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 System

Transaction-based System

Semi-/Un-Structured System

Collect Data

Data ETL

Data Warehouse

Organize Data

Data Mart

Analysis Cube

Data Visualization

Data Exploration

Data Mining

Statistical Analysis

Machine Learning

Analyze Data

Collect Data

Organize Data

Analyze Data
Sensor-based Systems

Sensor-based System

Transaction-based System

Semi-/Un-Structured System

Data ETL

Data Warehouse

Data Mart

Analysis Cube

Statistical Analysis

Data Visualization

Data Exploration

Data Mining

Machine Learning

Collect Data

Organize Data

Analyze Data
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
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

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

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
Transaction-based Systems

Sensor-based System
Transaction-based System
Semi-/Un-Structured System

Collect Data
Organize Data
Analyze Data

Data ETL
Data Warehouse
Data Mart
Analysis Cube

Statistical Analysis
Data Visualization
Data Exploration
Data Mining
Machine Learning
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

<table>
<thead>
<tr>
<th>ID</th>
<th>Date</th>
<th>Customer</th>
<th>Product</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2012-10-27</td>
<td>John</td>
<td>Pizza</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2012-10-27</td>
<td>John</td>
<td>Soda</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2012-10-27</td>
<td>Jill</td>
<td>Salad</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2012-10-27</td>
<td>Bob</td>
<td>Milk</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2012-10-28</td>
<td>Sue</td>
<td>Soda</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2012-10-28</td>
<td>Bob</td>
<td>Pizza</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>2012-10-28</td>
<td>Jill</td>
<td>Pizza</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2012-10-28</td>
<td>Jill</td>
<td>Soda</td>
<td>3</td>
</tr>
</tbody>
</table>
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

- Sensor-based System
- Transaction-based System
- Semi-/Unstructured System

Data ETL → Data Warehouse

- Data Mart
- Analysis Cube

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

Collect Data → Organize Data → Analyze Data
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
Semi-Structured Data

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

- Sensor-based System
- Transaction-based System
- Semi- / Un-Structured System

Data ETL → Data Warehouse

- Data Mart
- Analysis Cube

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

Collect Data → Organize Data → Analyze Data
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., amount = price * 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
Popular ETL Software

- IBM – Information Server (Datastate)
- 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

• 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

Time Dimension
- Order Date
- Year
- Quarter
- Month

Product Dimension
- Name
- Category
- Price

Fact Table
- Total
- Quantity
- Freight
- Discount

Customer Dimension
- Name
- Address
- City
- Zip

Employee Dimension
- Name
- Supervisor
- Department
- Region
- Territory

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 (3rd-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

Sensor-based System

Transaction-based System

Semi- / Un-Structured System

Data ETL

Data Warehouse

Data Mart

Analysis Cube

Statistical Analysis

Data Visualization

Data Exploration

Data Mining

Machine Learning

Collect Data

Organize Data

Analyze Data
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

Sensor-based System

Transaction-based System

Semi-/Un-Structured System

Data ETL

Data Warehouse

Data Mart

Analysis Cube

Statistical Analysis

Data Visualization

Data Exploration

Data Mining

Machine Learning

Collect Data

Organize Data

Analyze Data
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
Cube Operations

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

Analysis Cube Query Languages

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

Sample MDX Query:

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

Query Results:

<table>
<thead>
<tr>
<th>Year</th>
<th>Store Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>32663.74</td>
</tr>
<tr>
<td>2012</td>
<td>65303.44</td>
</tr>
</tbody>
</table>
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

Sensor-based System

Transaction-based System

Semi-/Un-Structured System

Data ETL

Data Warehouse

Data Mart

Analysis Cube

Statistical Analysis

Data Visualization

Data Exploration

Data Mining

Machine Learning

Collect Data

Organize Data

Analyze Data
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

- Data analysis using statistical methods
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: 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

Sensor-based System

Transaction-based System

Semi-/Un-Structured System

Data ETL

Data Warehouse

Data Mart

Analysis Cube

Statistical Analysis

Data Visualization

Data Exploration

Data Mining

Machine Learning

Collect Data

Organize Data

Analyze Data
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

<table>
<thead>
<tr>
<th>ID</th>
<th>Date</th>
<th>Customer</th>
<th>Product</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2012-10-27</td>
<td>John</td>
<td>Pizza</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2012-10-27</td>
<td>John</td>
<td>Soda</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2012-10-27</td>
<td>Jill</td>
<td>Salad</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2012-10-27</td>
<td>Bob</td>
<td>Milk</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2012-10-28</td>
<td>Sue</td>
<td>Soda</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>2012-10-28</td>
<td>Bob</td>
<td>Pizza</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>2012-10-28</td>
<td>Jill</td>
<td>Pizza</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2012-10-28</td>
<td>Jill</td>
<td>Soda</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Milk</th>
<th>Pizza</th>
<th>Salad</th>
<th>Soda</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Jill</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>John</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Sue</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>8</td>
<td>15</td>
</tr>
</tbody>
</table>
Charts and Graphs

- Visual representation of multivariate data
- Discrete and continuous data representations
- Common examples:
  - Bar/Column Chart
  - Line Graph
  - Pie Chart
  - Scatter Plot
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 represent location and spatial distribution of data
  • Choropleth
    • Colors represent values of data within a boundary

Source: USDA
Report

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

- Sensor-based System
- Transaction-based System
- Semi-/Un-Structured System

Data ETL ➔ Data Warehouse

Data Mart ➔ Analysis Cube

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

Collect Data ➔ Organize Data ➔ Analyze Data
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

![Pivot Table Example](image)

![Pivot Chart Example](image)

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

• 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

Source: Tableau
Cluster Analysis

• Assigns a set of objects into groups of similar properties

• Applications:
  • Market Segmentation
  • Image Recognition
  • Crime Analysis

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

- Sensor-based System
- Transaction-based System
- Semi-/Un-Structured System

Collect Data

Organize Data

Analyze Data

Data ETL

Data Warehouse

- Data Mart
- Analysis Cube

- Statistical Analysis
- Data Visualization
- Data Exploration
- Data Mining
- 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, Har, 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?

<table>
<thead>
<tr>
<th>RAIN</th>
<th>SPRINKLER</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0.4</td>
</tr>
<tr>
<td>T</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RAIN</th>
<th>SPRINKLER</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RAIN</th>
<th>GRASS WET</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RAIN</th>
<th>GRASS WET</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RAIN</th>
<th>GRASS WET</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>0.99</td>
</tr>
</tbody>
</table>

Neural Networks

- Mathematical model inspired by biological neural networks
- Nodes have summation and activation functions
- May contain one or more hidden layers
- Backpropogation 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

Sensor-based System

Transaction-based System

Data ETL

Data Warehouse

Data Mart

Analysis Cube

Statistical Analysis

Data Visualization

Data Exploration

Data Mining

Machine Learning

Growing Trends

Collect Data

Organize Data

Analyze Data
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
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
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 2,500,000,000,000,000,000,000 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

Sensor-based System

Transaction-based System

Semi-/Un-Structured System

Data ETL

Data Warehouse

Data Mart

Analysis Cube

Statistical Analysis

Data Visualization

Data Exploration

Data Mining

Machine Learning

Collect Data

Organize Data

Analyze Data
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

matthew@renzeconsulting.com

Renze Consulting

www.renzeconsulting.com