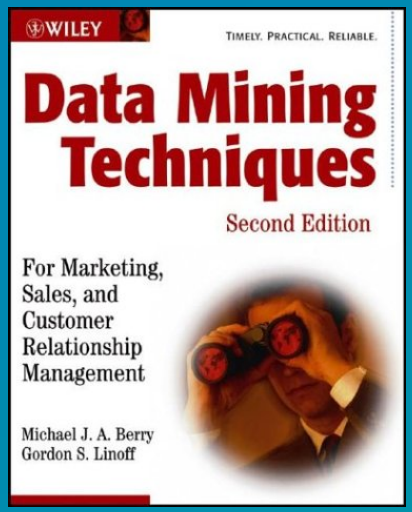
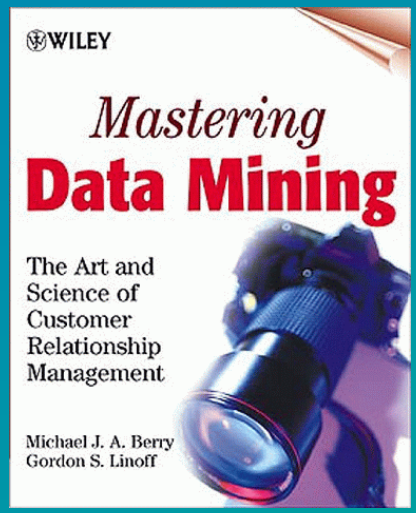
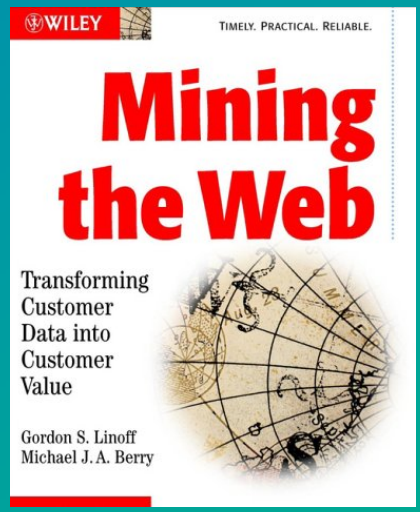






Survival Data Mining

Gordon S. Linoff

Founder

Data Miners, Inc.

gordon@data-miners.com

Data Miners

What to Expect from this Talk

- Background on survival analysis from a data miner's perspective
- Introduction to key ideas in survival analysis
  - hazards
  - survival
  - competing risks
- Lots of examples
  - Stratification
  - Quantifying Loyalty Effort
  - Voluntary and Involuntary Churn
  - Forecasting
  - Time to Reactivation and Re-purchase

Data Miners

Who Am I?

- I am not a statistician
- Adept with databases and advanced algorithms
- Founded Data Miners with Michael Berry in 1998
- We have written three books on data mining
- Have become very interested in survival analysis for mining customer data – survival data mining

Data Miners

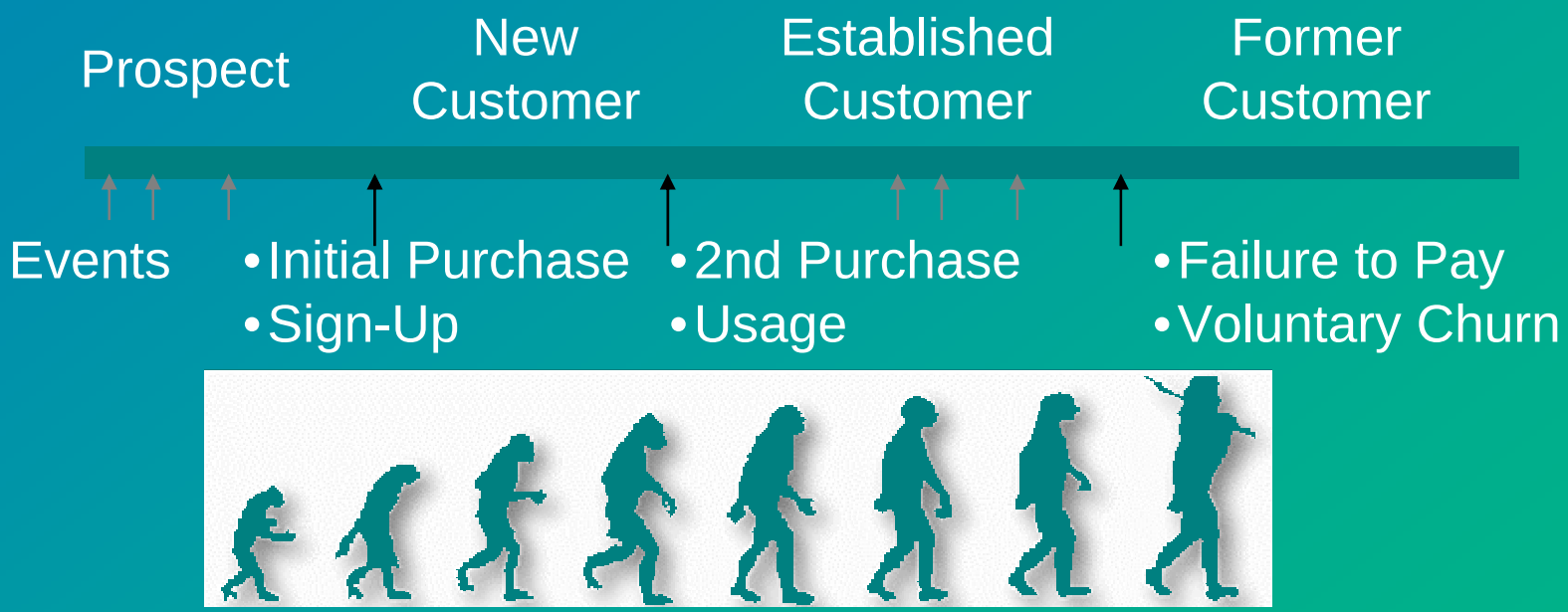
What Does Data Mining Really Do?

- Provides ways to *quantitatively* measure what business users know or should know *qualitatively*
- Connects data to business practices
- Used to understand customers
- Occasionally, produces interesting predictive models

Data Miners

Data Mining Is About Customers

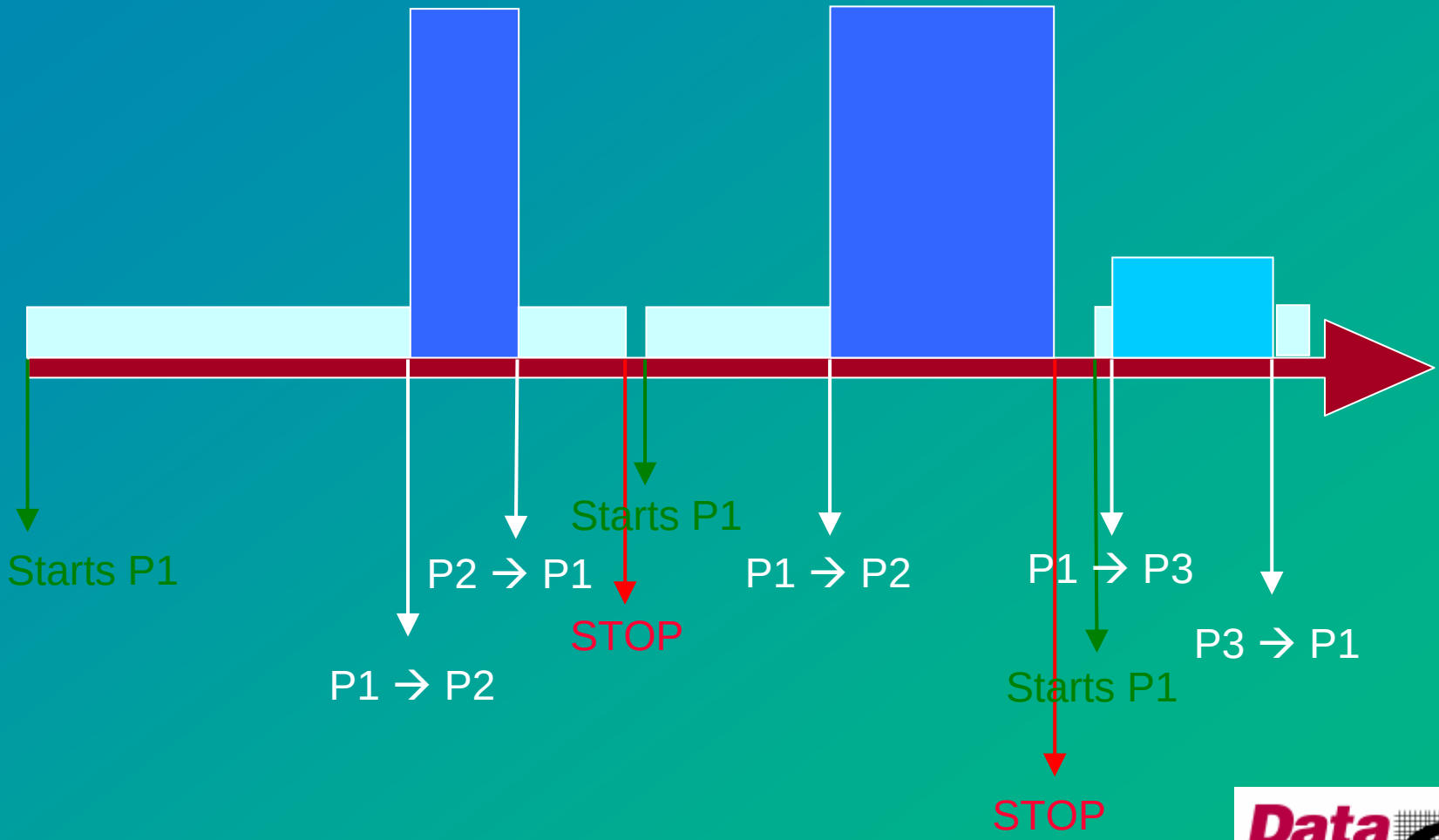
Customer Relationship Lifetime



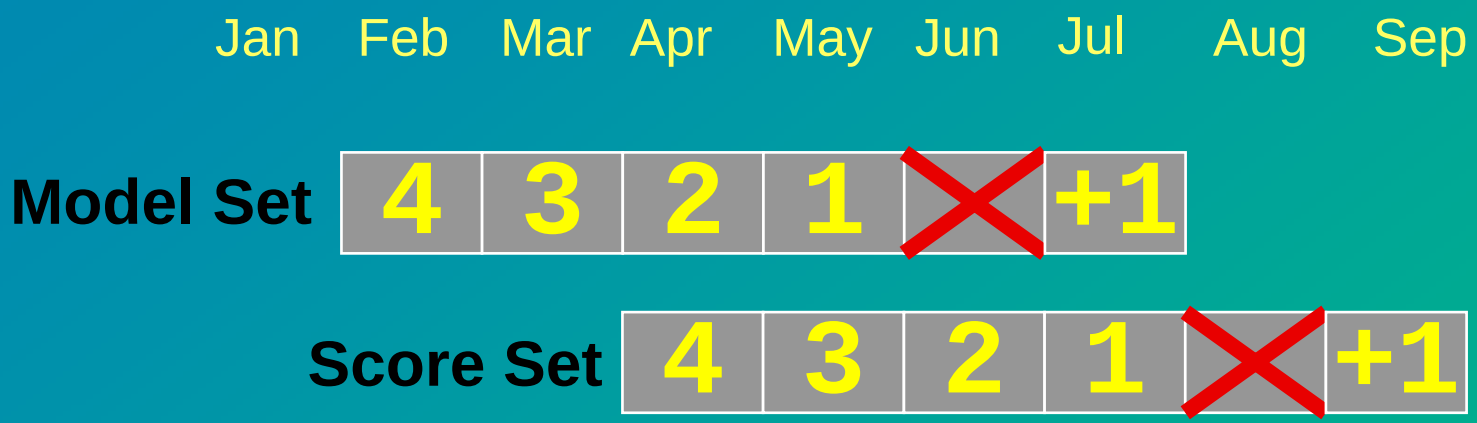
Customers Evolve Over Time

Data Miners

The State of the Customer Relationship Changes Over Time



Traditional Approach to Data Mining Uses Predictive Modeling



- Data to build the model comes from the past
- Predictions for some fixed period in the future
- Present when new data is scored
- Models built with decision trees, neural networks, logistic regression, regression, and so on

Data Miners

Survival Data Mining Adds the Element of When Things Happen

- Time-to-event analysis
- Terminology comes from the medical world
  - which patients survive a treatment, which patients do not
- Can measure effects of variables (initial covariates or time-dependent covariates) on survival time
- Natural for understanding customers
- Can be used to quantify marketing efforts

Data Miners

Example Results (Made Up)

- 99.9% of Life bulbs will last 2000 hours
- Mean-time-to-failure for hard disk is 500,000 hours
- *"A recent study published in the American Journal of Public Health found that '**life expectancy**' among smokers who quit at age 35 exceeded that of continuing smokers by 6.9 to 8.5 years for men and 6.1 to 7.7 years for women."* [www.med.upenn.edu/tturc/pdf/benefits.pdf]
  - Example of initial covariate
- Stopping smoking before age 50 increases lifespan by one year for every decade before 50

Data Miners

Original Statistics

- Life table methods used by actuaries for a long, long time
  - These the are the methods we will be focused on
- Applied with vigor to medicine in the mid-20th Century
- Applied with vigor to manufacturing during 20th Century as well
- Took off with Sir David Cox's Proportion Hazards Models in 1972 that provide effects of initial covariates

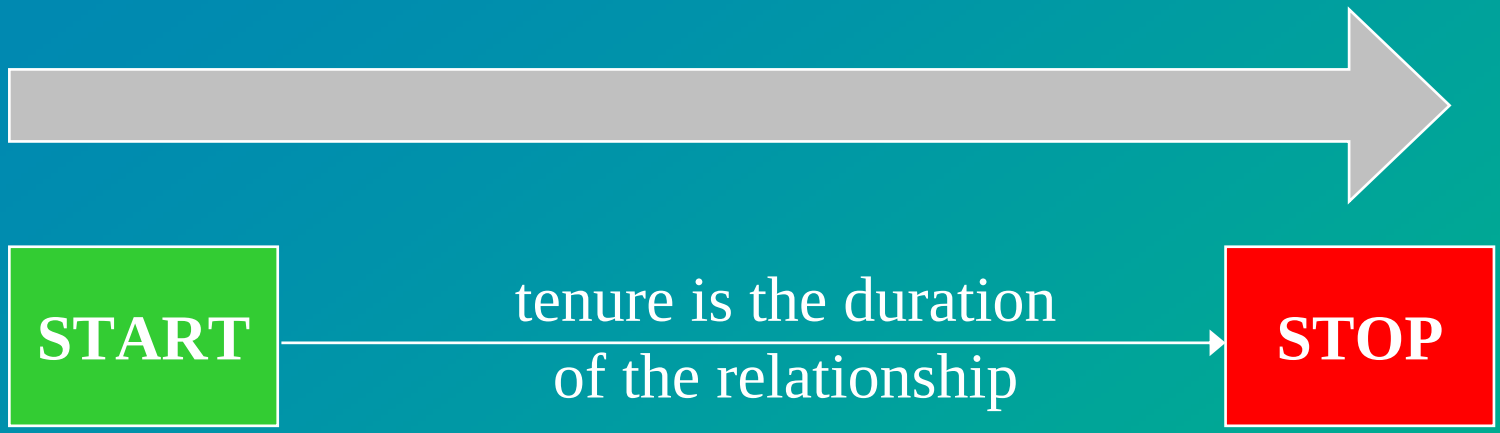
Data Miners

Survival for Marketing Has Some Differences

- We are happy with discrete time (probability vs rate)
  - Traditional survival analysis uses continuous time
- Marketers have hundreds of thousands or millions of examples
  - Traditional survival analysis might be done on dozens or hundreds of participants
- We have the benefit of a wealth of data
  - Traditional survival analysis looks at factors incorporated into the study
- We have to deal with “window” effects due to business practices and database reality
  - Traditional survival ignores left truncation

Data Miners

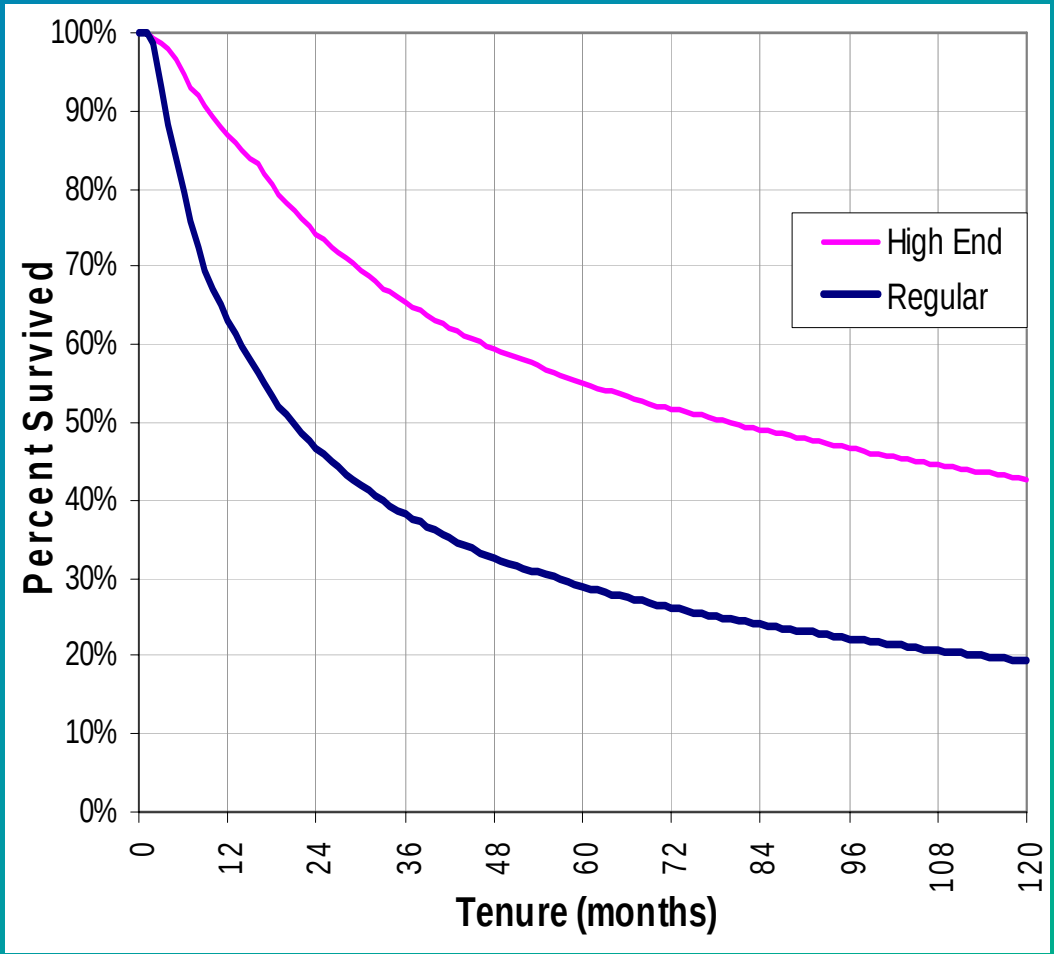
To Understand the Calculation, Let's Focus on the End of the Relationship



- Challenge defining beginning of customer relationship
- Challenge defining end of customer relationship
- Challenge finding either in customer databases

Data Miners

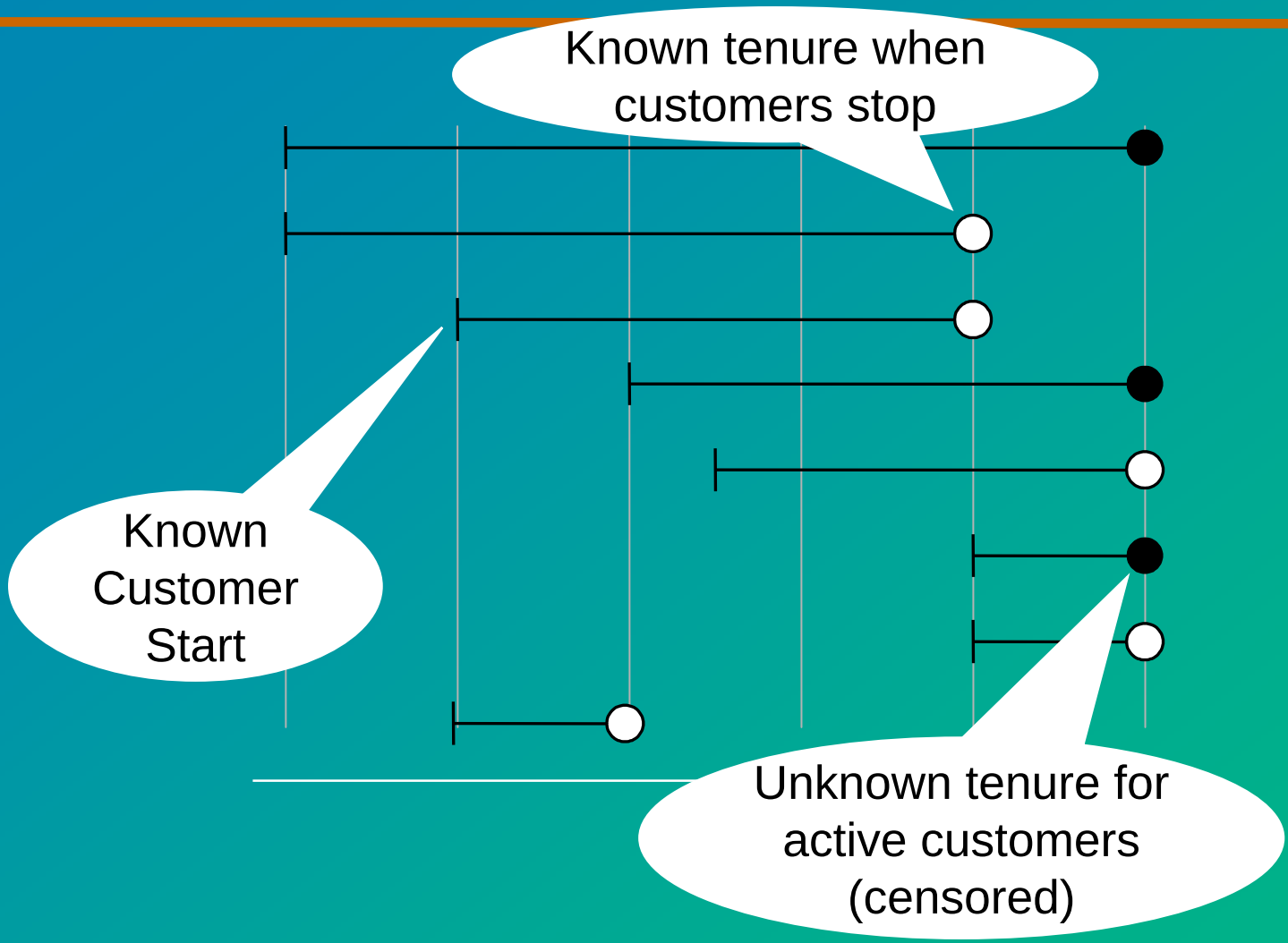
How Long Will A Customer Survive?



- Survival always starts at 100% and declines over time
- If everyone in the model set stopped, then survival goes to 0; otherwise it is always greater than 0
- Survival is useful for comparing different groups

Data Miners

Key Idea: Data is *Censored*



Data Miners

Use Censored Data to Calculate Hazard Probabilities

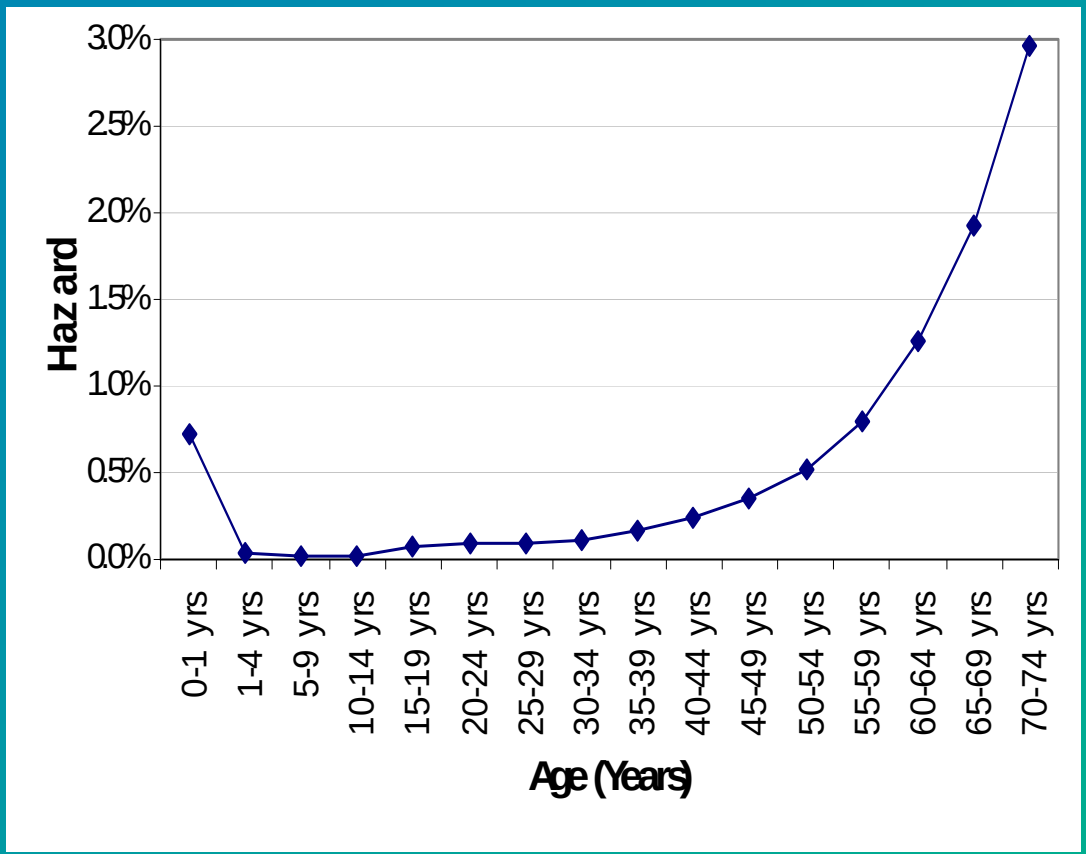
- Hazard, h(t), at time *t* is the probability that a customer who has survived to time *t* will not survive to time *t+1*

$$h(t) = \frac{\#\ customers\ who\ stop\ at\ exactly\ time\ t}{\#\ customers\ at\ risk\ of\ stopping\ at\ time\ t}$$

- Value of hazard depends on units of time – days, weeks, months, years
- Differs from traditional definition because time is discrete – hazard probability not hazard rate

Data Miners

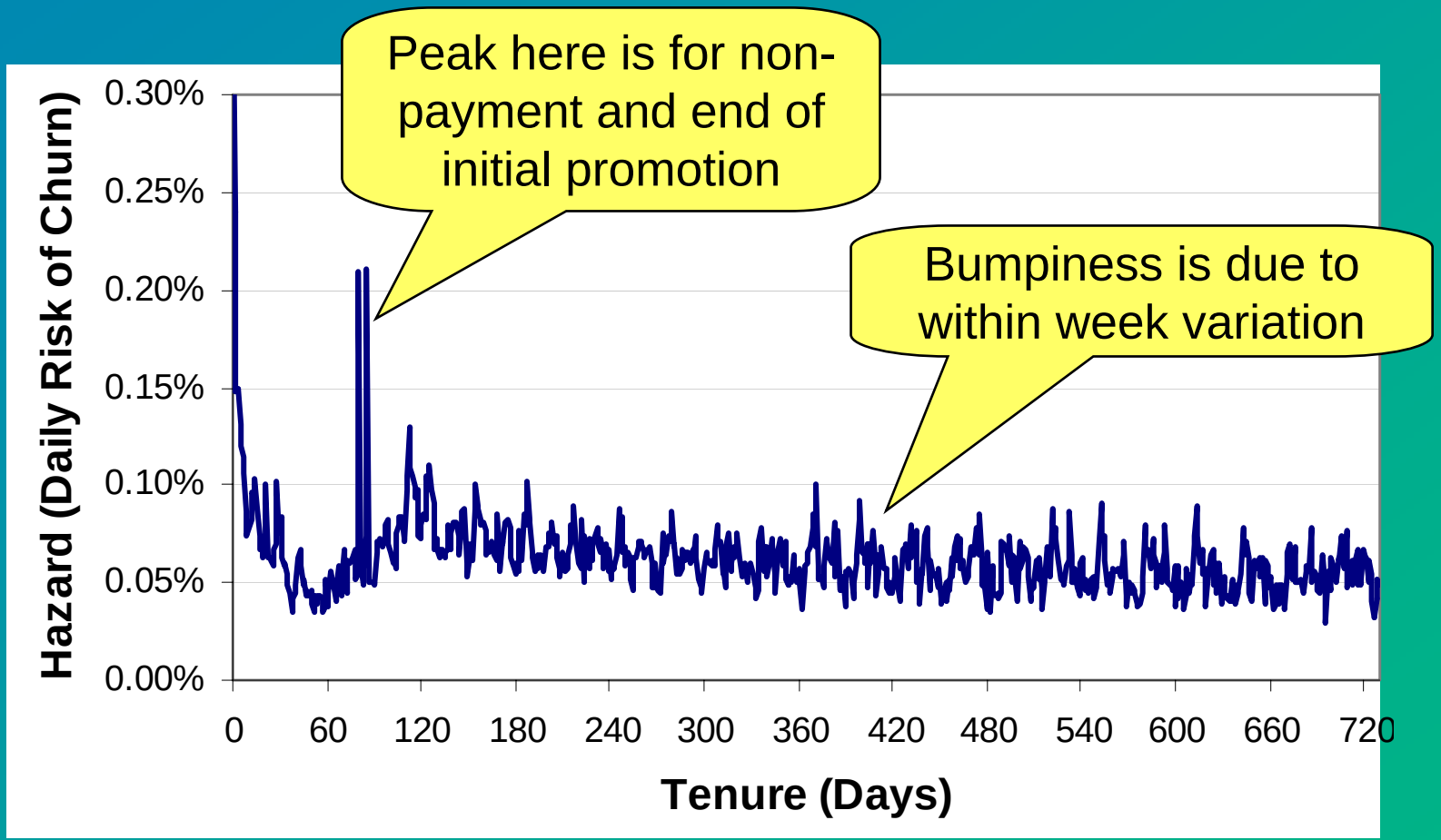
Bathtub Hazard (Risk of Dying by Age)



- Bathtub hazard starts high, goes down, and then increases again
- Example from US mortality tables shows probability of dying at a given age

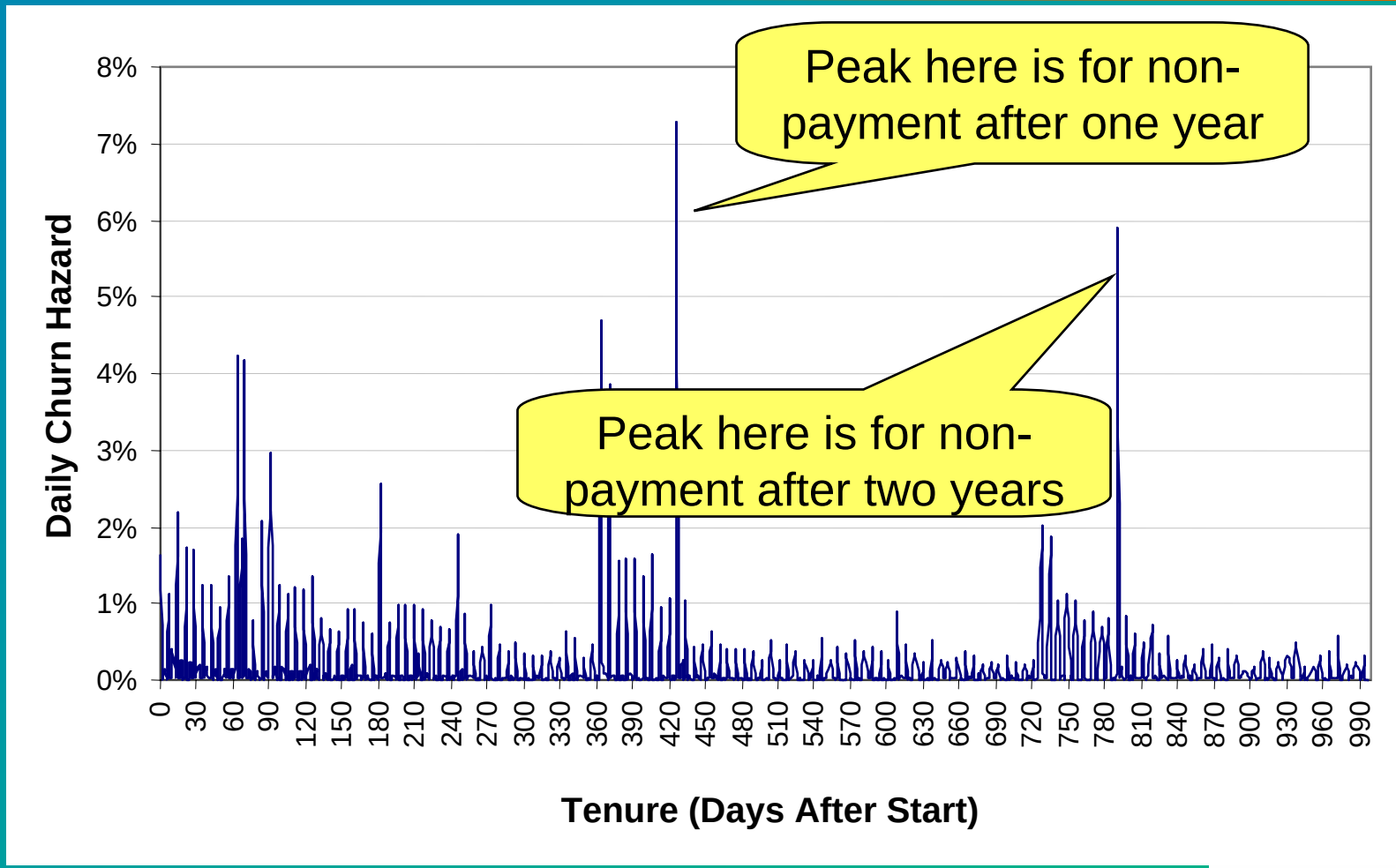
Data Miners

Hazards Are Like An X-Ray Into Customers



Data Miners

Hazards Can Show Interesting Features of the Customer Lifecycle



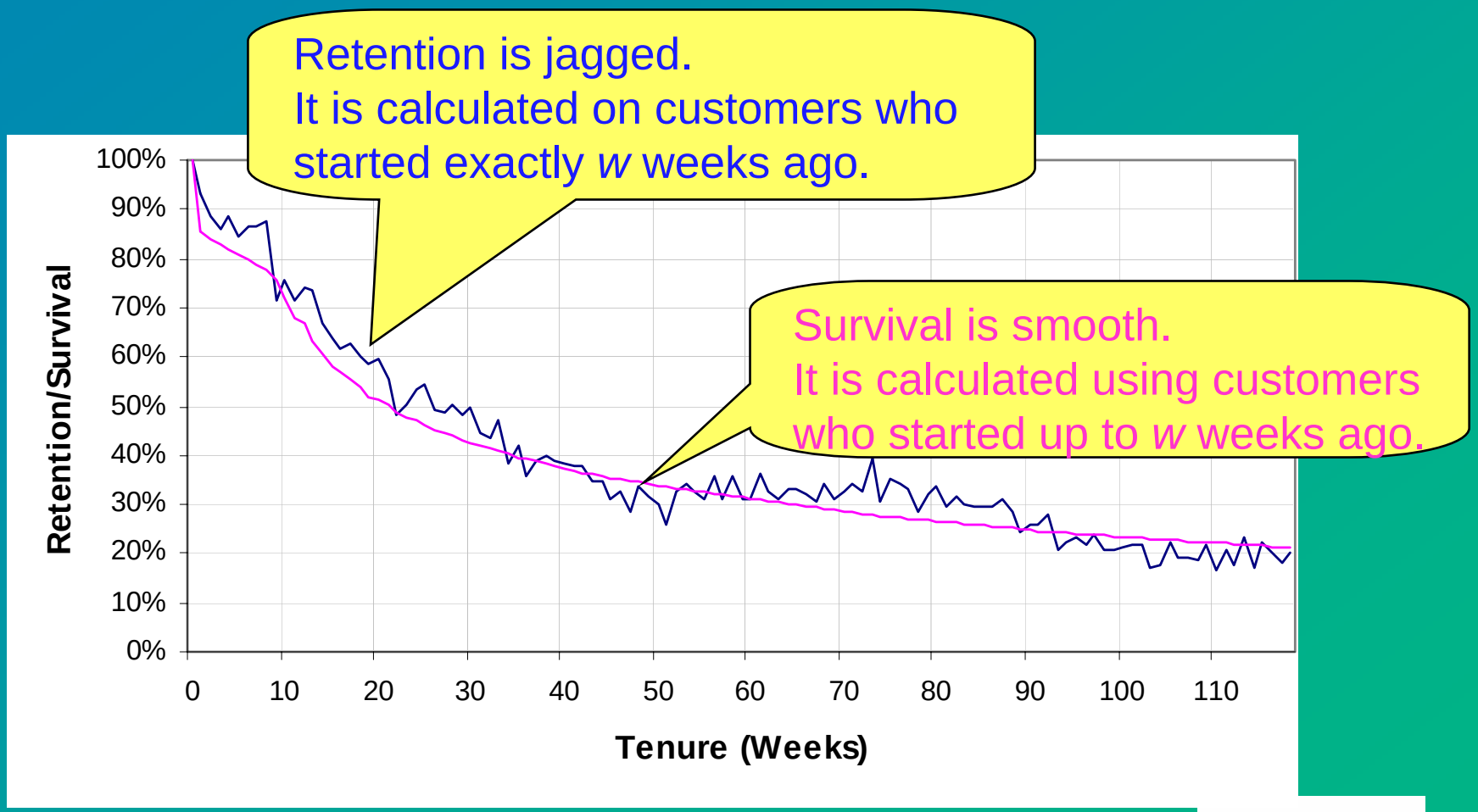
Data Miners

How Long Will A Customer Survive?

- Survival analysis answers the question by rephrasing it slightly:
  - What proportion of customers survive to time *t*?
- Survival at time t, S(t), is the probability that a customer will survive exactly to time t
- Calculation:
  - $S(t) = S(t-1) * (1 - h(t-1))$
  - $S(0) = 100\%$
- Given a set of hazards by time, survival can be easily calculated in SAS or a spreadsheet

Data Miners

Survival is Similar to Retention



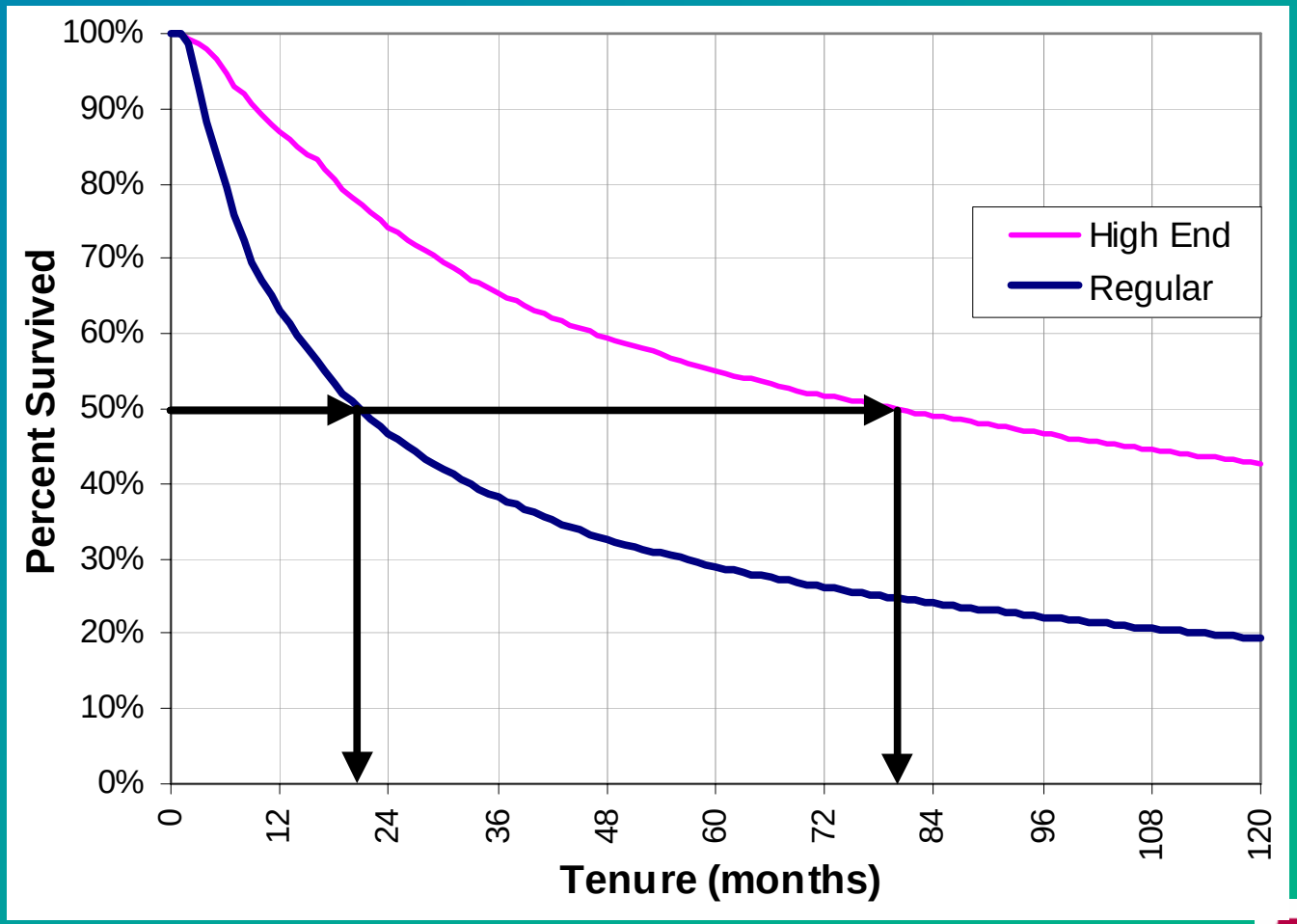
Data Miners

Why Survival is Useful

- Compare hazards for different groups of customers
- Calculate median time to event
  - How long does it take for half the customers to leave?
- Calculate truncated mean tenure
  - What is the average customer tenure for the first year after starting? For the first two years?
  - Can be used to quantify effects

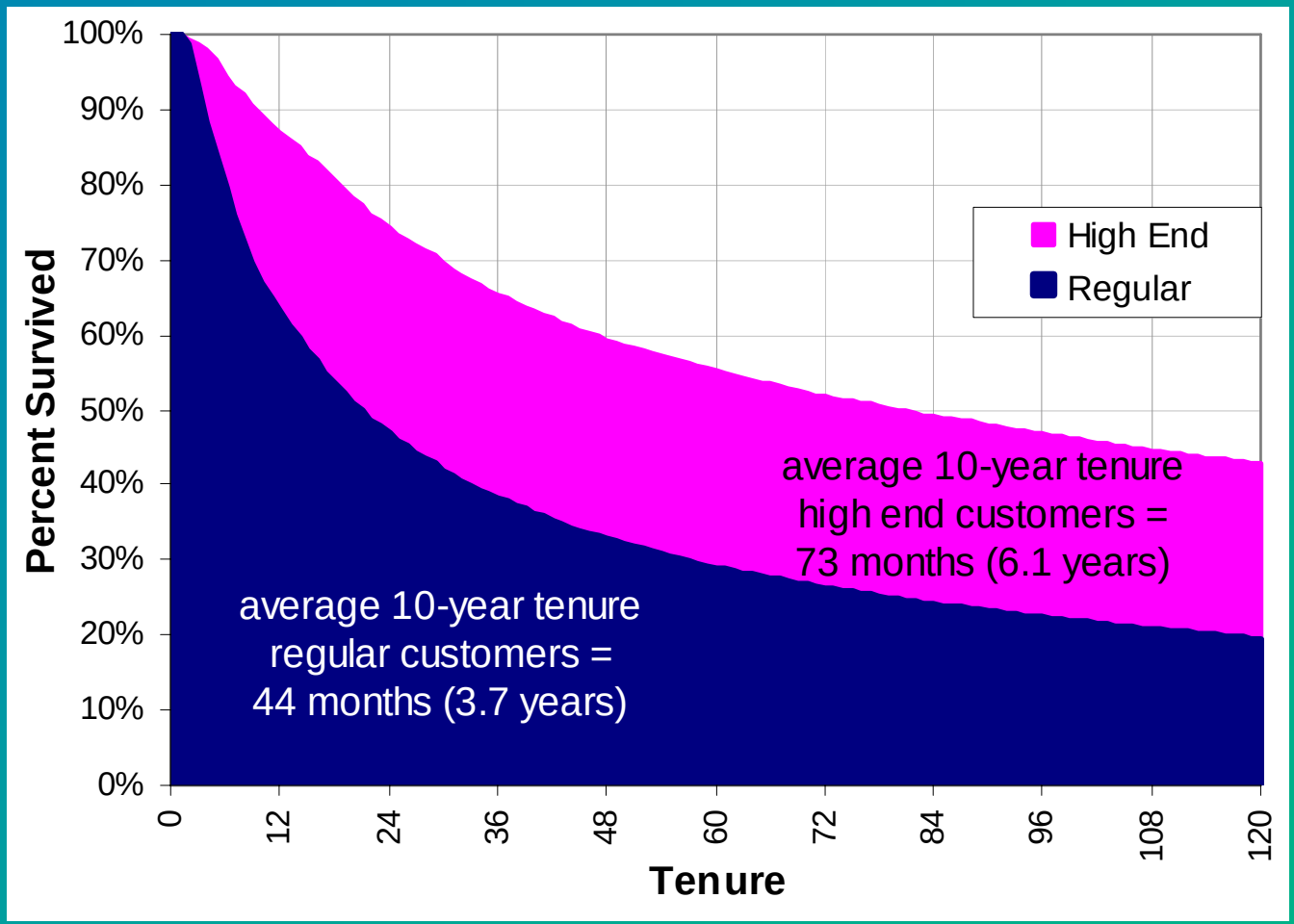
Data Miners

Median Lifetime is the Tenure Where Survival = 50%



Data Miners

Average Tenure Is The Area Under The Curves



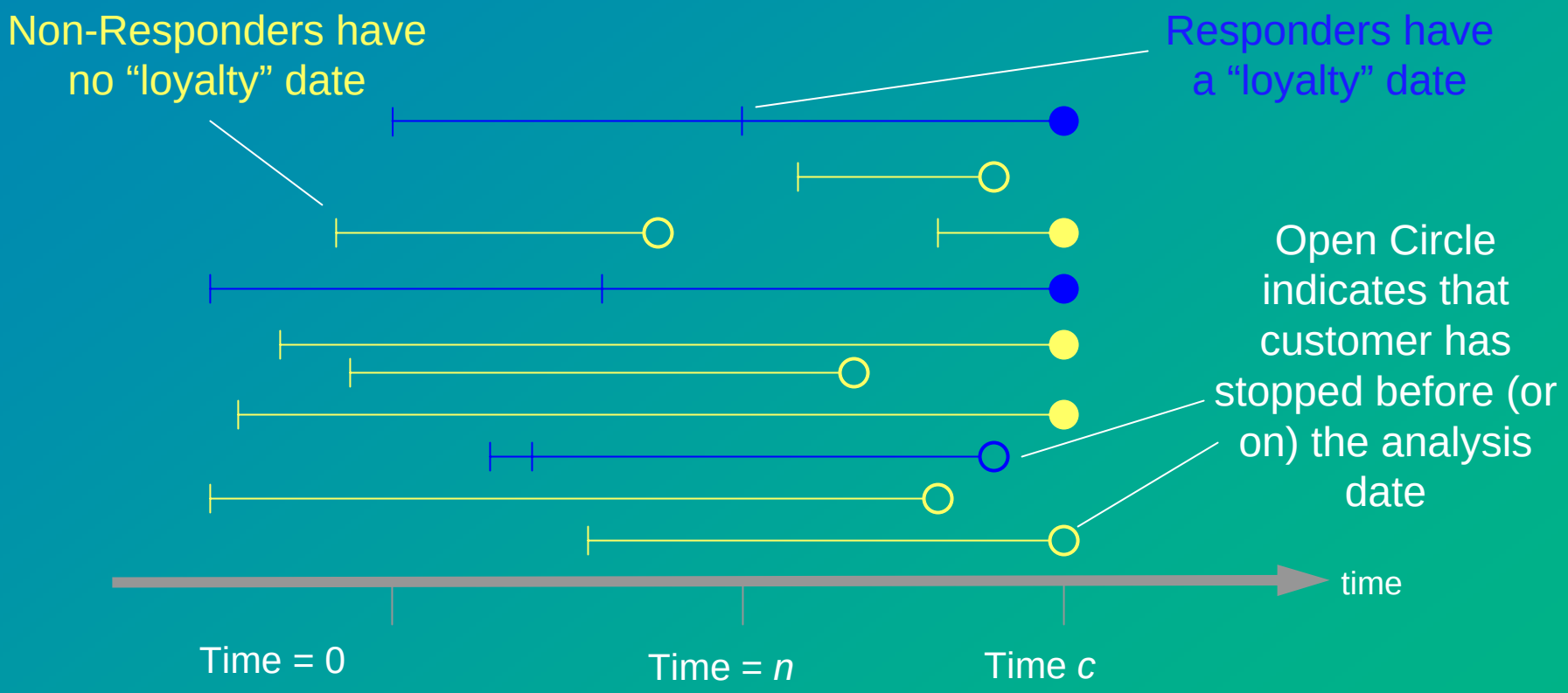
Data Miners

Survival to Quantify Marketing Efforts

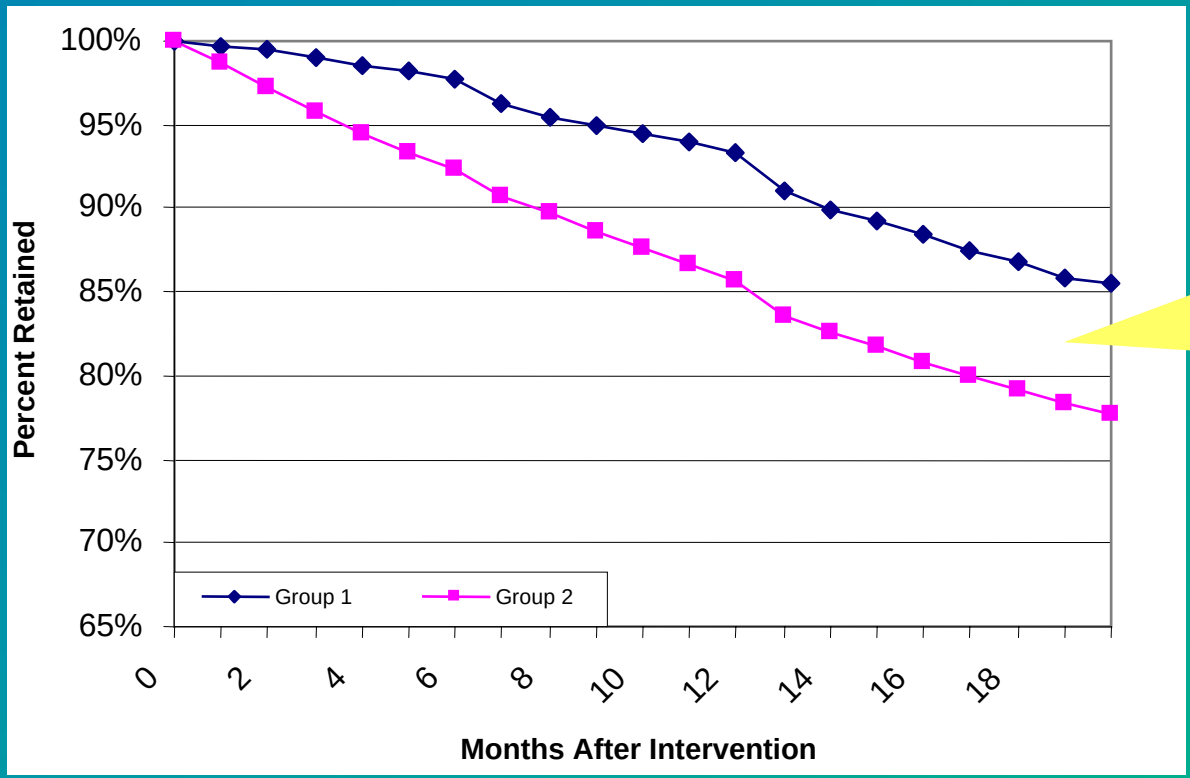
- A company has a loyalty marketing effort designed to keep customers
- This effort costs money
- What is the value of the effort as measured in increased customer lifetimes?
- SOLUTION: survival analysis
  - How much longer to customers survive after accepting the offer?
  - How to quantify this in dollars and cents?

Data Miners

Loyalty Offers is an Example of a Time-Dependent Covariate



We Can Use Area to Quantify Results

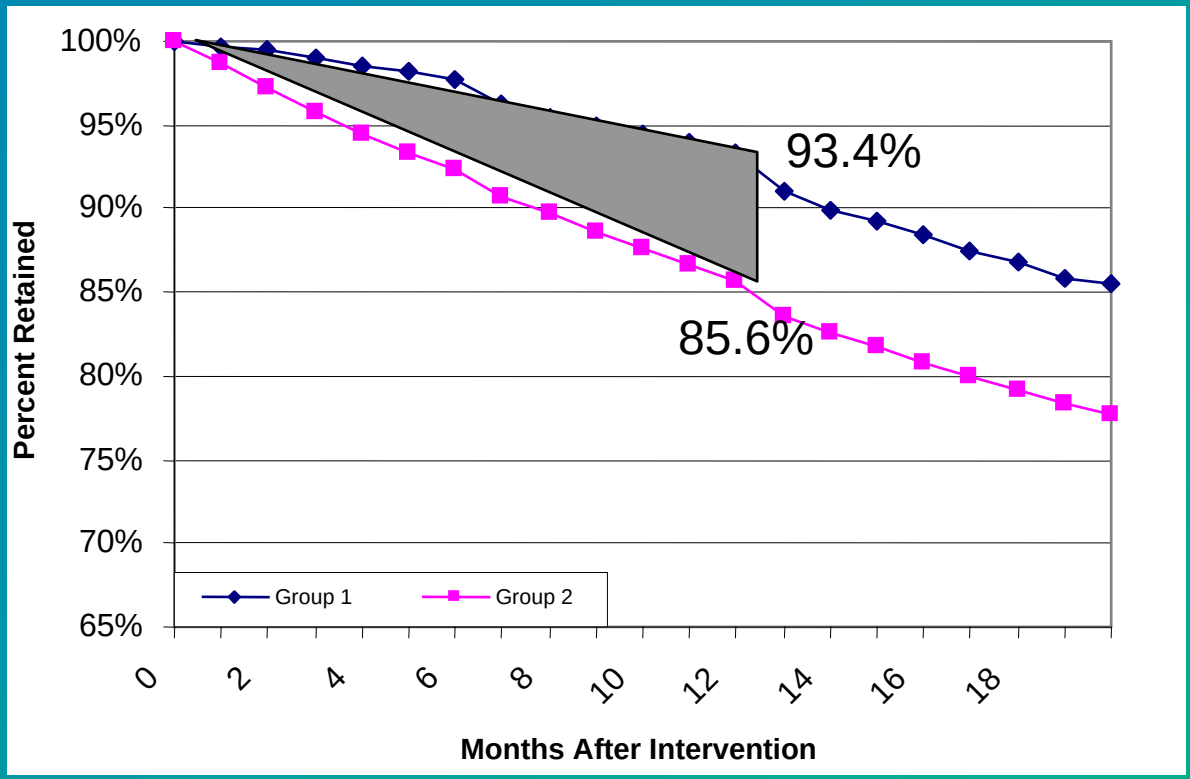


How much is this difference worth?

Chart shows survival after loyalty offer acceptance compared to similar group not given offer

Data Miners

We Can Use Area to Quantify Results



- Increase in survival is given by the area between the curves.
- For the first year, area of triangle is a good enough estimate

Note: there are easy ways to calculate the exact value

Data Miners

Loyalty-Responsive Customers Are Doing Better Than Others

- Survival for first year after loyalty intervention
  - Group 1: 93.4%
  - Group 2: 85.6%
  - Increase for Group 1: 7.8%
- Average increase in lifetime for first year is 3.9% (assuming the 7.8% would have stayed, on average, 6 months)
- Assuming revenue of $400/year, loyalty responsive contribute an additional revenue of $15.60 during the first year
- This actually compares favorably to cost of loyalty program

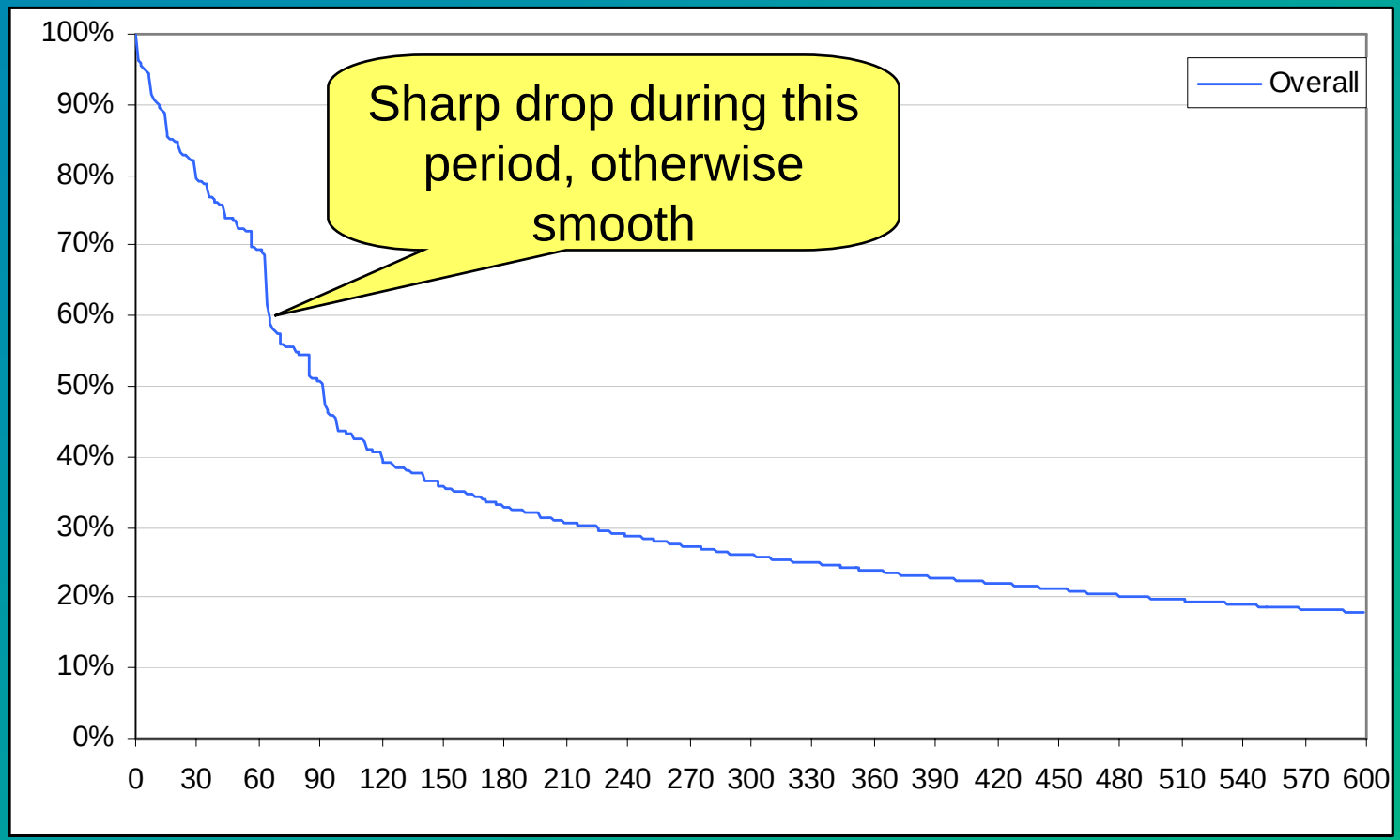
Data Miners

Competing Risks: Customers May Leave for Many Reasons

- Customers may cancel
  - Voluntarily
  - Involuntarily
  - Migration
- Survival Analysis Can Show Competing Risks
  - overall S(t) is the product of $S_r(t)$ for all risks
- We'll walk through an example
  - Overall survival
  - Competing risks survival
  - Competing risks hazards
  - Stratified competing risks survival

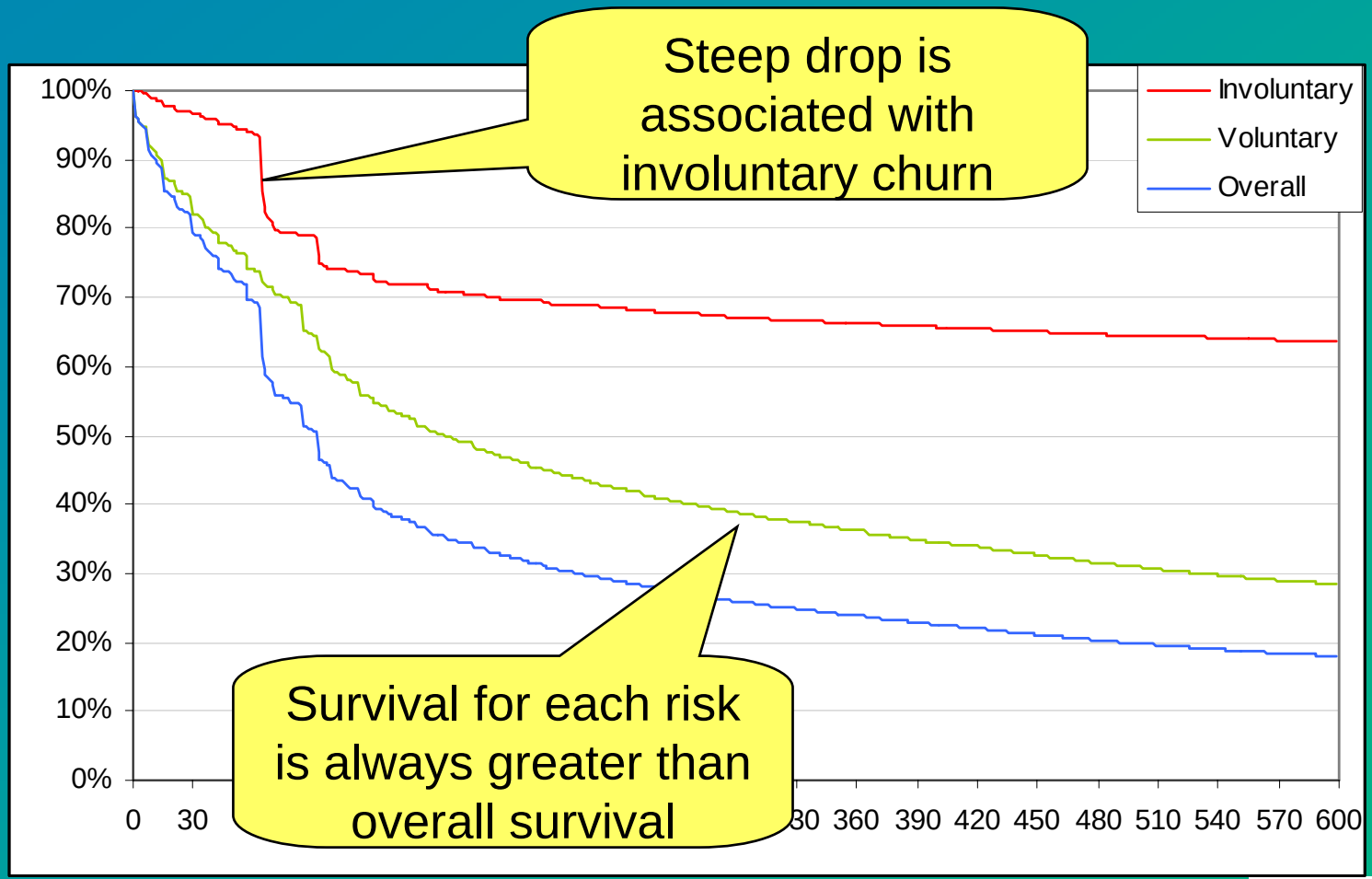
Data Miners

Overall Survival for a Group of Customers



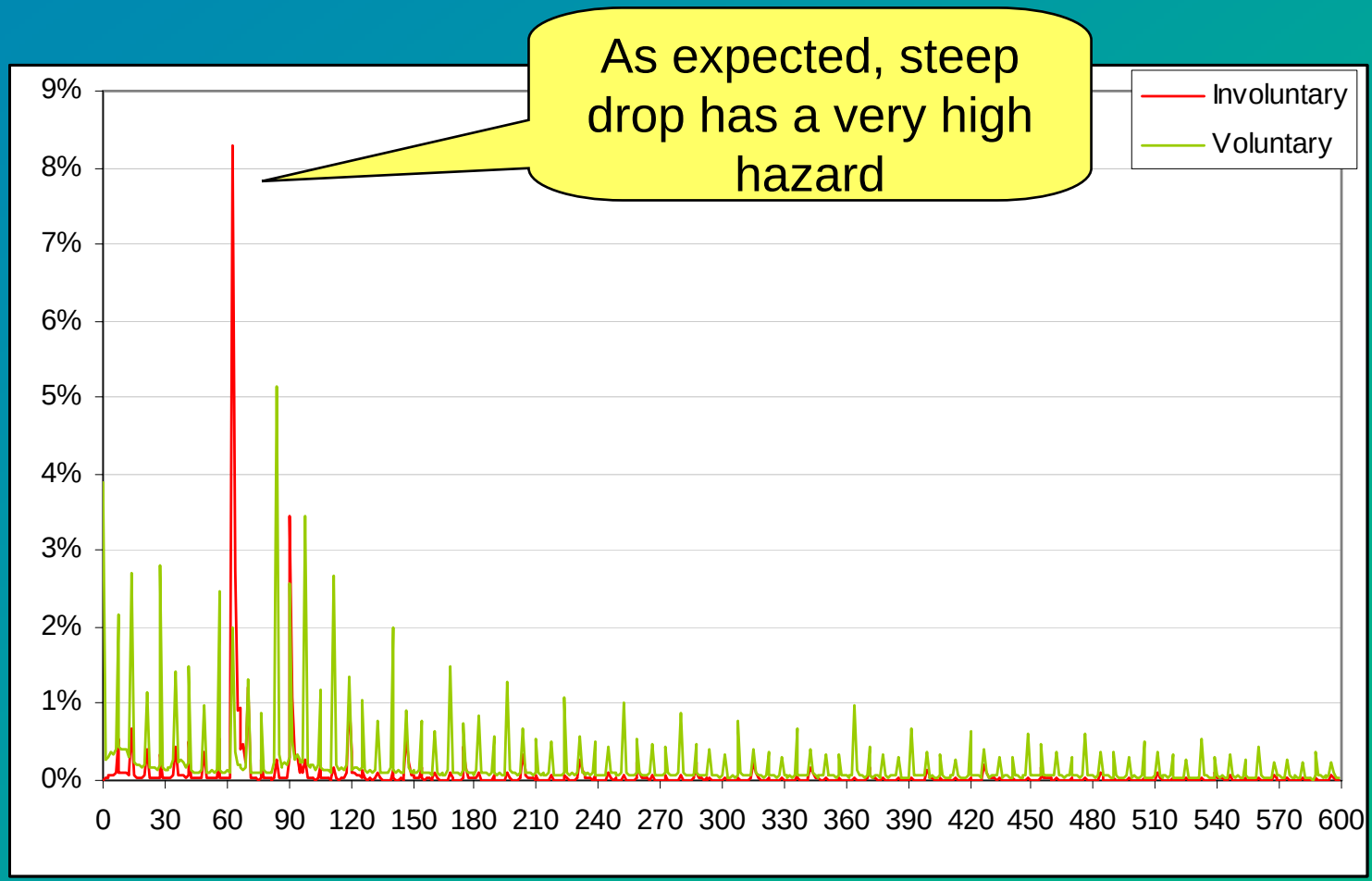
Data Miners

Competing Risks for Voluntary and Involuntary Churn



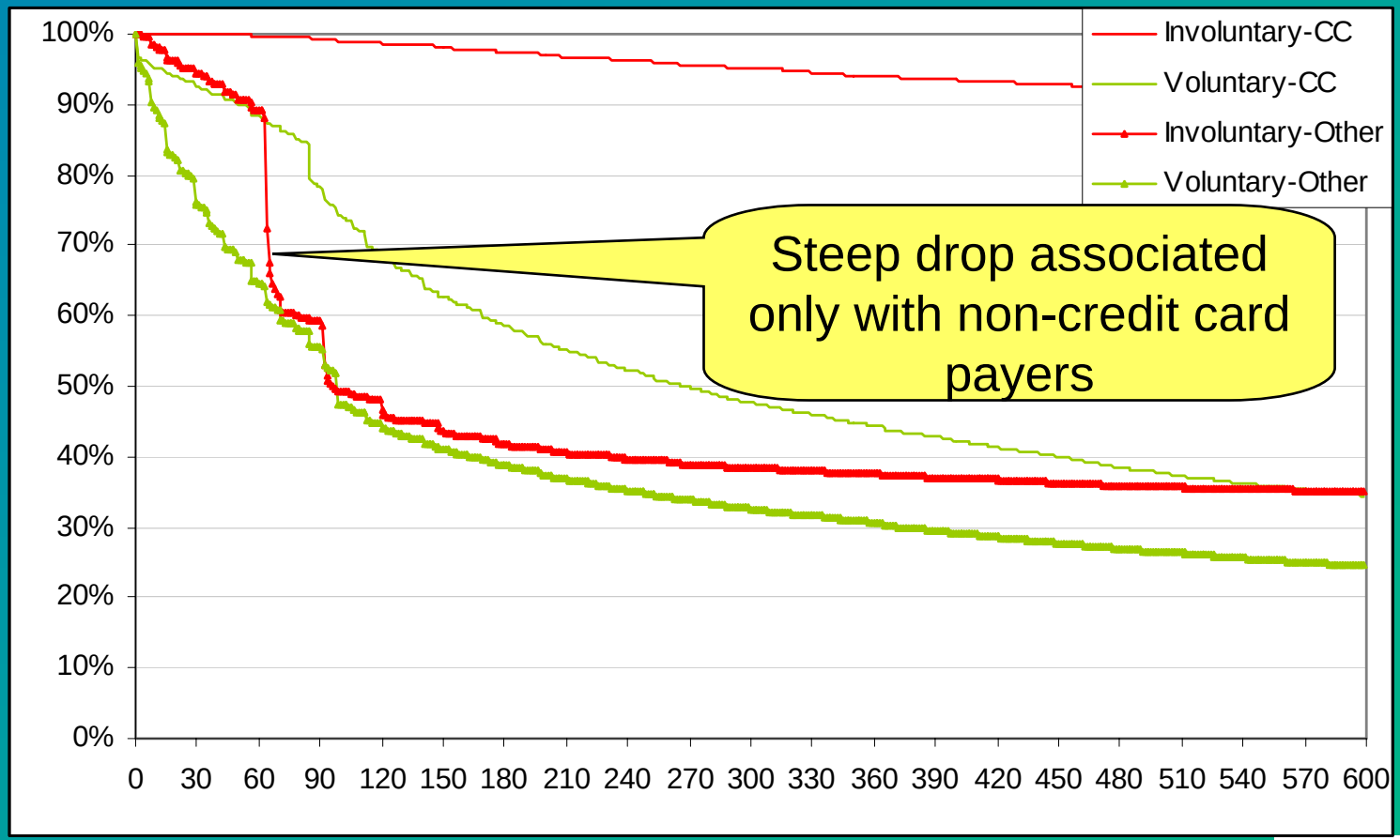
Data Miners

What the Hazards Look Like



Data Miners

Most Involuntary Churn is from Non Credit Card Payers

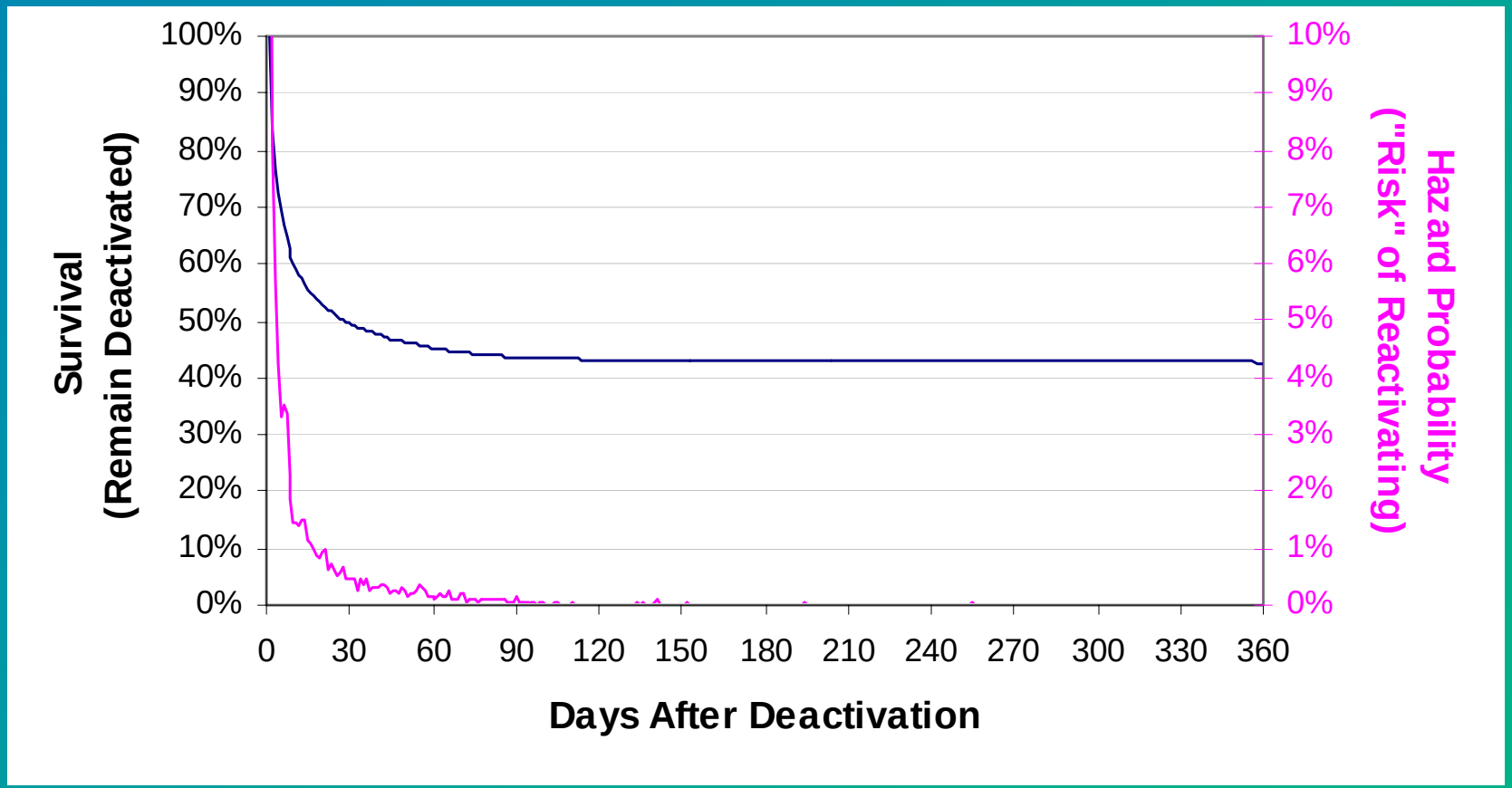


Time to Reactivation

- When a customer stops, often they come back – this is winback or reactivation
- In this case, the “initial condition” is the stop
- The “final condition” is the restart
- Not all customers restart

Data Miners

Survival Can Be Applied to Other Events: Reactivation



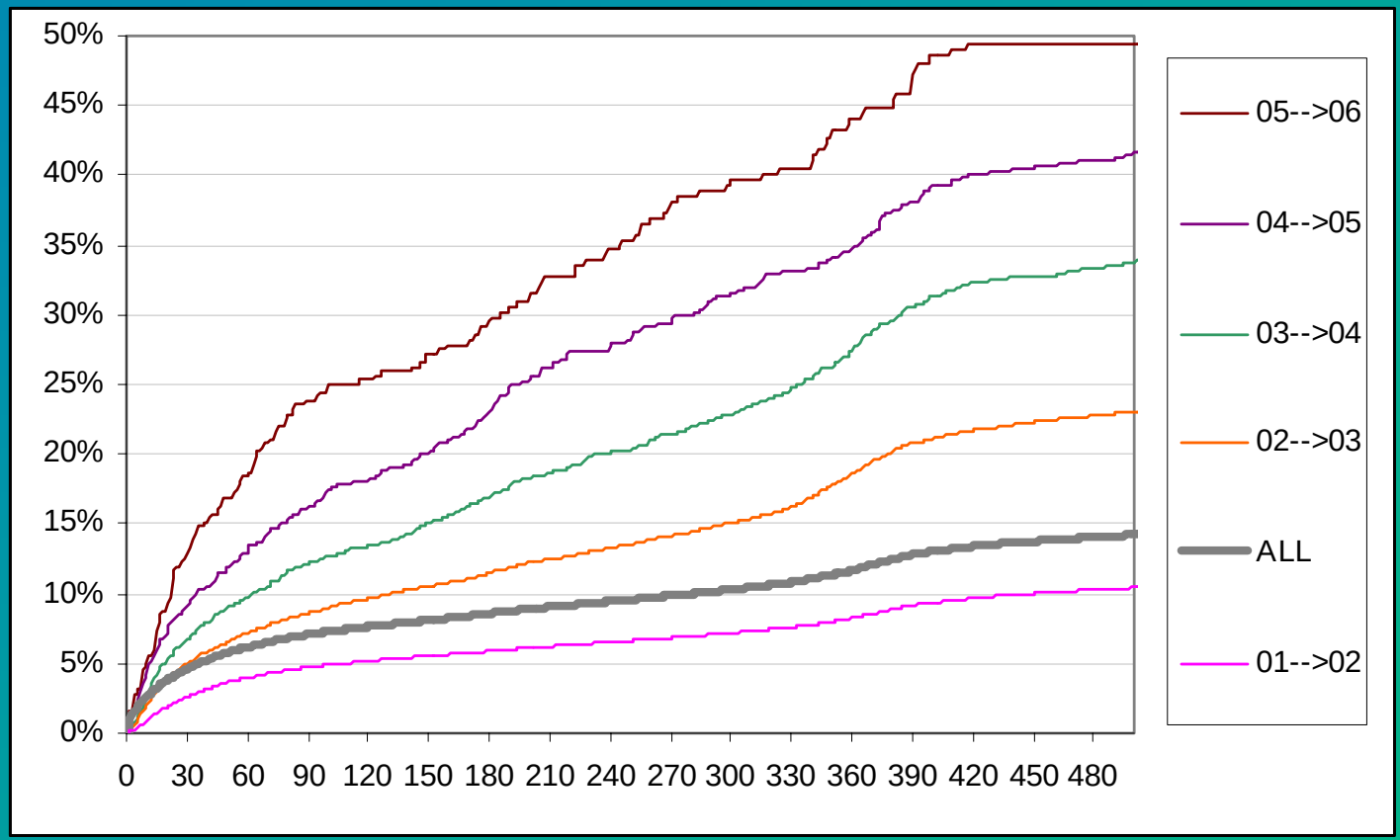
Data Miners

Time to Next Order

- In retailing type businesses, customers make multiple purchases
- Survival analysis can be applied here, too
- The question is: how long to the next purchase?
- Initial state is: date of purchase
- Final state is: date of next purchase
- It is better to look at 1 – survival rather than survival

Data Miners

Time to Next Purchase, Stratified by Number of Previous Purchases



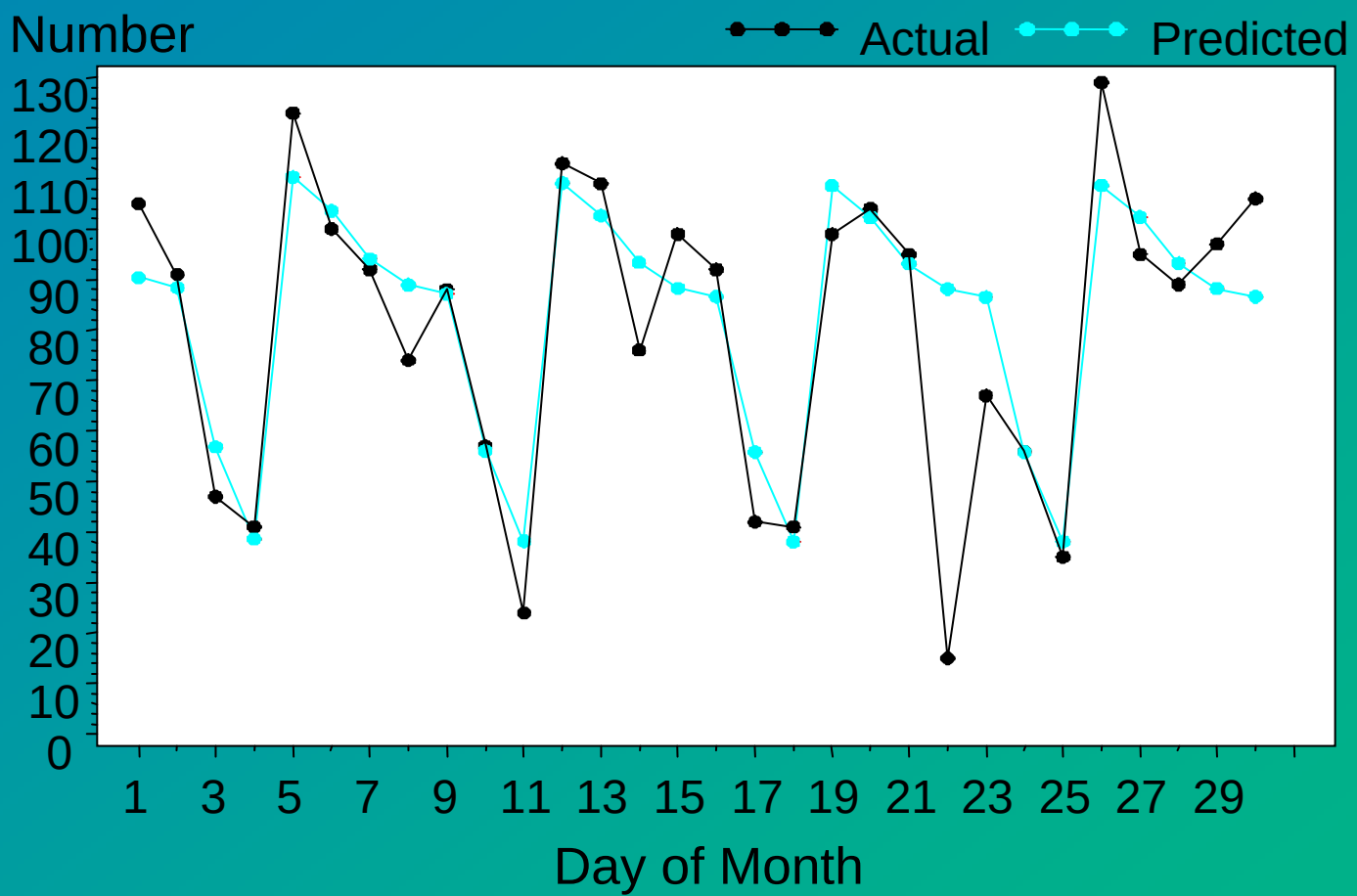
Data Miners

Customer-Centric Forecasting Using Survival Analysis

- Forecasting customer numbers is important in many businesses
- Survival analysis makes it possible to do the forecasts at the finest level of granularity – the customer level
- Survival analysis makes it possible to incorporate customer-specific factors
- Can be used to estimate restarts as well as stops

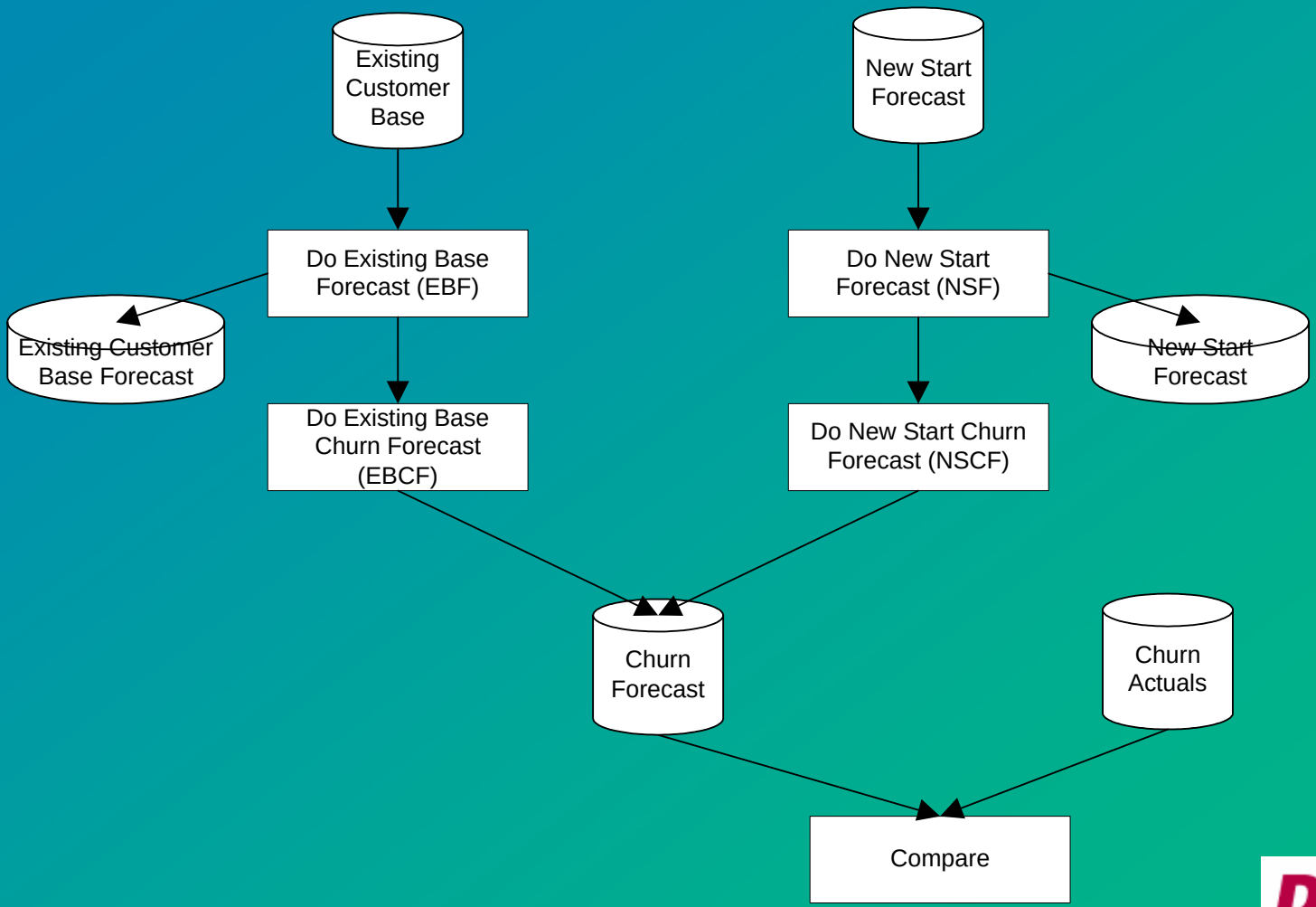
Data Miners

Using Survival for Customer Centric Forecasting



Data Miners

The Forecasting Solution is a Bit Complicated



Survival Data Mining Connects Data to Business Needs

- Provides ways to *quantitatively* measure what business users know or should know *qualitatively*
- Connects data to business practices
- Techniques such as survival analysis provide new ways of looking at customers
- Techniques such as customer-centric forecasting integrate data mining with business processes such as forecasting

Data Miners