Beyond Structured Prediction: Inverse Reinforcement Learning

Hal Daumé III

Computer Science
University of Maryland

me@hal3.name

A Tutorial at ACL 2011
Portland, Oregon

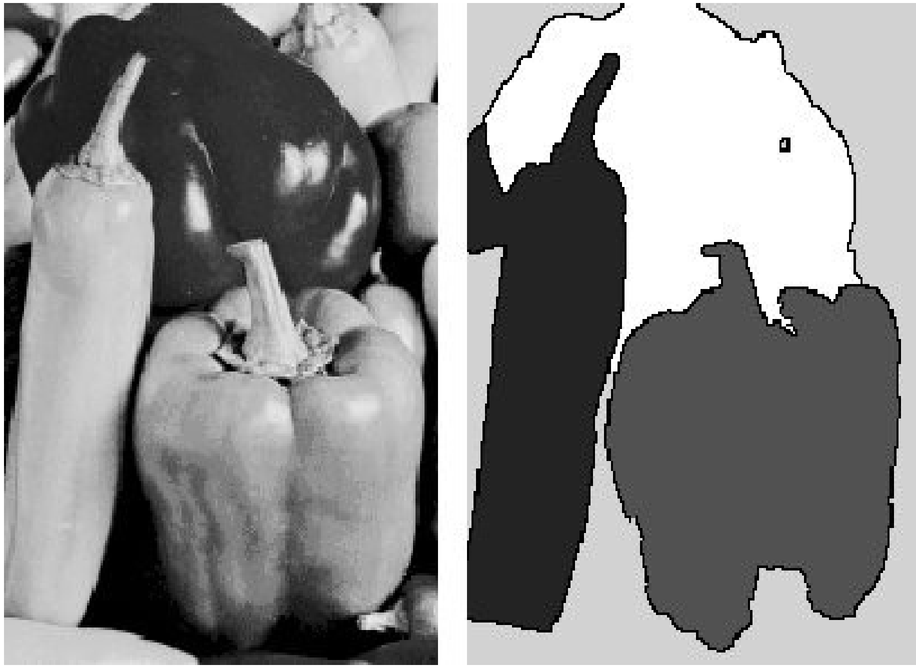
Sunday, 19 June 2011

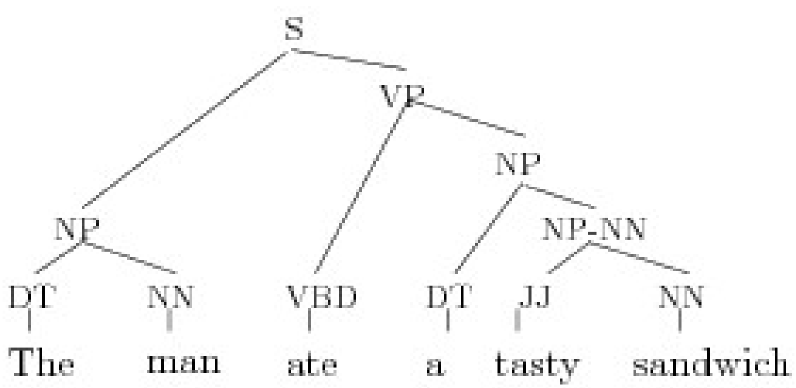
Acknowledgements

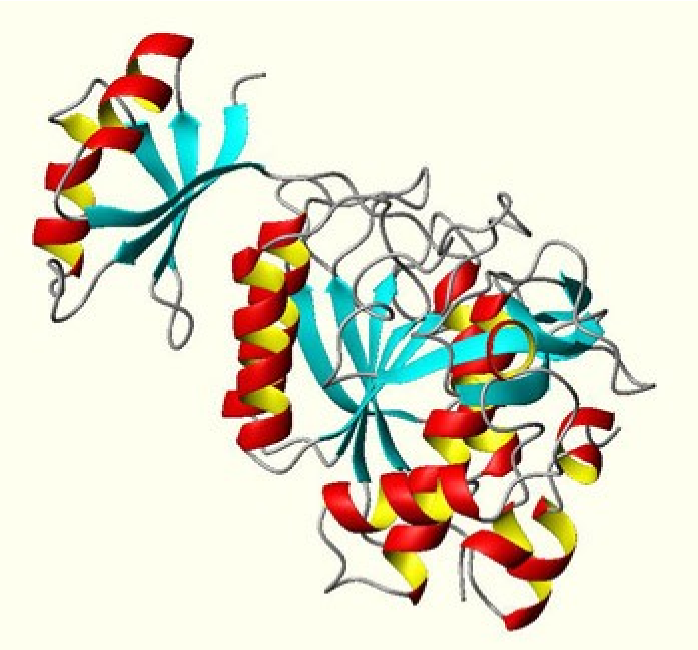
Some slides:
Stuart Russell
Dan Klein
J. Drew Bagnell
Nathan Ratliff
Stephane Ross

Discussions/Feedback:
MLRG Spring 2010

Examples of structured problems



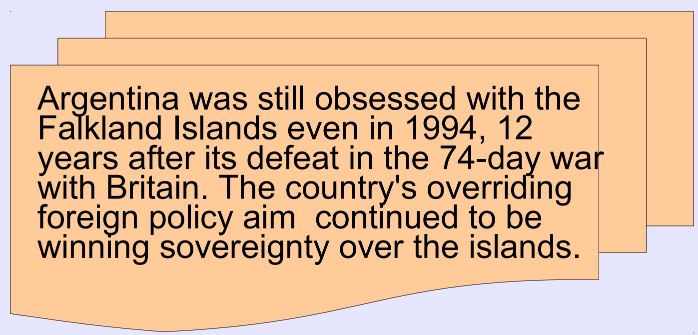
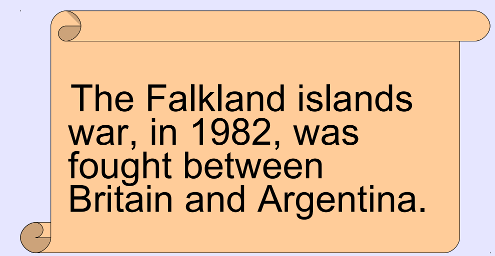
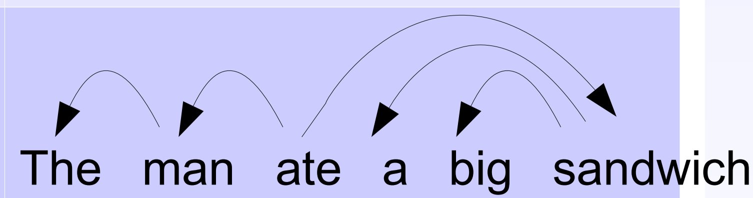




Examples of demonstrations

Examples of demonstrations

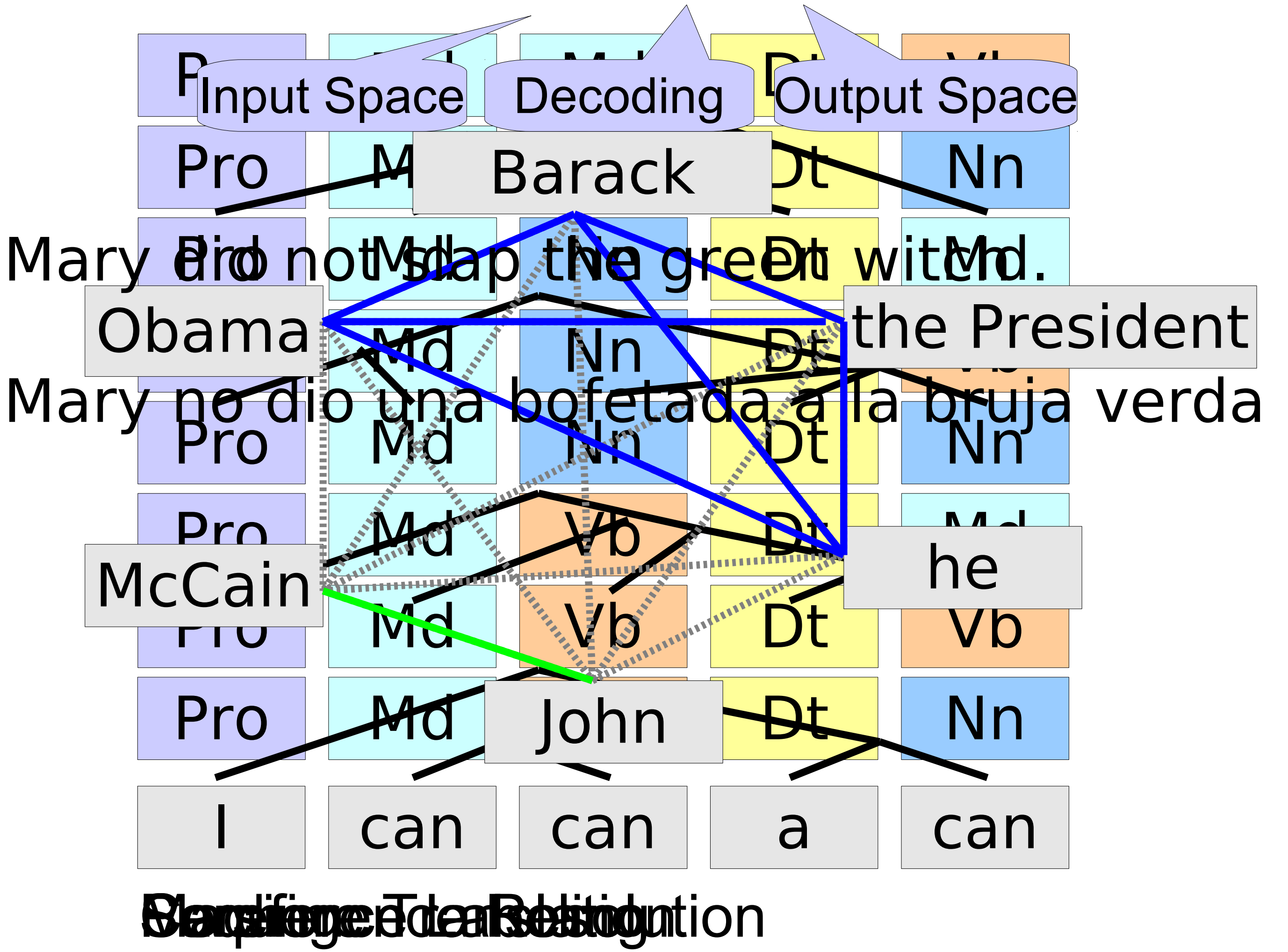
NLP as transduction

<table>
<tr><th>Task</th><th>Input</th><th>Output</th></tr>
<tr><td>Machine Translation</td><td>Ces deux principes se tiennent à la croisée de la philosophie, de la politique, de l’économie, de la sociologie et du droit.</td><td>Both principles lie at the crossroads of philosophy, politics, economics, sociology, and law.</td></tr>
<tr><td>Document Summarization</td><td>
</td><td>
</td></tr>
<tr><td>Syntactic Analysis</td><td>The man ate a big sandwich.</td><td>
</td></tr>
<tr><td>...many more...</td><td></td><td></td></tr>
</table>

Structured prediction 101

Learn a function mapping inputs to complex outputs:

$$f : X \rightarrow Y$$

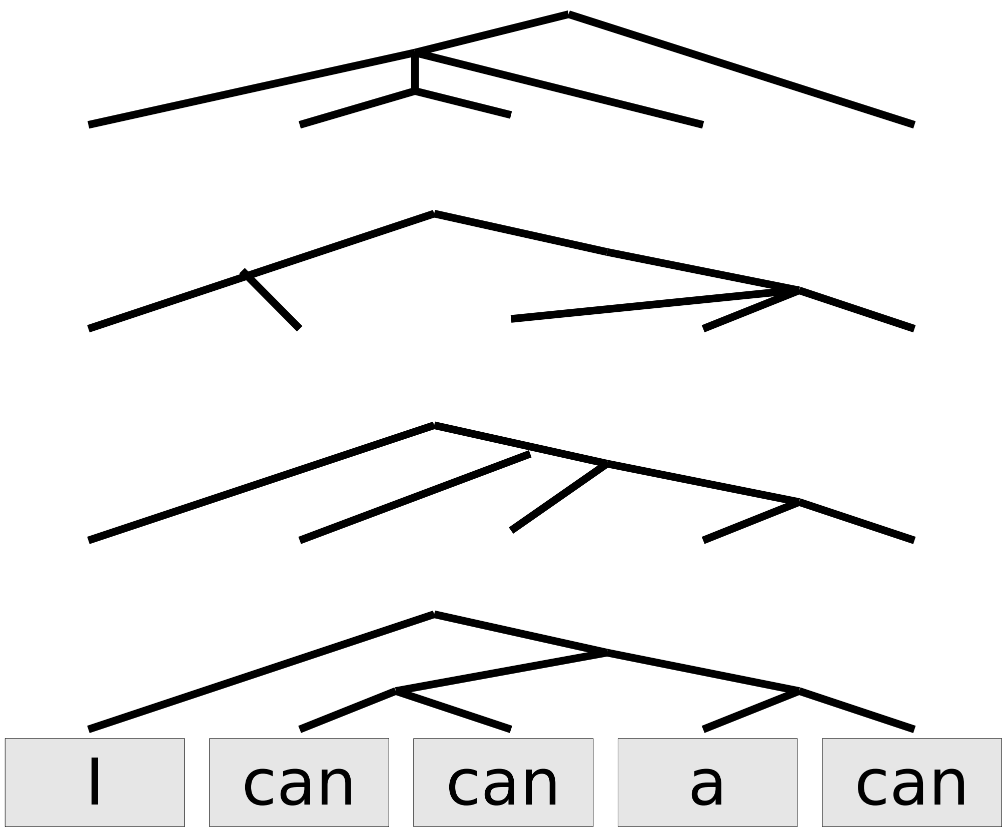


Structured prediction 101

Learn a function mapping inputs to complex outputs:

$$f : X \to Y$$

Input Space
Decoding
Output Space



Why is structure important?

- Correlations among outputs
  - Determiners often precede nouns
  - Sentences usually have verbs
- Global coherence
  - It just *doesn't make sense* to have three determiners next to each other
- My objective (aka "loss function") forces it
  - Translations should have good sequences of words
  - Summaries should be coherent

Outline: Part I

- What is Structured Prediction?
- Refresher on Binary Classification
  - What does it mean to learn?
  - Linear models for classification
  - Batch versus stochastic optimization
- From Perceptron to Structured Perceptron
  - Linear models for Structured Prediction
  - The “argmax” problem
  - From Perceptron to margins
- Structure without Structure
  - Stacking
  - Structure compilation

Outline: Part II

- Learning to Search
  - Incremental parsing
  - Learning to queue
- Refresher on Markov Decision Processes
- Inverse Reinforcement Learning
  - Determining rewards given policies
  - Maximum margin planning
- Learning by Demonstration
  - Searn
  - Dagger
- Discussion

Refresher on Binary Classification

What does it mean to learn?

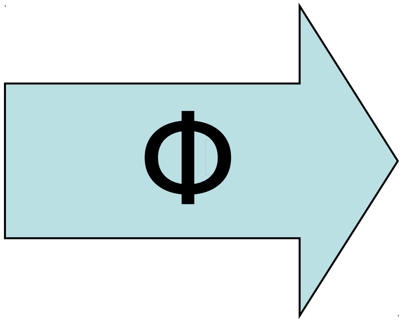
- Informally:
  - to predict the future based on the past
- Slightly-less-informally:
  - to take *labeled examples* and construct a function that will label them as a human would
- Formally:
  - Given:
    - A fixed unknown distribution D over X*Y
    - A loss function over Y*Y
    - A finite sample of (x,y) pairs drawn i.i.d. from D
  - Construct a function f that has low expected loss with respect to D

Feature extractors

- A feature extractor Φ maps examples to vectors

```
Dear Sir.

First, I must solicit
your confidence in
this transaction,
this is by virtue of
its nature as being
utterly confidencial
and top secret. …
```

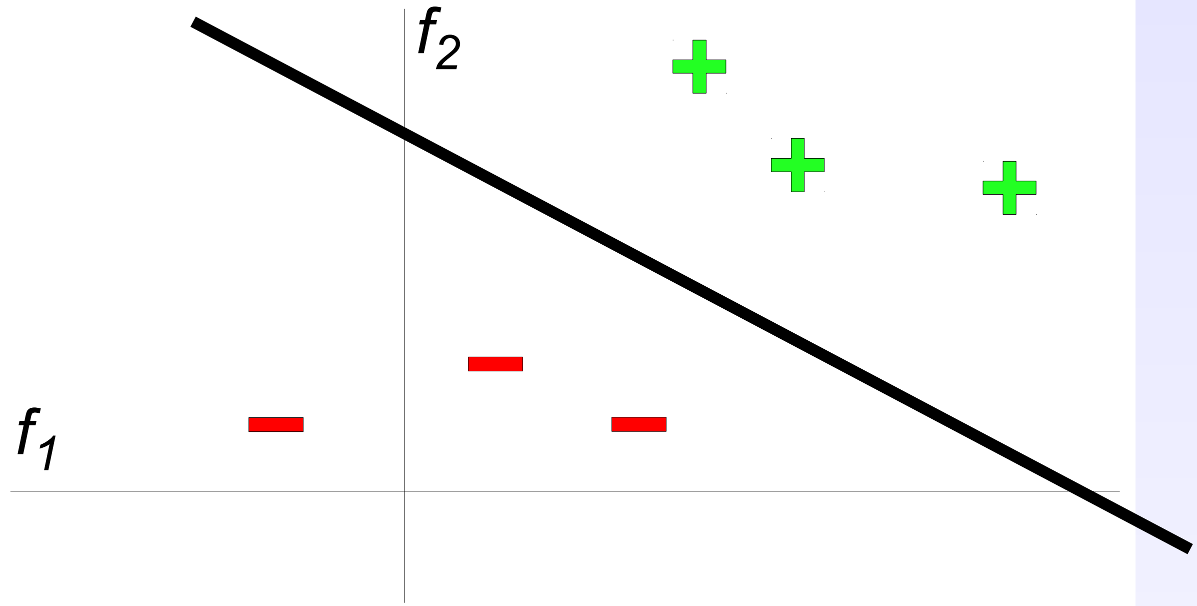



```
W=dear      :  1
W=sir       :  1
W=this      :  2
...
W=wish      :  0
...
MISSPELLED  :  2
NAMELESS    :  1
ALL_CAPS    :  0
NUM_URLS    :  0
...
```

- Feature vectors in NLP are frequently sparse

Linear models for binary classification

- Decision boundary is the set of “uncertain”points
- Linear decision boundaries are characterized by weight vectors

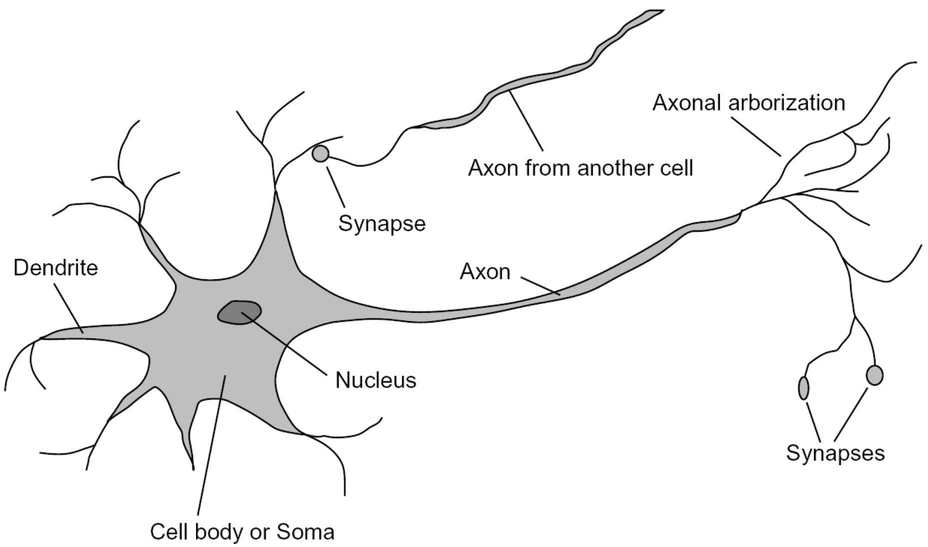


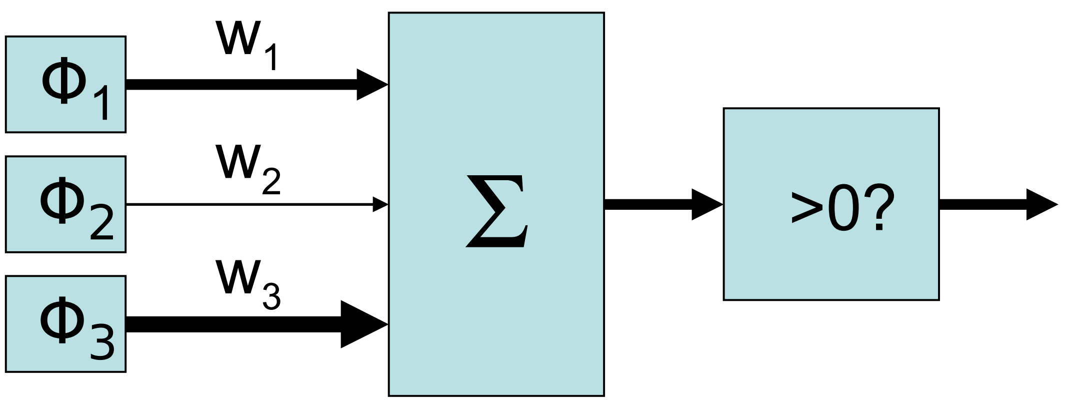
x	$\Phi(x)$	w	$\sum_i w_i \Phi_i(x)$
“free money”	BIAS : 1 free : 1 money : 1 the : 0 ...	BIAS : -3 free : 4 money : 2 the : 0 ...	$(1)(-3) +$ $(1)(4) +$ $(1)(2) +$ $(0)(0) +$ ... $= 3$

The perceptron

- Inputs = feature values
- Params = weights
- Sum is the response

- If the response is:
  - Positive, output +1
  - Negative, output -1





- When training, update on errors:

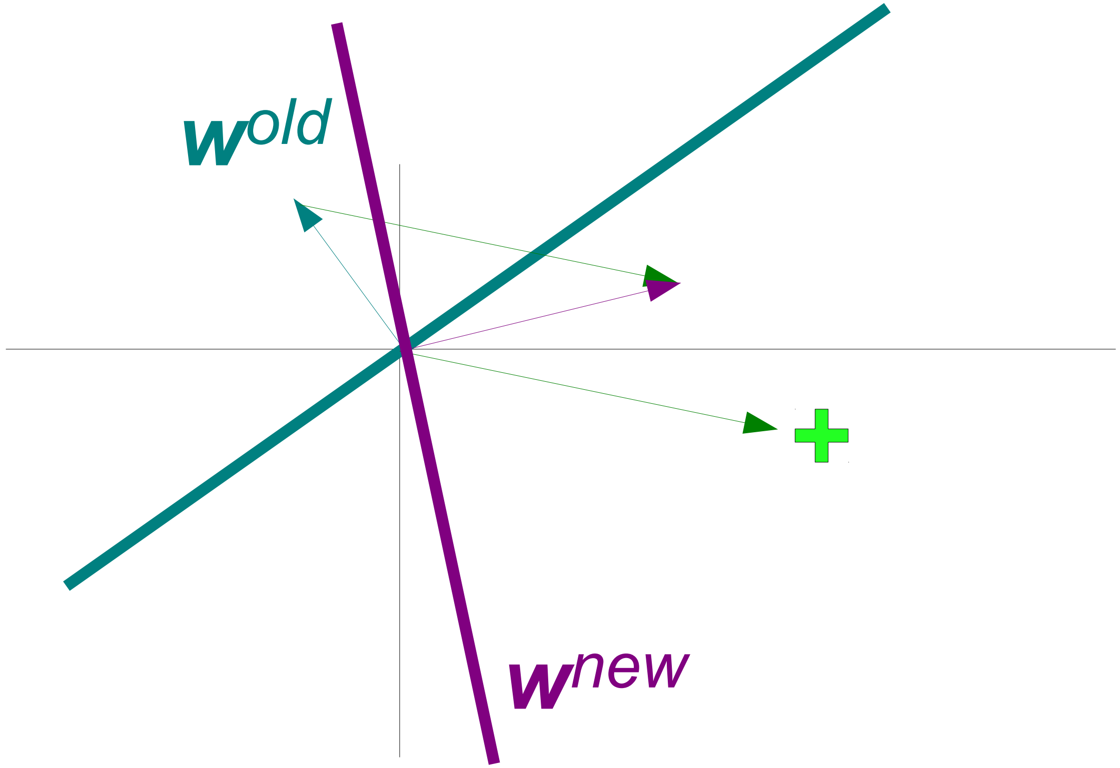
$$\mathbf{w} = \mathbf{w} + y\,\phi(x)$$

"Error" when:

$$y\,\mathbf{w}\cdot\phi(x) \leq 0$$

Why does that update work?

- When $y\,\mathbf{w}^{old}\cdot\phi(x)\leq 0$, update: $\mathbf{w}^{new}=\mathbf{w}^{old}+y\,\phi(x)$

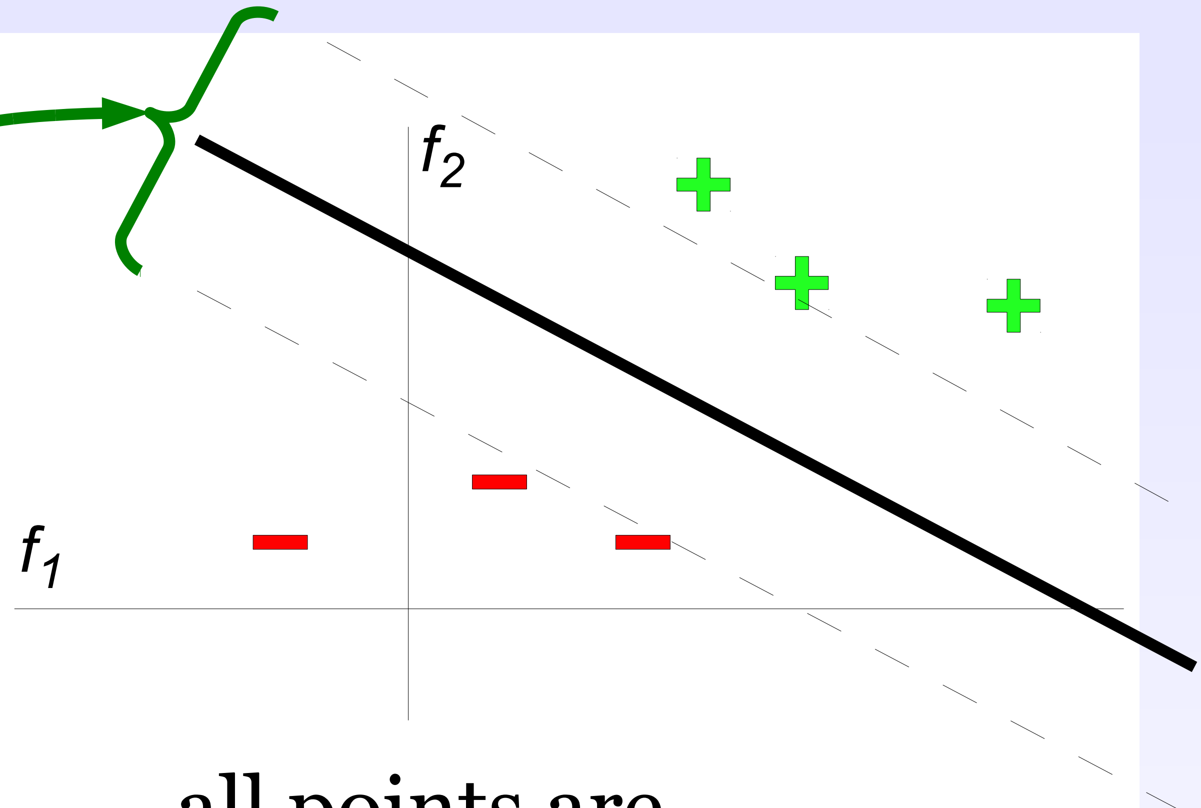


$$y\,\mathbf{w}^{new}\phi(x)=y\left(\mathbf{w}^{old}+y\,\phi(x)\right)\phi(x)$$

$$=\underbrace{y\,\mathbf{w}^{old}\phi(x)}_{<0}+\underbrace{yy\,\phi(x)\phi(x)}_{>0}$$

Support vector machines

- Explicitly optimize the ***margin***
- Enforce that all training points are correctly classified



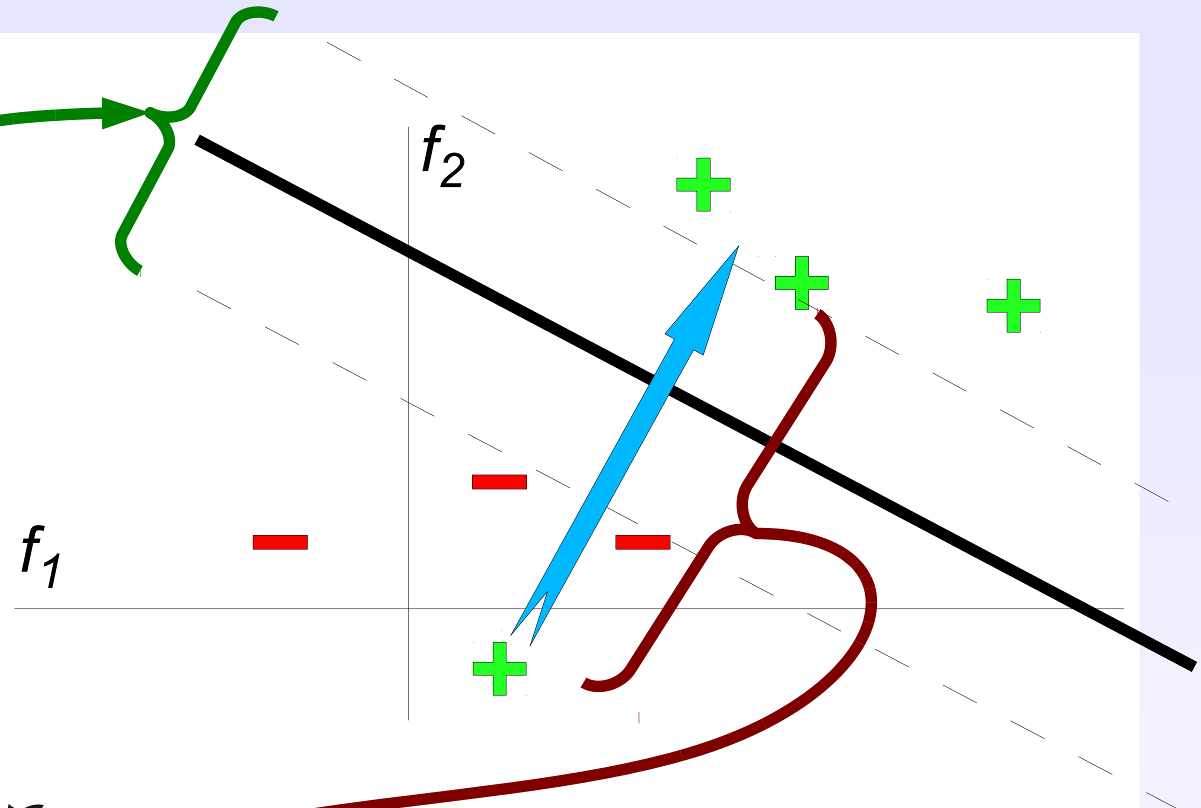
$$\max_{\mathbf{w}} \ \text{margin} \quad s.t. \quad \text{all points are correctly classified}$$

$$\max_{\mathbf{w}} \ \text{margin} \quad s.t. \quad y_n \mathbf{w} \cdot \phi(x_n) \geq 1 \ , \ \forall n$$

$$\min_{\mathbf{w}} \ \|\mathbf{w}\|^2 \quad s.t. \quad y_n \mathbf{w} \cdot \phi(x_n) \geq 1 \ , \ \forall n$$

Support vector machines with *slack*

- Explicitly optimize the ***margin***
- Allow some “noisy” points to be misclassified



$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_n \xi_n$$

$$s.t. \quad y_n \mathbf{w}\cdot\phi(x_n) + \boxed{\xi_n} \geq 1 \ , \quad \forall n$$

$$\xi_n \geq 0 \ , \quad \forall n$$

Batch versus stochastic optimization

- Batch = read in all the data, then process it
- Stochastic = (roughly) process a bit at a time

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_n \xi_n$$

$$s.t. \quad y_n \mathbf{w}\cdot\phi(x_n) + \xi_n \geq 1 \quad , \quad \forall n$$

$$\xi_n \geq 0 \quad , \quad \forall n$$

- For n=1..N:
  - If $y_n \mathbf{w}\cdot\phi(x_n) \leq 0$
    - $\mathbf{w} = \mathbf{w} + y_n \phi(x_n)$

Stochastically optimized SVMs

SVM Objective

SOME MATH

- For n=1..N:
  - If $y_n \mathbf{w} \cdot \phi(x_n) \leq 1$
    - $\mathbf{w} = \mathbf{w} + y_n \phi(x_n)$
  - $\mathbf{w} = \left(1 - \frac{1}{CN}\right) \mathbf{w}$

Implementation Note:

Weight shrinkage is *SLOW.* Implement it lazily, at the cost of double storage.

For n=1..N:

- If $y_n \mathbf{w} \cdot \phi(x_n) \leq 0$
  - $\mathbf{w} = \mathbf{w} + y_n \phi(x_n)$

From Perceptron to Structured Perceptron

Perceptron with multiple classes

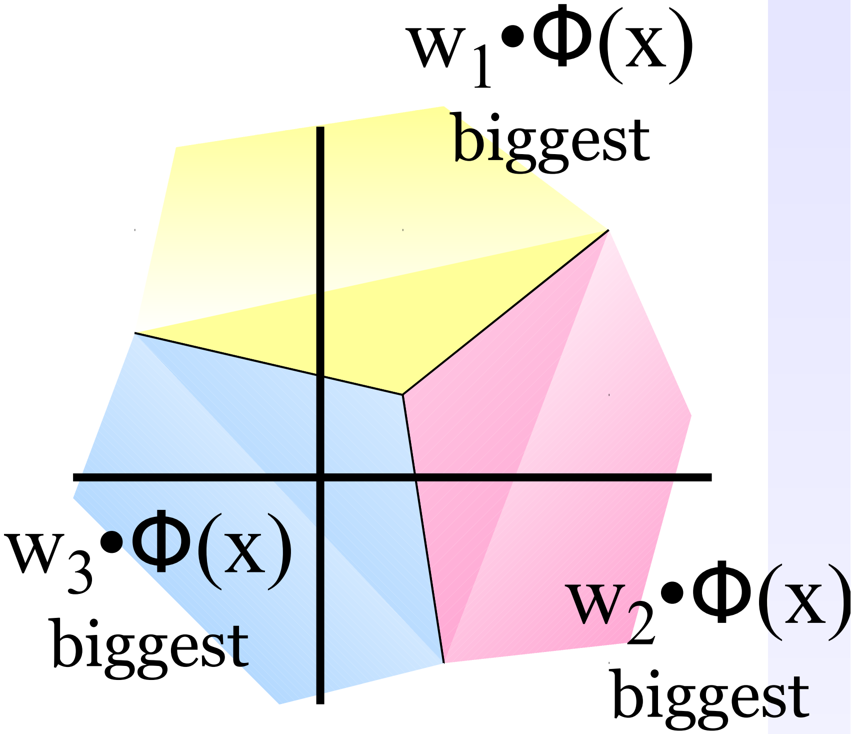
- Store separate weight vector for each class $w_1, w_2, \ldots, w_K$
- For n=1..N:
  - Predict:

$$\hat{y} = \arg\max_k \boldsymbol{w_k} \cdot \phi(x_n)$$

  - If $\hat{y} \neq y_n$:

$$\boldsymbol{w}_{\hat{\boldsymbol{y}}} = \boldsymbol{w}_{\hat{\boldsymbol{y}}} - \phi(x_n)$$

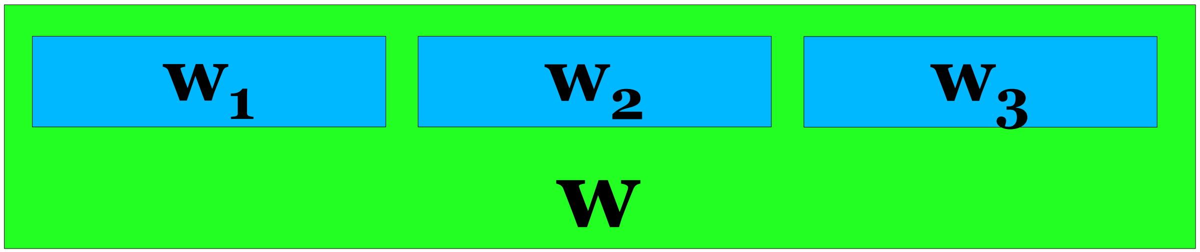
$$\boldsymbol{w}_{\boldsymbol{y_n}} = \boldsymbol{w}_{\boldsymbol{y_n}} + \phi(x_n)$$



?! Why does this do the right thing?

Perceptron with multiple classes v2

- Originally:



- For n=1..N:
  - Predict:

$$\hat{y} = arg\,max_k\, \mathbf{w_k} \cdot \phi(x_n)$$

  - If $\hat{y} \neq y_n$:

$$\mathbf{w}_{\hat{\mathbf{y}}} = \mathbf{w}_{\hat{\mathbf{y}}} - \phi(x_n)$$
$$\mathbf{w}_{\mathbf{y_n}} = \mathbf{w}_{\mathbf{y_n}} + \phi(x_n)$$

- For n=1..N:
  - Predict:

$$\hat{y} = arg\,max_k\, \mathbf{w} \cdot \phi(x_n, k)$$

  - If $\hat{y} \neq y_n$:

$$\mathbf{w} = \mathbf{w} - \phi(x_n, \hat{y})$$
$$+ \phi(x_n, y_n)$$

Perceptron

- Originally:

$\mathbf{w_1}$

x	Φ(x,1)	Φ(x,2)
"free money"	spam_BIAS : 1 spam_free : 1 spam_money : 1 spam_the : 0 ...	ham_BIAS : 1 ham_free : 1 ham_money : 1 ham_the : 0 ...

- For n=1..N:
  - Predict:

$$\hat{y} = \arg\max_k \mathbf{w_k} \cdot \phi(x_n)$$

  - If $\hat{y} \neq y_n$:

$$\mathbf{w}_{\hat{\mathbf{y}}} = \mathbf{w}_{\hat{\mathbf{y}}} - \phi(x_n)$$
$$\mathbf{w}_{\mathbf{y_n}} = \mathbf{w}_{\mathbf{y_n}} + \phi(x_n)$$

- For n=1..N:
  - Predict:

$$\hat{y} = \arg\max_k \mathbf{w} \cdot \phi(x_n, k)$$

  - If $\hat{y} \neq y_n$:

$$\mathbf{w} = \mathbf{w} - \phi(x_n, \hat{y}) + \phi(x_n, y_n)$$

Features for structured prediction

- Allowed to encode *anything* you want

Pro	Md	Vb	Dt	Nn
I	can	can	a	can

$$\phi(\boldsymbol{x}, y) =$$

```
I_Pro      : 1    <s>-Pro   :  1   has_verb    :  1
can_Md     : 1    Pro-Md    :  1   has_nn_lft  :  0
can_Vb     : 1    Md-Vb     :  1   has_n_lft   :  1
a_Dt       : 1    Vb-Dt     :  1   has_nn_rgt  :  1
can_Nn     : 1    Dt-Nn     :  1   has_n_rgt   :  1
...               Nn-</s>   :  1   ...
                  ...
```

- Output features, Markov features, other features

Structured perceptron

Enumeration over 1..K

- For n=1..N:
  - Predict:

$$\hat{y} = arg\,max_k\, \mathbf{w} \cdot \phi(x_n, k)$$

  - If $\hat{y} \neq y_n$:

$$\mathbf{w} = \mathbf{w} - \phi(x_n, \hat{y}) + \phi(x_n, y_n)$$

Enumeration over all outputs

- For n=1..N:
  - Predict:

$$\hat{y} = arg\,max_k\, \mathbf{w} \cdot \phi(x_n, k)$$

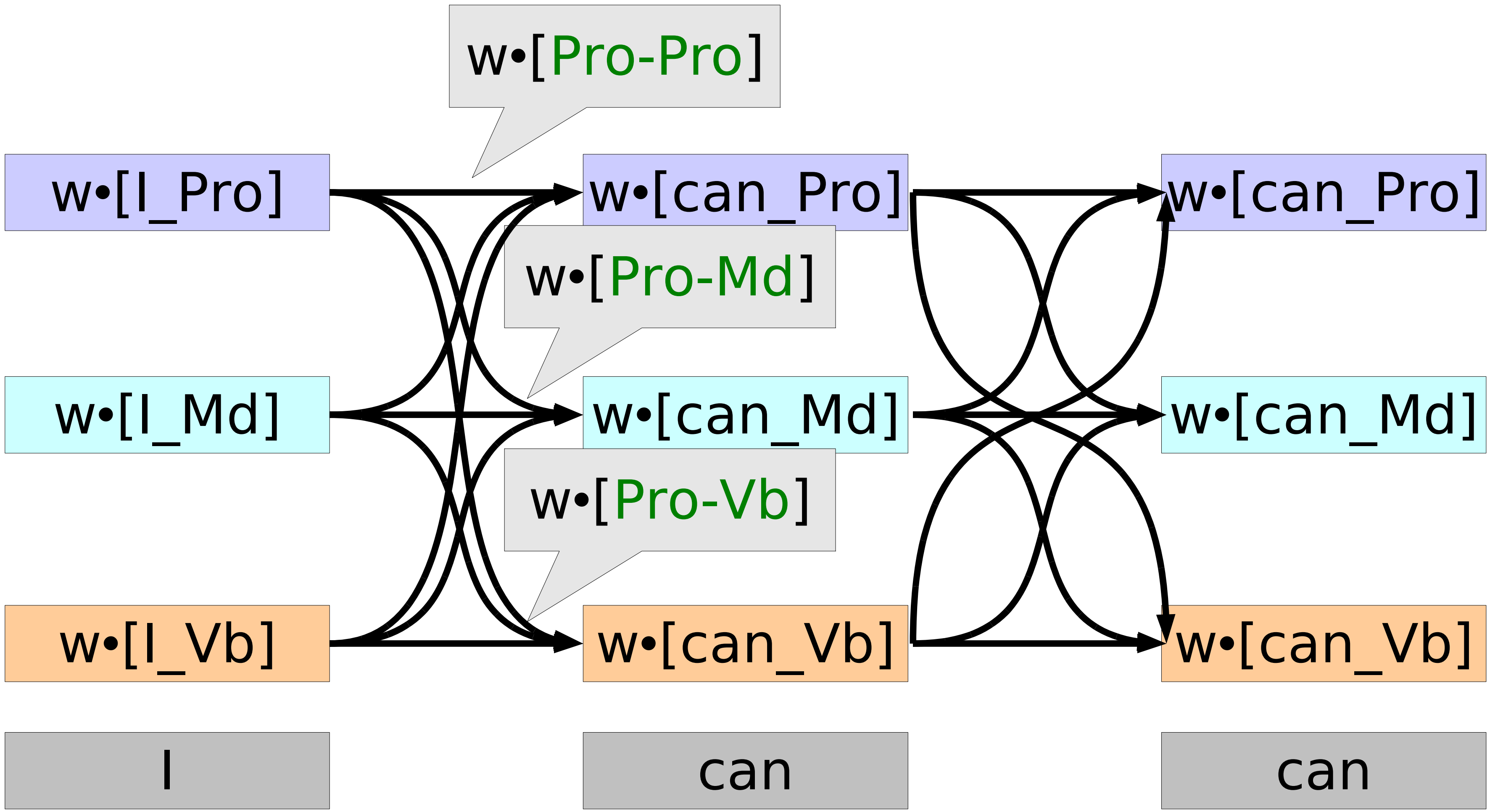
  - If $\hat{y} \neq y_n$:

$$\mathbf{w} = \mathbf{w} - \phi(x_n, \hat{y}) + \phi(x_n, y_n)$$

[Collins, EMNLP02]

Argmax for sequences

- If we only have output and Markov features, we can use Viterbi algorithm:



(plus some work to account for boundary conditions)

Structured perceptron as ranking

- For n=1..N:
  - Run Viterbi: $\hat{y} = arg\,max_k\, \mathbf{w} \cdot \phi(x_n, k)$
  - If $\hat{y} \neq y_n$: $\mathbf{w} = \mathbf{w} - \phi(x_n, \hat{y}) + \phi(x_n, y_n)$

- When does this make an update?

Pro	Md	Vb	Dt	Nn
Pro	Md	Md	Dt	Vb
Pro	Md	Md	Dt	Nn
Pro	Md	Nn	Dt	Md
Pro	Md	Nn	Dt	Nn
Pro	Md	Vb	Dt	Md
Pro	Md	Vb	Dt	Vb

I	can	can	a	can

From perceptron to margins

Maximize Margin

Minimize Errors

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_n \xi_n$$

$$s.t. \quad y_n \boldsymbol{w}\cdot\phi(x_n) + \xi_n \geq 1, \quad \forall n$$

Each point is correctly classified, modulo ξ

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_n \xi_{n,\hat{y}}$$

Response for truth

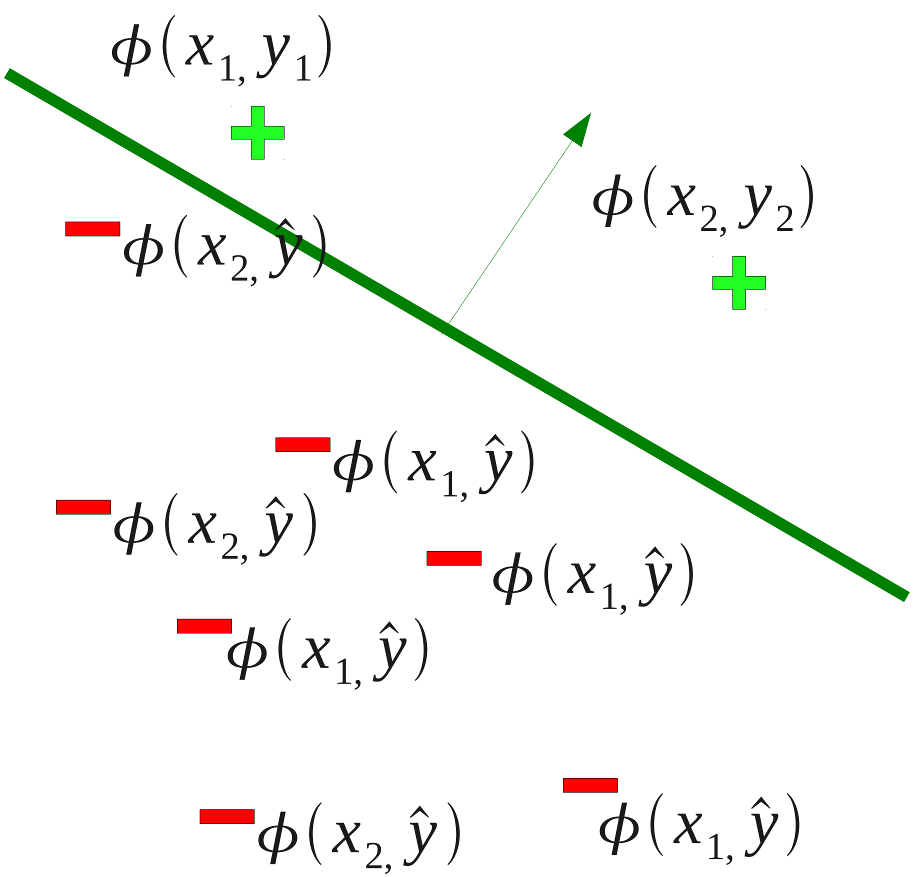
Response for other

$$s.t. \quad \boldsymbol{w}\cdot\phi(x_n, y_n) - \boldsymbol{w}\cdot\phi(x_n, \hat{y}) + \xi_n \geq 1, \forall n, \hat{y} \neq y_n$$

Each true output is more highly ranked, modulo ξ

[Taskar+al, JMLR05; Tsochandaritis, JMLR05]

From perceptron to margins



$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_n \xi_{n,\hat{y}}$$

Response for truth

Response for other

$$s.t. \quad \mathbf{w}\cdot\phi(x_n, y_n) - \mathbf{w}\cdot\phi(x_n, \hat{y}) + \xi_n \geq 1, \forall n, \hat{y} \neq y_n$$

Each true output is more highly ranked, modulo ξ

[Taskar+al, JMLR05; Tsochandaritis, JMLR05]

Ranking margins

- Some errors are worse than others...

Pro	Md	Vb	Dt	Nn

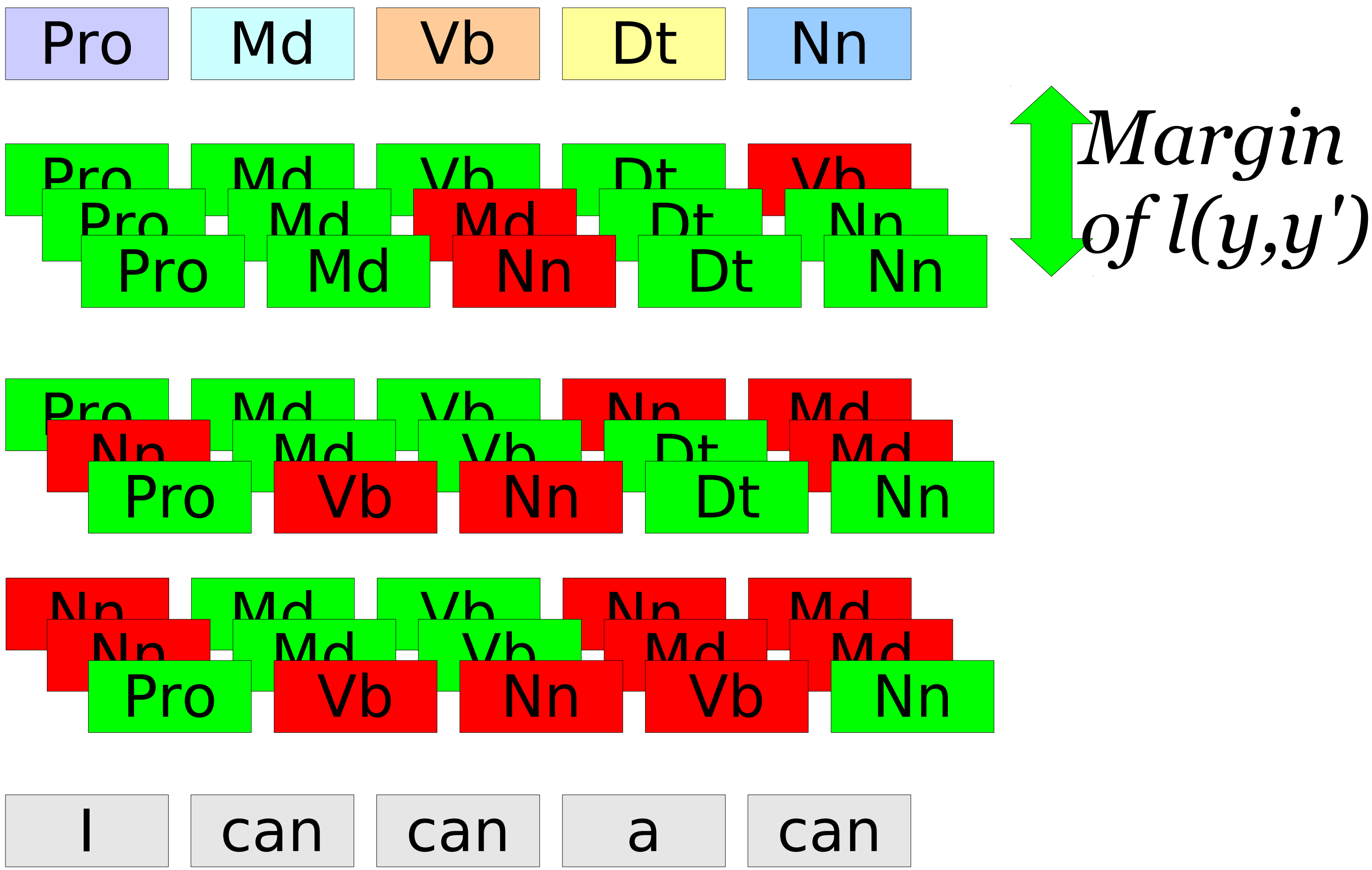
Margin of one

Pro	Md	Md	Dt	Vb
Pro	Md	Md	Dt	Nn
Pro	Md	Nn	Dt	Md
Pro	Md	Nn	Dt	Nn
Pro	Md	Vb	Dt	Md
Pro	Md	Vb	Dt	Vb
I	can	can	a	can

[Taskar+al, JMLR05; Tsochandaritis, JMLR05]

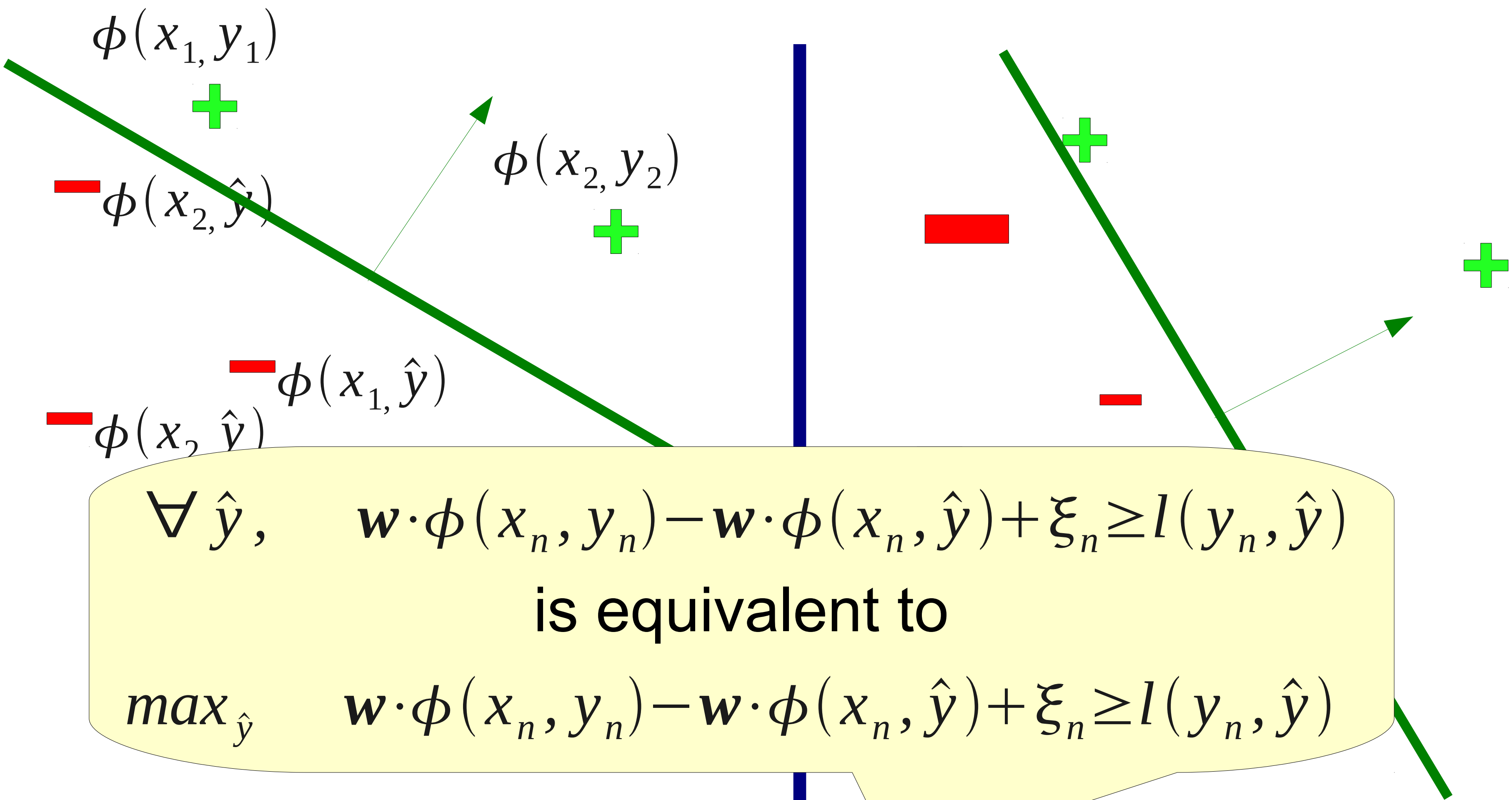
Accounting for a loss function

- Some errors are worse than others...



[Taskar+al, JMLR05; Tsochandaritis, JMLR05]

Accounting for a loss function



$$\forall \hat{y}, \quad \boldsymbol{w}\cdot\phi(x_n, y_n) - \boldsymbol{w}\cdot\phi(x_n, \hat{y}) + \xi_n \geq l(y_n, \hat{y})$$

is equivalent to

$$max_{\hat{y}} \quad \boldsymbol{w}\cdot\phi(x_n, y_n) - \boldsymbol{w}\cdot\phi(x_n, \hat{y}) + \xi_n \geq l(y_n, \hat{y})$$

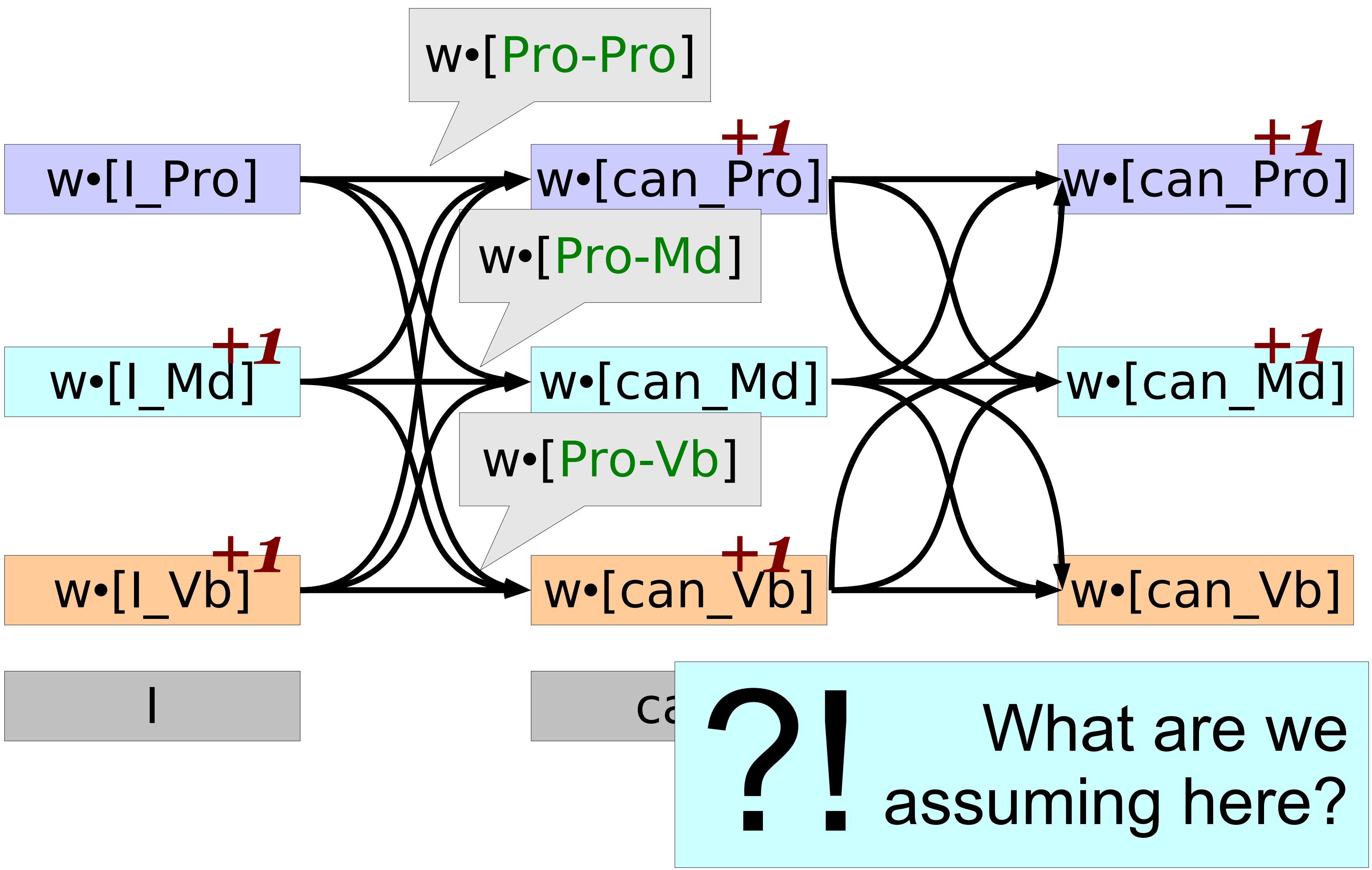
$$\boldsymbol{w}\cdot\phi(x_n, y_n) - \boldsymbol{w}\cdot\phi(x_n, \hat{y}) + \xi_n \geq 1$$

$$\boldsymbol{w}\cdot\phi(x_n, y_n) - \boldsymbol{w}\cdot\phi(x_n, \hat{y}) + \xi_n \geq l(y_n, \hat{y})$$

[Taskar+al, JMLR05; Tsochandaritis, JMLR05]

Augmented argmax for sequences

- Add “loss” to each wrong node!



[Taskar+al, JMLR05; Tsochandaritis, JMLR05]

Stochastically optimizing Markov nets

M3N Objective

SOME MATH

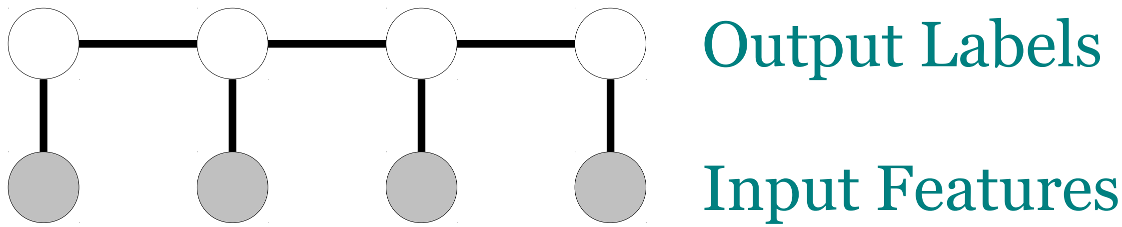
- For n=1..N:
  - Augmented Viterbi:
    $$\hat{y} = arg\,max_k\, \mathbf{w}\cdot\phi(x_n, k) + l(y_n, k)$$
  - If $\hat{y} \neq y_n$:
    $$\mathbf{w} = \mathbf{w} - \phi(x_n, \hat{y}) + \phi(x_n, y_n)$$
  - $$\mathbf{w} = \left(1 - \frac{1}{CN}\right)\mathbf{w}$$

- For n=1..N:
  - Viterbi:
    $$\hat{y} = arg\,max_k\, \mathbf{w}\cdot\phi(x_n, k)$$
  - If $\hat{y} \neq y_n$:
    $$\mathbf{w} = \mathbf{w} - \phi(x_n, \hat{y}) + \phi(x_n, y_n)$$

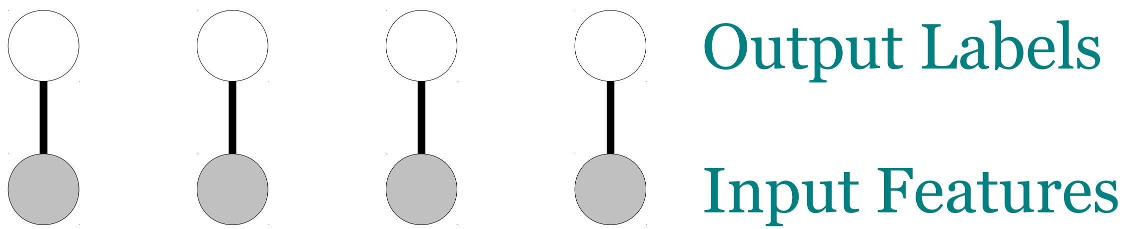
[Ratliff+al, AIStats07]

Stacking

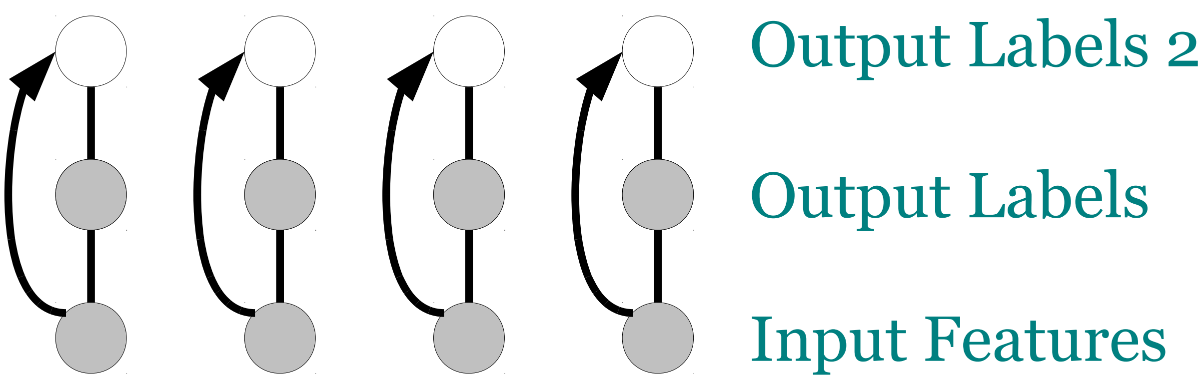
- **Structured models:** accurate but slow



- **Independent models:** less accurate but fast

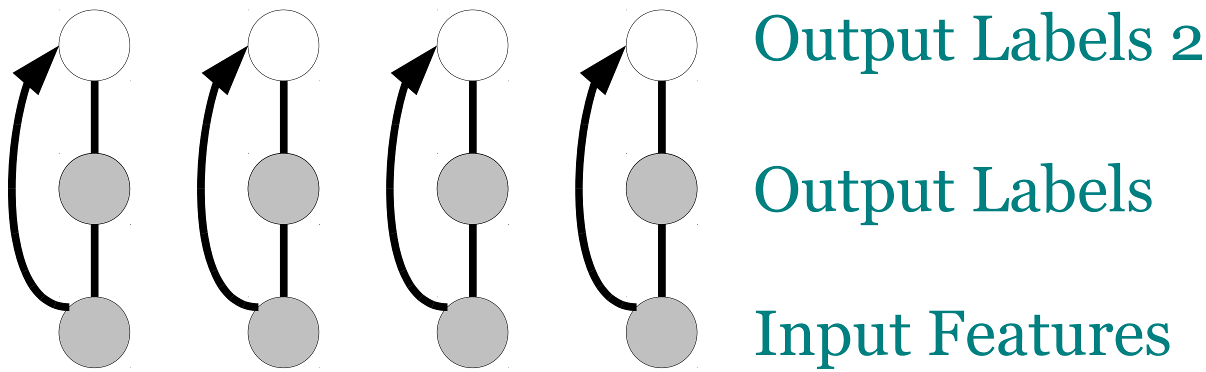


- **Stacking:** multiple independent models



Training a stacked model

- Train independent classifier f_1 on input features



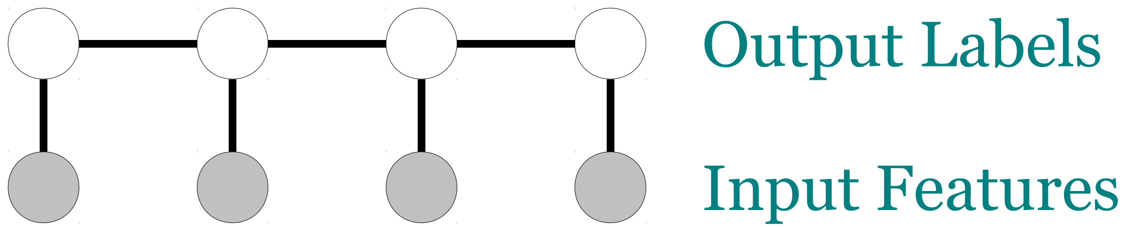
- Train independent classifier f_2 on input features + f_1's output

- **Danger:** overfitting!

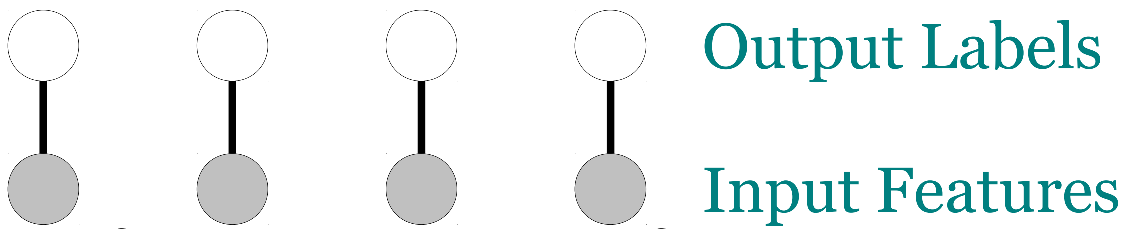
- **Solution**: cross-validation

Do we really need structure?

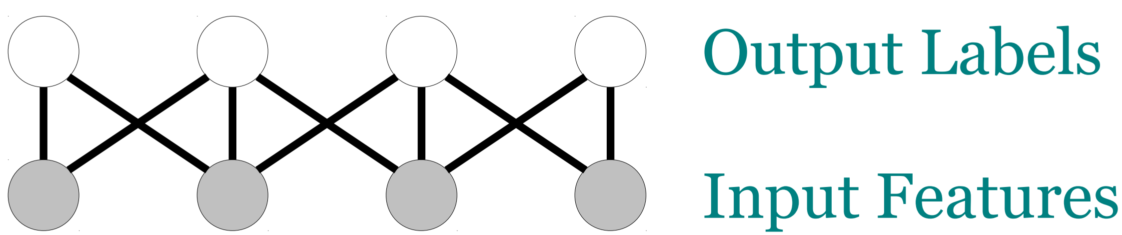
- **Structured models:** accurate but slow



- **Independent models:** less accurate but fast



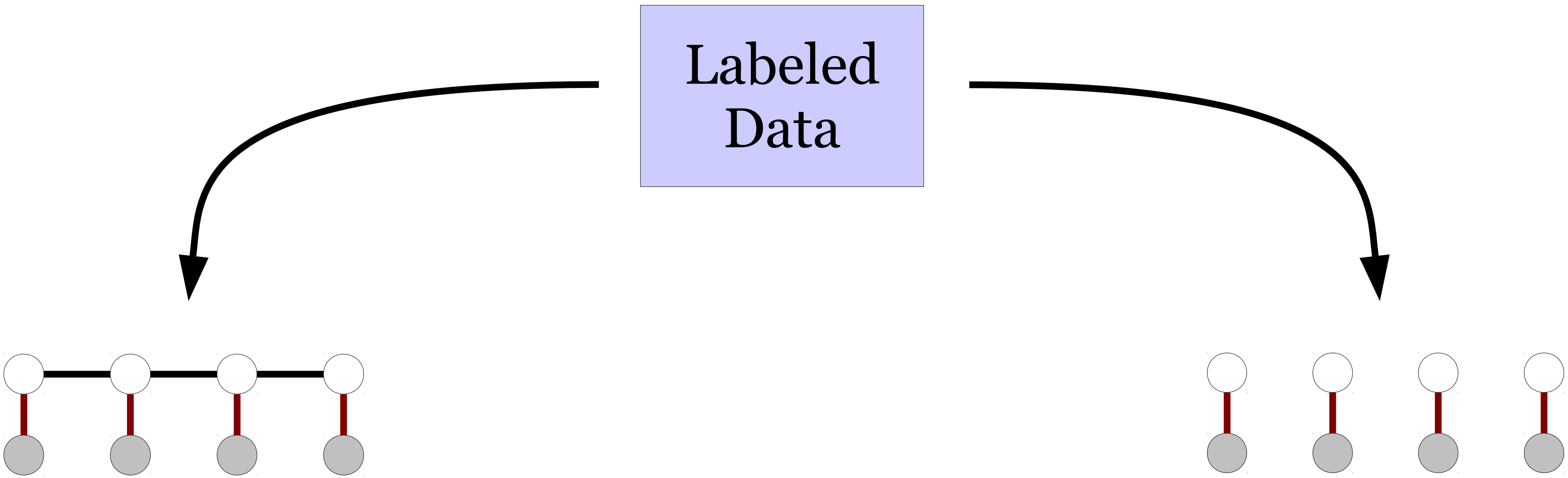
- **Goal:** transfer power to get fast+accurate



- **Questions:** are independent models...
  - … expressive enough? (approximation error)
  - … easy to learn? (estimation error)

[Liang+D+Klein, ICML08]

“Compiling” structure out



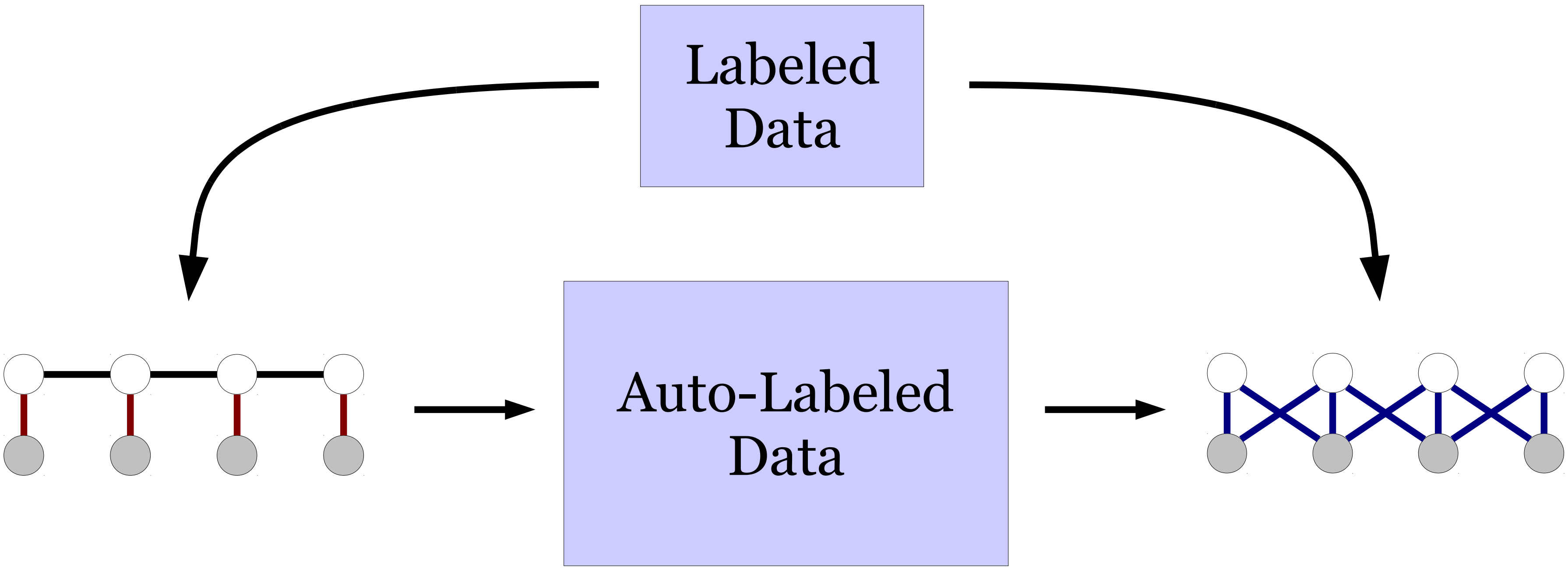
CRF(f_1) POS: 95.0% NER: 75.3%

IRL(f_1) POS: 91.7% NER: 69.1%

f_1 = words/prefixes/suffixes/forms

[Liang+D+Klein, ICML08]

“Compiling” structure out



CRF(f_1) POS: 95.0% NER: 75.3%

IRL(f_1) POS: 91.7% NER: 69.1%

IRL(f_2) POS: 94.4% NER: 66.2%

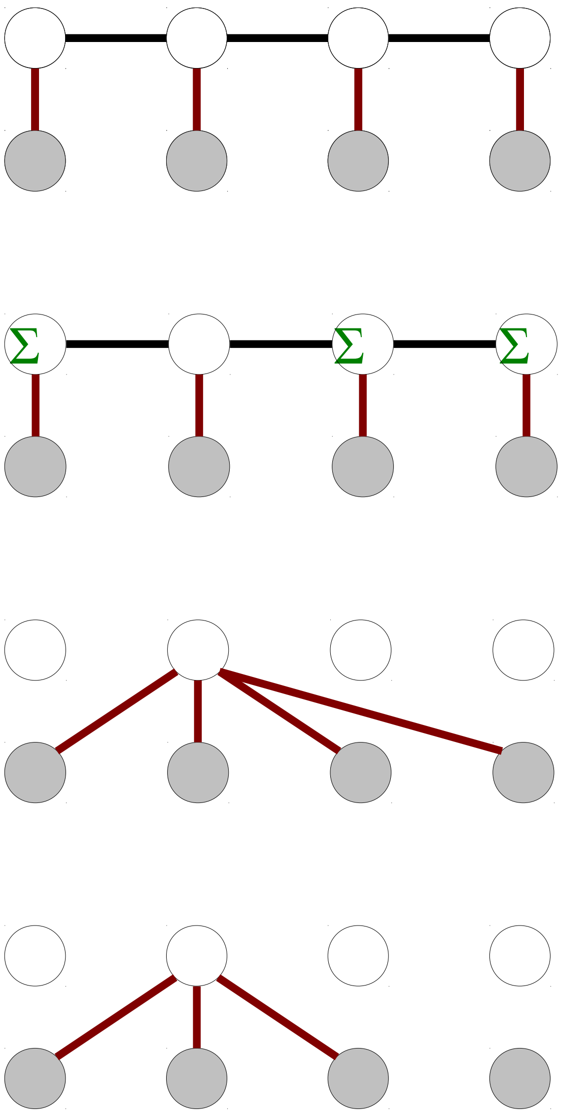
CompIRL(f_2) POS: 95.0% NER: 72.7%

f_1 = words/prefixes/suffixes/forms

f_2 = f_1 applied to a larger window

[Liang+D+Klein, ICML08]

Decomposition of errors



CRF(f_1): p_C

Sum of MI on edges
POS=.003 (95.0% → 95.0%)
NER=.009 (76.3% → 76.0%)

coherence

marginalized CR

Train a truncated CRF
NER: 76.0% → 72.7%

nonlinearities

IRL(f_∞): p_{A^*}

Train a marginalized CRF
NER: 76.0% → 76.0%

global information

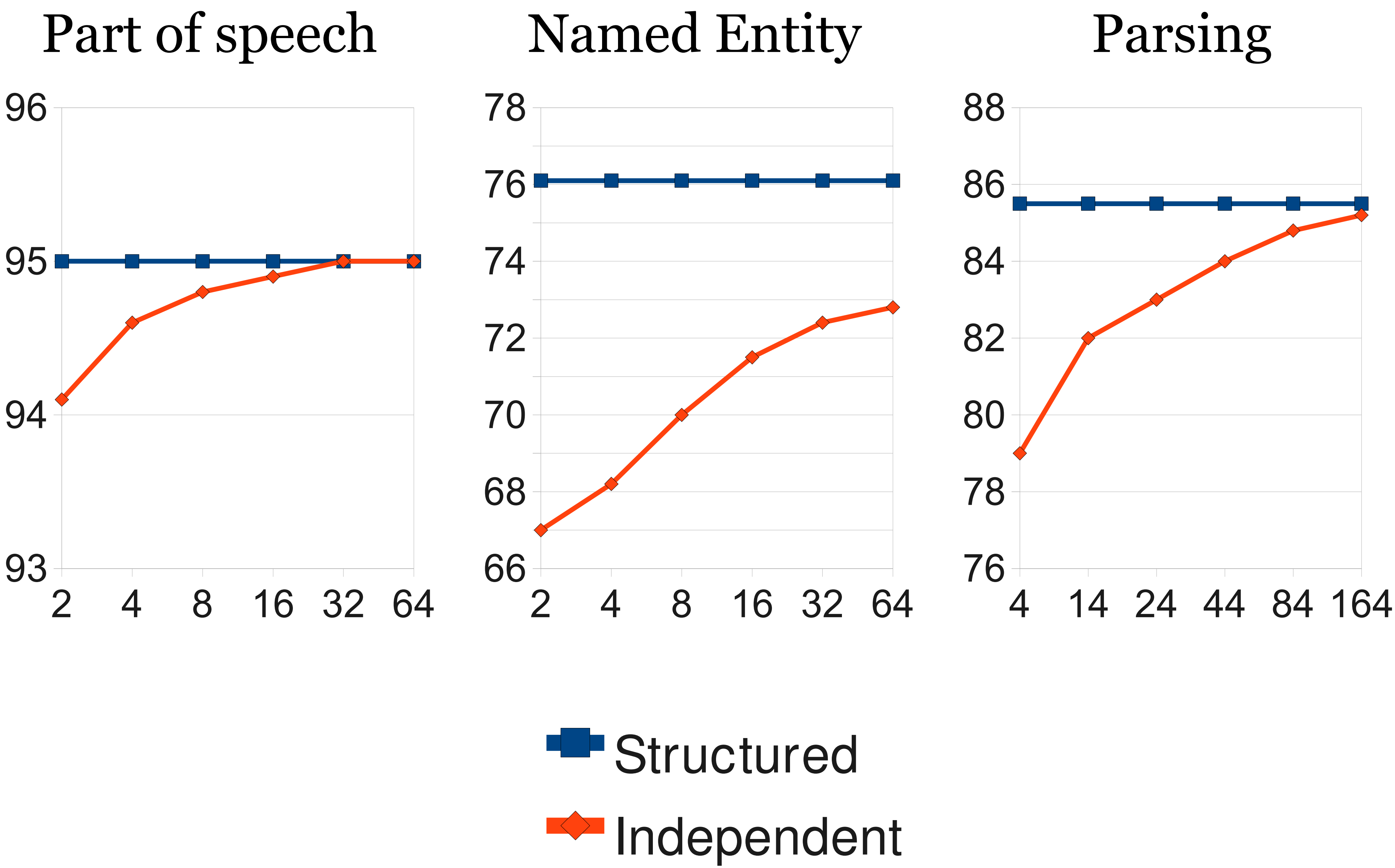
IRL(f_2): p_{1^*}

Theorem:

$$\mathrm{KL}(p_C \,||\, p_{1^*}) = \mathrm{KL}(p_C \,||\, p_{MC}) + \mathrm{KL}(p_{MC} \,||\, p_{A^*}) + \mathrm{KL}(p_{A^*} \,||\, p_{1^*})$$

[Liang+D+Klein, ICML08]

Structure compilation results



[Liang+D+Klein, ICML08]

Coffee Break!!!

Outline: Part I

- What is Structured Prediction?
- Refresher on Binary Classification
  - What does it mean to learn?
  - Linear models for classification
  - Batch versus stochastic optimization
- From Perceptron to Structured Perceptron
  - Linear models for Structured Prediction
  - The “argmax” problem
  - From Perceptron to margins
- Structure without Structure
  - Stacking
  - Structure compilation

Outline: Part II

- Learning to Search
  - Incremental parsing
  - Learning to queue
- Refresher on Markov Decision Processes
- Inverse Reinforcement Learning
  - Determining rewards given policies
  - Maximum margin planning
- Learning by Demonstration
  - Searn
  - Dagger
- Discussion

Learning to Search

Argmax is *hard!*

- Classic formulation of structured prediction:

$$score(x,y) = \text{something we learn to make “good” } x,y \text{ pairs score highly}$$

- At test time:

$$f(x) = argmax_{y \in Y} score(x,y)$$

- Combinatorial optimization problem
  - Efficient only in very limiting cases
  - Solved by heuristic search: beam + A* + local search

Argmax is *hard!*

- Classic

  sc

- At test

  $f($

Order these words: bart better I madonna say than ,

[Soricut, PhD Thesis, USC 2007]

- Combinatorial optimization problem
  - Efficient only in very limiting cases
  - Solved by heuristic search: beam + A* + local search

Argmax is *hard!*

- Classic
  sc
- At test
  f (
- Combinatorial optimization problem
  - Efficient only in very limiting cases
  - Solved by heuristic search: beam + A* + local search

Order these words:		bart better I madonna say than ,
Best search	*(32.3):*	I say better than bart madonna ,
Original	*(41.6):*	better bart than madonna , I say

[Soricut, PhD Thesis, USC 2007]

Argmax is *hard!*

- Classic
  - sc
- At test
  - f (

Order these words: bart better I madonna say than ,
Best search (32.3): I say better than bart madonna ,
Original (41.6): better bart than madonna , I say

Best search (51.6): and so could really be a neural apparently thought things as dissimilar firing two identical

[Soricut, PhD Thesis, USC 2007]

- Combinatorial optimization problem
  - Efficient only in very limiting cases
  - Solved by heuristic search: beam + A* + local search

Argmax is *hard!*

- Classic

  sc

- At test

  f (

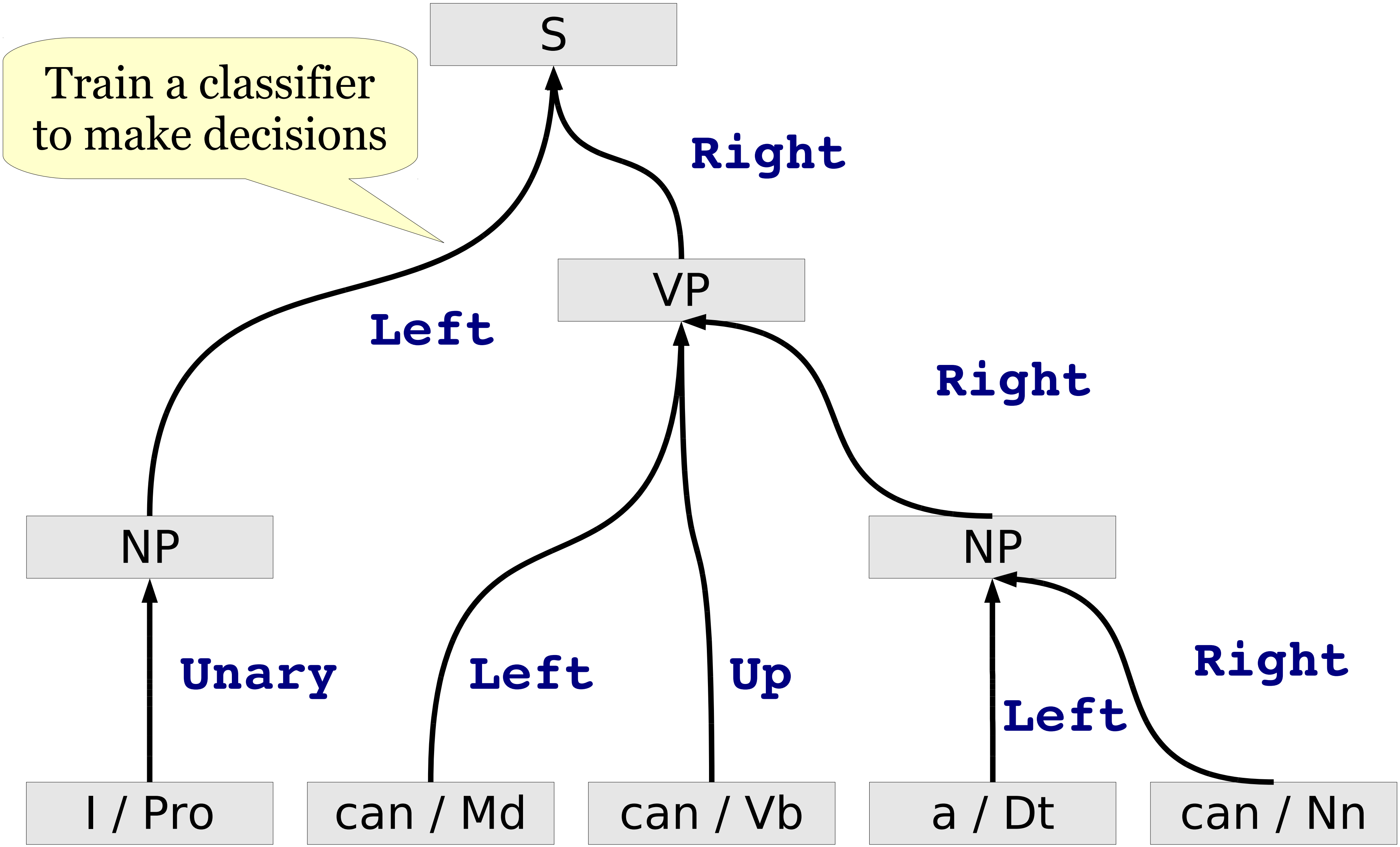
Order these words: bart better I madonna say than ,
Best search (32.3): I say better than bart madonna ,
Original (41.6): better bart than madonna , I say

Best search (51.6): and so could really be a neural apparently thought things as dissimilar firing two identical
Original (64.3): could two things so apparently dissimilar as a thought and neural firing really be identical

[Soricut, PhD Thesis, USC 2007]

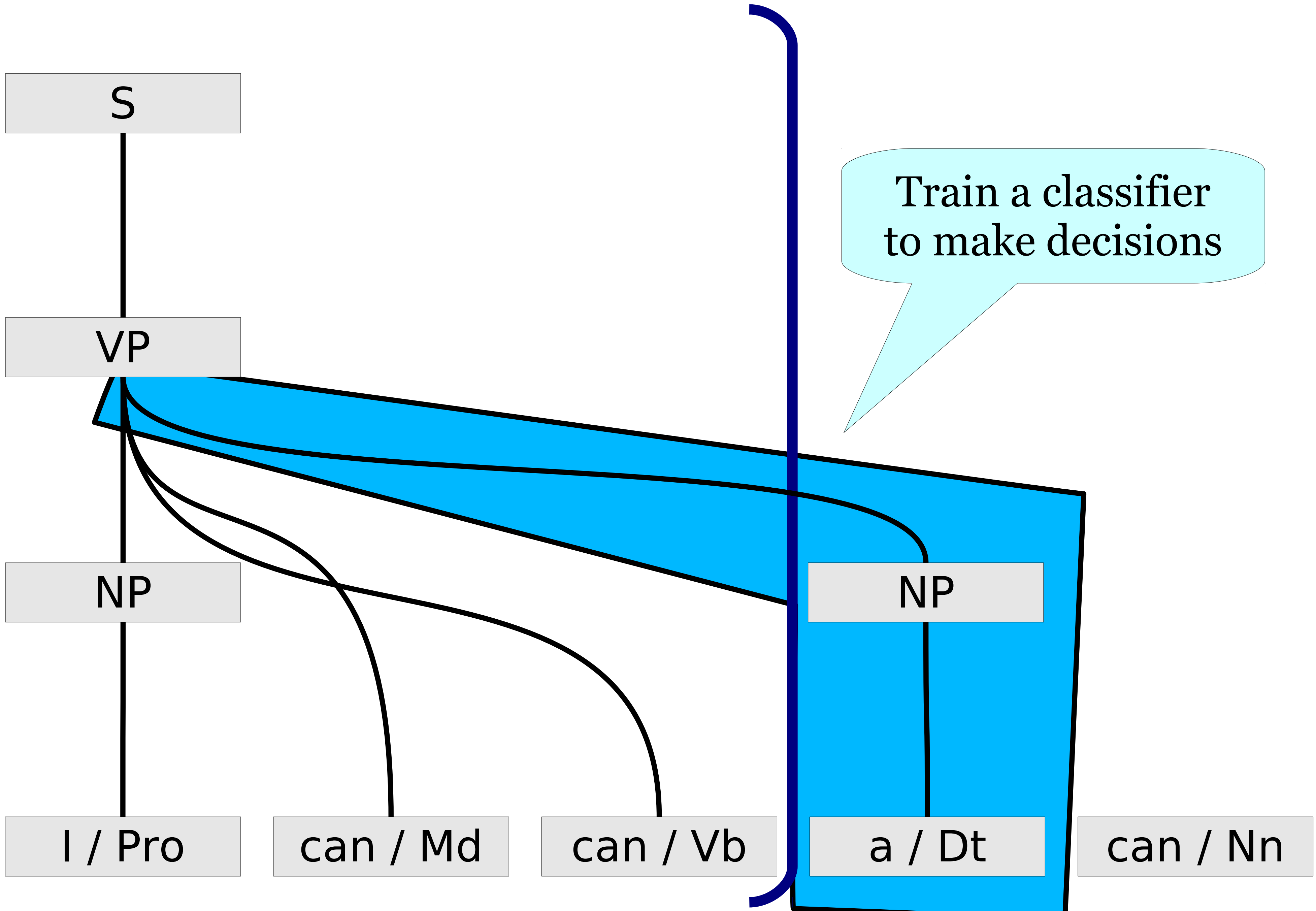
- Combinatorial optimization problem
  - Efficient only in very limiting cases
  - Solved by heuristic search: beam + A* + local search

Incremental parsing, early 90s style



[Magerman, ACL95]

Incremental parsing, mid 2000s style



[Collsns+Roark, ACL04]

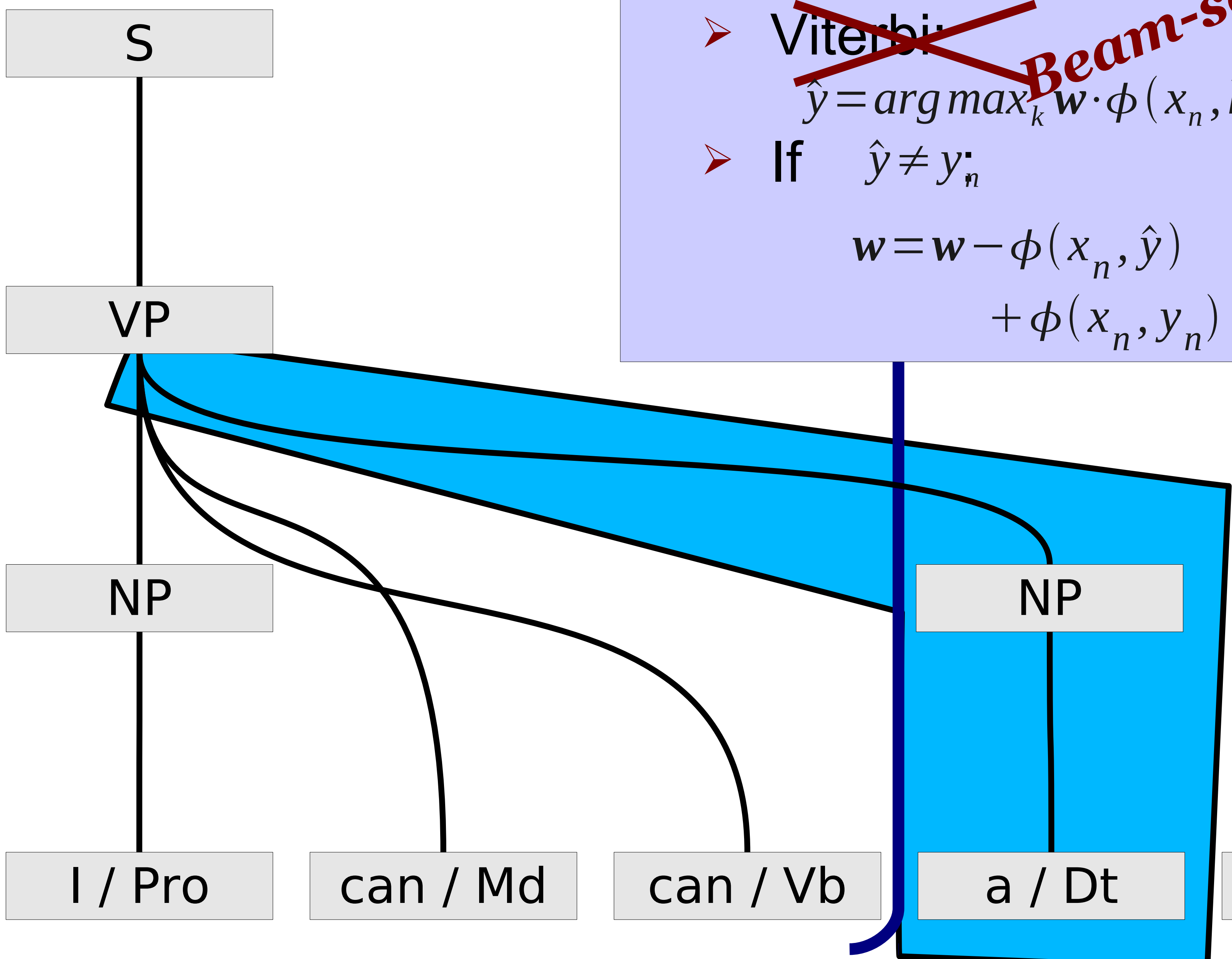
Learning to beam-search

- For n=1..N:
  - ~~Viterbi:~~ *Beam-search*

    $$\hat{y} = \arg\max_k \mathbf{w} \cdot \phi(x_n, k)$$

  - If $\hat{y} \neq y_n$:

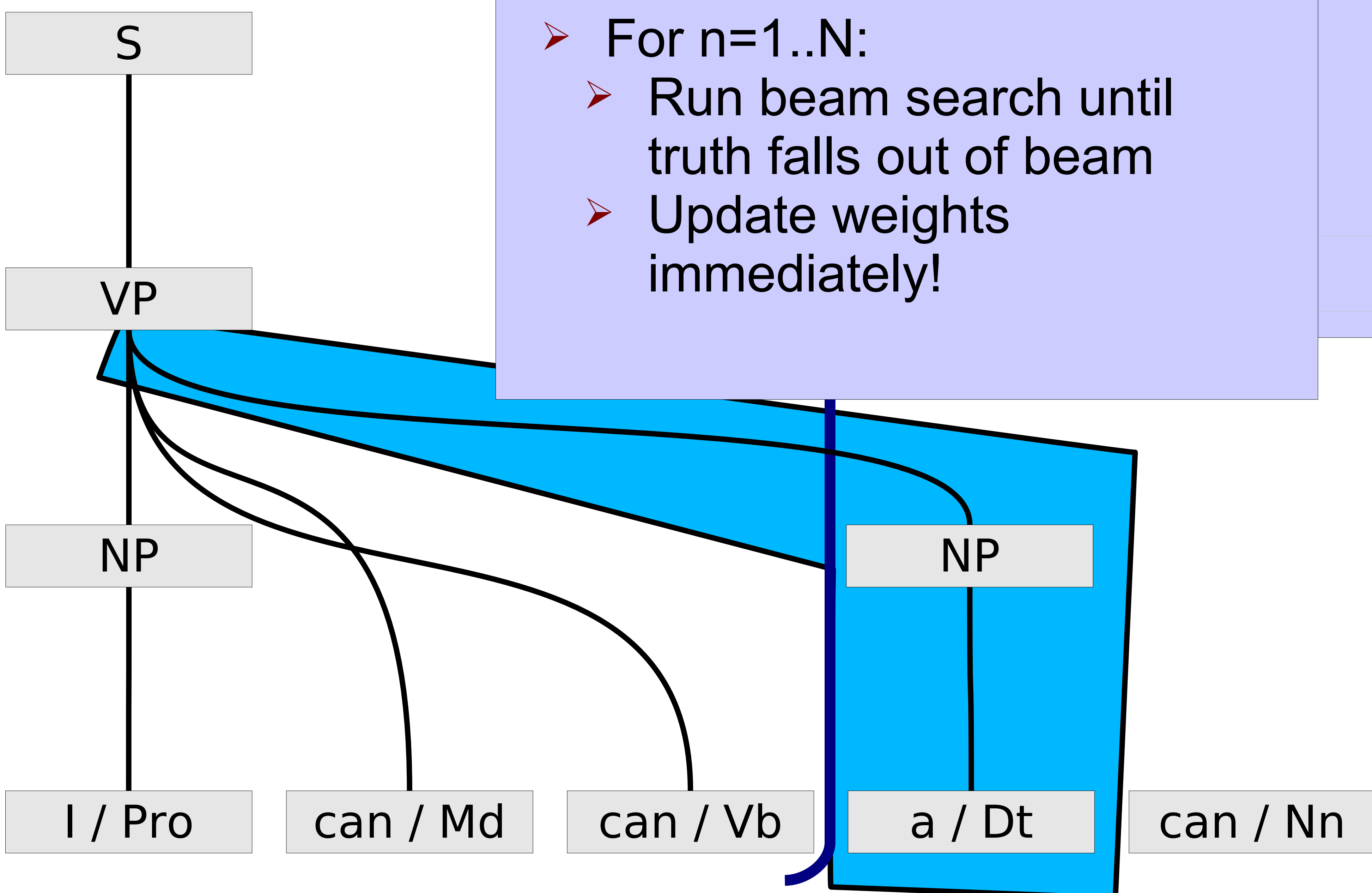
    $$\mathbf{w} = \mathbf{w} - \phi(x_n, \hat{y}) + \phi(x_n, y_n)$$



[Collins+Roark, ACL04]

Learning to beam-search

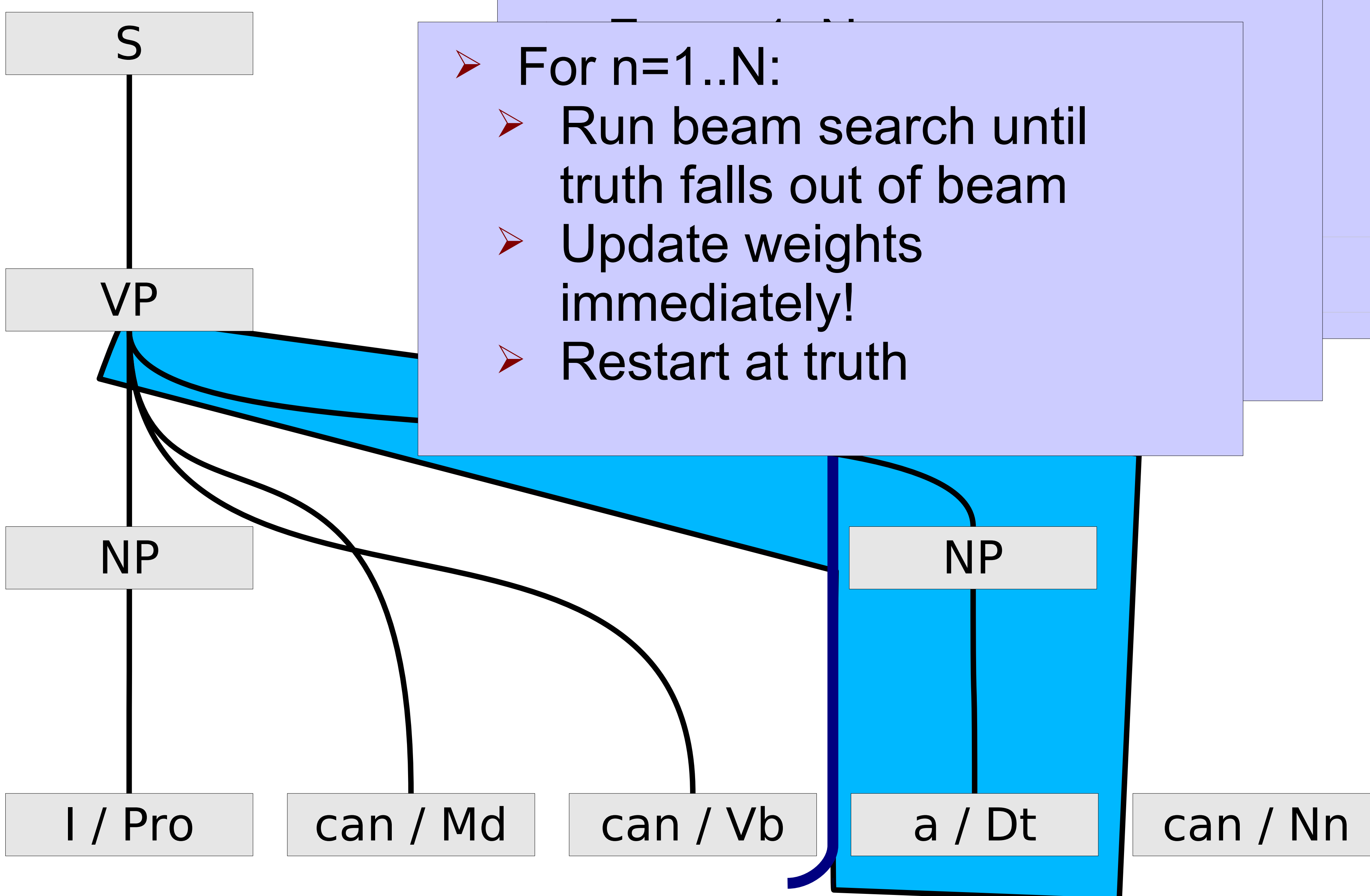
- For n=1..N:
  - Run beam search until truth falls out of beam
  - Update weights immediately!



[Collins+Roark, ACL04]

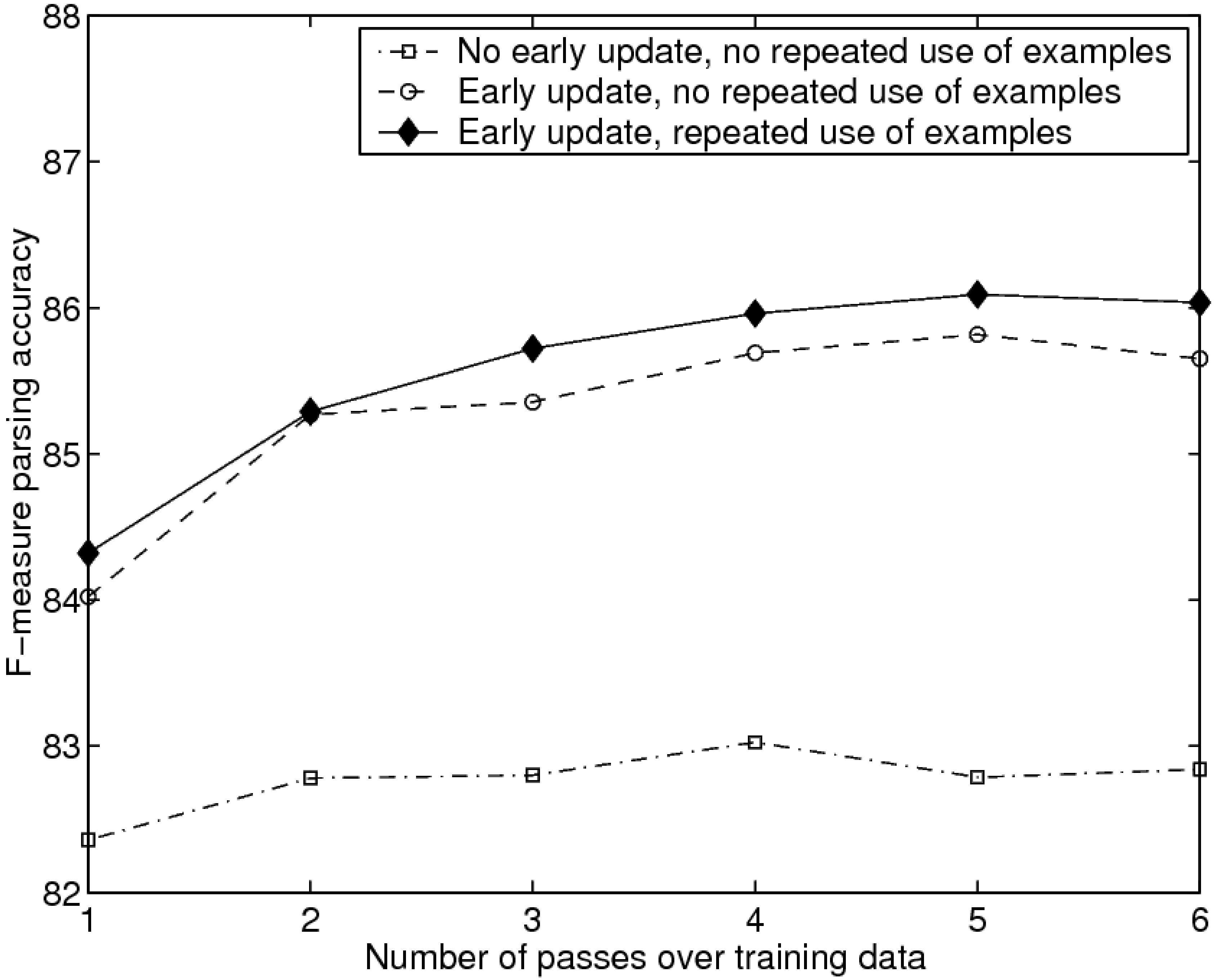
Learning to beam-search

- For n=1..N:
  - Run beam search until truth falls out of beam
  - Update weights immediately!
  - Restart at truth



[D+Marcu, ICML05; Xu+al, JMLR09]

Incremental parsing results



[Collins+Roark, ACL04]

Generic Search Formulation

- Search Problem:
  - Search space
  - Operators
  - Goal-test function
  - Path-cost function
- Search Variable:
  - Enqueue function

Varying the **Enqueue** function can give us DFS, BFS, beam search, A* search, etc...

- nodes := MakeQueue(S0)
- **while** nodes is not empty
  - node := RemoveFront(nodes)
  - **if** node is a goal state **return** node
  - next := Operators(node)
  - nodes := Enqueue(nodes, next)
- **fail**

[D+Marcu, ICML05; Xu+al, JMLR09]

Online Learning Framework (LaSO)

- nodes := MakeQueue(S0)
- **while** nodes is not empty
  - node := RemoveFront(nodes)
  - **if** none of {node} ∪ nodes is y-good **or** node is a goal **&** not y-good
    - sibs := siblings(node, y)
    - w := update(w, x, sibs, {node} ∪ nodes)
    - nodes := MakeQueue(sibs)
  - **else**
    - **if** node is a goal **return** w
    - next := Operators(node)
    - nodes := Enqueue(nodes, next)

Monotonicity: for any node, we can tell if it can lead to the correct solution or not

If we erred...

Where should we have gone?

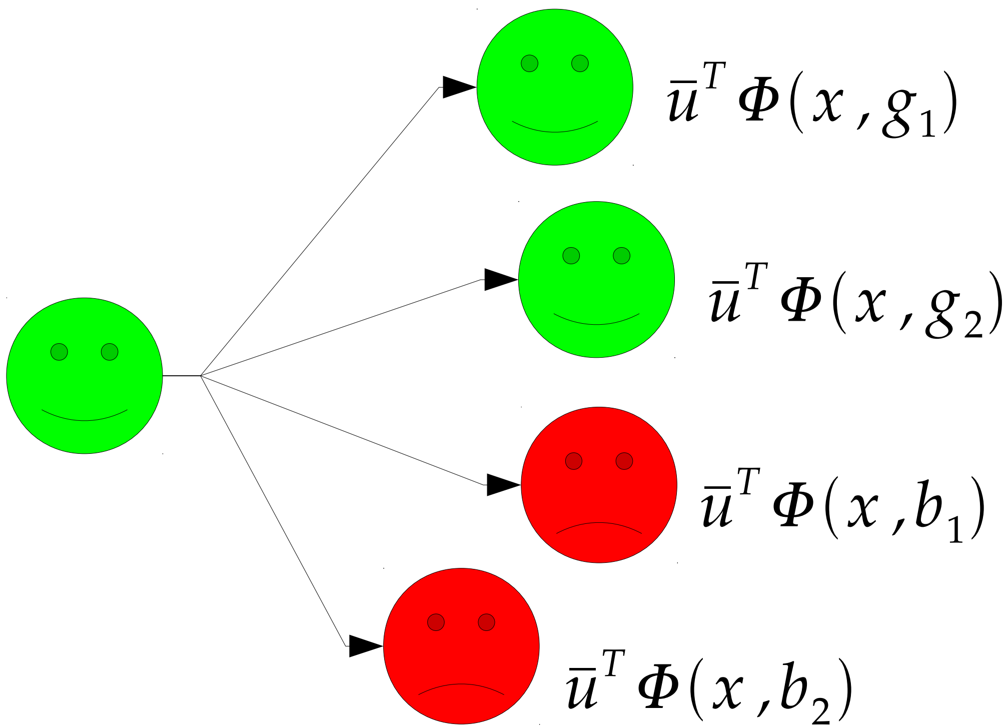
Update our weights based on the good and the bad choices

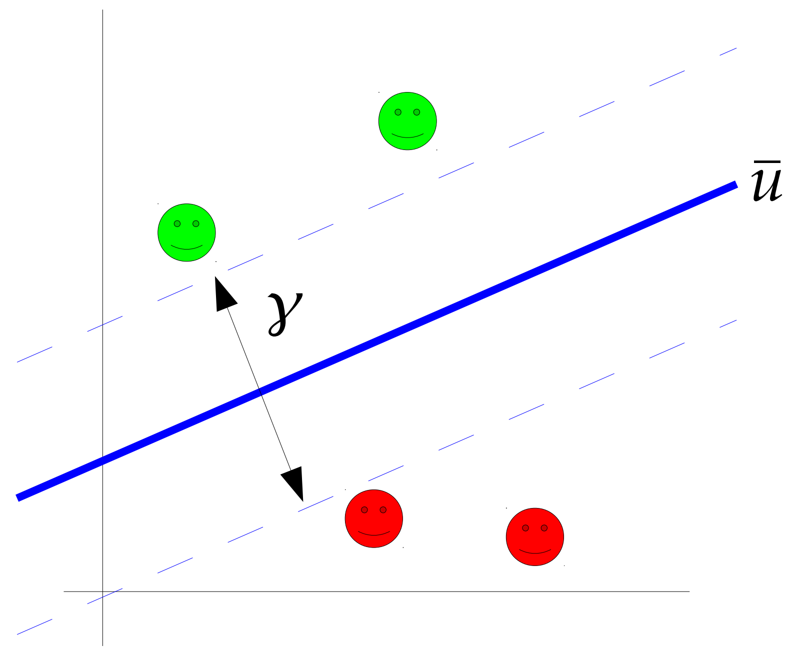
Continue search...

[D+Marcu, ICML05; Xu+al, JMLR09]

Search-based Margin

- The *margin* is the amount by which we are correct:

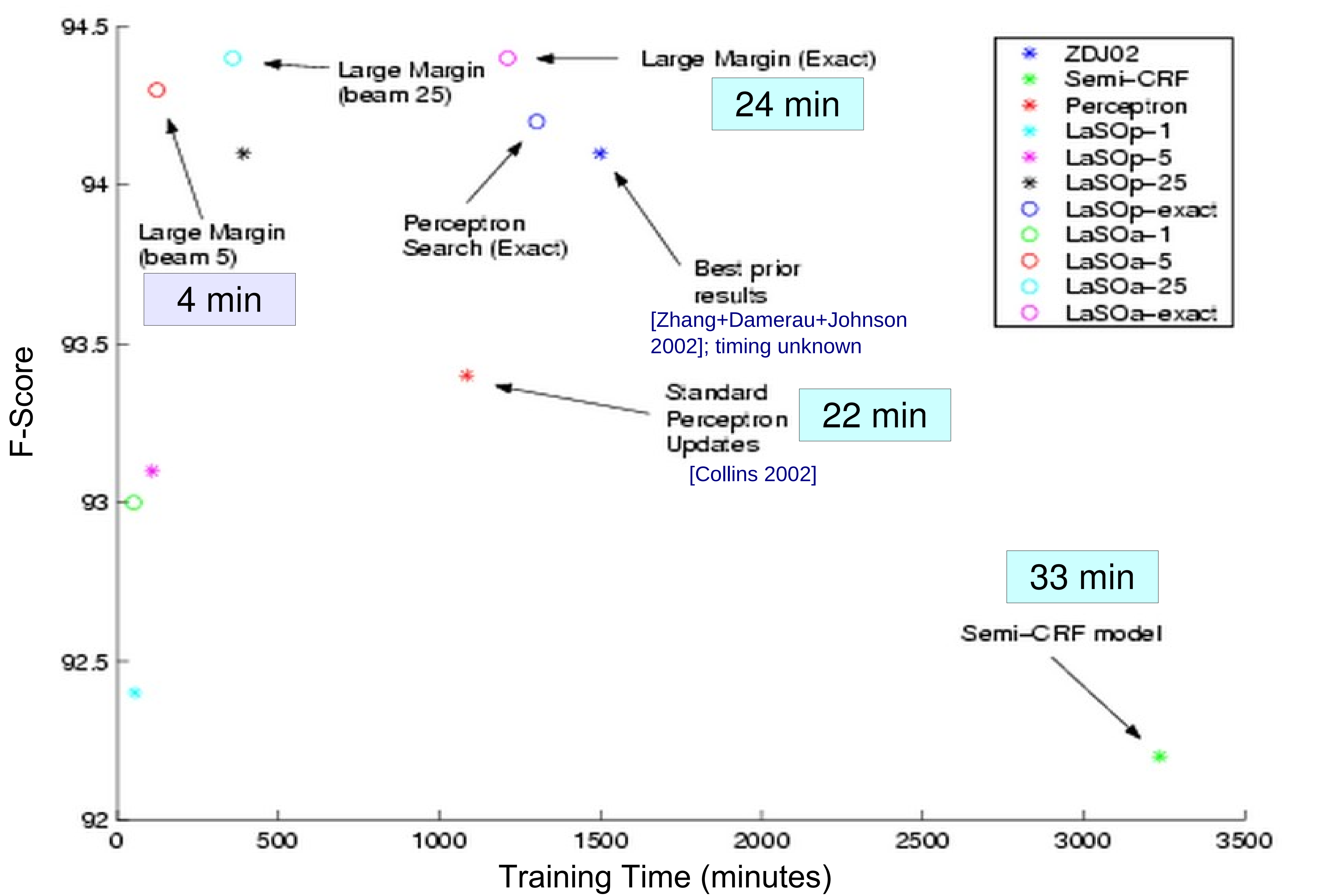




- Note that the *margin* and hence *linear separability* is also a function of the *search algorithm!*

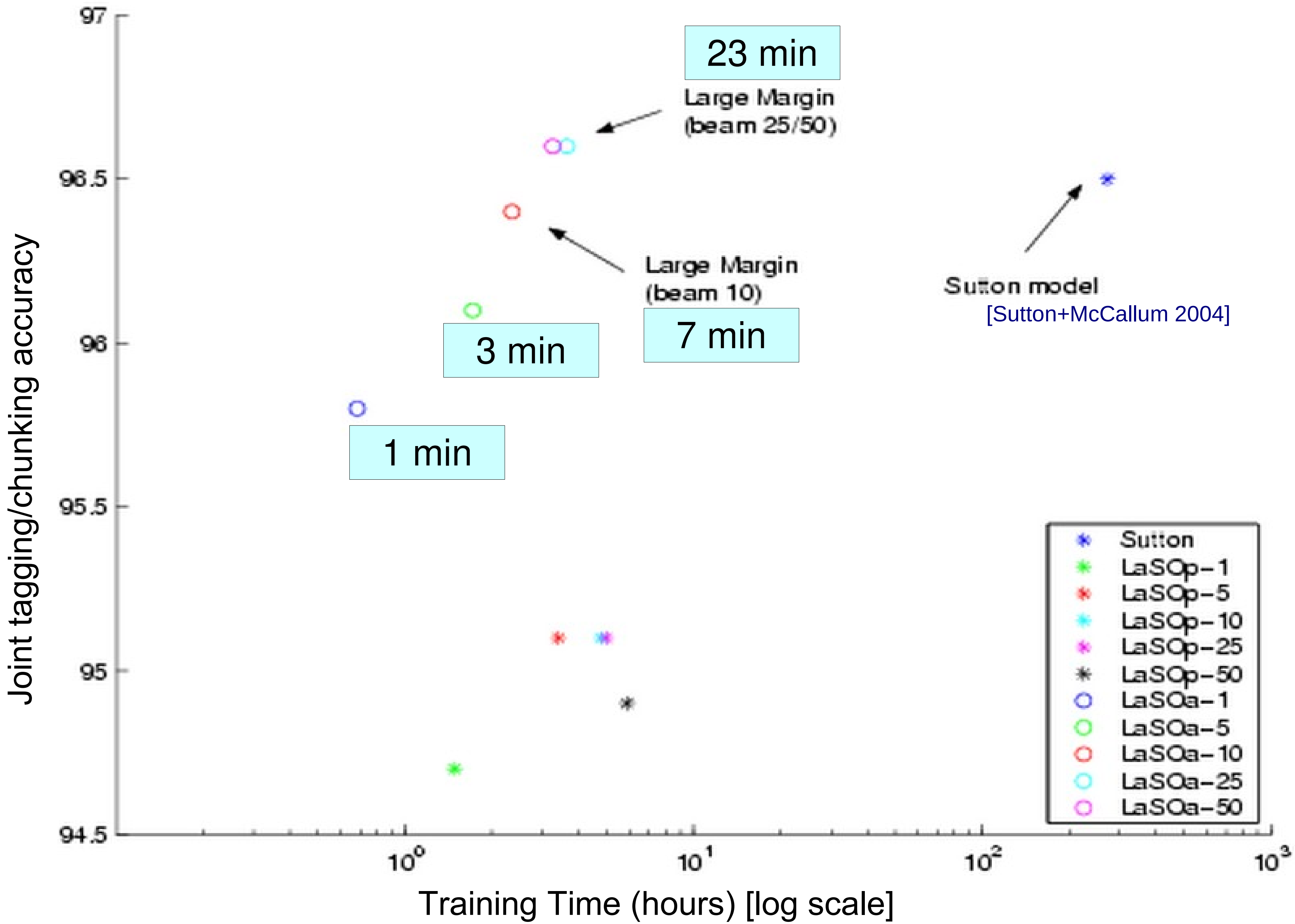
[D+Marcu, ICML05; Xu+al, JMLR09]

Syntactic chunking Results



[D+Marcu, ICML05; Xu+al, JMLR09]

Tagging+chunking results



[D+Marcu, ICML05; Xu+al, JMLR09]

Variations on a beam

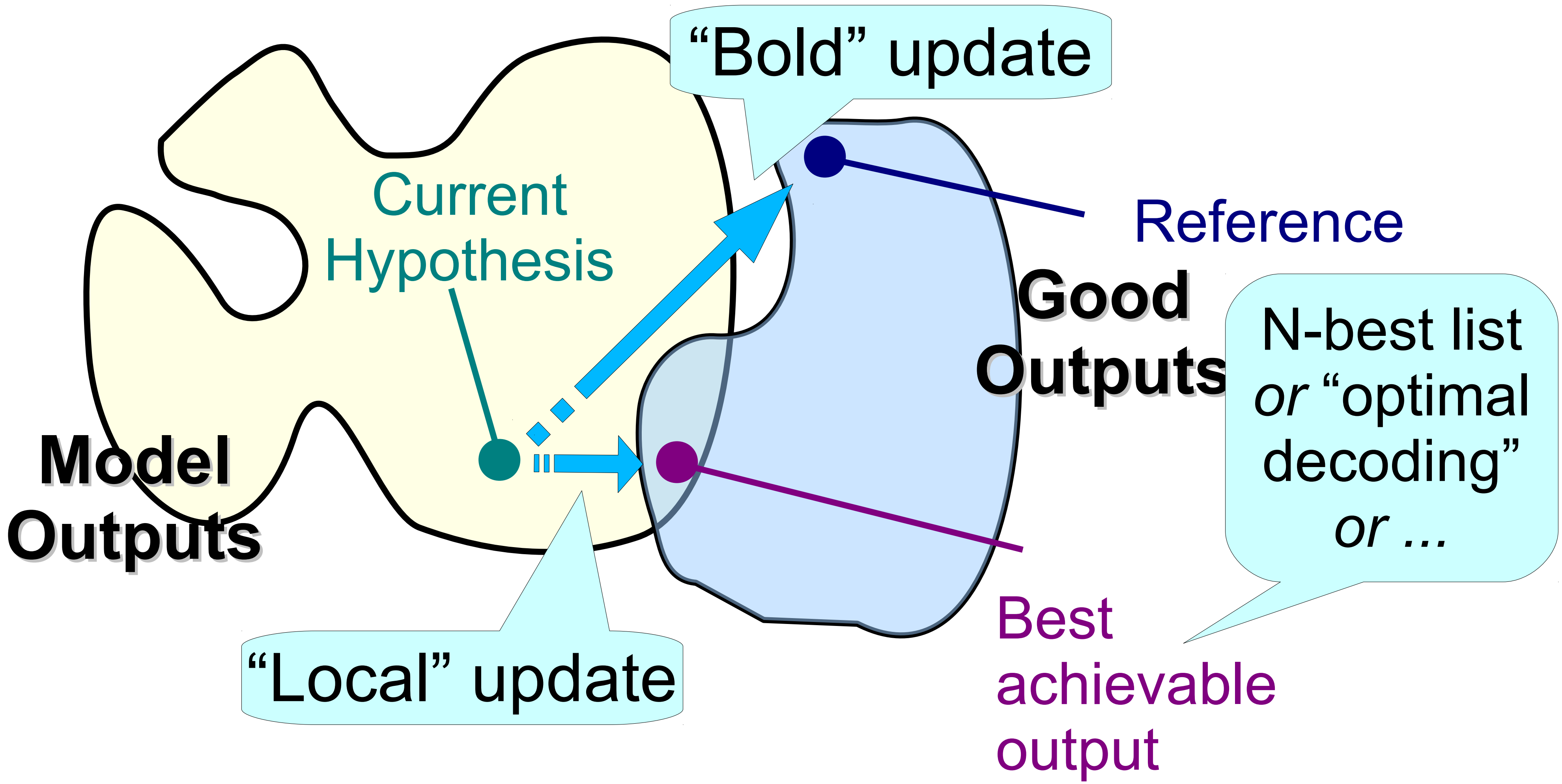
- Observation:
  - We needn't use the same beam size for training and decoding
  - Varying these values independently yields:

Training Beam	Decoding Beam: 1	5	10	25	50
1	93.9	92.8	91.9	91.3	90.9
5	90.5	94.3	94.4	94.1	94.1
10	89.5	94.3	94.4	94.2	94.2
25	88.7	94.2	94.5	94.3	94.3
50	88.4	94.2	94.4	94.2	94.4

[D+Marcu, ICML05; Xu+al, JMLR09]

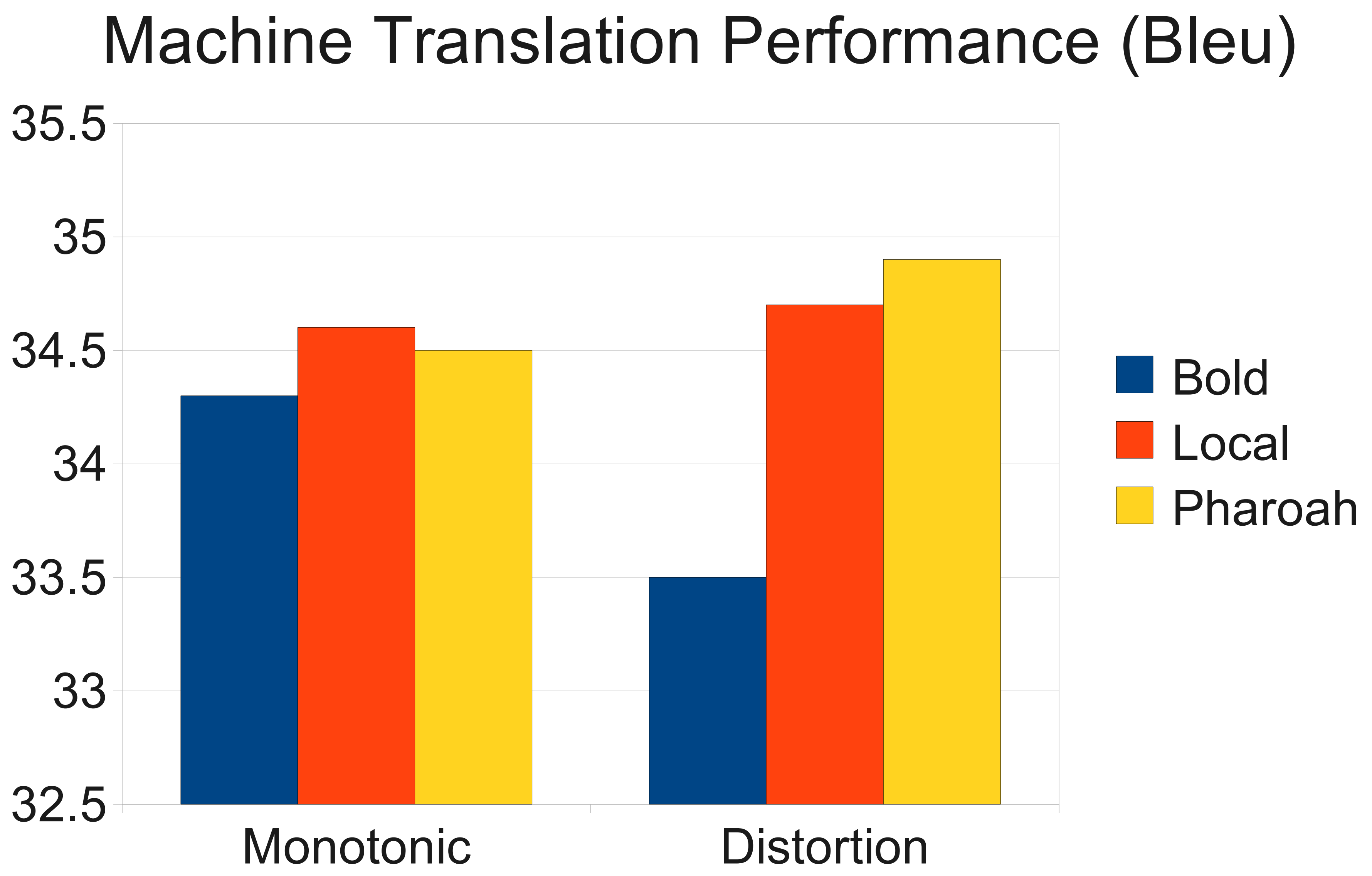
What if our model sucks?

- Sometimes our model *cannot* produce the “correct” output
  - canonical example: machine translation



[Och, ACL03; Liang+al, ACL06]

Local versus bold updating...



Refresher on Markov Decision Processes

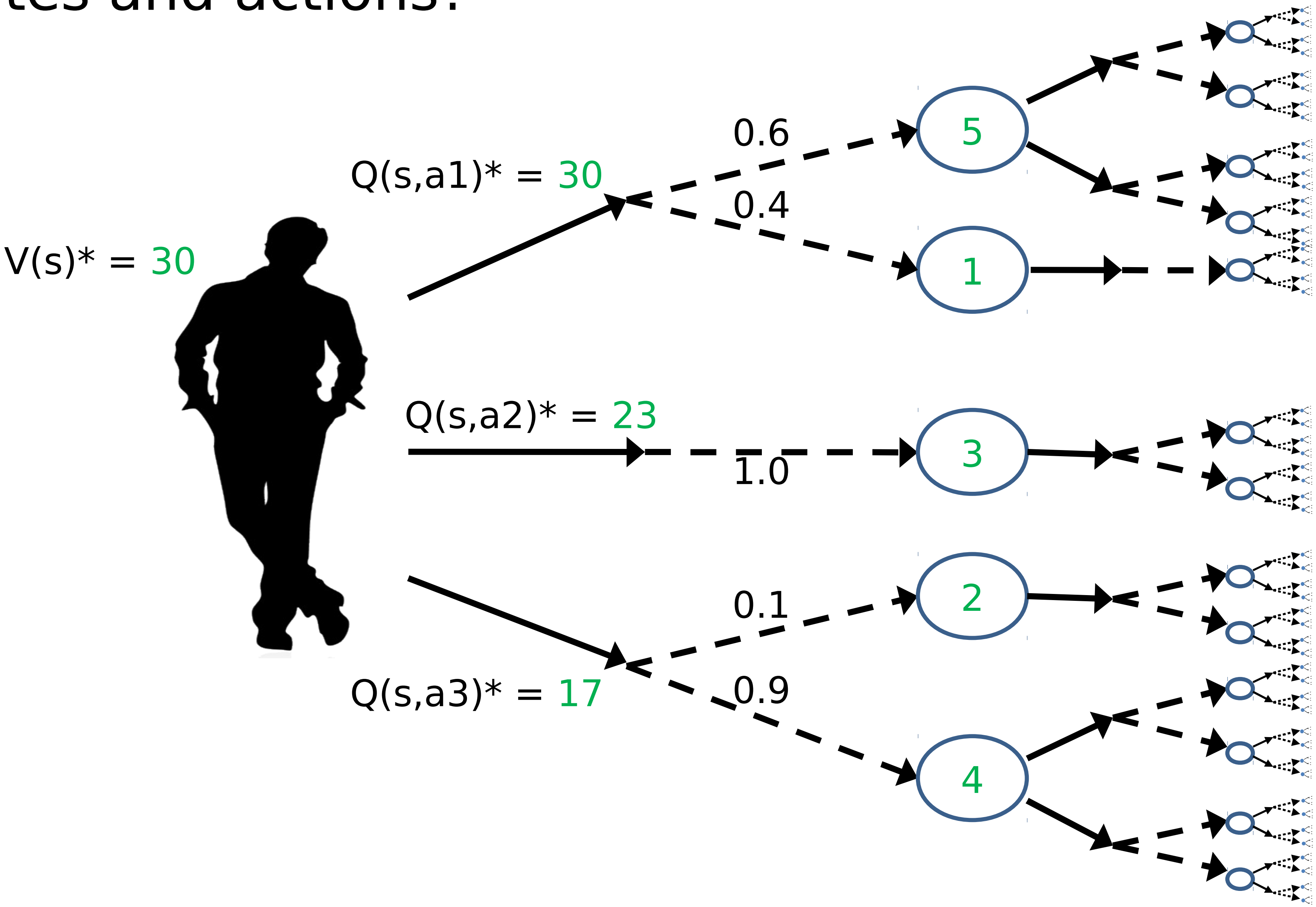
Reinforcement learning

- Basic idea:
  - Receive feedback in the form of rewards
  - Agent's utility is defined by the reward function
  - Must learn to act to maximize expected rewards
  - Change the rewards, change the learned behavior

- Examples:
  - Playing a game, reward at the end for outcome
  - Vacuuming, reward for each piece of dirt picked up
  - Driving a taxi, reward for each passenger delivered

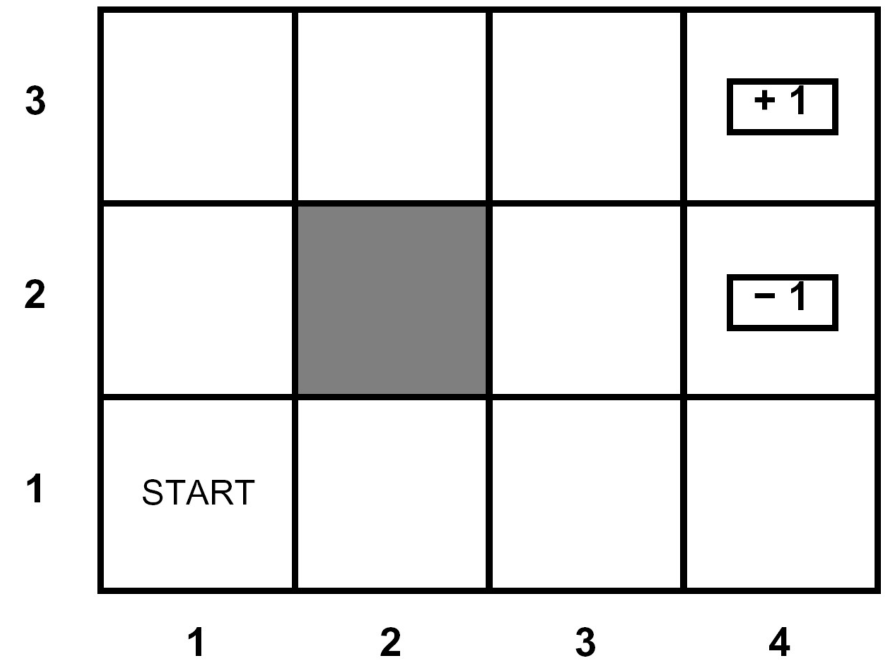
Markov decision processes

What are the values (expected future rewards) of states and actions?



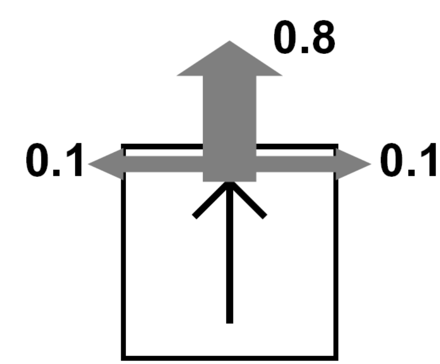
Markov Decision Processes

- An MDP is defined by:
  - A set of states s ∈ S
  - A set of actions a ∈ A
  - A transition function T(s,a,s')
    - Prob that a from s leads to s'
    - i.e., P(s' | s,a)
    - Also called the model
  - A reward function R(s, a, s')
    - Sometimes just R(s) or R(s')
  - A start state (or distribution)
  - Maybe a terminal state



- MDPs are a family of non-deterministic search problems
- Total utility is one of:

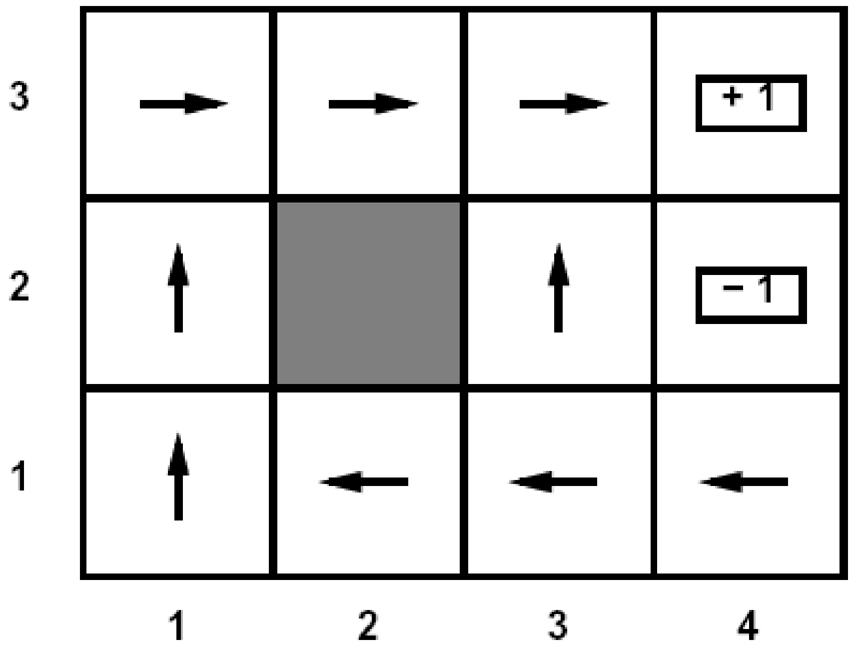
$$\sum_t r_t \text{ or } \sum_t \gamma^t r_t$$



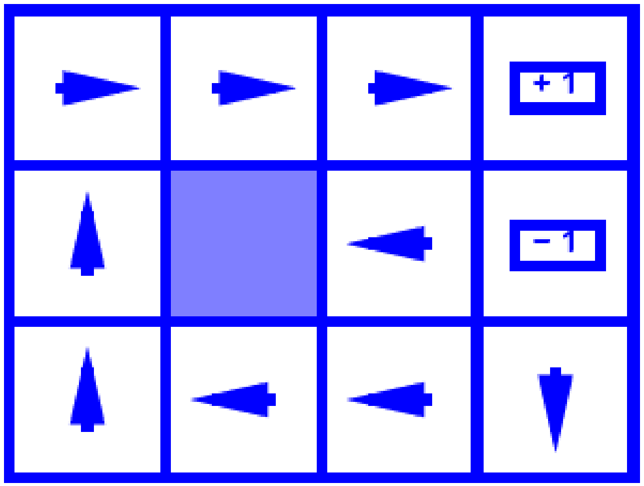
Solving MDPs

- In deterministic single-agent search problem, want an optimal plan, or sequence of actions, from start to a goal
- In an MDP, we want an optimal policy $\pi(s)$
  - A policy gives an action for each state
  - Optimal policy maximizes expected if followed
  - Defines a reflex agent

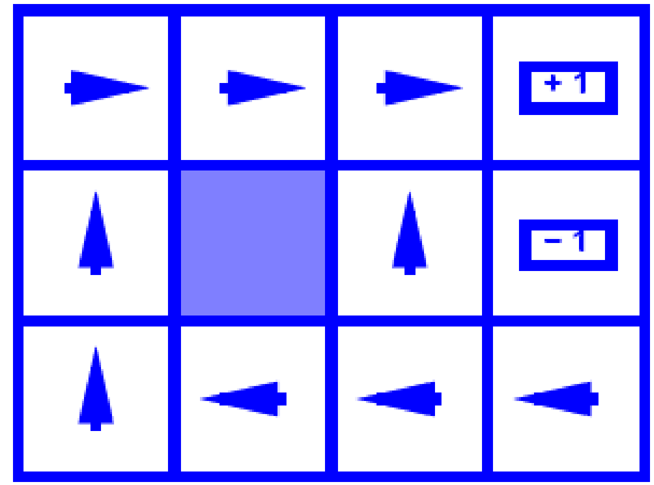
Optimal policy when R(s, a, s') = -0.04 for all non-terminals s



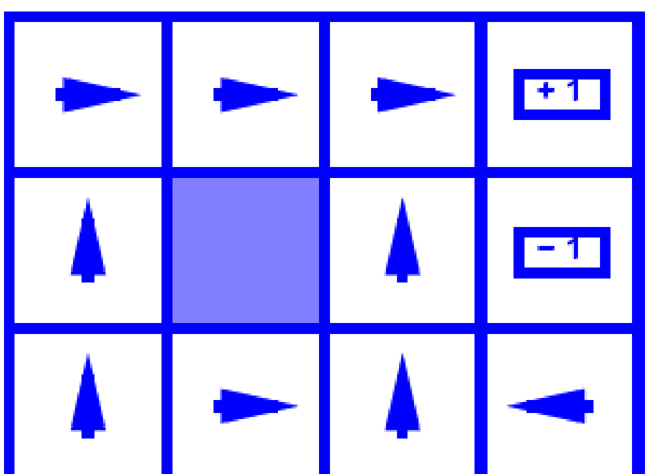
Example Optimal Policies



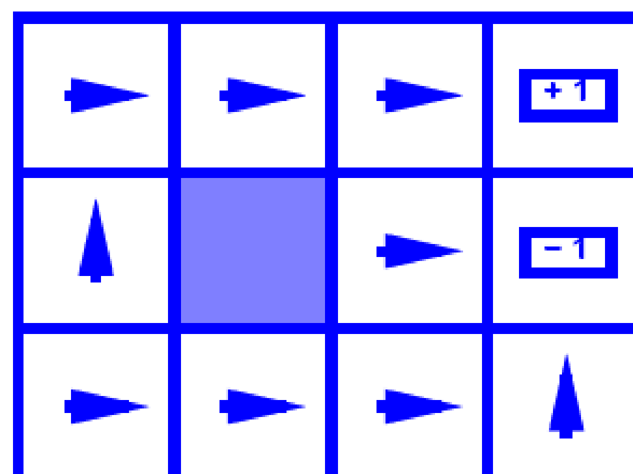
R(s) = -0.01



R(s) = -0.03



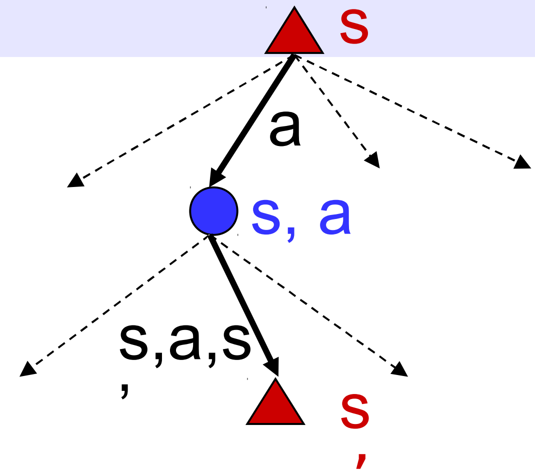
R(s) = -0.4



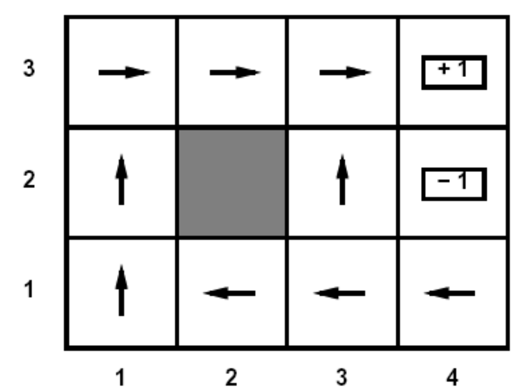
R(s) = -2.0

Optimal Utilities

- Fundamental operation: compute the optimal utilities of states s (all at once)
- Why? Optimal values define optimal policies!
- Define the utility of a state s:
  $V^*(s)$ = expected return starting in s and acting optimally
- Define the utility of a q-state (s,a):
  $Q^*(s,a)$ = expected return starting in s, taking action a and thereafter acting optimally
- Define the optimal policy:
  $\pi^*(s)$ = optimal action from state s



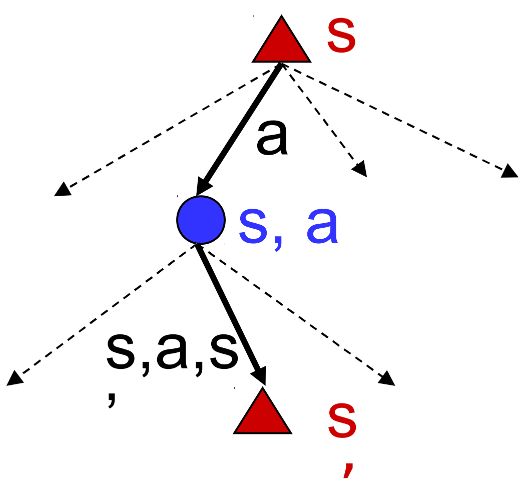
	1	2	3	4
3	0.812	0.868	0.912	+1
2	0.762		0.660	−1
1	0.705	0.655	0.611	0.388



The Bellman Equations

- Definition of utility leads to a simple one-step lookahead relationship amongst optimal utility values:

  Optimal rewards = maximize over first action and then follow optimal policy



- Formally:

$$V^*(s) = \max_a Q^*(s,a)$$

$$Q^*(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s') \right]$$

$$V^*(s) = \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^*(s') \right]$$

Solving MDPs / memoized recursion

- Recurrences:

$$V_0^*(s) = 0$$

$$V_i^*(s) = \max_a Q_i^*(s,a)$$

$$Q_i^*(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_{i-1}^*(s') \right]$$

$$\pi_i(s) = \arg\max_a Q_i^*(s,a)$$

- Cache all function call results so you never repeat work
- What happened to the evaluation function?

Q-Value Iteration

- Value iteration: iterate approx optimal values
  - Start with $V_0^*(s) = 0$, which we know is right (why?)
  - Given V_i^*, calculate the values for all states for depth i+1:

$$V_{i+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_i(s')\right]$$

- But Q-values are more useful!
  - Start with $Q_0^*(s,a) = 0$, which we know is right (why?)
  - Given Q_i^*, calculate the q-values for all q-states for depth i+1:

$$Q_{i+1}(s,a) \leftarrow \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q_i(s',a')\right]$$

RL = Unknown MDPs

- If we *knew* the MDP (i.e., the reward function and transition function):
  - Value iteration leads to optimal values
  - Will always converge to the truth

- Reinforcement learning is what we do when we *do not know* the MDP
  - All we observe is a *trajectory*
  - $(s_1,a_1,r_1,\quad s_2,a_2,r_2,\quad s_3,a_3,r_3,\quad \ldots)$

- Many algorithms exist for this problem; see Sutton+Barto's excellent book!

Q-Learning

- Learn Q*(s,a) values
  - Receive a sample **(s,a,s',r)**
  - Consider your old estimate: $Q(s,a)$
  - Consider your new sample estimate:

$$Q^*(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q^*(s',a') \right]$$

  - Incorporate the new estimate into a running average:

$$sample = R(s,a,s') + \gamma \max_{a'} Q(s',a')$$

$$Q(s,a) \leftarrow (1-\alpha) Q(s,a) + (\alpha) \left[sample \right]$$

Exploration / Exploitation

- Several schemes for forcing exploration
  - Simplest: random actions (ε greedy)
    - Every time step, flip a coin
    - With probability ε, act randomly
    - With probability 1-ε, act according to current policy
  - Problems with random actions?
    - You do explore the space, but keep thrashing around once learning is done
    - One solution: lower ε over time
    - Another solution: exploration functions

Q-Learning

- In realistic situations, we cannot possibly learn about every single state!
  - Too many states to visit them all in training
  - Too many states to hold the q-tables in memory

- Instead, we want to generalize:
  - Learn about some small number of training states from experience
  - Generalize that experience to new, similar states:

$$Q(s,a) = w_1 f_1(s,a) + w_2 f_2(s,a) + \ldots + w_n f_n(s,a)$$

  - Very simple stochastic updates:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\,[error]$$

$$w_i \leftarrow w_i + \alpha\,[error]\, f_i(s,a)$$

Inverse Reinforcement Learning

(aka Inverse Optimal Control)

Inverse RL: Task

- Given:
  - measurements of an agent's behavior over time, in a variety of circumstances
  - if needed, measurements of the sensory inputs to that agent
  - if available, a model of the environment.
- Determine: the reward function being optimized
- Proposed by [Kalman68]
- First solution, by [Boyd94]

Why inverse RL?

- Computational models for animal learning
  - "In examining animal and human behavior we must consider the reward function as an unknown to be ascertained through empirical investigation."
- Agent construction
  - "An agent designer [...] may only have a very rough idea of the reward function whose optimization would generate 'desirable' behavior."
  - eg., "Driving well"
- Multi-agent systems and mechanism design
  - learning opponents' reward functions that guide their actions to devise strategies against them

IRL from Sample Trajecto

Warning: need to be careful to avoid trivial solutions!

- Optimal policy available through sa
(eg., driving a car)

- Want to find *Reward* function that makes this policy look *as good as possible*

- Write $R_{\mathbf{w}}(s) = \mathbf{w}\phi(s)$ so the reward is linear

and $V_{\mathbf{w}}^{\pi}(s_0)$ be the value of the starting state

$$\max_{\mathbf{w}} \sum_{k=1}^{K} f\left(V_{\mathbf{w}}^{\pi^*}(s_0) - V_{\mathbf{w}}^{\pi_k}(s_0)\right)$$

How good does the "optimal policy" look?

How good does the some other policy look?

[Ng+Russell, ICML00]

Apprenticeship Learning via IRL

- For $t = 1,2,\ldots$
  - Inverse RL step:
    Estimate expert's reward function $R(s)= w^T\phi(s)$ such that under $R(s)$ the expert performs better than all previously found policies $\{\pi_i\}$.
  - RL step:
    Compute optimal policy π_t for the estimated reward w

[Abbeel+Ng, ICML04]

Car Driving Experiment

- No explicit reward function at all!
- Expert demonstrates proper policy via 2 min. of driving time on simulator (1200 data points).
- 5 different "driver types" tried.
- Features: which lane the car is in, distance to closest car in current lane.
- Algorithm run for 30 iterations, policy hand-picked.
- Movie Time! (Expert left, IRL right)

[Abbeel+Ng, ICML04]

“Nice” driver

“Evil” driver

Maxent IRL

Distribution over trajectories:

$$P(\zeta)$$

Match the reward of observed behavior:

$$\sum_\zeta P(\zeta)\, f_\zeta = f_{\mathbf{dem}}$$

Maximizing the **causal entropy** over trajectories given stochastic outcomes:

$$\max \mathbf{H}(\mathbf{P}(\zeta)||\mathbf{O})$$

(Condition on random uncontrolled outcomes, but only **after** they happen)

As uniform as possible

[Ziebart+al, AAAI08]

Data collection

Length
Speed
Road Type
Lanes

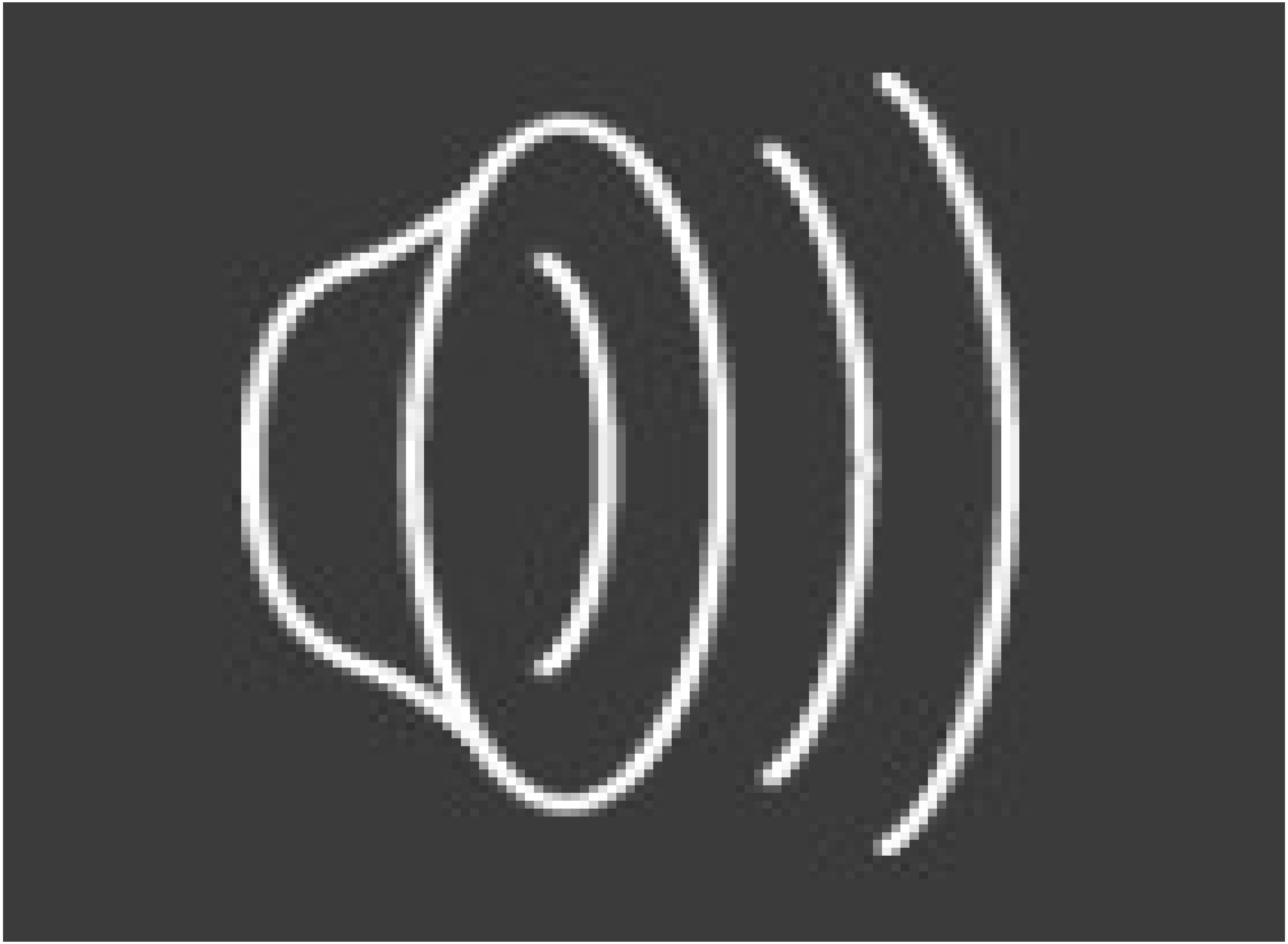
Accidents
Construction
Congestion
Time of day

25 Taxi Drivers

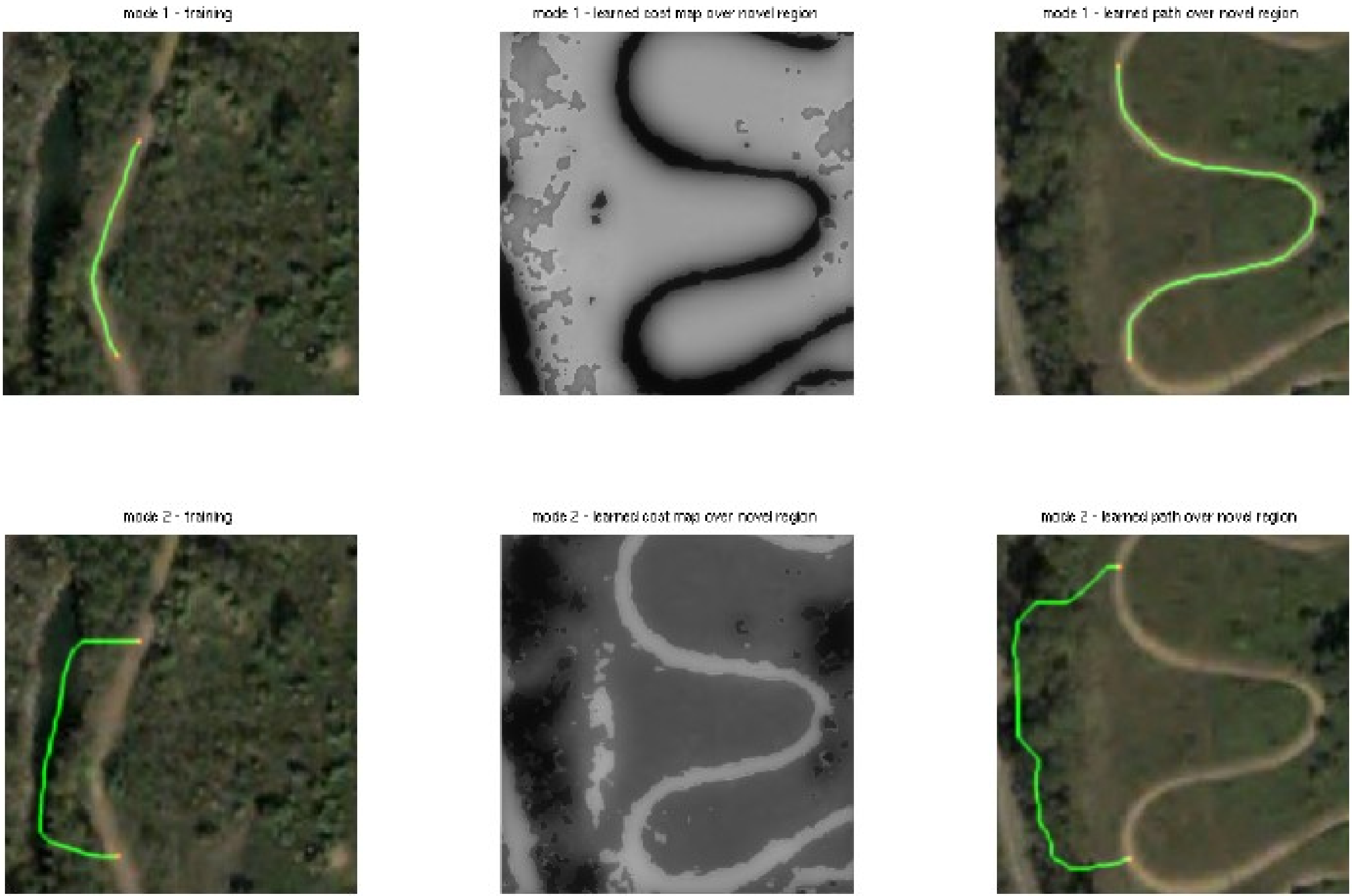
Over 100,000 miles

[Ziebart+al, AAAI08]

Predicting destinations....



Planning as structured prediction



[Ratliff+al, NIPS05]

Maximum margin planning

- Let μ(s,a) denote the probability of reaching q-state (s,a) under current model w

$$\max_{\mathbf{w}} \quad \text{margin} \quad s.t. \quad \text{planner run with w yields human output}$$

Q-state visitation frequency by human

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|^2 \quad s.t. \quad \mu(s,a)\mathbf{w}\cdot\phi(x_n,s,a) - \hat{\mu}(s,a)\mathbf{w}\cdot\phi(x_n,s,a) \geq 1 \quad , \quad \forall n,s,a$$

Q-state visitation frequency by planner

All trajectories, and all q-states

[Ratliff+al, NIPS05]

Optimizing MMP

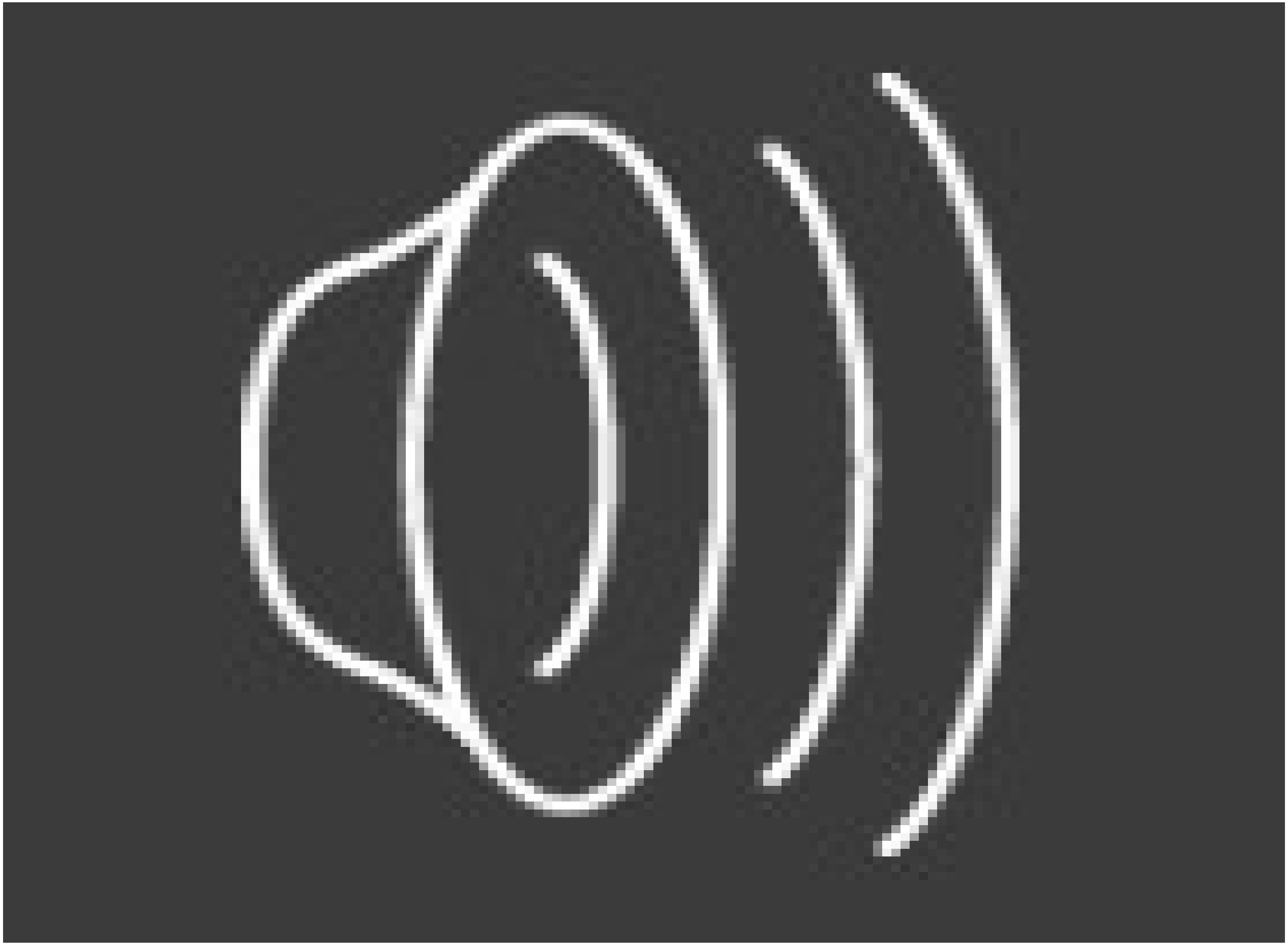
MMP Objective

SOME MATH

- For n=1..N:
  - Augmented planning:
    Run A* on current (augmented) cost map
    to get q-state visitation frequencies $\mu(s,a)$
  - Update: $\mathbf{w} = \mathbf{w} + \sum_s \sum_a \left[\hat{\mu}(s,a) - \mu(s,a)\right] \phi(x_n, s, a)$
  - Shrink: $\mathbf{w} = \left(1 - \frac{1}{CN}\right)\mathbf{w}$

[Ratliff+al, NIPS05]

Maximum margin planning movies



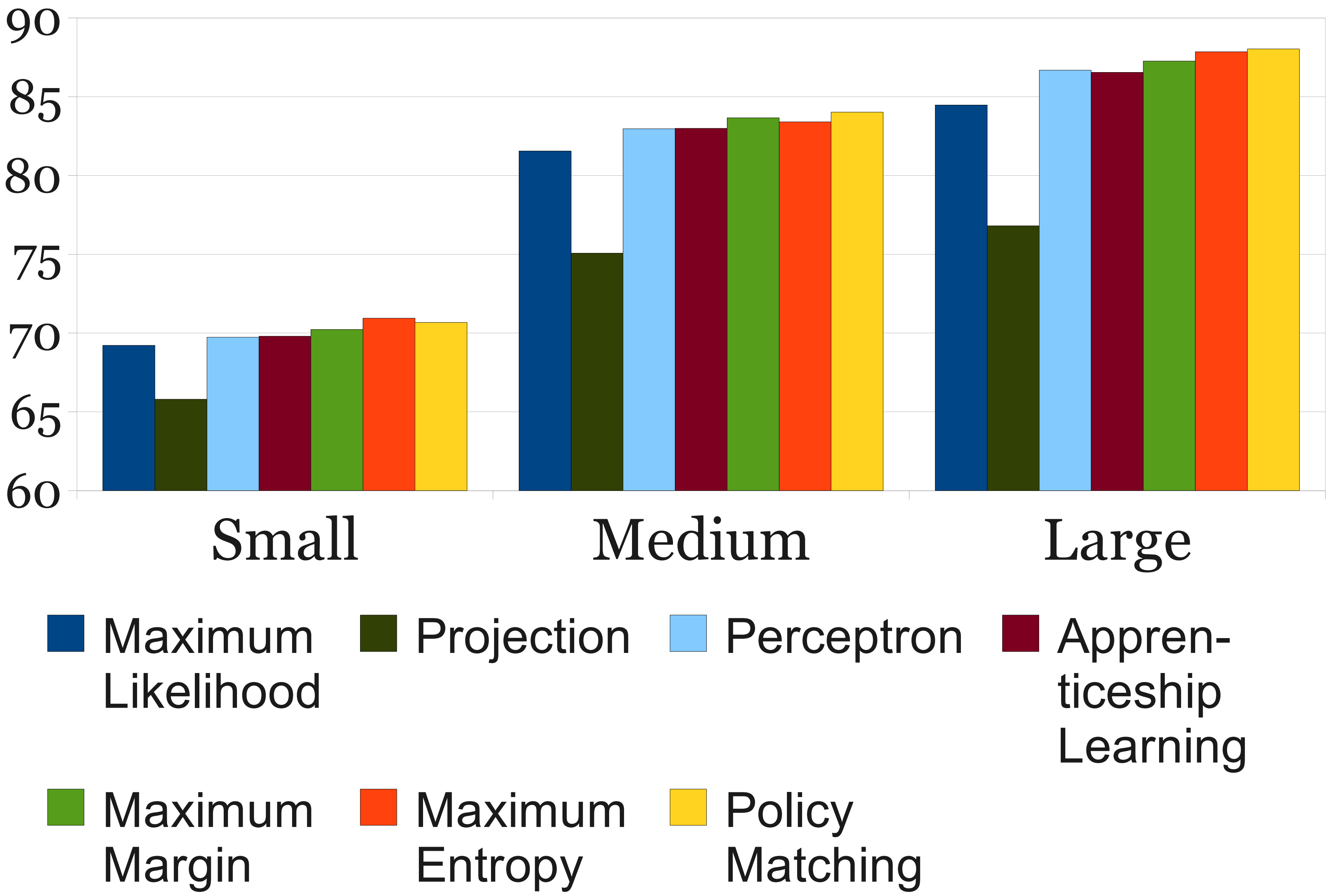
[Ratliff+al, NIPS05]

Parsing via inverse optimal control

- State space = all partial parse trees over the full sentence labeled “S”
- Actions: take a partial parse and split it anywhere in the middle
- Transitions: obvious
- Terminal states: when there are no actions left
- Reward: parse score at completion

[Neu+Szepevari, MLJ09]

Parsing via inverse optimal control



[Neu+Szepevari, MLJ09]

Learning by Demonstration

Integrating search and learning

Input: Le homme mange l' croissant.
Output: The man ate a croissant.

Hyp: The man ate
Cov: Le homme mange
l' croissant.

Classifier 'h'

Hyp: The man ate a croissant
Cov: Le homme mange
l' croissant.

Hyp: The man ate a fox
Cov: Le homme mange
l' croissant.

Hyp: The man ate happy
Cov: Le homme mange
l' croissant.

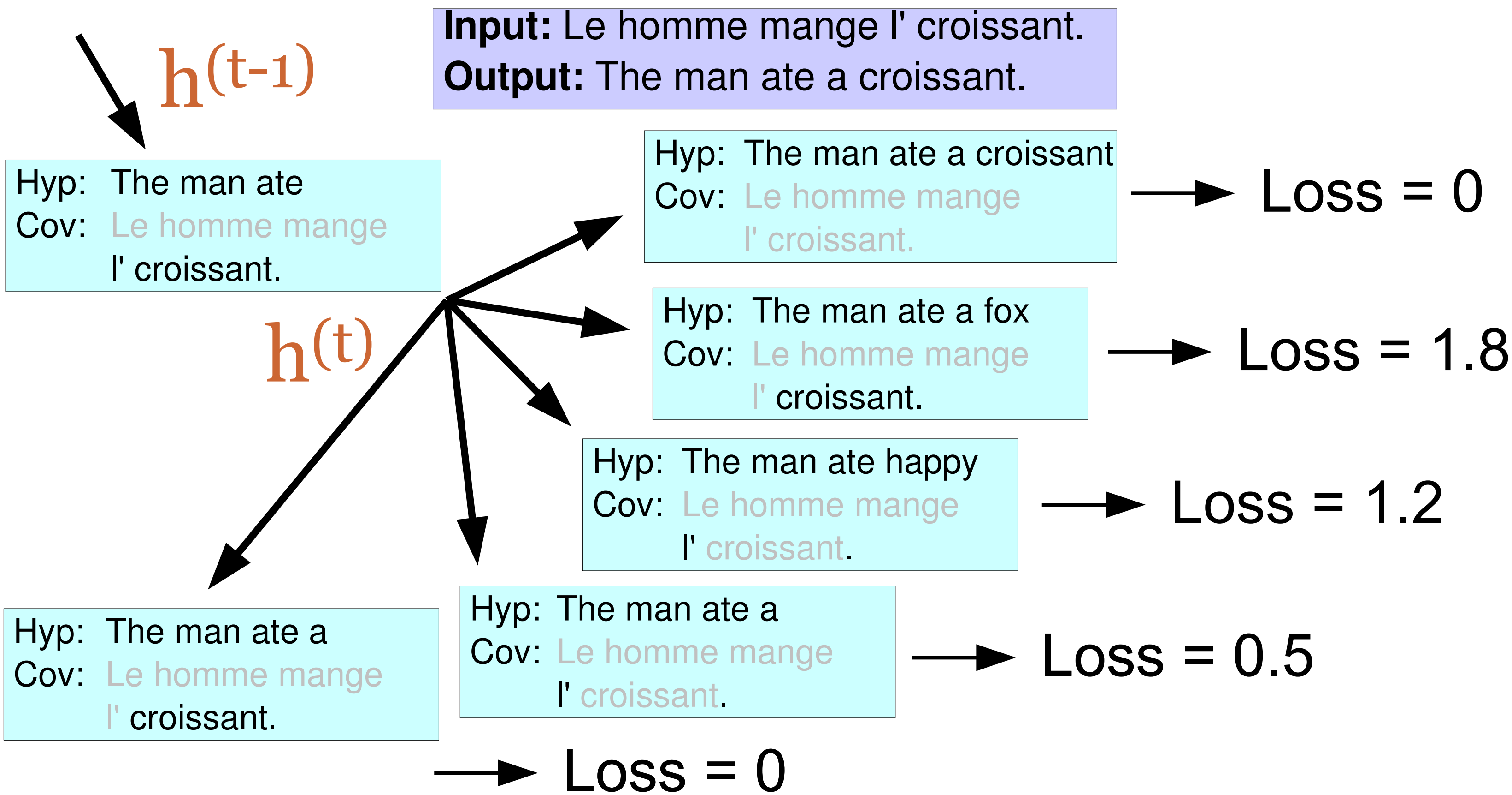
Hyp: The man ate a
Cov: Le homme mange
l' croissant.

Hyp: The man ate a
Cov: Le homme mange
l' croissant.

[D+Marcu, ICML05; D+Langford+Marcu, MLJ09]

Reducing search to classification

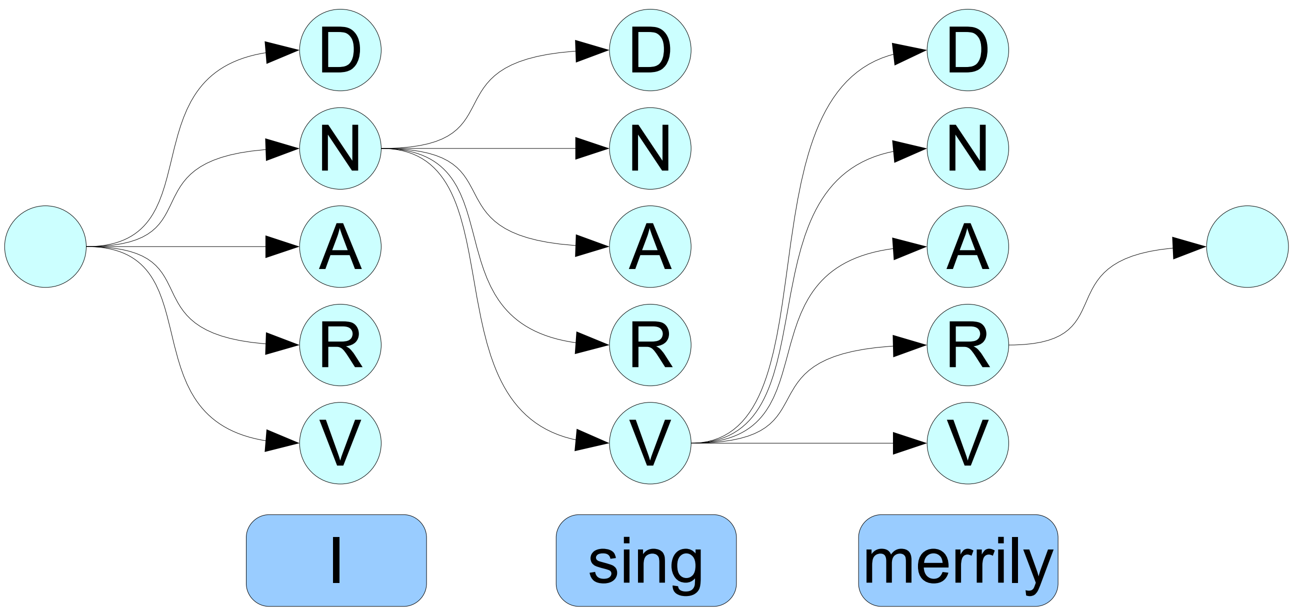
- Natural chicken and egg problem:
  - Want *h* to get low expected future loss
  - … on future decisions made by *h*
  - … and starting from states visited by *h*
- Iterative solution



[D+Langford+Marcu, MLJ09]

Reduction for Structured Prediction

- Idea: view structured prediction in light of search



Each step here looks like it could be represented as a weighted multi-class problem.

Can we formalize this idea?

Loss function:

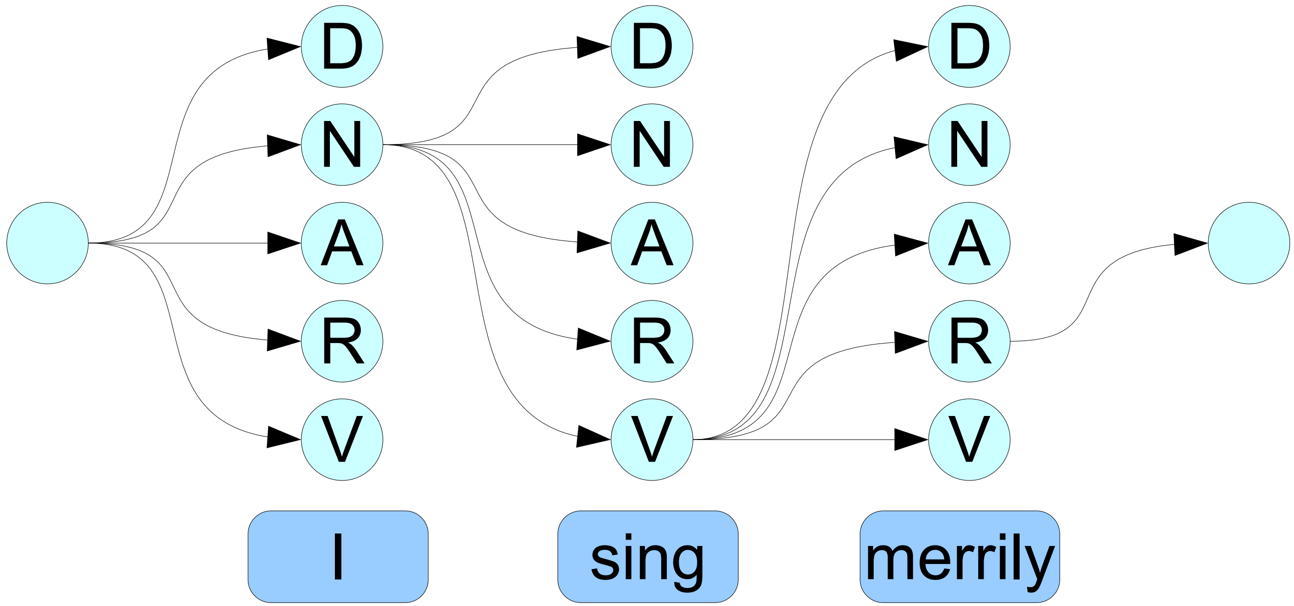
L([N V R], [N V R]) = 0

L([N V R], [N V V]) = 1/3

...

[D+Langford+Marcu, MLJ09]

Reducing Structured Prediction



Desired: good *policy* on test data (i.e., given only input string)

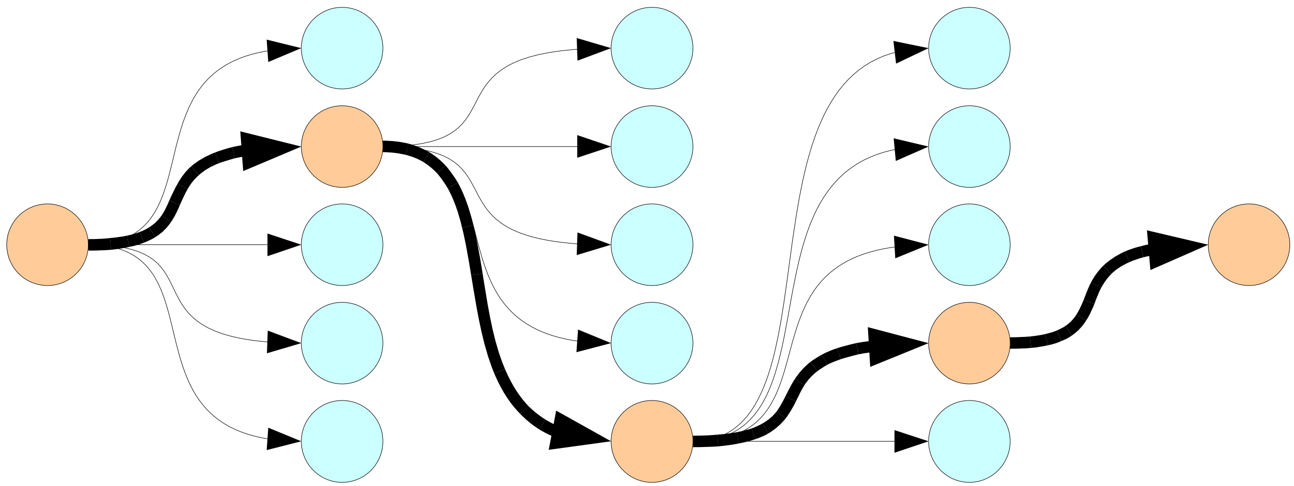
Key Assumption: *Optimal Policy for training data*

Given: input, true output and state;
Return: best successor state

Weak!

[D+Langford+Marcu, MLJ09]

How to Learn in Search

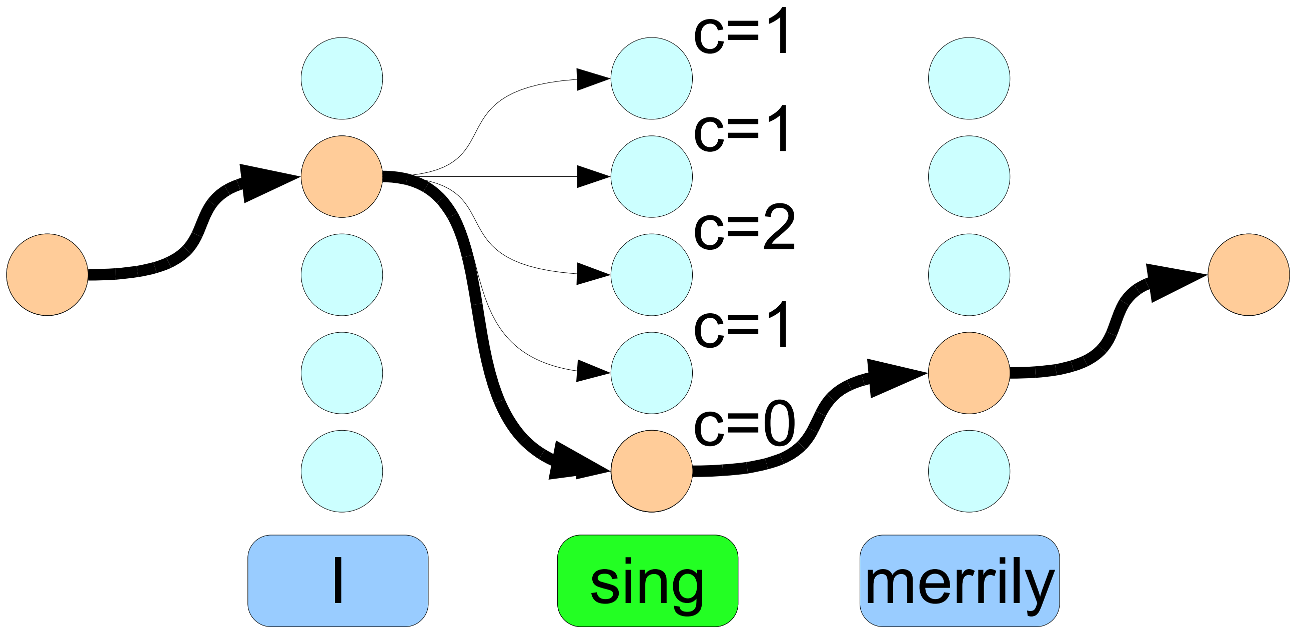


[D+Langford+Marcu, MLJ09]

Idea: Train based only on optimal path (ala MEMM)

Better Idea: Train based only on optimal policy,
then train based on optimal policy + a little learned policy
then train based on optimal policy + a little more learned policy
then ...
eventually only use learned policy

How to Learn in Search



- Translating D^{SP} into Searn(D^{SP}, loss, π):
  - Draw $x \sim D^{SP}$
  - Run π on x, to get a path
  - Pick position uniformly on path
  - Generate example with costs given by expected (wrt π) completion costs for "loss"

[D+Langford+Marcu, MLJ09]

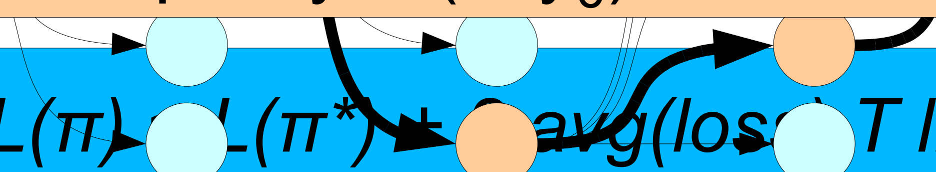
Searn

Algorithm: Searn-Learn(A, D^{SP}, loss, π^*, β)

1: *Initialize*: $\pi = \pi^*$
2: **while** not converged **do**
3: *Sample*: D ~ Searn(D^{SP}, loss, π)
4: *Learn*: h ← A(D)
5: *Update*: $\pi \leftarrow (1-\beta)\,\pi + \beta\, h$
6: **end while**
7: **return** π *without* π^*

Ingredients for Searn:

Input space (*X*) and output space (*Y*), data from *X*
Loss function (*loss(y, y')*) and features
"Optimal" policy $\pi^*(x, y_0)$



Theorem: $L(\pi) \le L(\pi^*) + 2\, avg(loss)\, T \ln T + c(1+\ln T) / T$

[D+Langford+Marcu, MLJ09]

But what about demonstrations?

- What did we assume before?

Key Assumption:
Optimal Policy for training data

Given: input, true output and state;
Return: best successor state

- We can have a *human* (or system) demonstrate, thus giving us an *optimal policy*

3d racing game (TuxKart)

Input:

Resized to 25x19 pixels (1425 features)

Output:

Steering in [-1,1]

[Ross+Gordon+Bagnell, AIStats11]

DAgger: Dataset Aggregation

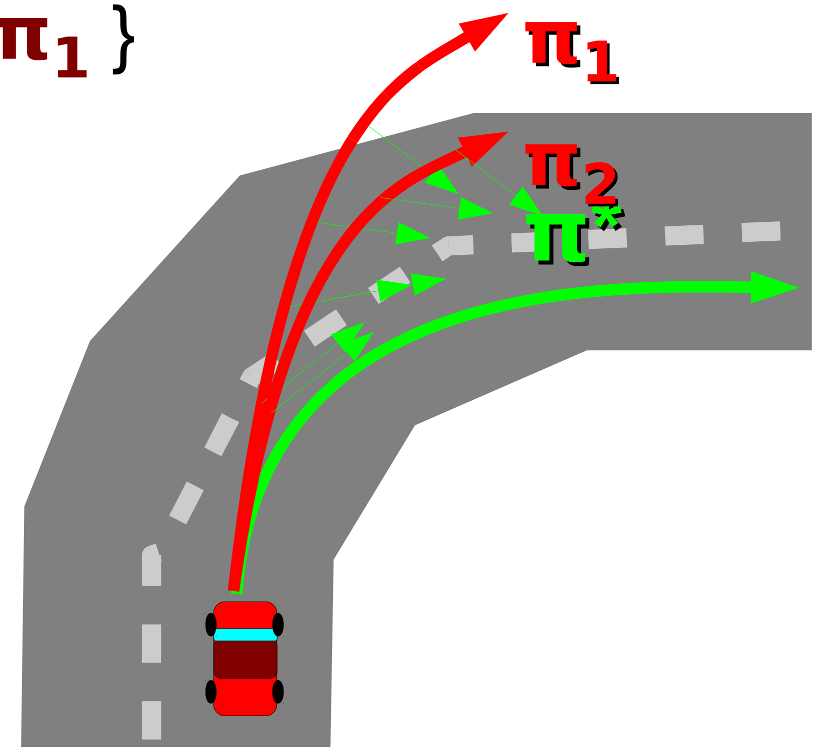
- Collect trajectories from expert π^*
- Dataset $\mathbf{D_0} = \{ (s, \pi^*(s)) \mid s \sim \pi^* \}$
- Train π_1 on $\mathbf{D_0}$
- Collect new trajectories from π_1
  - But let the *expert* steer!
- Dataset $\mathbf{D_1} = \{ (s, \pi^*(s)) \mid s \sim \pi_1 \}$
- Train π_2 on $\mathbf{D_0} \cup \mathbf{D_1}$

- In general:
  - $\mathbf{D_n} = \{ (s, \pi^*(s)) \mid s \sim \pi_n \}$
  - Train π_n on $\cup_{i<n} \mathbf{D_i}$

If $N = T \log T$,

$$\mathbf{L(\pi_n) < T\, \epsilon_N + O(1)}$$

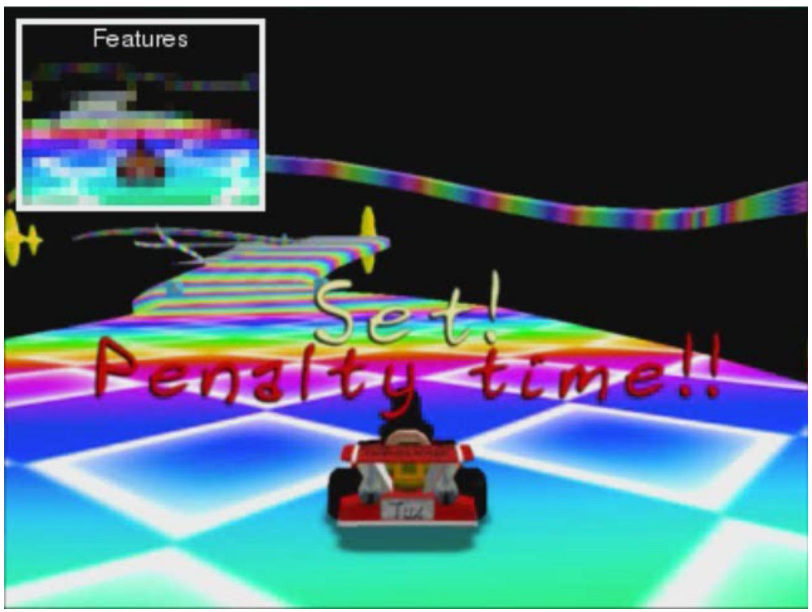
for some n



[Ross+Gordon+Bagnell, AIStats11]

Experiments: Racing Game

Input:



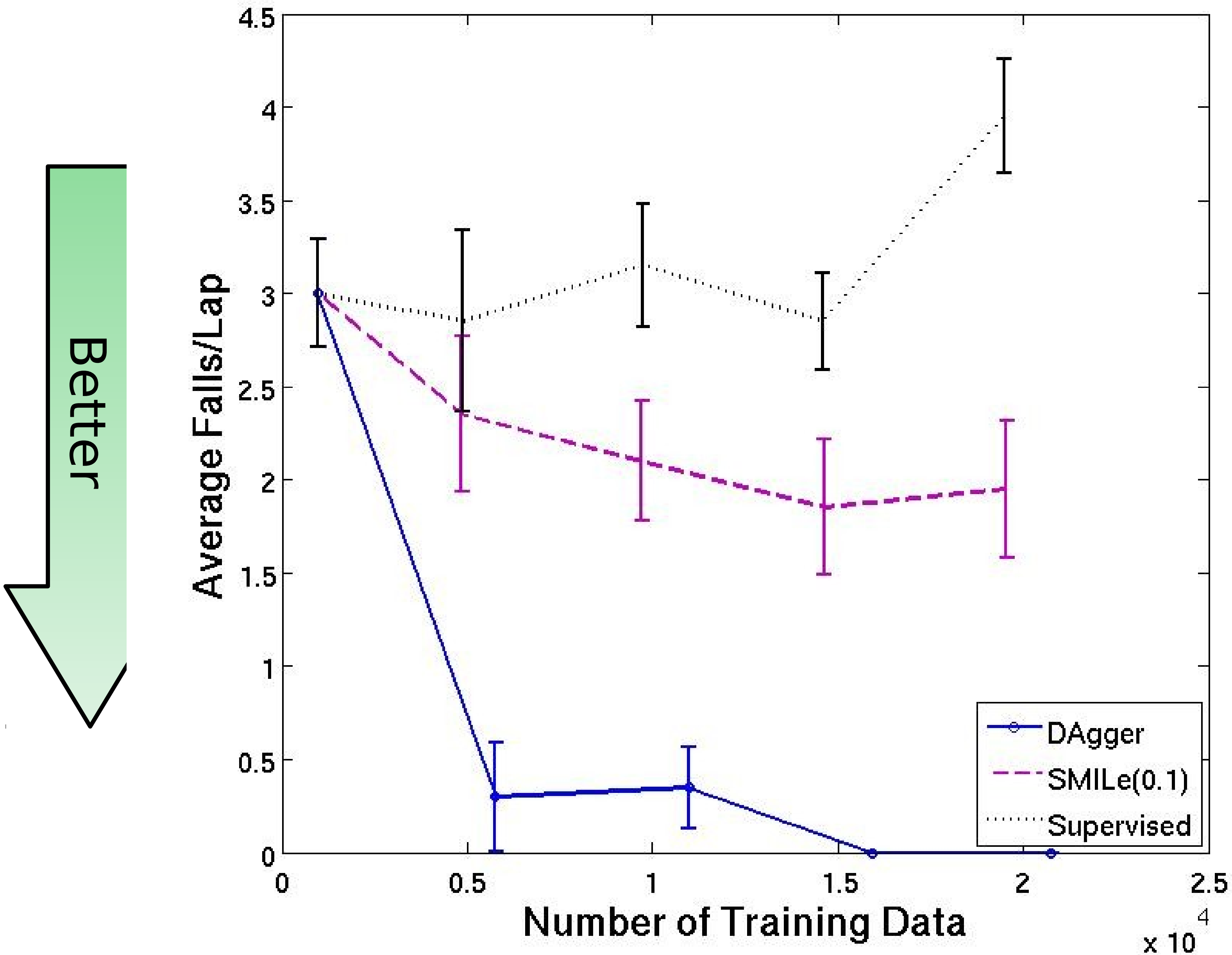
Resized to 25x19 pixels (1425 features)

Output:

Steering in [-1,1]

[Ross+Gordon+Bagnell, AIStats11]

Average falls per lap

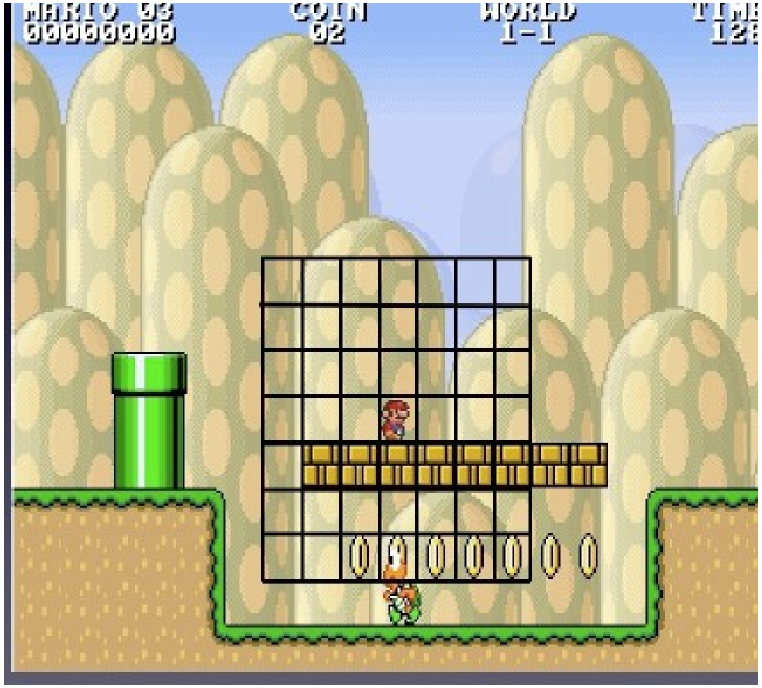


[Ross+Gordon+Bagnell, AIStats11]

Super Mario Bros.

From Mario AI competition 2009

Input:



Output:

Jump in {0,1}
Right in {0,1}
Left in {0,1}
Speed in {0,1}

Extracted 27K+ binary features
from last 4 observations
(14 binary features for every cell)

[Ross+Gordon+Bagnell, AIStats11]

Training (expert)

[Ross+Gordon+Bagnell, AIStats11]

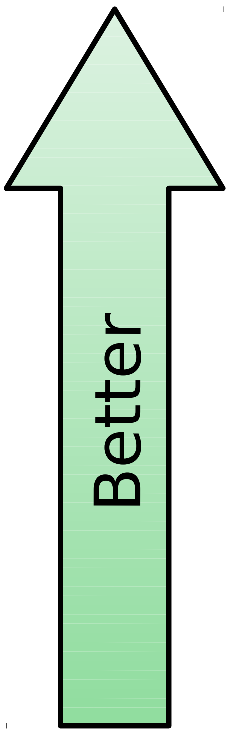
Test-time execution (classifier)

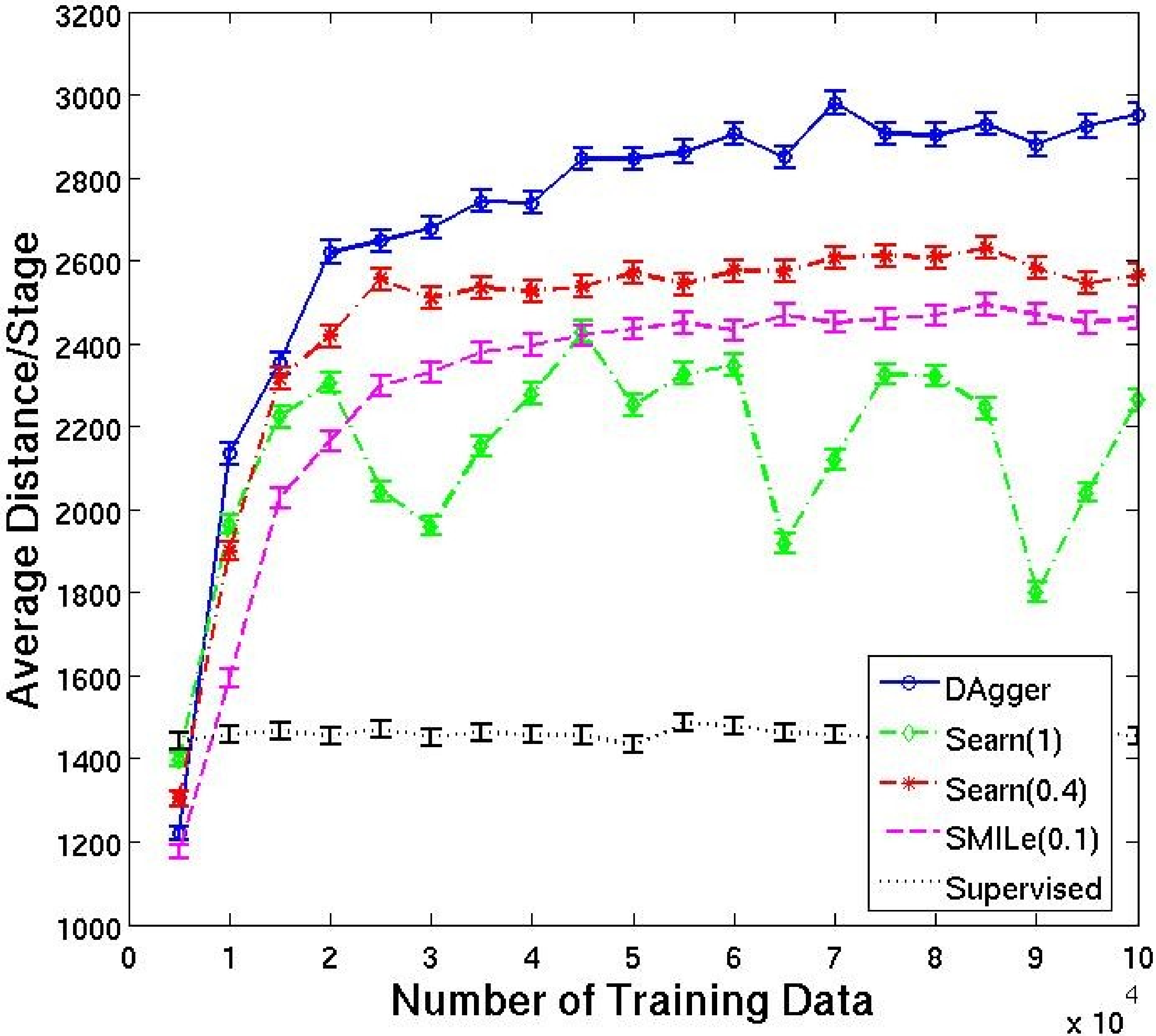
[Ross+Gordon+Bagnell, AIStats11]

Test-time execution (Dagger)

[Ross+Gordon+Bagnell, AIStats11]

Average distance per stage





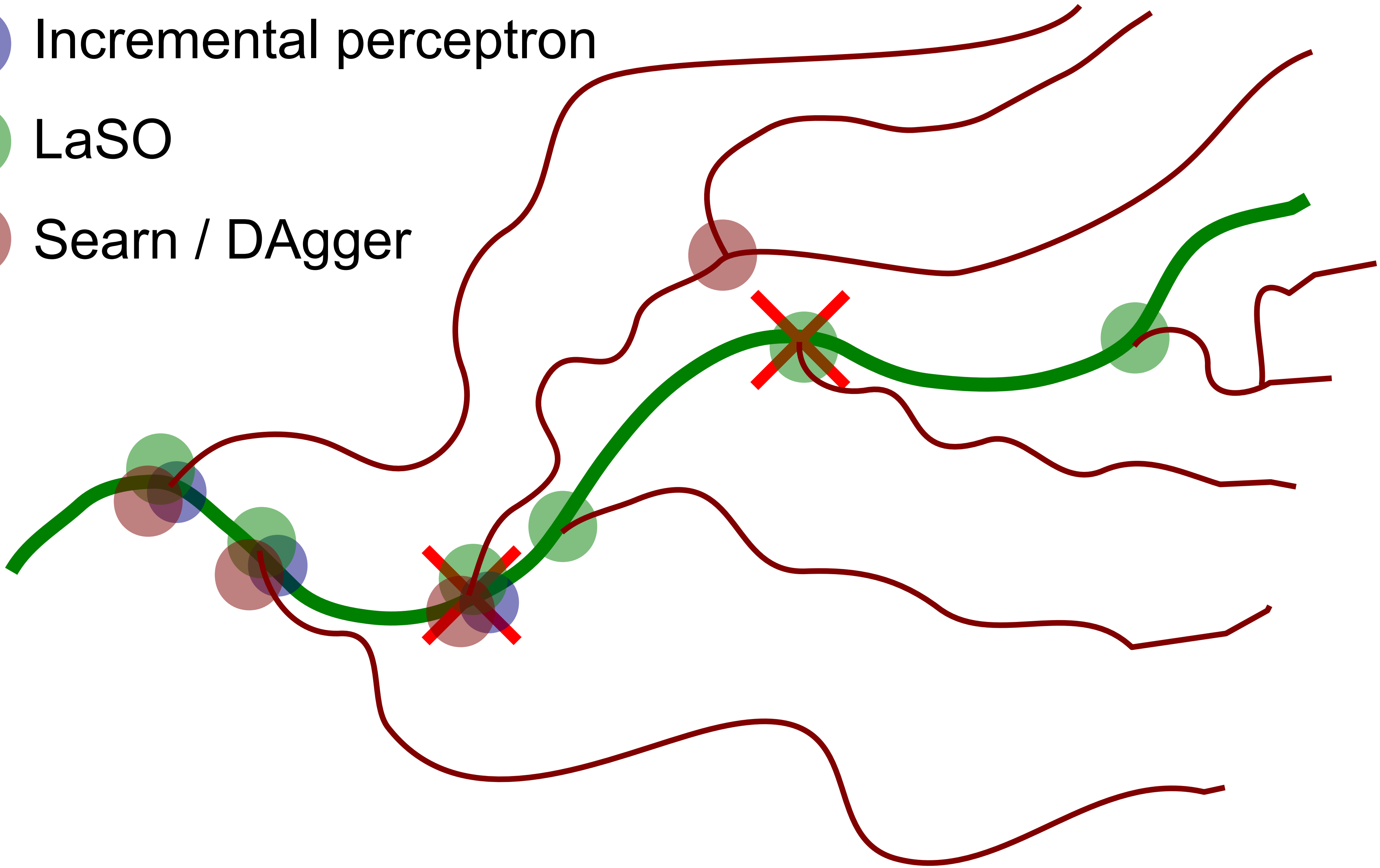
[Ross+Gordon+Bagnell, AIStats11]

Perceptron vs. LaSO vs. Searn

Incremental perceptron

LaSO

Searn / DAgger



Un-learnable decision

Discussion

Relationship between SP and IRL

- Formally, they're (nearly) the same problem
  - See humans performing some task
  - Define some loss function
  - Try to mimic the humans

- Difference is in philosophy:
  - (I)RL has little notion of beam search or dynamic programming
  - SP doesn't think about separating reward estimation from solving the prediction problem
  - (I)RL has to deal with stochastiticity in MDPs

Important Concepts

- Search and loss-augmented search for margin-based methods
- Bold versus local updates for approximate search
- Training on-path versus off-path
- Stochastic versus deterministic worlds
- Q-states / values
- Learning reward functions vs. matching behavior

Hal's Wager

- Give me a structured prediction problem where:
  - Annotations are at the lexical level
  - Humans can do the annotation with reasonable agreement
  - You give me a few thousand labeled sentences
- Then I can learn reasonably well...
  - ...using one of the algorithms we talked about
- Why do I say this?
  - Lots of positive experience
  - I'm an optimist
  - I want your *counter-examples!*

Open problems

- How to do SP when argmax is intractable....
  - Bad: simple algorithms diverge [Kulesza+Pereira, NIPS07]
  - Good: some work well [Finley+Joachims, ICML08]
  - And you can make it fast! [Meshi+al, ICML10]
- How to do SP with delayed feedback (credit assignment)
  - Kinda just works sometimes [D, ICML09; Chang+al, ICML10]
  - Generic RL also works [Branavan+al, ACL09; Liang+al, ACL09]
- What role does structure actually play?
  - Little: only constraints outputs [Punyakanok+al, IJCAI05]
  - Little: only introduces non-linearities [Liang+al, ICML08]
- Role of experts?
  - what if your expert isn't actually optimal?
  - what if you have more than one expert?
  - what if you only have trajectories, not the expert?

Things I have no idea how to solve...

all : (a → Bool) → [a] → Bool

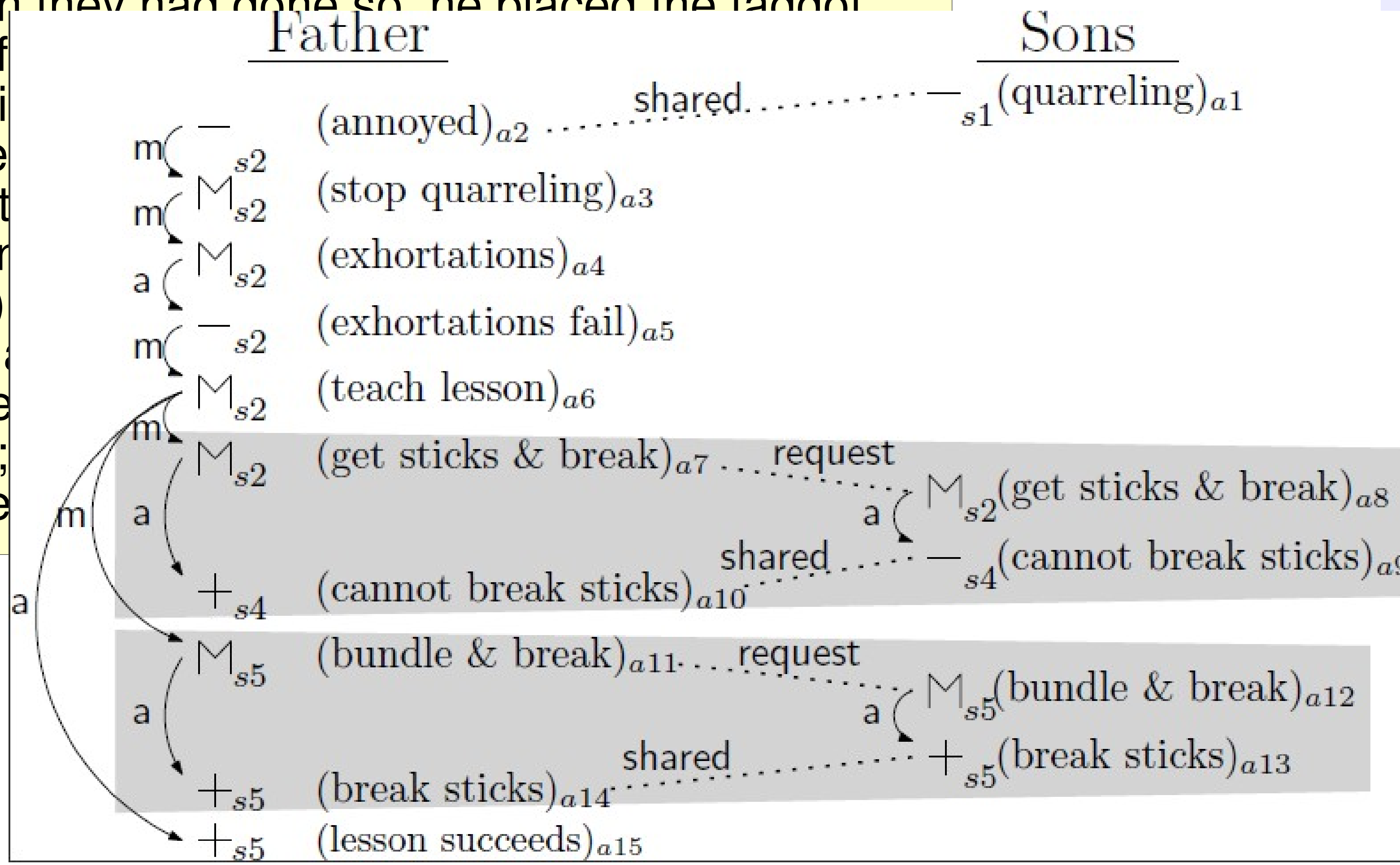
Applied to a predicate and a list, returns `True' if all elements of the list satisfy the predicate, and `False' otherwise.

```
%module main:MyPrelude
  %data main:MyPrelude.MyList aadj =
    {main:MyPrelude.Nil;
     main:MyPrelude.Cons aadj ((main:MyPrelude.MyList aadj))};
  %rec
  {main:MyPrelude.myzuall :: %forall tadA . (tadA ->
                                              ghczmprim:GHCziBool.Bool)
                                            ->
                                            (main:MyPrelude.MyList tadA) ->
                                            ghczmprim:GHCziBool.Bool =
     \ @ tadA
       (padk::tadA -> ghczmprim:GHCziBool.Bool)
       (dsddE::(main:MyPrelude.MyList tadA)) ->
         %case ghczmprim:GHCziBool.Bool dsddE
         %of (wildB1::(main:MyPrelude.MyList tadA))
           {main:MyPrelude.Nil ->
              ghczmprim:GHCziBool.True;
            main:MyPrelude.Cons
            (xadm::tadA) (xsadn::(main:MyPrelude.MyList tadA)) ->
              %case ghczmprim:GHCziBool.Bool (padk xadm)
              %of (wild1Xc::ghczmprim:GHCziBool.Bool)
                {ghczmprim:GHCziBool.False ->
                   ghczmprim:GHCziBool.False;
                 ghczmprim:GHCziBool.True ->
                   main:MyPrelude.myzuall @ tadA padk xsadn}}};
```

```
all p
all p
  if p
     th
     el
```

Things I have no idea how to solve...

(s1) A father had a family of sons who were perpetually quarreling among themselves. (s2) When he failed to heal their disputes by his exhortations, he determined to give them a practical illustration of the evils of disunion; and for this purpose he one day told them to bring him a bundle of sticks. (s3) When they had done so, he placed the faggot
into the hands of
them to break it i
strength, and we
the faggot, took t
again put them in
them easily. (s6)
"My sons, if you a
other, you will be
of your enemies;
you will be broke



Software

- Sequence labeling
  - Mallet http://mallet.cs.umass.edu
  - CRF++ http://crfpp.sourceforge.net

- Search-based structured prediction
  - LaSO http://hal3.name/TagChunk
  - Searn http://hal3.name/searn

- Higher-level “feature template” approaches
  - Alchemy http://alchemy.cs.washington.edu
  - Factorie http://code.google.com/p/factorie

Summary

- Structured prediction is *easy* if you can do argmax search (esp. loss-augmented!)
- Label-bias can kill you, so iterate (Searn/Dagger)
- Stochastic worlds modeled by MDPs
- IRL is all about learning reward functions
- IRL has fewer assumptions
  - More general
  - Less likely to work on easy problems
- We're a long way from a complete solution
- Hal's wager: we can learn pretty much anything

Thanks! Questions?

References

See also:

http://www.cs.utah.edu/~suresh/mediawiki/index.php/MLRG
http://braque.cc/ShowChannel?handle=P5BVAC34

Stuff we talked about explicitly

- *Apprenticeship learning via inverse reinforcement learning, P. Abbeel and A. Ng. ICML, 2004.*
- *Incremental parsing with the Perceptron algorithm. M. Collins and B. Roark. ACL 2004.*
- *Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms.* M. Collins. EMNLP 2002.
- *Search-based Structured Prediction.* H. Daumé III, J. Langford and D. Marcu. Machine Learning, 2009.
- *Learning as Search Optimization: Approximate Large Margin Methods for Structured Prediction.* H. Daumé III and D. Marcu. ICML, *2005.*
- *An End-to-end Discriminative Approach to Machine Translation.* P. Liang, A. Bouchard-Côté, D. Klein, B. Taskar. ACL 2006.
- *Statistical Decision-Tree Models for Parsing. D. Magerman. ACL 1995.*
- *Training Parsers by Inverse Reinforcement Learning. G. Neu and Cs. Szepesvári. Machine Learning 77, 2009.*
- *Algorithms for inverse reinforcement learning, A. Ng and A. Russell. ICML, 2000.*
- *(Online) Subgradient Methods for Structured Prediction.* N. *Ratliff, J. Bagnell, and M. Zinkevich. AIStats 2007.*
- *Maximum margin planning.* N. Ratliff, J. Bagnell and M. Zinkevich. ICML, 2006.
- *Learning to search: Functional gradient techniques for imitation learning.* N. *Ratliff, D. Silver, and J. Bagnell. Autonomous Robots, Vol. 27, No. 1, July, 2009.*
- *Reduction of Imitation Learning to No-Regret Online Learning.* S. Ross, G. Gordon and J. Bagnell. AIStats 2011.
- *Max-Margin Markov Networks.* B. Taskar, C. Guestrin, V. Chatalbashev and D. Koller. JMLR 2005.
- *Large Margin Methods for Structured and Interdependent Output Variables.* I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. JMLR 2005.
- *Learning Linear Ranking Functions for Beam Search with Application to Planning. Y.* Xu, A. Fern, and S. Yoon. JMLR 2009.
- *Maximum Entropy Inverse Reinforcement Learning.* B. Ziebart, A. Maas, J. Bagnell, and A. Dey. AAAI 2008.

Other good stuff

- *Reinforcement learning for mapping instructions to actions.* S.R.K. Branavan, H. Chen, L. Zettlemoyer and R. Barzilay. ACL, 2009.
- *Driving semantic parsing from the world's response.* J. Clarke, D. Goldwasser, M.-W. Chang, D. Roth. CoNLL 2010.
- *New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. M.*Collins and N. Duffy. *ACL 2002.*
- *Unsupervised Search-based Structured Prediction.* H. Daumé III. ICML 2009.
- *Training structural SVMs when exact inference is intractable.* T. Finley and T. Joachims. ICML, 2008.
- *Structured learning with approximate inference.* A. Kulesza and F. Pereira. NIPS, 2007.
- *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* J. Lafferty, A. McCallum, F. Pereira. ICML 2001.
- *Structure compilation: trading structure for features.* P. Liang, H. Daume, D. Klein. ICML 2008.
- *Learning semantic correspondences with less supervision.* P. Liang, M. Jordan and D. Klein. ACL, 2009.
- *Generalization Bounds and Consistency for Structured Labeling.* D. McAllester. In *Predicting Structured Data, 2007.*
- *Maximum entropy Markov models for information extraction and segmentation.* A. McCallum, D. Freitag, F. Pereira. ICML 2000.
- *FACTORIE: Efficient Probabilistic Programming for Relational Factor Graphs via Imperative Declarations of Structure, Inference and Learning.* A. McCallum, K. Rohanemanesh, M. Wick, K. Schultz, S. Singh. NIPS Workshop on Probabilistic Programming, 2008
- *Learning efficiently with approximate inference via dual losses.* O. Meshi, D. Sontag, T. Jaakkola, A. Globerson. ICML 2010.
- *Learning and inference over constrained output.* V. Punyakanok, D. Roth, W. Yih, D. Zimak. IJCAI, 2005.
- *Boosting Structured Prediction for Imitation Learning.* N. Ratliff, D. Bradley, J. Bagnell, and J. Chestnutt. NIPS 2007.
- *Efficient Reductions for Imitation Learning.* S. Ross and J. Bagnell. AISTATS, 2010.
- *Kernel Dependency Estimation*. J. Weston, O. Chapelle, A. Elisseeff, B. Schoelkopf and V. Vapnik. NIPS 2002.