Using Predictions to Power the Business

Wayne Eckerson
Director of Research and Services, TDWI
February 18, 2009

TDWI RESEARCH

tdwi
THE DATA WAREHOUSING INSTITUTE

Sponsor

Speakers

Wayne Eckerson
Director, TDWI Research

Caryn A. Bloom
Data Mining Specialist, IBM

Agenda

- Understanding Predictive Analytics
- Trends
- Business Challenges
- Technical Challenges
- Q&A

What is Predictive Analytics??

Data Mining

Machine Learning

Supervised Learning

Knowledge Discovery

Artificial Intelligence

Statistical Modeling

Predictive Analytics

Advanced Analytics

Unsupervised Learning

Classification

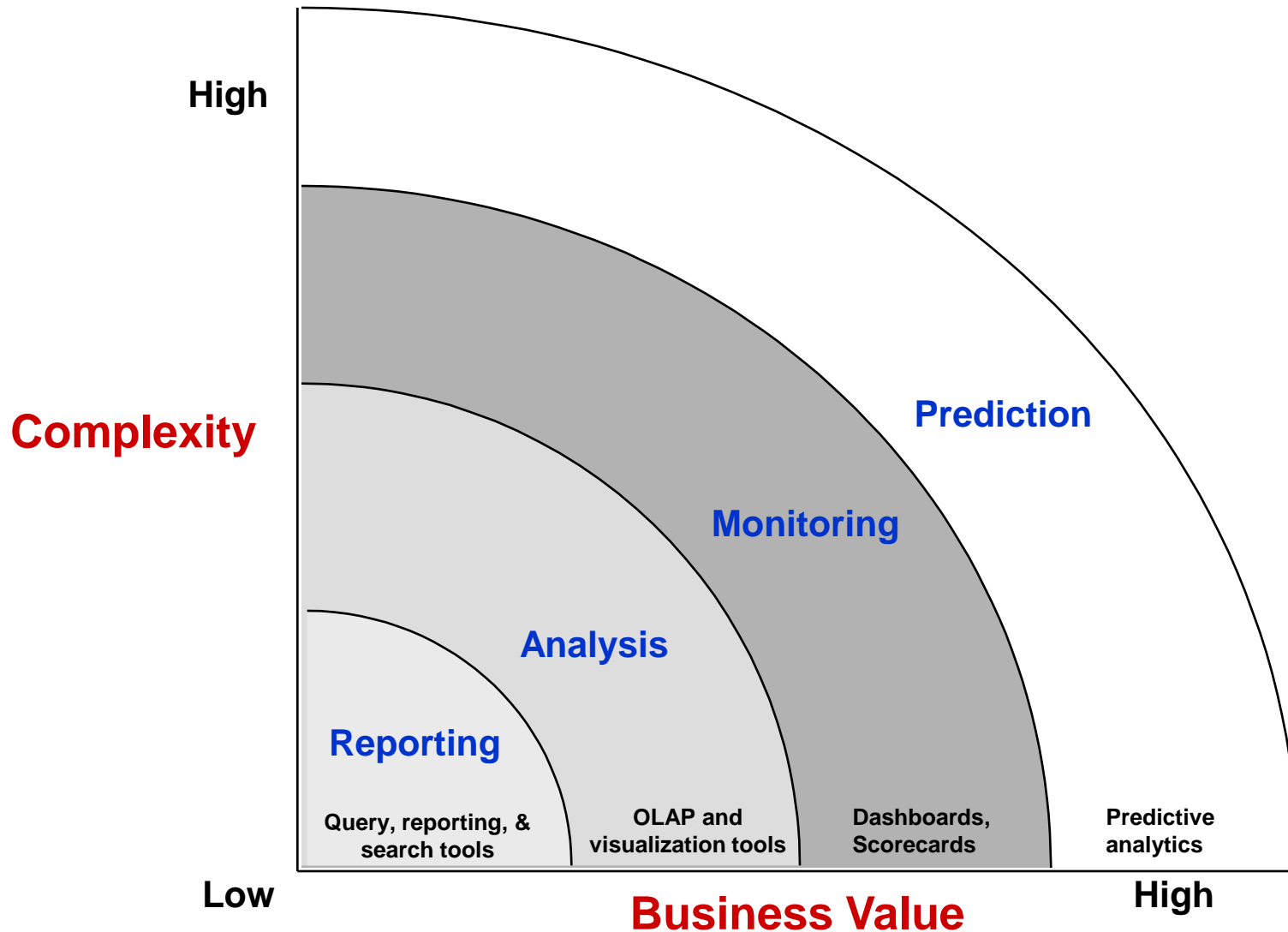Neural networks

Decision Trees

Clustering

Associations

What is Predictive Analytics?

A set of BI technologies that uncovers relationships and patterns within large volumes of data that can be used to predict behavior and events, or to optimize activities.

Other Terms:

- Data mining

- Analytics

- Knowledge discovery

Range of BI technologies

Range of Analytic Users

- ## Statisticians
  - *Creates highly complex analytical models that drive an organization's core business*

- ## Business/data analysts
  - *Creates simple to moderately complex models for departmental usage*

- ## Business users
  - *Run predefined models within applications and view and act on reports that incorporate model results*
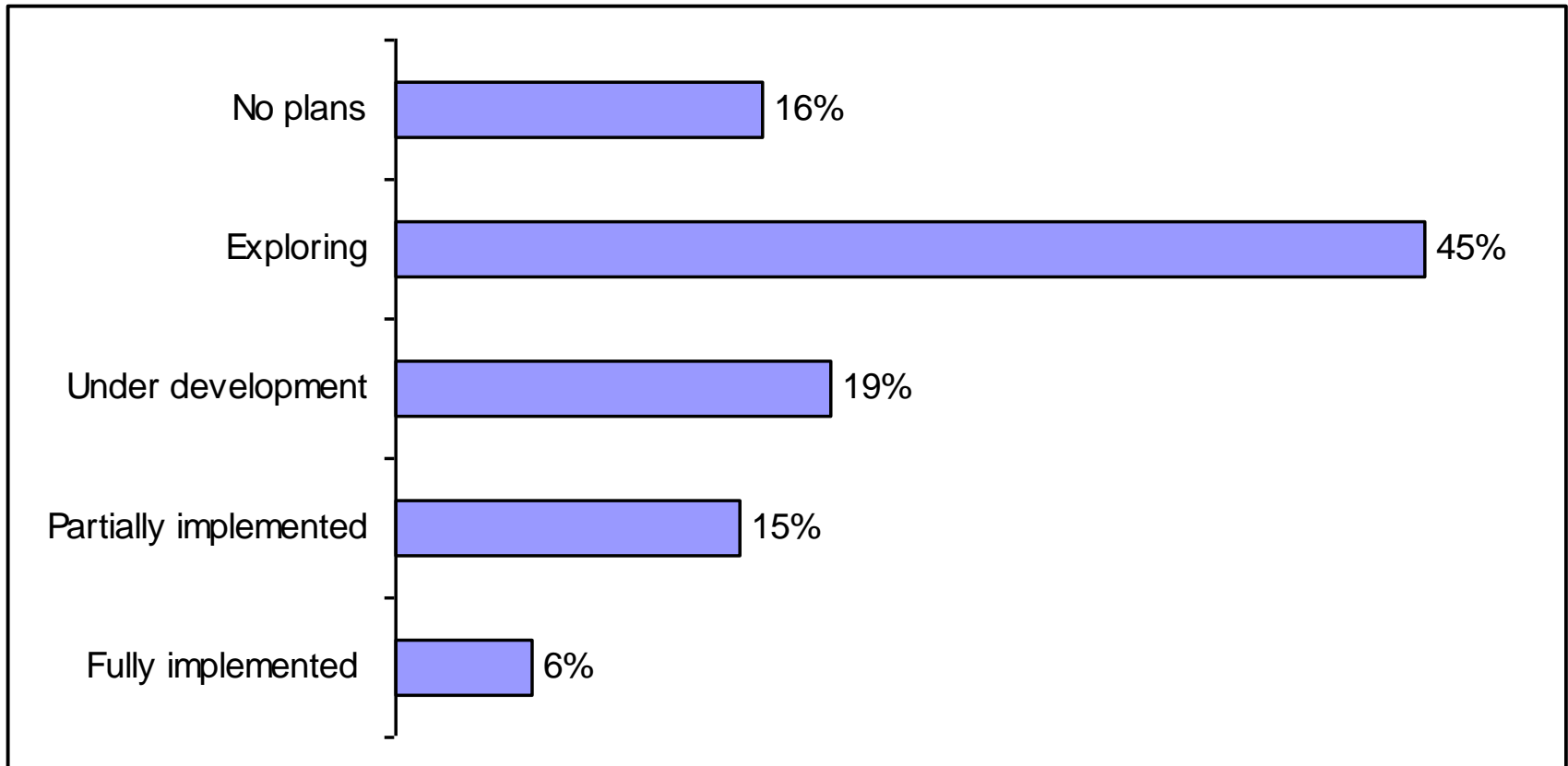
Brian Siegel, VP of Marketing Analytics

Creates predictive models to improve response rates to marketing campaigns

1. Identifies data sources
2. Cleans, standardizes, and formats data
3. Applies predictive algorithms
4. Identifies the most predictive variables
5. Creates and tests the model
6. Delivers target list

"Our response modeling efforts are worth millions. We would not be able to acquire a new client .. Without the lift provide by our predictive models."
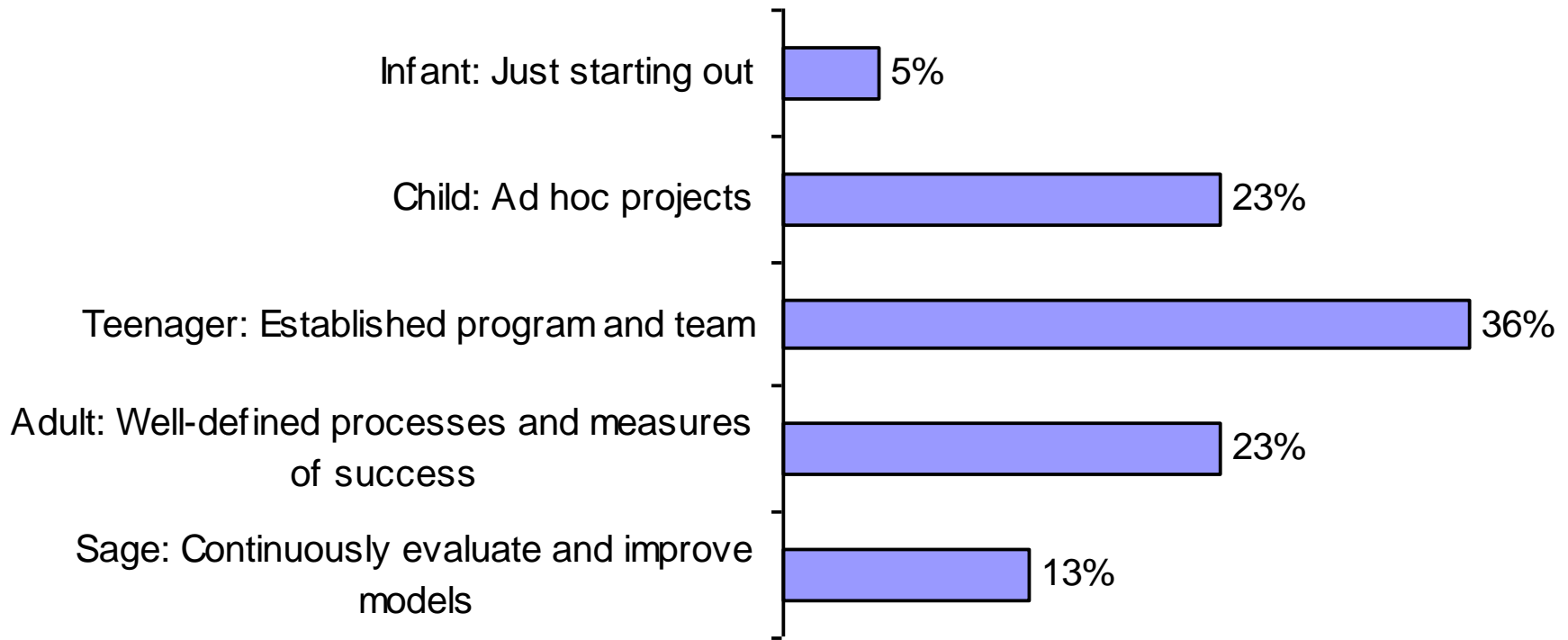*-- Brian Siegel*

Status of Predictive Analytics

No plans — 16%
Exploring — 45%
Under development — 19%
Partially implemented — 15%
Fully implemented — 6%

Based on 833 responses, Wayne Eckerson, "Predictive Analytics: Extending the Value of Your DW Investment," TDWI Research, 2007.

Maturity

Describe the Maturity of Predictive Analytics in Your Group

Category	Percentage
Infant: Just starting out	5%
Child: Ad hoc projects	23%
Teenager: Established program and team	36%
Adult: Well-defined processes and measures of success	23%
Sage: Continuously evaluate and improve models	13%

Based on 166 responses, Wayne Eckerson, "Predictive Analytics: Extending the Value of Your DW Investment," TDWI Research, 2007.

What is the Return?

- Average ROI – TDWI Survey
  - *Average investment: $1.36M*
  - *Average payback: 11.2 mos*
  - *Only 24% conducted ROI study*

- IDC Study in 2002: "Financial Impact of Business Analytics"
  - *Average ROI – 431% with 1.65 year payback*
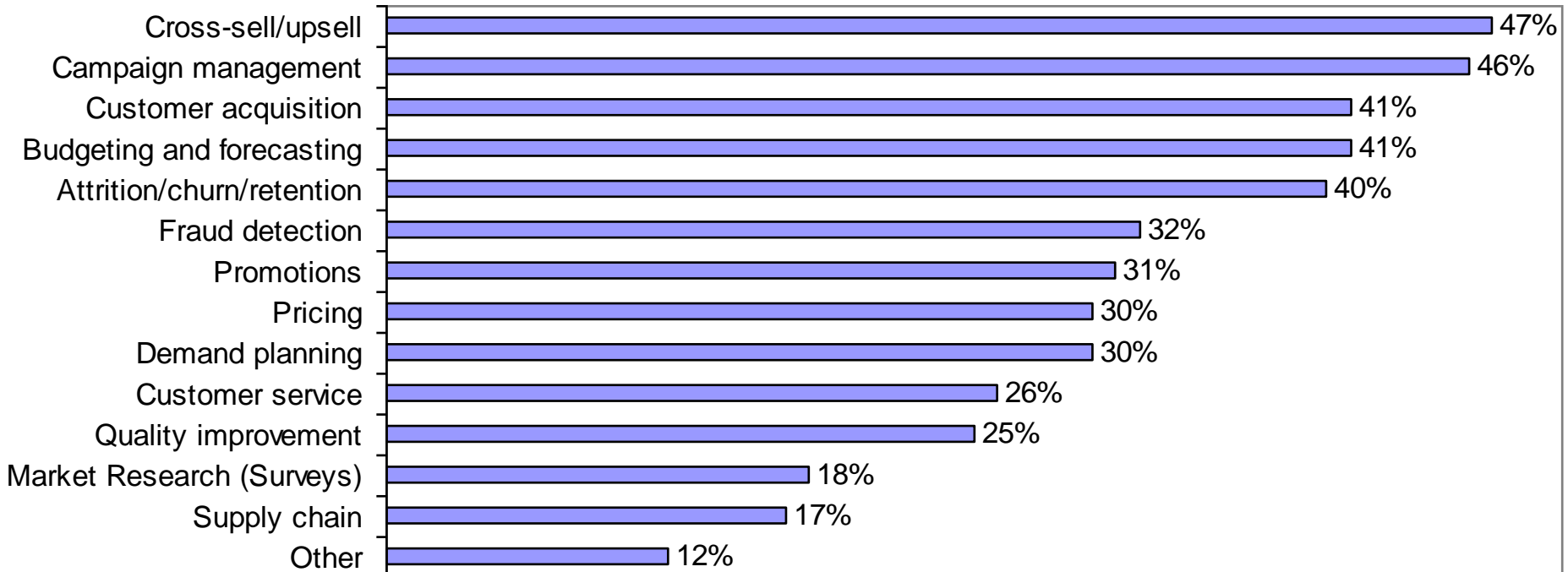 - *Median ROI – 112% with 1 year payback*

Business Challenges

- What applications is it suited for?

- What will it cost?

- How do I set up and organize a predictive analytics practice?

What Applications?

- Target applications have:
  - *Complex processes with multiple variables*
  - *Lots of good historical data*
- Examples:
  - ***Retail****: Design store layouts to optimize profits*
  - ***Trucking****: Schedule deliveries to optimize truck carrying capacity and on-time arrivals*
  - ***Banking****: Set prices to optimize profits without losing customers*
  - ***Insurance****: Identify fraudulent claims*
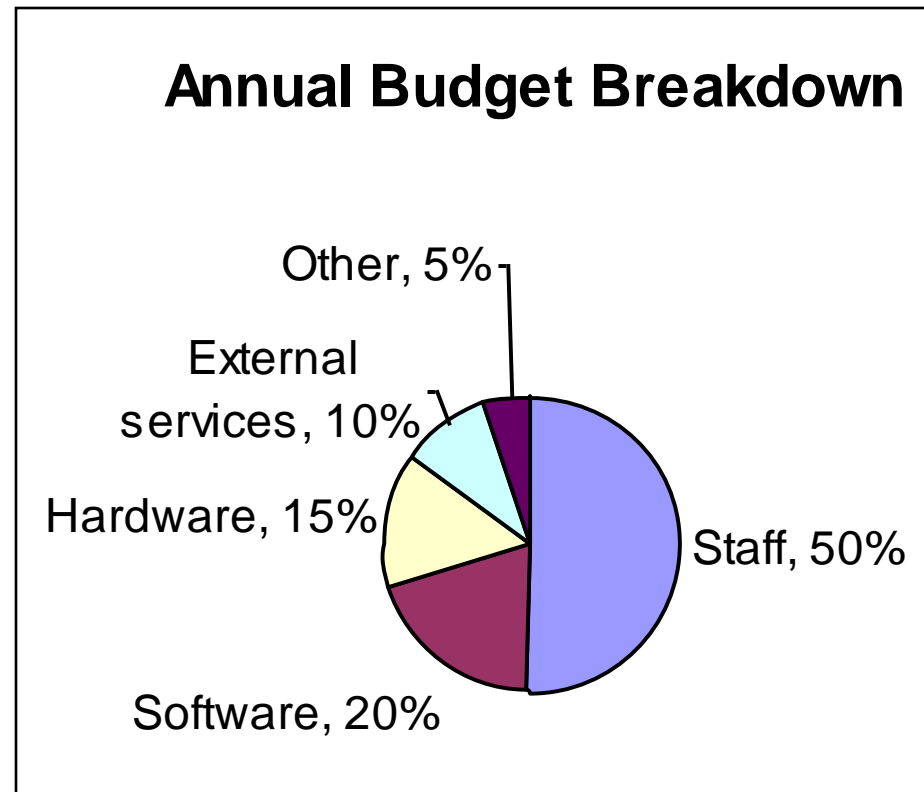 - ***Higher Ed****: Which admitted students will enroll*

What Applications?

Application	Percentage
Cross-sell/upsell	47%
Campaign management	46%
Customer acquisition	41%
Budgeting and forecasting	41%
Attrition/churn/retention	40%
Fraud detection	32%
Promotions	31%
Pricing	30%
Demand planning	30%
Customer service	26%
Quality improvement	25%
Market Research (Surveys)	18%
Supply chain	17%
Other	12%

Based on 166 responses, Wayne Eckerson, "Predictive Analytics: Extending the Value of Your DW Investment," TDWI Research, 2007.

TDWI RESEARCH

15

What Does it Cost?

- Annual budget:
  - *$600,000 median*
  - *$1M for "high value" programs*
- Staff:
  - *Average: 6.5*
  - *Median: 3.5*

Annual Budget Breakdown

- Other, 5%
- External services, 10%
- Hardware, 15%
- Staff, 50%
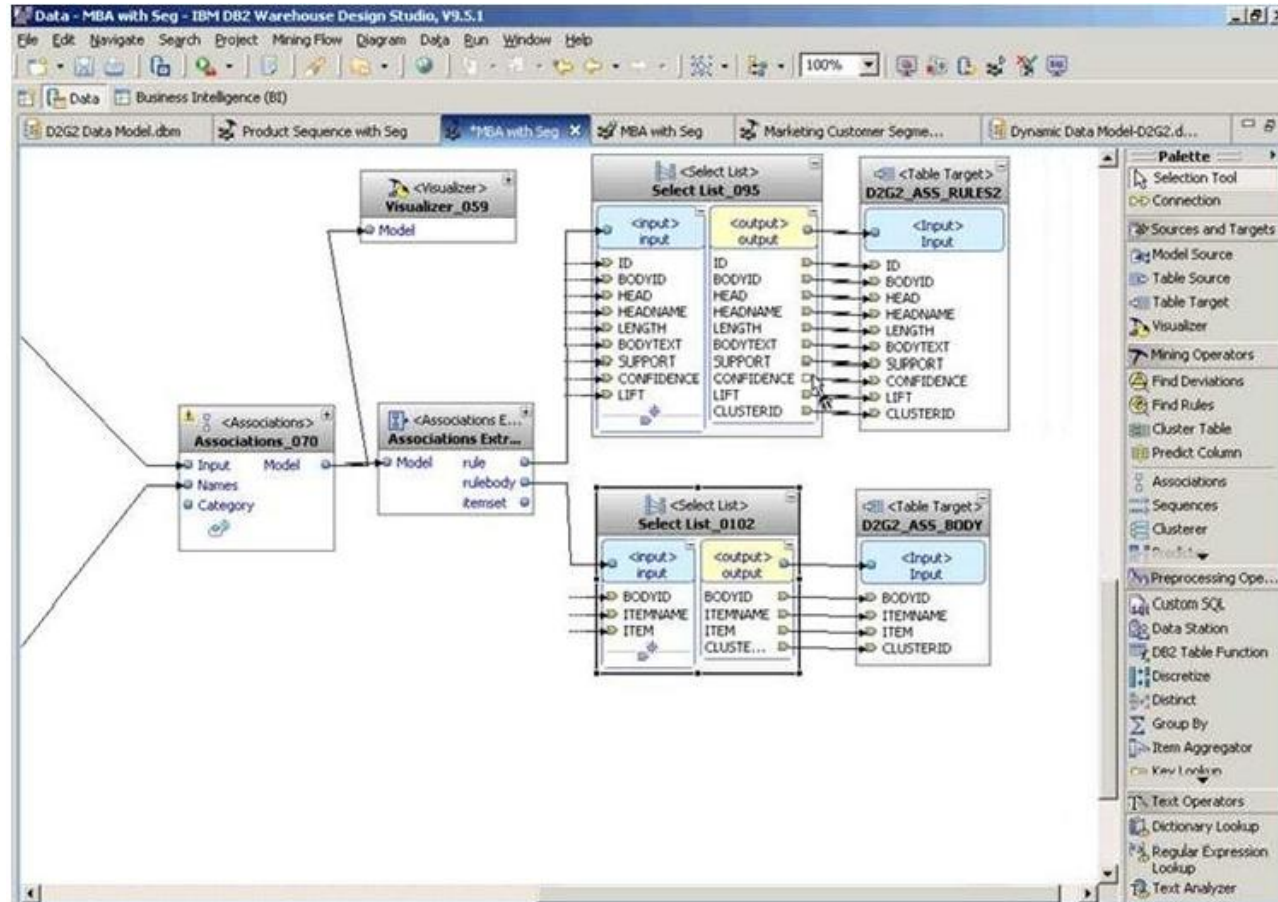- Software, 20%

How to Set Up an Analytics Practice?

1. Hire business-savvy analysts to create models
2. Create a rewarding environment to retain them
3. Fold predictive analytics into a central team

Technical Challenges

- "Analytics Bottleneck"
  - *Complexity*
  - *Data volumes*
  - *Processing expense*
  - *Pricing*
  - *Interoperability*
  - *Dissemination*

Analytical Workbenches

- Graphical
- Integrated
- Automated
- Client/server

# Integration with BI Tools

In-database Analytics

- Leading databases offer specialized functions for creating and scoring analytical models:
    - *Profile data*
    - *Transform, reformat, or derive columns.*
    - *Restructure tables, create data partitions*
    - *Generate samples*
    - *Apply analytical algorithms*
    - *Visualize model results*
    - *Score models*
    - *Store results*

Leverage the Data Warehouse

- Saves time
  - *Sourcing data from multiple locations*
  - *Cleaning, transforming, and formatting data*
- Don't have to move data
  - *To prepare data, create models, score models*
  - *Avoids clogging networks and slowing queries*

Outboard Analytic Data Marts
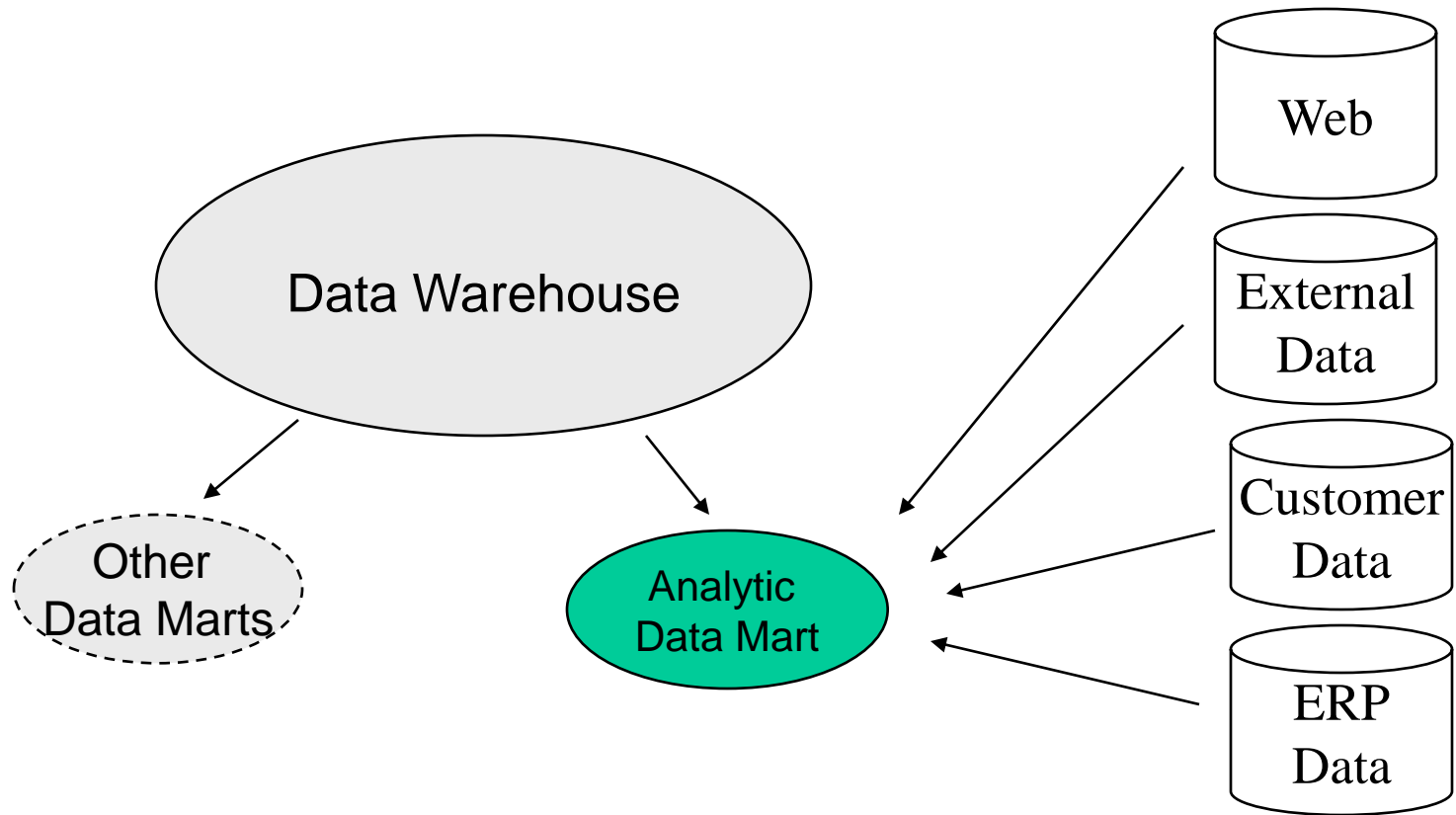
Logical Analytical Data Marts
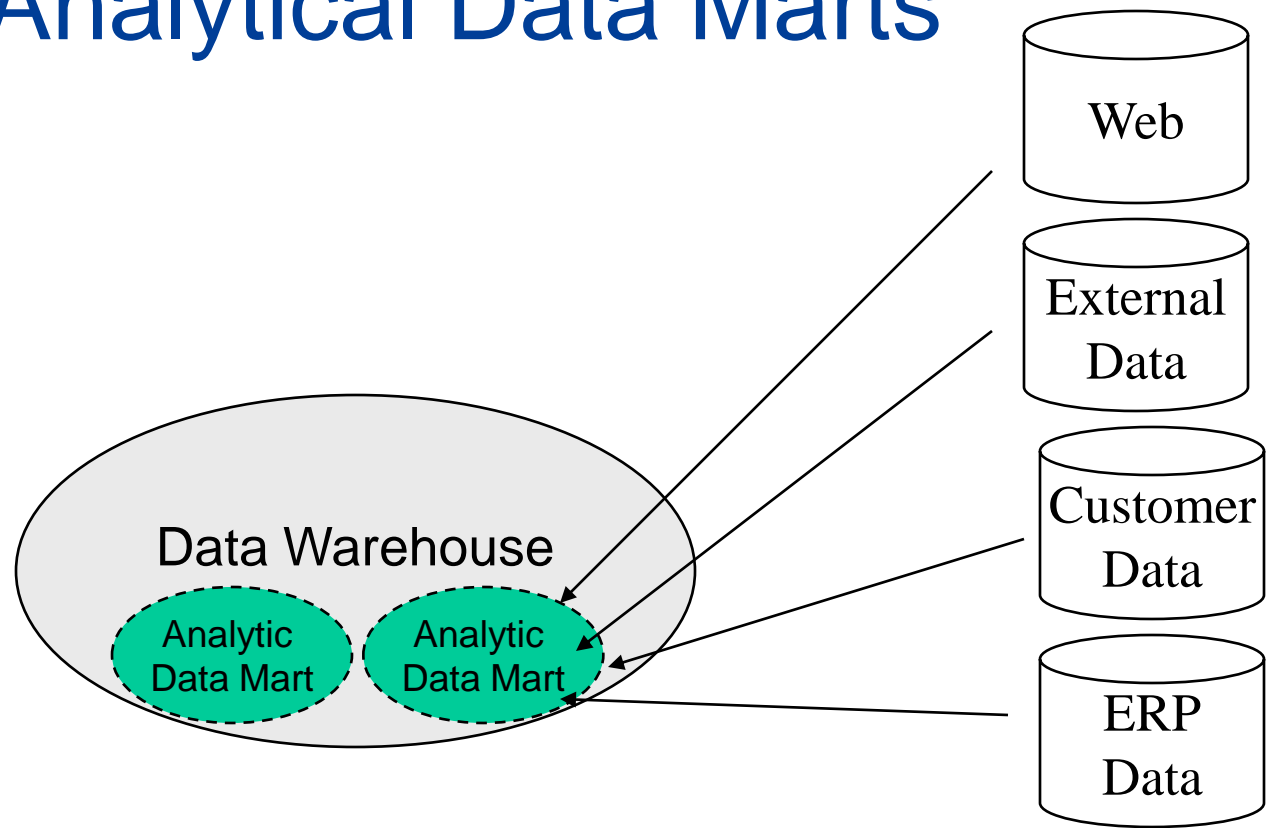
Summary

- Predictive analytics is the next wave in BI
  - *Intimidating jargon, but high business value*
- Business strategy
  - *Find and retain analytical modelers*
- Technical strategy
  - *Reduce analytical bottleneck*

Using Predictions to Power the Business

Analytics Best Practices Team

Information Management software

Data Mining Functional Usage Spectrum

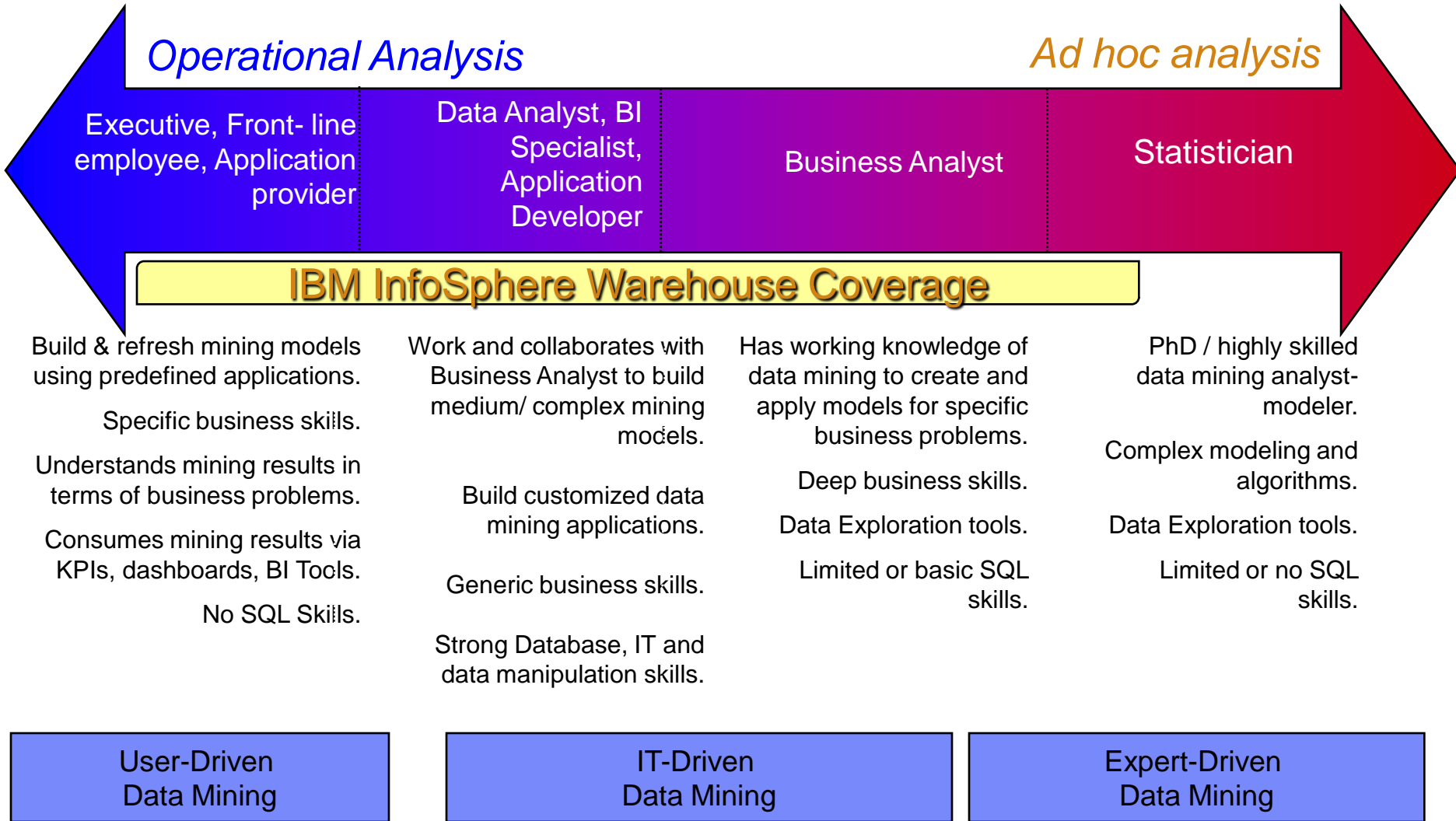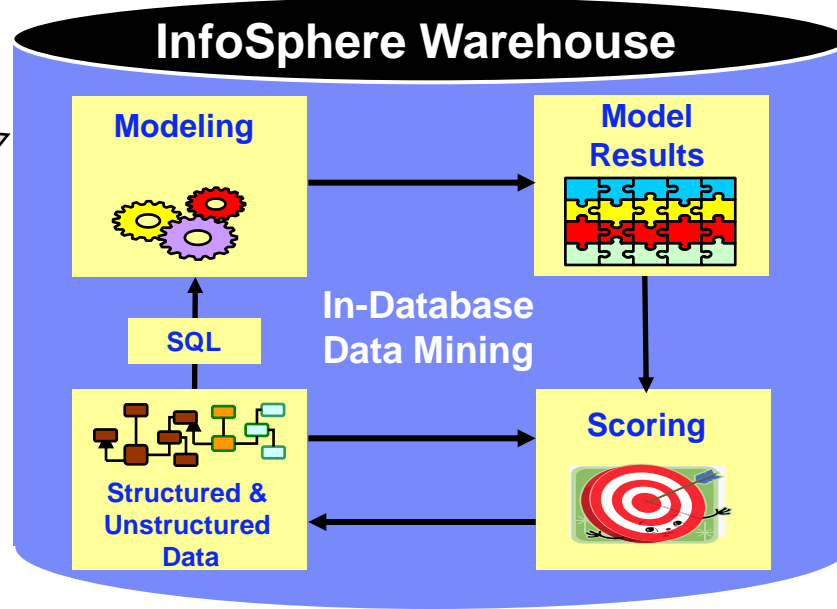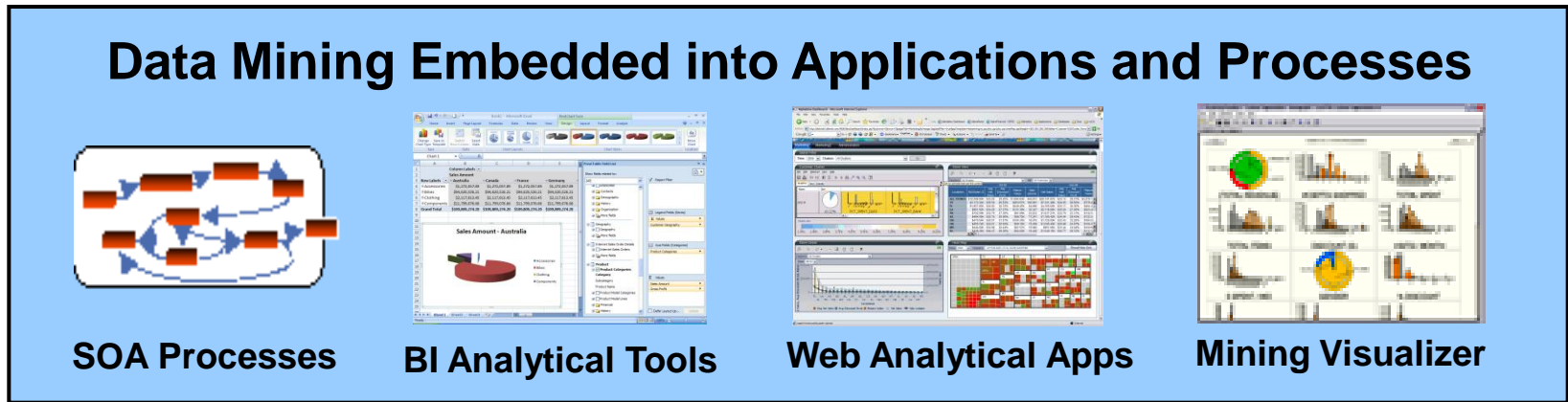*Operational Analysis* → *Ad hoc analysis*

Executive, Front- line employee, Application provider	Data Analyst, BI Specialist, Application Developer	Business Analyst	Statistician

IBM InfoSphere Warehouse Coverage

Build & refresh mining models using predefined applications.	Work and collaborates with Business Analyst to build medium/ complex mining models.	Has working knowledge of data mining to create and apply models for specific business problems.	PhD / highly skilled data mining analyst- modeler.
Specific business skills.		Deep business skills.	Complex modeling and algorithms.
Understands mining results in terms of business problems.	Build customized data mining applications.	Data Exploration tools.	Data Exploration tools.
Consumes mining results via KPIs, dashboards, BI Tools.	Generic business skills.	Limited or basic SQL skills.	Limited or no SQL skills.
No SQL Skills.	Strong Database, IT and data manipulation skills.		

User-Driven Data Mining	IT-Driven Data Mining	Expert-Driven Data Mining

Operational Data Mining with InfoSphere Warehouse

Data Mining Embedded into Applications and Processes

SOA Processes **BI Analytical Tools** **Web Analytical Apps** **Mining Visualizer**

SQL Interface

InfoSphere Warehouse

- *Enterprise-Level Data Mining*
- *High-Speed, In-Database Scoring*

Modeling

Model Results

SQL

In-Database Data Mining

Structured & Unstructured Data

Scoring

Best Practices for Healthcare Analytics Evolution - Payers

Revolutionary Technology

Volume

Throughput Analytics

Automated Systems

Information Correlation

First Generation

Predictive Health Care
- Detect fraud
- Predict treatment costs
- Future charge predictions

Translational Medicine
- Estimate future costs
- Accurate treatment coding
- New provider segments

Payers Today
- Provider
- Claim

Complexity

- Cost group analysis
- Predictive underwriting

Data and Systems Integration

Non-specific   *Organized*   *Personalized*

Evolutionary Practices

Best Practices for Healthcare Analytics Evolution - Providers

Volume

Complexity

Revolutionary Technology

Automated Systems

Information Correlation

First Generation

Throughput Analytics

Predictive Health Care
- Lifetime treatment
- Pre-symptomatic treatment
- Computer aided diagnosis

Translational Medicine
- Molecular medicine
- Genetic predisposition testing
- Clinical Genomics

Healthcare Today
- Digital imaging
- Episodic treatment

Electronic health records

Artificial Expert Systems

Data and Systems Integration

Non-specific   *Organized*   *Personalized*

Evolutionary Practices

Provider Effectiveness: Average Time in ESRD

- **Business problem**

  – Healthcare payer wants to identify "best practices" of physicians who are most successful in treating End Stage Renal Disease (ESRD)

- **Analytical approach**

  – Focus on physicians treating patients who have reached end stage

    • Length of time that a physician's renal disease patients remain in end stage before dying

    • Demographic attributes of their patients (age, gender)

    • Clinical practice attributes (treatment protocols followed)

  – Predict the average number of days that each physician's renal-disease patients remain in end stage

Provider Predictive Analytics

ESRD Member Cluster Summary

Cluster ID	Cluster Size	Top Fields	Field Importance	Field Significance	Significance
1	12.25% of 1,746 Members — 1,532 / 214 — This Cluster / Other Clusters	DAYS_IN_5850	1	Much HIGHER than expected	
		DIABETES_IND	2	POST-CKD DIABETES => is much HIGH	
		TOTAL_DIAGNOSIS_ALL	3	Much HIGHER than expected	
		TOTAL_DAYS_HEMATOP	4	Expected distribution	
		DAYS_TO_STAGE6	5	HIGHER than expected	
		TOTAL_CLAIMS_ALL	6	Much HIGHER than expected	
		PHOSPHATE_IND	7	Y => is much HIGH	
		TOTAL_DAYS_VITAMIN_D	8	Expected distribution	
		DAYS_IN_STAGE6	9	Much HIGHER than expected	
		TOTAL_DAYS_PHOSPHATE	10	HIGHER than expected	
2	7.33% of 1,746 Members — 1,618 / 128 — This Cluster / Other Clusters	TOTAL_DIAGNOSIS_ALL	1	Much HIGHER than expected	
		CONGESTIVE_HEART_FAILURE_IND	2	POST-CKD CONGTV_HEART_FAILURE => is much HIGH	
		DAYS_IN_STAGE4	3	Expected distribution	
		CORNARY_ATHERO_IND	4	POST-CKD CORNARY_ATHERO => is much HIGH	
		TOTAL_DAYS_HEMATOP	5	Expected distribution	
		TOTAL_PROVIDERS_ALL	6	Much HIGHER than expected	
		CITY	7	Dallas => is AVERAGE	
		DAYS_IN_STAGE5	8	HIGHER than expected	
		TOTAL_PROVIDERS_ESRD	9	Much HIGHER than expected	
		DIABETES_IND	10	PRE-CKD DIABETES => is much HIGH	

Sep 3, 2008   1   12:46:47 PM

≍ Top ≙ Page up ⬇ Page down ≍ Bottom
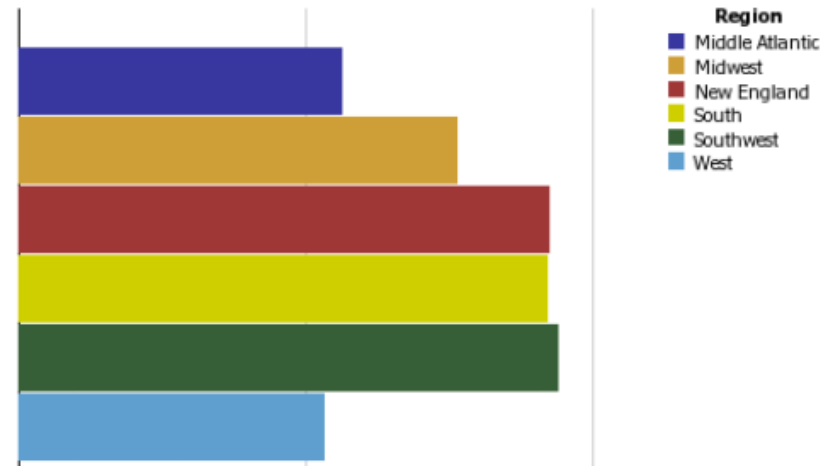
Payer Predictive Analytics

End Stage Renal Disease (ESRD) Analysis

Region	ESRD Member Count	AVG Cost per ESRD Member	Total Cost for ESRD Members	AVG Days in Stage6
Middle Atlantic	296	$45,037.84	$13,331,201.82	562
Midwest	464	$61,021.65	$28,314,046.22	547
New England	96	$73,904.29	$7,094,812.00	548
South	528	$73,599.30	$38,860,432.40	570
Southwest	318	$75,114.20	$23,886,315.48	482
West	44	$42,471.14	$1,868,730.14	409
Summary	**1,746**	**$64,922.99**	**$113,355,538.06**	**520**
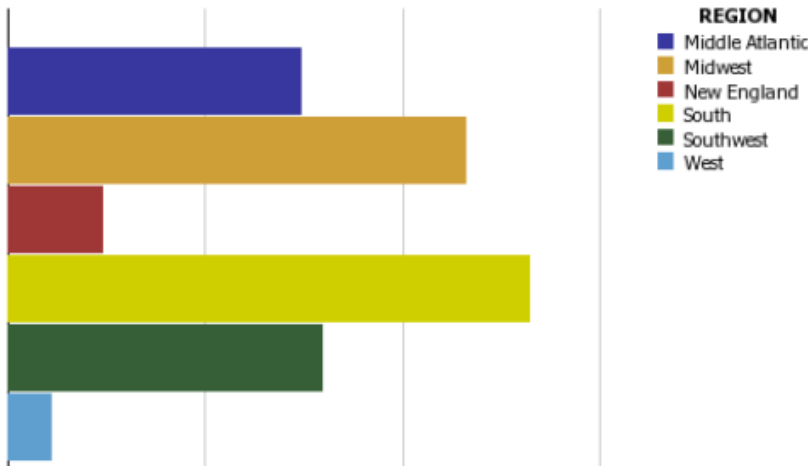
Dictionary Definition in InfoSphere Warehouse

Unfavorable sentiments detect in the call center logs

Importance of unstructured variables for the Decision Tree model
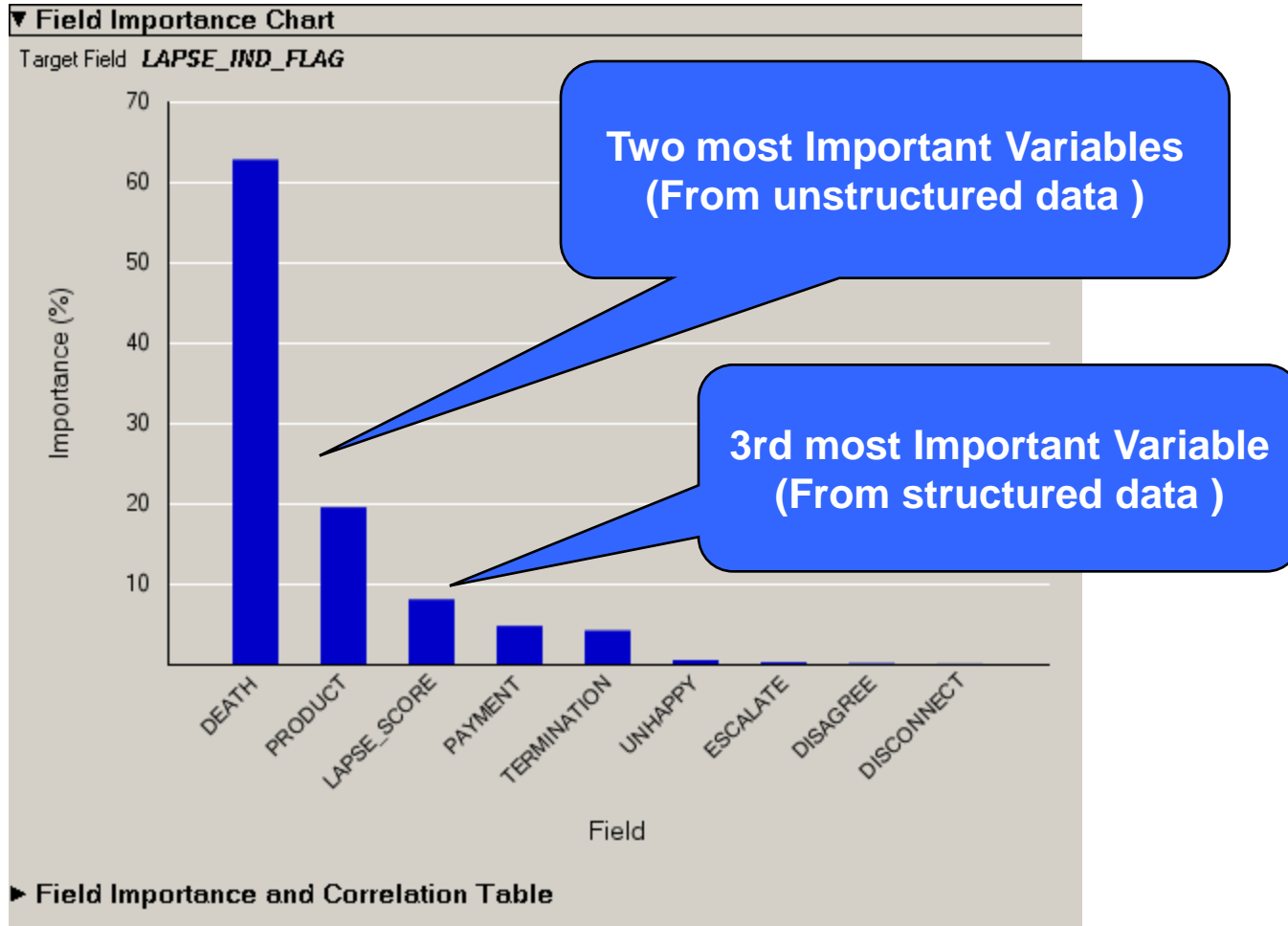
Decision Tree Gains Chart: With vs Without Text

Target Field **LAPSE_IND_FLAG** Target Value Y ▼

Gains

% of "Y" cases (y-axis)

% Records (x-axis)

73% of cases

60% of cases

50% of the records
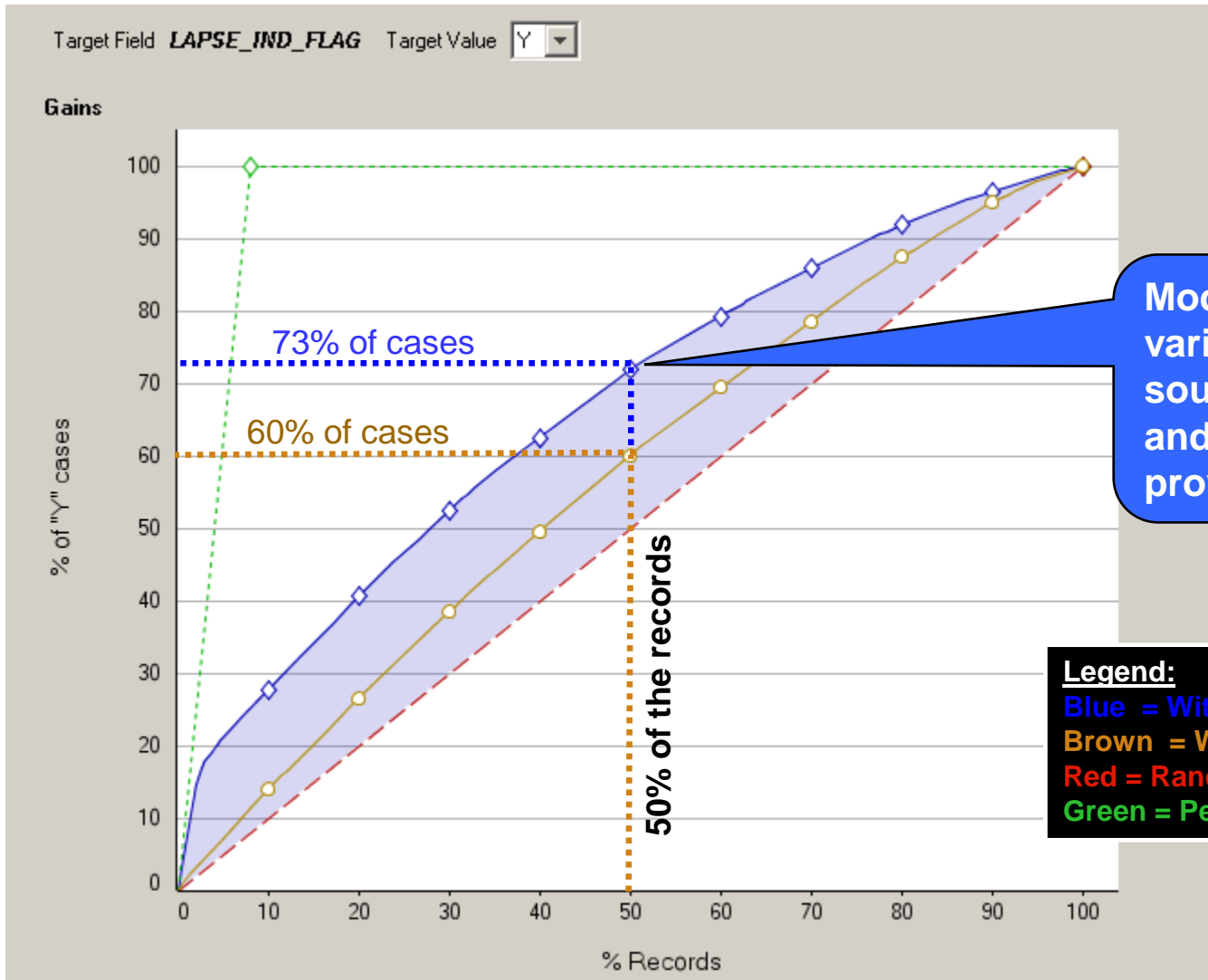
Model using variables from both sources (structured and unstructured) provides better ROI

Legend:
Blue = With Text Analytics
Brown = Without Text Analytics
Red = Random
Green = Perfect Model

Questions?

Contact Information

- If you have further questions or comments:

  Wayne Eckerson, TDWI
    weckerson@tdwi.org

  Caryn Bloom, IBM
    cbloom@us.ibm.com

Please visit the following resources after the webinar for additional information on this topic:

InfoSphere Warehouse Product Web Site:
www.ibm.com/software/data/infosphere/warehouse/

InfoSphere Warehouse Data Sheet:
http://download.boulder.ibm.com/ibmdl/pub/software/data/sw-library/infosphere/datasheets/IMD10900-USEN-01.pdf

Embedded Analytics Solution Brochure:
ftp://ftp.software.ibm.com/software/data/db2/warehouse/IMF14002-USEN-01.pdf

Redbook - Dynamic Warehousing Data Mining Made Easy:
www.redbooks.ibm.com/abstracts/sg247418.html

Technical Whitepaper - Data Mining for Everyone:
http://www-01.ibm.com/software/sw-library/en_US/detail/Y815951M69194W67.html