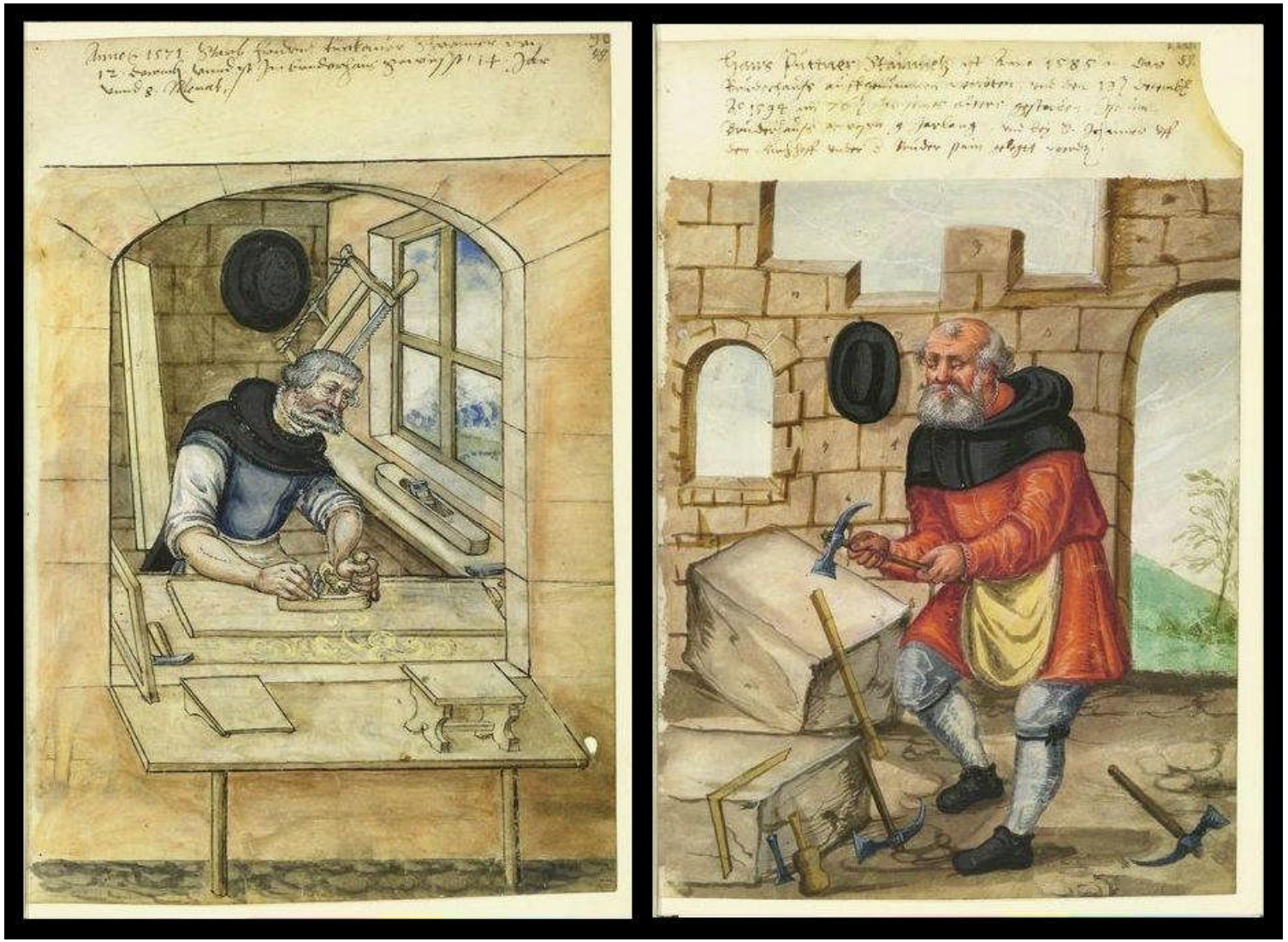
The Craft of Credit Scoring:
Data Mining Applied to Retail Finance

Ross Gayler

The applied data miner as craftsperson



Reused from http://www.flickr.com/photos/bibliodyssey/3085763753/sizes/o/in/photostream/ with permission of peacay

The applied data miner as craftsperson

The applied setting (at least in credit scoring) creates certain emphases:

- The importance of expertise
- The importance of pragmatics
- The environment as bottleneck

Leading to:

- The (relative) unimportance of new methods
- The (relative) importance of the analyst

Where and how does the analyst add value?

What is credit scoring?

What is credit scoring?

- Predictive modelling of operational outcomes in mass-market credit
- Used to automate operational decision making
  - Make decisions on the basis of predicted outcomes

Why use credit scoring?

- Better decision making
  - More accurate predictions than subjective judgment
- Automation
  - Cost saving
  - Faster service
- Objectivity
- Repeatability
- Controllability

Potential value of scoring

Consider credit card application processing for a large Australian lender (order of magnitude estimate)

- Total new exposure
  - 1k applications / day
  - 70% approval
  - $5k average credit limit
  - ~ $900M p.a. new exposure
- Total credit loss
  - 2.5% loss rate
  - ~ $23M p.a. loss
- Value of improvement (relative to subjective)
  - 15% decrease in loss ~ $3.4M p.a.

Total consumer debt exceeds commercial debt

History of credit scoring

The first scorecards

Terminology: scorecard = predictive model

- First used in 1946
- Lightly used in the 1950s and early 1960s
- Took off with credit cards and computing in the late 1960s and early 1970s
- First applied to application processing
  - Predictors from application form and credit bureau
  - Predicted outcome is failure to repay
  - Decision is accept/reject the application

Constraints on early scorecards

- Minimal computing power available to build models
  - Simple modelling techniques
    - Linear regression or discriminant analysis
- Models applied manually
  - Simple functional form required (scorecards)

Example application scorecard

Time at Job (yrs)	< 0.5 **5**	0.5 to 1.4 **14**	1.5 to 6.4 **20**	6.5 to 10.5 **27**	> 10.5 **39**	
Time at Address (yrs)	< 1 **-11**	1 to 2.4 **0**	2.5 to 6 **8**	> 6 **17**		
Residential Status	Own or Buy **40**	Rent **19**	Other **26**			
Occupation	Retired **41**	Professional **36**	Clerical **27**	Sales **18**	Service **12**	Other **27**
Age of Applicant (yrs)	18 to 25 **19**	26 to 31 **14**	32 to 34 **22**	35 to 51 **26**	52 to 61 **34**	> 61 **40**
Bureau Enquiries	No record **-12**	1 to 2 **0**	3 to 5 **-7**	> 5 **-32**		
Bureau Defaults	No record **0**	0 **0**	1 **-57**	>1 **-126**		

Adapted from E.M. Lewis (1992) "An introduction to credit scoring"

Later scorecards

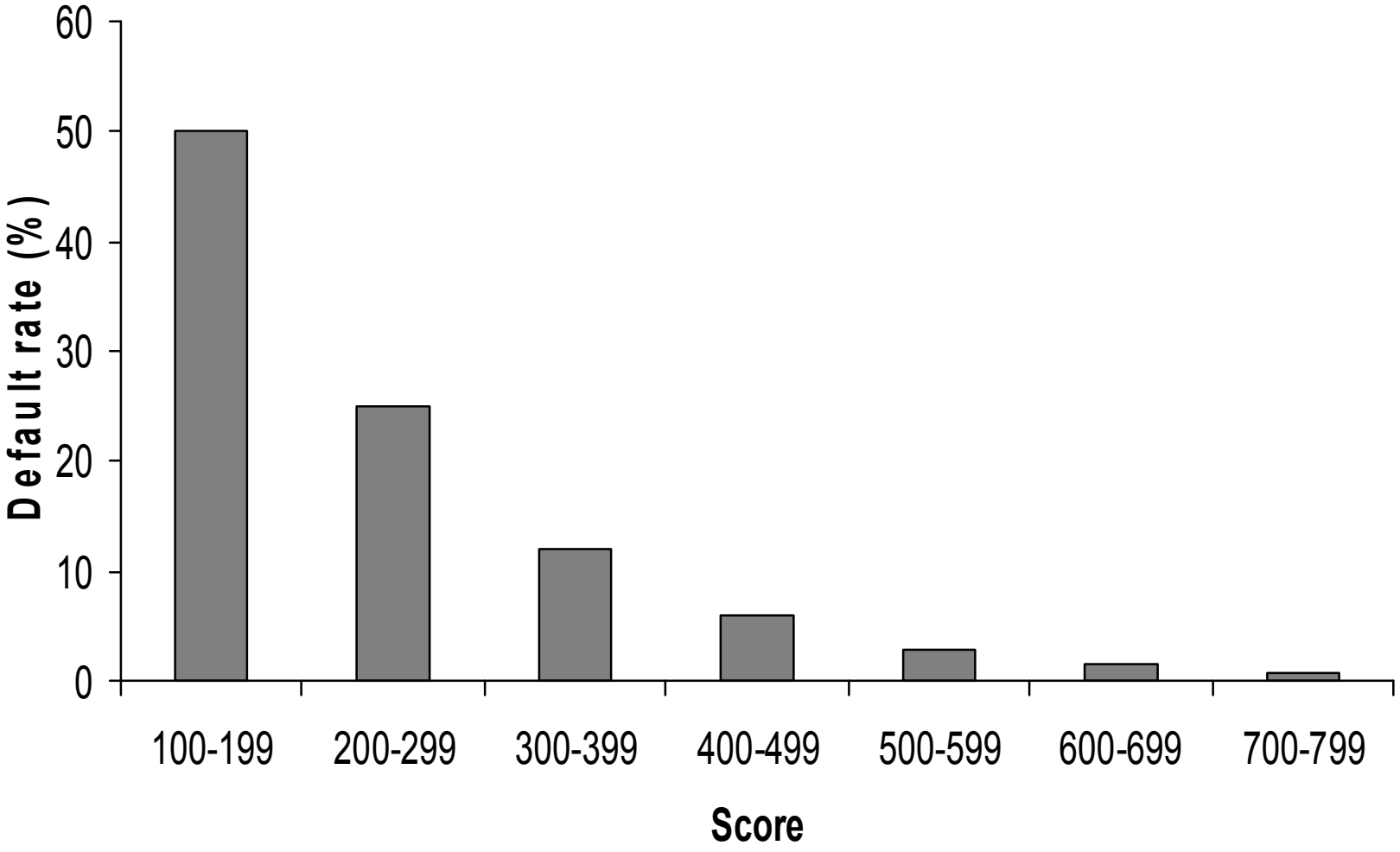
- Predict different outcomes
  - Response to mailed offer (marketing)
  - Account closure (account management)
- Use different predictors
  - Geodemographics
  - Account transaction history

- Modelling method has not changed much

  Now mostly logistic regression

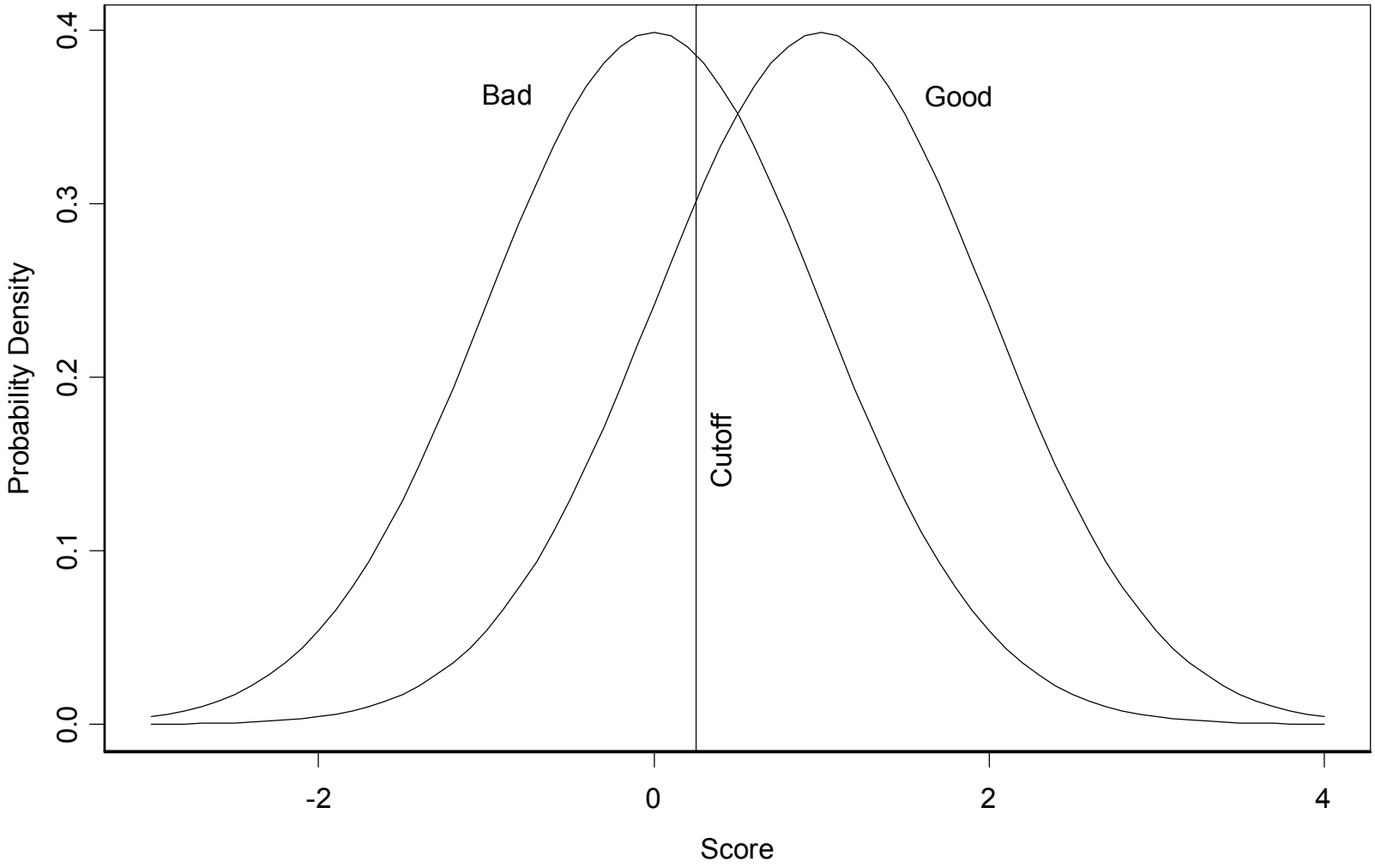
How well does scoring work?

Relationship between score and outcome



Distributions of application scores

d’ = separation (in standard deviations)



How well does scoring work?

- More predictive than subjective estimates for application processing
  - Typically 15% - 25% lower default rate at the same accept rate
- Typical predictive performance
  - Application scoring (demographic predictors)
    - Extreme odds ratio ~ 50:1
    - AUC ~ 0.75
    - d' ~ 1 sd
  - Behaviour scoring (transactional predictors)
    - Extreme odds ratio ~ 200:1
    - AUC ~ 0.85
    - d' ~ 1.5 sd

An opportunity for new methods?

Argument for sophisticated data mining

- Credit scoring is simple and modestly predictive
- Why not use:
  - Neural networks
  - Random forests
  - Support vector machines
  - …
- Academic research has concentrated on more sophisticated modelling techniques
  - Rarely adopted in practice in credit scoring

Why credit scoring hasn't changed much

Possible reasons for lack of change:

- Methodological inertia
  - People aren't trained in new methods
- Systems inertia
  - Implementation is slow and expensive
- Lender conservatism
  - Transparency of methods for confidence
  - Regulatory requirements
  - Management expertise

- Mostly - lack of benefit

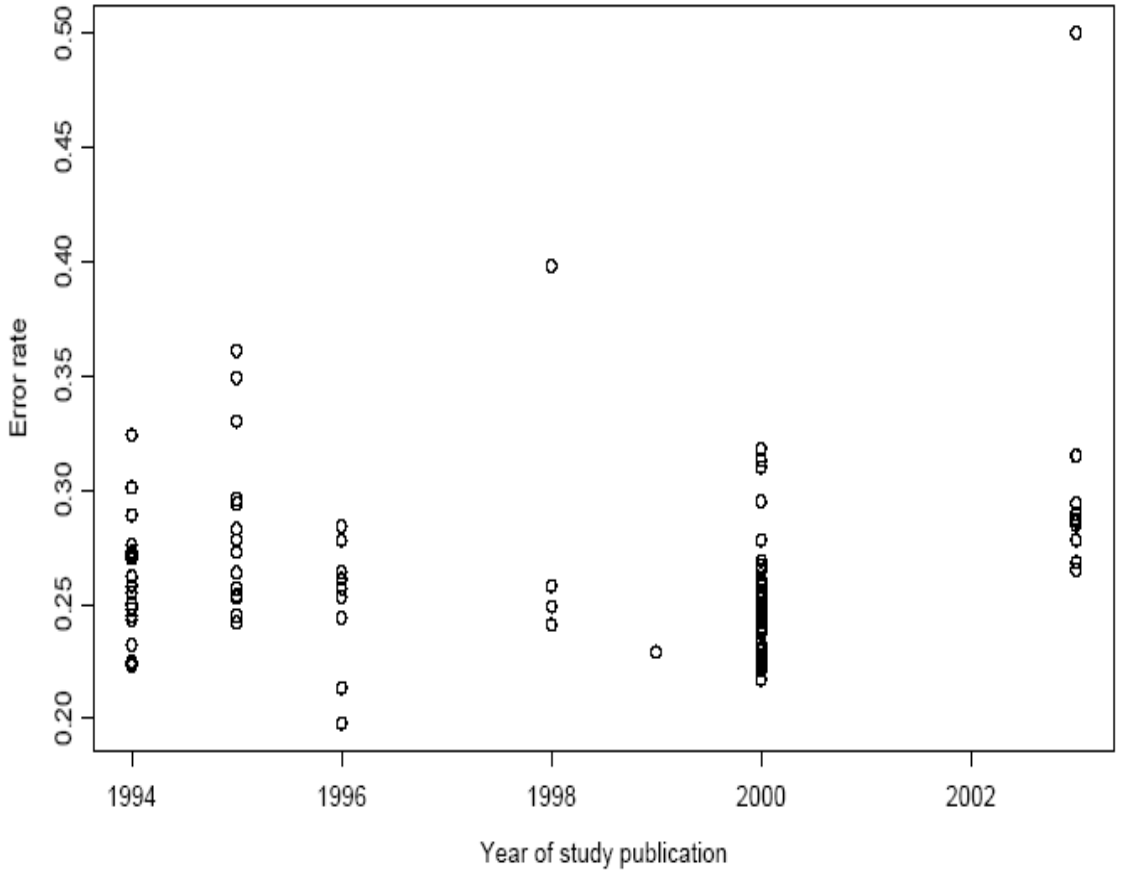
Is data mining becoming more predictive?
(Generally? In credit scoring?)

No improvement over 10 years

Error rates vs paper publication date

Pima Indian data (UCI ML repository)

Hand (2003) Banff Credit Scoring Workshop



Theoretical benefit of new methods likely to be dominated by other issues

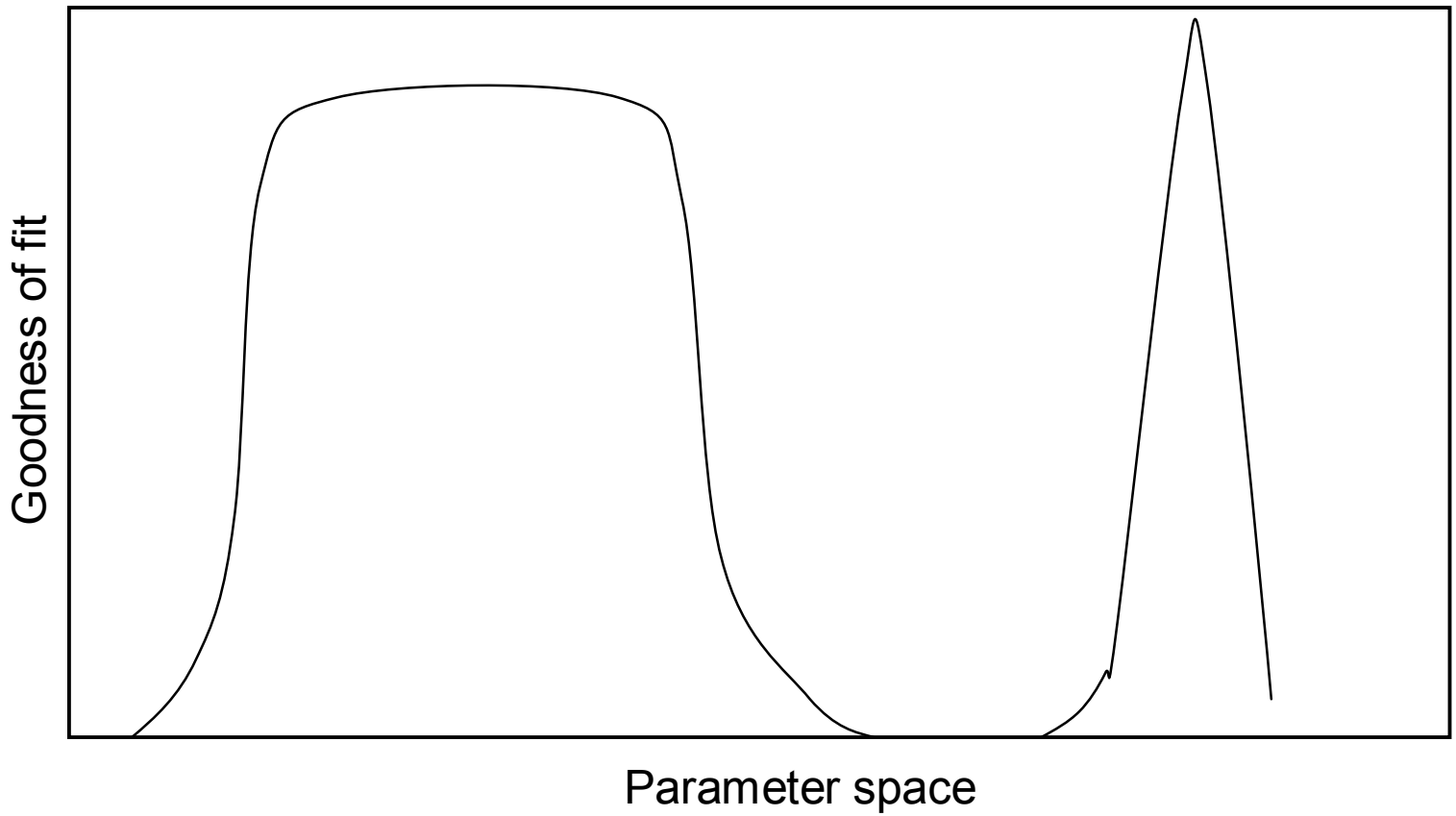
Issues

- Functional form of scorecards
  - Standard regression is not really inflexible
  - Every surface looks flat with enough noise
  - Diminishing returns in incremental models
  - Flat maximum effect is valuable
- Data issues
  - Scoring models are not scientific models
    - Not causal models
    - Population drift & jump
  - Poor quality data (modelling the defects)
  - Arbitrary outcome definitions

Flat maximum effect gives flexibility

What is the flat maximum effect?

- Distribution of goodness of fit in model parameter space
- Many approximately equivalent models



Properties of flat maximum effect

- Arises from
  - Additive form of standard regression
  - Conditional monotone relationship between predictors and outcome
- Benefits
  - Models don't fail suddenly
  - Models can be cross-applied
  - Models can be selected on a basis other than goodness of fit

Assumptions of predictive modelling

Basic assumptions of predictive modelling

- Similar cases behave similarly
- The future is like the past
- In practice, your data are the past

Caveats

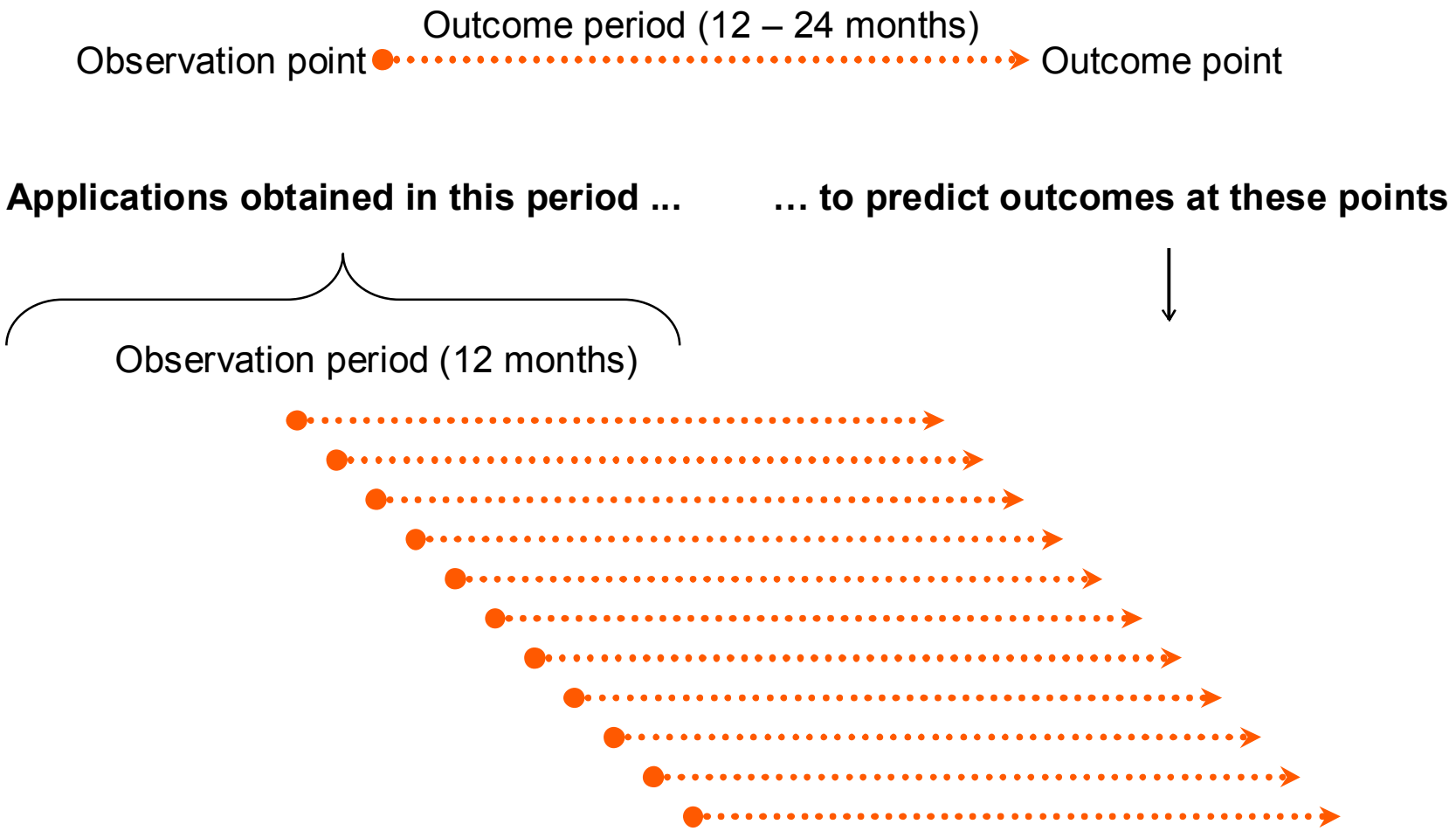
- Available data does not address the (stable) causal mechanisms of customer behaviour
- Important predictors are not available
- Important predictors do not vary in the data
- Data is always out of date

The future is not like the past

- New types of customers
- New systems
- New operational procedures
- New competitors
- …

Time frame of modelling

Model based on data 2 ~ 3 years old

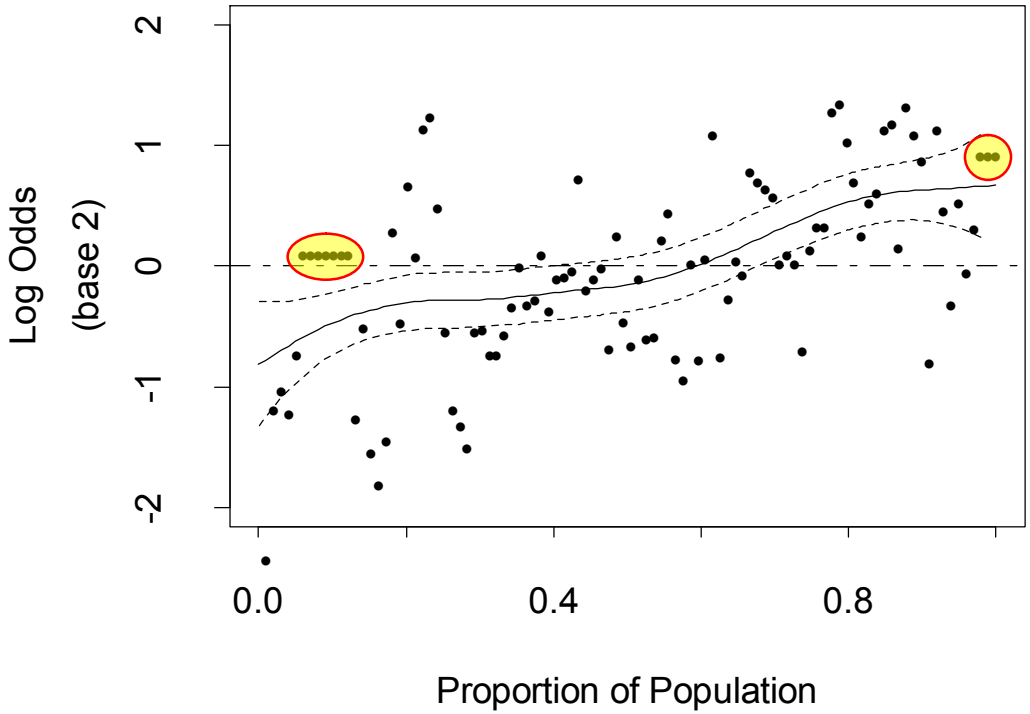


The truth is in the data?

- Poor data quality
  - The data are from operational systems not designed to support statistical data collection
- Data not representative of the future
  - Introducing a new credit product, so no data available for that specific credit product
  - The world will change after data collection
- Puts a premium on being able to understand the models
- Puts a premium on being able to subjectively modify the models

Operational changes

Portfolio acquisitions

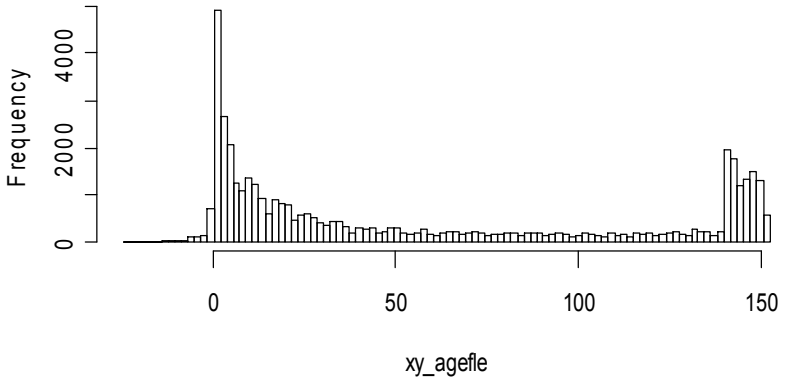


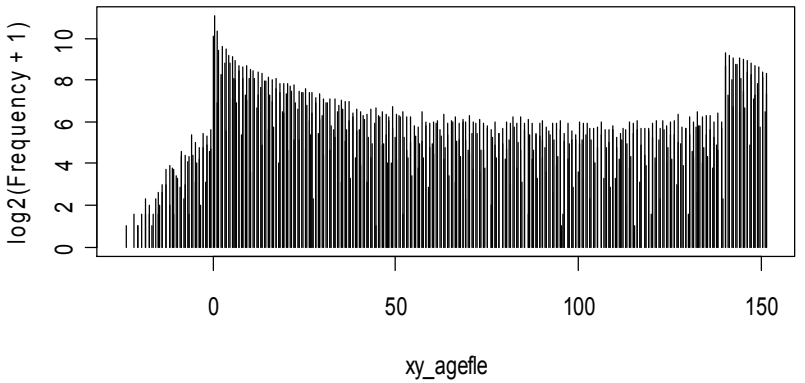
- Small off-trend groups at 5th & 97th percentiles
- Identified as cases from an acquired portfolio
- Exclude them?
- Correct for them?

System changes

System changes

- Encoding of dates was changed
- Location of cut-over point will change over time
- Better to see the issue and deal with it than to lose it in an automated process



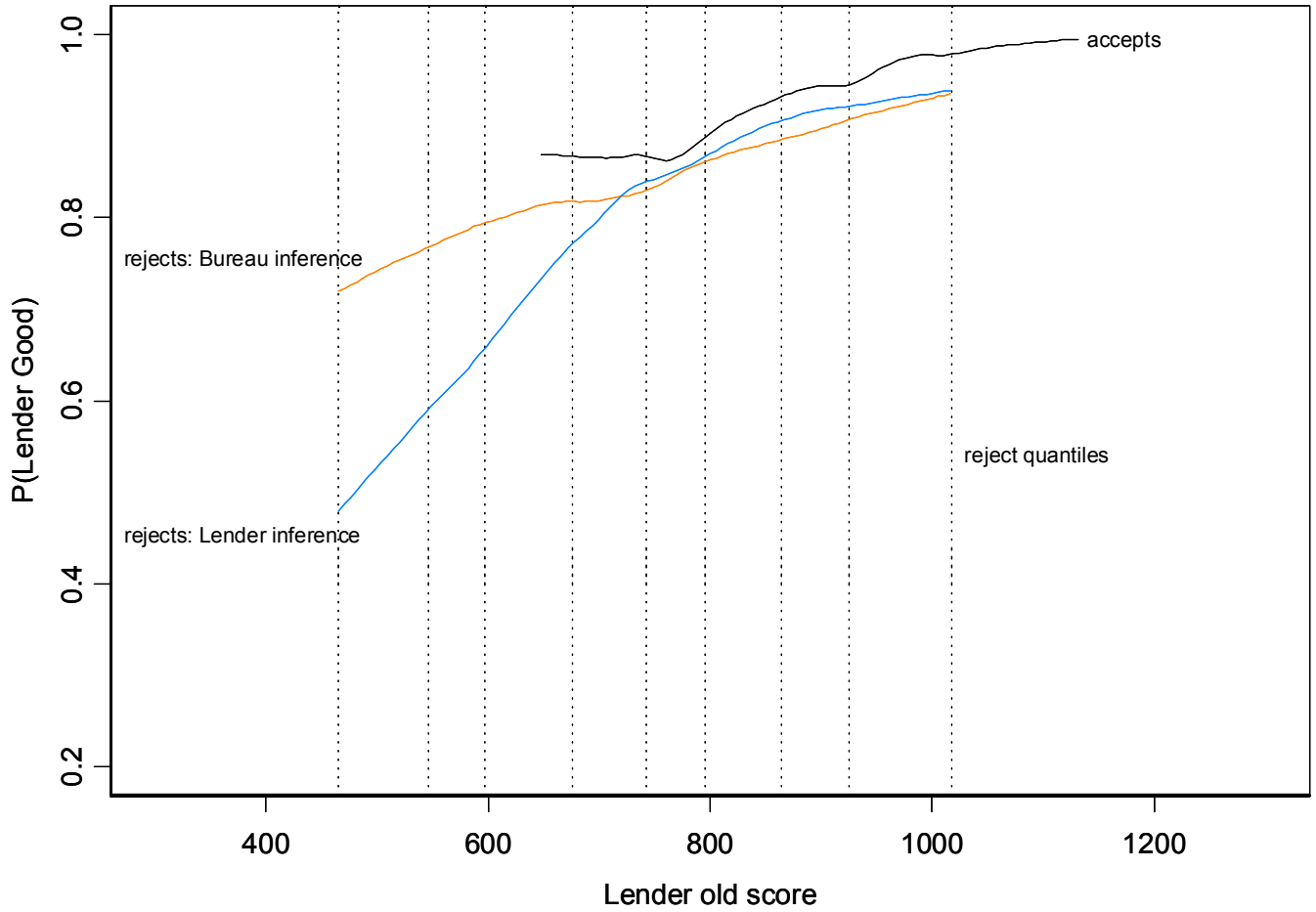


Reject inference

Reject inference

- Want to apply the model to all applications
- Have outcome data for accepted applications
- Accepts are a systematically biased sample
- Model will be biased if built on accepts
- Need to infer the outcome for the rejects
- Two broad approaches to reject inference:
  1. Extrapolate from the accepts (no new information)
  2. Use proxy outcome information for all applications
- Uncertainty of inferred outcome may dominate change in model due to improved techniques

Comparison of actual and inferred outcomes



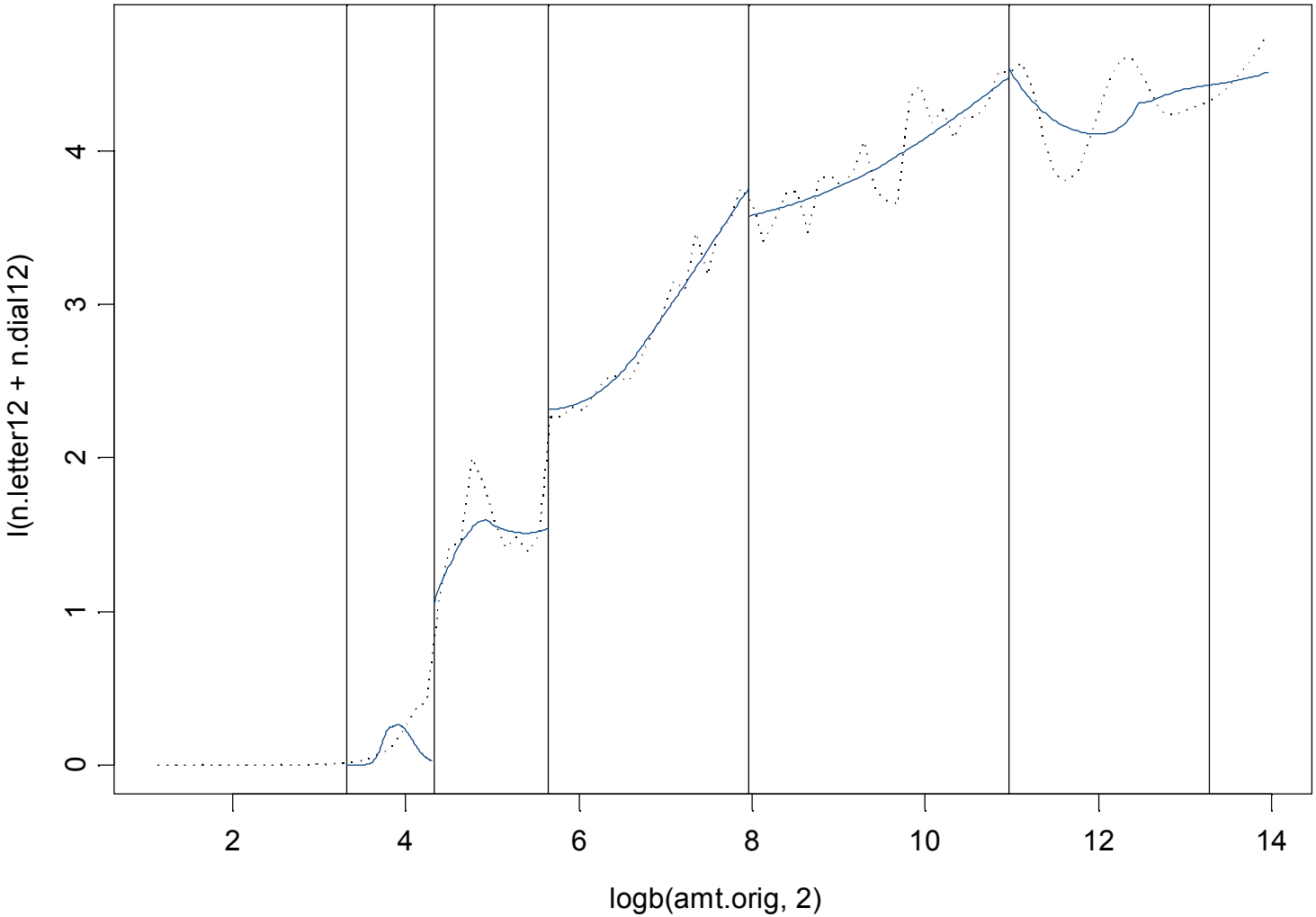
Account management actions

Impact of effort on outcome

- Collections outcome probably depends on collections effort
- Collections effort is systematically allocated via collection plans
- Collection plans are systematically allocated based on predictive characteristics
- Modelling the raw outcome may be modelling the effects of past plan allocations

- Aim for “all other things being equal” model

Maximum effort function



Population volatility

Volatility of creditor mix

Debt-collection company

- Debt from new creditors can be loaded at the drop of a hat
- Relationship between characteristics and outcome may vary
- Creditors vary widely in size (and impact on portfolio)

- Try to exclude characteristics that show a volatile relationship

Volatility of complex effects

Volatility of complex effects

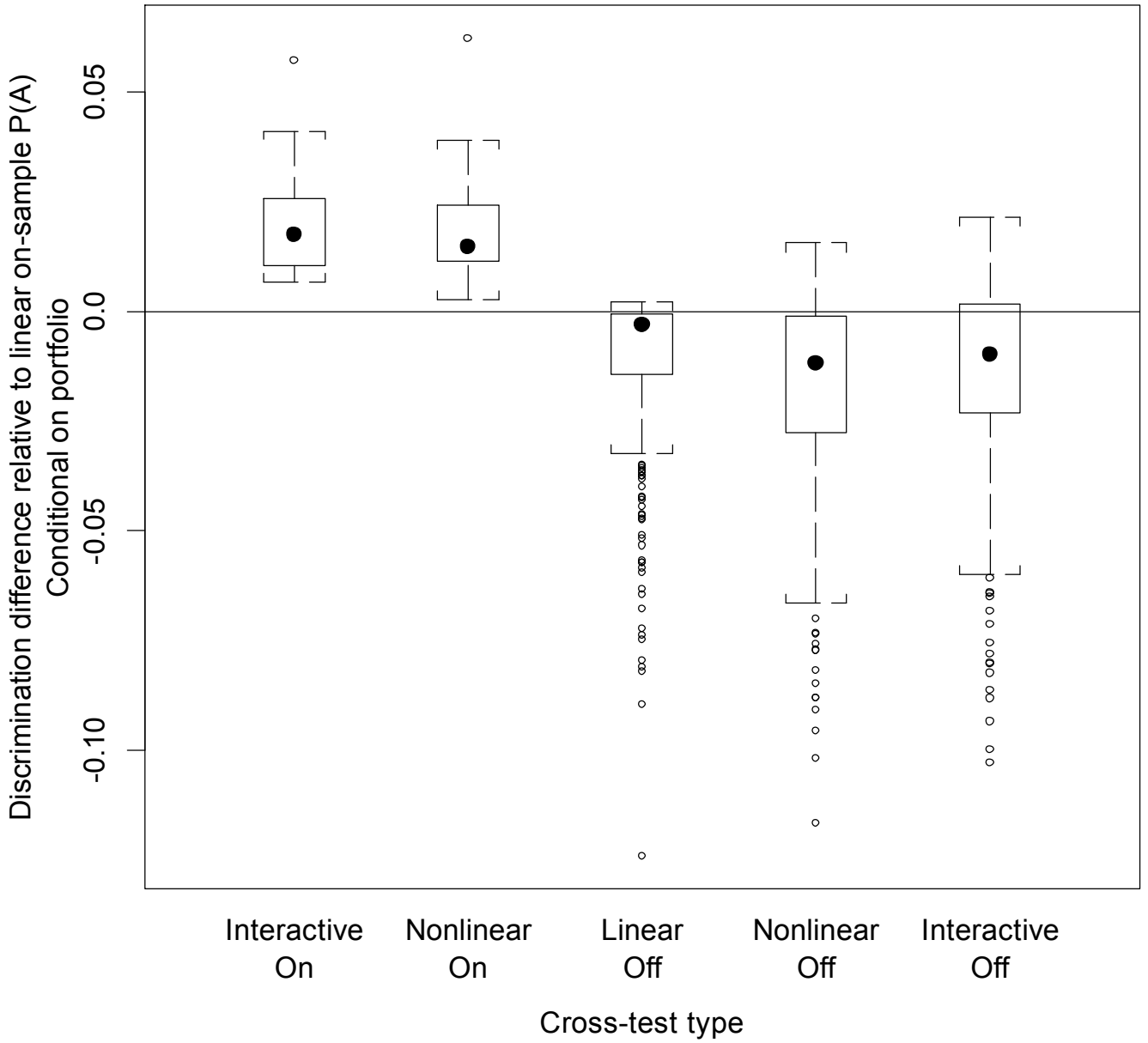
- Advanced techniques get increased prediction from modelling complex effects (nonlinearities and interactions)
- Credit scorers believe that unmotivated complex effects are more likely to be volatile
- Do not include complex effects unless there is external evidence for their reality

- 11 data sets from a range of countries, type of credit provider and credit product

Volatility study

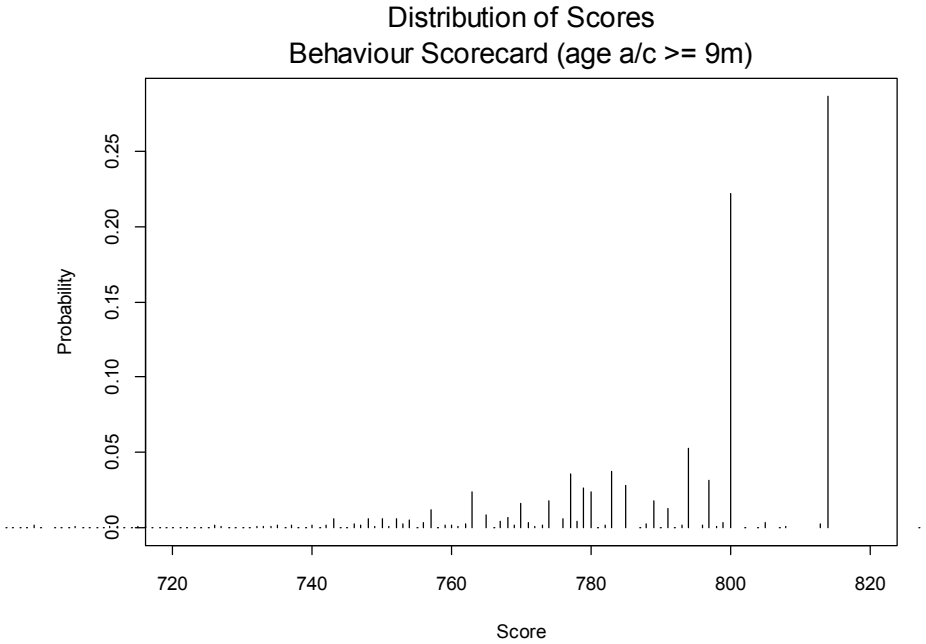
- 11 data sets from a range of countries, type of credit provider and credit product
- Three predictor variables (Age, Time in Employment, Time at Address) taken 2 at a time
- Three regression models fitted to each combination of data set and predictor variables
  - LINEAR = additive, no further transformation of predictors
  - NONLINEAR = nonparametric optimal transformation of predictors (GAM)
  - INTERACTIVE = locally weighted regression (loess; span = 0.5, degree = 2)

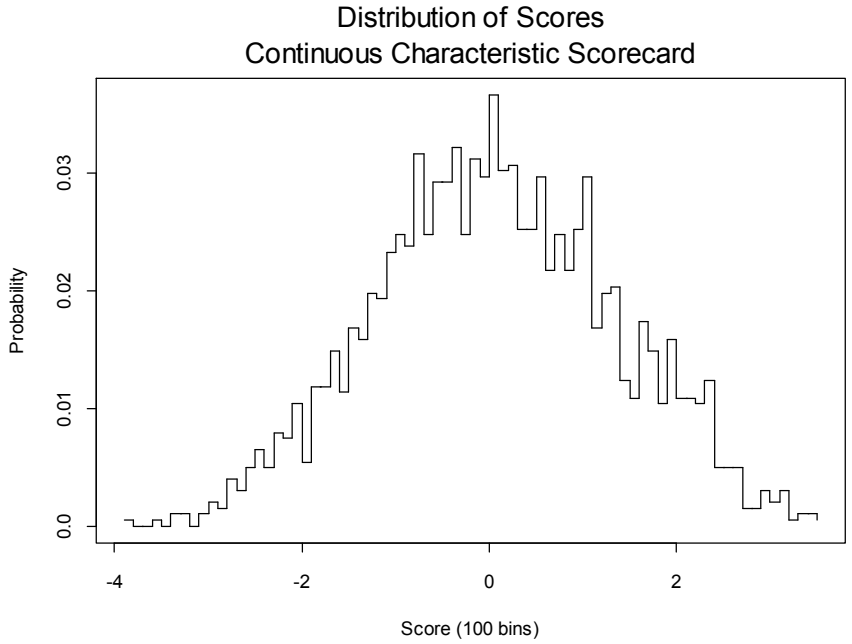
Cross-testing models



Implementation pragmatics - granularity

Granularity of score distributions

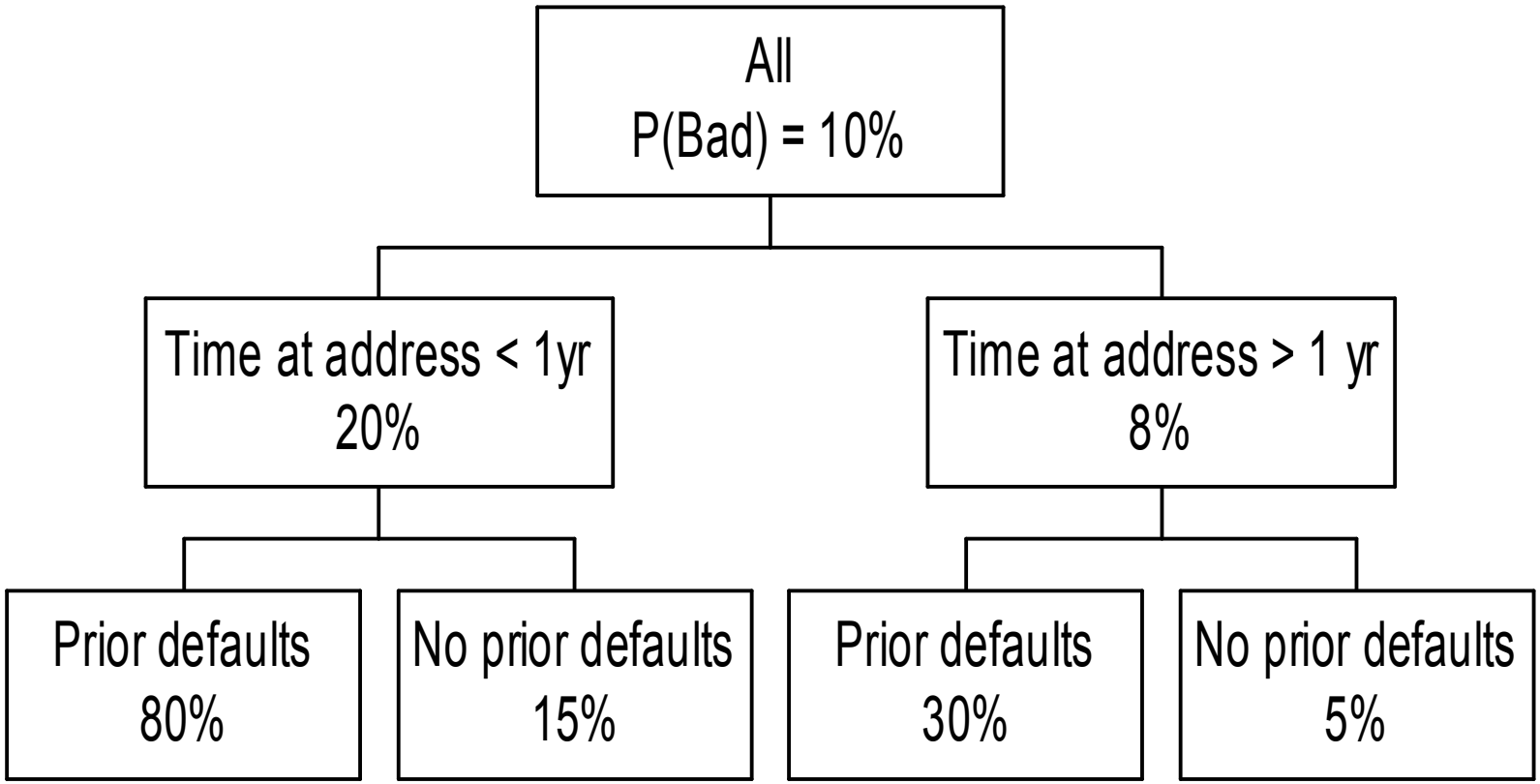




- More fine-grained scores are easier to control
  - Decision cut-offs can be placed anywhere
- Direct classification output is not controllable

Predictions are inherently probabilistic

- Scoring is **not** a classification problem



Implementation pragmatics – resistance to gaming

Application fraud detection

Much hand-crafting, including:

- Emphasise predictors that are harder to fake
  - Gender of applicant at point of sale
  - Bureau inquiries > 12 months ago
- Emphasise predictors that work in the lender's favour if gamed
  - Applicant income

Comparison of fraud models

- Hand-crafted model
  - Nominal predictor df of 3 models: ~400
  - Effective predictor df: < ~60
- Alternative “standard practice” model
  - Nominal predictor df of 1 model: ~100
- Performance
  - Equivalent predictive power at development
  - Alternative model significantly worse after a few months
  - Crafted model’s predictive power unaltered after a few months
  - Crafted model still in use > 5 years after implementation!

Conclusion

Conclusions

Credit scoring is not just fitting a model to data.

There are many reasons why the available data do not represent the future population.

Techniques for better fitting the data at hand are unlikely to yield practical benefits

Problems for automated modelling:

- Principled incorporation of external knowledge
- Optimisation for robustness
- Action-conditional predictions in the absence of true experimental design
- Dealing with causal loops

Meta-Conclusions

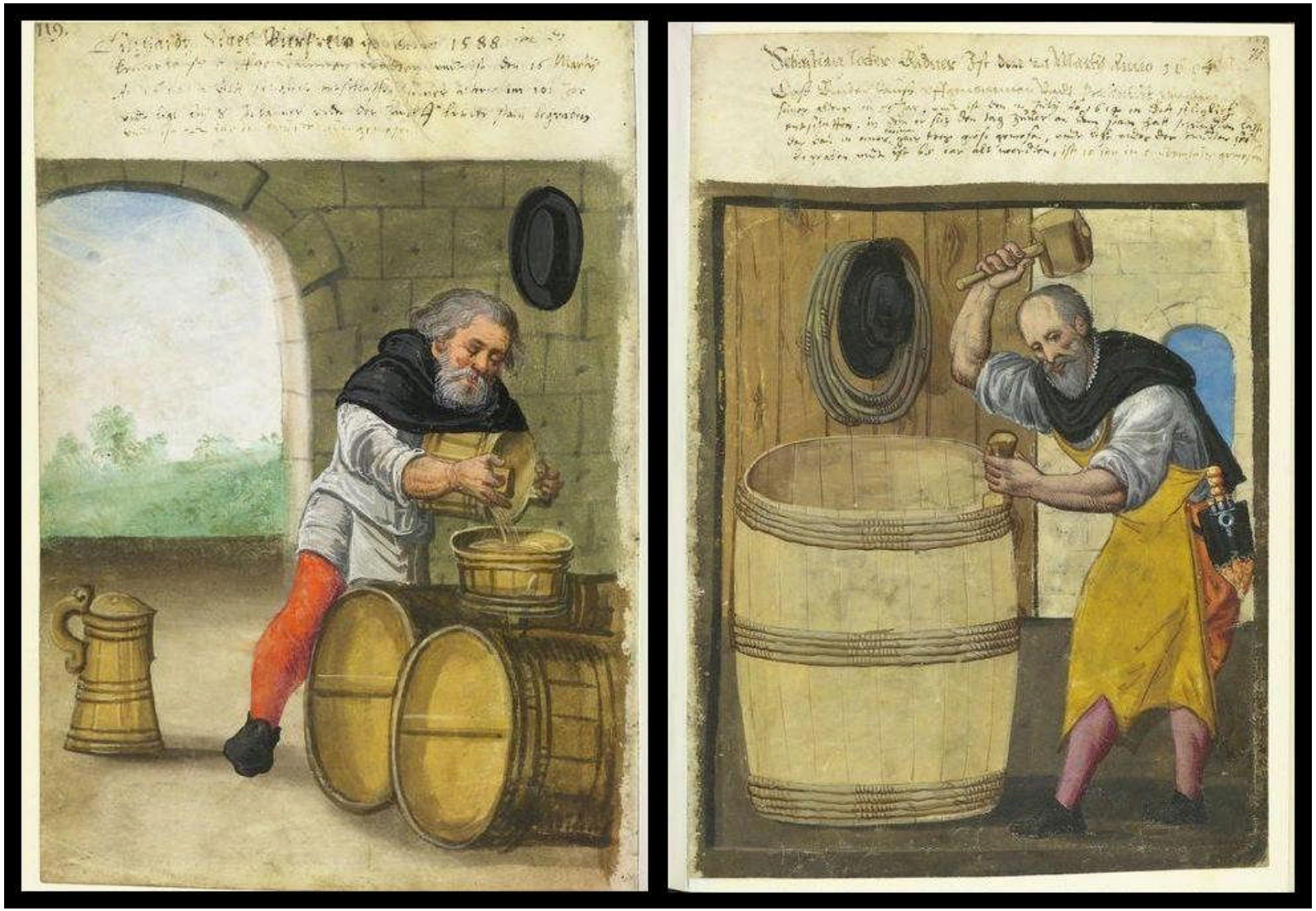
The analyst is very important!

The analyst is a mechanism for incorporating all the effects that are not present in the data now (but may be later)

The analyst is a means for understanding the environmental constraints and incorporating them in the model

The environment may reduce/remove the theoretical advantage of more sophisticated techniques

Your brain is your most powerful tool



Reused from http://www.flickr.com/photos/bibliodyssey/3085763437/sizes/o/in/photostream/ with permission of peacay