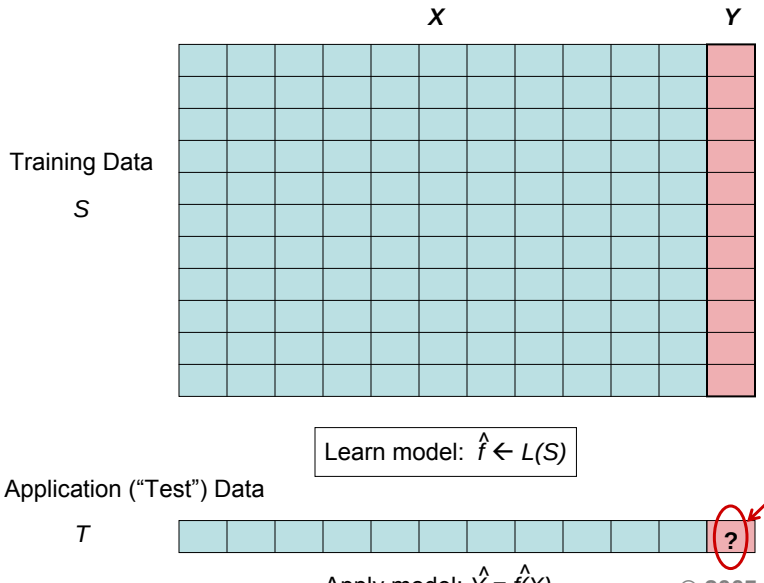
Toward Economic Machine Learning and Utility-based Data Mining

Foster Provost
New York University

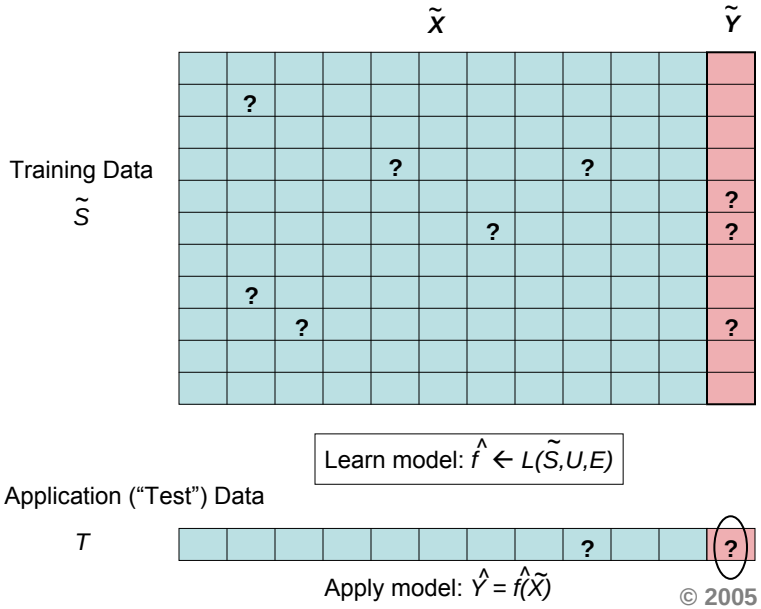
First International Workshop on
Utility-Based Data Mining
August 2005

Thanks to many … especially Maytal Saar-Tsechansky.

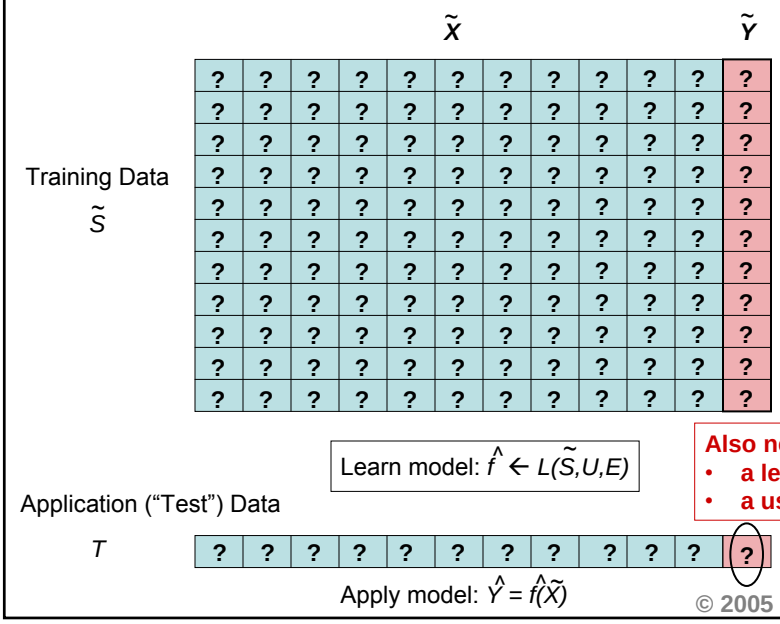
Traditional Supervised Learning



Utility-based Data Mining

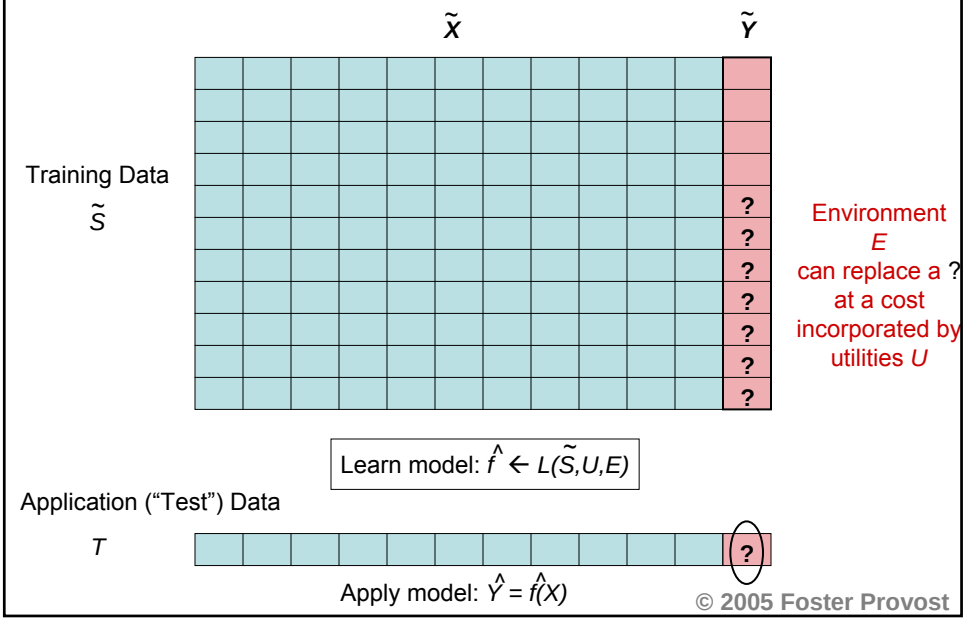


Utility-based Data Mining

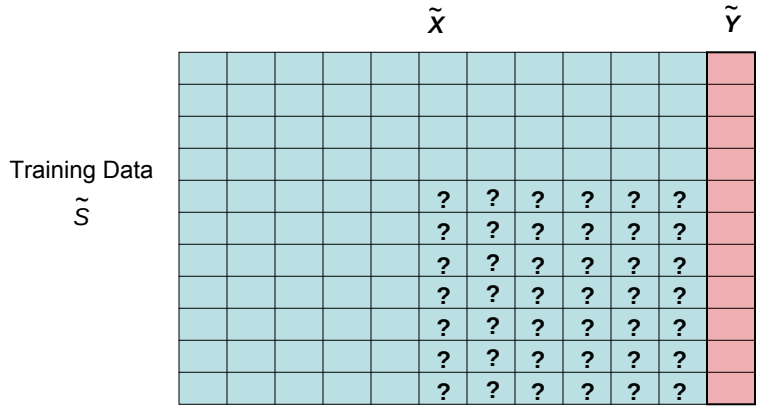


Issue #1: What information may be missing and acquirable at a cost?

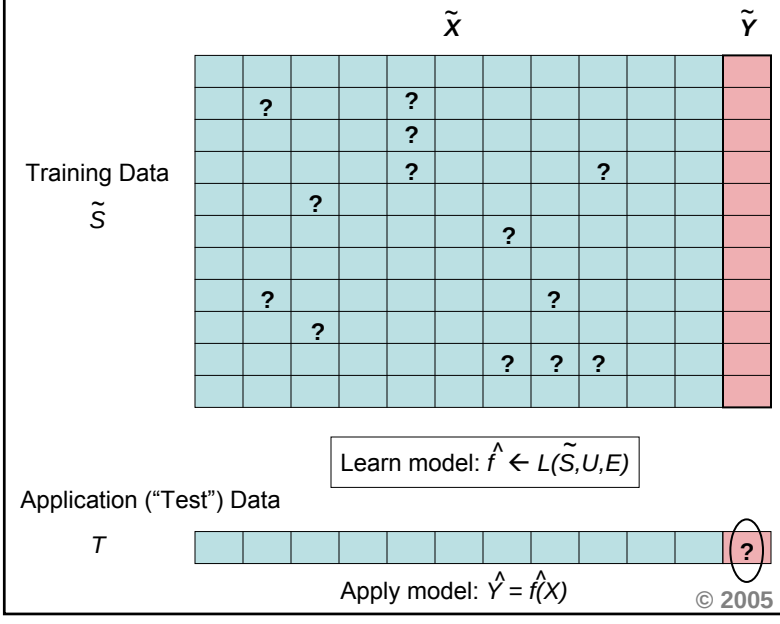
Utility-based Data Mining - Example 1



Utility-based Data Mining - Example 2



Utility-based Data Mining - Example 3



Issue #1′ : What acquisition actions does the environment support?

- $a(y_i)$ – traditional active learning
- $a(x_{ij})$ – general active feature acquisition (Melville et al., 2005)
- $a(x_{i*})$ – instance completion (Zheng & Padmanabhan 2002) (Melville et al. 2004)
- $a(x_{?j}|y_? = c)$ – "budgeted learning" a la (Lizotte et al. 2003)
- $a(x_i y_i)$ – progressive sampling (Provost et al. 1999)
- $a(x_{i*}|y_i = c)$ – "budget-sensitive" progressive sampling (Weiss & Provost 2003)
- $a(y_i|x_i)$ – learning with membership queries (Angluin 1988)
- $a(x_{ij})$ (with $x_{ik+1} = y_i$) general active learning (Somebody 2006)

Other settings?

- $a(e_{i*})$ – secondary data access for network learning (Macskassy & Provost 2005)
- s(i′) – costly feature construction (Somebody 2007)
- s(κ) – background knowledge acquisition (Somebody 2008)
- etc.

Issue #2: How to decide what information should be acquired (next)?

- Common strategy: *estimate uncertainty*
  - in order to reduce it
  - most common strategy for active learning
    - e.g., uncertainty sampling, query by committee, etc.
- Limitations?
  - why/when does it work?
    - e.g., training-set outliers may have high uncertainty
  - unclear how to apply with different sorts of information

- One general strategy: *maximize expected utility*
- Applies to general setting with different sorts of information
- Examples:
  - active learning (e.g., (Roy & McCallum 2001) and others)
  - active feature acquistion (e.g., this workshop)
- Limitations? (Challenges)
  - computational expense
  - estimation accuracy!
  - myopia vs. all possible info combinations
  - need to build models of various probabilities (e.g.. features vs. target)

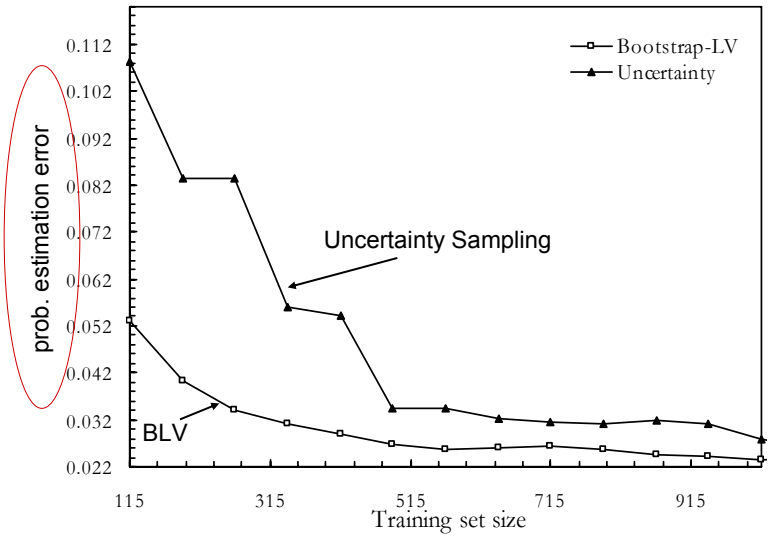
Issue #3: Is the acquisition directed by the (right) goal of the learning?

Examples:

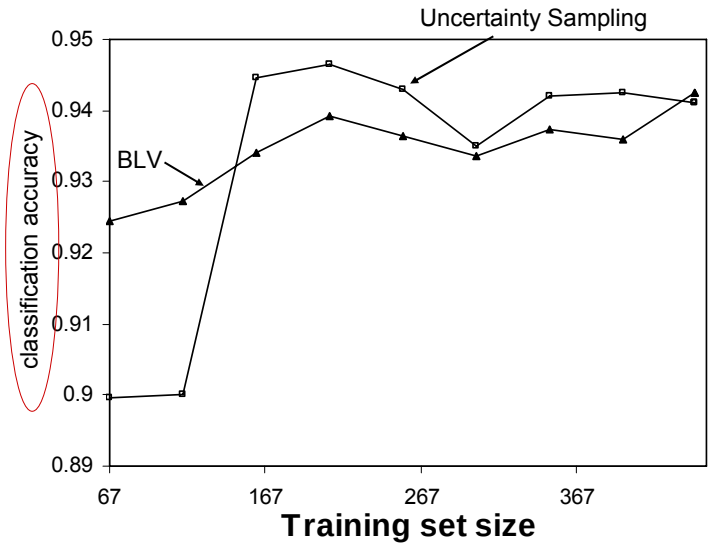
- minimize classification error
- minimize prob. estimation error
- maximize utilility!
  - for some specific problem

  → need to take decision-making into account
- on-line utility maximization
  - learning while acting
  - cf. bandit problems, seq. analysis, reinforcement learning
  - I won't have time to talk about today, but Naoki will…?

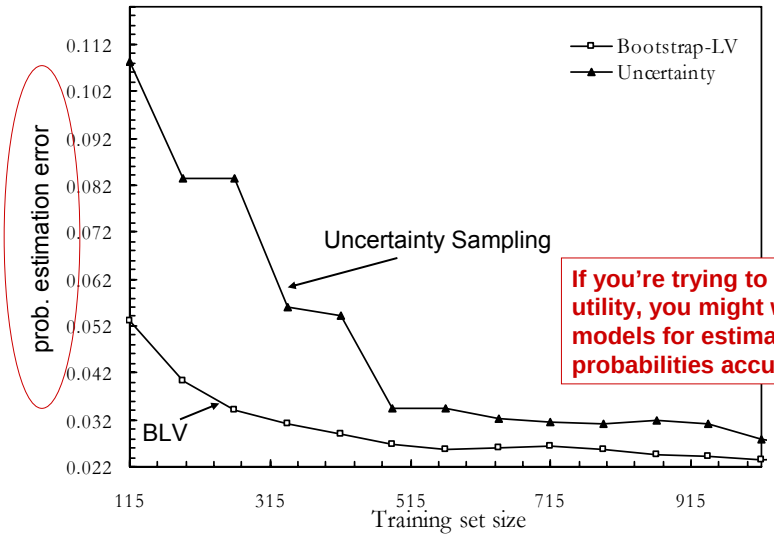
(Saar-Tsechansky & Provost, 2004a)



(Saar-Tsechansky & Provost, 2001)

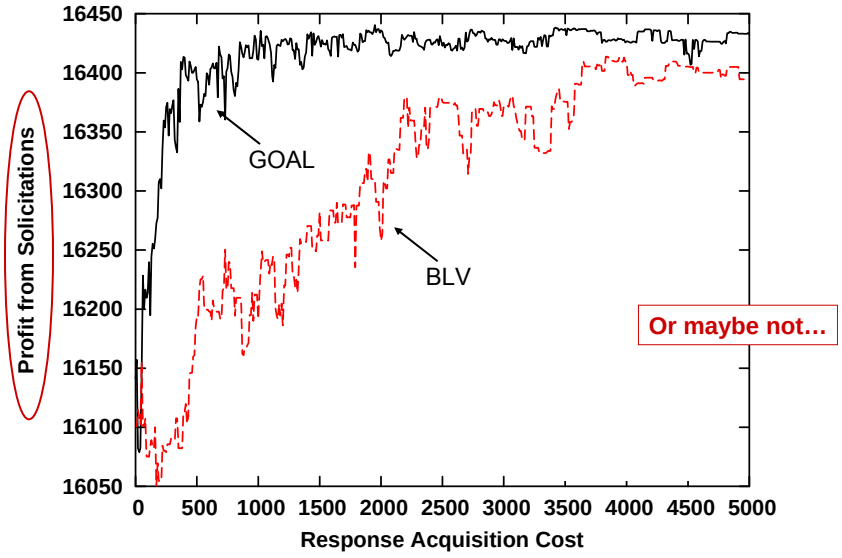


(Saar-Tsechansky & Provost, 2004a)



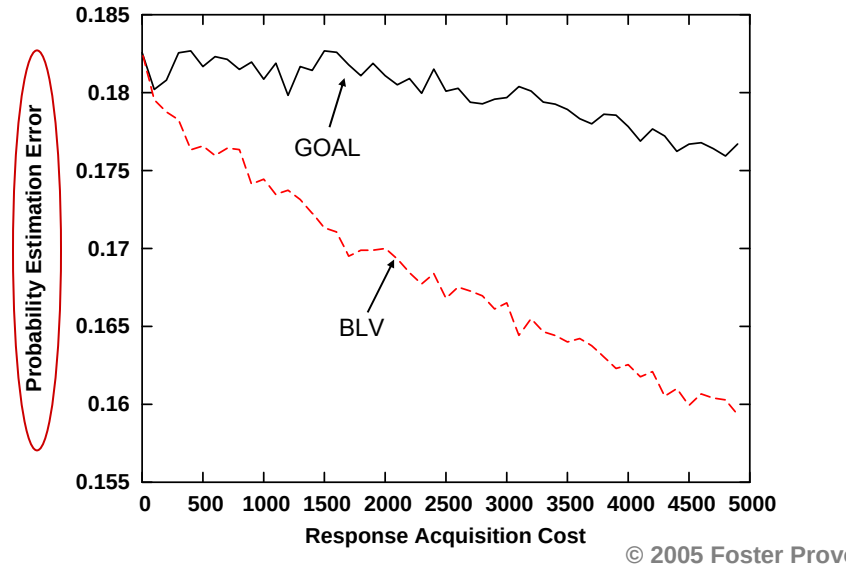
If you're trying to maximize utility, you might want models for estimating probabilities accurately…

(Saar-Tsechansky & Provost, 2004b)



Or maybe not…

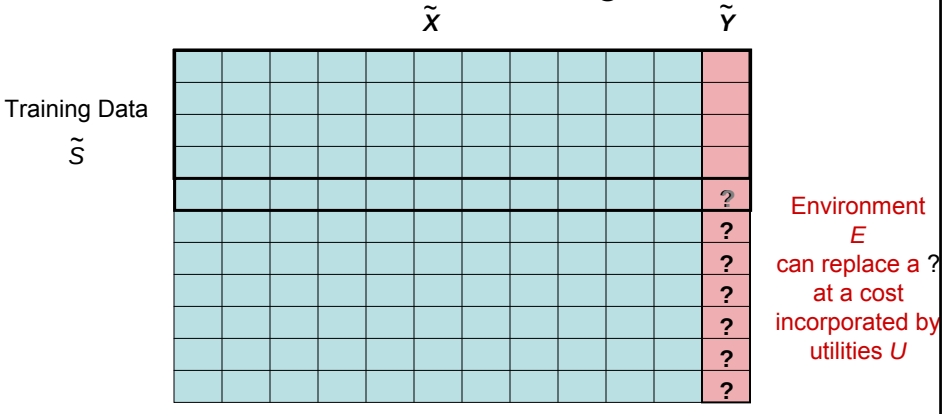
(Saar-Tsechansky & Provost, 2004b)



Issue #4: Maximize utility as compared to what?

- ignoring missing information

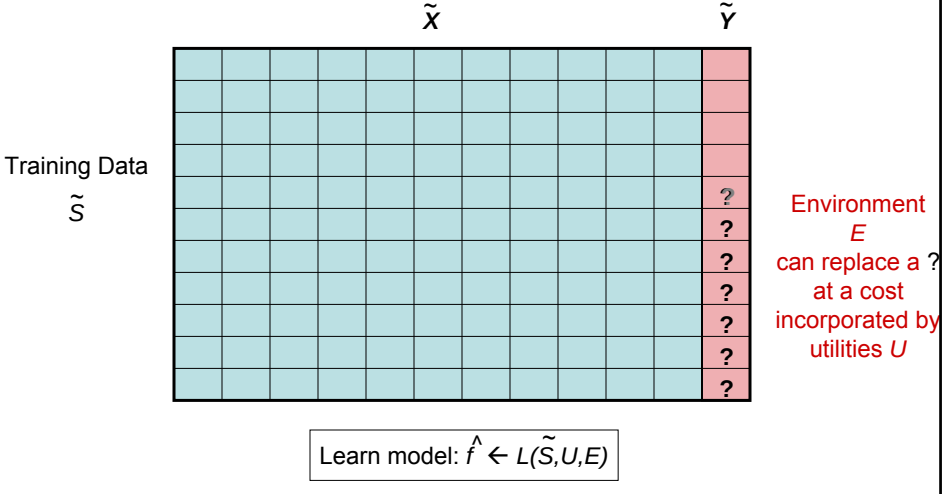
Utility-based Data Mining: Traditional Active Learning *Revisited*



Issue #4: Maximize utility as compared to what?

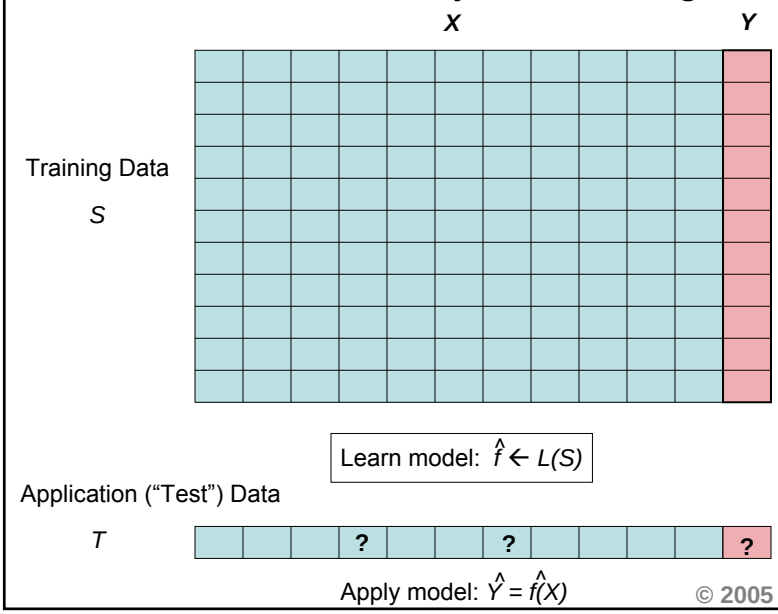
- ignoring missing information
- best alternative treatment for missing information!
  - potential info may look valuable in isolation, but marginal value may be small
- nobody has done this?
- we don't even know what are the "best alternative treatments"

Utility-based Data Mining: Traditional Active Learning *Revisited*



Should estimate not just the increase in expected value over ignoring the cases with no labels, but instead the increase in expected value over (say) the best semi-supervised learning alternative

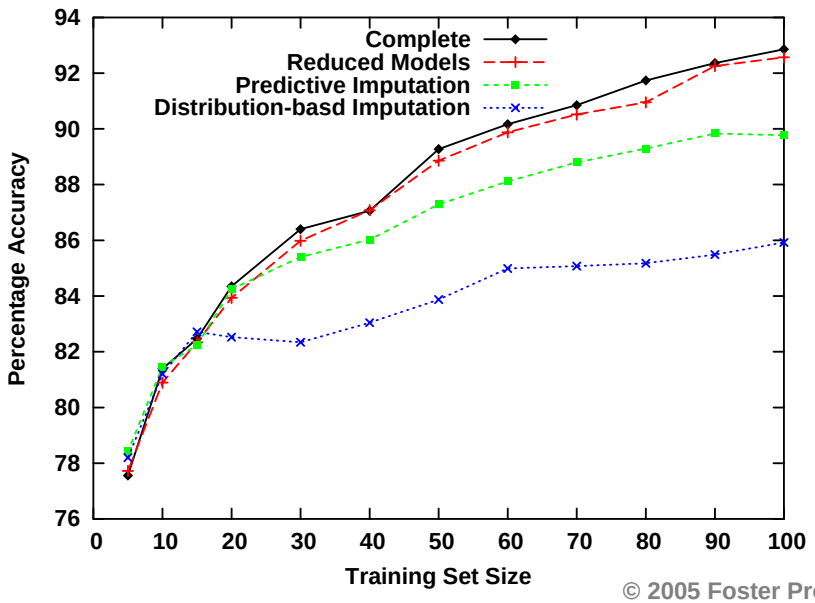
One last simple example: When should we buy the missing value(s)?



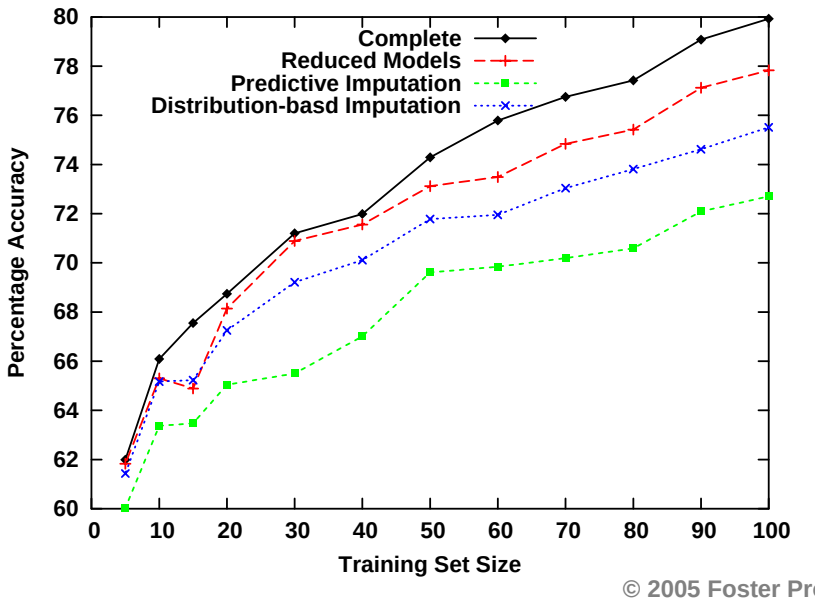
Alternatives to buying missing values at prediction time

- (ignore the case)
- estimate the value (imputation)
  - mean, mode imputation
  - use a predictive model
- estimate value distribution and combine probabilistically
  - e.g., C4.5's technique
- build a "reduced" model without the missing values!
  - lazy learning
  - store multiple models

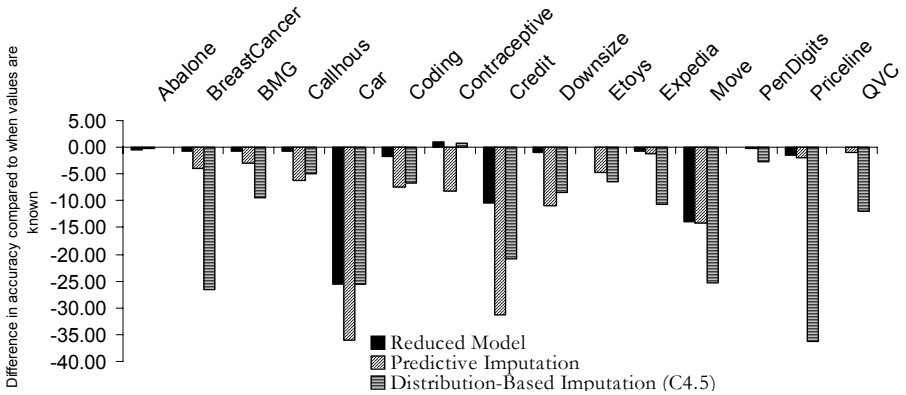
(Saar-Tsechansky & Provost, 2005)



(Saar-Tsechansky & Provost, 2005)



(Saar-Tsechansky & Provost, 2005)



which missing-value treatment is used can make a huge difference

…for many domains, with reduced models no advantage to buying information!

Summary: Issues & Challenges

- Issue #1: What information may be missing?
- Issue #1′ : What are supported acquisition actions?
  - **proposal**: a general UBDM framework
  - **open question**: *can a general framework work as well on special cases?*
- Issue #2: What information should be acquired?
  - **proposal**: general expected-utility estimation
  - **open question**: *how to deal with challenges like*
    - *computational expense*
    - *estimation accuracy*
    - *myopia vs. all possible info combinations*
    - *need to build models of various probabilities (e.g.. features vs. target)*
- Issue #3: Is the acquisition directed by the (right) goal?
  - **proposal**: goal should be factored into the utility calculations
  - **open question**: do we know the costs/benefits well enough?
- Issue #4: Maximize utility as compared to what?
  - **proposal**: should compare to best alternative
  - **open question**: what is best alternative & can its performance be estimated?

thanks!