



# 2007 Caribbean Actuarial Association Conference

---

## **Predictive Modeling – An Introduction**

Russell H. Greig, CFA, FCAS, MAAA

December 6, 2007

# Background of Predictive Modeling via Generalized Linear Models (GLMs)

---

- Traditional ratemaking methods were not statistically sophisticated
  - Evolved to use classical linear models and minimum bias procedures
- GLMs are an extension of traditional linear models that allow the mean of a population to depend on a predictor through a non-linear link function and allows the response variable distribution to be any member of the exponential family of distributions
  - Allows for explicit assumptions to be made (and validated) about the nature of the insurance data and its relationship with the predictive variables
  - Provide statistical diagnostics for selecting predictive variables and validating model assumptions
  - Primary applications are in rate making and underwriting

## Linear Models are a special case of GLMs

---

- Linear models characterize  $Y$  as the sum of its mean,  $\mu$  and a random variable  $\varepsilon$
- $Y = \mu + \varepsilon$ , assuming the  $E(Y) = \mu$  can be formulated as a linear combination of the covariates  $X$  or

$$\mu = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

and the error term,  $\varepsilon$  is  $N(0, \sigma^2)$

- Linear Model's key assumptions are often violated with insurance data

## GLM assumptions and statistical framework

---

- Random component:  $\underline{Y}$  is assumed to be a member of the exponential family of distributions
- Systematic component: The covariates, or  $\underline{X}$  are combined to give linear predictor  $\eta$ :

$$\underline{\eta} = \underline{X}\underline{\beta}$$

- The relationship between  $\underline{Y}$  and  $\underline{X}$  is specified via a link function,  $g$  that is differentiable and monotonic

$$E[\underline{Y}] = \underline{\mu} = g^{-1}(\underline{\eta})$$

- $\text{Var}[\underline{Y}]$  varies with  $E[\underline{Y}]$

## GLMs include a wide variety of models from the exponential family of distributions

---

- The exponential family of distributions is a 2-parameter family defined as

$$f_i(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}$$

$$E(Y_i) = b'(\theta) = \overset{\text{and}}{\mu_i},$$

$$\text{Var}(Y_i) = b''(\theta) \cdot a_i(\phi) = \phi \cdot V(\mu_i) / \omega_i$$

The exponential family of distributions include: Normal, Poisson, Gamma, Negative Binomial, Binomial, Inverse Gaussian and Tweedie

<u>Y</u>	Claim frequencies	Average claim amounts	Probability (e.g. of renewing)
Link function $g(x)$	$\ln(x)$	$\ln(x)$	$\ln(x/(1-x))$
Error	Poisson	Gamma	Binomial
Scale parameter $\phi$	1	Estimated	1
Variance function $V(x)$	$x$	$x^2$	$x(1-x)^*$
Prior weights $\omega$	exposure	# of claims	1

\*Where the number of trials = 1, or  $y(t-x)/t$  where the number of trials = t

## Theoretical advantages of GLMs include

---

- Fewer theoretical restrictions
- $\underline{Y}$  can model pure premiums, frequencies, severities, probability of renewing, loss ratios, etc.
- $\underline{Y}$  can be non-linear function of  $\underline{X}$
- Reflect correlations and interaction effects of  $\underline{X}$
- Have solid statistical foundation and theory and do offer a practical method for insurance companies to attain a competitive advantage
- Use maximum likelihood for parameter estimates

## GLMs in Practice

---

- A GLM analysis usually contains the following steps
  - Pre-modeling analysis, reconnaissance and roadmap – considers data sources and preparation, competitors, clear articulation of desired pricing/rating objectives
  - Data gathering and cleansing
  - Analysis and model development
  - Refinement of pricing or underwriting approach
  - Implementation and monitoring

# GLMs in practice data requirements and issues

---

- Overall structure of dataset consists of linking policy and claims information at the individual risk level
  - Raw explanatory variables – internal and external to insurance company
  - Dummy variables – standardize time-related effects, geographic, historical underwriting effects and other category effects
  - Loss experience – frequency and severity information
  - Premium data – attributable to policy record
  - Exposure data – attributable to policy record

## What are the concerns for a smaller company considering predictive modeling?

---

“If we do nothing, adverse selection will happen.”

“Do we have enough data to do this type of analysis?”

“What kind of analysis can we do with the data we do have?”

## Some specific concerns among smaller companies

---

- Current trends are in the wrong direction
  - Drop off in policies written
  - Low retention on parts of the book
  - Deteriorating loss ratios
- Some evidence of adverse selection
  - Drift toward worse credit-based insurance scores
- Frustration among producers
  - You always used to be competitive for a multi-car, multi-line, clean driving record

## A smaller company may have some advantages over larger companies

---

- Better local market knowledge
  - Agents may be a good source of information
- Implementation may not draw too much attention
  - Larger competitors may not care
  - Regulatory objections may be less frequent
- May be more nimble, and able to implement changes faster
  - More institutional knowledge — in a smaller group of people
  - Better communication between departments
  - More streamlined management decision-making
  - One wildcard is capacity/capability of policy processing systems

# Exploration is fun, but it's better to have a plan . . .

---

- Identify business goals
  - Rating plans: Better accuracy, new structure, new variables
  - Retention/elasticity: Problem areas, impact of planned changes
- Identify limitations
  - Timeline
  - Budget
  - Data availability
- Develop your plan
  - Broad questions → more leeway on data and analysis
  - Study of all coverages / perils combined or separately?
  - Will a frequency study suffice?
  - Will the same factor apply to all coverages?

## A smaller company should probably not just...

---

- Add a credit score (or some other variable) based on a competitor's rating plan, on top of the existing rating plan
  - This over-discounts some classes, and under-discounts others
  - May have implications for new business growth
- Attempt to create a full-blown class plan on their own data
  - Less reliable results if data are too thin
    - Variables may not be significant
    - Levels within variables may be volatile
  - Models can be unstable as variables are added or dropped

## There are things that can be done — subject to limitations

---

- Given many records and data variables
  - Analyze main rating variables and rate relativities
  - Explore new variables or variable interactions
  - Analyze territory boundaries and relativities
    - Be cautious about calculating directly in a GLM
- Fewer records and/or fewer variables suggest simpler analyses
  - Tier definitions and tier relativities
  - Refine major risk factors (subdivide some categories)
  - Underwriting and/or schedule debit/credit guidance
  - Rely more heavily on competitive analysis for other key variables

## Don't forget about retention models

---

- Identify groups with better / worse than average retention
- Identify specific events which might influence retention
  - Look at policy change like change or add car, change or add drivers, etc.
  - Rate changes
  - Isolate voluntary vs. involuntary attrition
- Effects of competition

## Match model complexity to data and goals

---

- Underwriting models or tier analysis
  - Probably sufficient to control for main rating characteristics, but not necessarily use results to change those factors
  - Initially, want direction and magnitude, not precision
- Retention — you may want to isolate company actions
  - Insurance to value program, changes in some billing options, etc.
    - Again, control for main characteristics
    - Add appropriate indicator variables
- Elasticity/rate impact analysis adds other twists
  - May want to restrict to short intervals
    - Management may want quick feedback: how did rate change affect mono-line vs. multi-line risks?
  - Lots of exogenous variables change quickly – e.g. competitive information

## How many records will I need to model loss data?

---

- Think claims, not exposures
  - For a relatively high frequency line of business, fewer exposures are needed
- Would prefer 5,000+ claims
  - This is a very rough rule of thumb
  - A lot depends on what you're trying to analyze
- If you're short on claims, pursue more years of data

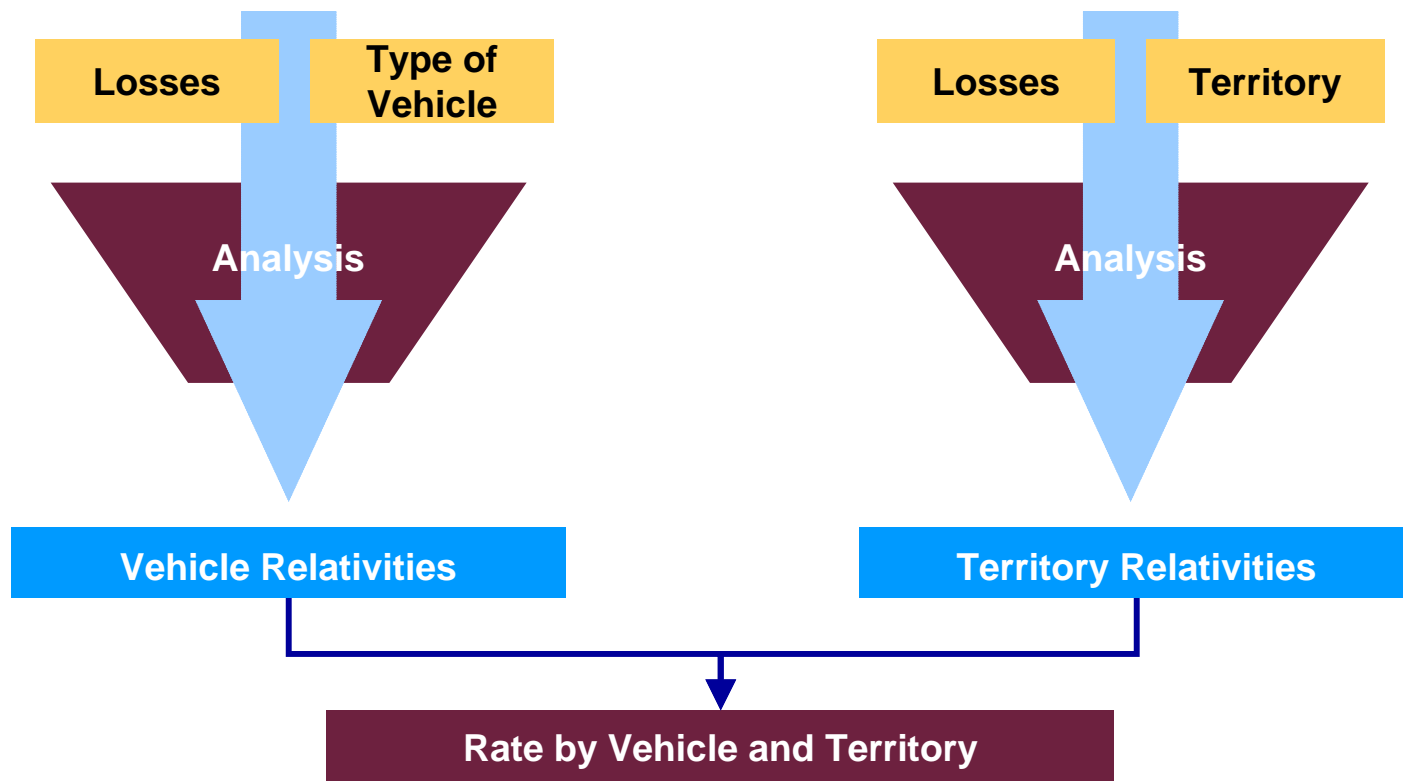
## Identifying variables to include

---

- Brute force searches using predictive modeling is not a good substitute for subject matter expertise
- Speak with underwriters and claim adjusters
  - They'll possibly have a sense for important variables
  - They may have an idea of magnitude
- Understand competitors' approaches (don't reinvent the wheel)

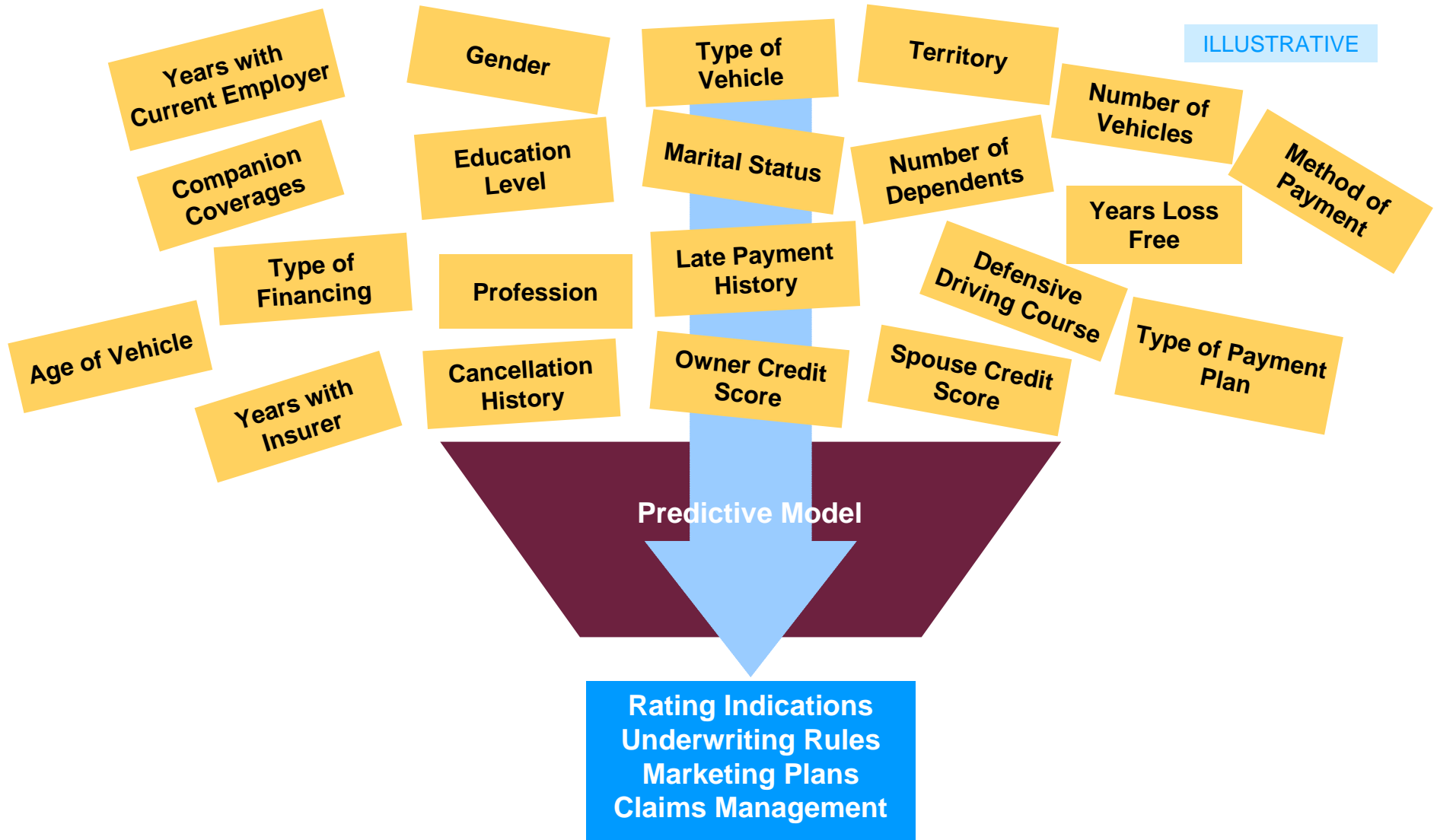
# Traditional vehicle and territory approach: One characteristic at a time

ILLUSTRATIVE



Repeat for all other rating factors

# Predictive modeling: All relevant characteristics in combination, including non-insurance characteristics



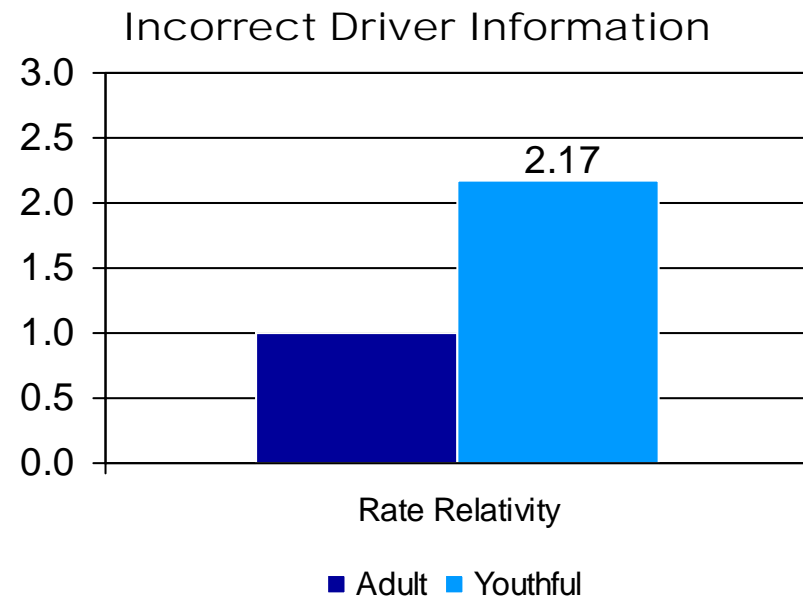
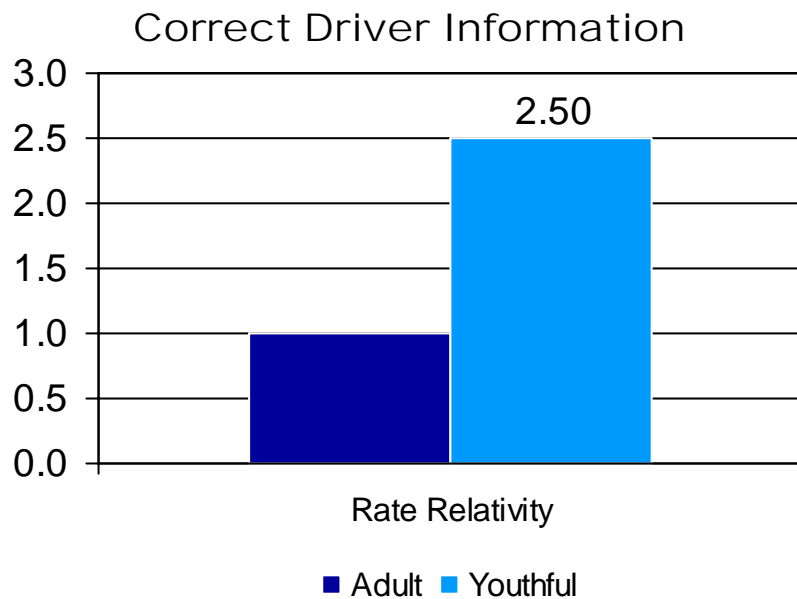
# Data preparation is the key to success

---

- Ideally want near transaction-level data
  - At the policy (vehicle) level
  - A separate record representing the risk characteristics in effect for that period
    - When a variable changes, add a new record
- Prefer to have only one claim per policy exposure segment
  - If more than one claim, can split the premium record into finer segments
  - Can usually aggregate claims to match to a given period
    - Parameter estimates are unchanged
    - Standard errors are understated

# Messy data causes problems

- Model may not converge (you'll have to fix the data anyway!)
- Poor coding will suppress differences between classes
- Hypothetical example - 20% of vehicles not coded as youthful
  - Rate relativities are understated for young drivers



## Other potential data problems

- Severity analyses – very small claims
  - Make sure they're not miscodes
  - Possibly an issue with salvage/subrogation
- Frequency analyses – claims on very small exposure periods
  - Check to see if there are system issues
- Miscoded/missing data
  - Often highly correlated for several independent variables
  - “Near-aliasing” can result, and cause convergence problems

Construction / Hurricane Protection	1-4	5-8	9	10	Miscoded Hurricane Protection
Hurricane Resistive	3000	2000	800	500	0
All Other Valid Types	6000	4500	1600	1000	5
Miscoded Construction	0	0	0	0	395

- One option – force “All Other Valid Types”/”Miscoded Hurricane Protection” to “Miscoded Construction”

## Potential ways to simplify data preparation

---

- Accumulate data on a policy year basis
- Take snapshot at beginning of period
- Take snapshots quarterly
- Worry about “important” policy changes
  - Change in vehicles
  - Change in driver
  - Add/drop coverage

## Do you already have what you need?

---

- If your current data system has:
  - Merged premium and loss data partially summarized
  - Includes all major characteristics
  - Individual claim data is separately available with risk characteristics attached
  - You could:
    - Build a frequency model on aggregated results
    - Do a separate claim severity model
    - Combine frequency and severity factors manually
- If individual claim data are not available
  - You could do pure premium analysis directly using a Tweedie distribution
- It may not be possible to add new variables when using summarized data

## Initial model design

---

- Although it might seem counter-intuitive, consider separate frequency and severity models
  - Frequency is often the predominate contributor to cost differences
  - Standard errors are usually tighter – more variables survive vetting
  - Isolates claim size volatility (often end up with simpler severity models)
  - Then combine into pure premium relativities

## Modeling issues

---

- Big danger is over-fitting
  - Find signal, not noise
  - Judgment will often be needed in face of volatility
  - When possible, split the data (out of sample predictive validation)
  - Make use of competitor information to check magnitude and direction of results
- General understanding of modeling will help maximize value of your data
  - Data volume
  - Model design

# Summary: modeling for small companies v. larger companies

- Keep in mind that modeling may be harder for smaller companies
- Don't try to do more than is realistic for the amount of data you have

Business Goal	Likelihood of Success
Develop full class plan	✓
Refine granularity for a given variable	✓✓✓✓
Explore new variables	✓✓✓
Explore variable interactions	✓✓✓✓✓
Develop tier definitions/factors	✓✓✓
Develop underwriting models	✓✓✓
Analyze retention	✓✓✓✓

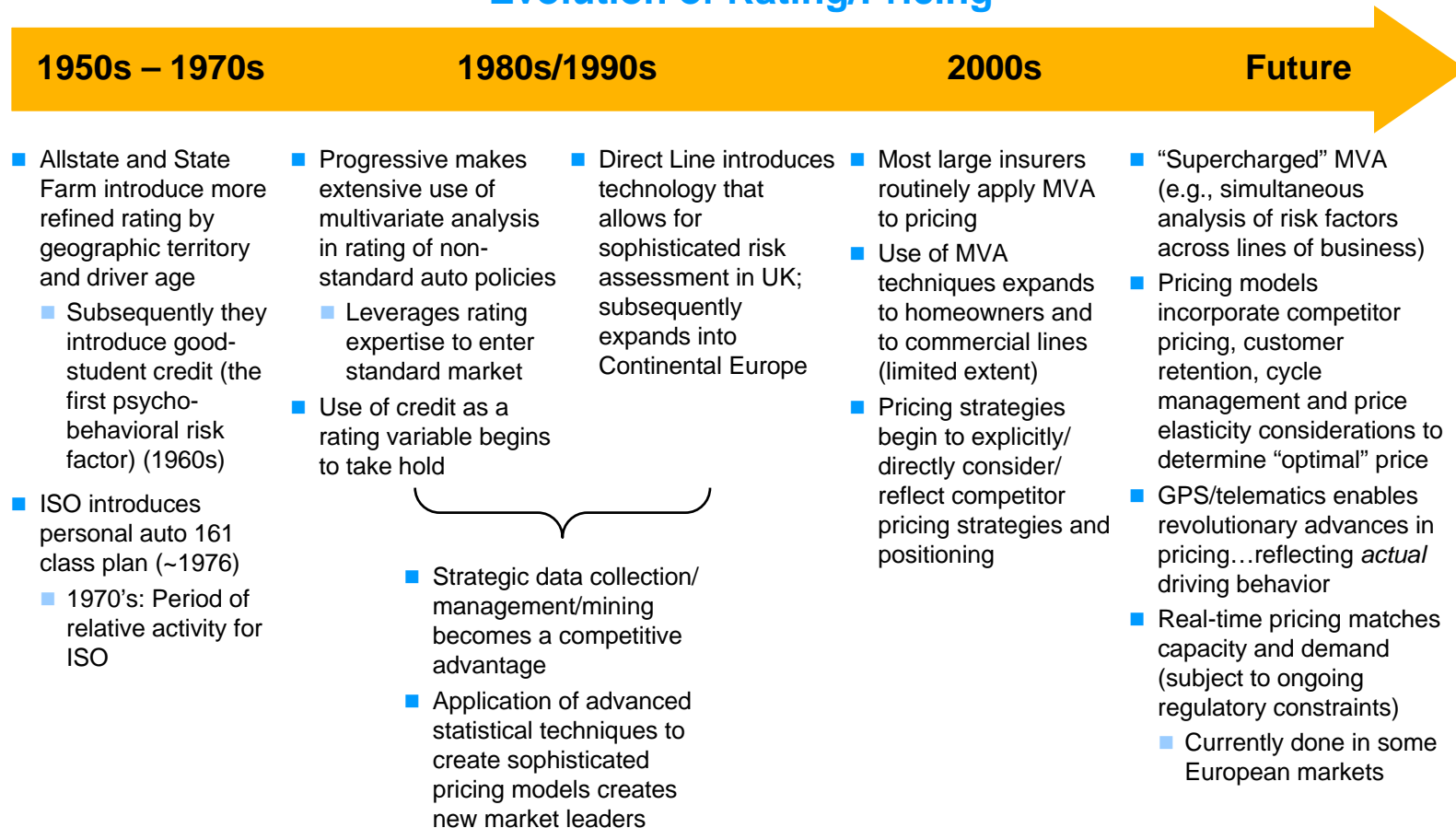
✓ Low likelihood



High Likelihood  
(Large Company)

# Developments in rating/pricing have fundamentally changed the competitive environment

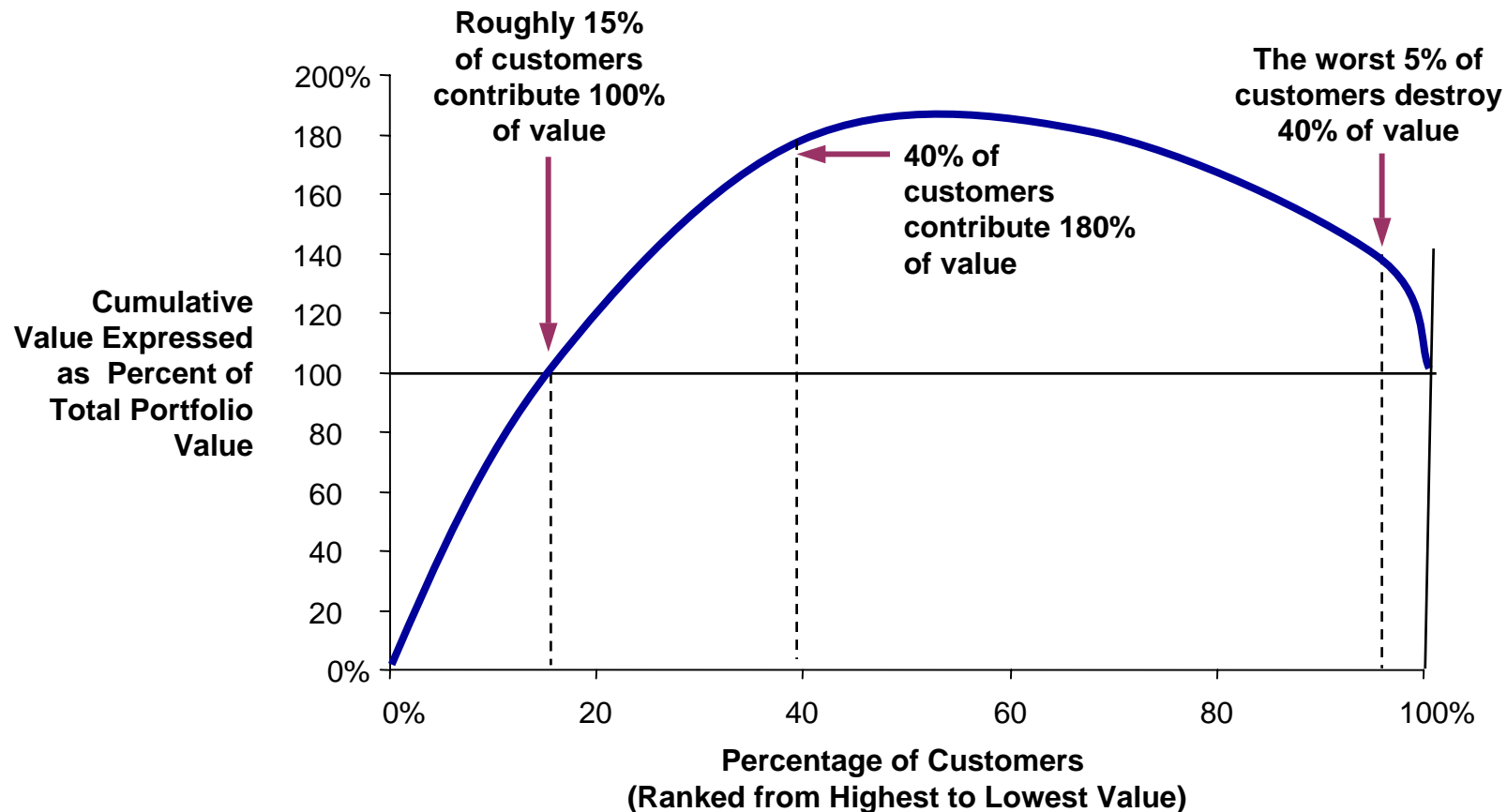
## Evolution of Rating/Pricing



Profound changes in market leadership occurred because of dramatic advances in risk assessment, coupled with the courage to implement that knowledge via more accurate pricing

Market leaders took advantage of the prevailing economic dynamics to change the rules of the game...

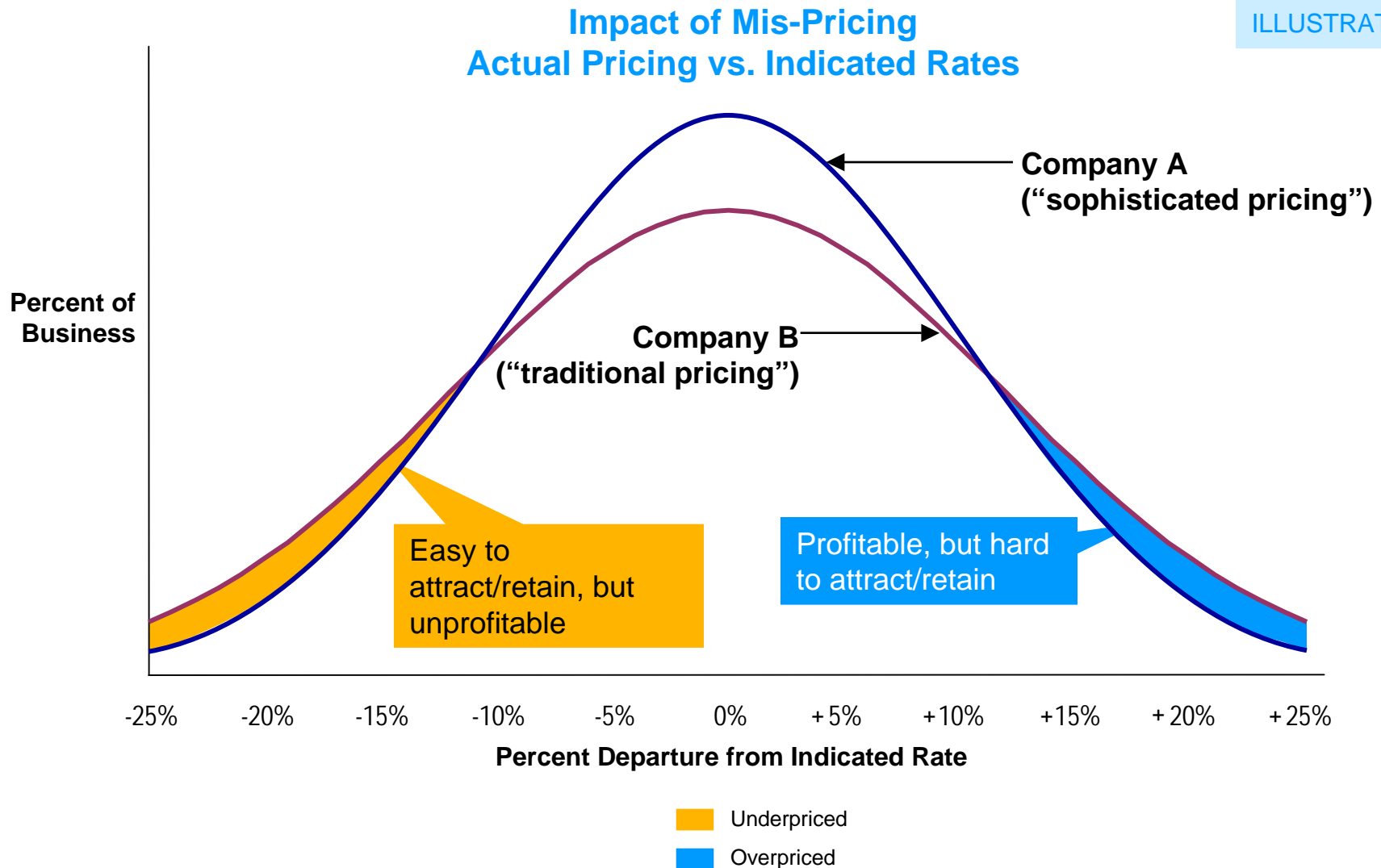
ILLUSTRATIVE



What if you could predict which policyholder would have losses better than your competitors?

To illustrate, inaccurate rates make it difficult to attract and retain the business you want

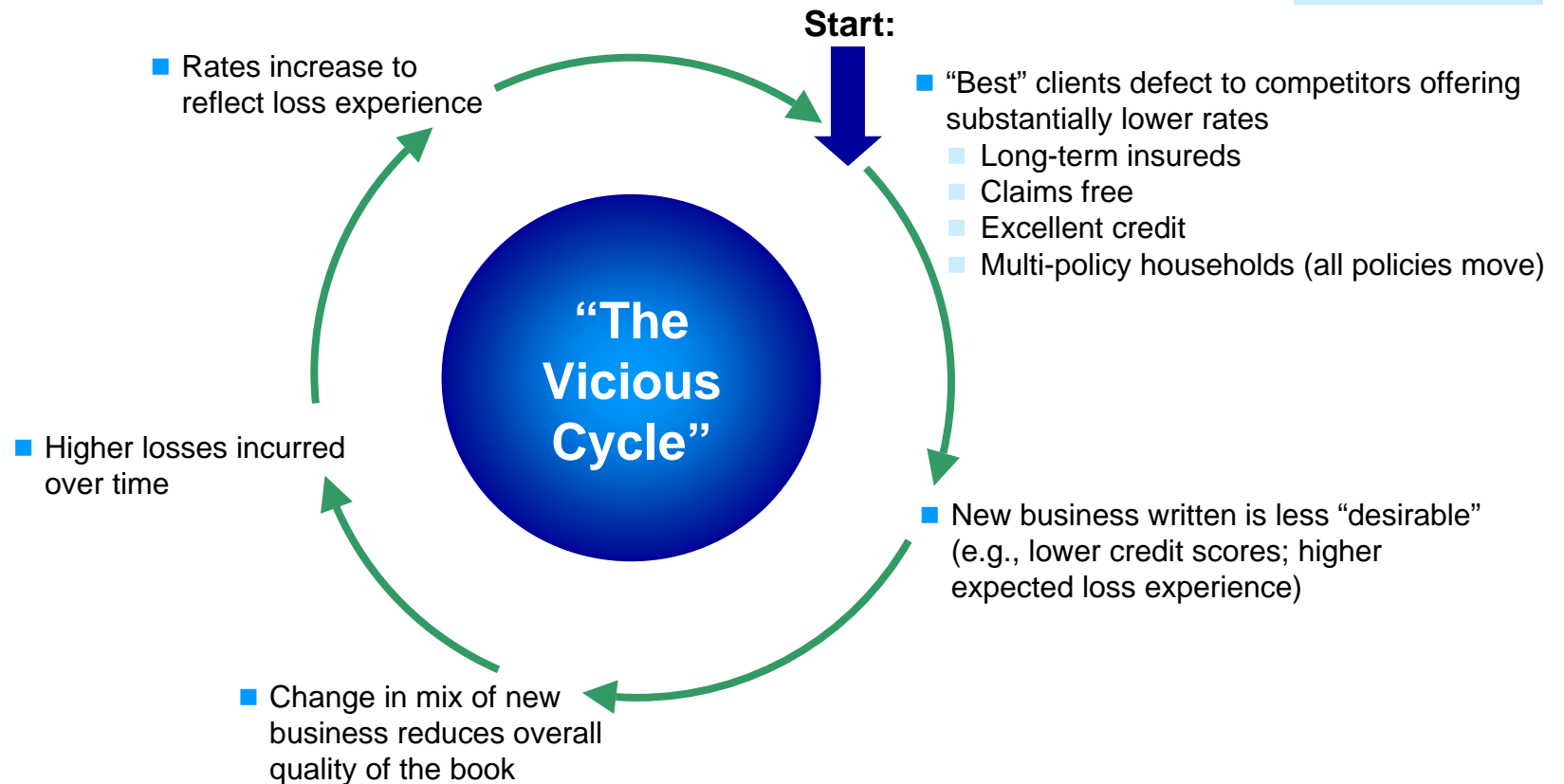
ILLUSTRATIVE



Companies that don't play — or that don't play aggressively — are fundamentally disadvantaged and subject to adverse selection

- Multi-line exclusive agent: *"I'm seeing defections of my best clients coupled with an inability to attract desirable new business"*

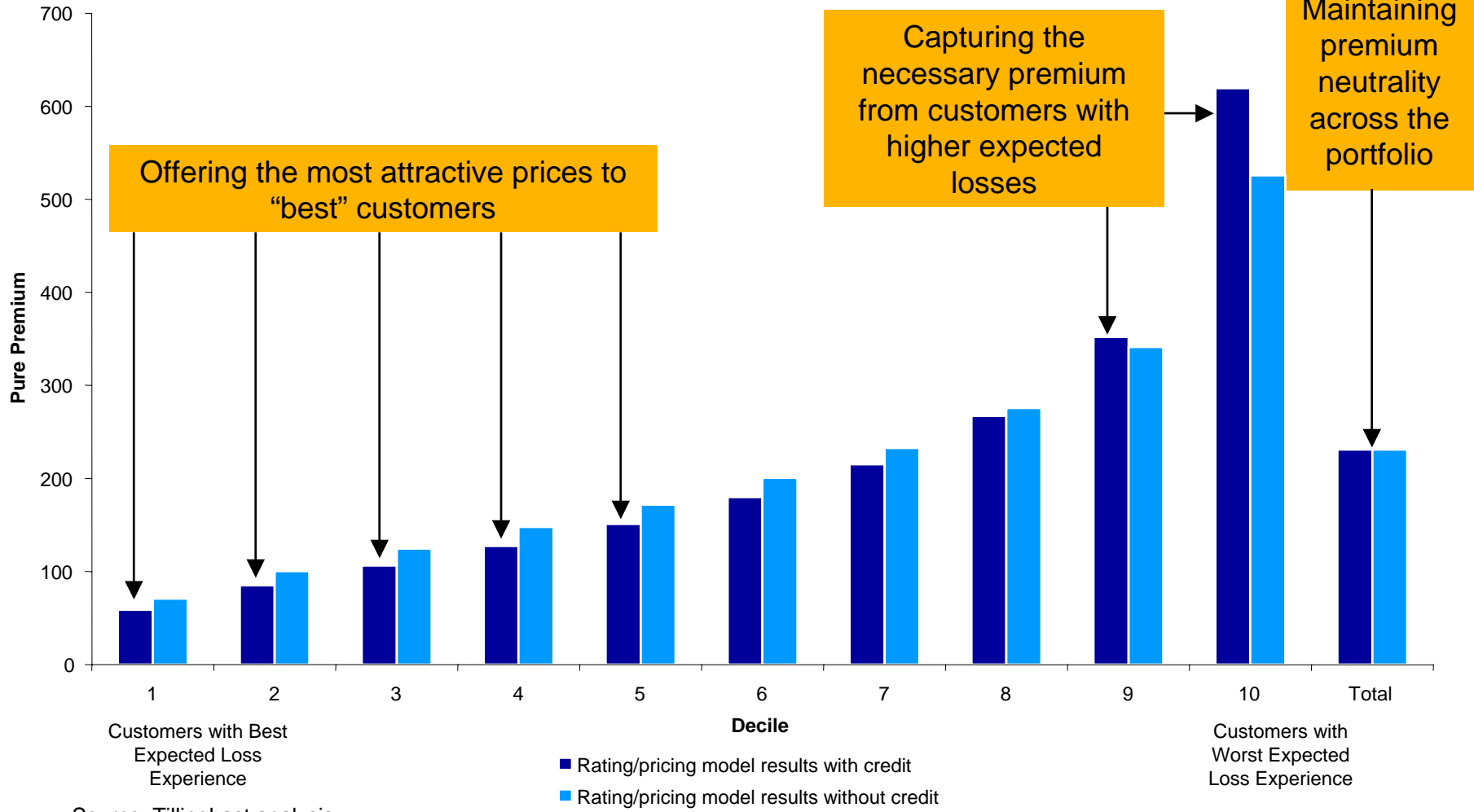
ILLUSTRATIVE



The use of credit scoring — itself a powerful predictor of loss — exemplifies the opportunity to properly align pricing

### Premium Impact of Using Credit Scores Across Policyholder Segments

ILLUSTRATIVE



Source: Tillinghast analysis.

There always will be competitors that are running faster — you don't need to be the fastest, just faster than others

- The “fast runners” are seeking to accurately rate all segments of their books — they seek more granularity and refinement
- The greater the pricing accuracy, the less concern over adverse selection and changes in the business mix

In today's market, sophisticated pricing is “needed to play”

# The benefits of pricing enhancements fall into two main areas

---

## Top-Line Growth

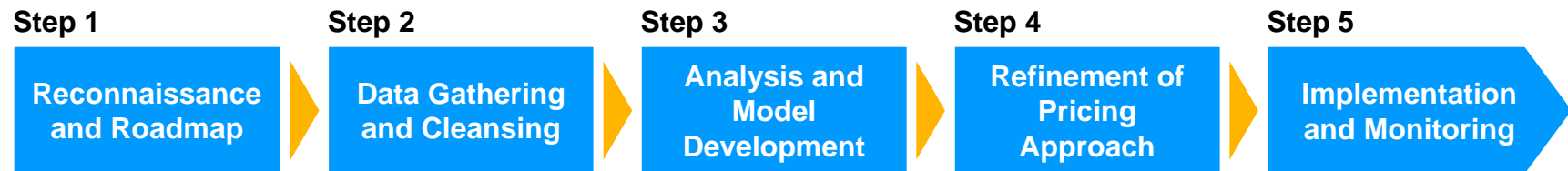
- Opportunity to grow premium and market share (and/or prevent deterioration in market share/position)
  - Partly reflects ability to compete across broader spectrum of risks

## Loss Ratio Improvement

- Our experience suggests the potential for a 2% to 4% improvement in loss ratio

A disciplined/systematic approach  
to enhance pricing sophistication is essential

### Overview of Basic Approach



#### Primary Outputs/Deliverables

- A clear articulation of the desired pricing/rating/positioning
  - A statistical model for determining price relativities
  - An updated rating plan that reflects new variables, interactions, tiering, revised territory definitions, etc.
  - High-level implementation (and change management) plan
  - Recommended priorities for next-generation enhancements
- 
- The duration of these engagements can vary considerably and particularly depends on data quality and availability — a “typical” assignment might take six to nine months to get to the point of implementation

## Closing thoughts

---

- Predictive modeling can help you make better business decisions
  - Even simple analyses can be better than traditional approaches
  - At a minimum, it can help you convince other stakeholders when contemplating something new
- Leverage what you've learned from PM
  - Develop monitoring / early warning reports
    - Monthly retention reports for desired/undesirable segments
    - Policies written or quotes made by segment
- Develop a strategy for the future
  - Plan to move beyond 'one-off' database construction
  - Identify data variables which you might have available while competitors do not
  - Make sure any newly identified variables are available for easy merge with traditional data

## Contact information

---

Russell Greig

404-365-1707

[russell.greig@towersperrin.com](mailto:russell.greig@towersperrin.com)