

# Data Mining

---

CS57300

Purdue University

December 7, 2010

# Announcements

---

- Qualifier: Dec 13 9-10pm (after the final)
- Final report: deadline extended to Dec 15th 4pm
- Please complete online student evaluations!

Data mining systems

# How to choose a data mining system

---

- Commercial data mining systems have little in common
  - Different data mining functionality or methodology
  - May even work with completely different kinds of data
- Need to consider multiple dimensions in selection
  - Data types: relational, transactional, sequential, spatial?
  - Data sources: ASCII text files? multiple relational data sources? support open database connectivity (ODBC) connections?
  - System issues: running on only one or on several operating systems? a client/server architecture? provide Web-based interfaces and allow XML data as I/O?

# Choosing a system

---

- Dimensions (cont):
  - Data mining functions and methodologies
    - One vs. multiple data mining functions
    - One vs. variety of methods per function
    - More functions and methods per function provide the user with greater flexibility and analysis power
  - Coupling with DB and/or data warehouse systems
    - Four forms of coupling: no coupling, loose coupling, semitight coupling, and tight coupling
    - Ideally, a data mining system should be tightly coupled with a database system

# Choosing a system

---

- Dimensions (cont):
  - Scalability: Row-based (or database size)? Column-based (or dimension)
    - Curse of dimensionality: it is much more challenging to make a system column scalable than row scalable
  - Visualization tools
    - “A picture is worth a thousand words”
    - Data visualization, mining result visualization, mining process visualization, and visual data mining
  - Data mining query language and graphical user interface
    - Easy-to-use and high-quality graphical user interface
    - Essential for user-guided, highly interactive data mining

# Example data mining systems

---

- IBM InfoSphere Warehouse
  - Wide range of data mining algorithms
  - Scalable mining algorithms
  - Toolkits: OLAP, data preparation, data visualization tools, unstructured data analysis
  - Tight integration with IBM's DB2 relational db system
- SAS Enterprise Miner
  - A variety of statistical analysis tools
  - Data warehouse tools and multiple data mining algorithms
  - Easy to use GUI

# Example systems

---

- Microsoft SQL Server 2008
  - Integrate DB and OLAP with multiple mining methods
  - Supports Object Linking and Embedding Database (OLEDB) -- access to wider formats of data than just ODBC
- Vero Insight MineSet
  - Multiple data mining algorithms and advanced statistics
  - Advanced visualization tools (originally developed by Silicon Graphics)
- PASW Modeler (SPSS)
  - Integrated data mining development environment for end-users and developers
  - Multiple data mining algorithms and visualization tools

# Example systems

---

- DBMiner (developed by Jiawei Han at SFU)
  - Multiple data mining modules: discovery-driven OLAP analysis, association, classification, and clustering
  - Efficient, association and sequential-pattern mining functions, and visual classification tool
  - Mining both relational databases and data warehouses

# Top Ten Data Mining Mistakes

(source: *John Edler, Edler Research*)

# You've made a mistake if you...

---

- Lack data
- Focus on training
- Rely on one technique
- Ask the wrong question
- Listen (only) to the data
- Accept leaks from the future
- Discount pesky case
- Extrapolate
- Answer every inquiry
- Sample casually
- Believe the best model

# 0: Lack data

---

- Need labeled cases for best gains
- Interesting known cases may be exceedingly rare
- Should not proceed until enough critical data is gathered to make analysis worthwhile
- Example: credit scoring
  - Company randomly gave credit to thousands of applicants who were risky by conventional scoring method, and monitored them for two years
  - Then they estimated risk using what was known at the start
  - This large investment in creating relevant data paid off

# 1: Focus on training

---

- Only out-of-sample results matter
- Example: cancer detection
  - MD Anderson doctors and researchers (1993), using neural networks, surprised to find that longer training (week vs. day) led to only slightly improved training results, and much worse evaluation results.
- Sampling (bootstrap, cross-validation, jackknife, leave-one-out...) is an essential tool for evaluation
- Note that resampling no longer tests a single model, but a model class, or a modeling process

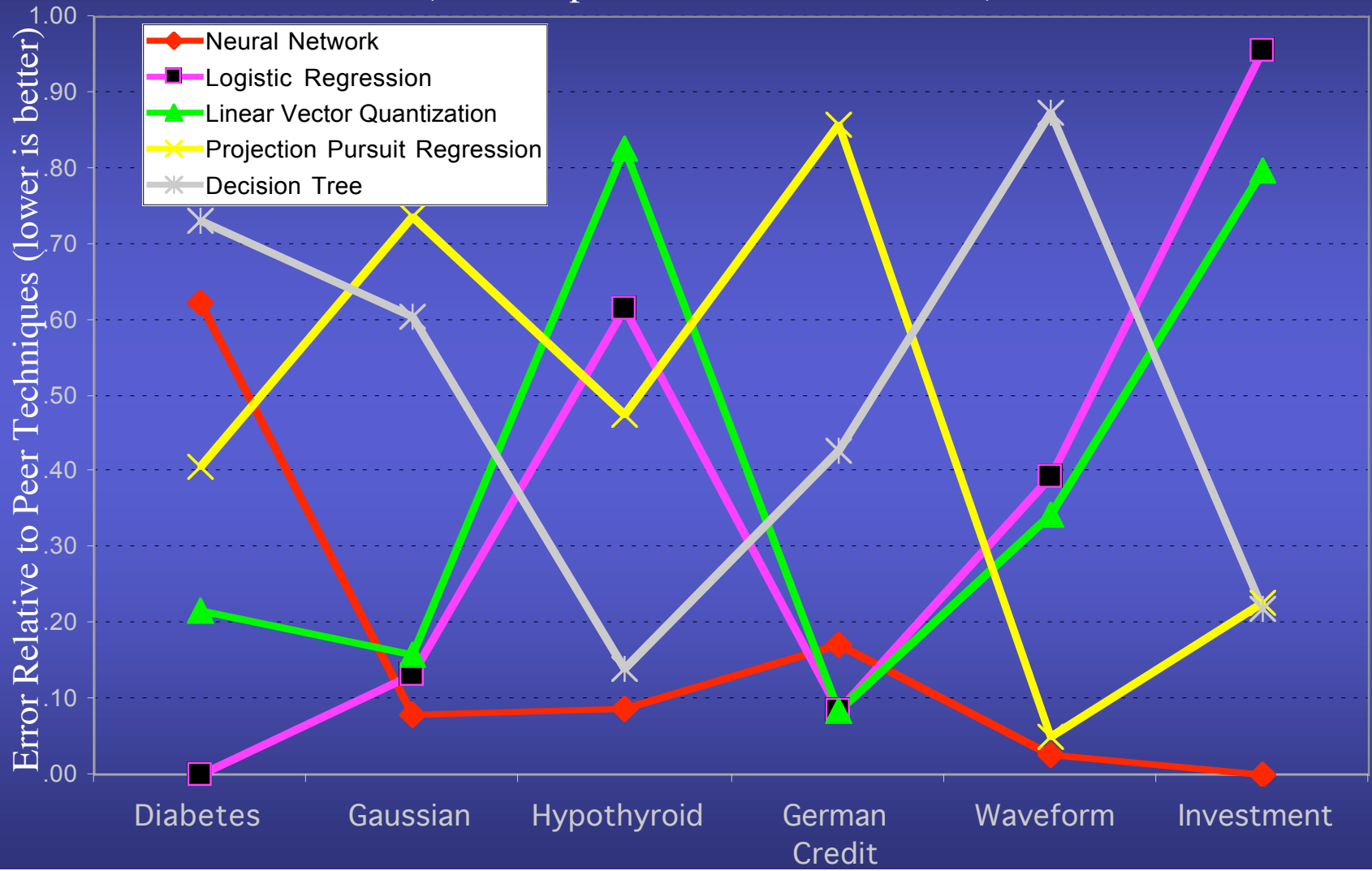
## 2: Rely on one technique

---

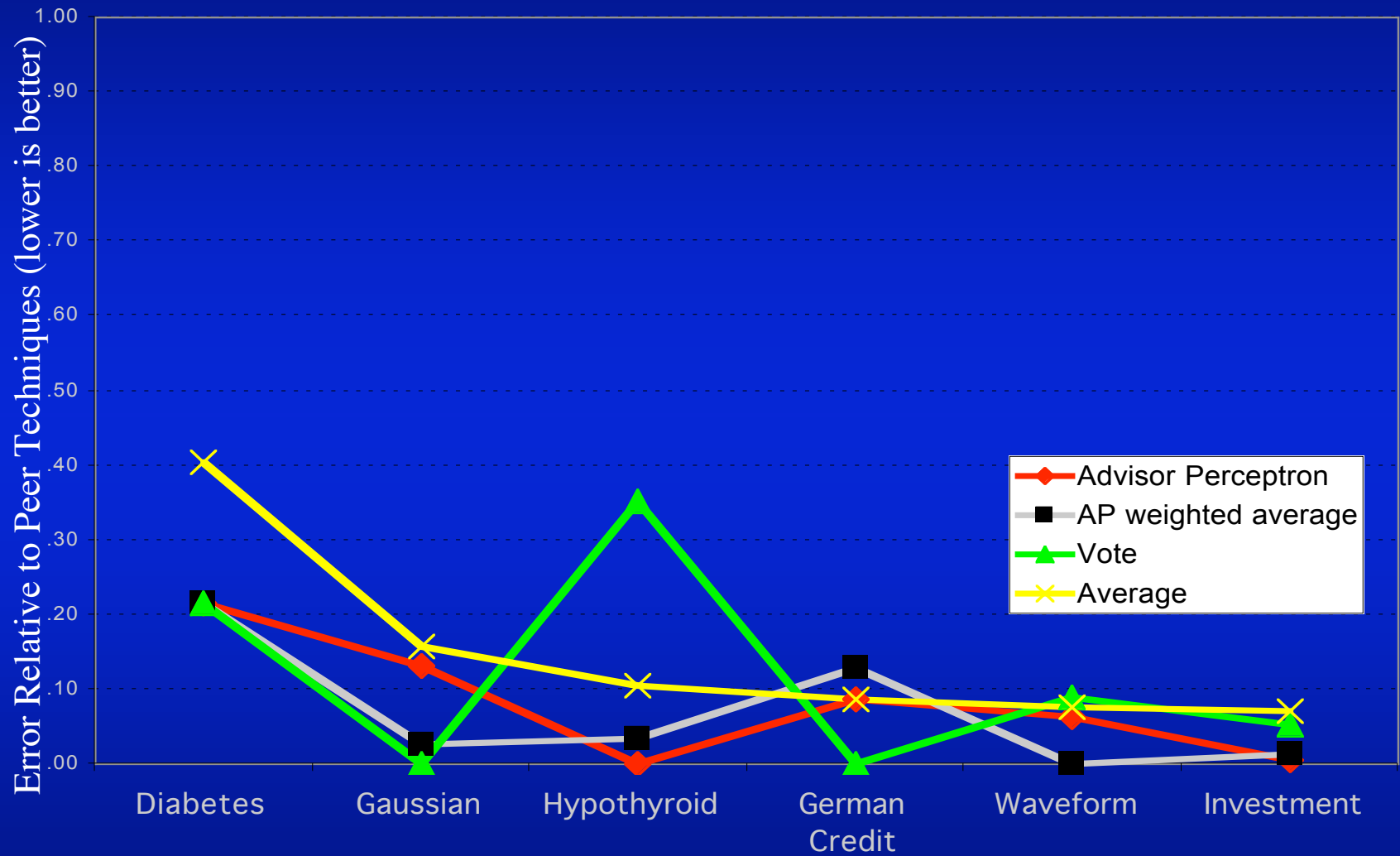
- "To a person with a hammer, all the world's a nail."
- For best work, need a whole toolkit.
- At very least, compare your method to a conventional one (e.g., naive Bayes, logistic regression)
- It's somewhat unusual for a particular modeling technique to make a big difference, and when it is will be hard to predict.
- Best approach: use a handful of good tools (Each adds only 5-10% effort)

# Relative Performance Examples: 5 Algorithms on 6 Datasets

(with Stephen Lee, U. Idaho, 1997)



## Essentially every Bundling method improves performance



# 3: Ask the wrong question

---

- Project Goal: Aim at the right target
  - Fraud Detection at AT&T Labs: predict fraud in international calls
  - Didn't attempt to classify fraud/nonfraud for general call, but characterized normal behavior for each account, then flagged outliers → a brilliant success.
- Model Goal: Evaluate appropriately
  - Most researchers use squared error or accuracy for their convenience
  - Ask the algorithm to do what's most helpful for the system, not what's easiest for it

## 4: Listen (only) to the data

---

- Opportunistic data:
  - If data are show self-selection bias, nothing inside the data protects the analyst from finding significant, but wrong results
- Designed experiments:
  - Example 1:
    - Distinguish between tanks and background using neural networks
    - Great results on out-of-sample portion of database but results were found to depend on random pixels
    - Why? Tanks photographed on sunny days, background only on cloudy days

## 4: Listen (only) to the data

---

- Designed experiments:
  - Example 2:
    - Tanks and Trucks on rotating platforms, to train to discriminate at different angles
    - Used radar, Fourier transforms, principle components, and polynomial networks
    - But, *platform corner* was determined to be the key signal
    - And, discriminated between the two classes primarily using *bushes*

## 5: Accept leaks from the future

---

- Example:
  - Forecasting interest rate at Chicago Bank
  - Neural network was 95% accurate, but *output* was a candidate input
- Example 2:
  - Used moving average of 3 days, but centered on *today*
- Look for variables which work (too) well
  - Example: Insurance code associated with 25% of purchasers turned out to describe type of cancellation
- Need domain knowledge about collection process

## 6: Discount pesky cases

---

- Outliers may be skewing results (e.g. decimal point error on price) or be the whole answer (e.g. Ozone hole), so examine carefully!
- The most exciting phrase in research isn't "Aha!" but "That's odd..."
- Inconsistencies in the data may be clues to problems with the information flow process
  - Example: Direct mail
    - Persistent questioning of oddities found errors in the merge-purge process and was a major contributor to doubling sales per catalog

# 7: Extrapolate

---

- Tend to learn too much from first few experiences
- Hard to "erase" findings after an upstream error is discovered
- Curse of Dimensionality: low-dimensional intuition is useless in high dimensions
- Human and computer strengths are more complementary than alike

## 8: Answer every inquiry

---

- "Don't Know" is a useful model output state
- Could estimate the uncertainty for each output (a function of the number and spread of samples near  $X$ )
- However, few algorithms provide an estimate of uncertainty along with their predictions

## 9: Sample without care

---

- Example: Down sampling
  - MD Direct Mailing firm had too many non-responders (NR) for model (about 99% of >1M cases)
  - Took all responders, and every 10th NR to create a more balanced database of 100K cases
  - Model predicted that everyone in Ketchikan, Wrangell, and Ward Cove Alaska would respond
  - Why? Sorted data, by zipcode and 100Kth case drawn before reaching bottom of file (i.e., 999<sup>\*\*</sup>)
- Solution: Add random variables to candidate list
  - Use as "canaries in the mine" to signal trouble

## 9: Sample without care

---

- Example: Up sampling in credit scoring
  - Paucity of interesting cases led to quintupling them
  - Cross-validation employed with many techniques and modeling cycles → results tended to improve with the complexity of the models (instead of the reverse)
  - Noticed that rare cases were better estimated by complex models but others were worse
  - Had duplicated cases in each set by upsampling before splitting → need to split first!
- It's hard to beat a stratified sample (a proportional sample from each group of interest)

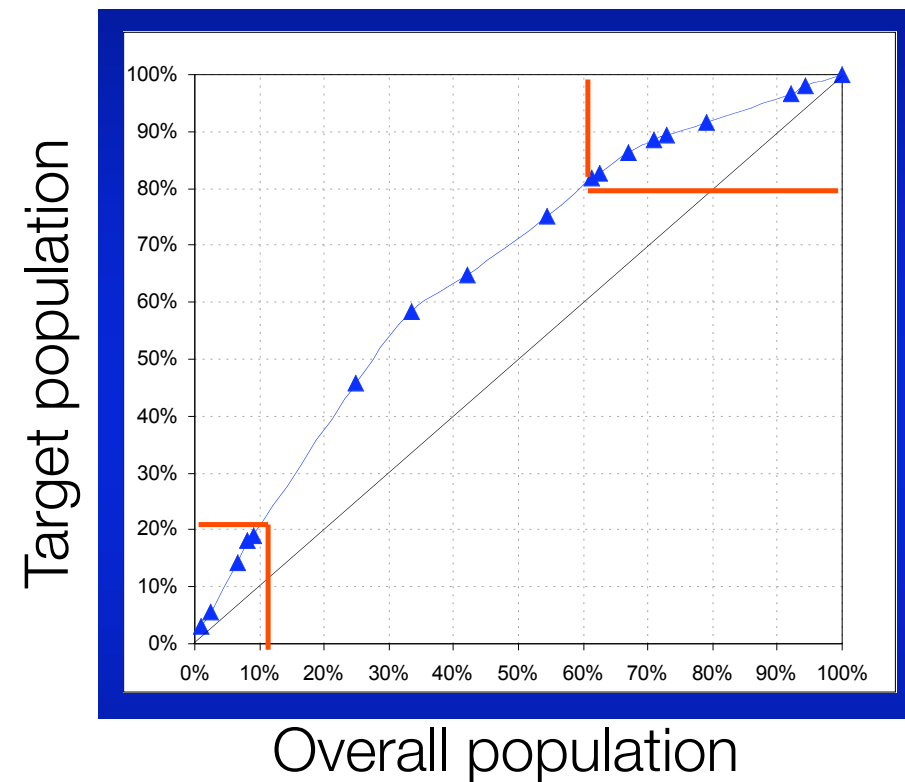
# 10: Believe the best model

---

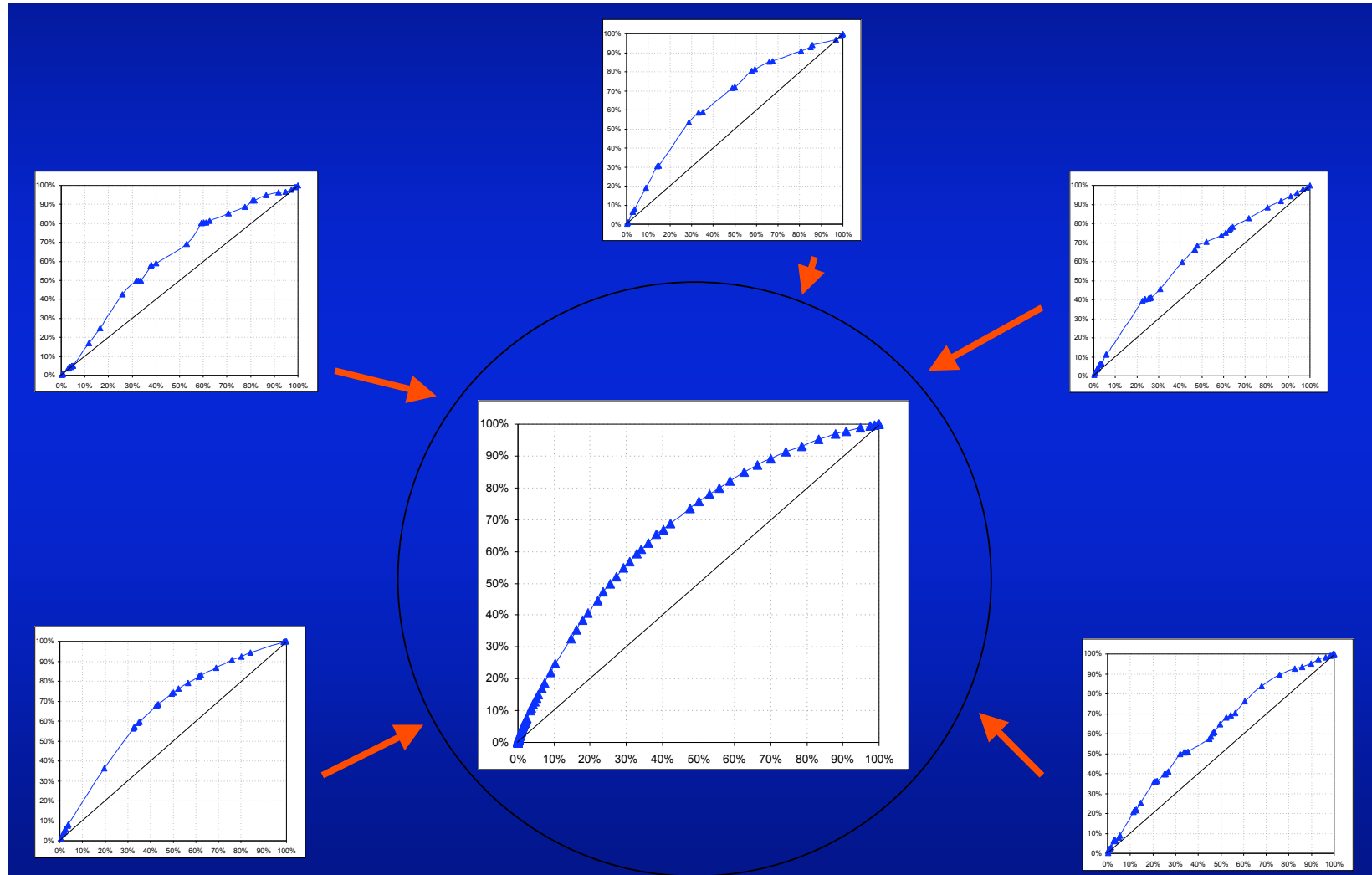
- Interpretability is not always necessary
  - Model can be useful without being "correct"
  - In practice there are often many very similar variables available and the selected variables may have only barely won out
  - And structural similarity is different from functional similarity -- competing models often look different, but act the same
- Best estimator is likely to be an ensemble of models

# Example: Lift chart

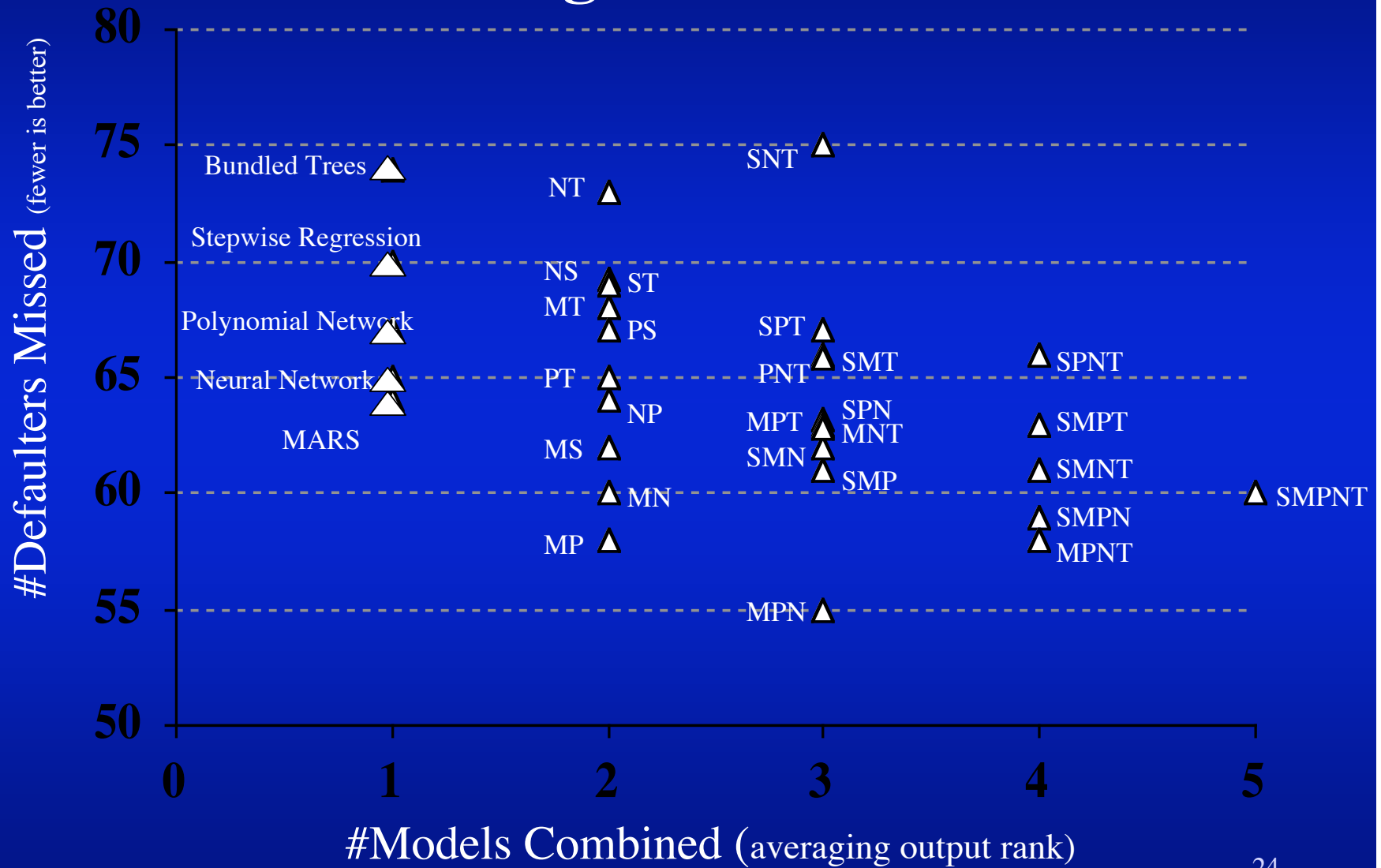
- Last quintile of customers are 4X more expensive to obtain than first quintile
- Decision tree provides relatively few decision points



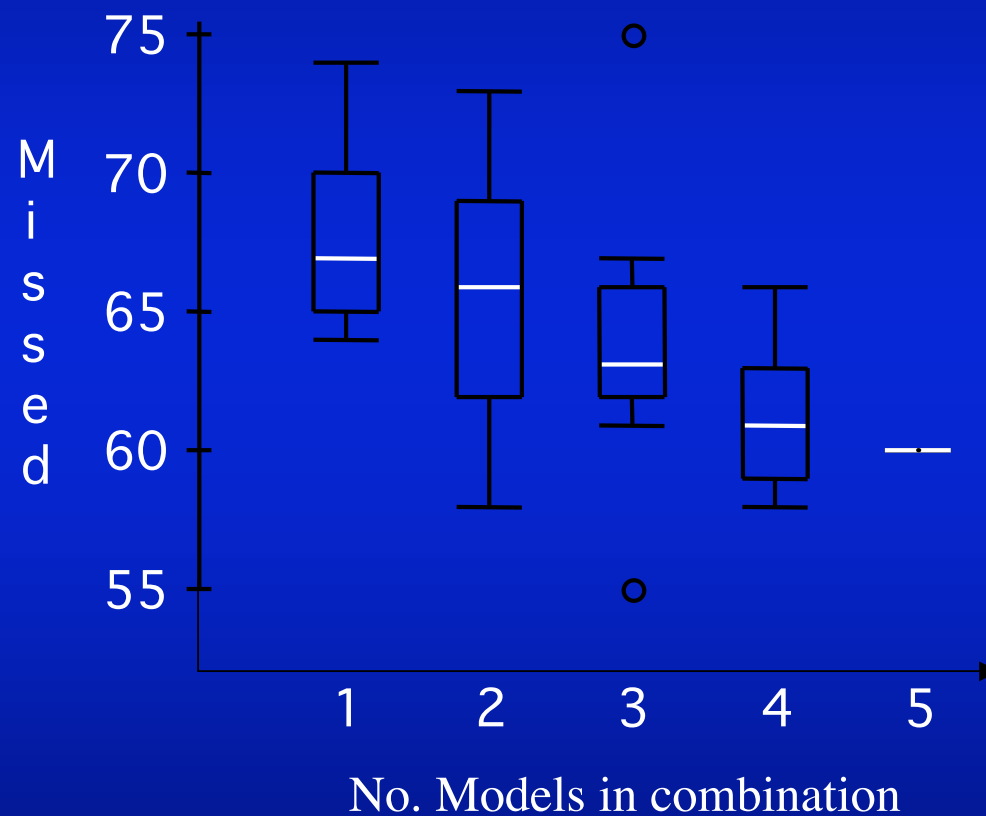
# Ensemble of 5 trees



# Credit Scoring Model Performance



# Median (and Mean) Error Reduced with each Stage of Combination



# How to succeed?

---

- More complex tools and harder problems → more ways to make mistakes
- Don't expect too much of technology alone!
- Success ← Learning ← Experience ← Mistakes
  - Persistence: Attack repeatedly, from different angles
  - Collaboration: Domain and statistical experts need to cooperate
  - Humility: Learning from mistakes requires vulnerability

# Myths and pitfalls of data mining

(source: *Tom Khabaza, DMReview*)

# Myth #1

---

- Data mining is all about algorithms
  - Data mining is a process consisting of many elements, such as formulating business goals, mapping business goals to data mining goals, acquiring, understanding and preprocessing the data, evaluating and presenting the results of analysis and deploying these results to achieve business benefits
  - A problem occurs when data miners focus too much on the algorithms and ignore the other 90-95 percent of the data mining process

# Myth #2

---

- Data mining is all about predictive accuracy
  - Predictive models should have some degree of accuracy because this demonstrates that it has truly discovered patterns in the data
  - However, the usefulness of an algorithm or model is also determined by a number of other properties, one of which is understandability
  - This is because the data mining process is driven by business expertise -- it relies on the input and involvement of non-technical business professionals in order to be successful

# Myth #3

---

- Data mining requires a data warehouse
  - Data mining can benefit from warehoused data that is well organized, relatively clean and easy to access
  - But warehoused data may be less useful than the source or operational data -- in the worst case, warehoused data may be completely useless (e.g. if only summary data is stored)
- Data mining benefits from a properly designed data warehouse and constructing such a warehouse often benefits from doing some exploratory DM

## Myth #4:

---

- Data mining is all about vast quantities of data
  - Focusing on the size of data stores provides a convenient introduction to the topic, but subtly misrepresents the nature of data mining
  - Many useful data mining projects are performed on small or medium-sized data sets (e.g., containing only a few hundreds or thousands of records)
  - Usability and scalability are both important attributes -- whether the datasets are large or small, organizations should choose a data mining tool that optimizes the analyst's time

# Myth #5

---

- Data mining should be done by a technology expert
  - When performed without business knowledge, data mining can produce nonsensical or useless results so it is essential that data mining be performed by someone with extensive knowledge of the business problem
  - It is the responsibility of data mining tool providers to ensure that tools are accessible to business users

# Pitfalls

---

## 1. Buried under mountains of data

- Do not always need to build models from millions of examples just because the data are available

## 2. The Mysterious Disappearing Terabyte

- For a given data mining problem, the amount of available and *relevant* data may be much less than initially supposed

# Pitfalls

---

## 3. Disorganized Data Mining

- Data mining can occasionally, despite the best of intentions, take place in an ad hoc manner, with no clear goals and no idea of how the results will be used -- this leads to wasted time and unusable results

## 4. Insufficient Business Knowledge

- Business knowledge is critical -- without it, organizations can neither achieve useful results nor guide the data mining process towards them

# Pitfalls

---

## 5. Insufficient Data Knowledge

- In order to perform data mining, we must be able to answer questions such as: What do the codes in this field mean, and can there be more than one record per customer in this table and more? In some cases, this information is surprisingly hard to come by

## 6. Erroneous Assumptions (courtesy of experts)

- Business and data experts are crucial resources, but this does not mean that the data miner should unquestioningly accept every statement they make

# Pitfalls

---

## 7. Incompatibility of Data Mining Tools

- No toolkit will provide every possible capability, especially when the individual preferences of analysts are taken into account, so the toolkit should interface easily with other available tools and third-party options

## 8. Locked in the Data Jail House

- Some tools require the data to be in a proprietary format that is not compatible with commonly used database systems
- This can result in high overhead costs and create difficulty in deployment into an organization's system