:
  - Random Forest Classifier
  - IBM's Watson

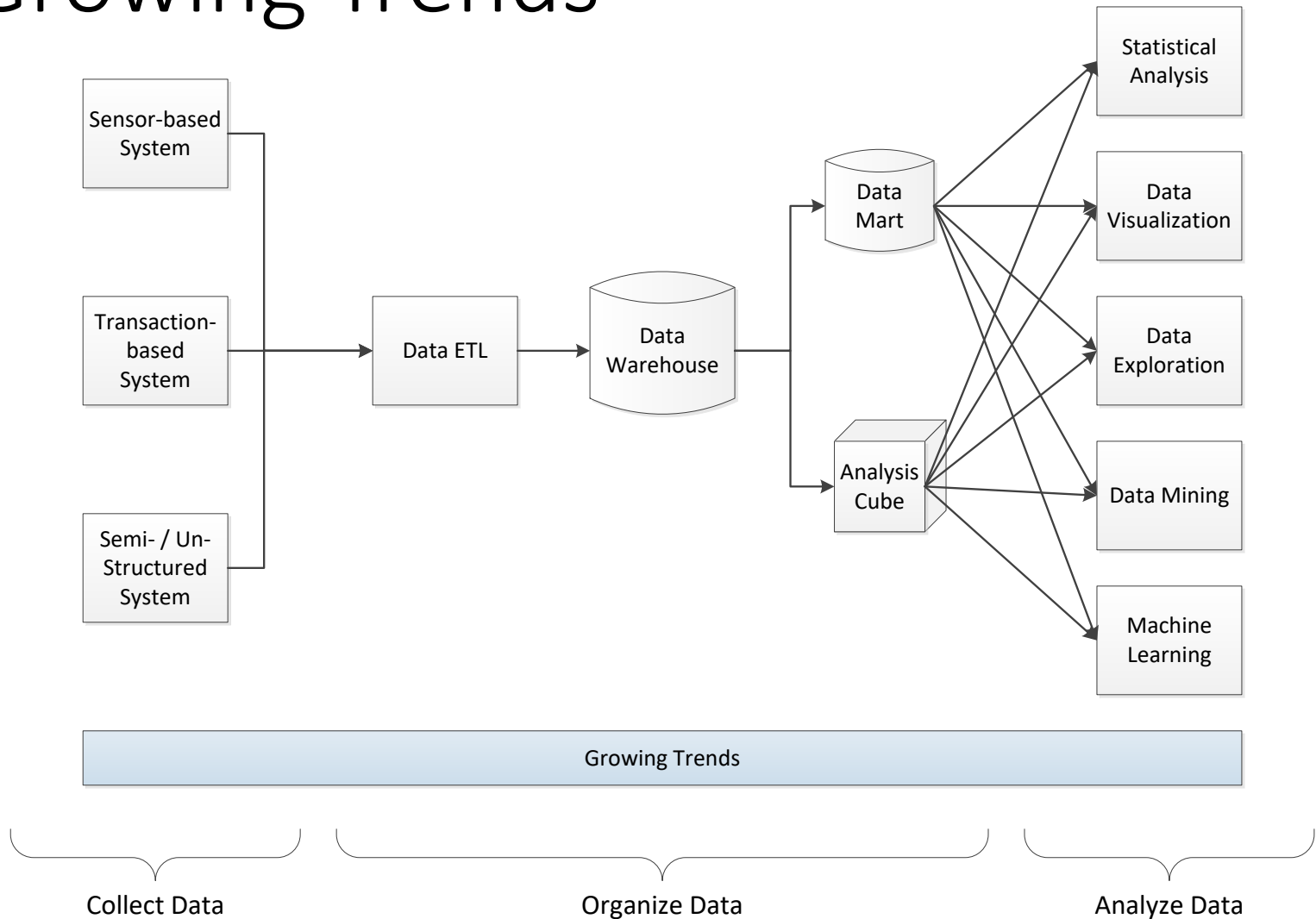


# Machine Learning Toolkit

- Provides a set of machine learning algorithms
- Popular ML toolkits:
  - Apache Mahout
  - KMINE
  - Rapid Miner
  - Weka



# Growing Trends



# NoSQL Databases

- Growing trend of non-relational databases
  - Does not use SQL as a query language
  - Optimized for retrieval and append operations
  - Data is typically stored in key-value pairs, XML, documents, or graphs
  - Distributed across multiple machines
  - Elastic scaling (scale out vs. scale up)
  - Uses “eventual consistency” rather than ACID

# In-Memory Analytics

- Growing trend of storing data for analysis in-memory rather than on on-disk
- Up to a million times faster than on-disk solutions
- Types of In-Memory Analytic Tools:
  - In-Memory ROLAP (Relational)
  - In-Memory MOLAP (Cubes)
  - In-Memory Inverted Index
  - In-Memory Associative Index
  - In-Memory Spreadsheet

# Column-Store Database

- Tabular data is stored by columns instead of rows
- Can be orders of magnitude faster than row-oriented databases for analytic queries
- Typically used for data marts

ID	Date	Customer	Product	Quantity
1	2012-10-27	John	Pizza	2
2	2012-10-27	John	Soda	2
3	2012-10-27	Jill	Salad	1
4	2012-10-27	Bob	Milk	1
5	2012-10-28	Sue	Soda	3
6	2012-10-28	Bob	Pizza	2
7	2012-10-28	Jill	Pizza	1
8	2012-10-28	Jill	Soda	3



IDs	Date
1-4	2012-10-27
5-8	2012-10-28

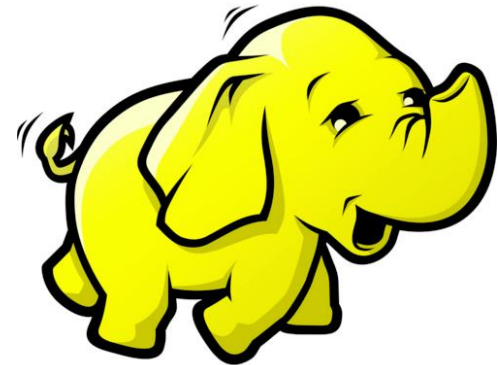
IDs	Produce
4	Milk
1,7,8	Pizza
3	Salad
2,5,8	Soda

IDs	Customer
4,6	Bob
3,7,8	Jill
1-2	John
5	Sue

IDs	Quantity
3,4,7	1
1,2,6	2
5,8	3

# Hadoop

- Used for data-intensive distributed applications
- Highly distributed (i.e., many nodes)
- Massively parallel processing
- Consists of three components
  - Hadoop Kernel
  - Hadoop Distributed File System (HDFS)
  - MapReduce



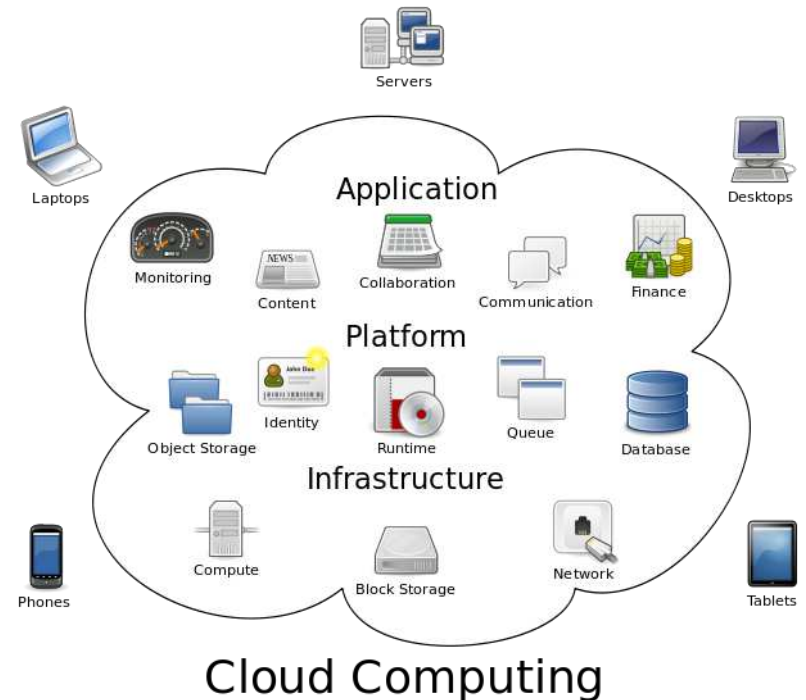
Source: Hadoop

# Growing Analysis Trends

- Predictive Analytics
  - Uses existing data to predict future events
  - Exploits relationships between explanatory variables and predictor variables to predict future values
- Sentiment Analysis
  - Detects subtle emotional content in text to determine if the content is favorable or unfavorable towards the subject of the text

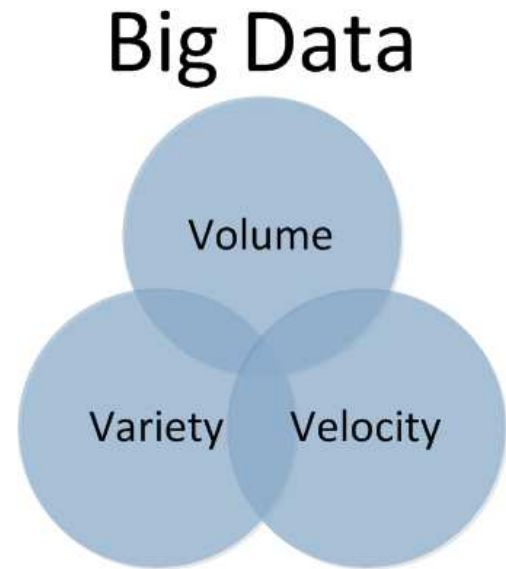
# Cloud Analytics

- Analytics being offered as a cloud service
  - Elastic scalability
  - Lower cost of ownership
- Driving new interest in functional languages
  - Scheme
  - F#



# Big Data

- Data that are difficult to process using conventional data processing means
- Three Vs of Big Data:
  - Volume – quantity of data
  - Velocity – speed that data must be processed
  - Variety – semi-structured and unstructured data



# Big Data

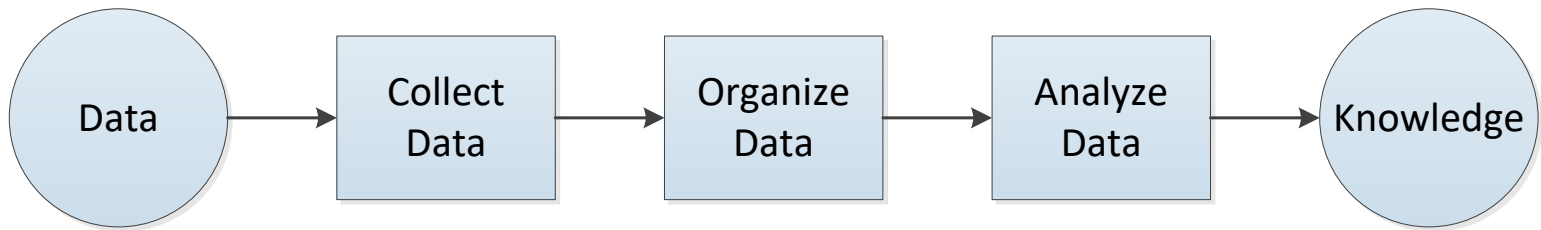
- What is fueling the big data movement?
  - Sensors are everywhere and growing fast
  - Human interaction with devices is increasing
  - Machines are generating lots of data as well
  - 90% of world's data was created in last 2 years
  - We are creating 2.5 exabytes of data daily
    - that's 2,500,000,000,000,000,000 bytes (source: IBM)
- Why is big data important?
  - More data => better knowledge => better decisions

# Where is this all going?

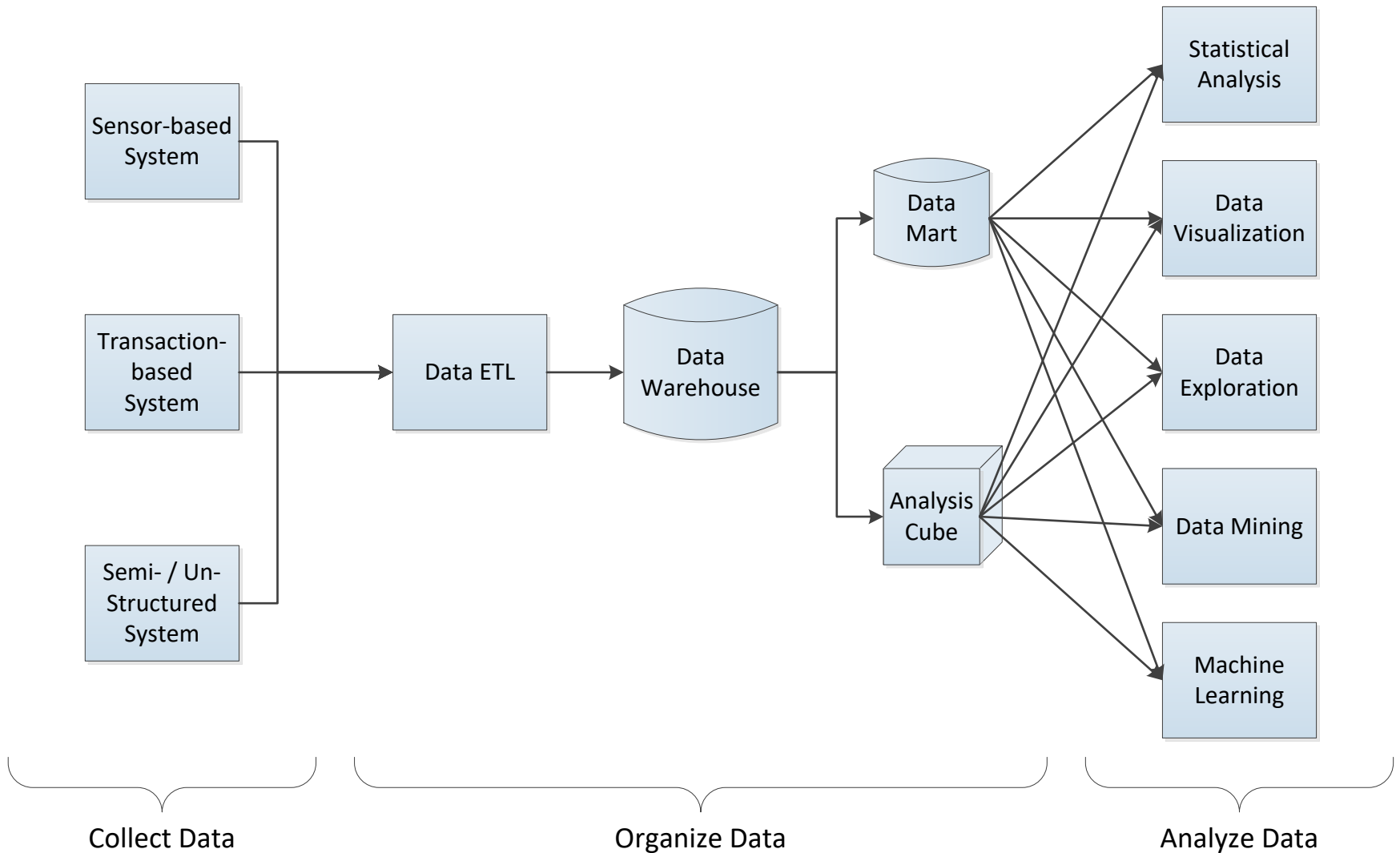
- Certain:
  - More data (doubling every 18-24 months)
  - More users performing data analysis
  - More machine decision making
- Probably:
  - Statistics will become the next hot profession
  - Data scientists will emerge
- Possibly:
  - Machines making scientific discoveries

# Conclusion

- How do we transform data into knowledge?
  1. Collect Data
  2. Organize Data
  3. Analyze Data



# Review



# Feedback

- Did you find this presentation valuable?
- What could I do to make the presentation better?
- What other presentations would you like to see?
  - Data Visualization
  - Data Exploration
  - Data Mining
  - Machine Learning

# Contact Info

Matthew Renze

[info@renzeconsulting.com](mailto:info@renzeconsulting.com)

Renze Consulting

[www.renzeconsulting.com](http://www.renzeconsulting.com)