

Business Intelligence: A Design Science Perspective

Salvatore T. March
David K Wilson Professor of Management
Owen Graduate School of Management
Vanderbilt University
<http://SalMarch.com>

Herb Simon

"The natural sciences are concerned with how things are."

"Design ... is concerned with how things ought to be, with devising artifacts to attain goals."

The Sciences of the Artificial

Agenda

Introduction and Overview

Case Studies

Data Warehouse Representations

Business Intelligence Tools

- Reporting

- OLAP

- Data Mining

- Predictive Analytics

Conclusions

Top 10 Business and Technology Priorities in 2010

Source: Gartner EXP (January 2010)

Business Priorities	Technology Priorities
Business process improvement	Virtualization
Reducing enterprise costs	Cloud computing
Increasing the use of information/analytics	Web 2.0
Improving enterprise workforce effectiveness	Networking, voice and data communications
Attracting and retaining new customers	Business Intelligence
Managing change initiatives	Mobile technologies
Creating new products or services (innovation)	Data/document management and storage
Targeting customers and markets more effectively	Service-oriented applications and architecture
Consolidating business operations	Security technologies
Expanding current customer relationships	IT management

Business Intelligence

Online Analytical Processing (OLAP)

- Reporting Tools
- Dashboards and Data Cubes
 - Performance Measures
 - Analysis Dimensions

Analytical Modeling

- Data Mining
 - Statistical
 - Artificial Intelligence
- Predictive Models

In 2008, the business intelligence (BI) tools market reached \$7.8 billion in software license and maintenance revenue. The market growth of 10.6% in 2008 surpassed previous IDC projections, as spending by organizations of all sizes continued. Organizations are focusing on BI and analytics projects that help reduce costs or retain customers. There is growing evidence that more pervasive BI and analytics have a direct impact on competitiveness. Better decision making is more important when resources become restricted during a recession, so BI and analytics projects will still appeal to management. However, justifying large capital outlays for software will be challenging unless short term benefits can be directly correlated with the investment. As more incremental projects are undertaken, it will be important to execute these projects within the long-term strategic plan of organization wide decision management.

Dan Vesset, Program Vice President, Business Analytics.

Amazon Business Intelligence is hiring.



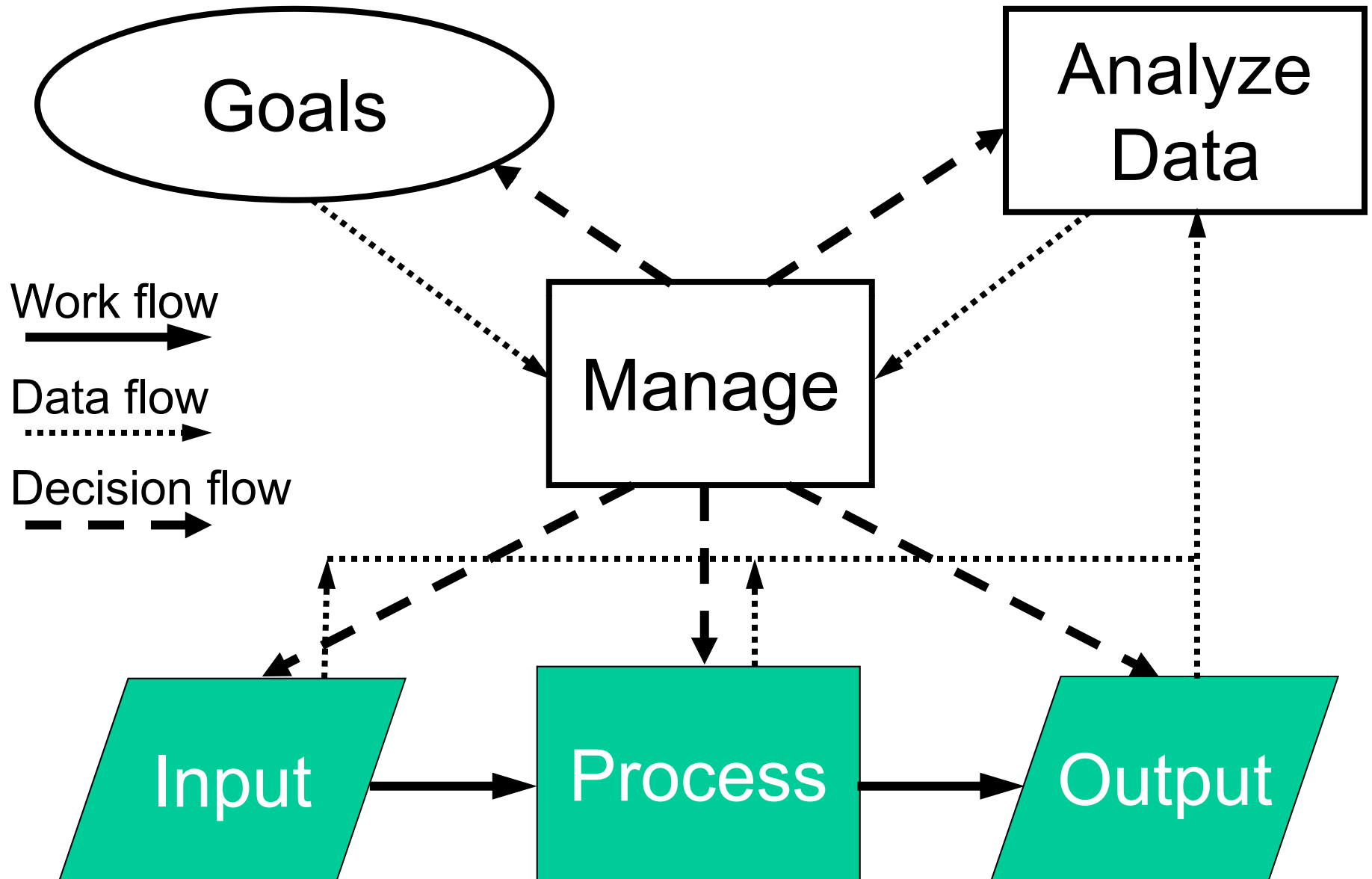
Come work on one of the world's largest and most sophisticated Data Warehouses.

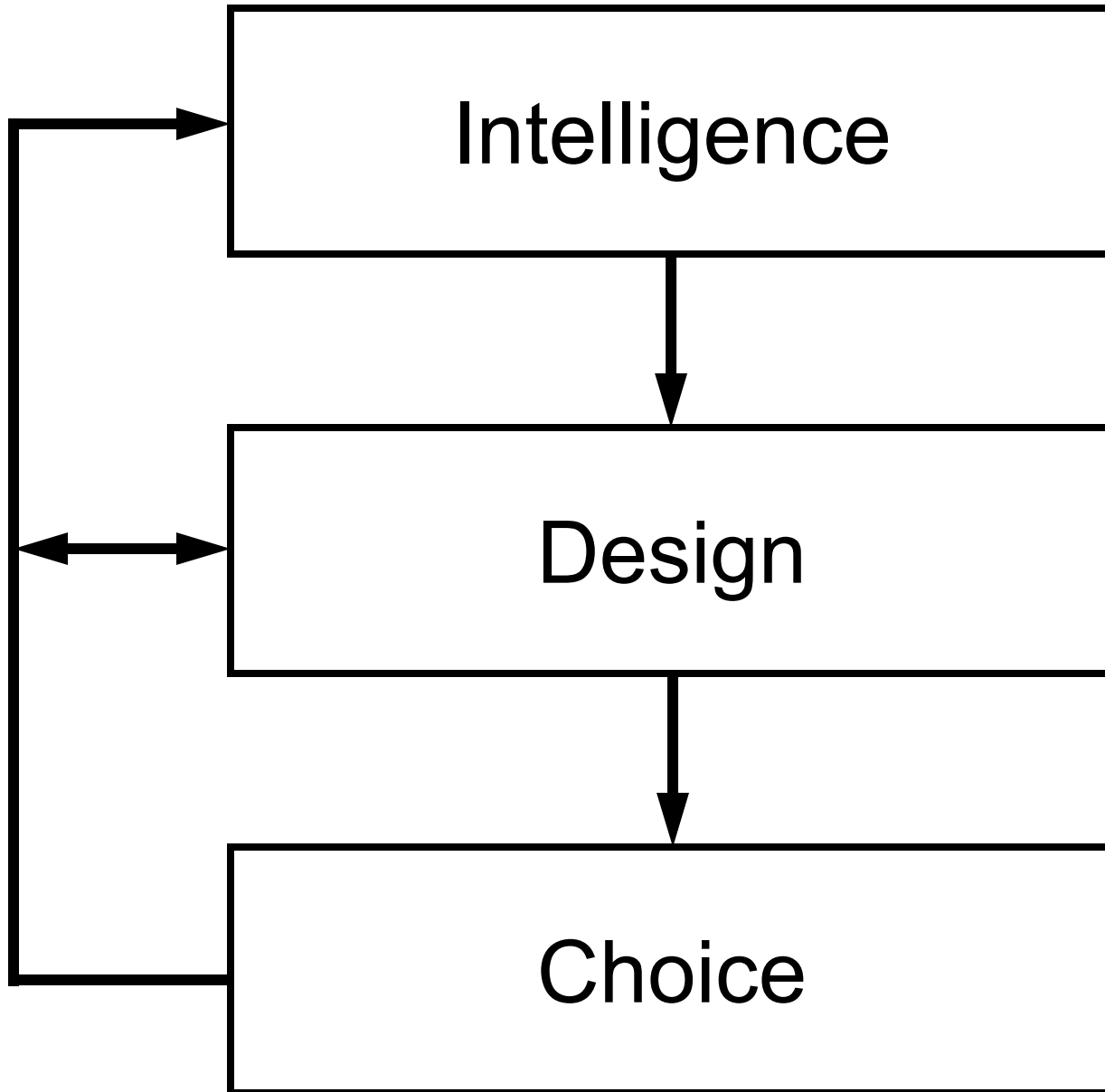
Visit biamazon.blogspot.com for more information.

Do you want run one of the world's largest data warehouses?

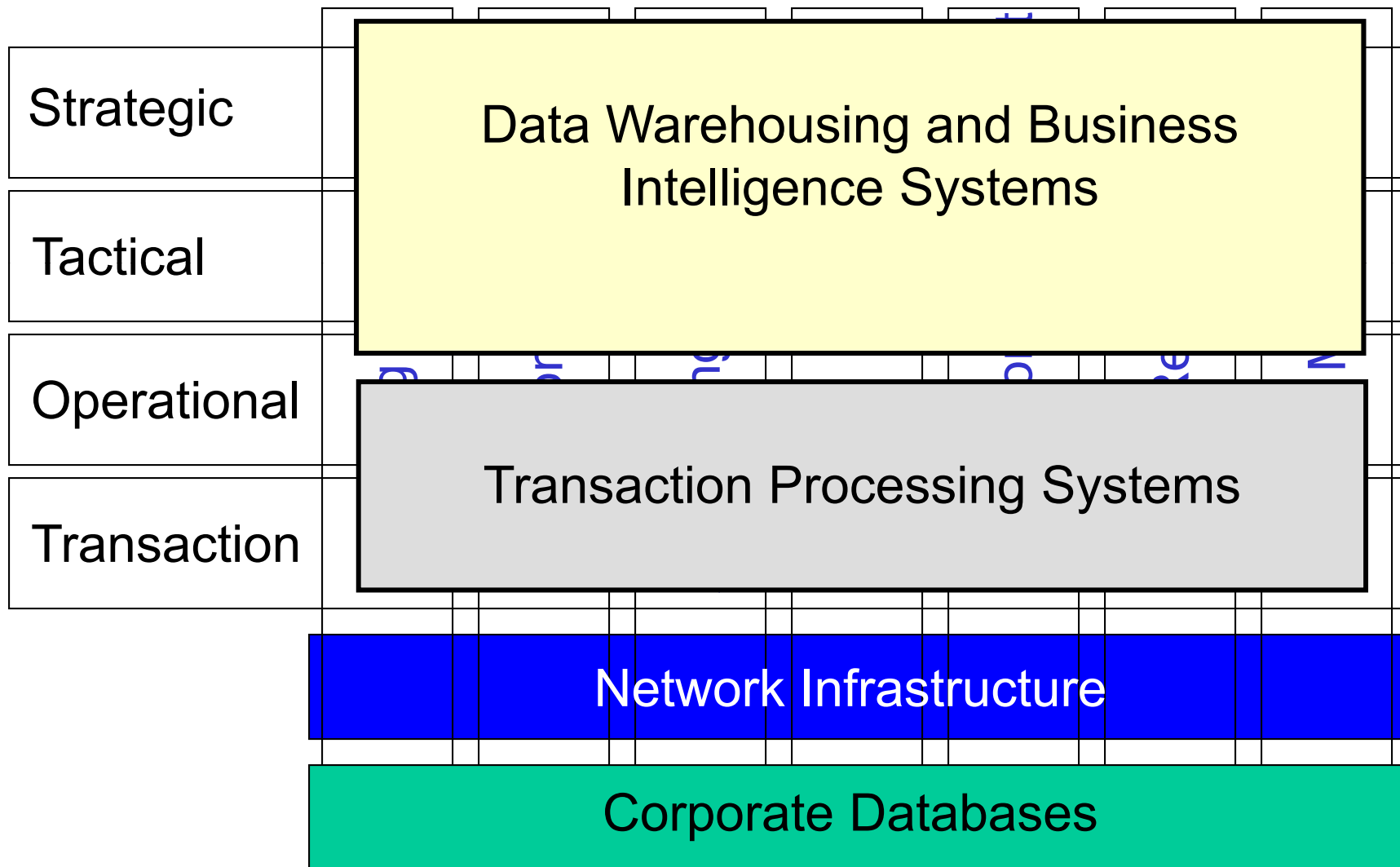
At Amazon.com data and analysis guide every business decision and deliver enormous business value. Amazon's Business Intelligence team is responsible for a business analytics platform that provides reporting, analysis, and data mining, and to over thousands of internal and external customers worldwide. To help deliver these core business metrics and decisions, we run one of the world's largest data warehouses. The BI team is recruiting for a Manager of Database Administration to help deliver the next generation of Amazon.com's data warehousing solution.

Business System Concepts





Conceptual MIS Structure



Red Flags

- Businesspeople debating the "correct" numbers coming from different reports.
- Businesspeople using spreadsheets to "adjust" the numbers to be "correct."
- Data shadow systems or "spreadmarts."
- Multiple pockets of IT resources developing, deploying, maintaining, upgrading and growing skills in different BI tools.
- Debates among different IT groups as to what data, metrics and algorithms should be used for different reports.

References

- Houghton, R. et al. "Vigilant Information Systems for Managing Enterprises in Dynamic Supply Chains: Real-time Dashboards At Western Digital," *MIS Quarterly Executive*, March 2004.
- McAfee, A. "Business Intelligence Software at SYSCO," *Harvard Business School Case Study*, September 2006.
- Wetherbe, J. C. "Executive Information Requirements: Getting it Right," *MIS Quarterly*, March 1991.
- Davenport, T. "Competing on Analytics," *Harvard Business Review*, Jan. 2006.

References: Yogi Berra

- If you don't know where you are going, you might wind up someplace else.
- When you come to a fork in the road, take it.
- You can observe a lot by just watching.
- If the world were perfect, it wouldn't be.
- If you ask me a question I don't know, I'm not going to answer.
- It's tough to make predictions, especially about the future.

Executive Information

Wetherbe (MISQ, March 1991)

- Articulate business objectives
- Analyze problems encountered
- Analyze decisions made
- Define CSFs (KPIs)
- Assess "Ends and Means"

Transform into information requirements using Joint Application Design principles.

Then assess information *importance* and *availability*.

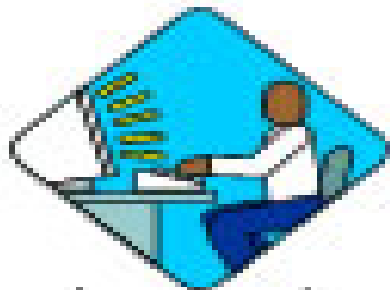
Western Digital

Western Digital (WD) is a \$3 billion global designer and manufacturer of high-performance hard drives for desktop personal computers, corporate networks, enterprise storage, and home entertainment applications. WD's top five business challenges:

1. Constantly changing customer requirements
2. A fiercely competitive global industry
3. Avoiding business disruption, product returns, excess inventory, and bad scheduling
4. Short product lifecycles
5. The need for extremely high quality and reliability

[Houghton, et al.]

Traditional
Information
System

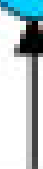
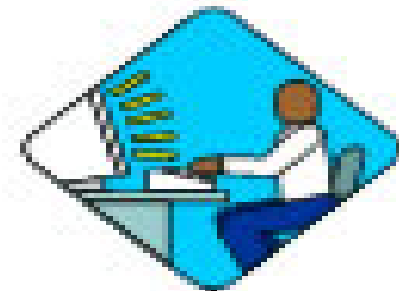


Passive Database

Data

Repetitive
Queries to
Discover
Status

Vigilant
Information
System

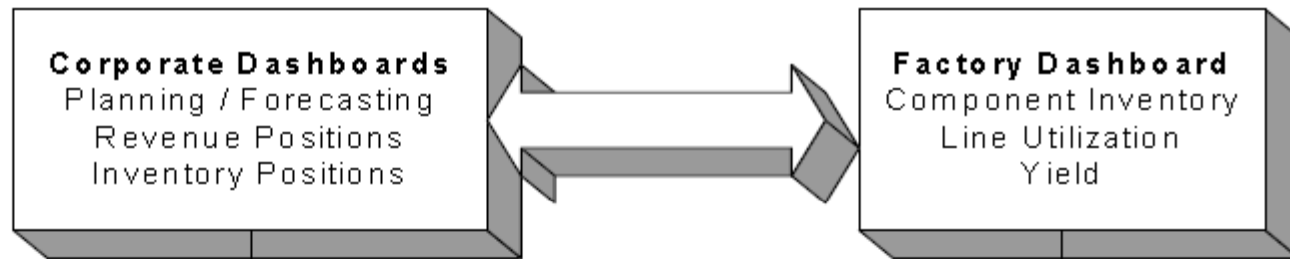


Active Database

Data

Single Alert
Notification
of Status
Triggered by
Data Update

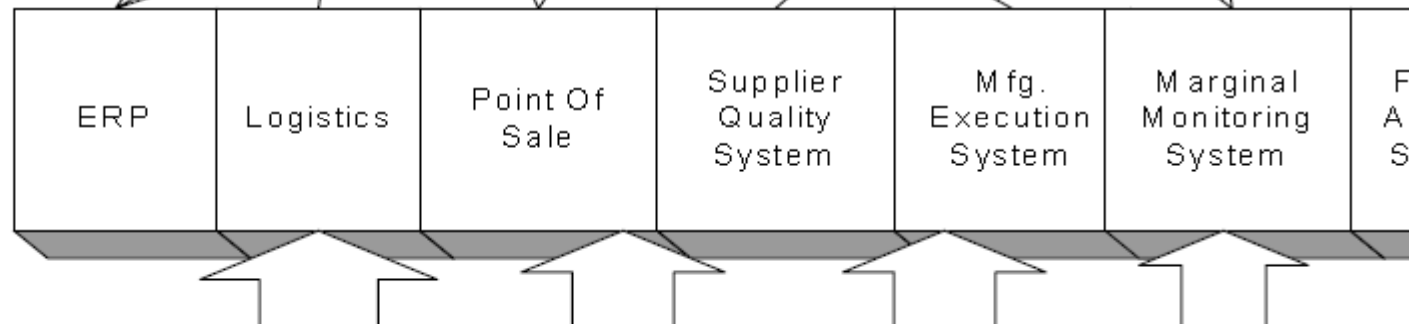
Dashboards
Highly Summarized
Key Metric Driven
Visualization and Alert



Business Intelligence
Cross Application Query / Data Mining
Statistical Analysis



Functional Applications
Transaction Based
Standard Reporting
Highly focused



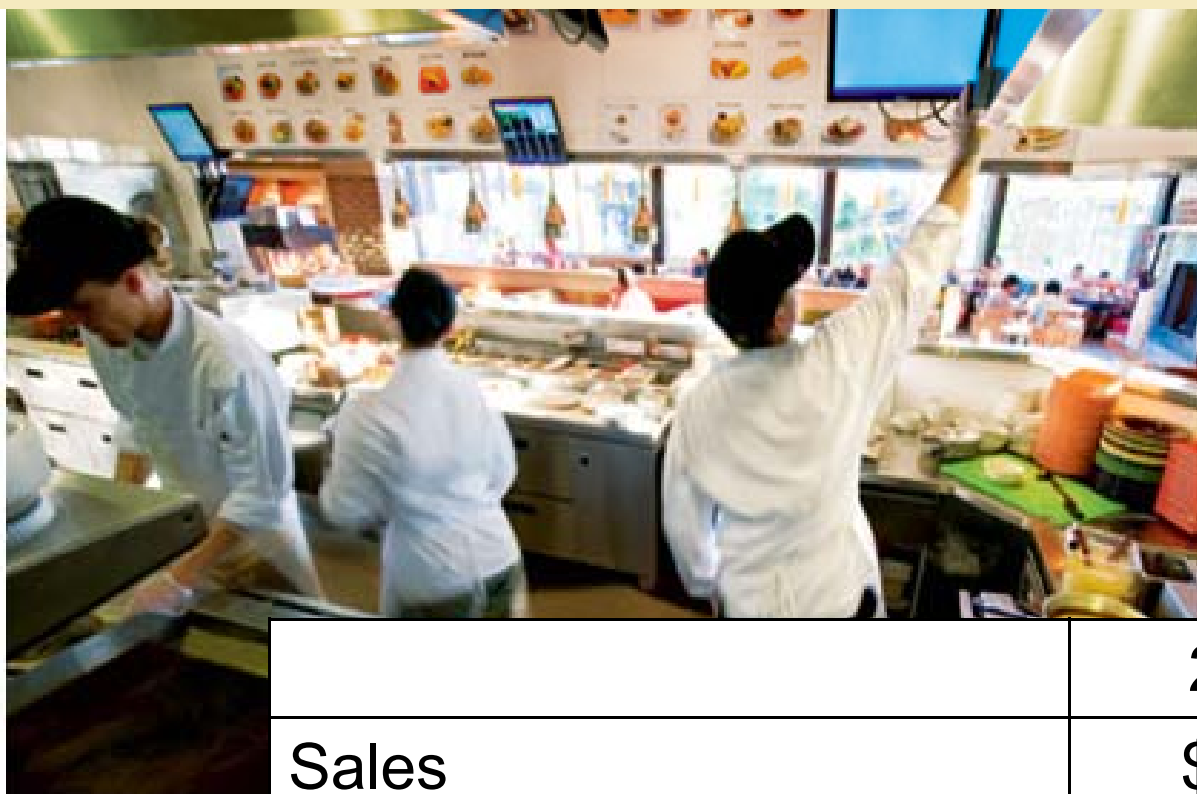
Factory Dashboards

Core requirements:

1. Show the health of the factory by providing near-real-time, graphical views of KPIs.
2. Show when a KPI goes below 2 sigma of its allowable value.
3. Give staff ways to drill down on each KPI to find the source of a problem.
4. Automatically issue alerts to the individuals responsible for a KPI so they can initiate damage control.

Corporate Dashboards

1. Billings and returns
2. Backlog
3. Outlook
4. Finished goods inventory
5. Distributor inventory and sell-through
6. Point of sale
7. Planned shipments
8. Finished goods in transit
9. Revenue recognition
10. Customer/channel status

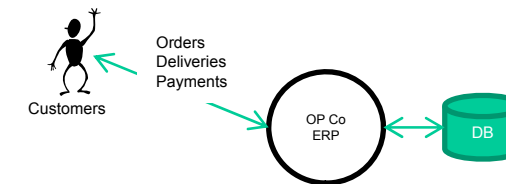
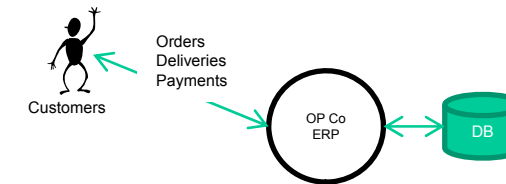
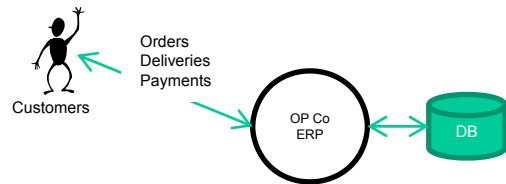
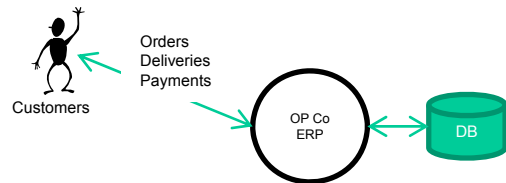
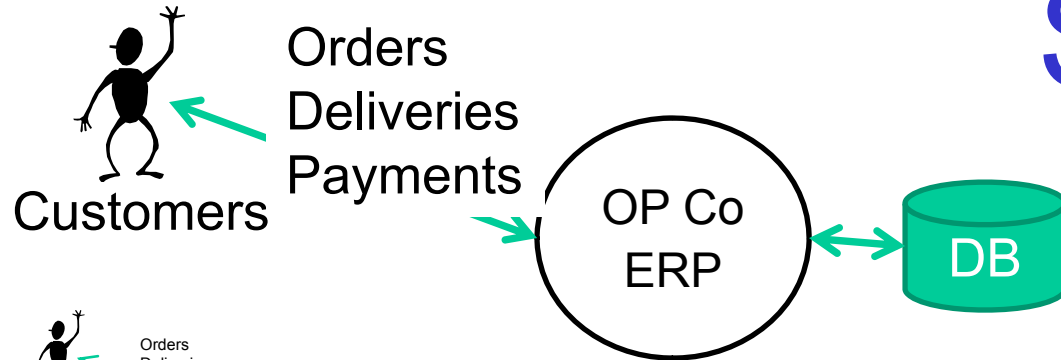


Welcome to Sysco

Sysco is the global leader in selling, marketing and distributing food products to restaurants, healthcare and educational facilities, lodging establishments and other customers who prepare meals away from home. Its family of products also includes equipment and supplies for the foodservice and hospitality industries.

	2009	2003
Sales	\$37B	\$26B
Net Earnings	\$1B	\$0.78B
Employees	47K	46K
Operating Companies	140	100

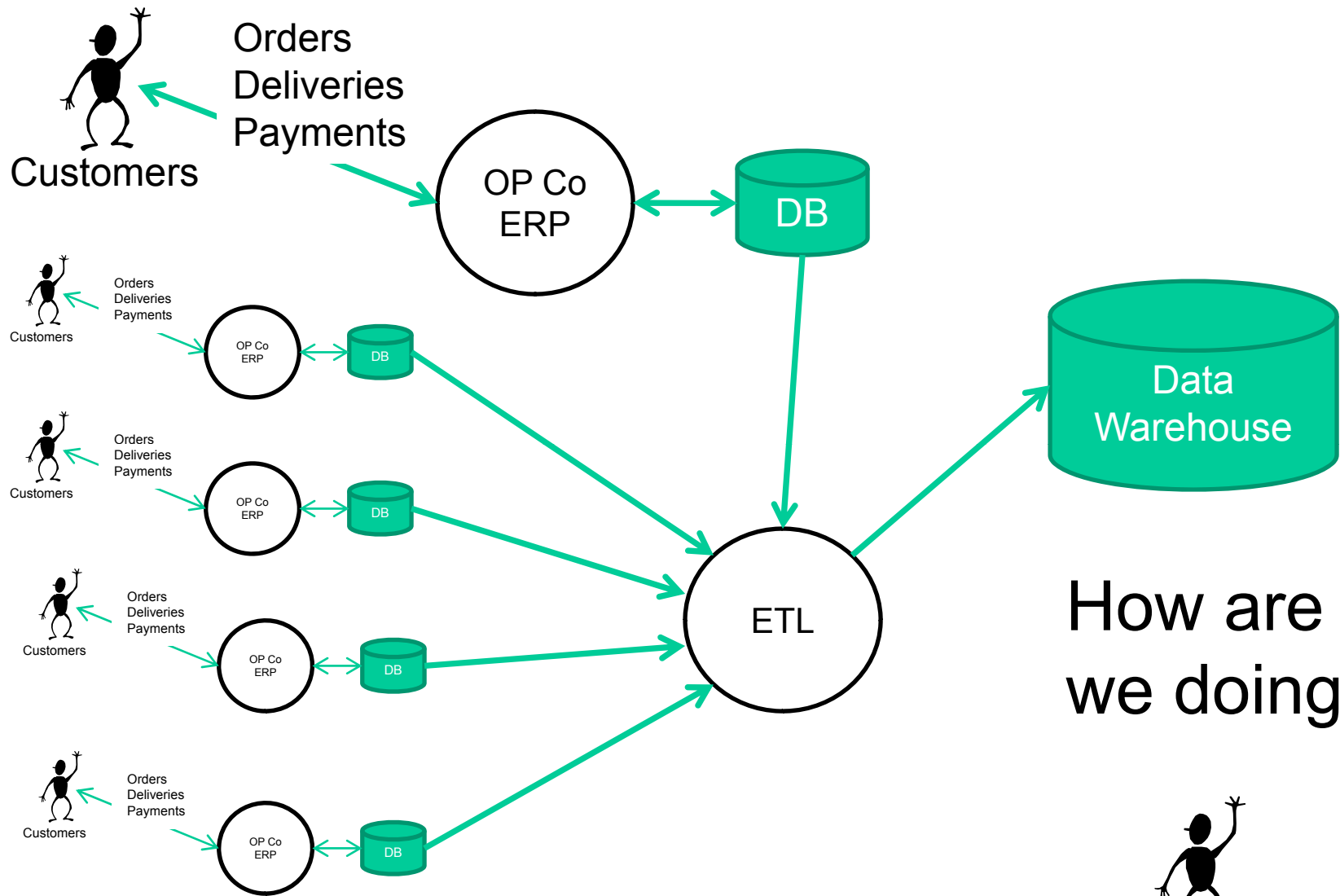
SYSCO



How are we doing?



Corporate
Headquarters



How are we doing?


Corporate Headquarters

Data Warehouse Structure

Mirror the Operational Database

- Easy to load data (copy from operational)
- Maximum flexibility and data content for existing data
- Massive storage and processing requirements
- Difficult or impossible to develop BI applications (if dimensions or performance measures are not included in operational databases)

Dimensional Data Model (Star Schema)

- More difficult to load data (must transform data from the operational databases and integrate with external data to represent performance measures and dimensions)
- Potentially less flexibility and data detail
- Reduced storage and processing requirements
- Easy to develop BI applications

Business Objects

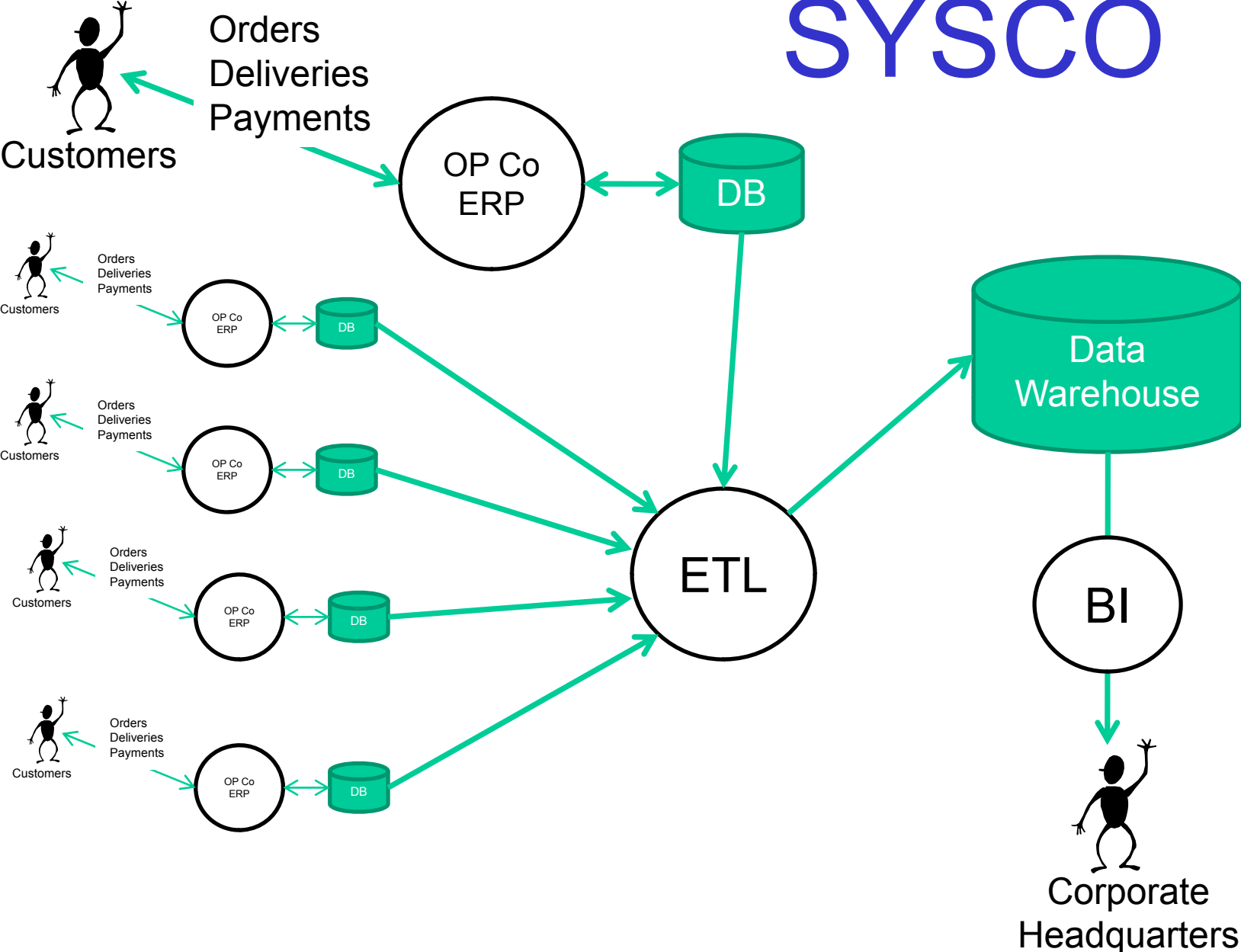
CSFs for Business Intelligence

1. Focus Your Efforts
2. Secure Executive Sponsorship
3. Build a Winning Project Plan
4. Make it Easy to Data Access
5. Make it Easy to Analyze Data
6. Make it Easy to Share Knowledge
7. Deliver Exceptionally Clean Data
8. Insist on Zero Client Administration
9. Implement Bullet-Proof Security
10. Plan for Growth

SYSCO BI Questions

- What additional products could we be selling to each of our customers?
 - Develop "typical purchase activity" by customer profile (size, type, geography, and so on)
- Which of our current customers are we most likely to lose?
 - Examine customers' ordering patterns over time (reduced volume may indicate high risk of loss)

SYSCO



Assess Costs and Benefits

Costs

- Hardware
- Software
- Personnel and Consulting

Benefits

- More and better information
- Improved decision-making
- Personnel savings
- Business process improvement
- Support for strategic objectives

Cost Scenario (\$000)	Bare Bones	Middle of the Road	Volume Discount
Module			
Query/Analysis	\$152	\$191	\$205
Performance Mgt	\$592	\$759	\$843
Report Creation	\$86	\$154	\$139
Report View	\$450	\$455	\$500
Analytical		\$170	\$205
Supply Chain			\$109
Total Software	\$1,280	\$1,729	\$2,001
Consulting	\$1,000	\$1,000	\$1,000
Maintenance	\$256	\$346	\$400
Total Cost	\$2,536	\$3,075	\$3,401

OLAP Cube

	Column Dimension Values							
Row Dimension Values								

Cells contain a cross-tabulation of performance measure values for each combination of Row and Column Dimension Values. Row and Column Dimensions typically support "drill-down" and "roll-up" capabilities.

What performance measures and dimensions would help SYSCO address its two questions?

OLAPQuery - Isys Data Warehouse

PivotTable To...

Home Create External Data Database Tools Add-Ins

View Show/Hide Selections Filter & Sort Data

Navigation Pane

ProductLine All

Region

	NJ	NY	PA	G
Industry	Sum of Revenue	Sum of Revenue	Sum of Revenue	S
CHEM	\$5,933.11			
CONT	\$148,272.65		\$211,977.59	
MNFR		\$78,870.98	\$26,499.55	
REFI	\$167,884.19	\$173,522.65	\$108,546.89	
UTIL		\$244,959.79		
Grand Total	\$322,089.95	\$497,353.42	\$347,024.03	

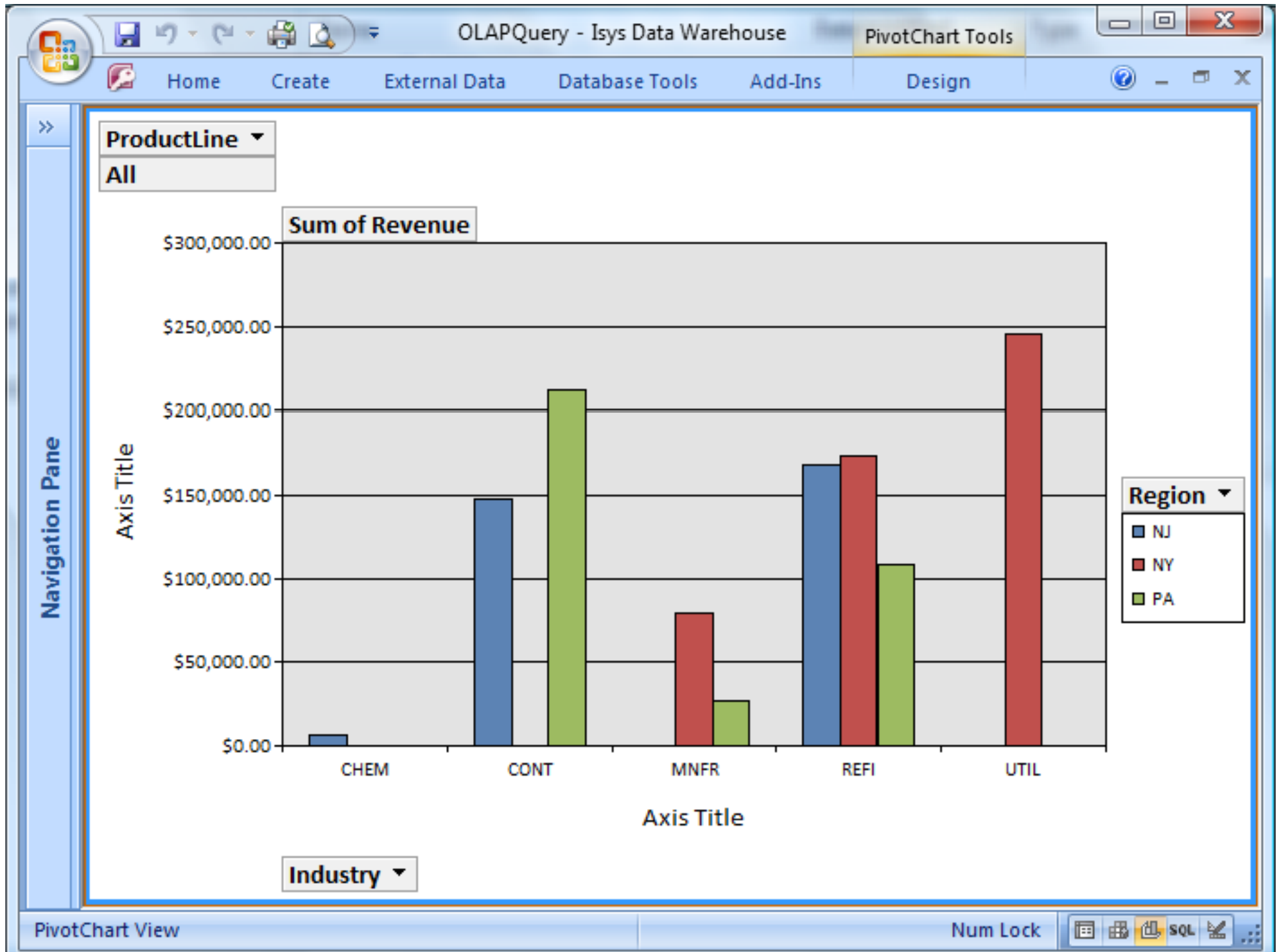
PivotTable Field List

Drag items to the PivotTable list

- OLAPQuery
 - Totals
 - Sum of Cost
 - Sum of Margin
 - Sum of Revenue
 - Average of Revenue
 - Average of Margin
 - ChannelType
 - PromotionType
 - ProductType
 - ProductLine**
 - ProductNo
 - ProductName
 - Industry**
 - Salesperson
 - Region**
 - Size
 - CustomerName
 - CustomerID
 - SaleDate
 - SaleDate By Week
 - SaleDate By Month

Add to Row Area

PivotTable View Num Lock



OLAPQuery - Isys Data Warehouse

Home Create External Data Database Tools Add-Ins

Navigation Pane

SaleDate By Month ▾
All

CustomerID ▾
1001 80998
+ - + -

Industry ▾	Region ▾	ProductLine ▾	Sum of Revenue	Sum of Revenue
☑ CONT	☑ NJ	10		\$133.20
		11	\$313.20	\$320.00
		21	\$6,358.80	
		31	\$30,747.20	\$572.95
		32	\$1,626.24	\$519.78
		34	\$3,650.40	\$961.70
		45	\$125.36	\$3,045.44
		51	\$29,849.61	\$3,728.28
		54	\$4,772.28	\$310.13
		60	\$8,935.92	\$473.28
		64	\$1,303.86	\$2,668.32
		80	\$2,270.84	\$2,149.29
		95	\$454.00	\$3,711.20
		F1	\$233.60	\$4,111.62
		Total	\$90,641.31	\$22,705.19
	Total		\$90,641.31	\$22,705.19
Grand Total			\$90,641.31	\$22,705.19

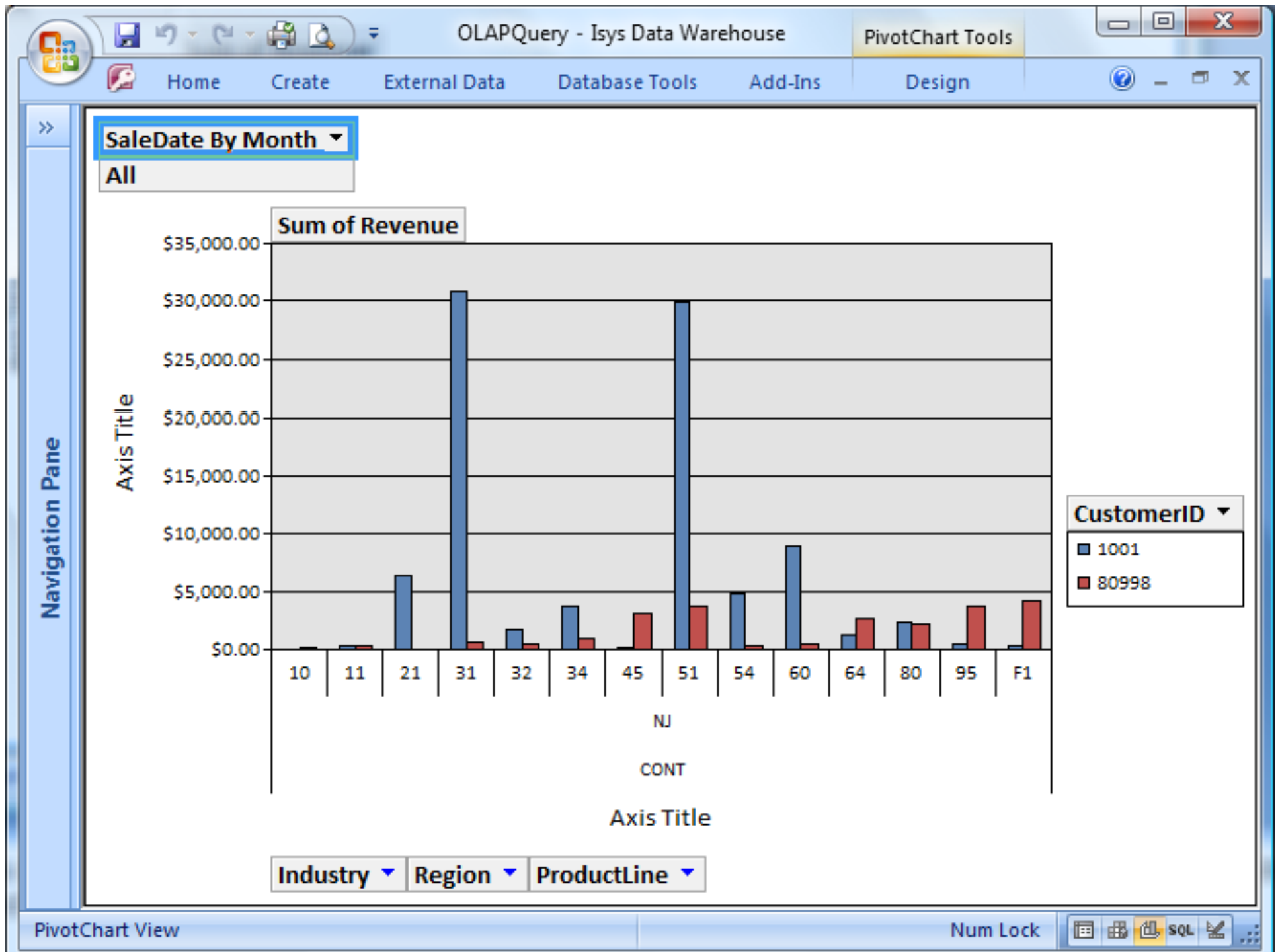
PivotTable View

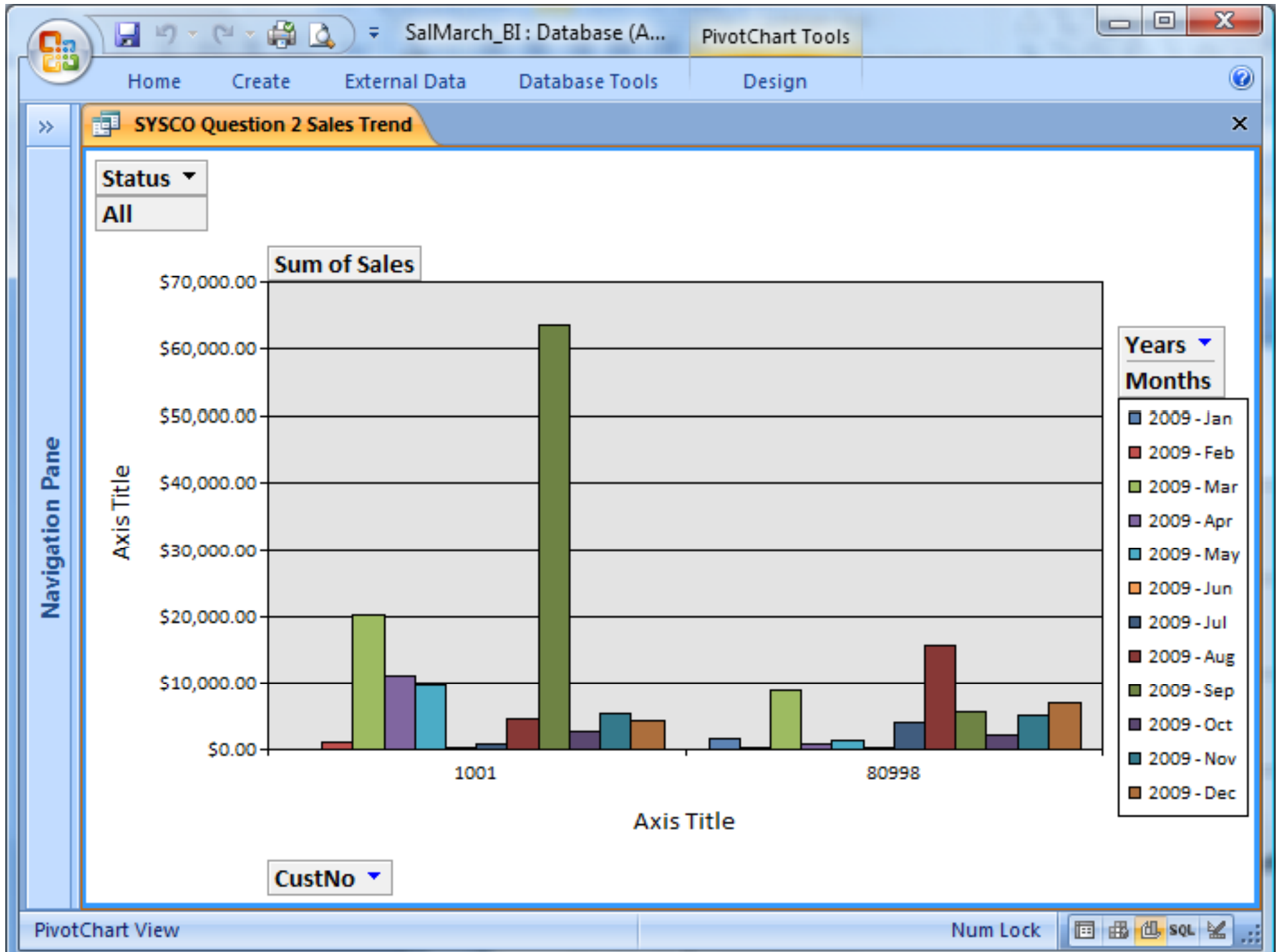
PivotTable Field List

Drag items to the PivotTable list

- OLAPQuery
 - Totals
 - Sum of Cost
 - Sum of Margin
 - Sum of Revenue
 - Average of Revenue
 - Average of Margin
 - ChannelType
 - PromotionType
 - ProductType
 - ProductLine**
 - ProductNo
 - ProductName
 - Industry**
 - Salesperson
 - Region**
 - Size
 - CustomerName
 - CustomerID**
 - SaleDate
 - SaleDate By Week
 - SaleDate By Month**
 - Revenue

Add to Row Area





Data Warehouse Structure

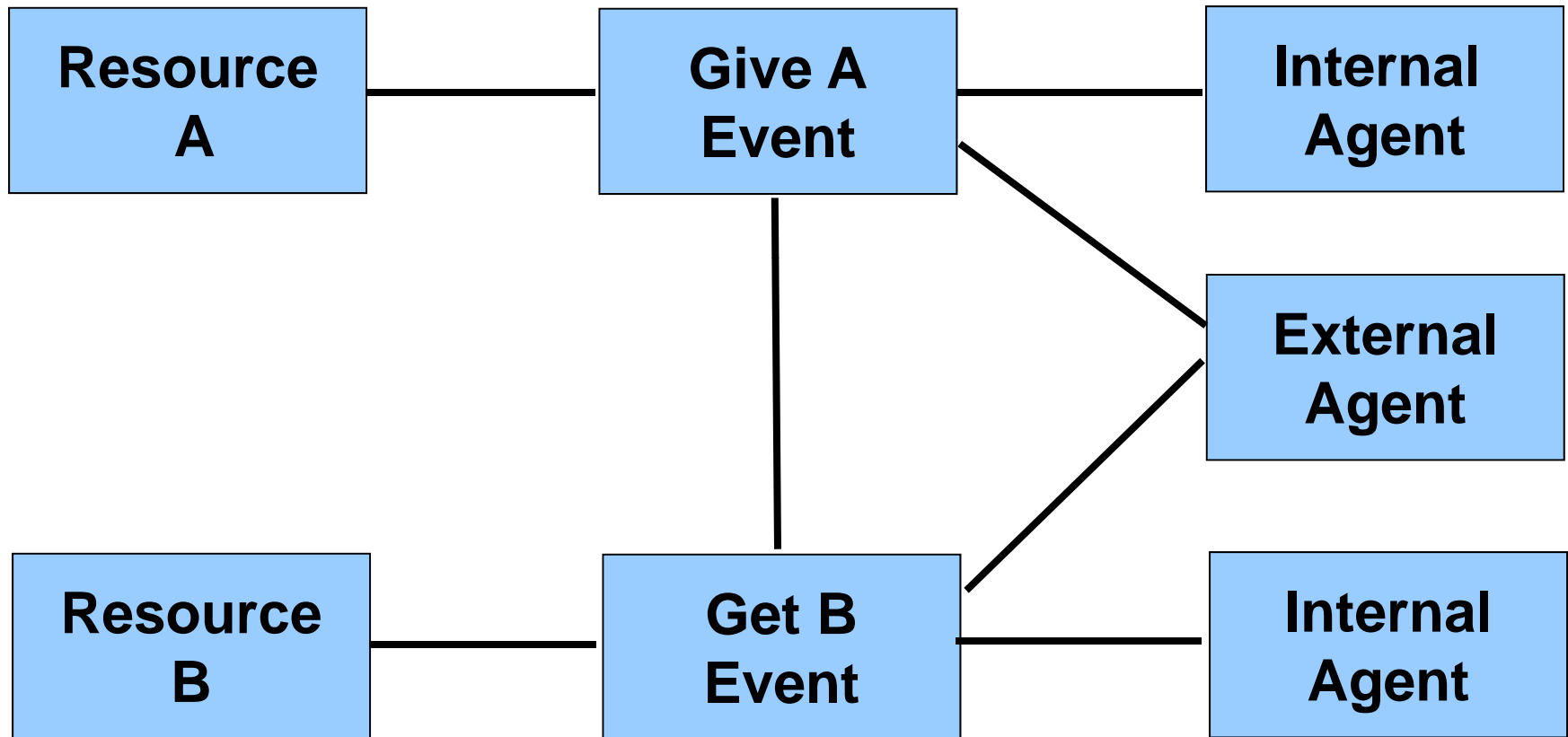
Mirror the Operational Database

- Easy to load data (copy from operational)
- Maximum flexibility and data content for existing data
- Massive storage and processing requirements
- Difficult or impossible to develop BI applications (if dimensions or performance measures are not included in operational databases)

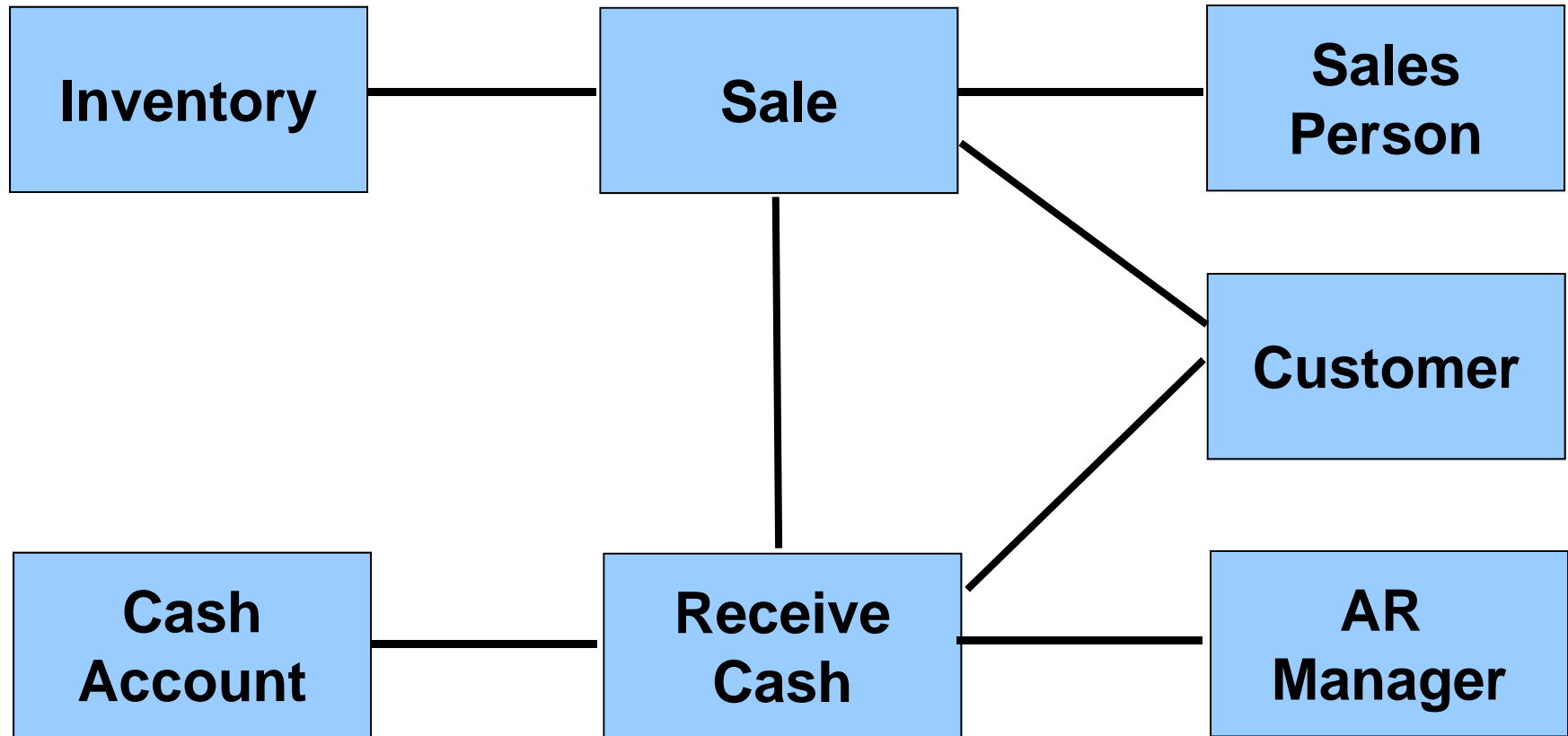
Dimensional Data Model (Star Schema)

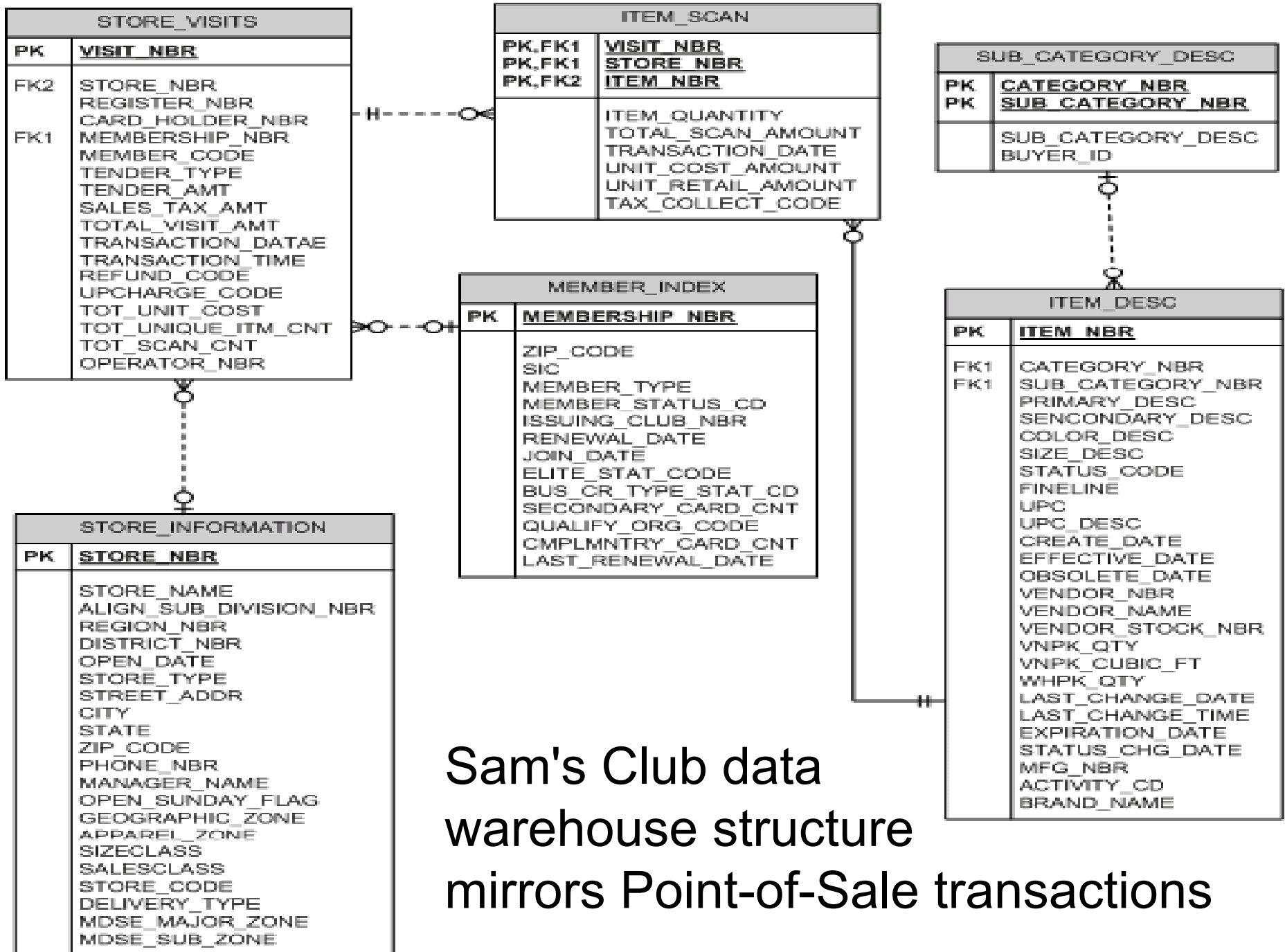
- More difficult to load data (must transform data from the operational databases and integrate with external data to represent performance measures and dimensions)
- Potentially less flexibility and data detail
- Reduced storage and processing requirements
- Easy to develop BI applications

Operational DB REA Model



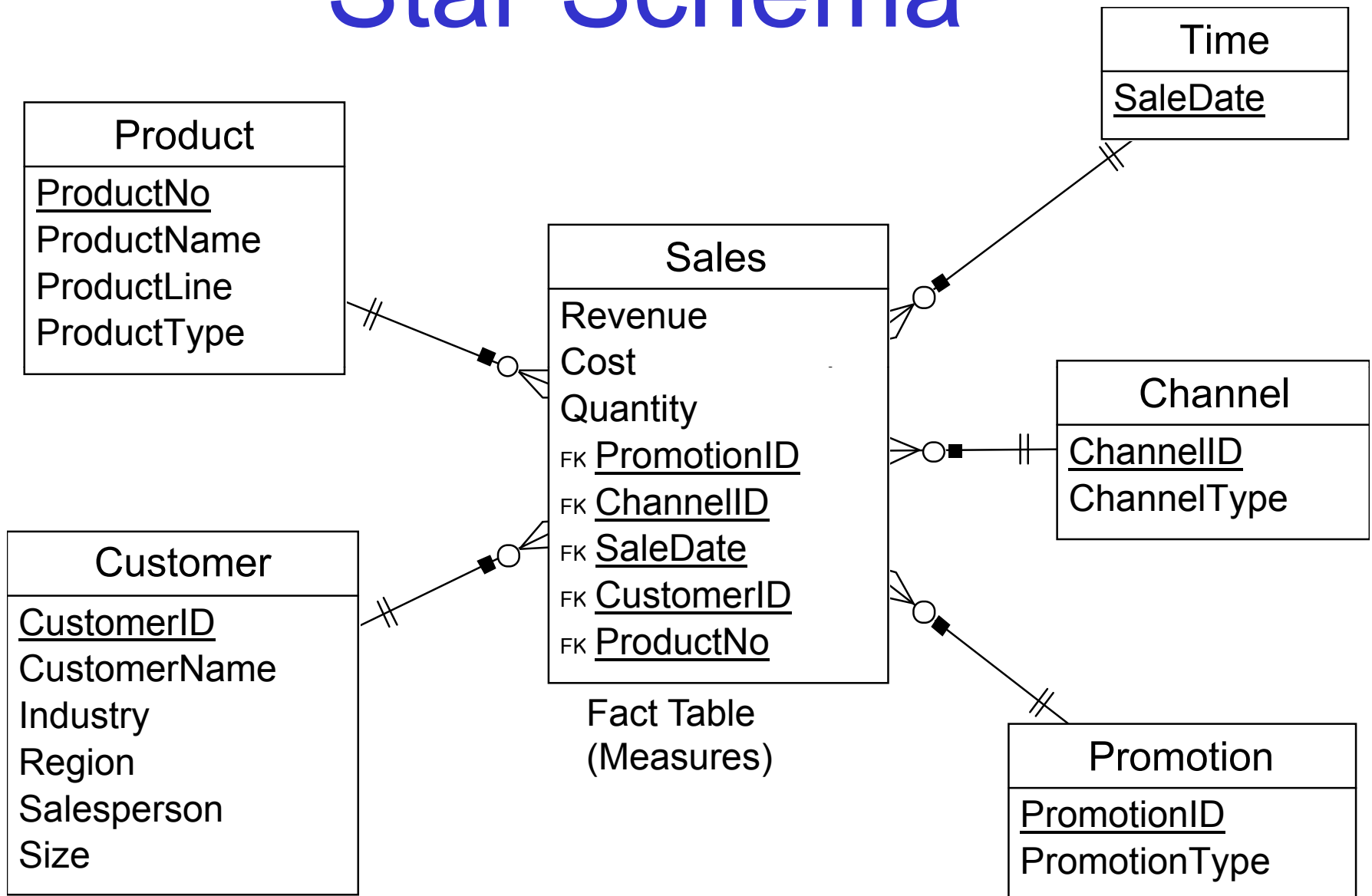
Revenue Cycle Model





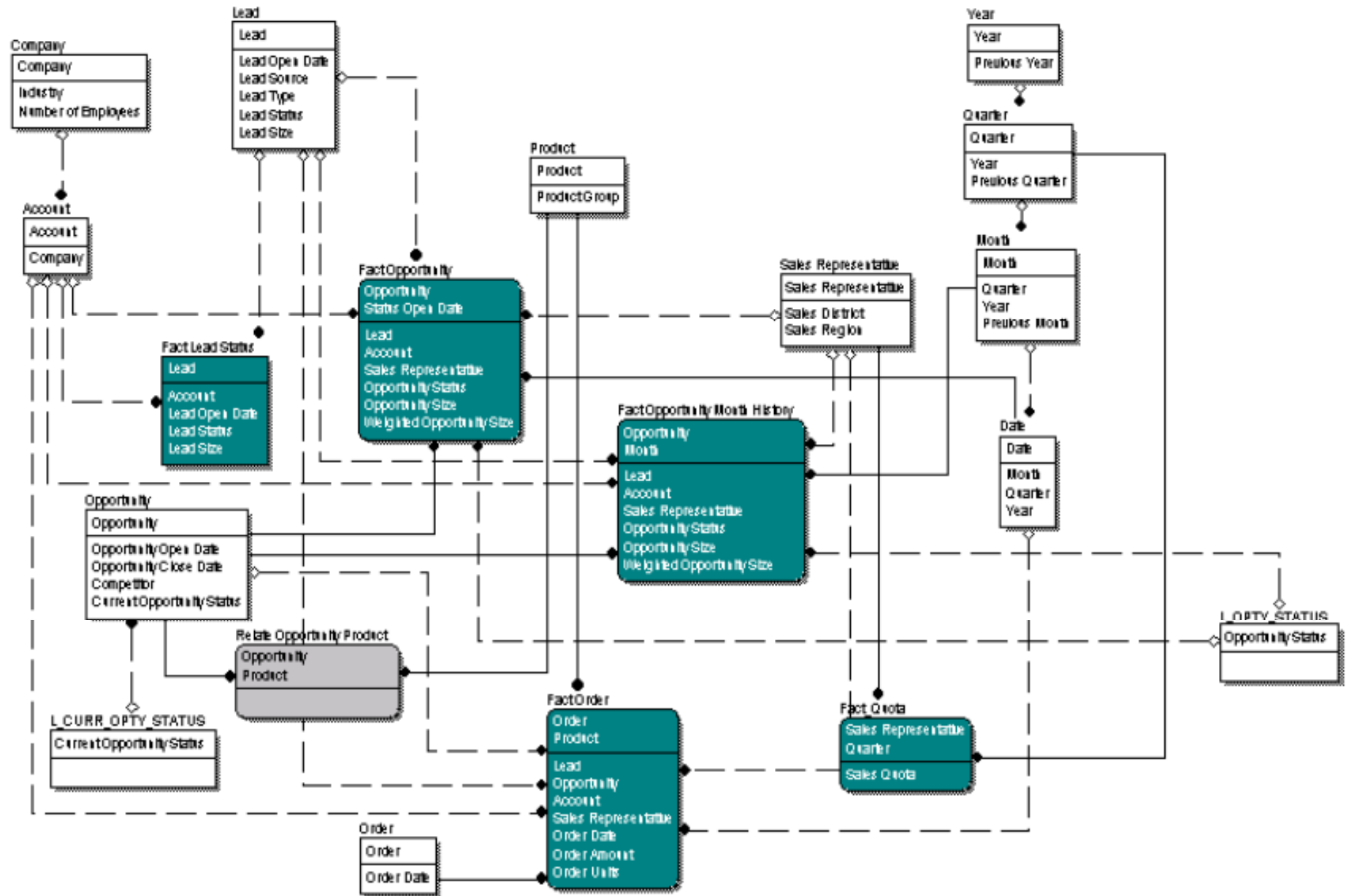
Sam's Club data
warehouse structure
mirrors Point-of-Sale transactions

Star Schema

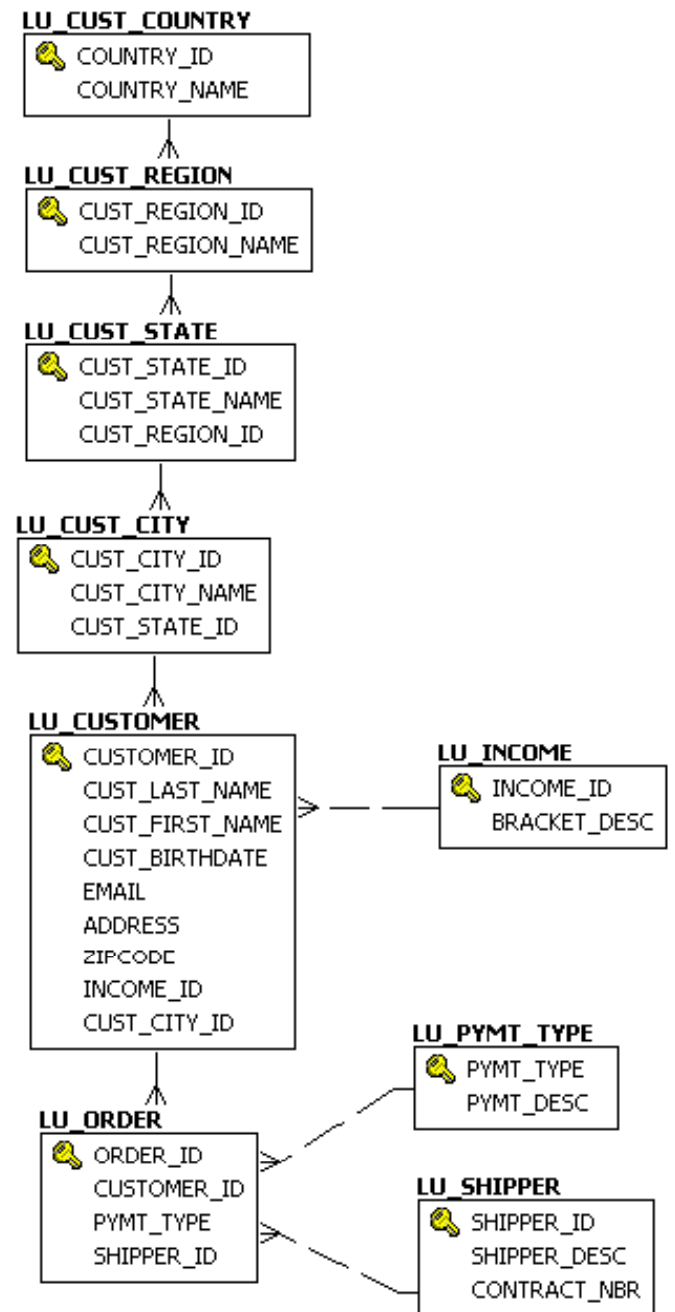
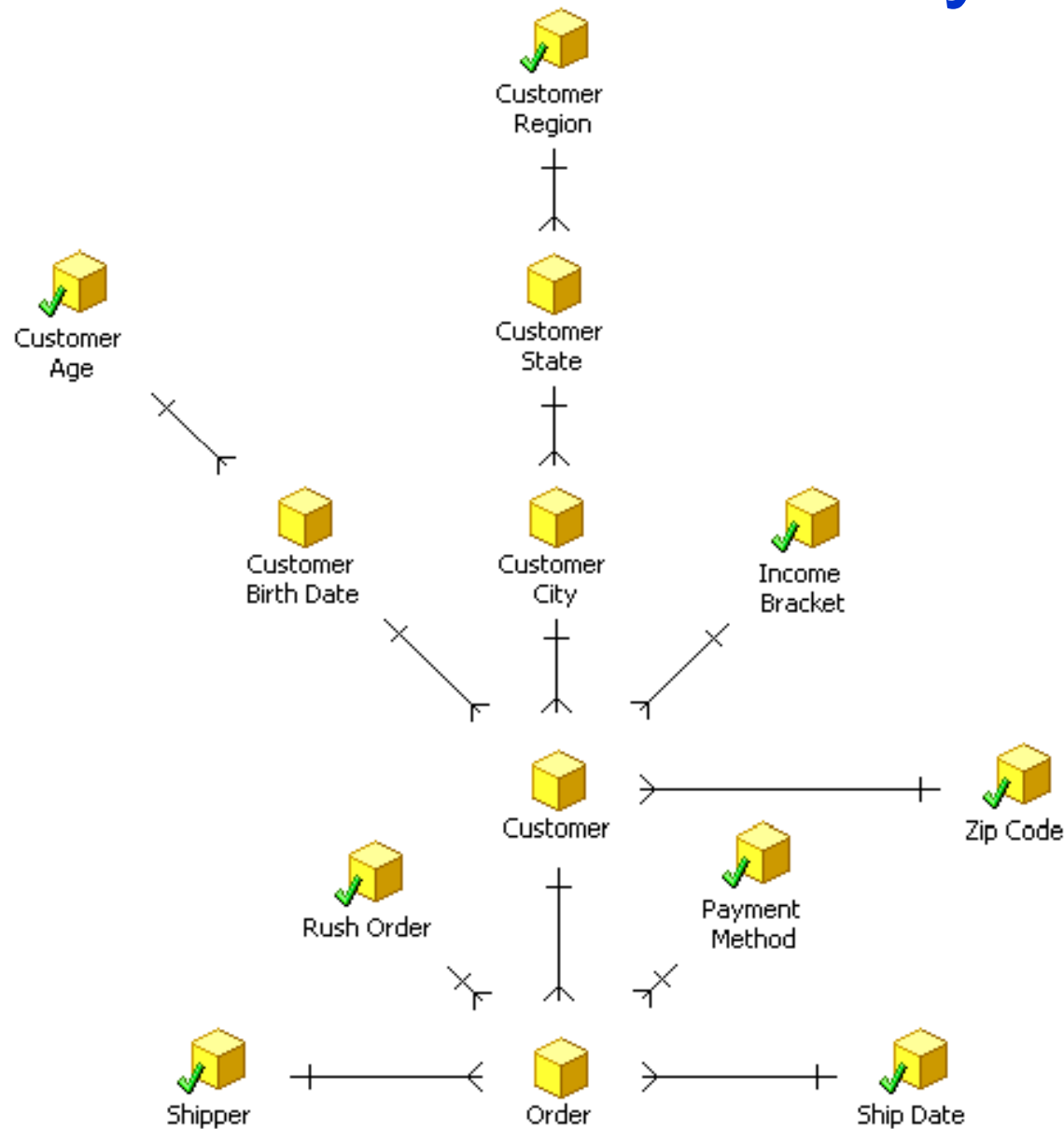


3.1 SALES – LOGICAL MODEL

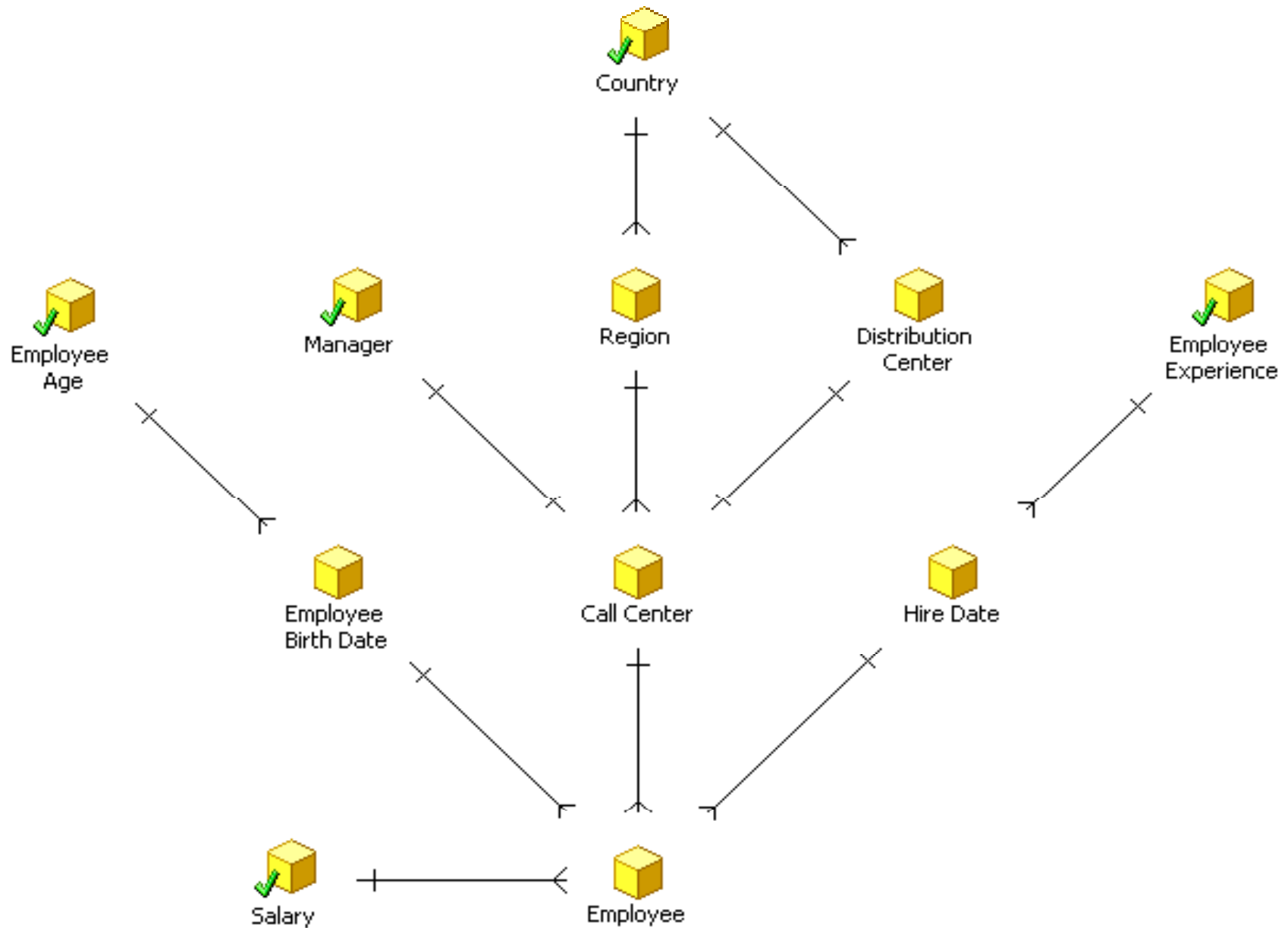
The following graphical view represents the logical model shipped with the module.



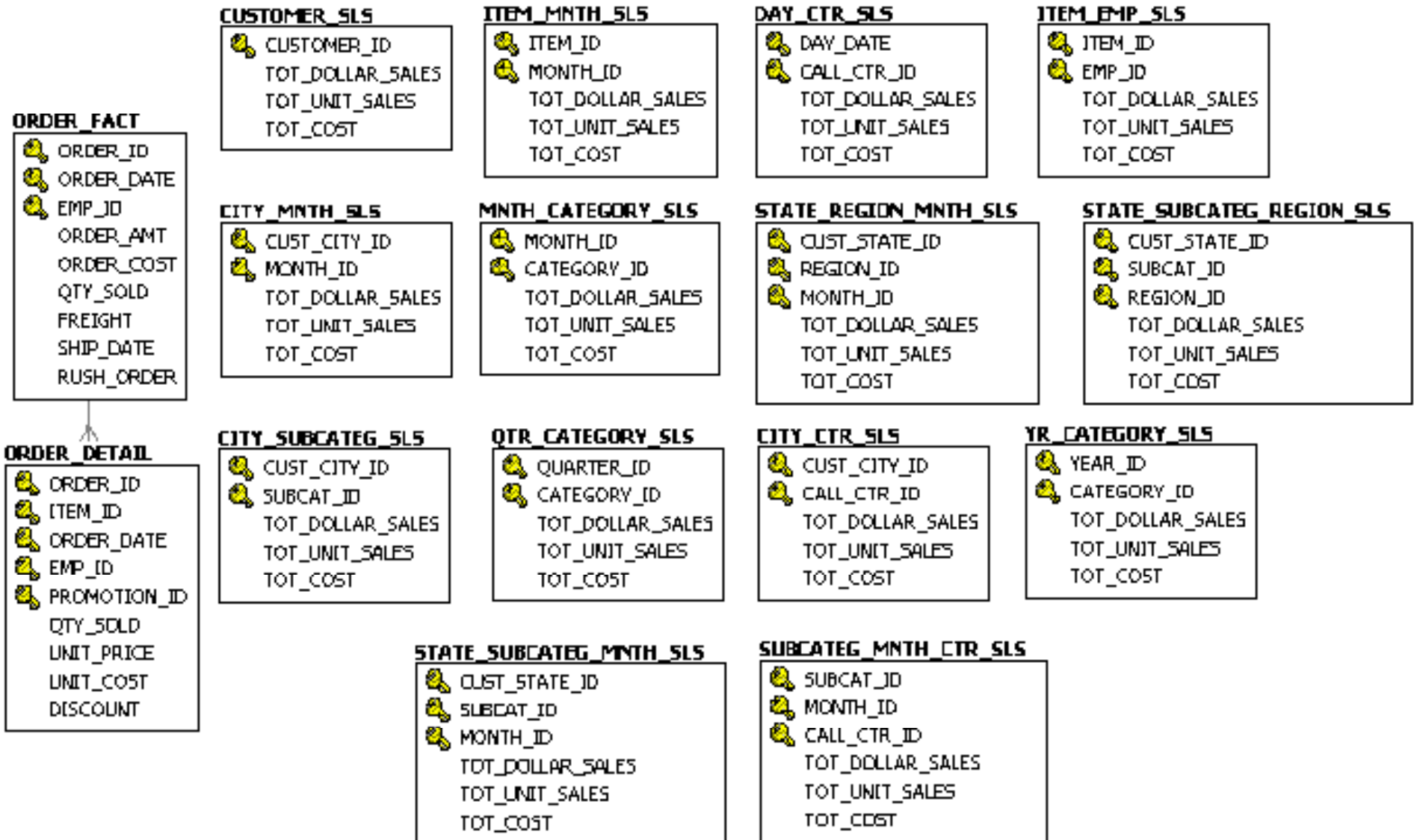
Customer Hierarchy



Geography Hierarchy



Fact Tables



Shared Reports. MicroStrategy 8 - Windows Internet Explorer

http://msi81web.teradata.ws/MicroStrategy/asp/Main.aspx?evt=3002&src=

Google

Shared Reports. MicroStrategy 8

My Reports Create Report Create Document History List Preferences Search

Sales Force Analysis Module > Shared Reports

Lead Analysis

Owner: Administrator
Modified: 1/14/04 7:16:11 AM
Lead Analysis reports provide insight into sales lead generation, lead qualification and lead conversion. These reports identify the most popular sources for leads, characteristics of leads with the highest ROI.

Pipeline Analysis

Owner: Administrator
Modified: 1/2/05 9:58:00 PM
Pipeline Analysis reports provide insight into all open opportunities and deals in the sales pipeline. These reports measure the current status of the sales pipeline, detect changing trends and key events, and identify key open opportunities.

Product Sales Analysis

Owner: Administrator
Modified: 1/14/04 7:14:10 AM
Product Sales Analysis reports provide insight into trends in product sales across the sales organization and customer base. These reports identify product sales momentum and the role of various products in key sales cycles.

Sales Performance Analysis

Owner: Administrator
Modified: 9/25/08 2:36:59 PM
Sales Performance Analysis reports provide insight into the current and historical performance of the sales organization. These reports identify potential problems in the sales organization so that timely corrective actions can be taken.

Scorecards

Owner: Administrator
Modified: 12/23/04 2:55:28 AM
This folder contains personalized reports and dashboards that users view

Internet | Protected Mode: On 125%

MicroStrategy BI Interface

Create a Report

Choose Template, Builder, or Wizard

Drag and drop report elements

The screenshot displays the MicroStrategy BI interface during the 'Create Report' process. The top navigation bar includes 'Shared Reports', 'My Reports', 'Create Report', 'Create Document', 'History List', and 'Preferences', along with a search field. The breadcrumb trail shows 'MicroStrategy Tutorial > Create Report > Design Mode: Customer Analysis'. Below the navigation, there are buttons for 'Run Report', 'Save Report', 'Cancel', and 'Edit Report Filter'. The 'OBJECT BROWSER' on the left lists 'Report Objects' and 'All Objects', with a list of 10 items: Customer City, Customer Region, Customer State, Month, Quarter, Year, Cost, Profit, Revenue, and Units Sold. The main workspace shows a table with a single column header 'Customer City' and a 'PAGE-BY: none' indicator. The text 'Customer Analysis Template' is overlaid at the bottom of the workspace.

Customer City

Customer Analysis Template

File View Data Format Last update: 9/15/09 9:17:51 AM

Agent

Row Axis Values Font Size

B I U \$ % ,

PAGE-BY: none ? X

1 2 3 4 5 of 9 page(s) Data rows: 1 - 50 of 436 Data columns: 1

Customer City	Customer Region	Customer State	Metrics	Avg Revenue per Customer
Addison	Central	Illinois		\$2,092
Akron	Central	Ohio		\$1,862
Albany	Northeast	New York		\$1,873
Albert City	Central	Iowa		\$2,004
Alexandria	Mid-Atlantic	Virginia		\$1,287
Allentown	Mid-Atlantic	Pennsylvania		\$1,442
Anderson	Mid-Atlantic	South Carolina		\$5,258
Annapolis	Mid-Atlantic	Maryland		\$1,689
Arden	Mid-Atlantic	North Carolina		\$1,343
Arlington Heights	Central	Illinois		\$1,569
Arlington	Mid-Atlantic	Virginia		\$2,160
Artesia	Southwest	New Mexico		\$1,951
Ashby	Central	Minnesota		\$1,704
Asheville	Mid-Atlantic	North Carolina		\$1,947
Ashland	South	Kentucky		\$1,502
Atlanta	Southeast	Georgia		\$1,783
Atlantic City	Mid-Atlantic	New Jersey		\$1,561
Auburn	Southeast	Georgia		\$1,655
Aurora	Southwest	Colorado		\$1,400
Austin	Southwest	Texas		\$2.521

Agenda

- ✓ Introduction and Overview
- ✓ Case Studies
- ✓ Data Warehouse Representations
- ✓ Business Intelligence Tools
 - ✓ Reporting
 - ✓ OLAP
 - Data Mining
 - Predictive Analytics

Conclusions

Data Mining

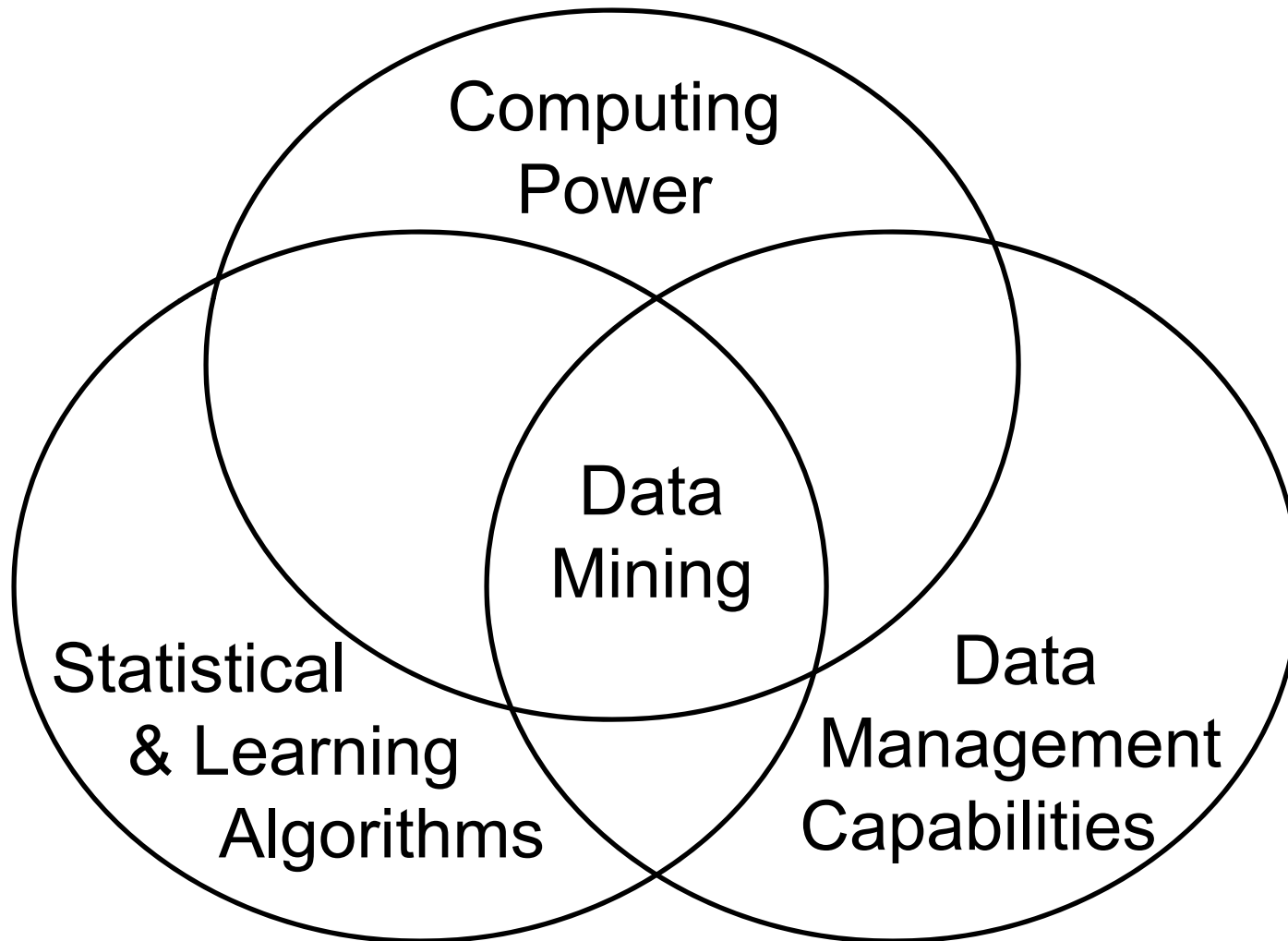
Is:

- Extracting useful information from large datasets.
- Exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.
- Discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques.

Is not:

- Identifying random correlations.

Key Technologies



Successful Applications

- Customer Segmentation
- Targeted Marketing (Predicting Response)
- Fraud Detection
- Predicting Customer Attrition
- Channel Optimization
- Predicting Loan Defaults
- Product Recommendations

Main Subdivisions

- **Supervised Learning**

The goal is to predict the value of an outcome measure based on a number of input measures (e.g., regression, logistic regression, discriminate analysis, neural networks).

- **Unsupervised Learning**

No output measure; the goal is to describe associations and patterns in the input measures (e.g., association rules, principal components, clustering).

Data Mining Challenges

- Personnel: domain experts, IT support, modelers
- Methodology: project management, problem definition, data acquisition, model development, knowledge deployment
- Technology Architecture: data warehouse, analytical tools

Major Players

- SAS (Enterprise Miner)
- Oracle (Darwin, Hyperion)
- IBM (Cognos, SPSS Clementine)
- SAP (Business Objects, Crystal Reports)
- Teradata Partners (Microstrategy, SAS, SAP, Microsoft, etc.)
- Microsoft (SQL Server, Excel, SharePoint, PowerPivot, Access)

Explore Data Clean Data Sample Data
Data Preparation

Classify Estimate Cluster Associate Forecast Advanced
Data Modeling

Accuracy Chart Classification Matrix Profit Chart Cross-Validation
Accuracy and Validation

A1 fx

	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										

Data Mining Process

1. develop an understanding of the problem
2. obtain the dataset
3. explore, clean and prepare the data (e.g. missing values, outliers)
4. reduce the dimensionality if necessary
5. determine the data mining task (classification, prediction, association rule discovery)
6. choose the data mining technique(s)
7. apply the technique(s), evaluate, compare and refine
8. interpret the results, choose the best model
9. deploy the selected model

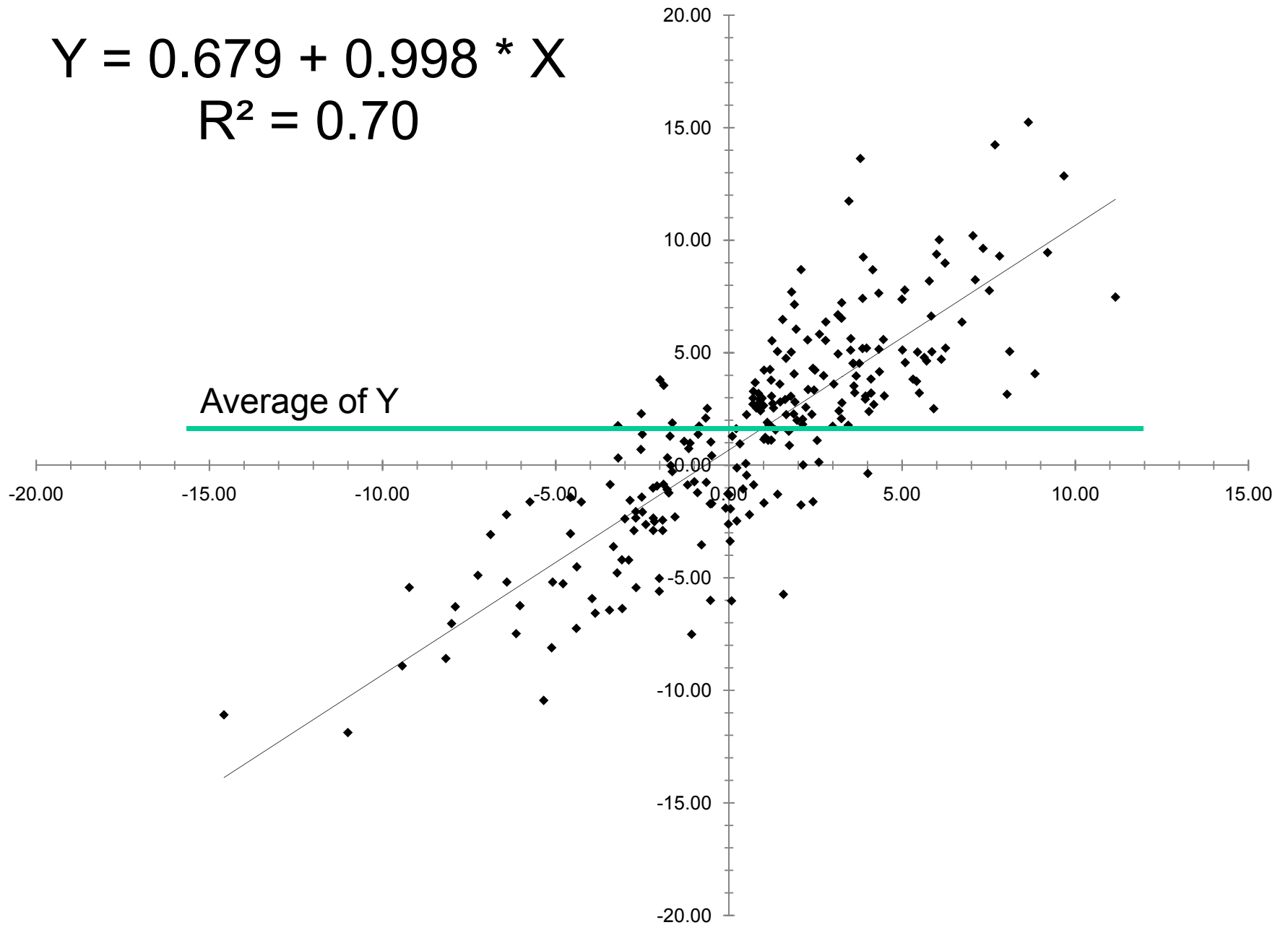
Data Mining Isn't a Good Bet For Stock-Market Predictions

"... data-mined numbers can be so irresistible that they are one of the leading causes of the evaporation of money [in the stock-market]."

Over a 13-year period annual butter production in Bangladesh "explained" 75% of the variation in the annual returns of the Standard & Poor's 500-stock index. Tossing in U.S. cheese production and the total population of sheep in both Bangladesh and the U.S. improved the "explanation" to 99%.

Wall Street Journal, August 8, 2009

$$Y = 0.679 + 0.998 * X$$
$$R^2 = 0.70$$

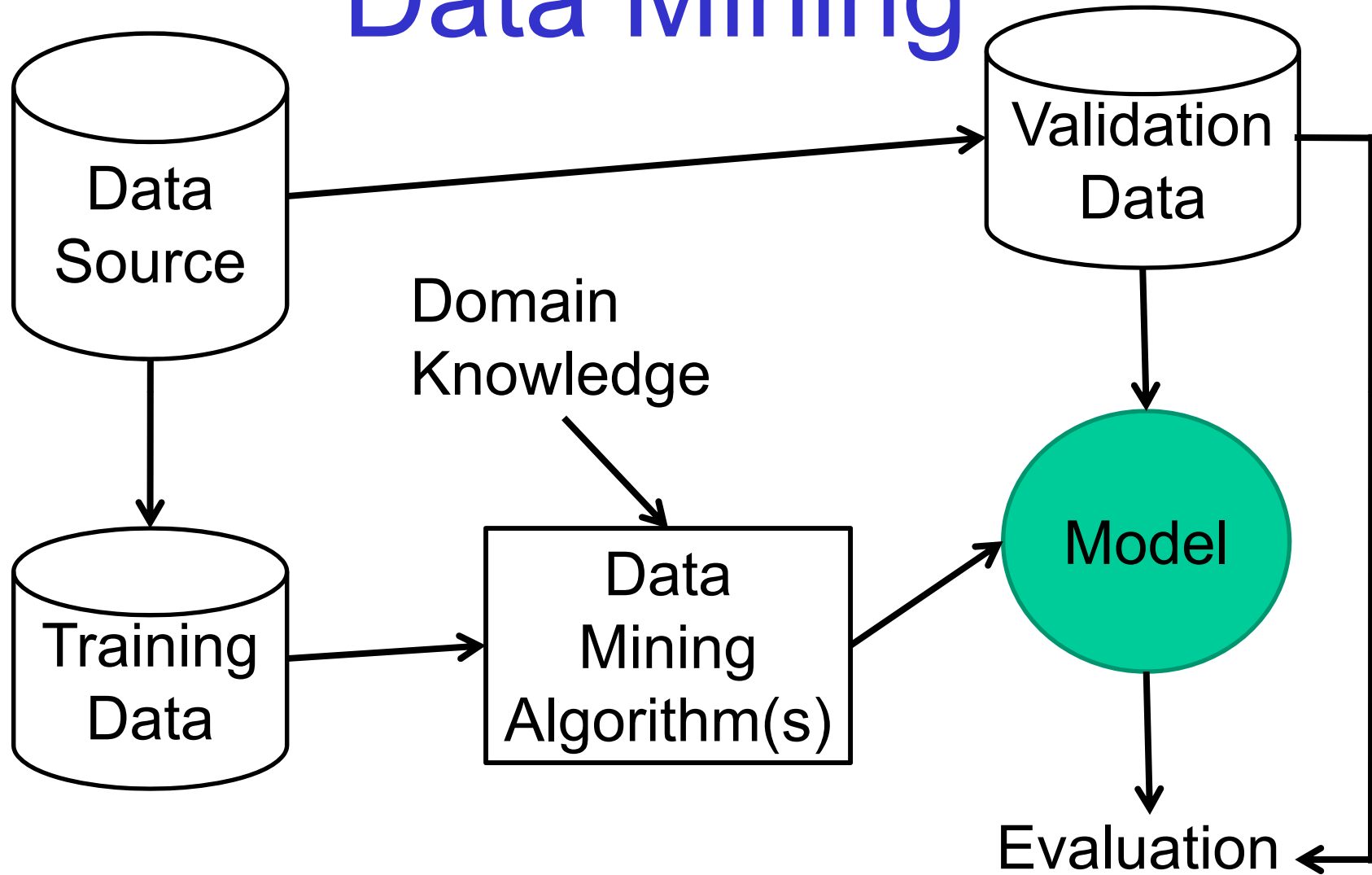


S&P 500 Returns 1950 - 2007					S&P 500 Returns		
				95% Conf Int	2008 - present		
	Mean	Std Dev	Lower	Upper	2008	2009	2010
Jan	1.36	4.67	0.13	2.59	-6.1%	-8.57%	-3.70%
Feb	-0.10	3.24	-0.95	0.75	-3.5%	-10.99%	2.85%
Mar	0.92	3.28	0.06	1.79	-0.6%	8.54%	5.88%
Apr	1.27	3.78	0.28	2.27	4.8%	9.39%	
May	0.30	3.57	-0.64	1.24	1.1%	5.31%	
Jun	0.19	3.35	-0.69	1.08	-8.6%	0.02%	
Jul	0.75	4.02	-0.31	1.80	-1.0%	7.41%	
Aug	0.00	4.67	-1.23	1.22	1.2%	3.36%	
Sep	-0.61	4.22	-1.72	0.50	-9.1%	3.57%	
Oct	0.93	5.14	-0.42	2.29	-16.9%	-1.98%	
Nov	1.58	4.37	0.43	2.73	-7.5%	5.74%	
Dec	1.61	3.18	0.78	2.45	0.8%	1.78%	

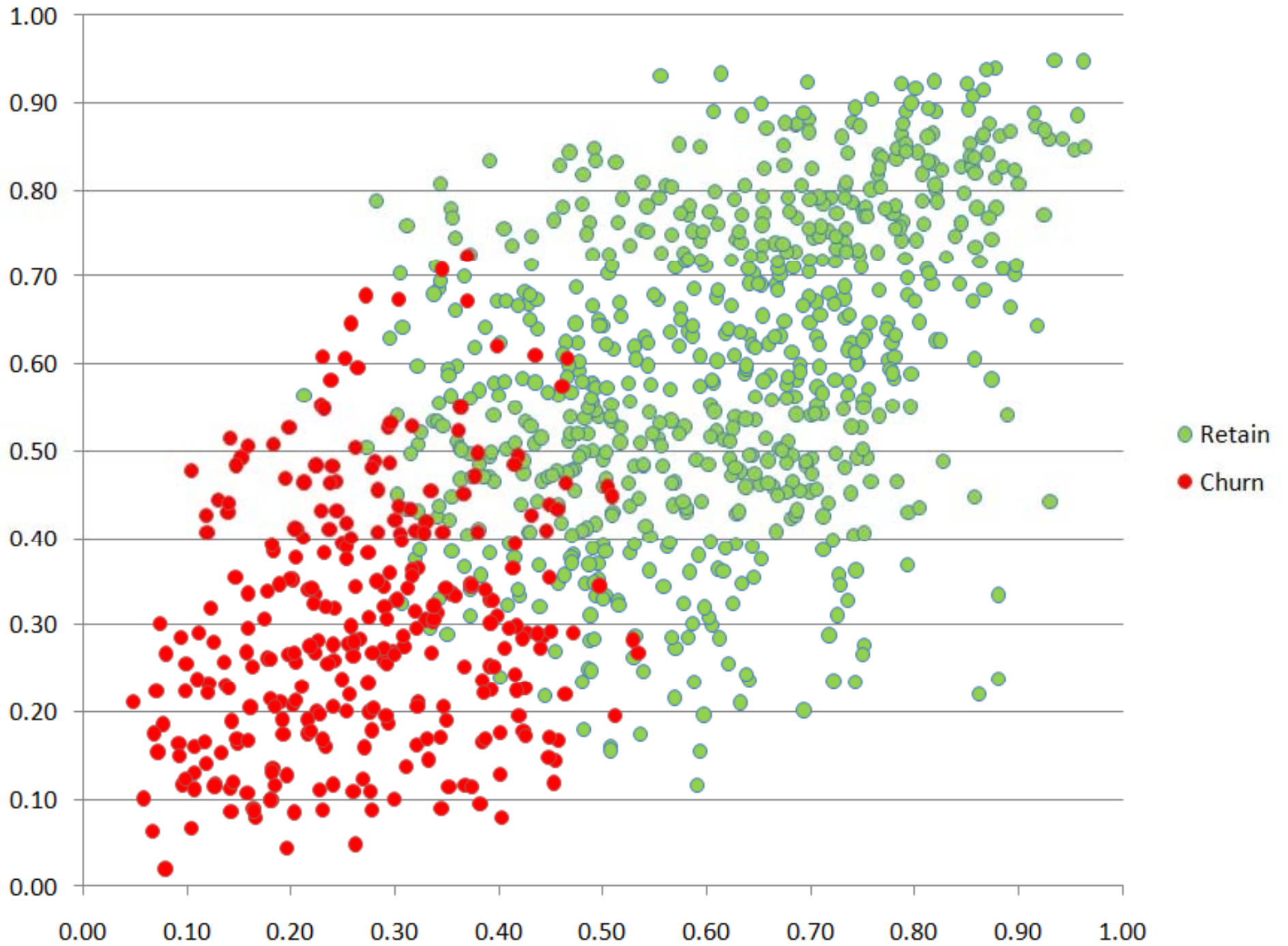
A \$1000 Investment

	2008	2009	2010	2008	2009	2010
Jan	\$939	\$833	\$1,026	\$939	\$641	\$904
Feb	\$939	\$833	\$1,026	\$939	\$641	\$904
Mar	\$933	\$905	\$1,086	\$933	\$696	\$957
Apr	\$978	\$990		\$978	\$761	
May	\$978	\$990		\$988	\$801	
Jun	\$978	\$990		\$903	\$802	
Jul	\$978	\$990		\$894	\$861	
Aug	\$978	\$990		\$905	\$890	
Sep	\$978	\$990		\$905	\$890	
Oct	\$978	\$990		\$752	\$872	
Nov	\$904	\$1,046		\$695	\$922	
Dec	\$912	\$1,065		\$701	\$939	

Data Mining



What data and algorithms would help SYSCO address its two questions?



Classification Techniques

Predict the best classification for an observation.

Example Business Tasks:

- Detect / Predict Fraud
- Predict Bankruptcy
- Predict Response to Marketing Promotion

Basic Techniques:

- Naïve Classifier (Predominant Class)
- Bayesian Classifier (Conditional Probability)
- K-Nearest Neighbors (Similarity)
- Classification Tree (Iterative Partitioning)

Cell Phone Insurance Claims

Carrier	Legitimate	Fraudulent	Total
A	174	19	193
B	79	22	101
C	26	15	41
D	135	61	196
E	522	86	608
Total	936	203	1,139

A sample of 1,139 out of 100,000 claims processed in a single month were investigated to determine if they were legitimate or fraudulent.

"Bayesian" Classifier

Carrier	Legitimate	Fraudulent	Total
A	90%	10%	100%
B	78%	22%	100%
C	63%	37%	100%
D	69%	31%	100%
E	86%	14%	100%
Total	82%	18%	100%

$$\Pr(\text{Fraudulent}) = 0.18$$

$$\Pr(\text{Fraudulent} \mid A) = 0.10$$

$$\Pr(\text{Fraudulent} \mid B) = 0.22$$

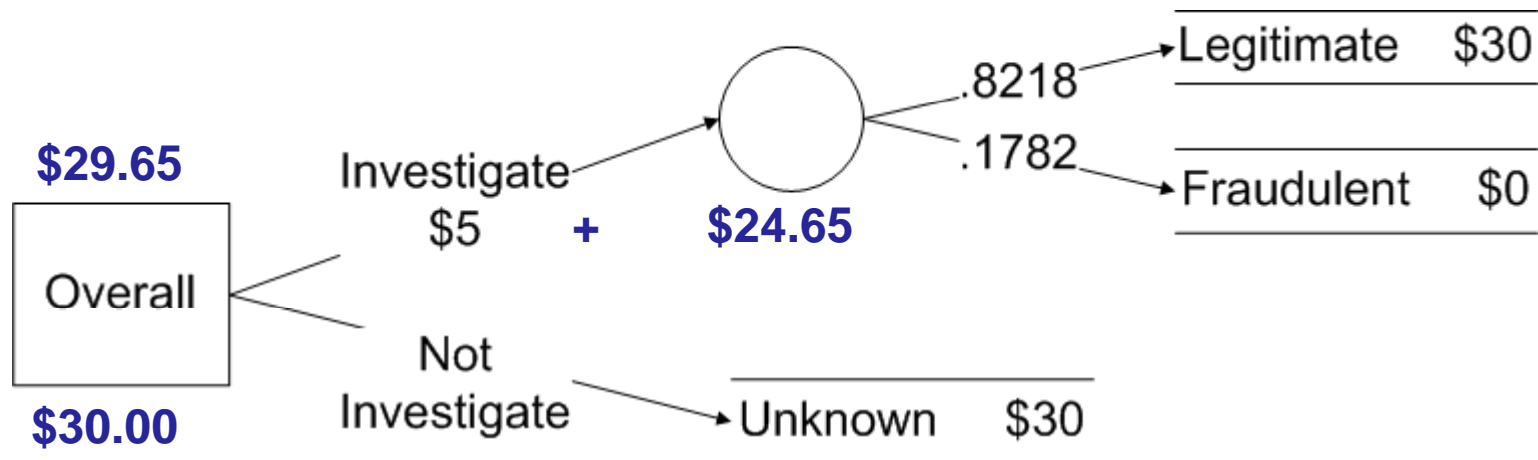
$$\Pr(\text{Fraudulent} \mid C) = 0.37$$

$$\Pr(\text{Fraudulent} \mid D) = 0.31$$

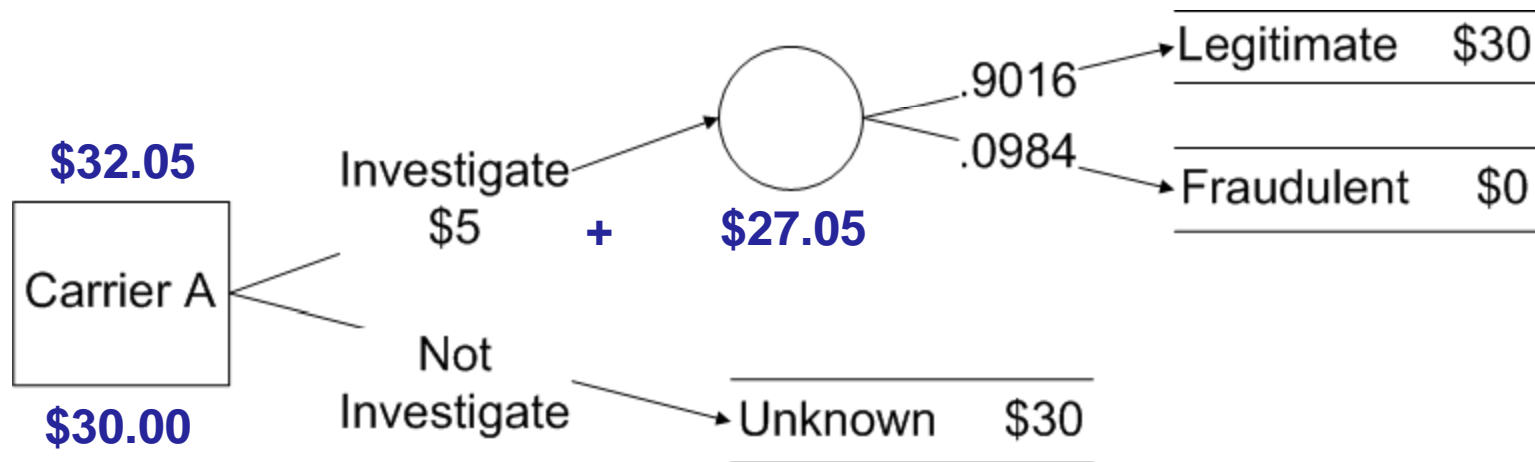
$$\Pr(\text{Fraudulent} \mid E) = 0.14$$

Claims Investigation

Suppose the cost to pay a legitimate claim is \$30 and the cost to investigate a claim is \$5. Then if no information about the carrier is used the conclusion is to investigate all claims.



Claims Investigation



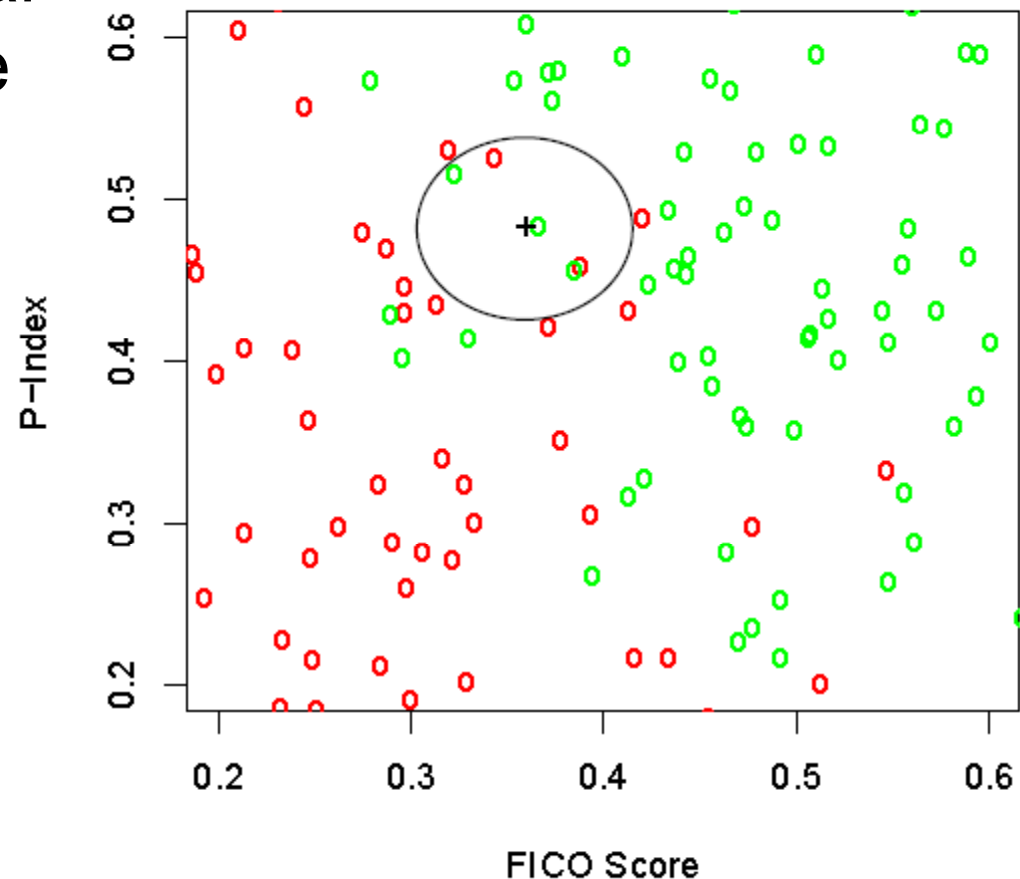
Including the *condition* that the claim came from Carrier A, the probability that the claim is legitimate changes, resulting in a different decision. What other factors change the probability of legitimacy?

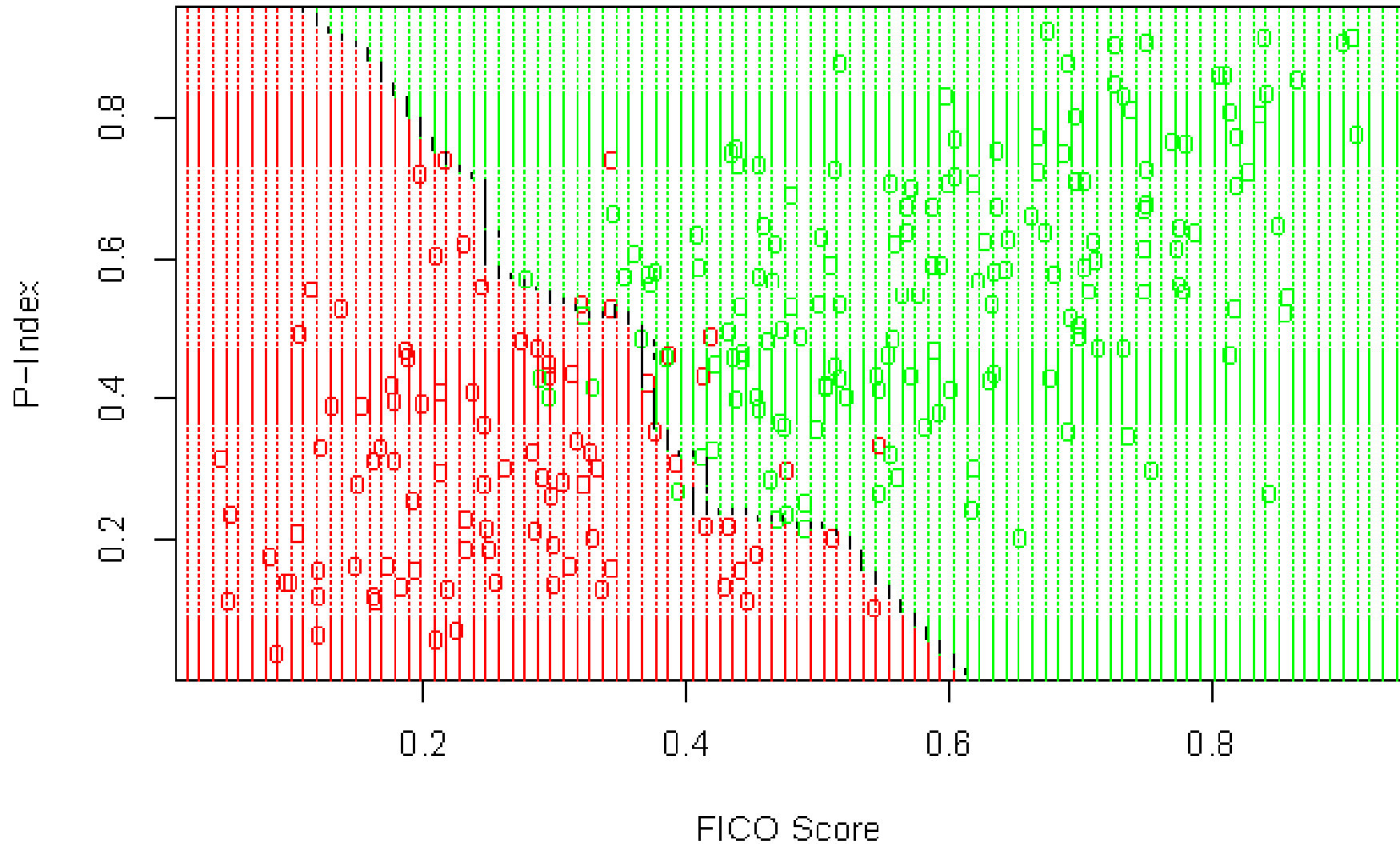
Credit Scoring

A bank would like to score loan applications based on the likelihood that the loan will be repaid. They plan to use two factors: FICO credit score purchased from Fair, Isaac & Company and a profitability index computed from a factors on the loan application such as the ratio of the loan amount to income and the interest rate of the loan. To develop the model they have gathered this data for 1000 past completed loans. Of these loans 700 have been paid in full (Default = 0), 700 have defaulted (Default = 1).

K-Nearest Neighbors

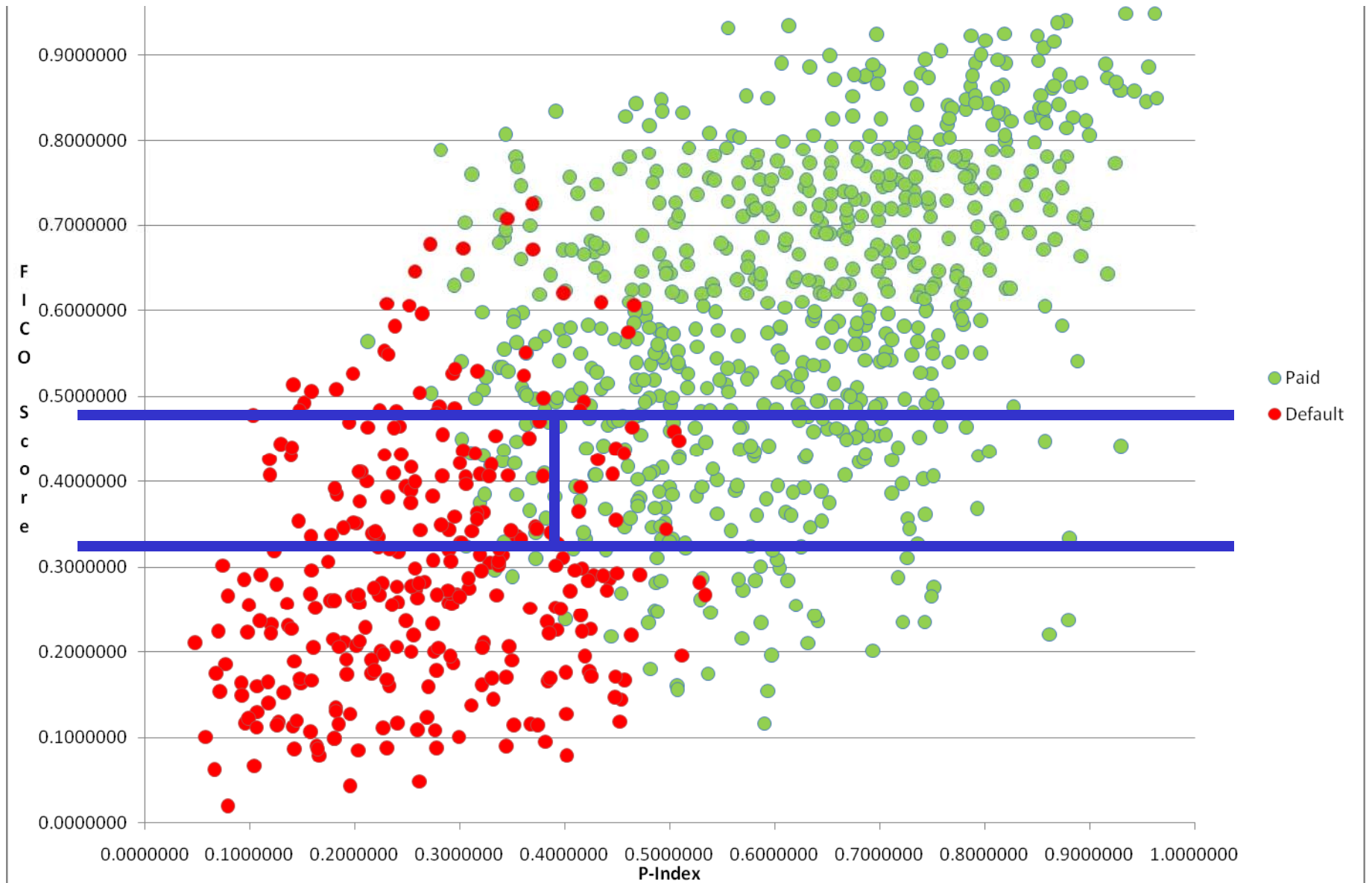
A k-nearest neighbors (k-NN) approach classifies a new observation according to the majority class of its k "most similar" observations in the training data set. "k" is a parameter selected during the training process. Small values of k are very sensitive to local neighbors.

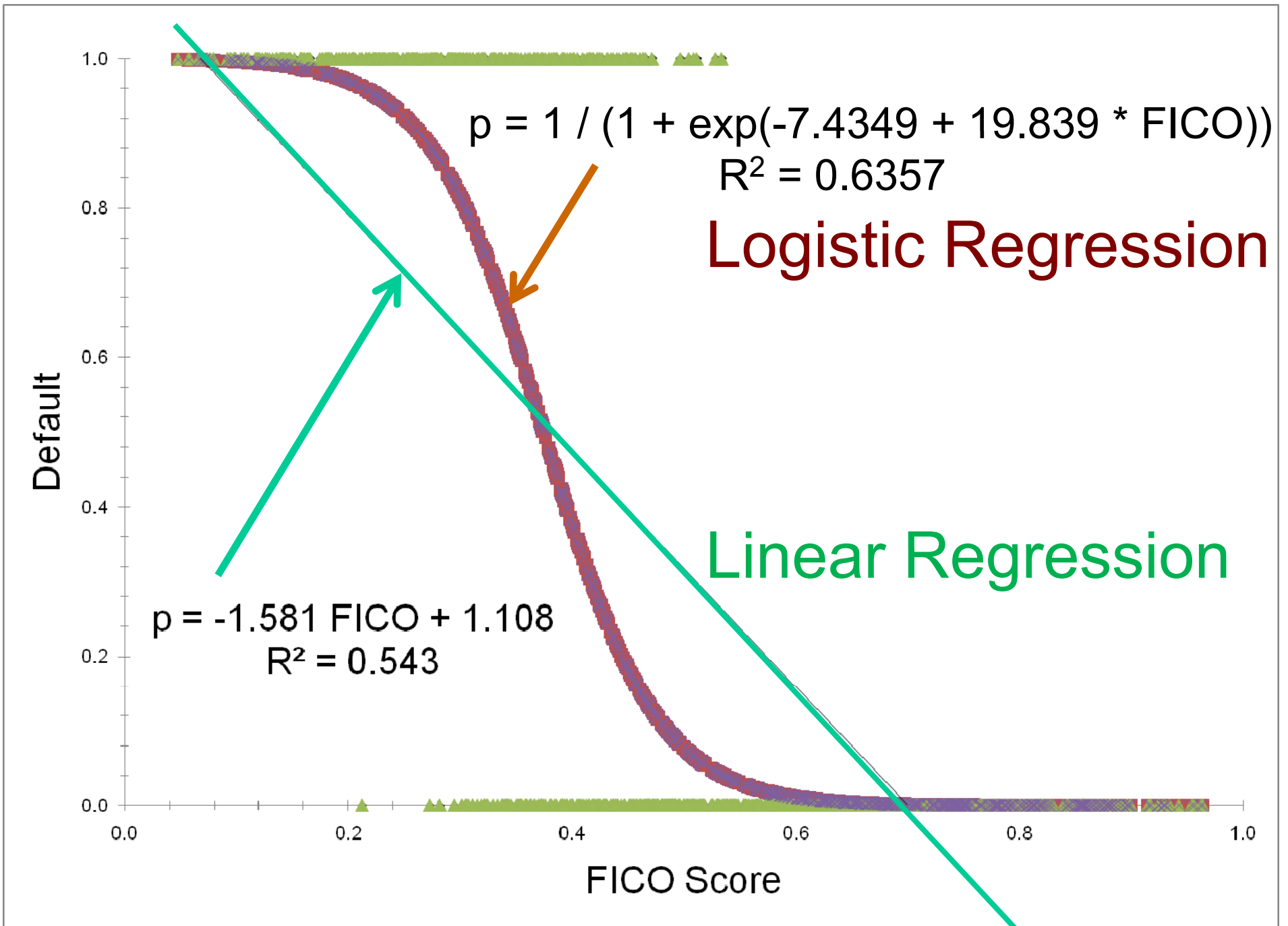




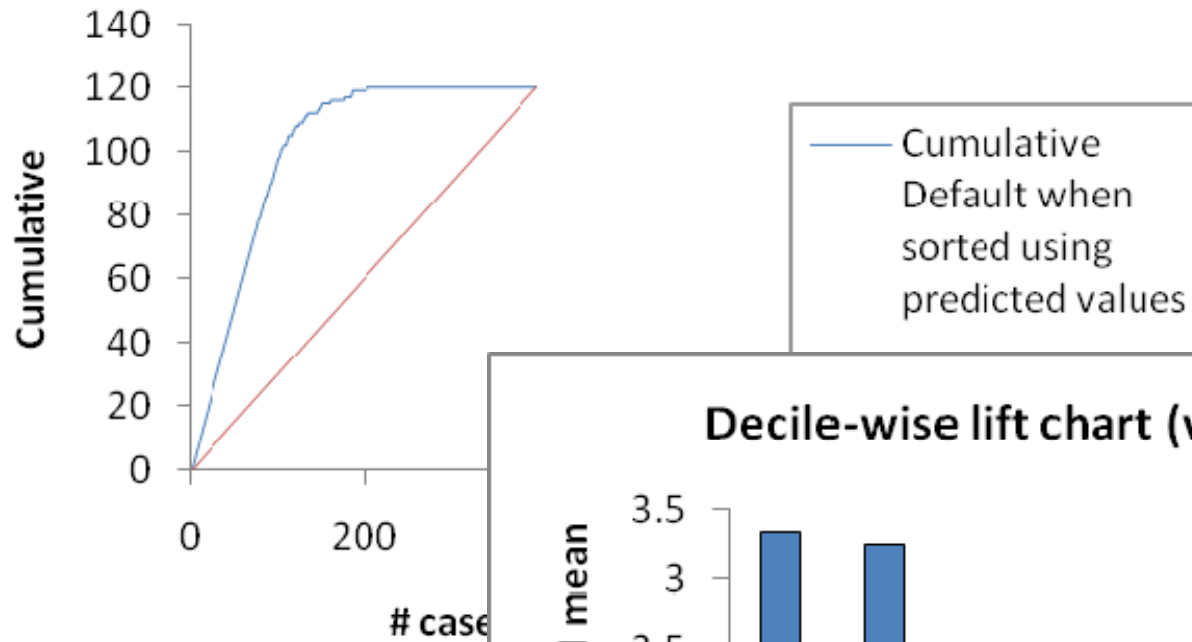
The 10-nearest neighbors procedure yields the above "classification frontier" for new observations.

Classification Tree

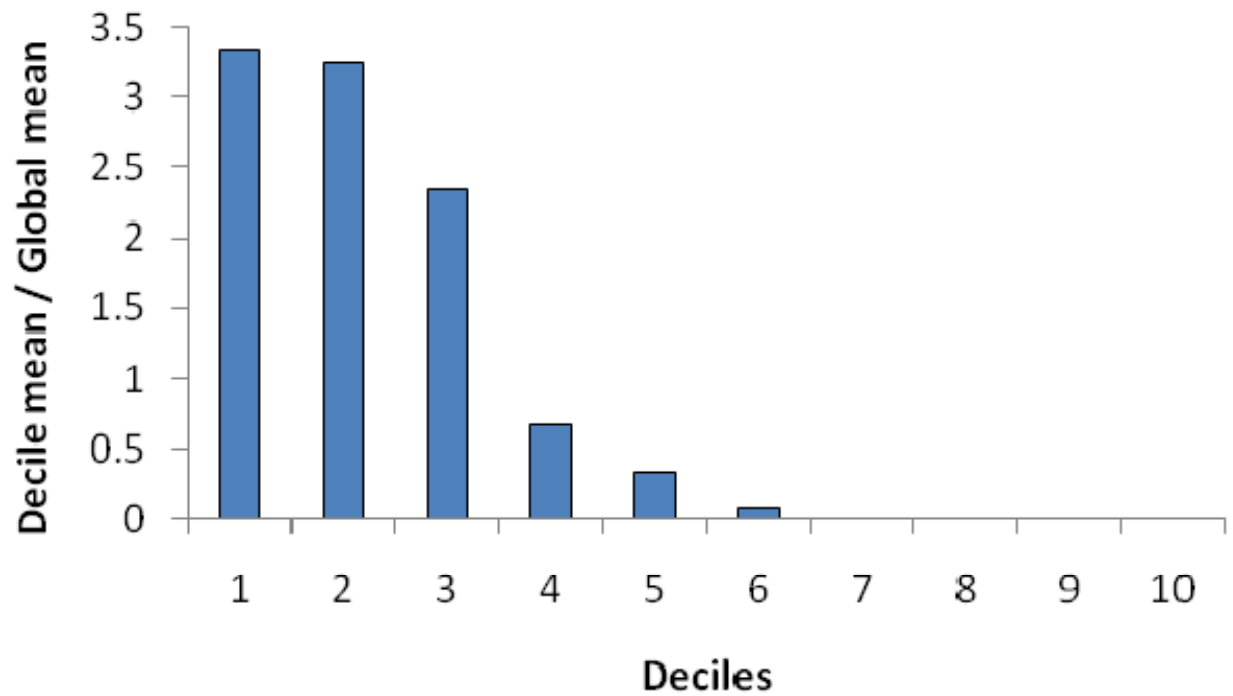




Lift chart (validation dataset)



Decile-wise lift chart (validation dataset)



Comparison of Methods

10-Nearest Neighbors Error Report			
Class	# Cases	# Errors	% Error
1	120	15	12.50
0	280	6	2.14
Overall	400	21	5.25

Classification Tree Error Report			
Class	# Cases	# Errors	% Error
1	120	14	11.67
0	280	9	3.21
Overall	400	23	5.75

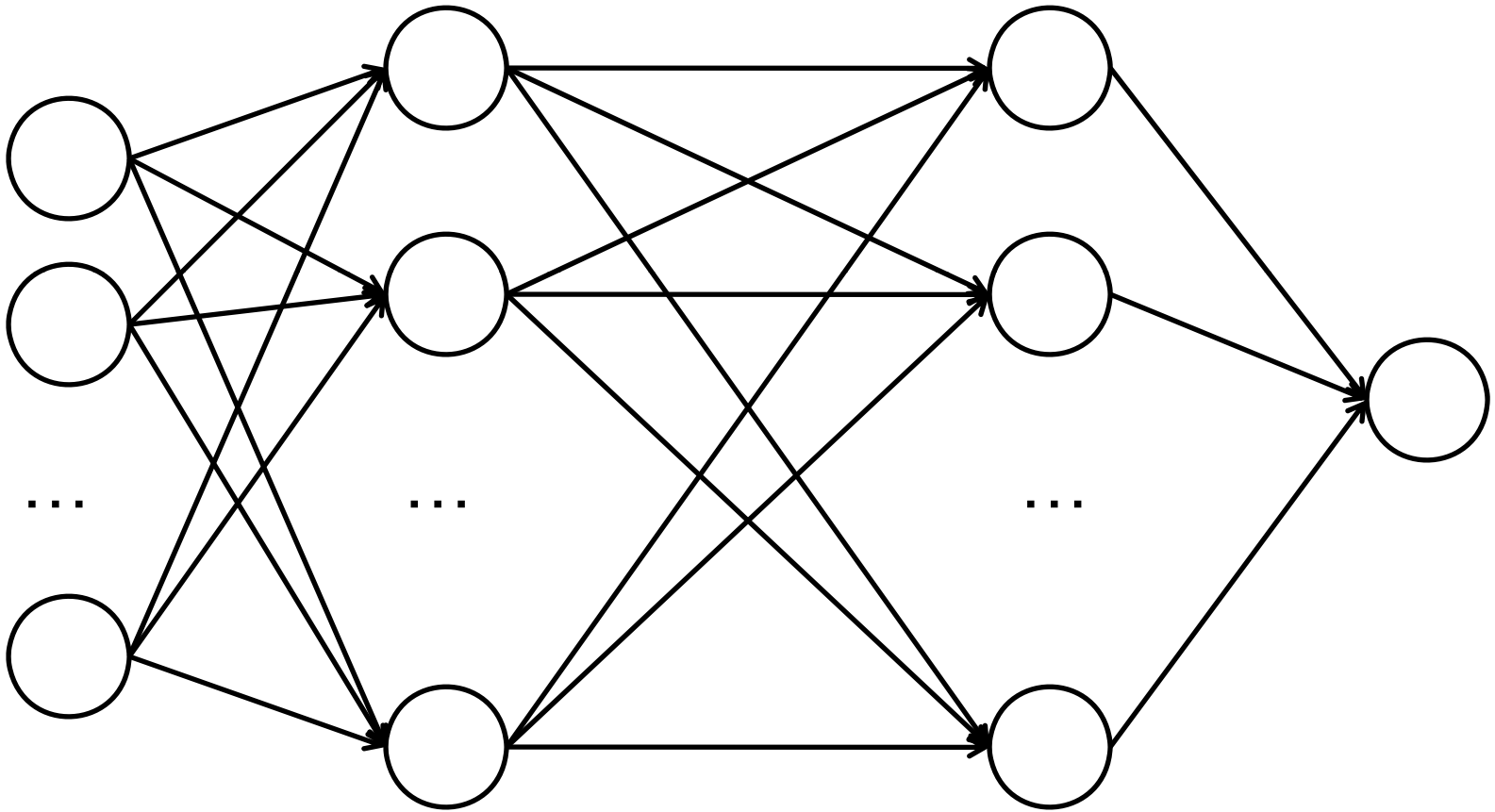
Logistic Regression Error Report			
Class	# Cases	# Errors	% Error
1	120	15	12.50
0	280	9	3.21
Overall	400	24	6.00

Advanced Classifiers

Neural Networks: simulation of an hypothesized model of human learning: interconnected neurons (nodes) that interact to transform inputs (factors) into an output response (prediction or classification).

Discriminant Analysis: uses a linear function of factors (similar to regression analysis) to score observations, separating (discriminating) observations based on the "statistical distance" between an observation and the centroid of a class.

Neural Networks



Input

Hidden

Output

Neural Network Process

1. Randomly initialize a weighted linear response function (β_{ki}) for each hidden and output node k : $S_k = \beta_{k0} + \beta_{k1} * X_1 + \dots + \beta_{kn} * X_n$.
2. Compute the output for an observation: The output from a node is its response function S_k used in a selected transfer function, typically a logistic / sigmoid function: $O_k = 1 / (1 + e^{-S_k})$
3. Adjust the weights for each node to reduce the error in the prediction for that observation.
4. Repeat 2 and 3 for the training data.
5. Repeat 4 until: no improvement or acceptable misclassification rate or maximum iterations.

2-Hidden Node Neural Network Error Report			
Class	# Cases	# Errors	% Error
1	120	20	16.67
0	280	5	1.79
Overall	400	25	6.25

3-Hidden Node Neural Network Error Report			
Class	# Cases	# Errors	% Error
1	120	20	16.67
0	280	5	1.79
Overall	400	25	6.25

25-Hidden Node Neural Network Error Report			
Class	# Cases	# Errors	% Error
1	120	18	15.00
0	280	7	2.50
Overall	400	25	6.25

Decile-Wise Lift Comparison (Testing Data)				
	Class Tree	Logistic Regr'n	Neural Net	Discriminant Analysis
Decile	Mean	Mean	Mean	Mean
1	0.9400	0.8100	0.9200	0.7800
2	0.0228	0.0800	0.0500	0.1300
3	0.0041	0.0400	0.0200	0.0600
4	0.0041	0.0400	0.0100	0.0200
5	0.0041	0.0100	0.0100	0.0100
6	0.0041	0.0200	0.0100	0.0100
7	0.0041	0.0100	0.0000	0.0100
8	0.0041	0.0100	0.0000	0.0000
9	0.0113	0.0000	0.0000	0.0000
10	0.0211	0.0000	0.0000	0.0000

Clustering Techniques

Find groups of observations with a high degree of intra-group similarity and a low degree of inter-group similarity. Understanding similarities and differences among groups enables the development of different strategies relative to those groups.

Such grouping techniques are widely applied in market segmentation, market structure analysis, industry analysis, and portfolio analysis.

What is "Similarity?"

Quantitative (Normalized)

- Euclidean Distance (k-NN)
- Statistical Distance (Discriminant)
- Manhattan Distance (Square Block)
- Maximum Co-ordinate Distance
- Correlation (Inverse of Distance)

Qualitative

- Proportion of Matches (0 or 1)
- Proportion of Positive Matches (1 only)

Mixed Quantitative and Qualitative

- Gower's Similarity Measure

Collaborative Filtering

Movie Rating	Gone With the Wind	Spaceballs	Star Wars	Men in Black
Sue Naumi	10	N/R	10	2
Jane Arnold	1	8	2	9
Joe Beans	8	2	9	2
Isabel Loud	2	9	1	8

Would Spaceballs be a good recommendation for Sue?

Collaborative Filtering

Movie Rating	Gone With the Wind	Star Wars	Spaceballs	Men in Black
Sue Naumi	10	10	N/R	2
Joe Beans	8	9	2	2
Jane Arnold	1	2	8	9
Isabel Loud	2	1	9	8

Permute rows and columns to identify "affinity groups" (similar segments). Spaceballs is not a good recommendation for Sue!

Recommendations

Product Customer				

Amazon uses data mining (clustering) algorithms to define affinity groups based on purchases and stated preferences.

Optimization Techniques

Select the set of values for specified *decision variables* that optimizes (maximizes or minimizes) an *objective function* subject to a set of *constraints*. The objective function and constraints are specified in terms of the decision variables and constants.

- Calculus
- Linear and Nonlinear Programming
- Dynamic Programming
- Simulation and Numerical Methods

Extent Decisions

- **Extent decisions** deal with the allocation of scarce resources that determine the "production" level (output produced by an activity).
- Of necessity the extent of one activity impacts the extent of **other activities** competing for those scarce resources.
- Scarce resources include **money** (capital), **people** (labor), **facilities** (equipment), **raw materials**, etc.

Beer or Ale?

Production Problem: At a price of \$23 per barrel for beer and \$13 per barrel for ale a small brewery can sell all of the beer and ale it can produce. A barrel of beer requires 15 lbs of corn, 4 oz of hops, and 20 lbs of barley malt. A barrel of ale requires 5 lbs of corn, 4 oz of hops, and 35 lbs of barley malt. It currently has 480 lbs of corn, 160 oz of hops, and 1,190 lbs of barley malt in raw material inventory (assume perishable). The corn was purchased for \$100, the barley malt for \$300, and the hops for \$80. Production capacity is 100 barrels. What quantities should be produced?

Beer or Ale?

Analysis

What are the decisions?

What is the objective (costs and benefits)?

What are the constraints?

Beer is "more profitable" than ale (\$23 vs. \$13).

Why would a firm ever produce ale?

Beer or Ale?

Analysis

What are the decisions?

Beer vs. Ale production

What is the objective (costs and benefits)?

Revenue = 23 Beer + 13 Ale

What are the constraints?

Corn: $15 \text{ Beer} + 5 \text{ Ale} \leq 480$

Hops: $4 \text{ Beer} + 4 \text{ Ale} \leq 160$

Barley: $20 \text{ Beer} + 35 \text{ Ale} \leq 1190$

Capacity: $\text{Beer} + \text{Ale} \leq 100$

Beer, Ale ≥ 0

LP_BlandBrewery.xls [Compatibility Mode] - Microsoft Excel

Home Insert Page Layout Formulas **Data** Review View Add-Ins Data Mining

Get External Data Refresh All Connections Sort & Filter Filter Sort Clear Reapply Advanced Data Tools Text to Columns Remove Duplicates Outline Analysis Data Analysis Solver

B9 f_x =B8*B2+C8*C2

	A	B	C	D	E
1		Beer	Ale		
2	Optimal	28	12		
3				Used	Max
4	Corn	420	60	480	480
5	Hops	112	48	160	160
6	Barley	560	420	980	1190
7	Capacity	28	12	40	100
8	Unit Revenue	\$23.00	\$13.00		
9	Total Revenue	\$800.00			

Sheet1 Sheet2 Sheet3

Ready 100%

Set Solver Parameters

Solver Parameters

Set Target Cell:

Equal To: Max Min Value of:

By Changing Cells:

Subject to the Constraints:

-
-
-
-

LP_BlandBrewery.xls [Compatibility Mode] - Microsoft Excel

Home Insert Page Layout Formulas **Data** Review View Add-Ins Data Mining

Get External Data Refresh All Connections Sort & Filter Filter Sort Clear Reapply Advanced Text to Columns Remove Duplicates Data Tools Outline Solver Data Analysis

B9 f_x =B8*B2+C8*C2

	A	B	C	D	E
1		Beer	Ale		
2	Optimal	28	12		
3				Used	Max
4	Corn	420	60	480	480
5	Hops	112	48	160	160
6	Barley	560	420	980	1190
7	Capacity	28	12	40	100
8	Unit Revenue	\$23.00	\$13.00		
9	Total Revenue	\$800.00			

Sheet1 Sheet2 Sheet3

Ready 100%

Conceptual MIS Structure

