Automation through Structured Risk Minimization

Robert Cooley, Ph.D.

VP Technical Operations

Knowledge Extraction Engines (KXEN), Inc.

Personal Motivation & Background

- **“When the solution is simple, God is answering”**
  - – Albert Einstein

  - ***1991 – Navy Nuclear Reactor Design***
    - Emphasis on simple but highly functional design

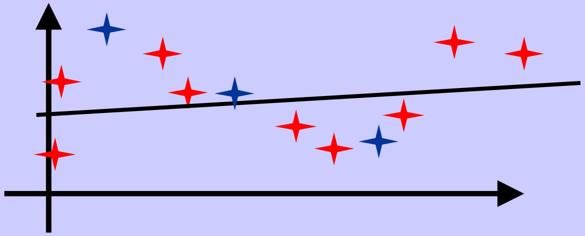
- **The key is to automate data mining as much as possible, but not more so**
  - – Gregory Piatetsky-Shapiro

  - ***1996 - Data mining research***
    - Emphasis on automation of knowledge discovery

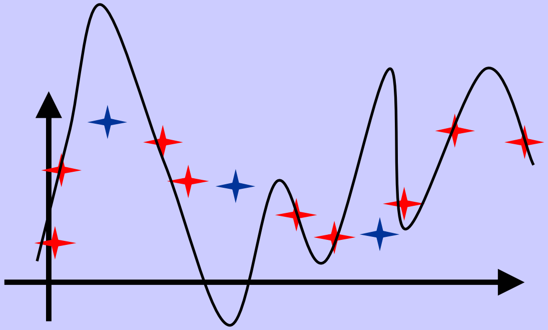
Personal Motivation & Background

- **"Solving a problem of interest, do not solve a more general problem as an intermediate step"**
  - – Vladimir Vapnik
  - ***1997 – Structured Risk Minimization (SRM)***
    - Emphasis on simplicity to manage generalization
  - ***1998 – Support Vector Machines (SVM) for Text Classification***
    - Hoping for better generalization, ended up with increased automation
- **When all you have is a hammer, every problem looks like a nail**
  - – Abraham Maslow
  - ***2001 – Joined KXEN***
    - Automation through SRM

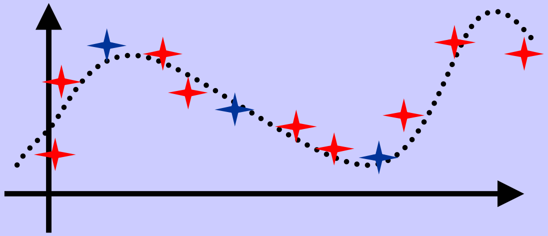
Learning / Generalization



Under Fit Model/High Robustness
(Training Error = Test Error)



Over Fit Model/Low Robustness
(No Training Error, High Test Error)



Robust Model
(Low Training Error≈Low Test Error)

What is “Robustness”?

- Statistical – Ability to generalize from a set of training examples
  - *Are the available examples sufficient?*
- Training – Ability to handle a wide range of situations
  - *Any number of variables, cardinality, missing values, etc.*
- Deployment – Ability to resist degradation under changing conditions
  - *Values not seen in training*
- Engineering – Ability to avoid catastrophic failure
  - *Return an answer without crashing under challenging conditions*

From W. of Ockham to Vapnik

If two models explain the data equally well then the simpler model is to be preferred

<u>1st discovery</u>: VC (Vapnik-Chervonenkis) dimension measures the complexity of a set of mappings.

From W. of Ockham to Vapnik

If two models explain the data equally well then the model with the smallest VC dimension is to be preferred

1st discovery: **VC (Vapnik-Chervonenkis) dimension measures the complexity of a set of mappings.**

2nd discovery: The VC dimension can be linked to generalization results (results on new data).

From W. of Ockham to Vapnik

If two models explain the data equally well then the model with the smallest VC dimension has better generalization performance

1st discovery: VC (Vapnik-Chervonenkis) dimension measures the complexity of a set of mappings.

2nd discovery: The VC dimension can be linked to generalization results (results on new data).

3rd discovery: Don't just observe differences between models, control them.

From W. of Ockham to Vapnik

To build the best model try to optimize the performance on the training data set,
AND minimize the VC dimension

SRM

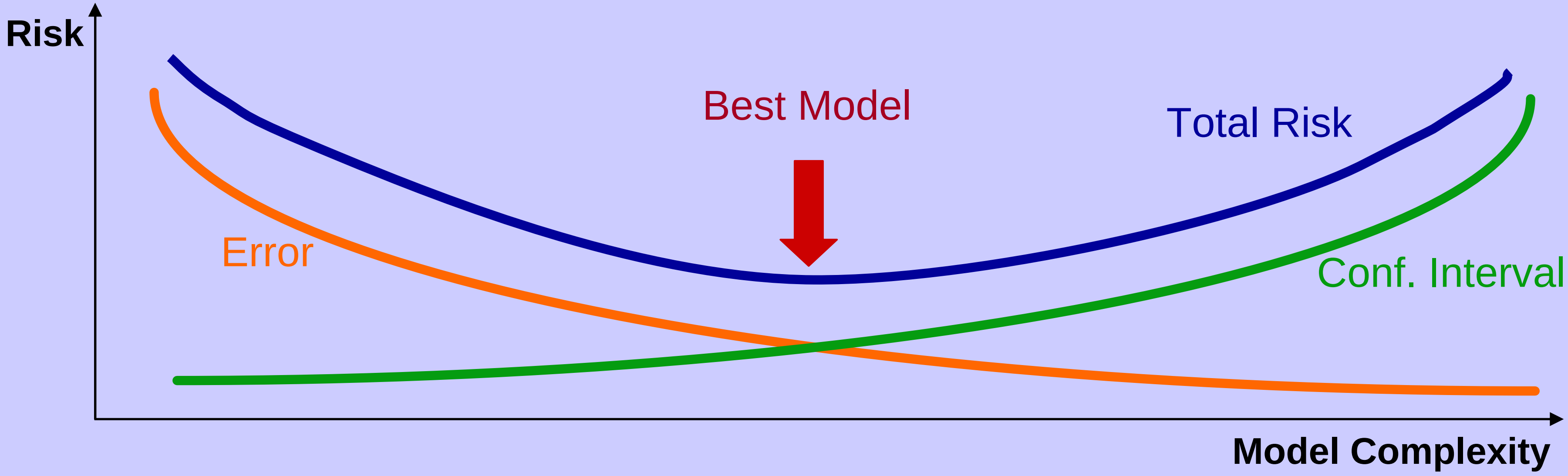
Structured Risk Minimization (SRM)

Quality:

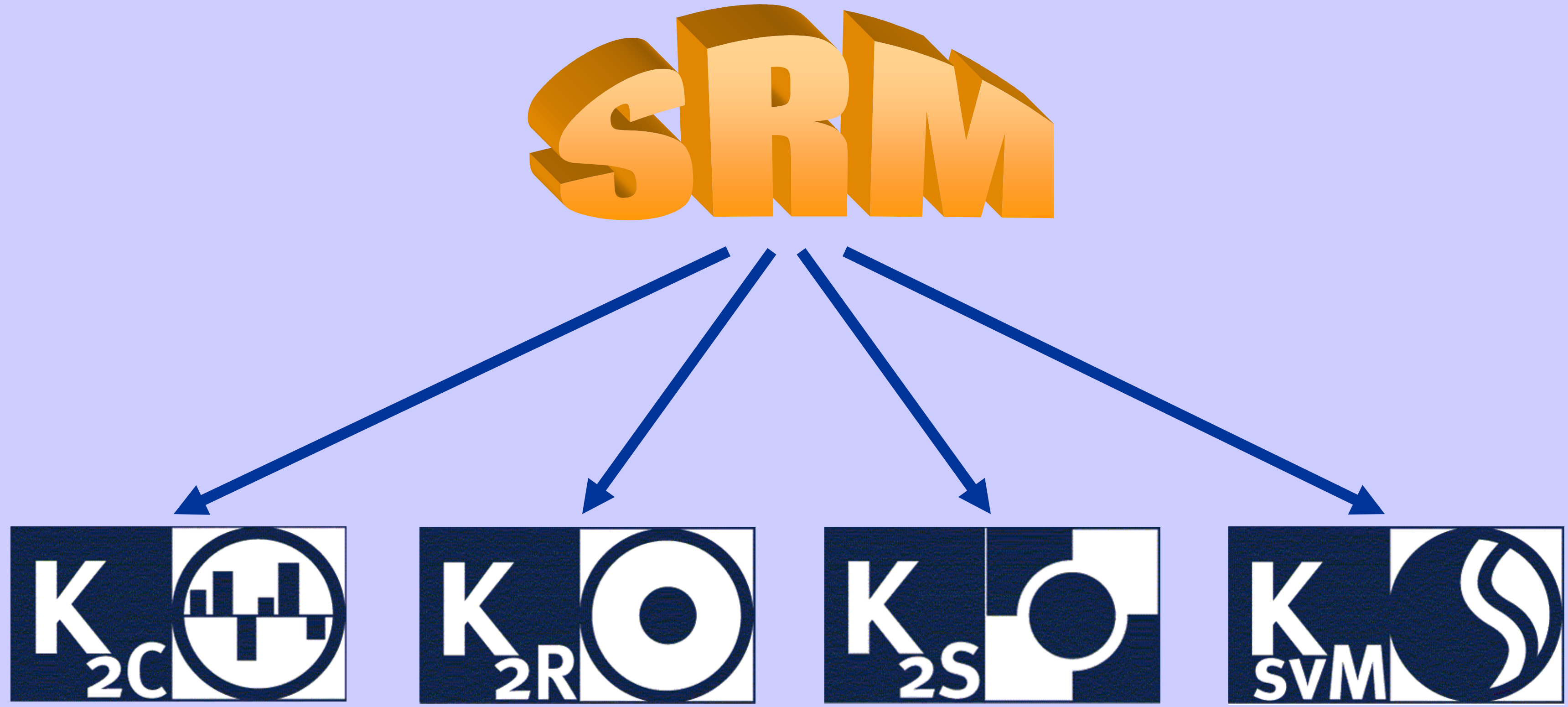
- How well does a model describe your existing data?
- Achieved by minimizing Error.

Reliability:

- How well will a model predict future data?
- Achieved by minimizing Confidence Interval.



SRM is a Principle, NOT an Algorithm



Consistent Coder

Robust Regression

Smart Segmenter

Support Vector Machine

Features of SRM

Adding Variables is Low Cost, Potentially High Benefit

- Does not cause over-fitting
- More variables can only add to the model quality
- Random variables do not harm quality or reliability
- Highly correlated variables do not harm the modeling process
- Efficient scaling with additional variables

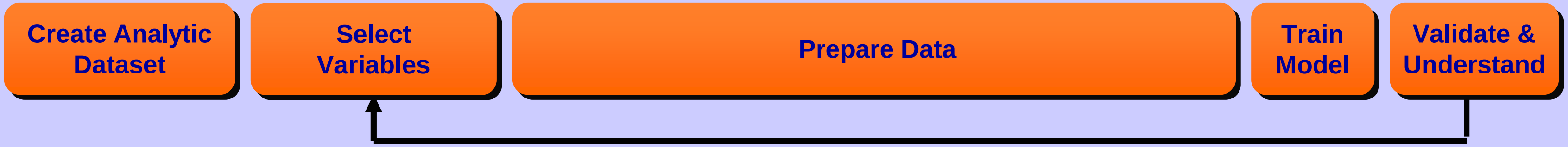
Free of Distribution Assumptions

- Normal distributions aren’t necessary
- Skewed distributions do not harm quality
- Resistant to outliers
- Independence of inputs isn’t necessary

Indicates Generalization Capacity of any Model

- Insufficient training examples to get a robust model
- Choose between several models with similar training errors

SRM enables Automation without sacrificing Quality & Reliability



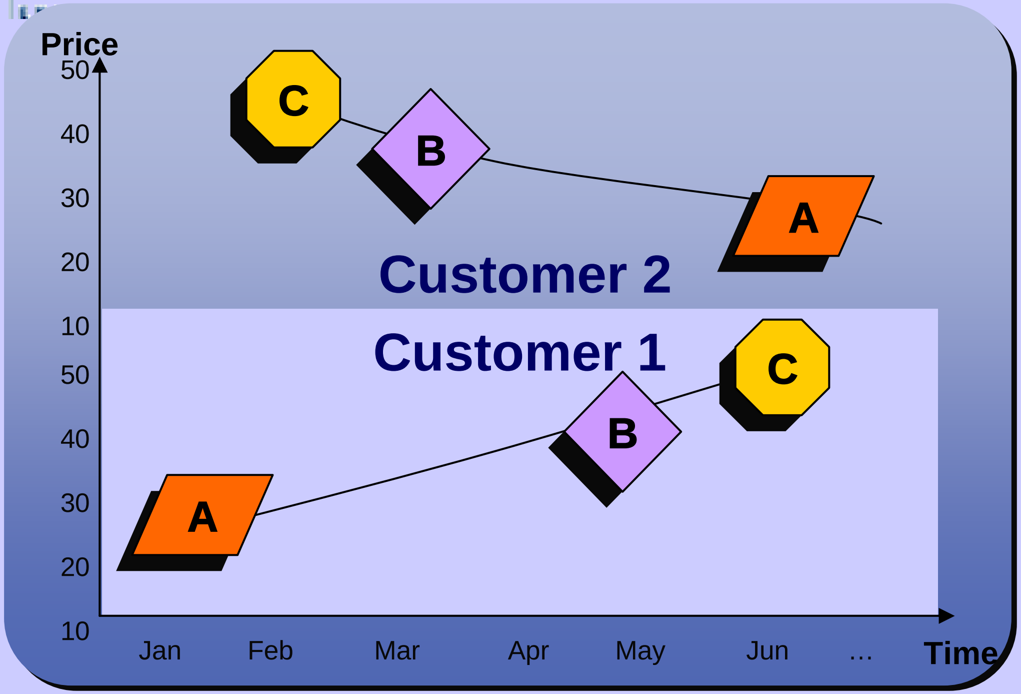
- Analytic Dataset Creation
- (Lack of) Variable Selection
- Data Preparation
- Model Selection
- Model Testing

Analytic Data Set Creation

Adding Variables is Low Cost, Potentially High Benefit

- “Kitchen Sink” philosophy for variable creation
- Don’t shoe-horn several pieces of information into a single variable (e.g. RFM)
- Break events out into several different time periods
- A single analytic data set with hundreds or thousands of columns can serve as input for hundreds of different models

Event Aggregation Example



RFM Aggregation

	RFM
Customer 1	132
Customer 2	132

Simple Aggregation

	A	B	C	Total purchase
Customer 1	1	1	1	105
Customer 2	1	1	1	105

Relative Aggregation

Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
$45	$35	0	0	$25	0
$25	0	0	0	$35	$45

Transitions

A → out	C → B	B → A	C → out	B → C	A → B
1	1	1	0	0	0
0	0	0	1	1	1

Data Preparation

Free of Distribution Assumptions

- Different encodings of the same information generally lead to the same predictive quality
- No need to check for co-linearities
- No need to check for random variables
- SRM can be used within the data preparation process to automate binning of values

Binning

- **Nominal – Group values**
  - ***TV, Chair, VCR, Lamp, Sofa → [TV,VCR], [Chair, Sofa], [Lamp]***
- **Ordinal – Group values, preserving order**
  - ***1, 2, 3, 4, 5 → [1, 2, 3], [4, 5]***
- **Continuous – Group values into ranges**
  - ***1.5, 2.6, 5.2, 12.0, 13.1 → [1.5 – 5.2], [12.0 – 13.1]***

Why Bin Variables?

- **Performance**
  - ***Captures non-linear behavior of continuous variables***
  - ***Minimizes impact of outliers***
  - ***Removes “noise” from large numbers of distinct values***
- **Explainability**
  - ***Grouped values are easier to display and understand***
- **Speed**
  - ***Predictive algorithms get faster as the number of distinct values decreases***

Binning Strategies

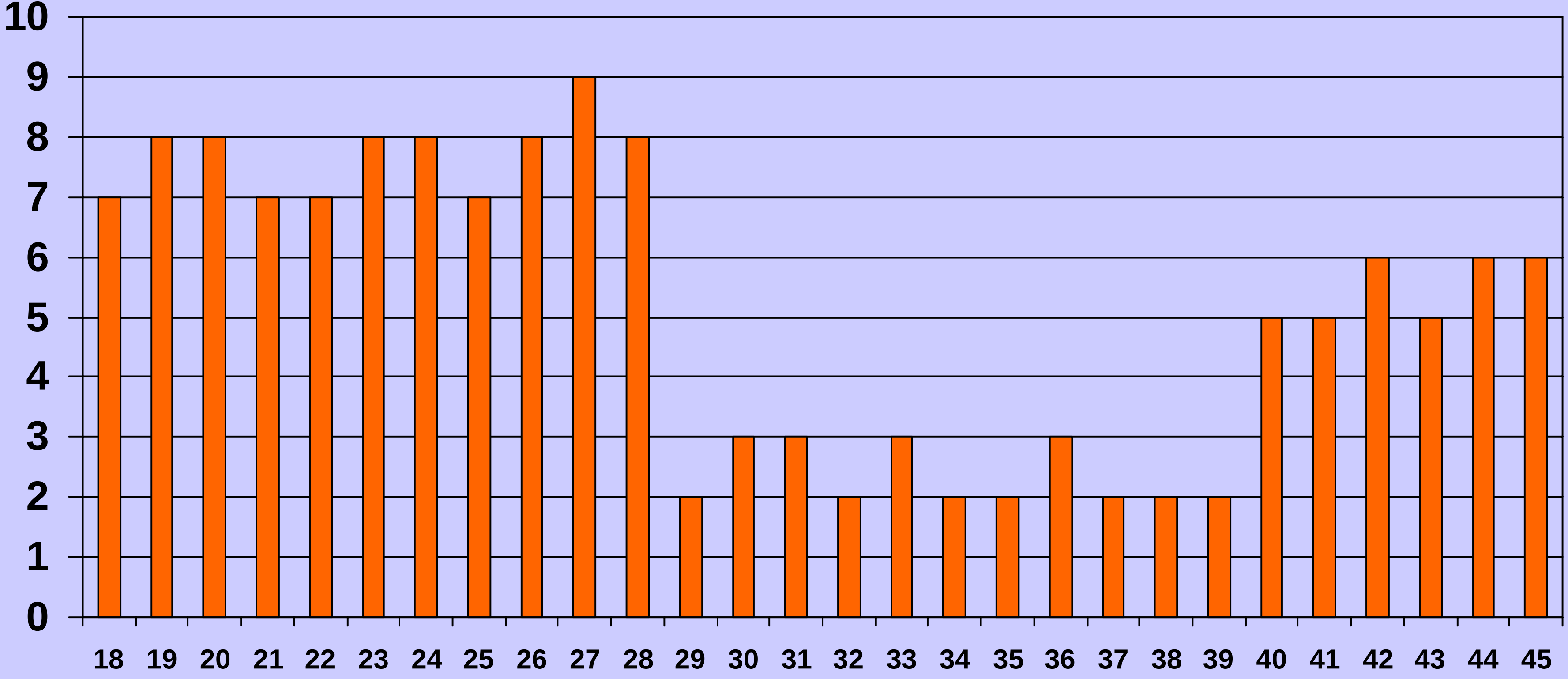
- **None – Leave each distinct value as a separate input**

- **Standardized – Group values according to preset ranges**
  - ***Age: [10 – 19], [20 – 29], [30 – 39], …***
  - ***Zip Code: [94100 – 94199], [94200 – 94299], …***

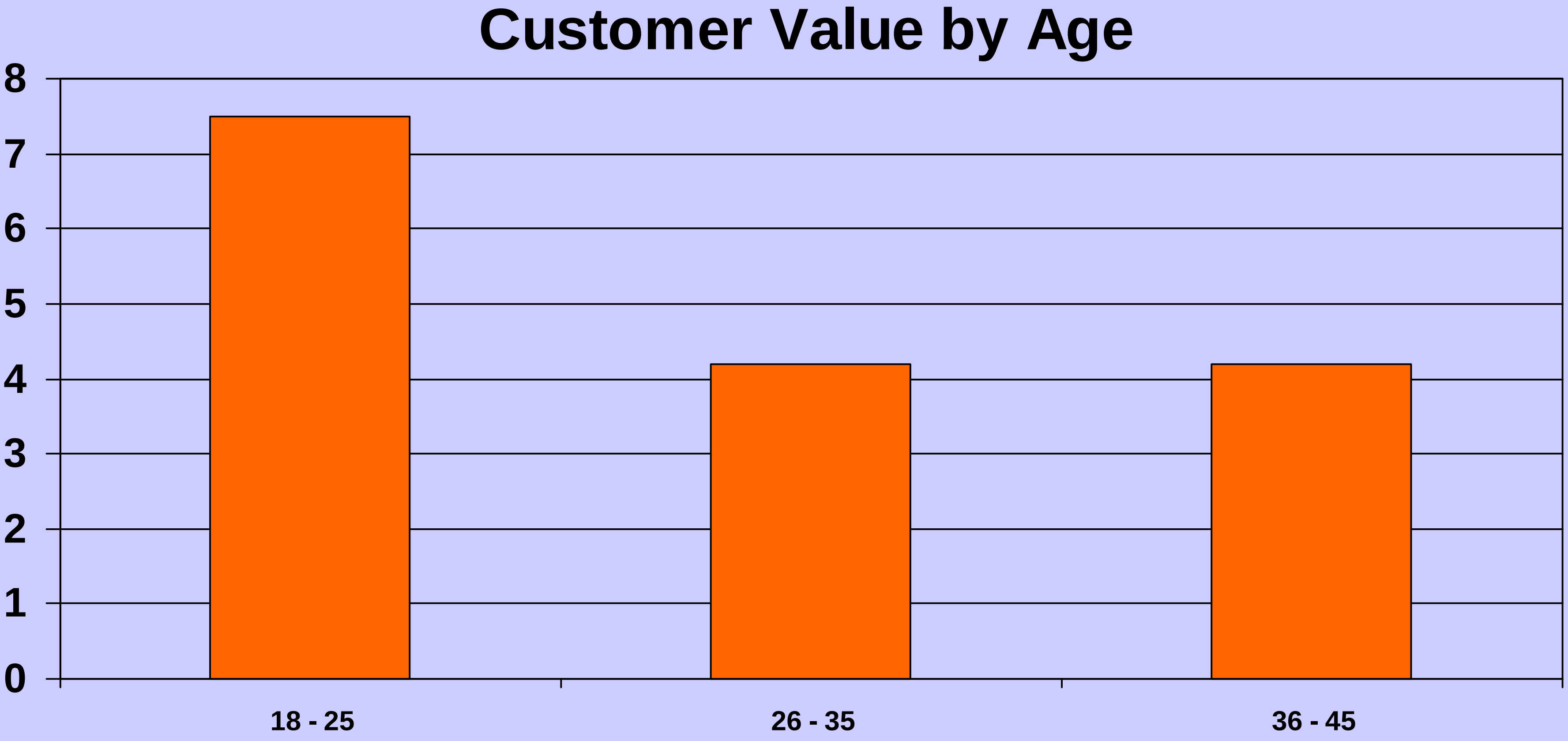
- **Target Based – Use advanced techniques to find the “right” bins for a given question. DEPENDS ON THE QUESTION!**

No Binning

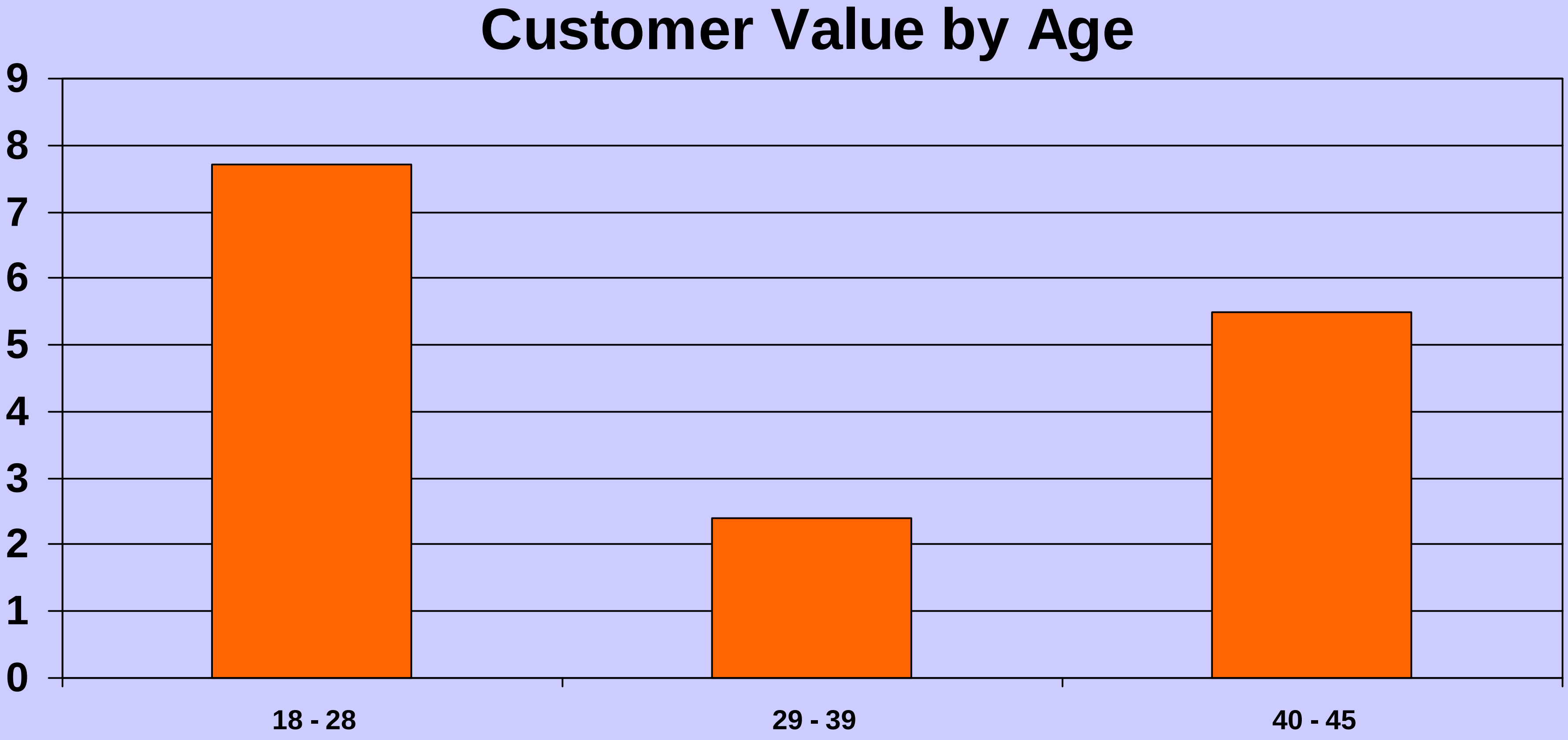
Customer Value by Age



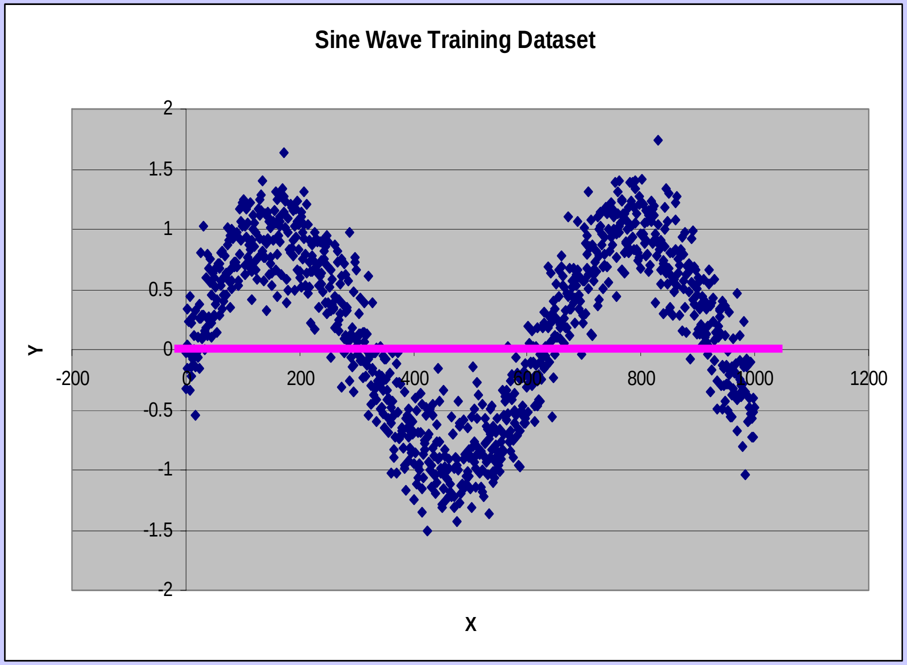
Standardized Binning

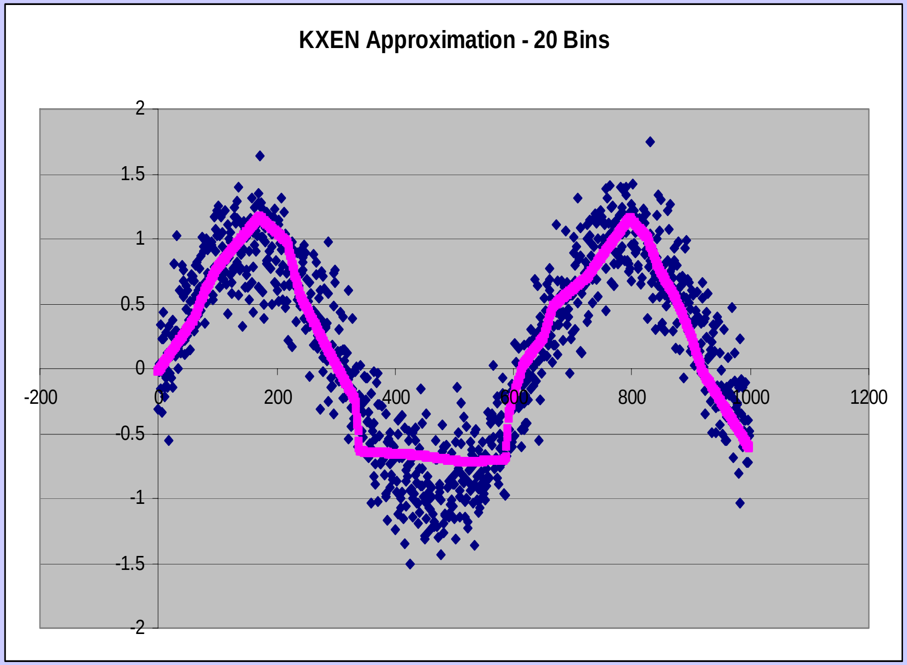


Target Based Binning



Predictive Benefit of Target Based Binning





Model Testing

Indicates Generalization Capacity of any Model

- It is not always possible to get a robust model from a limited set of data
- Determine if a given number of training examples is sufficient
- Determine amount of model degradation over time

Traditional Modeling Methodology

- CRISP-DM
  - *Business Understanding*
  - *Data Understanding*
  - *Data Preparation*
  - *Modeling*
  - *Evaluation*
  - *Deployment*

- SEMMA
  - *Sample*
  - *Explore*
  - *Modify*
  - *Model*
  - *Assess*

SRM Modeling Methodology

1. Business Understanding
2. Create Analytic Dataset
3. Modeling
4. Data Understanding
5. Deployment
6. Maintenance

Conclusions

- SRM is a general principle that can be applied to all phases of data mining

- The main benefit of applying SRM is increased automation, not increased predictive quality

- In order to take full advantage of the benefits, the overall modeling methodology must be modified

Thank You

Questions?

Contact Information:

rob.cooley@kxen.com

651-338-3880

Additional KXEN Information:

- **www.kxen.com**
- **sales-us@kxen.com**
- **5/25 NYC Discovery Workshop, 9am to 1pm**