The Role of Applications in the Science of Machine Learning

Foster Provost
New York University

Thanks to: John Aronis, Bruce Buchanan, Scott Clearwater, Andrea Danyluk, Tom Fawcett, Shawndra Hill, David Jensen, Ronny Kohavi, Andrew McCallum, Sofus Macskassy, Jen Neville, Claudia Perlich, Jeff Simonoff, Ben Taskar, Gary Weiss, and many others

ICML-2003
Washington, DC

Roles of applications

- Help to stimulate researchers
- Attract (good) graduate students
- Help to convince funding sources
- Provide tests of prior research results
  - of relevance
  - of necessity
  - of efficacy
- Highlight insufficiencies of the state of the art
  - → pointing to important areas for future research

"In the meantime, while the creative power of pure reason is at work, *the outer world* again comes into play, forces upon us new questions from actual experience, opens up new branches [of the science]"

David Hilbert, 1900

A spectrum of applicability

- "I make no mockery of honest ad hockery"
  attributed to John Tukey

- NB: research using benchmark (e.g., UCI) data sets is *not* applied research!

© 2003 Foster Provost

C4-LR

High separability
16-2

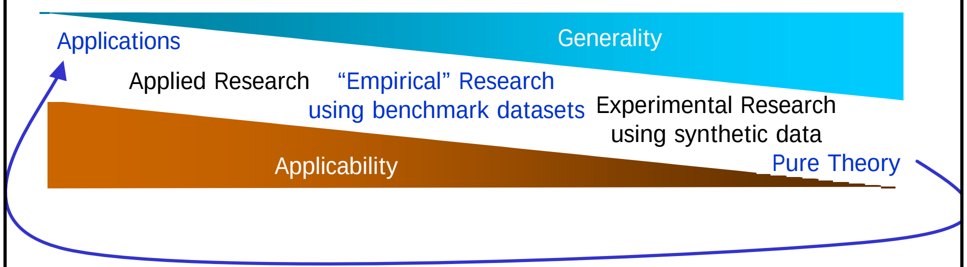
Low separability
1-8

Data set	Winner AUR	Winner Acc	Max AUR	Result
Nurse	none	none	1	Indistinguishable
Mushrooms	none	none	1	Indistinguishable
Optdigit	none	none	0.99	Indistinguishable
Letter–V	C4	C4	0.99	C4 dominates
Letter–A	C4	C4	0.99	C4 crosses
Intrusion	C4	C4	0.99	C4 dominates
DNA	C4	C4	0.99	C4 dominates
Covertype	C4	C4	0.99	C4 crosses
Telecom	C4	C4	0.98	C4 dominates
Pendigit	C4	C4	0.98	C4 dominates
Pageblock	C4	C4	0.98	C4 crosses
CarEval	none	C4	0.98	C4 crosses
Spam	C4	C4	0.97	C4 dominates
Chess	C4	C4	0.95	C4 dominates
CalHous	C4	C4	0.95	C4 crosses
Ailerons	none	C4	0.95	C4 crosses
Firm	LR	LR	0.93	LR crosses
Credit	C4	C4	0.93	C4 dominates
Adult	LR	C4	0.9	Mixed
Connects	C4	none	0.87	C4 crosses
Move	C4	C4	0.85	C4 dominates
Downsize	C4	C4	0.85	C4 crosses
Coding	C4	C4	0.85	C4 crosses
German	LR	LR	0.8	LR dominates
Diabetes	LR	LR	0.8	LR dominates
Bookbinder	LR	LR	0.8	LR crosses
Bacteria	none	C4	0.79	C4 crosses
Yeast	*none*	*none*	*0.78*	*Indistinguishable*
Patent	C4	C4	0.75	C4 crosses
Contra	*none*	*none*	*0.73*	*Indistinguishable*
IntShop	LR	LR	0.7	LR crosses
IntCensor	LR	LR	0.7	LR dominates
Insurance	*none*	*none*	*0.7*	*Indisting*
IntPriv	LR	none	0.66	LR cros
Mailing	LR	none	0.61	LR dom
Abalone	LR	LR	0.56	LR dom

Perlich et al., JMLR 2003
(cf. Lim, Loh, Shih MLJ 2001)

© 2003 Foster Provo

A spectrum of applicability



- "I make no mockery of honest ad hockery"

  John Tukey

- NB: research using benchmark (e.g., UCI) data sets is *not* applied research!

© 2003 Foster Provost

Example applications of ML

Business

- Credit scoring
- Customer retention
- Drug design
- Financial modeling
- Fraud detection ←
- Games ("video")
- Information retrieval
- Information monitoring
- Operations support
  - diagnosis/troubleshooting
  - monitoring
  - quality control
- Robotics
- Spam filtering
- Targeted marketing
- Web personalization
- and more …

Science/medicine/government

- Astronomic sky surveys
- Bioinformatics
- Biosurveillance
- Brain imaging
- Counterterrorism
- Crystallography
- Law enforcement
- Medical diagnosis
- Natural language processing
  - and other AI areas
- Planetary exploration
- Planetary image analysis
  - e.g., volcanoes on Venus
- Public health research
  - e.g., predicting carcinogenicity
- Weather forecasting
- and more …

© 2003 Foster Provost

Example: data mining for fraud detection

Can we identify which accounts have been compromised so as to take corrective action?

A typical frauded account

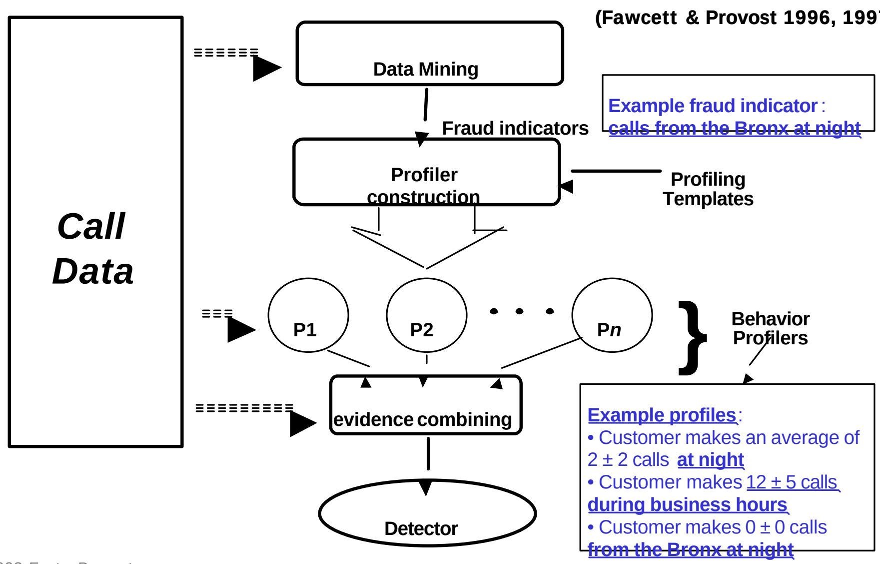
Date and Time	Day	Duration	From	To	Fraud
1/01/95 10:05:01	Mon	13 mins	Brooklyn, NY	Stamford, CT	
1/05/95 14:53:27	Fri	5 mins	Brooklyn, NY	Greenwich, CT	
1/08/95 09:42:01	Mon	3 mins	Bronx, NY	White Plains, NY	
1/08/95 15:01:24	Mon	9 mins	Brooklyn, NY	Brooklyn, NY	
1/09/95 15:06:09	Tue	5 mins	Manhattan, NY	Stamford, CT	
1/09/95 16:28:50	Tue	53 sec	Brooklyn, NY	Brooklyn, NY	
1/10/95 01:45:36	Wed	35 sec	Boston, MA	Chelsea, MA	YES
1/10/95 01:46:29	Wed	34 sec	Boston, MA	Yonkers, NY	YES
1/10/95 01:50:54	Wed	39 sec	Boston, MA	Chelsea, MA	YES
1/10/95 11:23:28	Wed	24 sec	White Plains,NY	Congers, NY	
1/11/95 22:00:28	Thu	37 sec	Boston, MA	EastBoston, MA	YES
1/11/95 22:04:01	Thu	37 sec	Boston, MA	EastBoston, MA	YES

© 2003 Foster Provost

Example: fraud detection

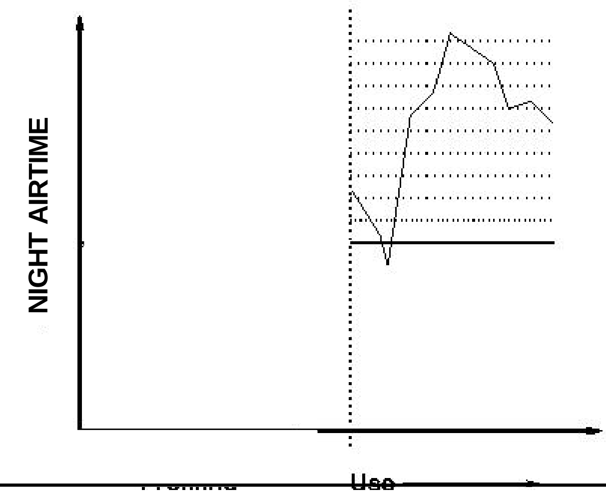
Profile customer behavior

(Fawcett & Provost 1996, 1997)



© 2003 Foster Provost

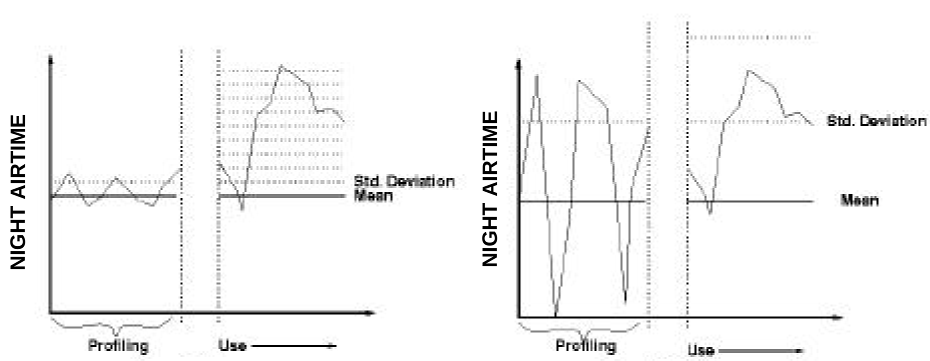
Is this behavior evidence of fraud?



©

Is this behavior evidence of fraud?

It depends on the customer.

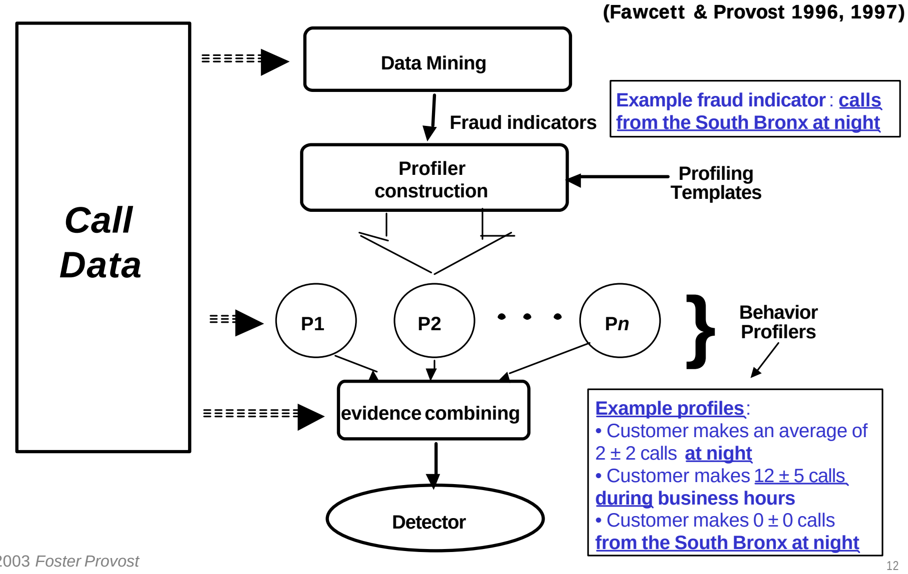


© 2003 *Foster Provost*

Example: fraud detection

Profile customer behavior

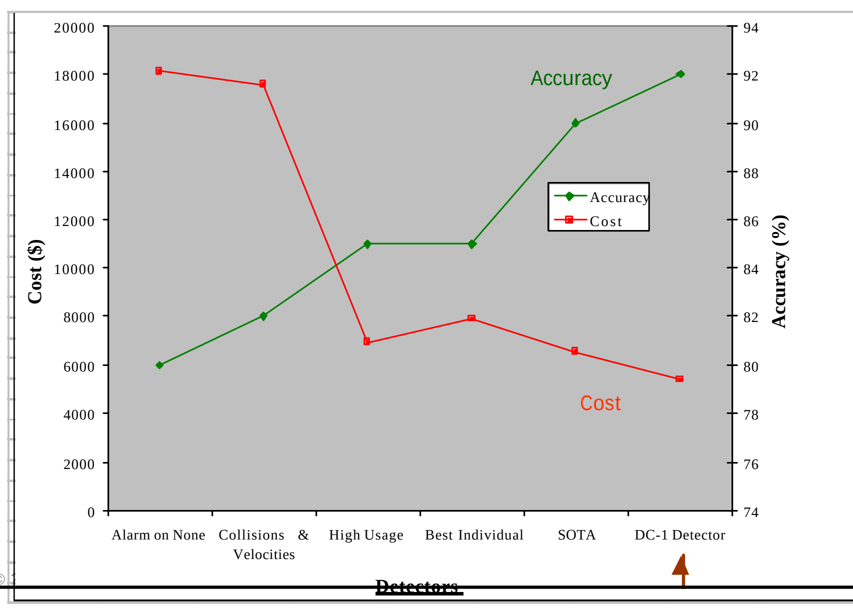
(Fawcett & Provost 1996, 1997)



© 2003 Foster Provost 12

Example: fraud detection

Evaluation of fraud detection methods



13

What “new” research issues emerge?

Today I’ll discuss two groups of issues, and associated research problems

- Costs
  – many different costs to consider
  – see (Turney, 2000) for an excellent overview
- Relational data
  – networked data
  – multi-table RDBs

© 2003 *Foster Provost*

What “new” research issues emerge?

Costs & machine learning
- e.g., Workshop @ ICML 2000

- Different misclassification costs
  – false-positive vs. false-negative errors
- Misclassification error cost varies case-by-case
  – e.g., (Zadrozny & Elkan ICML 2001, etc.) (Bennett, SIGIR 2003)
- Highly unbalanced class distribution
  – Workshops @ AAAI 2000, ICML 2003
- <u>Don’t know</u> target costs & class distributions
  – (Provost & Fawcett, KDD-97, MLJ 2001) (Latinne et al., ICML 2001)
- Procuring training data costly
  - budget-sensitive learning?
  – e.g., what proportion of positives and negatives for training?

© 2003 *Foster Provost*

Given an example procurement budget, what class distribution should be used?

- What does the manual tell us to do?

"if the sample size is fixed, a balanced sample will usually produce more accurate predictions than an unbalanced (one)" (SAS, 2001)

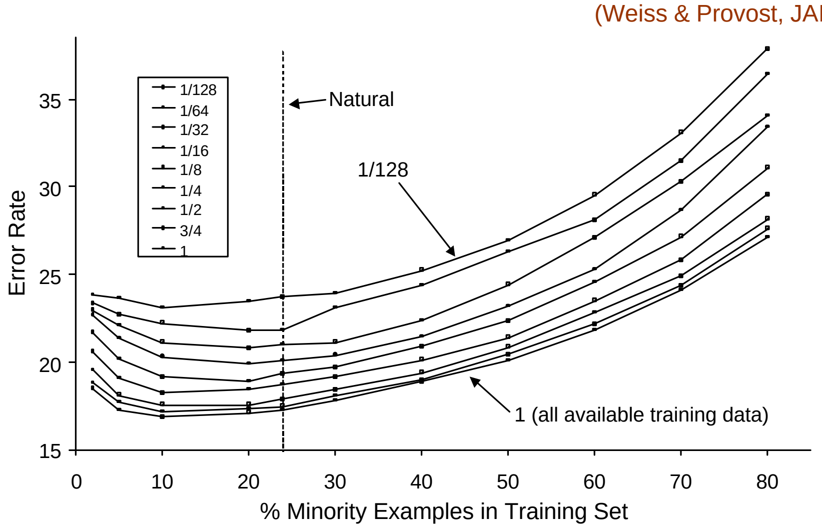
- General recommendations (based on C4.5):
  - For accuracy use natural distribution
    - but could do better
    - in 9 of 26 cases get burnt (with confidence)
    - *(balanced sample not recommended!)*
  - For AUC use balanced distribution
    - but could do better
    - in 7 of 26 cases get burnt (with confidence)

© 2003 Foster Provost

(Weiss & Provost, JAIR 2003) 16

Relative performance of different class distributions is remarkably consistent across training-set sizes

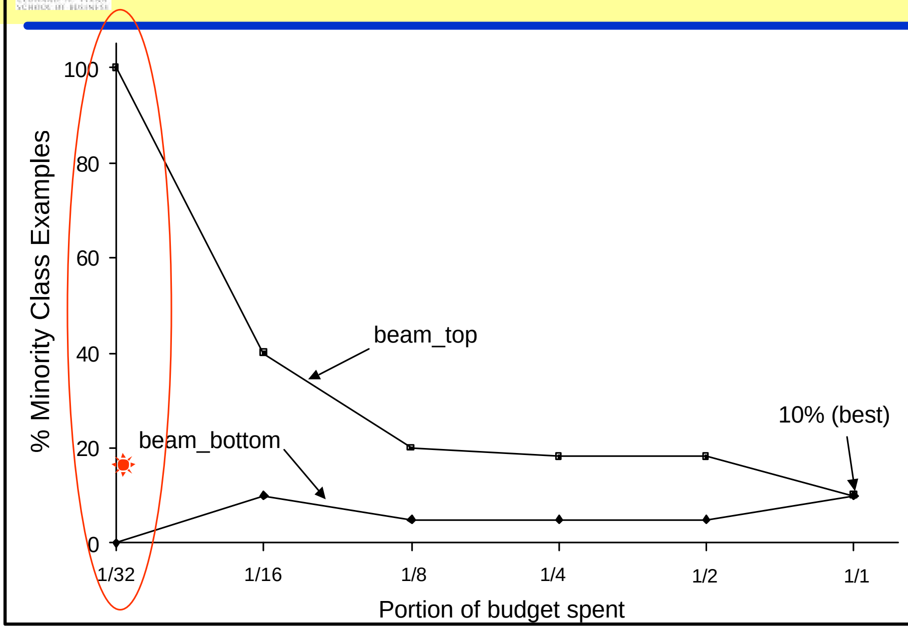
(Weiss & Provost, JAIR 2003)



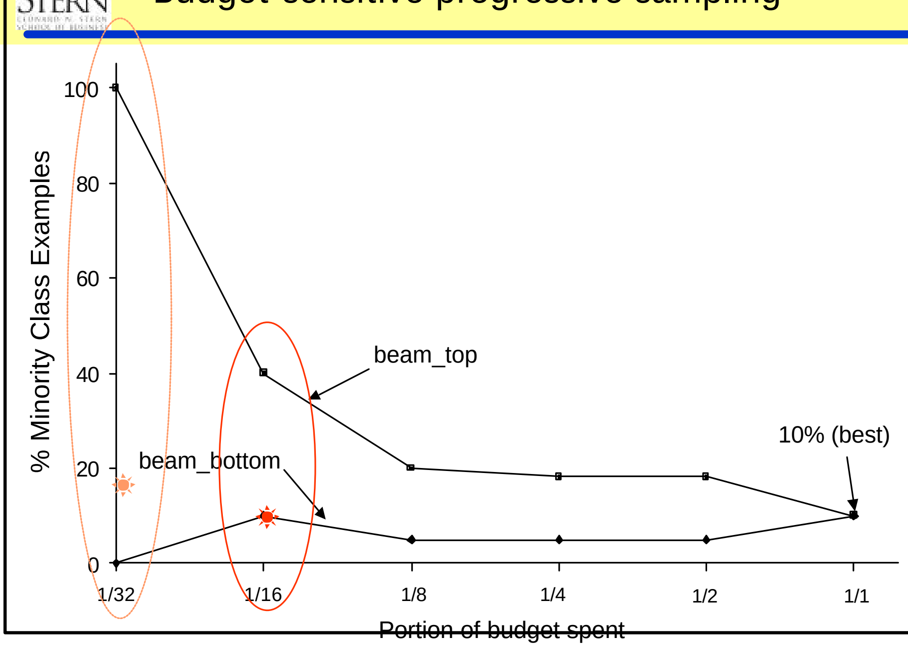
Adult Data Set

© 2003 Foster Provost 17

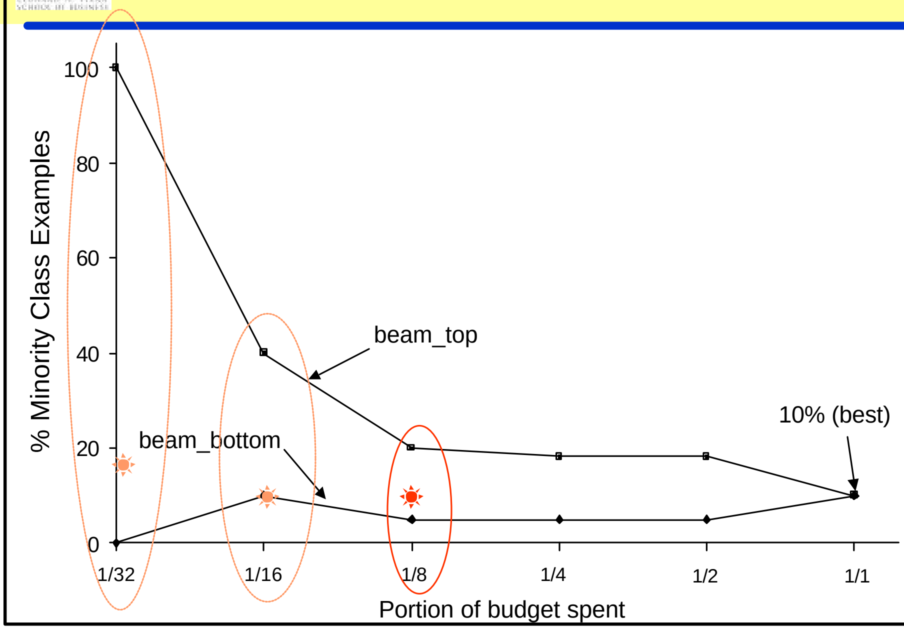
Budget-sensitive progressive sampling



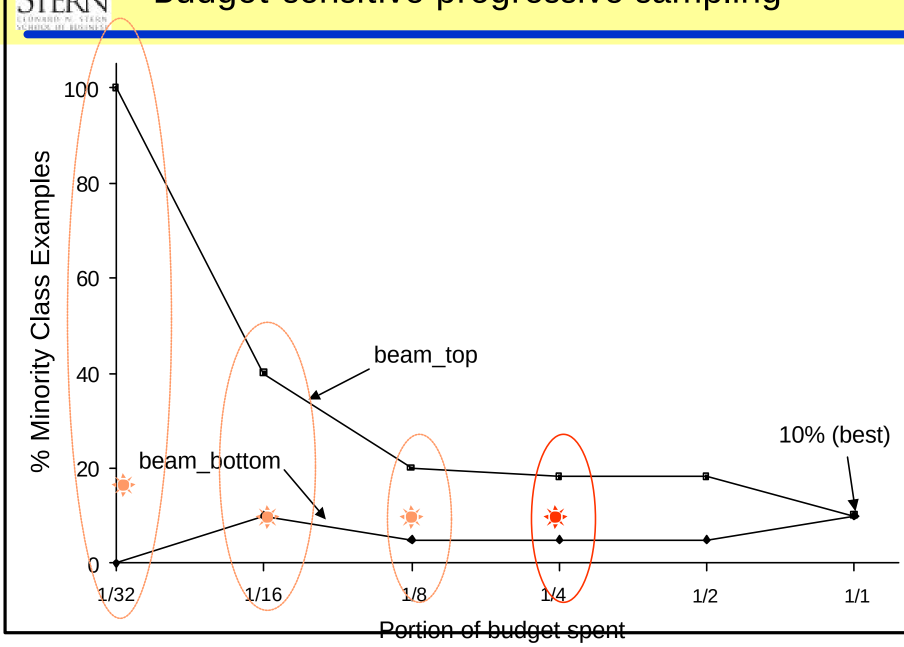
Budget-sensitive progressive sampling



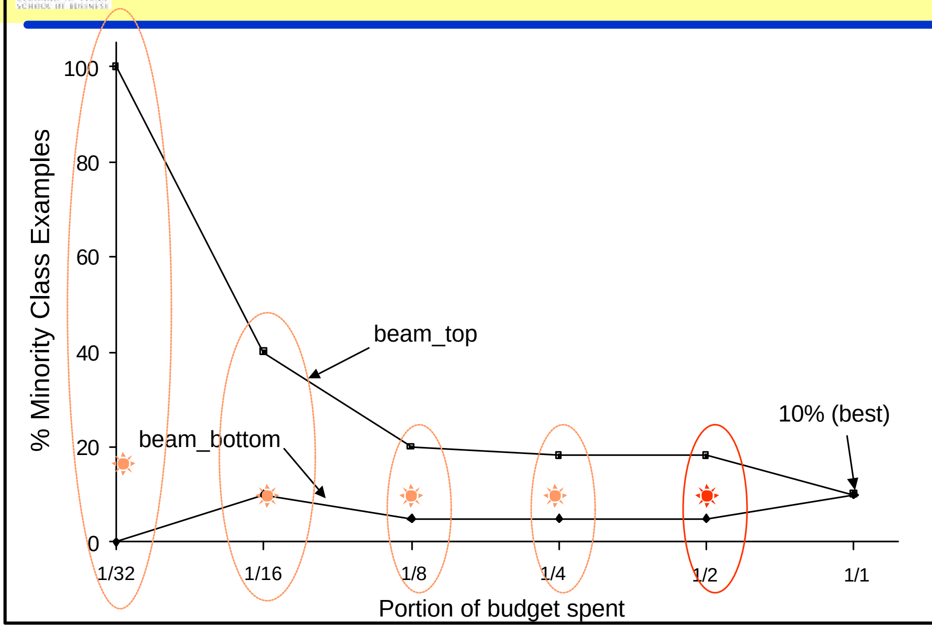
Budget-sensitive progressive sampling



Budget-sensitive progressive sampling

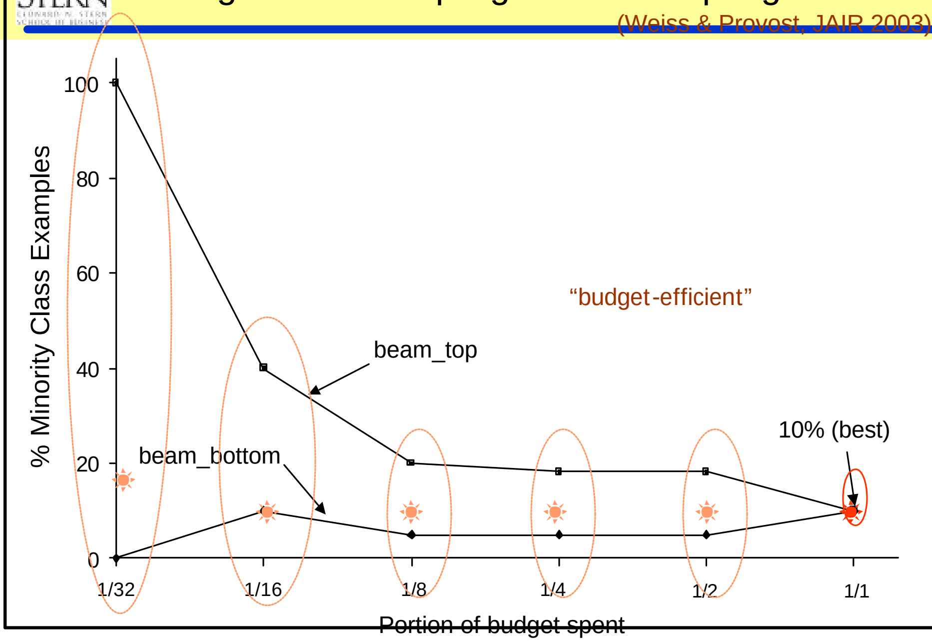


Budget-sensitive progressive sampling



Budget-sensitive progressive sampling

(Weiss & Provost, JAIR 2003)



What "new" research issues emerge?

Costs & machine learning

- e.g., Workshop @ ICML 2000

- Different misclassification costs
  - false-positive vs. false-negative errors
- Misclassification error cost varies case-by-case
  - e.g., (Zadrozny & Elkan ICML 2001, etc.) (Bennett, SIGIR 2003)
- Highly unbalanced class distribution
  - Workshops @ AAAI 2000, ICML 2003
- Don't know target costs & class distributions
  - (Provost & Fawcett, KDD-97, MLJ 2001) (Latinne et al., ICML 2001)
- Procuring training data costly
  - budget-sensitive learning
  - e.g., what proportion of positives and negatives for training?

© 2003 Foster Provost

What "new" research issues emerge?

- Learning from relational data

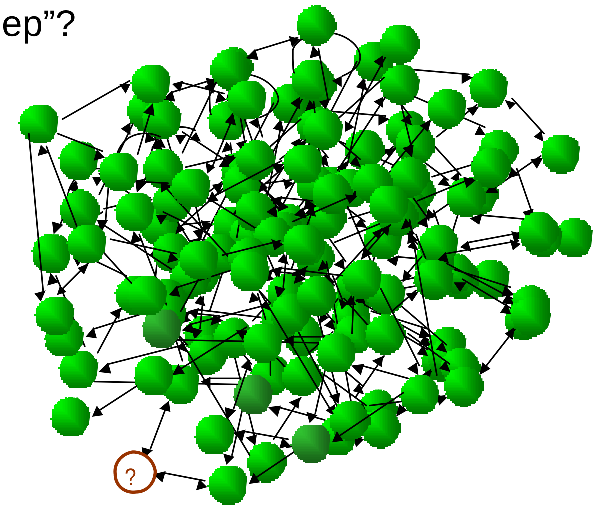
  Recently:
  - SRL Workshops @ AAAI 2000, IJCAI 2003
  - MRDM Workshops @ KDD 2002, KDD 2003

  - Networked data
  - Multirelational databases

© 2003 Foster Provost

Mining networked data

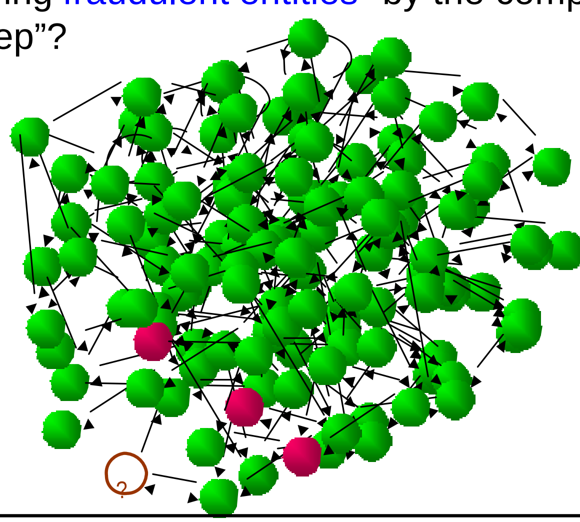
- Classifying fraudulent entities “by the company they keep”?



© 2003 Foster Provost

Mining networked data

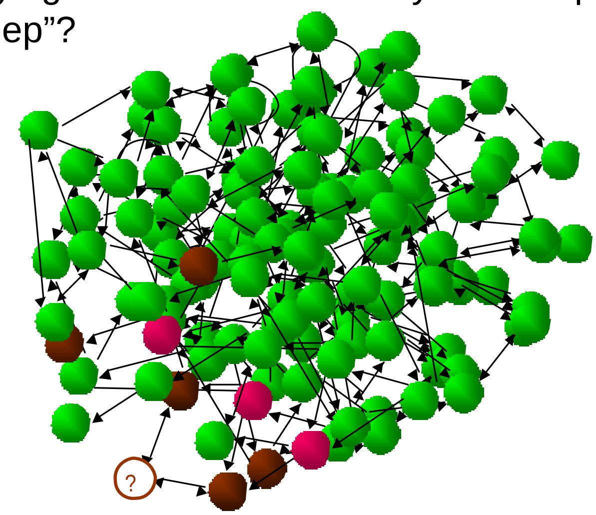
- Classifying fraudulent entities “by the company they keep”?



© 2003 Foster Provost

Mining networked data

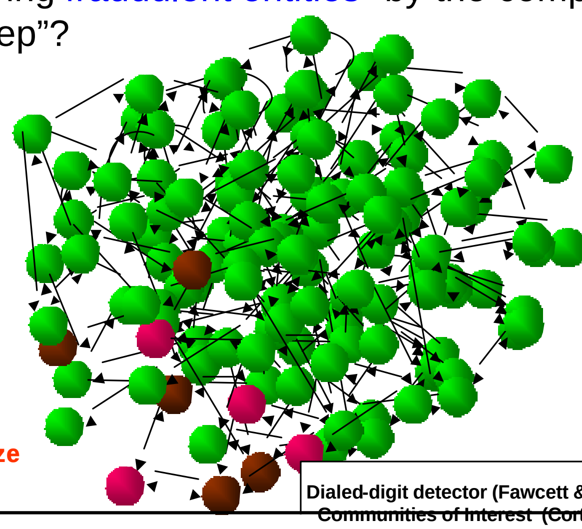
- Classifying fraudulent entities "by the company they keep"?



© 2003 *Foster Provost*

Mining networked data

- Classifying fraudulent entities "by the company they keep"?



Can we generalize this idea?

© 2003 *Foster Provost*

Dialed-digit detector (Fawcett & Provost, 1997)
Communities of Interest (Cortes et al. 2001)

A relational-neighbor classifier?

Is there theoretical justification?

Thanks to (McPherson, et al., 2001)

- *Birds of a feather, flock together*
  – attributed to Robert Burton (1577-1640)
- (*People) love those who are like themselves*
  -- Aristotle, *Rhetoric* and *Nichomachean Ethics*
- *Similarity begets friendship*
  -- Plato, *Phaedrus*
- *Hanging out with a bad crowd will get you into trouble*
  -- Mom

© 2003 Foster Provost

Relational neighbors and beyond

- Homophily
  - fundamental concept underlying social theories (e.g., Blau 1977)
  - one of the first features noticed by analysts of network structure
    - antecedents to SNA research from 1920's (Freeman 1996)
  - fundamental basis for links of many types in social networks
    (McPherson, et al., Annu. Rev. Soc. 2001) ←read this!
    Patterns of homophily:
    - *remarkably robust across widely varying types of relations*
    - *tend to get stronger as more relationships exist*
    - likely to be present for non-person entities?

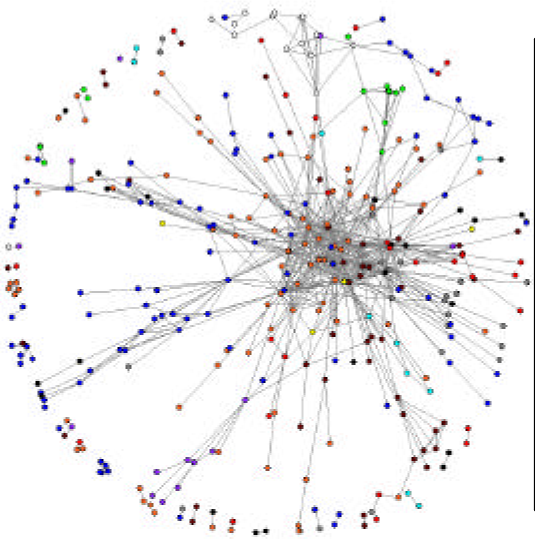
→ *Simple relational-neighbor classifier may be effective*

- *A baseline against which we should compare relational learning algorithms?*
- (More complex structure can be learned)
  - e.g., see Daphne's talk on Sunday

© 2003 Foster Provost

Classifying firms "by the company they keep" ...

Bernstein et al. 2003



Sector	Accuracy
Technology (orange)	0.78
Energy (purple)	0.76
Transportation (green)	0.74
Healthcare (red)	0.73
Utilities (white)	0.70
Financial (brown)	0.65
Services (blue)	0.64
ConsumerNonCyc (l.blue)	0.63
BasicMaterials (mag.)	0.45
ConsumerCyclical (black)	0.36
CapitalGoods (cyan)	0.23
Conglomerates (yellow)	0.21
Overall	**0.65**
base rate (Technology)	*0.28*

A "Relational Neighbor" (k-RN) Algorithm

Classifying movies by the company they keep...

Prior results: predicting movie success (AUC)

Relational Trees: 0.82

Relational Bayes: 0.85

IMDB data, using attributes on related entities (e.g., the most prevalent genre of a movie's studio).

RN Classifier:

RN link type	AUC	stdev
actor	0.766	0.003
director	0.658	0.007
producer	0.850	0.005
production company	0.862	0.003

© 2003 Foster Provost

Iterative/collective classification using belief propagation

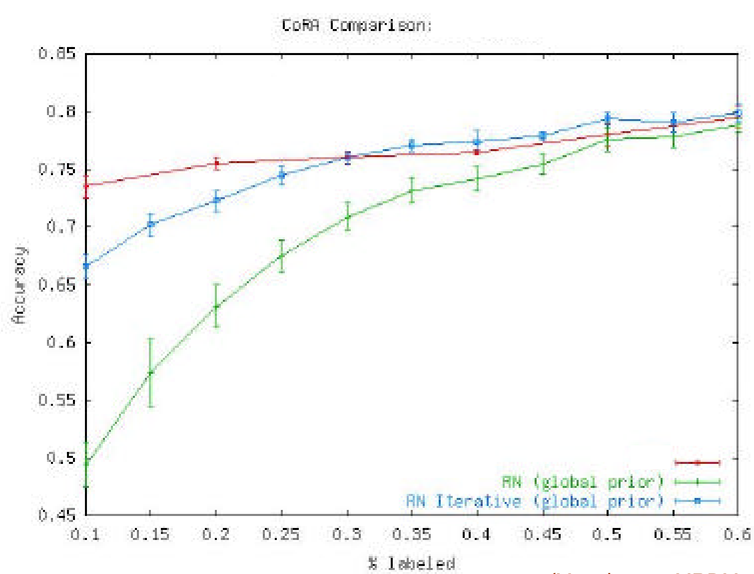
Idea: *belief about class membership can be propagated through a network of linked examples, to help to classify unlabeled/uncertain examples.*

– Probabilistic belief propagation (e.g., Pearl, 1988)
– Relaxation labeling (e.g., Chakrabarti et al., SIGMOD 1998)
– PRMs (e.g., Koller et al. AAAI-98, IJCAI-99, ICML-01, IJCAI-01, UAI-02, etc.)
– Iterative classification (Neville & Jensen, 2000, Jensen et al. KDD-2003)
– Mining viral marketing (Domingos & Richardson, KDD-2001)

- *Simple* belief-propagation version of RN?
  - Useful when:
    - few labeled data
    - linked to unlabeled data
  - *A baseline against which we should compare relational learning algorithms?*
    - (Macskassy, MRDM -2003 next week)

© 2003 *Foster Provost*

Classifying research papers by the company they keep...



© 2003 Foster Provost
(Macskassy, MRDM -2003 next week)

... by the company they keep...

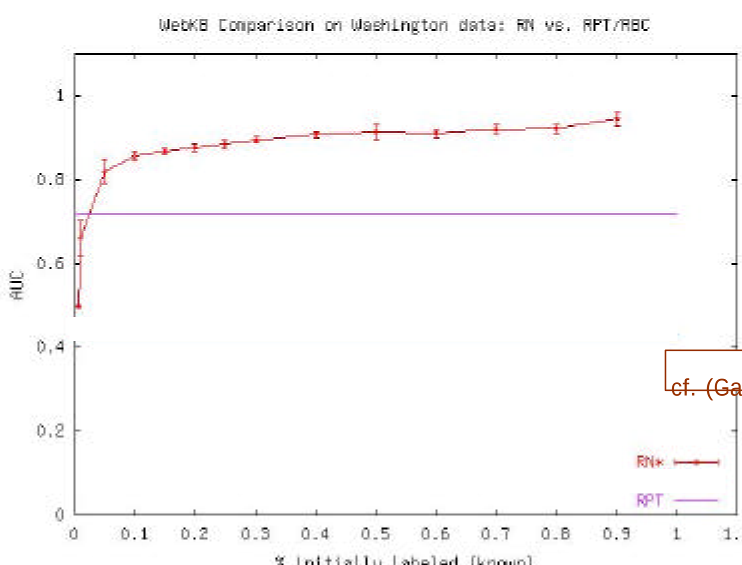
Mini-summary:

- people
- companies
- movies
- research papers
- …

Simple relational classifiers should be used as baselines in relational learning studies

© 2003 Foster Provost

Classifying web pages by the company they keep...



cf. (Gaines, ICML 1989)

(Macskassy, MRDM -2003 next week)

What “new” research issues emerge?

Today I’ll discuss two groups of issues, and associated research problems

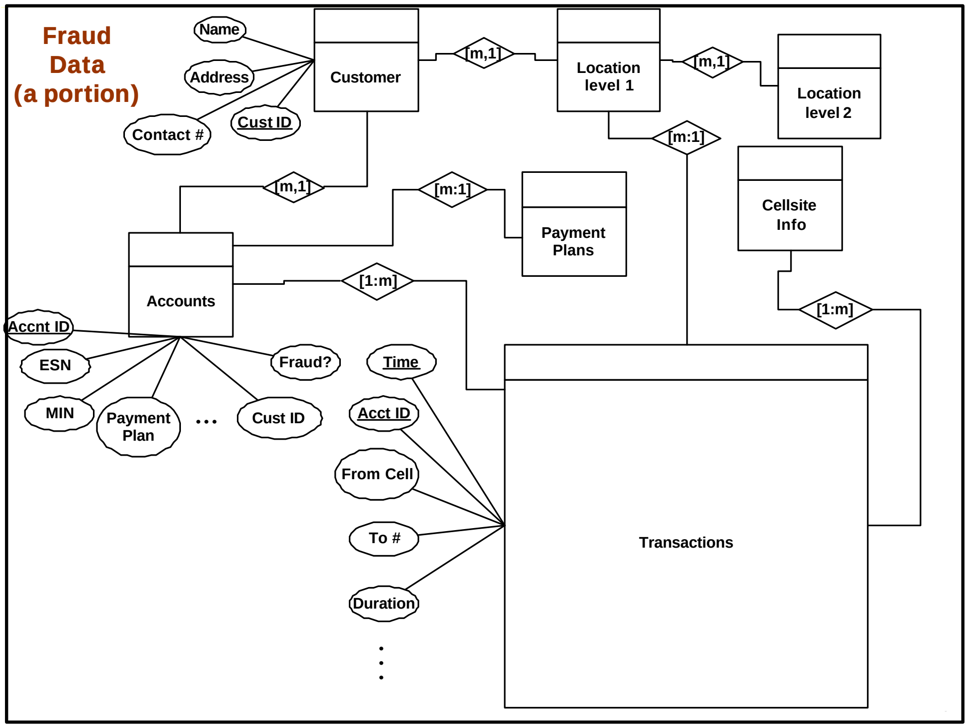
- Costs
  - many different costs to consider
  - see (Turney, 2000) for an excellent overview
- Relational data
  - networked data
  - multi-table RDBs

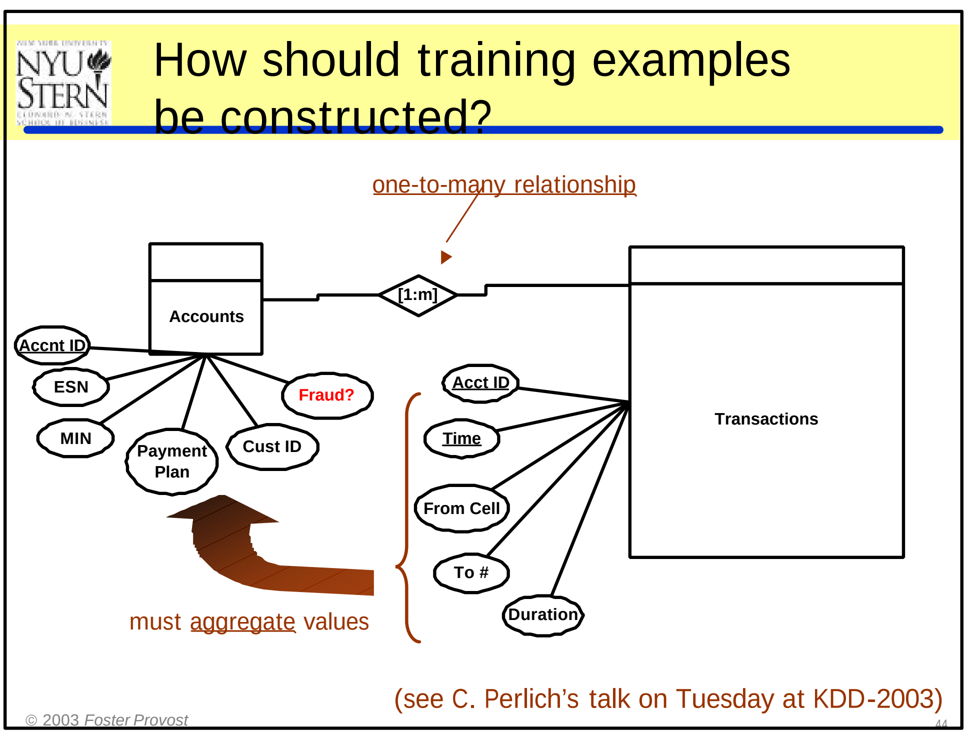
© 2003 Foster Provost

Data – traditional ML view

```
b,23.92,0.665,u,g,c,v,0.165,f,f,0,f,g,00100,0,+
a,25.75,0.5,u,g,c,h,0.875,t,f,0,t,g,00491,0,+
b,26.00,1,u,g,q,v,1.75,t,f,0,t,g,00280,0,+
b,37.42,2.04,u,g,w,v,0.04,t,f,0,t,g,00400,5800,+
b,34.92,2.5,u,g,w,v,0,t,f,0,t,g,00239,200,+
b,34.25,3,u,g,cc,h,7.415,t,f,0,t,g,00000,0,+
b,23.33,11.625,y,p,w,v,0.835,t,f,0,t,g,00160,300,+
b,23.17,0,u,g,cc,v,0.085,t,f,0,f,g,00000,0,+
b,44.33,0.5,u,g,i,h,5,t,f,0,t,g,00320,0,+
b,35.17,4.5,u,g,x,h,5.75,f,f,0,t,s,00711,0,+
b,43.25,3,u,g,q,h,6,t,t,11,f,g,00080,0,+
b,56.75,12.25,u,g,m,v,1.25,t,t,04,t,g,00200,0,+
b,31.67,16.165,u,g,d,v,3,t,t,09,f,g,00250,730,+
a,23.42,0.79,y,p,q,v,1.5,t,t,02,t,g,00080,400,+
a,20.42,0.835,u,g,q,v,1.585,t,t,01,f,g,00000,0,+
b,26.67,4.25,u,g,cc,v,4.29,t,t,01,t,g,00120,0,+
b,34.17,1.54,u,g,cc,v,1.54,t,t,01,t,g,00520,50000,+
a,36.00,1,u,g,c,v,2,t,t,11,f,g,00000,456,+
b,25.50,0.375,u,g,m,v,0.25,t,t,03,f,g,00260,15108,+
b,19.42,6.5,u,g,w,h,1.46,t,t,07,f,g,00080,2954,+
b,35.17,25.125,u,g,x,h,1.625,t,t,01,t,g,00515,500,+
b,32.33,7.5,u,g,e,bb,1.585,t,f,0,t,s,00420,0,-
b,34.83,4,u,g,d,bb,12.5,t,f,0,t,g,?,0,-
a,38.58,5,u,g,cc,v,13.5,t,f,0,t,g,00980,0,-
b,44.25,0.5,u,g,m,v,10.75,t,f,0,f,s,00400,0,-
b,44.83,7,y,p,c,v,1.625,f,f,0,f,g,00160,2,-
b,20.67,5.29,u,g,q,v,0.375,t,t,01,f,g,00160,0,-
b,34.08,6.5,u,g,aa,v,0.125,t,f,0,t,g,00443,0,-
```

© 2003 Foster Provost


Fraud
Data
(a portion)
Name
Address
Cust ID
Contact #
Customer
[m,1]
Location level 1
[m,1]
Location level 2
[m:1]
[m,1]
[m:1]
Payment Plans
Cellsite Info
Accounts
[1:m]
[1:m]
Accnt ID
ESN
MIN
Payment Plan
Cust ID
Fraud?
Time
Acct ID
From Cell
To #
Duration
Transactions

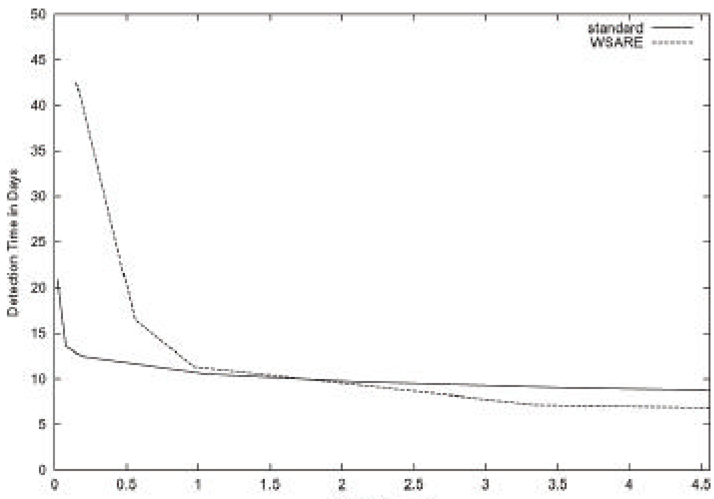

How should training examples be constructed?
one-to-many relationship
Accounts
[1:m]
Transactions
Accnt ID
ESN
MIN
Payment Plan
Cust ID
Fraud?
Acct ID
Time
From Cell
To #
Duration
must aggregate values
(see C. Perlich's talk on Tuesday at KDD-2003)
© 2003 Foster Provost

Future "new" research areas?

- Comprehensibility!
  (Pazzani, *IEEE Intelligent Systems*, 2000)
- Drift
  - concept (e.g., Lane & Brodley, KDD-1998; Bartlett et al. MLJ 2000)
  - population (e.g., Bartlett, CoLT-92; Long, CoLT-98)
- Activity monitoring
  - timeliness of detection vital
    - seldom taken into account by ML studies
    - (Weiss & Hirsh, KDD-98)
      - network monitoring
    - (Fawcett & Provost KDD-99)
  - biosurveillance …

© 2003 Foster Provost

ML for biosurveillance



(Wong, Moore, Cooper, Wagner, *J. Urban Health* 2003)

Sunday Afternoon: (Wong, Moore, Cooper, Wagner, ICML-2003)

© 2003 Foster Provost

Summary

Applications:

- help to stimulate researchers
  - and attract graduate students
- help to convince funding sources
- provide tests of prior research results
  - of relevance
  - of necessity
  - of efficacy
- highlight insufficiencies of the state of the art
  - →pointing to important areas for future research

"In the meantime, while the creative power of pure reason is at work, *the outer world* again comes into play, forces upon us new questions from actual experience, opens up new branches [of the science]"

David Hilbert, 1900

© 2003 *Foster Provost*

On writing, reviewing, editing ...

- Keep complexity of applications in mind
  - must ignore certain things
  - but can't "assume" them away
- Keep "spectrum" in mind
  - different sorts of contributions
- Have a clear focus
  - don't try to sneak your new X algorithm through a back door
    - where X is a well-known problem
- Argue for generality
- Don't forget the basics
  - what are the claimed contributions?
    - novel? important?
  - how are they supported?
  - how does the work fit in the context of prior work?
  - be scholarly

© 2003 *Foster Provost*

Potential contributions?

- In-depth descriptions of (new) practically important problems
  - that can't be solved with existing methods
- A solid case study
  - Feedback on the (lack of) utility of prior results
  - Areas of weakness in our knowledge
- A new benchmark data set
- Good novel algorithms/methods
  - maybe ad hoc
  - potential starting point for future studies

© 2003 *Foster Provost*

Summary

Applications:

- help to stimulate researchers
  - and attract graduate students
- help to convince funding sources
- provide tests of prior research results
  - of relevance
  - of necessity
  - of efficacy
- highlight insufficiencies of the state of the art
  - →pointing to important areas for future research

"In the meantime, while the creative power of pure reason is at work, *the outer world* again comes into play, forces upon us new questions from actual experience, opens up new branches [of the science]"
David Hilbert, 1900

© 2003 *Foster Provost*

References 1

- Bartlett, P. (1992). Learning with a slowly changing distribution. In *Proceedings of the 1992 Workshop on Computational Learning Theory*, 243-252.
- Bartlett, P., S. Ben-David, and S. Kulkarni (2000). Learning changing concepts by exploiting the structure of change. *Machine Learning* 41, 153-174, 2000.
- Bennett, P. (2003). Using Asymmetric Distributions to Improve Text Classifier Probability Estimates. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, July-August 2003. ACM Press.
- Bernstein, A., S. Clearwater, and F. Provost (2003). The Relational Vector-space Model and Industry Classification. In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data.*
- Blau, P. (1977). *Inequality and Heterogeneity: A Primitive Theory of Social Structure.* New York: Free Press.
- Chakrabarti, S., B. Dom, and P. Indyk (1998). Enhanced hypertext categorization using hyperlinks. In *Proceedings of the 1998 ACM International Conference on Management of Data* (SIGMOD'98), June 1998.
- Cortes, C., D. Pregibon, C. Volinsky (2001). Communities of Interest. In *Proceedings of IDA (Intelligent Data Analysis) 2001*: 105-114.
- Domingos, P. and M. Richardson (2001). Mining the Network Value of Customers. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining* (KDD-2001), 57-66.

© 2003 *Foster Provost*

References 2

- Fawcett, T. and F. Provost (1996). Combining Data Mining and Machine Learning for Effective User Profiling. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (KDD-96).
- Fawcett, T. and F. Provost (1997). Adaptive Fraud Detection. *Journal of Data Mining and Knowledge Discovery* 1 (3) 1997.
- Fawcett, T. and F. Provost (1999). Activity Monitoring: Noticing Interesting Changes in Behavior. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining* (KDD-99).
- Freeman, L. (1996). Some antecedents of social network analysis. *Connections* 19:39-42.
- Friedman, N., L. Getoor, D. Koller and A. Pfeffer (1999). Learning Probabilistic Relational Models. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence* (IJCAI), Stockholm, Sweden, August 1999, pages 1300--1307.
- Gaines, B. (1989). An Ounce of Knowledge is Worth a Ton of Data: Quantitative Studies of the Trade-off between Expertise and Data based on Statistically Well-founded Empirical Induction, In *Proceedings of the Sixth International Workshop on Machine Learning*, pp.156-159.
- Getoor, L., N. Friedman, D. Koller, and B. Taskar (2001). Learning Probabilistic Models of Relational Structure. In *Proceedings of the Eighteenth International Conference on Machine Learning* (ICML), Williams College, June 2001.

© 2003 *Foster Provost*

References 3

- Hilbert, D. (1900). Lecture delivered before the International Congress of Mathematicians at Paris.
- Jensen, D., M. Rattigan, and H. Blau (2003). Information Awareness: A Prospective Technical Assessment. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*
- Koller, D. and A. Pfeffer (1998). Probabilistic frame-based systems. *Proceedings of the 15th National Conference on Artificial Intelligence* (AAAI), Madison, Wisconsin, July 1998, pages 580--587.
- Lane, T, and C. Brodley (1998). Approaches to Online Learning and Concept Drift for User Identification in Computer Security. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (KDD-98), pp 259-263. 1998.
- Latinne, P., M. Saerens, C. Decaestecker (2001). Adjusting the Outputs of a Classifier to New a Priori Probabilities May Significantly Improve Classification Accuracy: Evidence from a multi-class problem in remote sensing. In *Proceedings of the Eighteenth International Conference on Machine Learning* (ICML-2001), 298-305.
- Lim, T.-S., W.-Y. Loh and Y.-S. Shih (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, *Machine Learning* 40, 203-228.
- Long, P. (1998). The complexity of learning according to two models of a drifting environment. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 116-125.

© 2003 *Foster Provost*

References 4

- McPherson, M., L. Smith-Lovin and J. Cook. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Socialogy,* 27:415-444:2001.
- Neville, J. and D. Jensen (2000). Iterative Classification in Relational Data. In *Papers of the AAAI-2000 Workshop on Learning Statistical Models from Relational Data.* AAAI Press.
- Pazzani, M. (2000). Knowledge discovery from data? *IEEE Intelligent Systems* 15(2): 10-13.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufman Publishers, San Francisco.
- Perlich, C., F. Provost, and J. Simonoff (2003). Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. *Journal of Machine Learning Research* 4(Jun):211-255, 2003.
- Perlich, C. and F. Provost (2003). Aggregation-based feature invention and relational concept classes. In *Proceedings of the Ninth SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD-2003).
- Provost, F. and T. Fawcett, (1997). Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (KDD-97).
- Provost, F. and T. Fawcett, (2001). "Robust Classification for Imprecise Environments." *Machine Learning* 42, 203-231, 2001.

© 2003 *Foster Provost*

References 5

- SAS (2001). Getting started with Enterprise Miner. Cary, N.C.: SAS Institute, Inc..
- Taskar, B., E. Segal, and D. Koller (2001). Probabilistic Classification and Clustering in Relational Data. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, Seattle, Washington, August 2001, pages 870--876.
- Taskar, B., P. Abbeel and D. Koller (2003). Discriminative Probabilistic Models for Relational Data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* (UAI02), Edmonton, Canada, August 2002.
- Turney, P. (2000). Types of cost in inductive concept learning. In *Proceedings Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning* (WCSL at ICML-2000), pages 15-21.
- Weiss, G. and H. Hirsh (1998). Learning to Predict Rare Events in Event Sequences. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (KDD-98), 359-363.
- Weiss, G. and Foster Provost (2003). Learning when Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research* (in press).
- Wong, W.-K., A. Moore, G. Cooper, and M. Wagner (2003a). What's Strange About Recent Events. *Journal of Urban Health* 80, June 2003, pp. i66-i75.

© 2003 *Foster Provost*

References 6

- Wong, W.-K., A. Moore, G. Cooper, and M. Wagner (2003b). Bayesian Network Anomaly Pattern Detection for Disease Outbreaks. In *Proceedings of the Twentieth International Conference on Machine Learning* (ICML 2003).
- Zadrozny, B., and C. Elkan (2001a). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning* (ICML'01).
- Zadrozny, B. and C. Elkan (2001b). Learning and Making Decisions When Costs and Probabilities are Both Unknown. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining* (KDD'01).

© 2003 *Foster Provost*