TOWERS PERRIN
TILLINGHAST

Predictive Modeling Lifecycle

A Practical Approach: What's Important & What's Hard

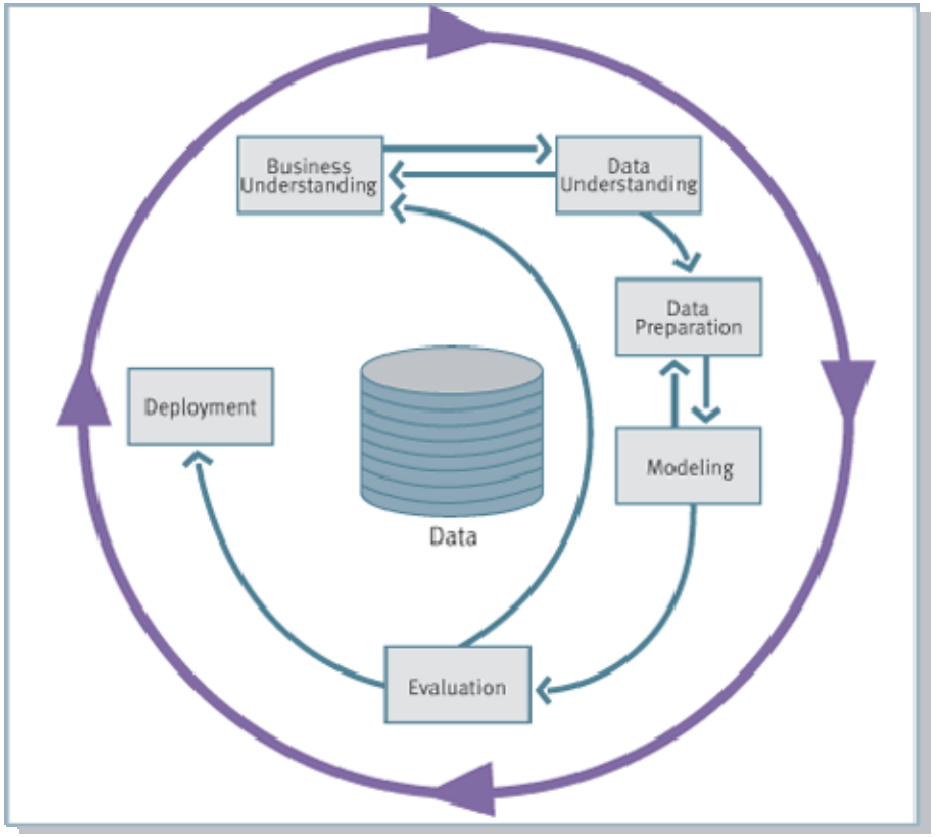
Data Understanding, Data Preparation, and Modeling

CAS Predictive Modeling Seminar
Las Vegas, NV
October 11-12, 2007

Martha A. Winslow, FCAS, MAAA

© 2007 Towers Perrin

Phases of the predictive modeling lifecycle



From CRISP_DM Process Model 1.0, 2000

© 2007 Towers Perrin

A modeler's view of project lifecycle

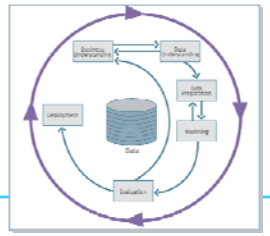
Modelers focus on the data and the modeling. They are notoriously poor project planners (a broad generalization)…and management is too optimistic.

Planned Gantt Chart

Actual Gantt Chart

© 2007 Towers Perrin

Data Understanding Phase



Tasks

- Produce a preliminary list of predictive data elements (internal and external)
- Review sources of internal data and identify data elements that should be captured in the future
- Identify potential external data sources and cost and identify data elements that should be acquire from outside sources
- Collect initial data
  - Acquire data dictionaries
- Explore data searching for trends and anomalies to gain understanding and ideas for the modeling phase
- Identify regulatory requirements/constraints in jurisdictions where the company operates
- Verify data quality
  - Does the data meet the business objectives?

What's Important

- Tying data element selection back to the business objectives
- Careful data element identification and exploration lays the groundwork for a successful model

What's Hard

- Valuable external data may be costly to acquire
- Regulators may disallow potentially highly predictive variables, e.g. credit score
- Identifying solutions to data quality problems

© 2007 Towers Perrin

Start identifying possible independent variables by brainstorming

- For example, what information might bear any statistical relationship to the likelihood, nature, and severity of a claim?
- At this stage, we should not judge any idea to be bad, unacceptable, or impractical

Brainstorming Flip Chart

Data Element	Why it might be good

© 2007 Towers Perrin

Perform a preliminary evaluation and initial culling of potential variables identified during the brainstorming step

Data Element	Potential Value		Acceptability		Ease of Gathering				Continue Investigation ?	Responsibility	
					Source	Electronic/ Manual	Existing/ New	Overall Avail-- ability			
	Hi/Lo	Comments	Hi/Lo	Comments					Yes/No	Who	When

1 = Good
5 = Bad

© 2007 Towers Perrin

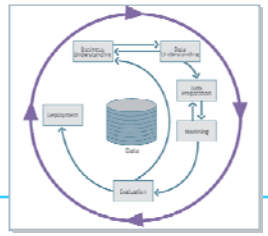
Complete a follow-up assessment for data elements that appear to have potential for the model

Ease of Gathering	**Data Element:** **Source:** **New or existing** **Currently obtained by your organization** **Electronic or paper** **Cost** **Timing** **United Fire systems implications**	
Acceptability	**Board** **Policyholders** **Future customers** **Regulators** **Overall evaluation**	

1 = Good
5 = Bad

© 2007 Towers Perrin

Data Preparation Phase



Tasks

- Select data for modeling and univariate analyses
- Clean data
- Derive new variables
- Merge/join tables and construct the modeling data
- Aggregate records to level to be used in modeling

What's Important

- Validate data elements and structure
- Matching data from various systems (e.g., policy issuance, claims, billing, etc.)

What's Hard

- Matching data from various systems (e.g., policy issuance, claims, billing, etc.)
- Merging data from external sources
- May discover unexpected data issues due to initial use of data elements

© 2007 Towers Perrin

Inevitably, there are numerous data issues to address/resolve

Typical Data Issues

- Data is usually in the wrong format for modeling
- Poor quality can cause model convergence problems and must be dealt with
- Many derived variables need to be added
- Missing characteristics for prior policy periods (e.g., insurance scores not ordered for all of historical data)

- The modeling effort should leverage a broad array of information sources/types

Drivers of Value — Automobile Insurance Customer

	Retention	Loss Experience
Credit history	✓	✓
Billing/pay plan information	✓	✓
Prior non-chargeable and comprehensive claims		✓
Cross-line policies and claims	✓	✓
Time on job and time at present address	✓	✓

The good news…you probably have, but may not use, much of the data you need

© 2007 Towers Perrin

ILLUSTRATIVE

Frequently, the company's basic data structure has to be reformatted

Likely Current Data Structure — Coverages in Rows

Policy Number	Policy Year	Coverage	Period Start Date	Risk Coding Variables	Earned Exp.	Claim Count	Incurred Loss
1	2003	BI	01/01/03	Age, sex, marital status, etc.	1.0	0	0
1	2003	PD	01/01/03		1.0	1	2500
1	2003	MED	01/01/03		0.5	0	0
1	2003	MED	01/01/03		0.5	1	250
Many more records…							

Advantages:

- Data are probably already stored this way
- Multiple records from mid-term changes only present for affected coverages
- For studying one coverage/peril at a time, file size can be smaller than alternative

Required Data Structure — Coverages in Columns

Policy Number	Policy Year	Period Start Date	Risk Coding Variables	Bodily Injury			Property Damage, etc.		
				Earned Exp.	Claim Count	Incurred Loss	Earned Exp.	Claim Count	Incurred Loss
1	2003	01/01/03	Age, sex, marital status, etc.	1.0	0	0	1.0	0	
2	2003	01/01/03		1.0	1	20,000	1.0	1	5,000
3	2003	01/01/03		0.5	0	0	0.5	0	
3	2003	01/01/03		0.5	1	1,250	0.5	1	500
Many more records…									

Advantages:

- Can combine "scored" results across coverages/perils
- Total file storage requirement could be smaller (risk variable coding not repeated)

Disadvantages:

- "Pivoting" the data is not always a trivial step
- Varying number of transactions and dates by coverage can complicate things

© 2007 Towers Perrin

ILLUSTRATIVE

More granular detail can highlight other data problems

Policy change endorsement records cause problems if done improperly

Policy Number	Policy Year	Policy Start Date	Transaction Date	Age	Limit	Premium
1	2003	01/01/03	01/01/03	39	100000	500
1	2003	01/01/03	05/01/03	40	100000	-250
1	2003	01/01/03	05/01/03	40	250000	300
Many more records…						

Example:

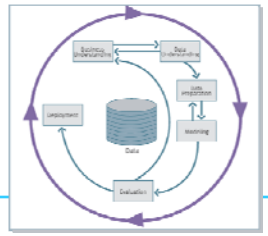
- By inappropriately incrementing driver age at endorsement time, records with negative values do not get aggregated
- Resulting negative values render that record unusable, and it is discarded
- Total premium and exposure for this policy are then overstated

Other Typical Challenges

- Earnings are inaccurate
- Policy-level calculated values are wrong (e.g., number of vehicles or drivers on the policy are inaccurate)
- Cancel-rewrites or policy transaction system changes
  - Policy tenure can be lost
  - Link to historical policy information and claim activity can be lost
- Cross-line information
  - Missing or inaccurate match-key fields
  - Non-aligned effective dates
- Claim data
  - Inadequate match-key data
  - Claim counts – one per event vs. one per claimant

© 2007 Towers Perrin

Modeling Phase



Tasks

- Perform initial univariate analysis
  - Evaluate results in light of business objectives to select/prioritize variables for multivariate analysis
- Conduct initial multivariate analysis
- Reduce data dimensions, eliminate redundant variables and group numeric variables (e.g. driver age)
- Build a series of models that will meet regulatory requirements in all jurisdictions
  - Note: Need to identify/confirm state regulatory variations
- Select desired variables for inclusion in rating formula in light of business objectives and with view of ease of implementation
- Finalize multivariate models

What's Important

- Consideration of regulatory acceptance of desired variables
- Consideration of agency acceptance of desired variables
- Design model output for users; e.g. reason codes

What's Hard

- Getting the actuarial relationships right
- Balancing level of complexity (number of tiers and introduction of new variables), which improves precision, with implementation realities
- Knowing when to stop, i.e. how many models to try

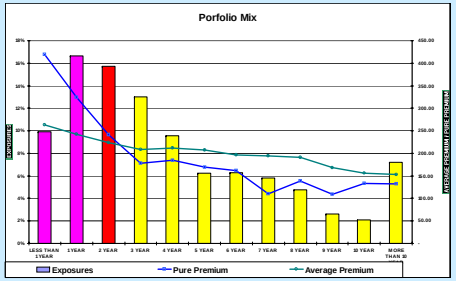
© 2007 Towers Perrin

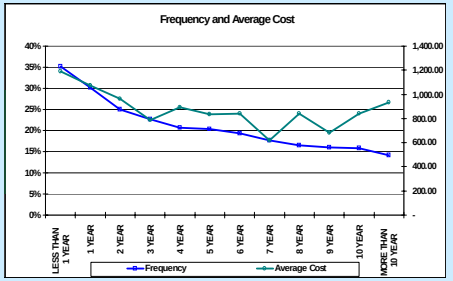
ILLUSTRATIVE

Univariate and multivariate analyses are the foundation for defining the rating variables

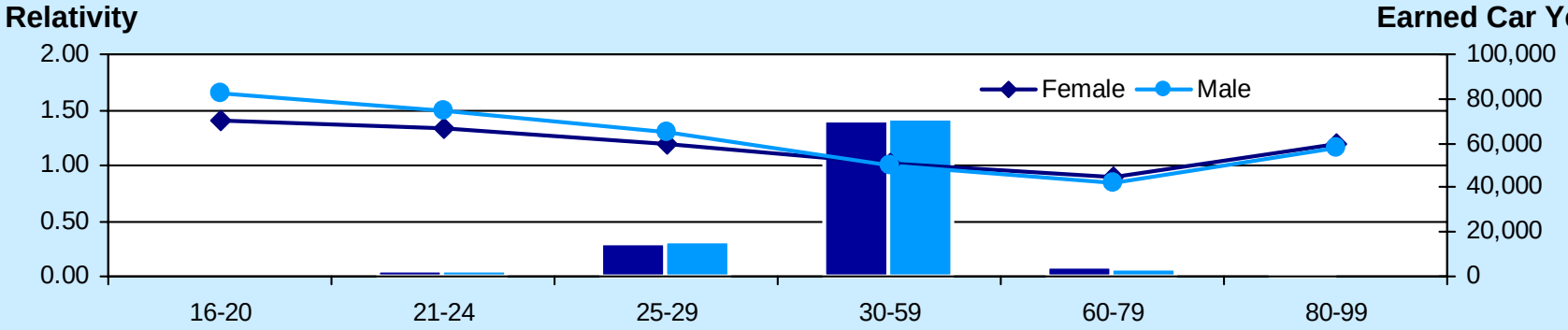
Univariate Analysis

Years in Force (1)	Exposures / Year (2)	Distribution Exposures (3)	Earned Premium (4)	Number Claims (5)	Incurred < 60000 (6)	Incurred > 60000 (7)	Total Incurred (8)	Frequency (9)	Average Cost (10)	Average Earned Premium (11)	Pure Premium Total (12)	Pure Premium Truncated (13)	Loss Ratio Total (14)	Loss Ratio Truncated (15)
LESS THAN 1 YEAR	63,171	9.9%	16,604,397	22,264	20,479,880	7,999,906	28,479,785	35.2%	1,279	263	451	324	171.5%	123.3%
1 YEAR	106,117	16.7%	25,772,766	31,978	27,388,881	9,308,005	36,696,887	30.1%	1,148	243	346	258	142.4%	106.3%
2 YEAR	100,285	15.7%	22,433,493	25,131	19,718,149	5,993,240	25,711,389	25.1%	1,023	224	256	197	114.6%	87.9%
3 YEAR	82,841	13.0%	17,254,497	18,725	12,823,144	2,539,342	15,362,485	22.6%	820	208	185	155	89.0%	74.3%
4 YEAR	61,054	9.6%	12,957,349	12,662	9,173,545	2,804,591	11,978,136	20.7%	946	212	196	150	92.4%	70.8%
5 YEAR	39,844	6.3%	8,265,338	8,091	5,606,850	1,518,850	7,125,700	20.3%	881	207	179	141	86.2%	67.8%
6 YEAR	40,084	6.3%	7,883,161	7,721	5,041,749	1,917,322	6,959,072	19.3%	901	197	174	126	88.3%	64.0%
7 YEAR	37,137	5.8%	7,219,076	6,588	3,972,052	141,202	4,113,254	17.7%	624	194	111	107	57.0%	55.0%
8 YEAR	30,303	4.8%	5,795,569	5,024	3,614,700	791,910	4,406,610	16.6%	877	191	145	119	76.0%	62.4%
9 YEAR	16,622	2.6%	2,804,579	2,662	1,575,950	317,357	1,893,308	16.0%	711	169	114	95	67.5%	56.2%
10 YEAR	13,372	2.1%	2,090,261	2,114	1,482,381	391,024	1,873,404	15.8%	886	156	140	111	89.6%	70.9%
MORE THAN 10 YEAR	46,011	7.2%	7,021,732	6,518	4,387,749	2,253,534	6,641,283	14.2%	1,019	153	144	95	94.6%	62.5%
Total	636,842	100.0%	136,102,218	149,478	115,265,028	35,976,285	151,241,313	23.5%	1,012	214	237	181	111.1%	84.7%





Multivariate Analysis

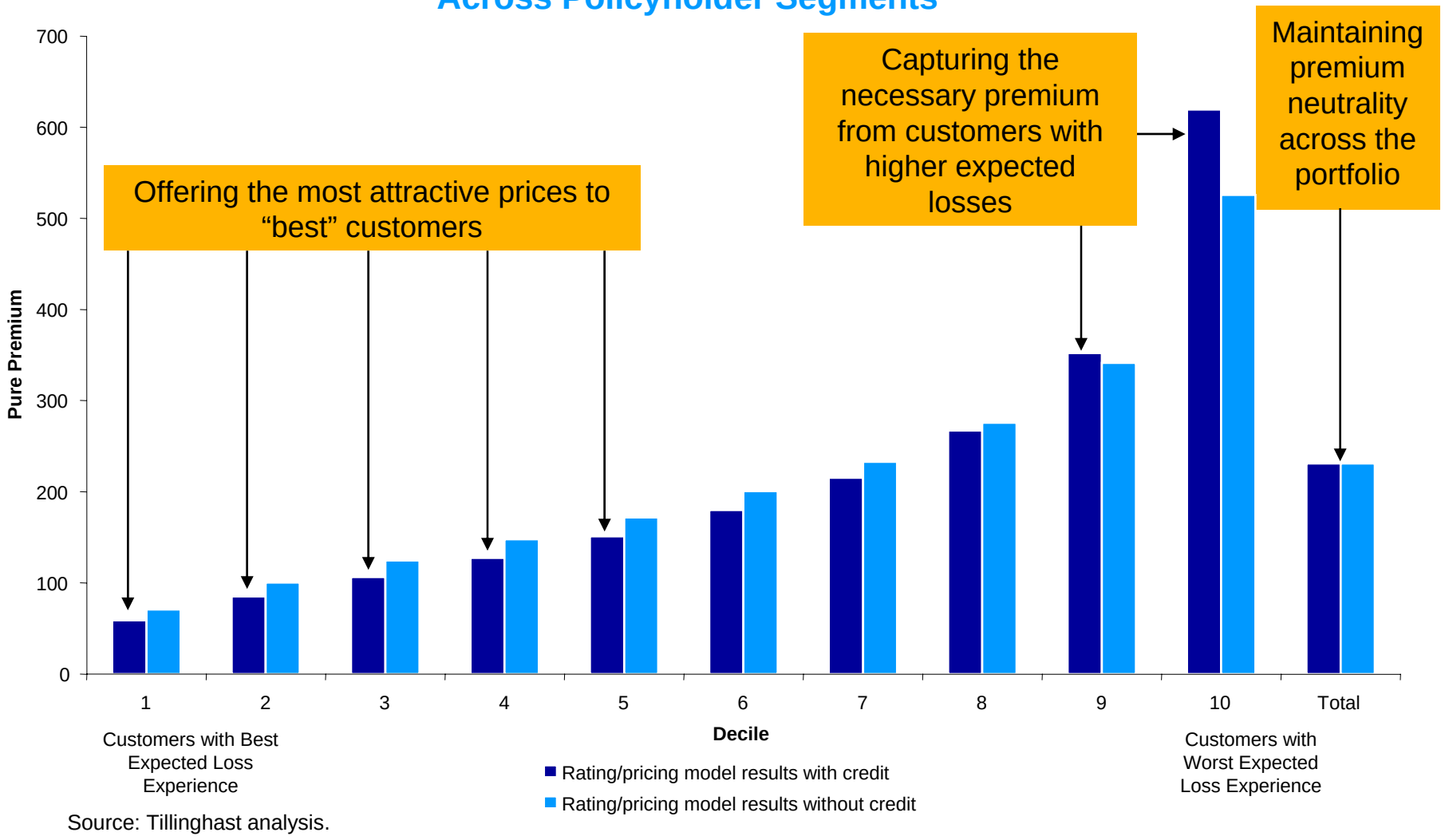


© 2007 Towers Perrin

Decisions about what variables will survive in the model must balance contribution to model "lift" and acceptability to stakeholders

Premium Impact of Using Credit Scores Across Policyholder Segments

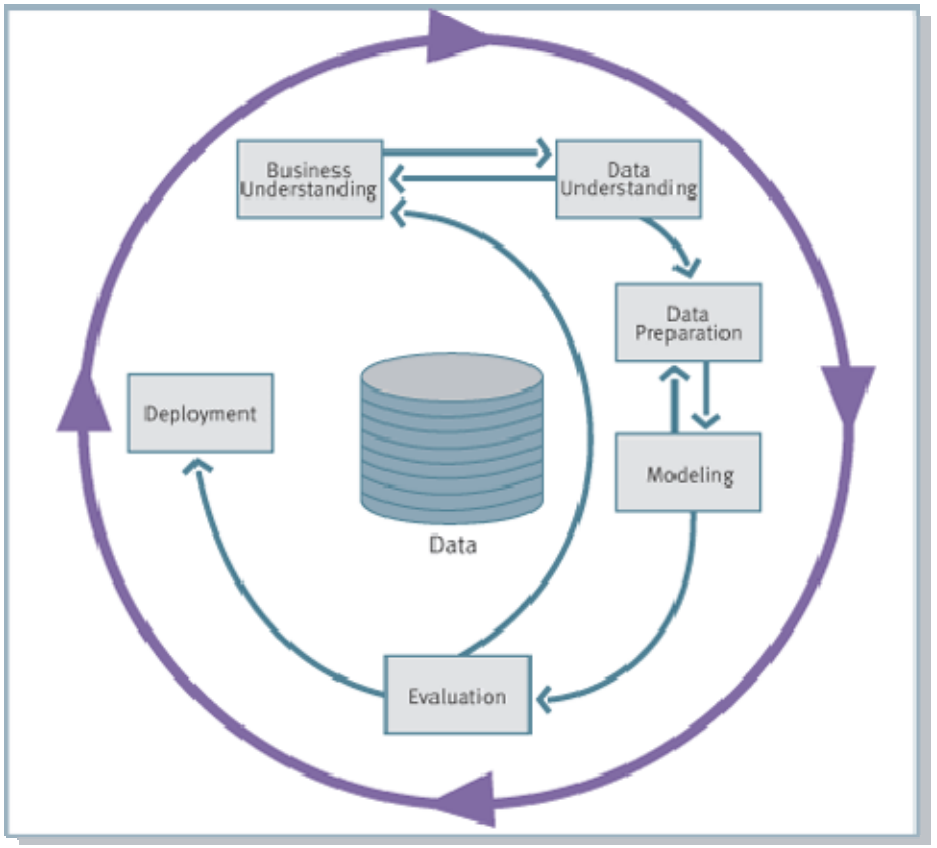
ILLUSTRATIVE



Source: Tillinghast analysis.

© 2007 Towers Perrin

Phases of the predictive modeling lifecycle



From CRISP_DM Process Model 1.0, 2000

© 2007 Towers Perrin

Speaking today

	Contact Information
 Martha Winslow	Senior Consultant Towers Perrin 7650 Edinborough Way Suite 500 Minneapolis, MN 55435-5978 Phone: +1 952 842 5627 Fax: +1 952 842 5666 E-mail: martha.winslow@towersperrin.com

© 2007 Towers Perrin