Ph.D. Thesis Oral Defense

Stanford Artificial Intelligence Laboratory (SAIL)

Valentin I. Spitkovsky

Grammar Induction and Parsing with Dependency-and-Boundary Models

Problem: Unsupervised Parsing (and Grammar Induction)

Problem: Unsupervised Parsing (and Grammar Induction)

- **Input: Raw Text**

  ... By most measures, the nation's industrial sector is now growing very slowly — if at all. Factory payrolls fell in September. So did the Federal Reserve ...

Problem: Unsupervised Parsing (and Grammar Induction)

- **Input: Raw Text (Sentences, Tokens and POS-tags)**

  ... By most measures, the nation's industrial sector is now growing very slowly — if at all. Factory payrolls fell in September. So did the Federal Reserve ...

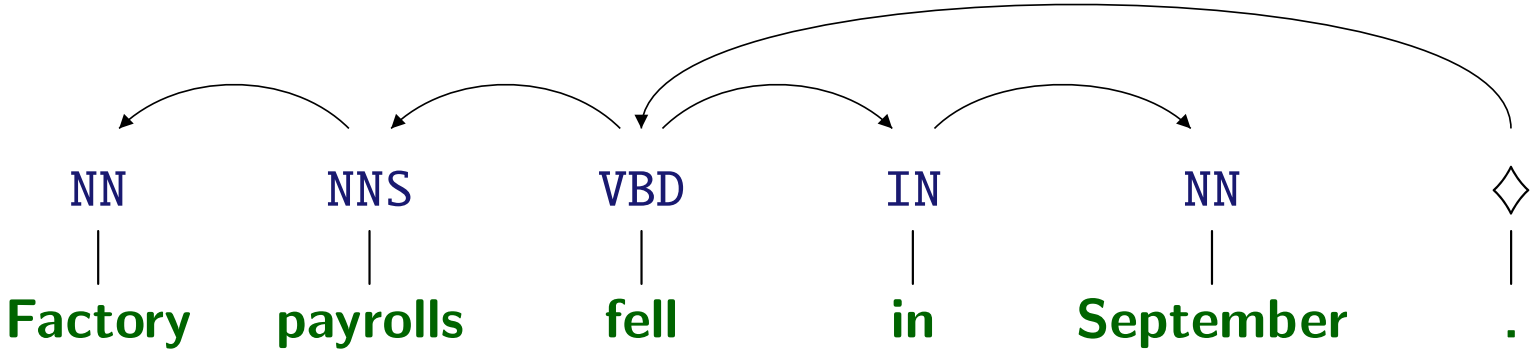
NN	NNS	VBD	IN	NN	◇
Factory	**payrolls**	**fell**	**in**	**September**	**.**

Problem: Unsupervised Parsing (and Grammar Induction)

- **Input: Raw Text (Sentences, Tokens and POS-tags)**

  ... By most measures, the nation's industrial sector is now growing very slowly — if at all. Factory payrolls fell in September. So did the Federal Reserve ...

- **Output: Syntactic Structures (and a Probabilistic Grammar)**



Motivation: Unsupervised (Dependency) Parsing

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**

Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**
- **... adjust for any data without reference tree banks: — i.e., understudied languages or genres (e.g., legal).**

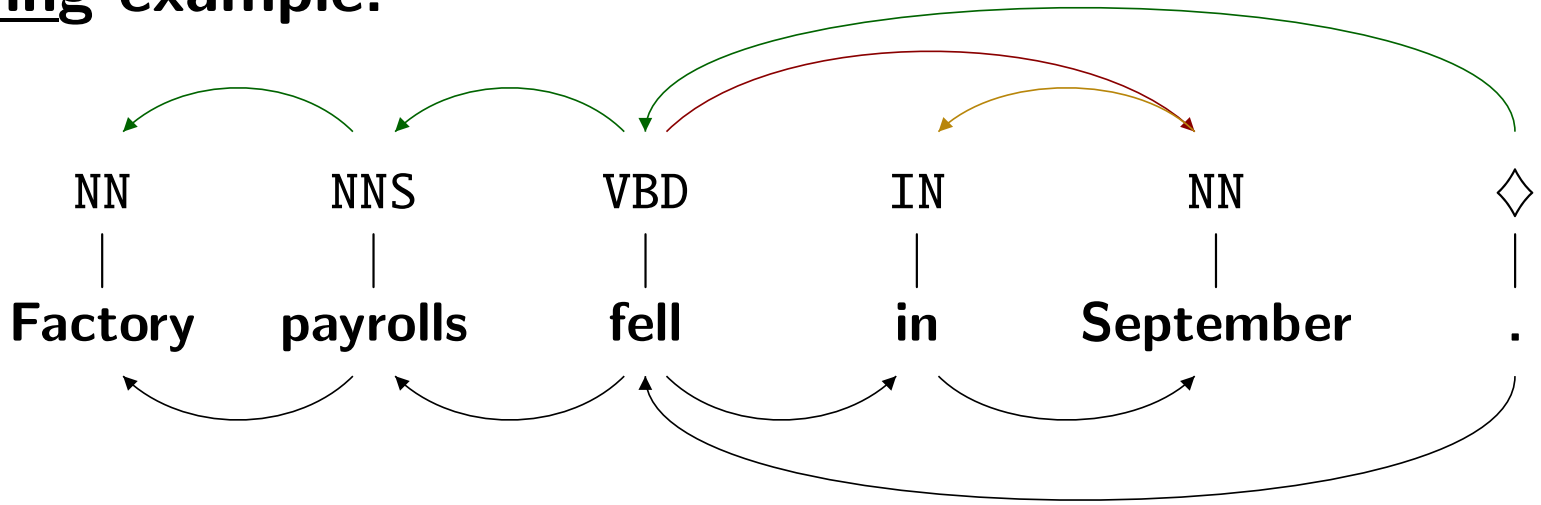
Motivation: Unsupervised (Dependency) Parsing

- **Insert your favorite reason(s) why you'd like to parse anything in the first place...**
- **... adjust for any data without reference tree banks: — i.e., understudied languages or genres (e.g., legal).**
- **Potential applications:**
  - **machine translation**
    — word alignment, phrase extraction, reordering;
  - **web search**
    — retrieval, query refinement;
  - **question answering, speech recognition, etc.**

Evaluation: Directed Dependency Accuracy

Evaluation: Directed Dependency Accuracy

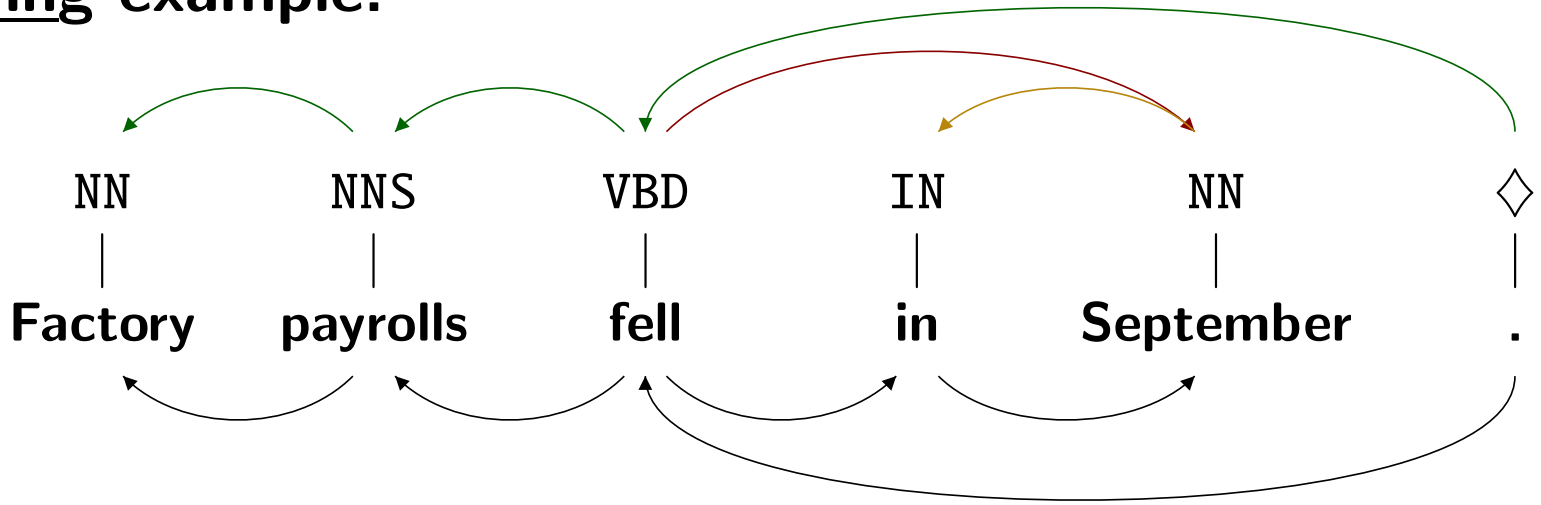
Scoring example:



Directed Score: $\frac{3}{5} = 60\%$ **(left/right-branching baselines:** $\frac{2}{5} = 40\%$**)**

Evaluation: Directed Dependency Accuracy

Scoring example:



Directed Score: $\frac{3}{5} = 60\%$ **(left/right-branching baselines:** $\frac{2}{5} = 40\%$**);**

Undirected Score: $\frac{4}{5} = 80\%$ **(left/right-branching baselines:** $\frac{4}{5} = 80\%$**).**

Prior Art: A Brief History

- **1992 — word classes (Carroll and Charniak)**
- **1998 — greedy linkage via mutual information (Yuret)**
- **2001 — iterative re-estimation with EM (Paskin)**

Prior Art: A Brief History

- 1992 — word classes (Carroll and Charniak)
- 1998 — greedy linkage via mutual information (Yuret)
- 2001 — iterative re-estimation with EM (Paskin)

- 2004 — right-branching baseline
  — valence (DMV) (Klein and Manning)

Prior Art: A Brief History

- 1992 — word classes (Carroll and Charniak)
- 1998 — greedy linkage via mutual information (Yuret)
- 2001 — iterative re-estimation with EM (Paskin)

- 2004 — right-branching baseline
  — valence (DMV) (Klein and Manning)

Prior Art: A Brief History

- 1992 — word classes (Carroll and Charniak)
- 1998 — greedy linkage via mutual information (Yuret)
- 2001 — iterative re-estimation with EM (Paskin)

- 2004 — right-branching baseline
  — valence (DMV) (Klein and Manning)

- 2004 — annealing techniques (Smith and Eisner)
- 2005 — contrastive estimation (Smith and Eisner)
- 2006 — structural biasing (Smith and Eisner)
- 2007 — common cover link representation (Seginer)
- 2008 — logistic normal priors (Cohen et al.)
- 2009 — lexicalization and smoothing (Headden et al.)
- 2009 — soft parameter tying (Cohen and Smith)

Prior Art: Dependency Model with Valence

Prior Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**

Prior Art: Dependency Model with Valence

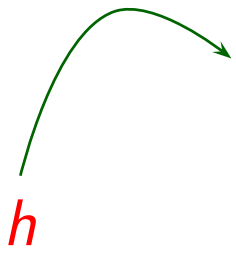
- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**

h

e.g.: verb

Prior Art: Dependency Model with Valence

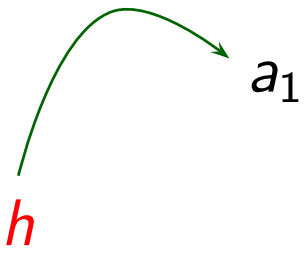
- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



e.g.: verb

Prior Art: Dependency Model with Valence

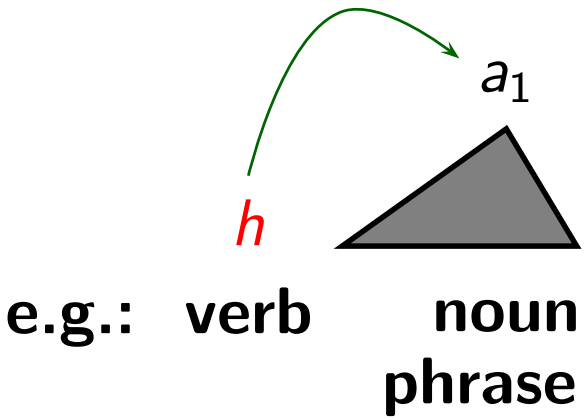
- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



e.g.: verb noun

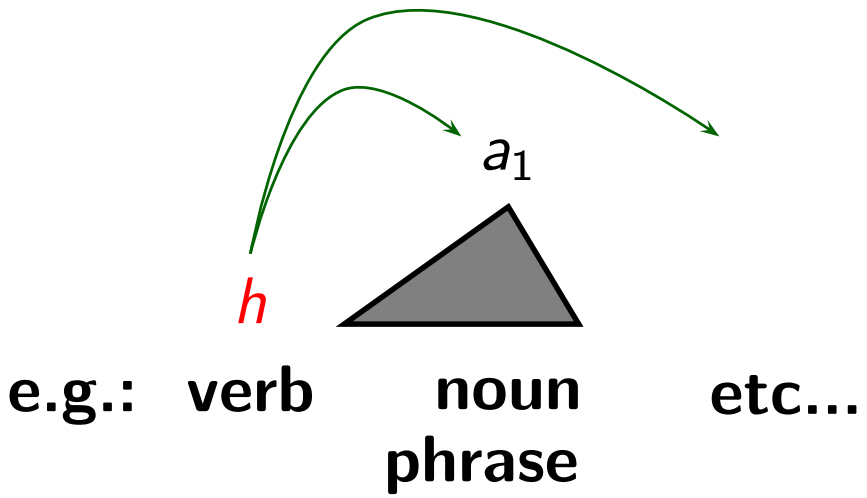
Prior Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



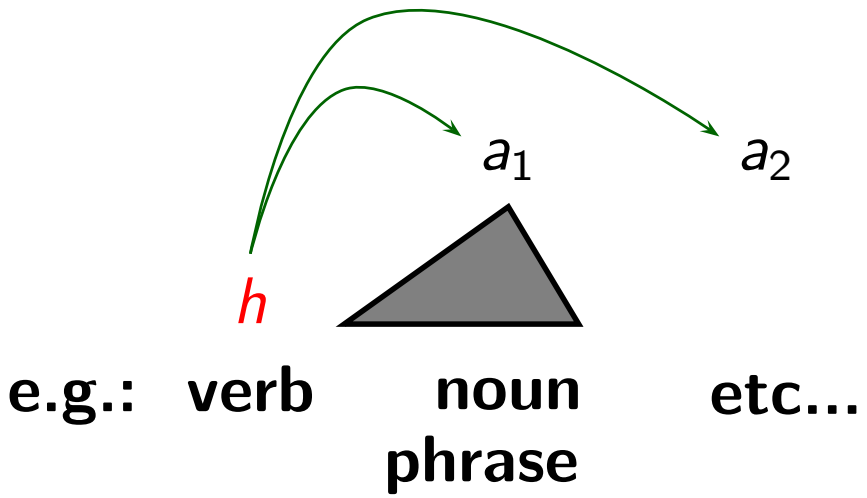
Prior Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



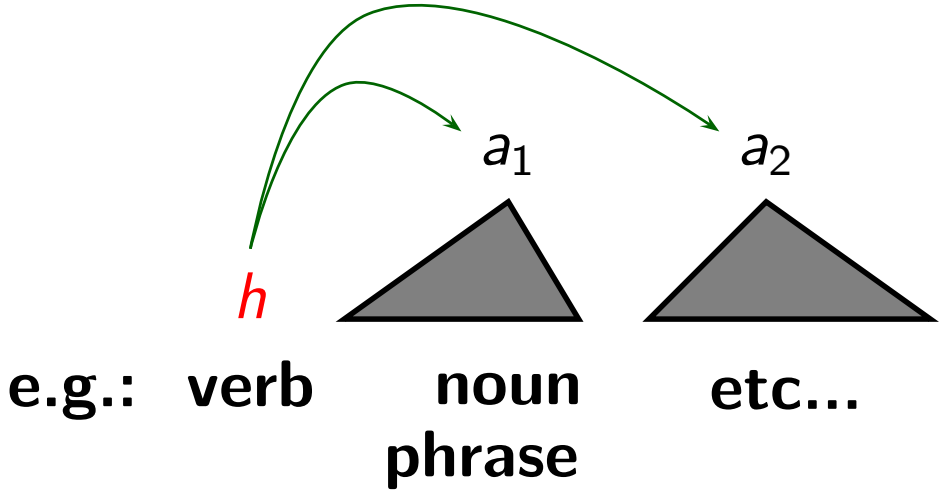
Prior Art: Dependency Model with Valence

- **a head-outward model, with word classes and valence/adjacency** **(Klein and Manning, 2004)**



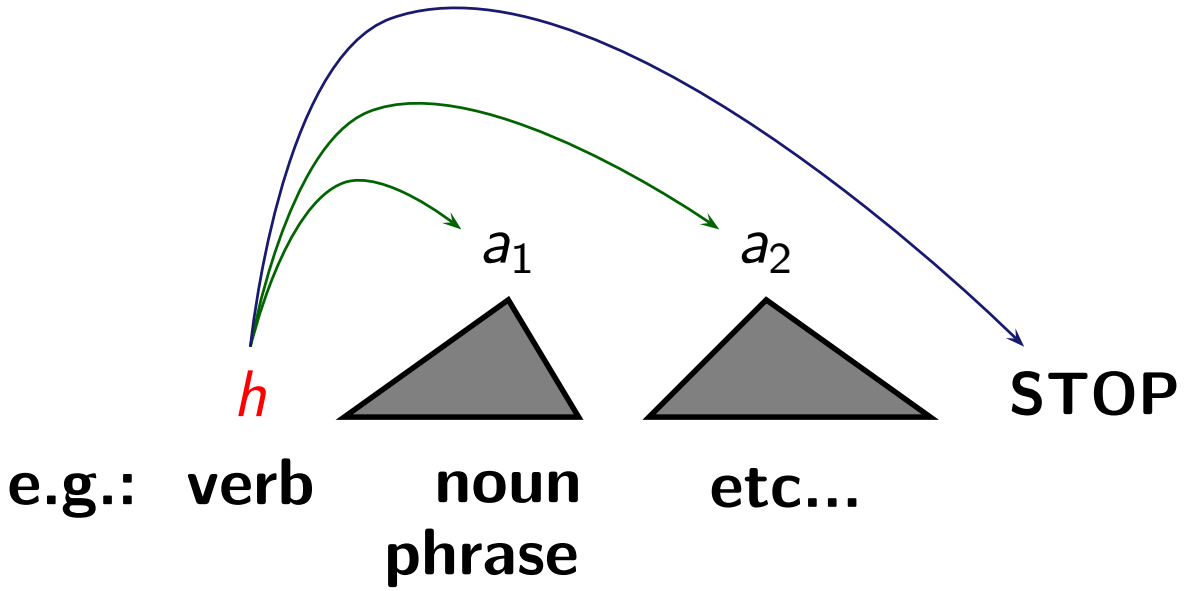
Prior Art: Dependency Model with Valence

- a **head-outward** model, with **word classes** and **valence/adjacency** (Klein and Manning, 2004)



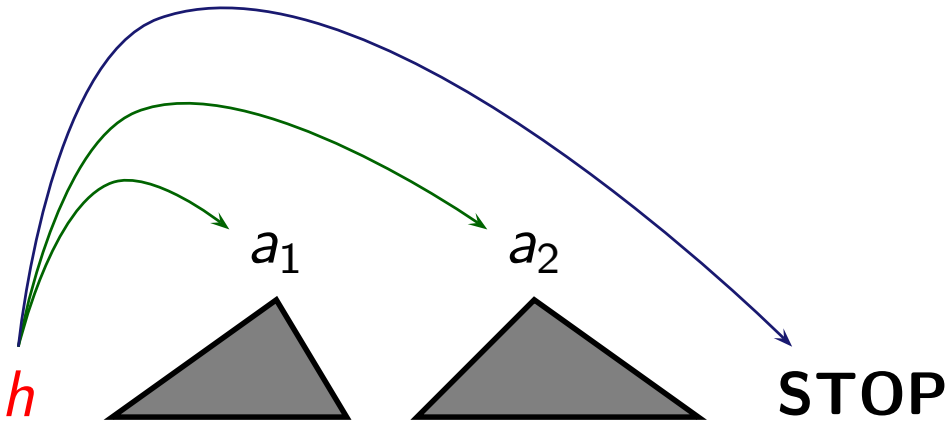
Prior Art: Dependency Model with Valence

- a **head-outward** model, with **word classes** and **valence/adjacency** (Klein and Manning, 2004)



Prior Art: Dependency Model with Valence

- a **head-outward** model, with **word classes** and **valence/adjacency** (Klein and Manning, 2004)



$$\mathbb{P}(t_h) = \prod_{dir \in \{\mathrm{L},\mathrm{R}\}} \left[\underline{\mathbb{P}_{\mathrm{STOP}}(c_h, dir, \overbrace{1_{n=0}}^{adj})} \prod_{i=1}^{n} \underline{\mathbb{P}(t_{a_i})}\, \mathbb{P}_{\mathrm{ATTACH}}(c_h, dir, c_{a_i}) \; \underline{(1 - \mathbb{P}_{\mathrm{STOP}}(c_h, dir, \overbrace{1_{i=1}}^{adj}))} \right]_{n=|args(h,dir)|}$$

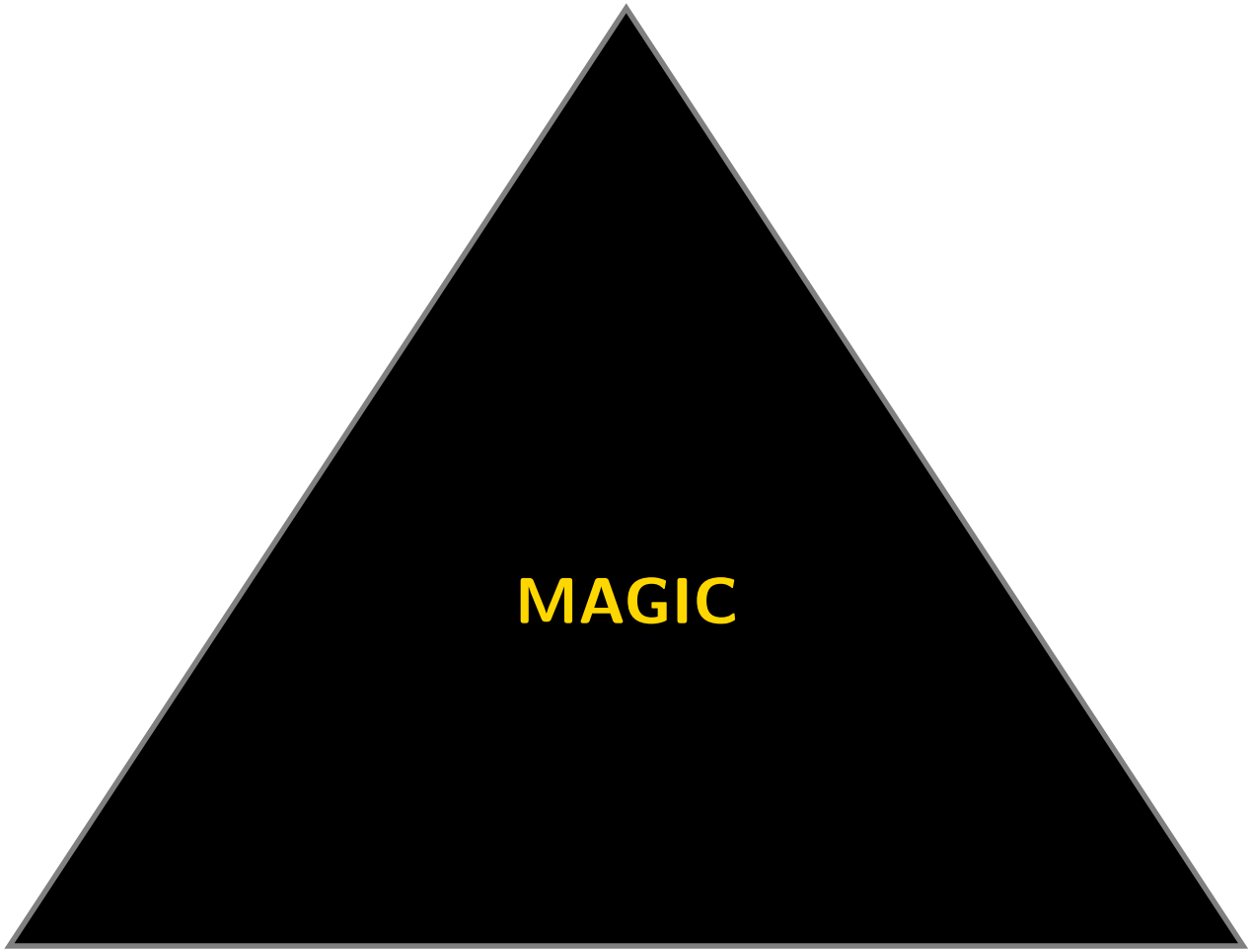
Prior Art: Unsupervised Learning Engine

Prior Art: Unsupervised Learning Engine

- **parameter fitting via expectation-maximization (EM)**

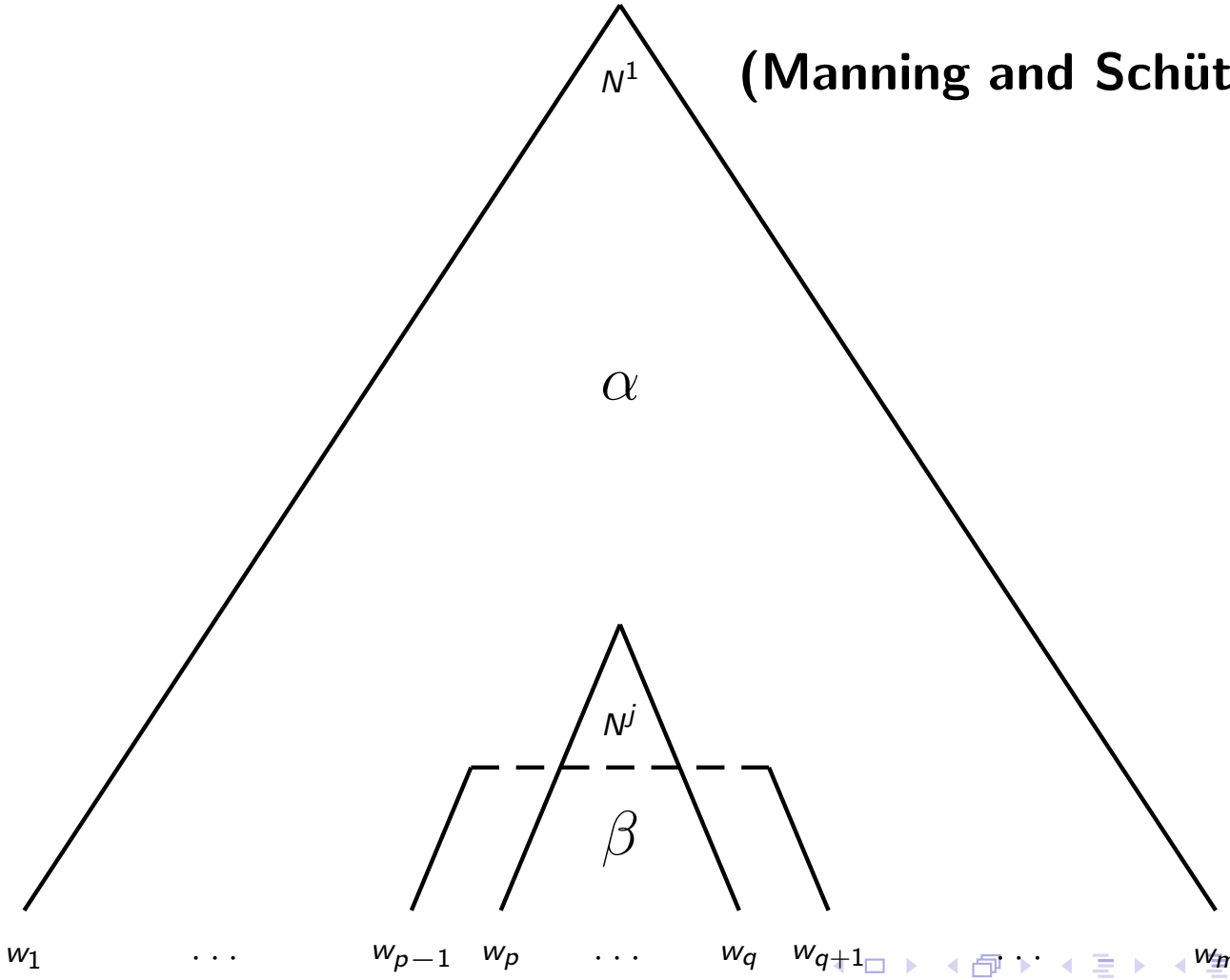
Prior Art: Unsupervised Learning Engine

- **parameter fitting via expectation-maximization (EM)**



Prior Art: Unsupervised Learning Engine

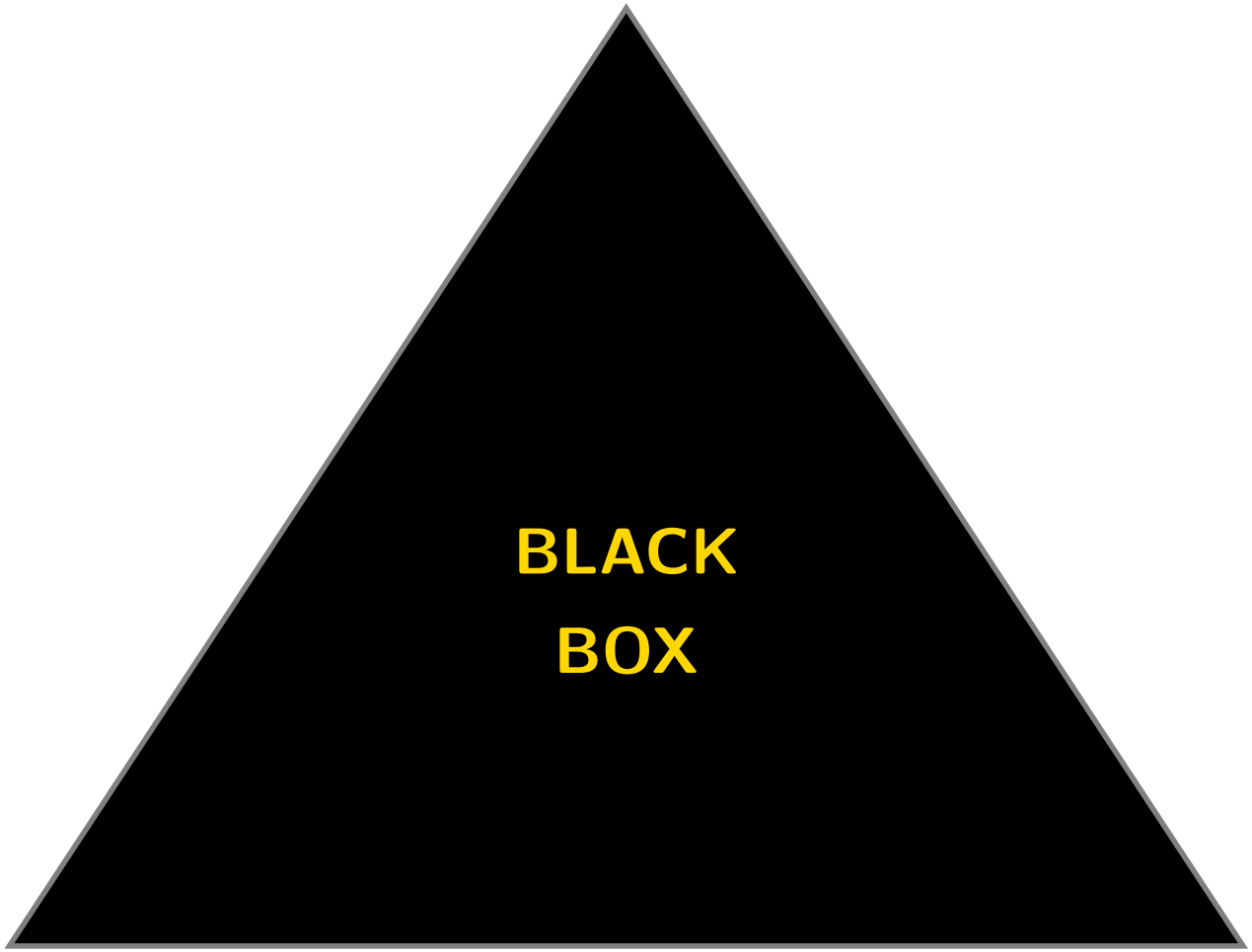
- **parameter fitting via expectation-maximization (EM)**
  - **by means of inside-outside re-estimation (Baker, 1979)**



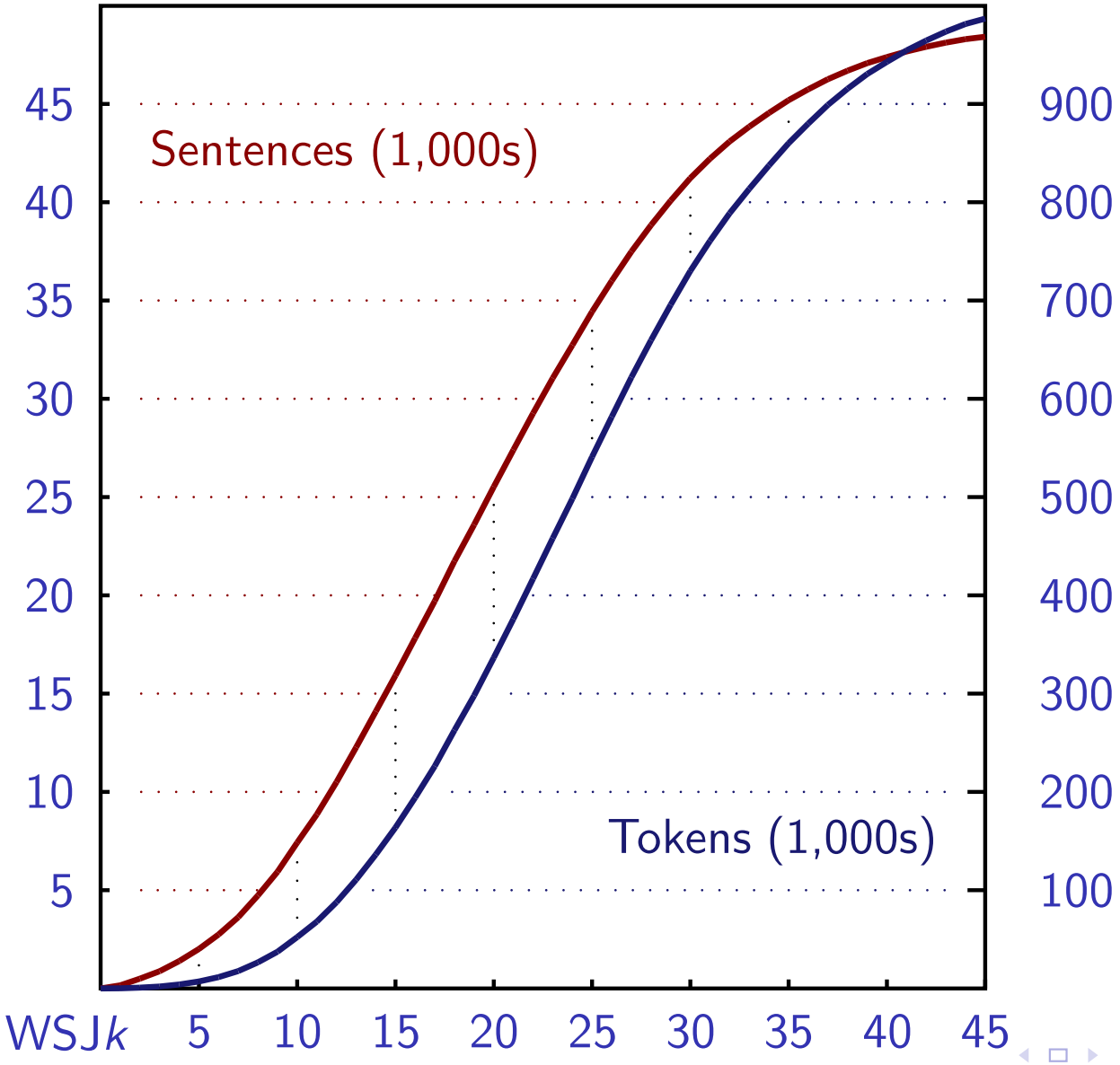
(Manning and Schütze, 1999)

Prior Art: Unsupervised Learning Engine

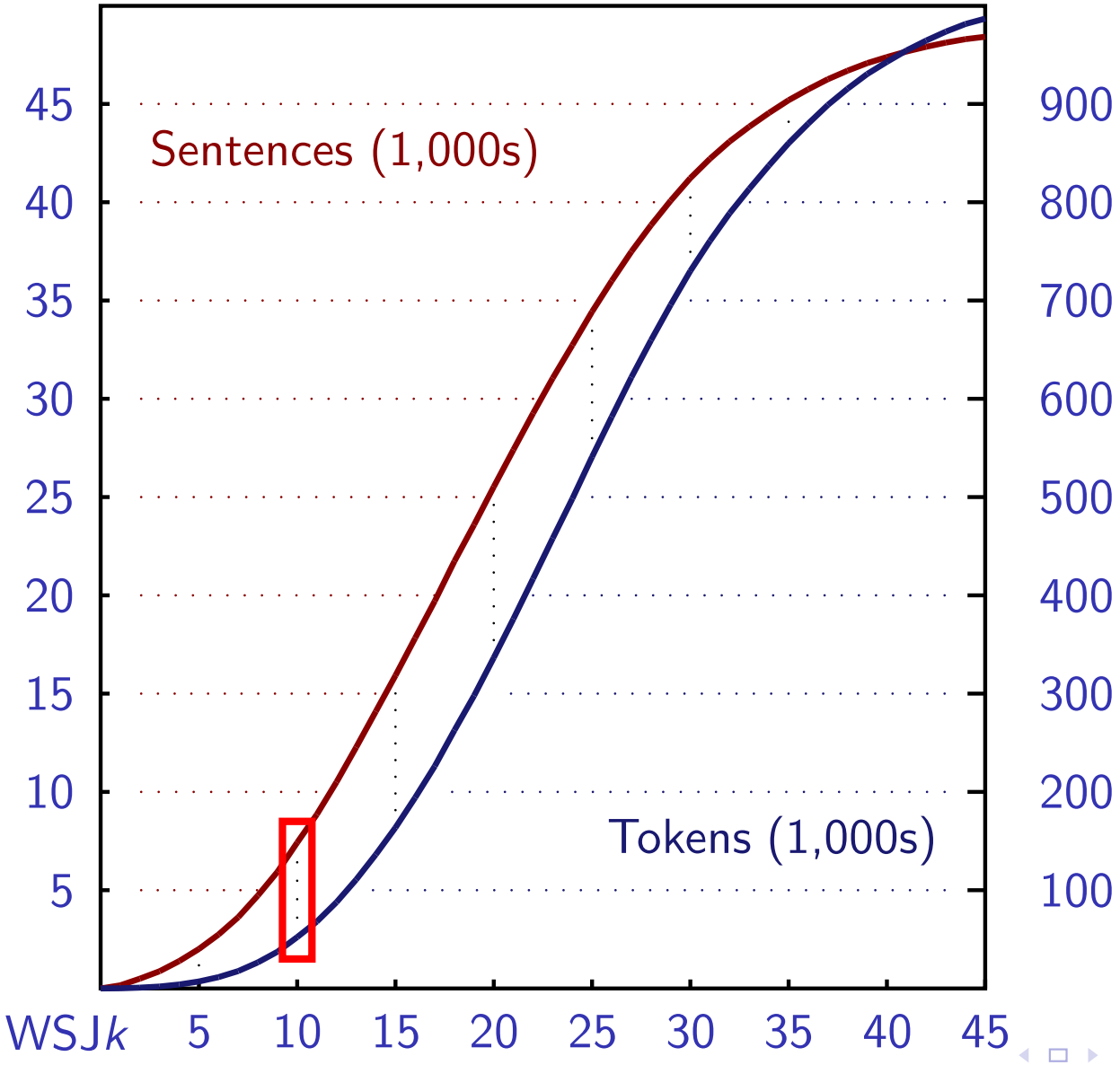
- **parameter fitting via expectation-maximization (EM)**



Prior Art: The Standard Corpus — WSJ



Prior Art: The Standard Corpus — WSJ10



Outline

- **Introduction to the Grammar Induction Problem**
  - **The Inputs and Outputs** (text $\rightarrow$ dependency parses)
  - **A Probabilistic Model** (DMV)
  - **Parameter Fitting** (EM)

Outline

- **Introduction to the Grammar Induction Problem**
  - **The Inputs and Outputs** (text $\rightarrow$ dependency parses)
  - **A Probabilistic Model** (DMV)
  - **Parameter Fitting** (EM)
- **Part I: Non-Convex Optimization**

Outline

- **Introduction to the Grammar Induction Problem**
  - **The Inputs and Outputs** (text $\rightarrow$ dependency parses)
  - **A Probabilistic Model** (DMV)
  - **Parameter Fitting** (EM)
- **Part I: Non-Convex Optimization**
- **Part II: Parsing Constraints**

Outline

- **Introduction to the Grammar Induction Problem**
  - **The Inputs and Outputs** (text $\rightarrow$ dependency parses)
  - **A Probabilistic Model** (DMV)
  - **Parameter Fitting** (EM)
- **Part I: Non-Convex Optimization**
- **Part II: Parsing Constraints**
- **Part III: Generative Models**

Outline

- **Introduction to the Grammar Induction Problem**
  - **The Inputs and Outputs** (text $\rightarrow$ dependency parses)
  - **A Probabilistic Model** (DMV)
  - **Parameter Fitting** (EM)
- **Part I: Non-Convex Optimization**
- **Part II: Parsing Constraints**
- **Part III: Generative Models**
- **State-of-the-Art Integrated Solution**

Outline

- **Introduction to the Grammar Induction Problem**
  - **The Inputs and Outputs** (text $\rightarrow$ dependency parses)
  - **A Probabilistic Model** (DMV)
  - **Parameter Fitting** (EM)
- **Part I: Non-Convex Optimization**
- **Part II: Parsing Constraints**
- **Part III: Generative Models**
- **State-of-the-Art Integrated Solution** (in submission)

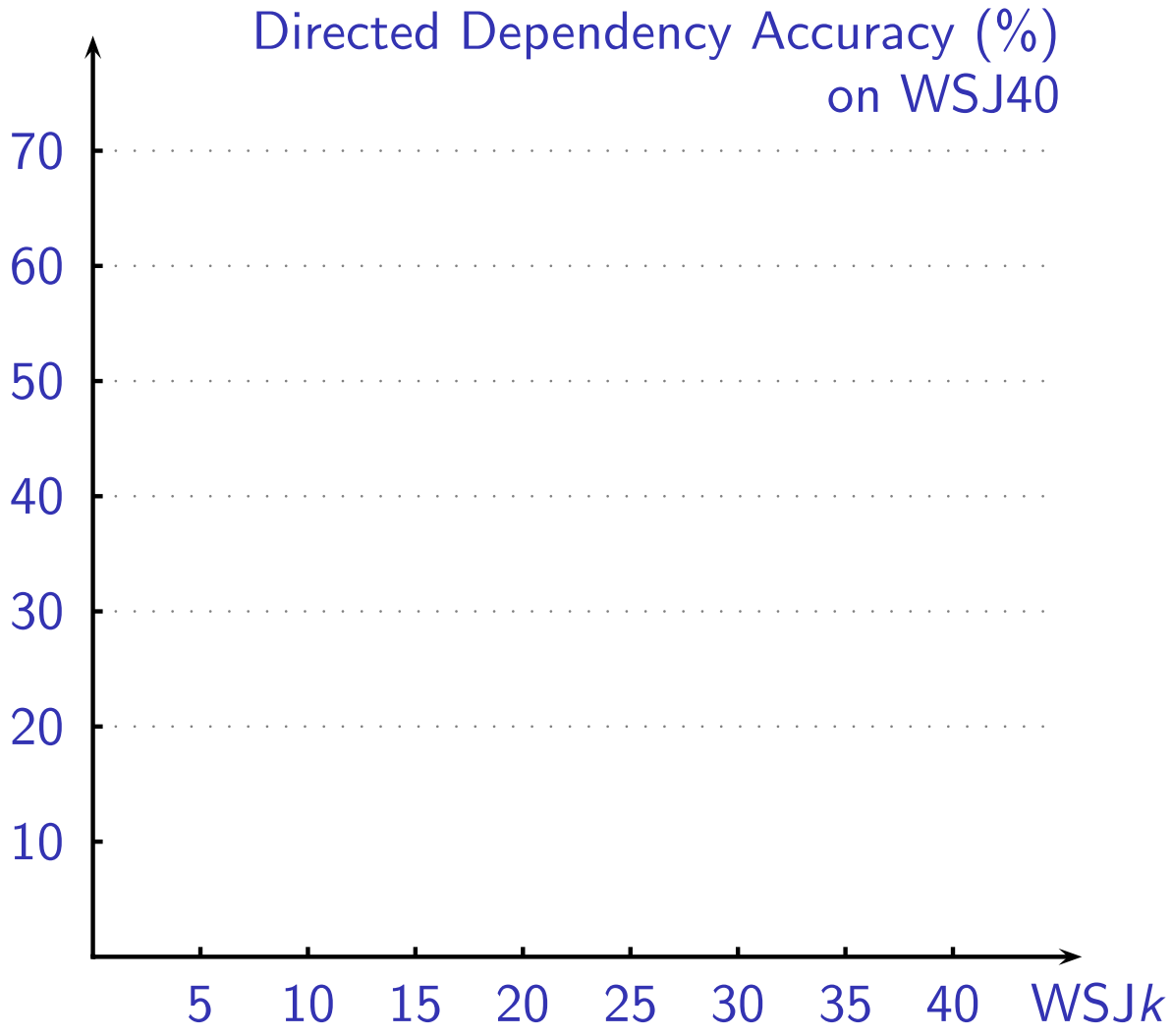
Optimization

Viterbi Training

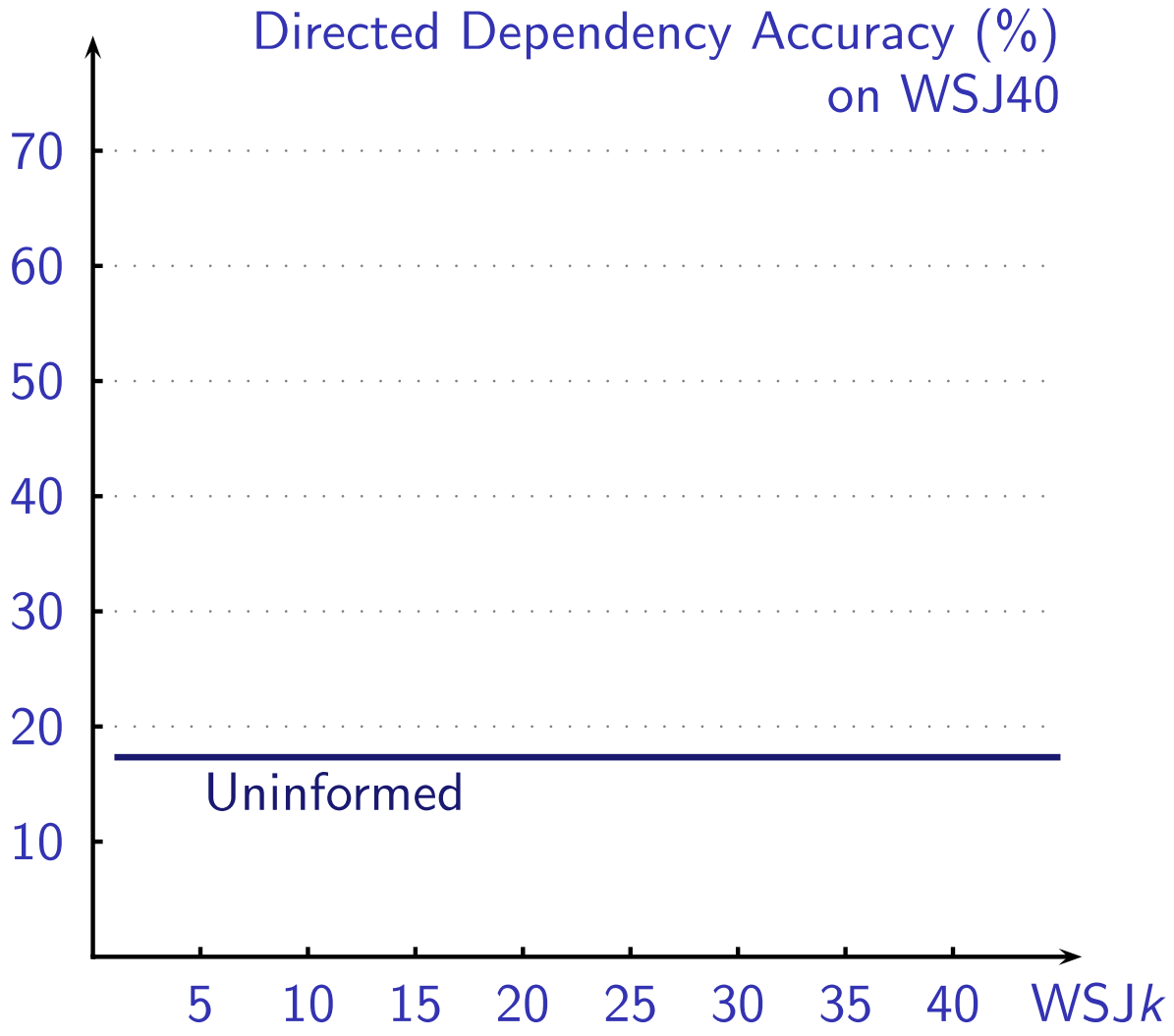
Baby Steps

Lateen EM

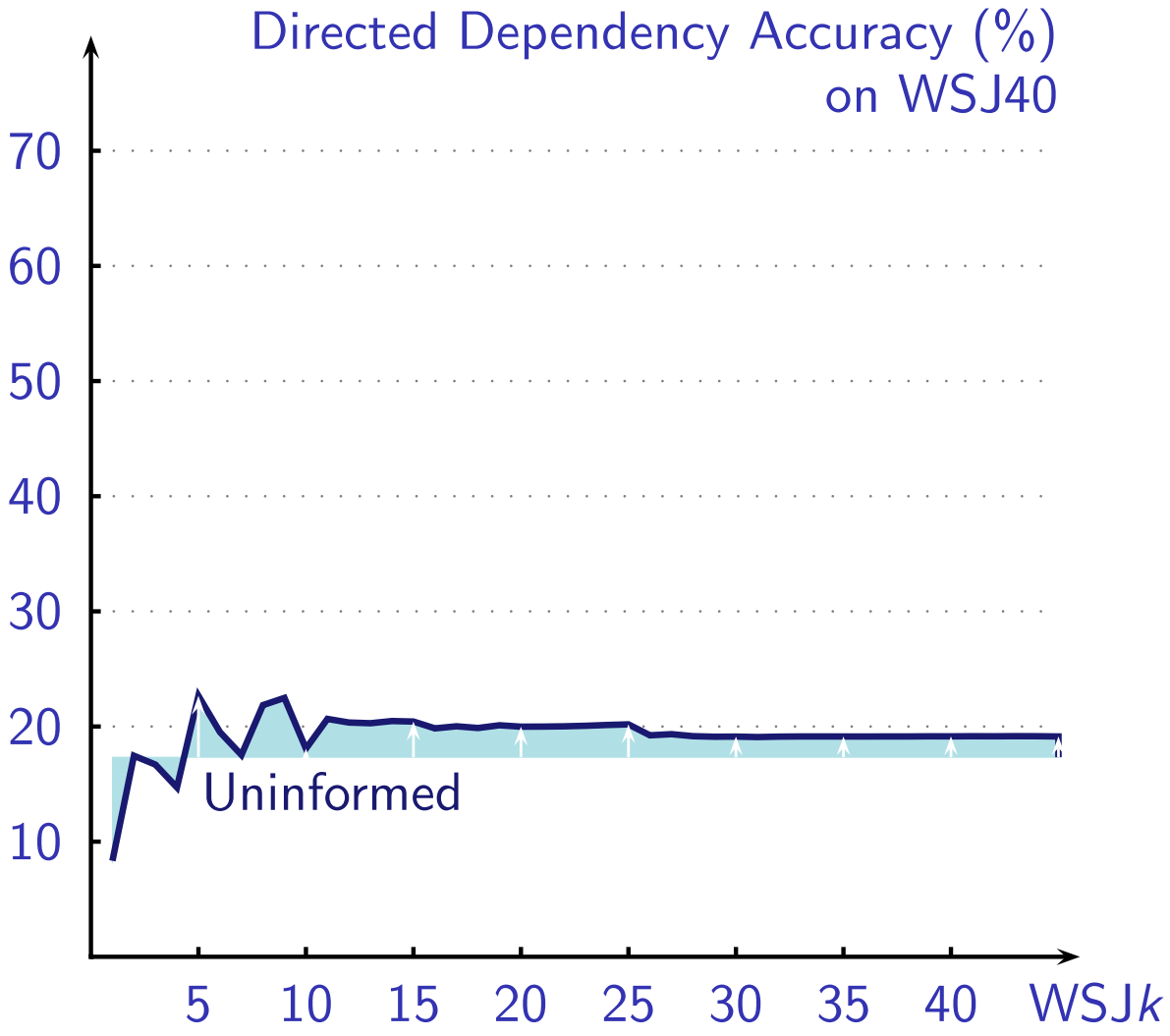
Issue I: Why so little data?



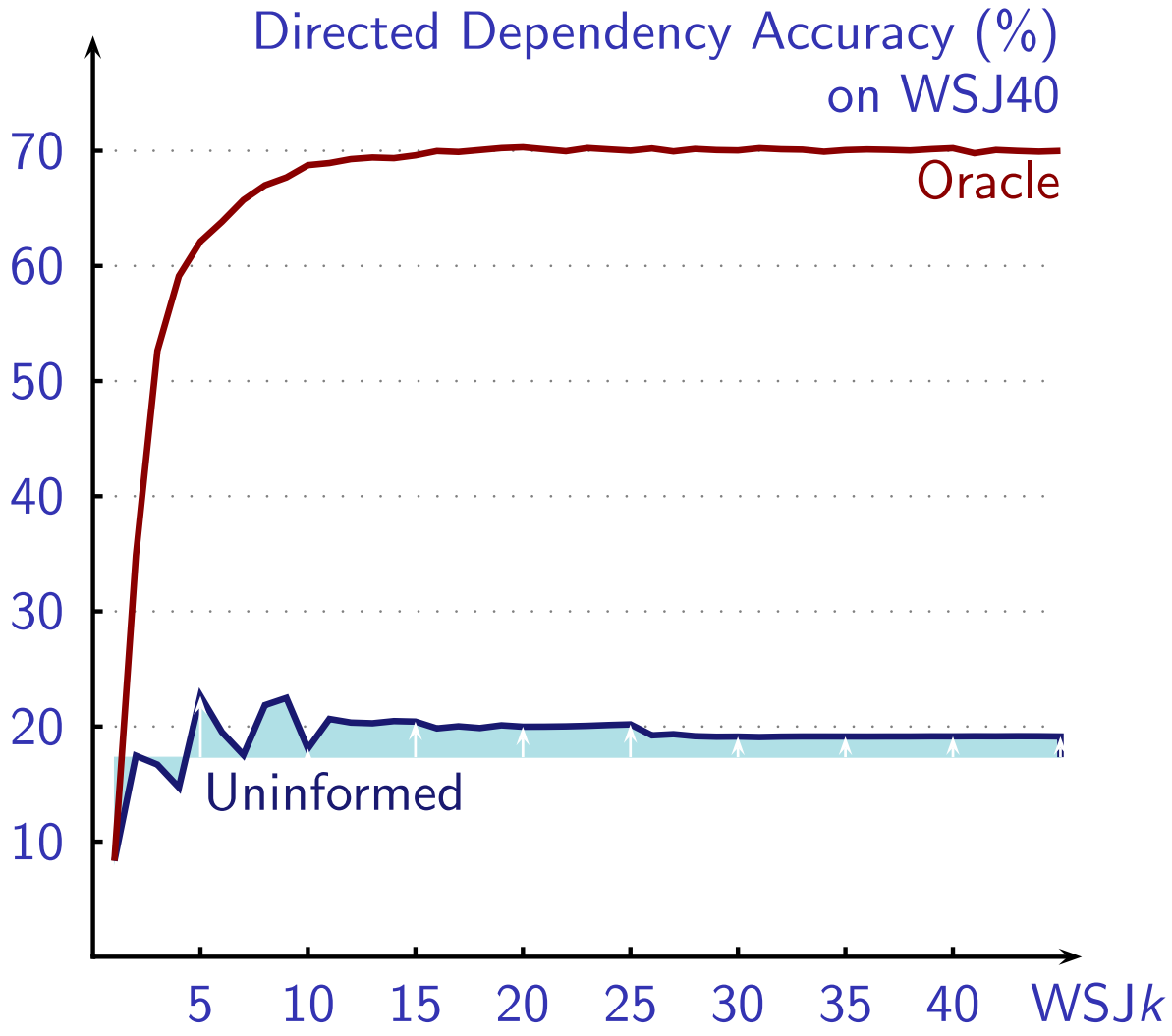
Issue I: Why so little data?



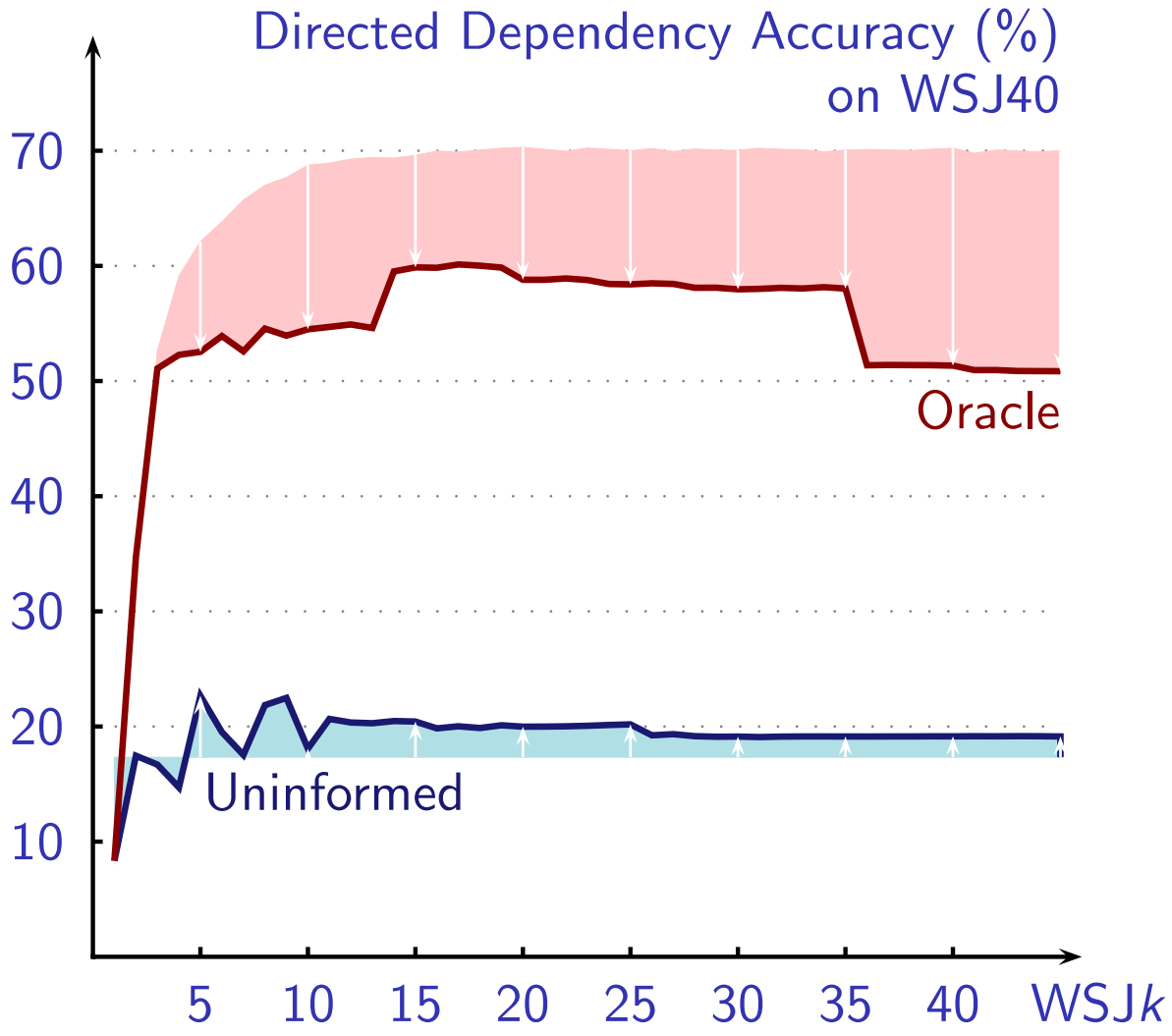
Issue I: Why so little data?



Issue I: Why so little data?



Issue I: Why so little data?



Issue II: Non-convex objectives...

- **maximizing the probability of data (sentence strings):**

$$\hat{\theta}_{\text{UNS}} = \arg\max_{\theta} \sum_{s} \log \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

Issue II: Non-convex objectives...

- **maximizing the probability of data (sentence strings):**

$$\hat{\theta}_{\textsf{UNS}} = \arg\max_{\theta} \sum_{s} \log \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

- **supervised objective would be convex:**

$$\hat{\theta}_{\textsf{SUP}} = \arg\max_{\theta} \sum_{s} \log \mathbb{P}_{\theta}(t^{*}(s))$$

Issue II: Non-convex objectives...

- **maximizing the probability of data (sentence strings):**

$$\hat{\theta}_{\text{UNS}} = \arg\max_{\theta} \sum_{s} \log \underbrace{\sum_{t \in T(s)} \mathbb{P}_{\theta}(t)}_{\mathbb{P}_{\theta}(s)}$$

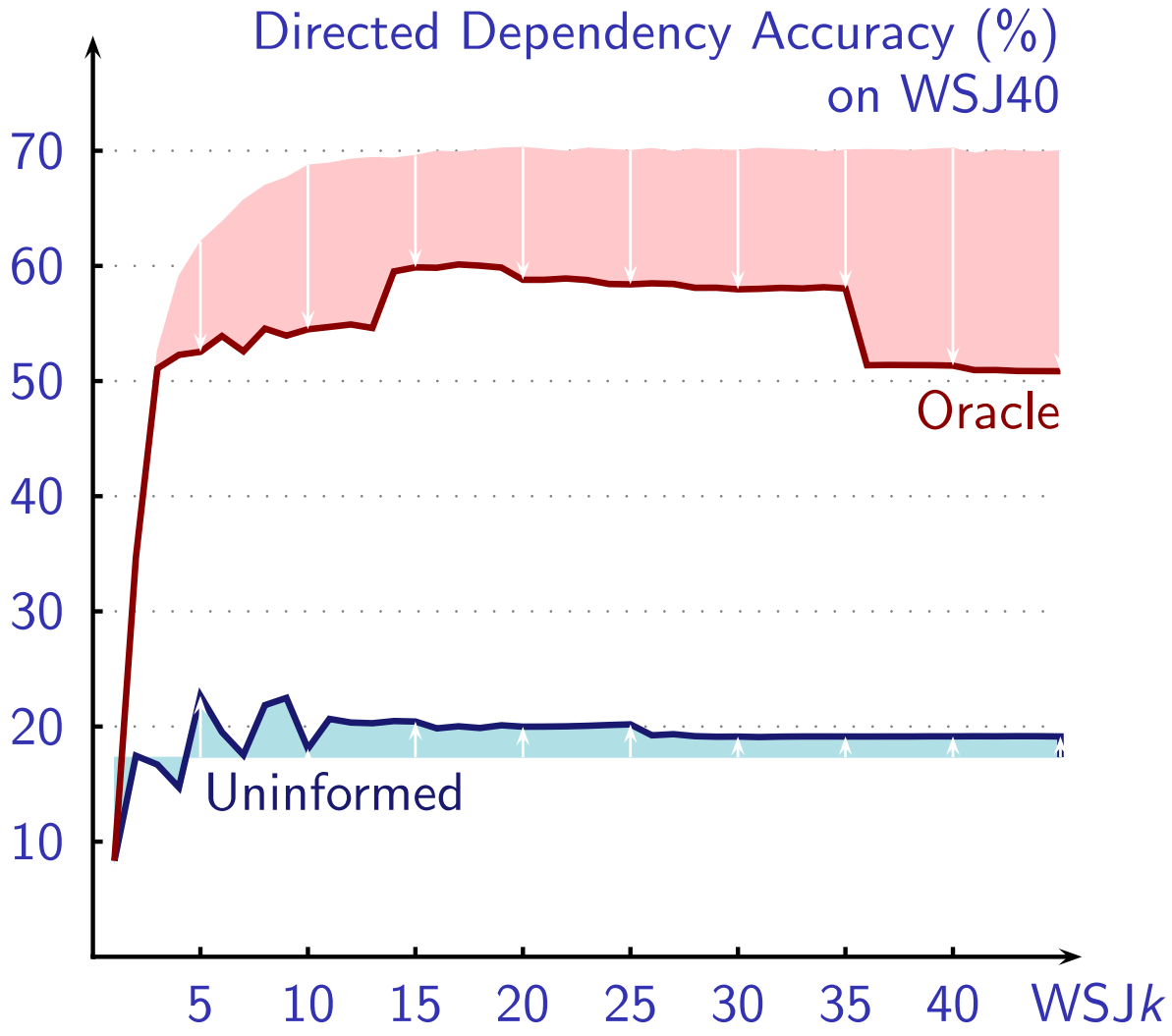
- **supervised objective would be convex:**

$$\hat{\theta}_{\text{SUP}} = \arg\max_{\theta} \sum_{s} \log \mathbb{P}_{\theta}(t^{*}(s))$$

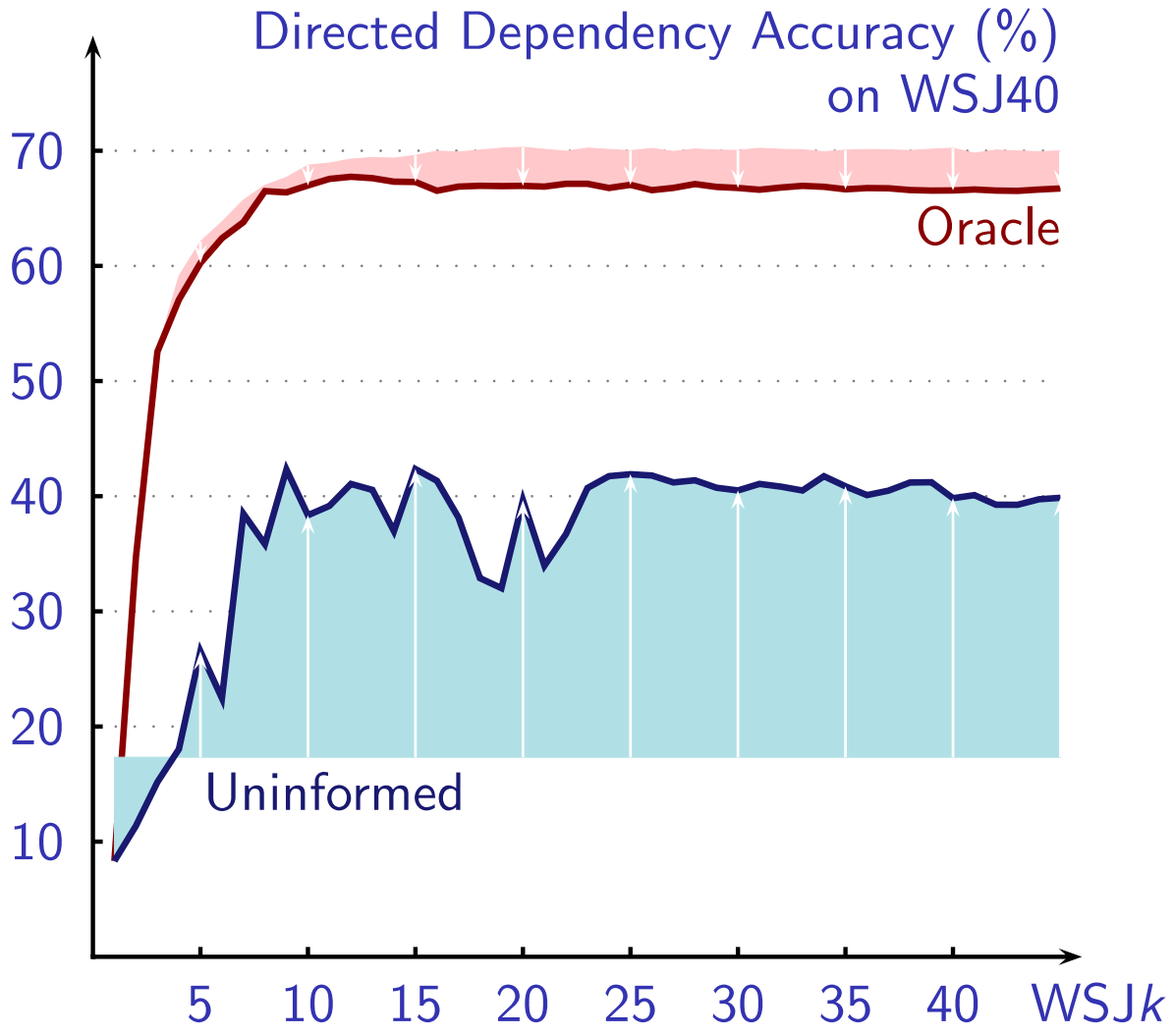
- **could optimize likeliest parse trees (also non-convex):**

$$\hat{\theta}_{\text{VIT}} = \arg\max_{\theta} \sum_{s} \max_{t \in T(s)} \log \mathbb{P}_{\theta}(t)$$

Classic EM:


Directed Dependency Accuracy (%)
on WSJ40
70
60
50
40
30
20
10
Oracle
Uninformed
5
10
15
20
25
30
35
40
WSJk

Viterbi EM:


Directed Dependency Accuracy (%)
on WSJ40
70
60
50
40
30
20
10
Oracle
Uninformed
5
10
15
20
25
30
35
40
WSJk

Hard vs. Soft EM:

- **Viterbi EM: zoom in on likeliest tree**

Hard vs. Soft EM:

- **Viterbi EM: zoom in on likeliest tree**
  - **does not degrade with input length**
  - **better retains supervised solutions**
  - **actually learns from scratch**

    — see also Cohen and Smith (2010)

Hard vs. Soft EM:

- **Viterbi EM: zoom in on likeliest tree**
  - **does not degrade with input length**
  - **better retains supervised solutions**
  - **actually learns from scratch**

    — see also Cohen and Smith (2010)

- **Classic EM: "focus across the board"**

Hard vs. Soft EM:

- **Viterbi EM: zoom in on likeliest tree**
  - **does not degrade with input length**
  - **better retains supervised solutions**
  - **actually learns from scratch**

    — see also Cohen and Smith (2010)

- **Classic EM: "focus across the board"**
  - **hard to see the trees for the forest...**

    ... but known to work with shorter inputs!

Idea I: Baby Steps ... as Non-convex Optimization

Idea I: Baby Steps ... as Non-convex Optimization

- **start with an easy (convex) case**

Idea I: Baby Steps ... as Non-convex Optimization

- **start with an easy (convex) case**
- **slowly extend it to the fully complex target task**

Idea I: Baby Steps ... as Non-convex Optimization

- **start with an easy (convex) case**
- **slowly extend it to the fully complex target task**
- **take tiny (cautious) steps in the problem space**

Idea I: Baby Steps ... as Non-convex Optimization

- **start with an easy (convex) case**
- **slowly extend it to the fully complex target task**
- **take tiny (cautious) steps in the problem space**
- **... try not to stray far from relevant neighborhoods in the solution space**

Idea I: Baby Steps ... as Non-convex Optimization

- **start with an easy (convex) case**
- **slowly extend it to the fully complex target task**
- **take tiny (cautious) steps in the problem space**
- **... try not to stray far from relevant neighborhoods in the solution space**

- **base case: sentences of length one (trivial — no init)**

Idea I: Baby Steps ... as Non-convex Optimization

- **start with an easy (convex) case**
- **slowly extend it to the fully complex target task**
- **take tiny (cautious) steps in the problem space**
- **... try not to stray far from relevant neighborhoods in the solution space**

- **base case: sentences of length one (trivial — no init)**
- **incremental step: smooth WSJk; re-init WSJ$(k+1)$**

Idea I: Baby Steps ... as Graduated Learning

Idea I: Baby Steps ... as Graduated Learning

- **WSJ1** — **Atone** (**verbs**!)

Idea I: Baby Steps ... as Graduated Learning

- **WSJ1** — **Atone** (**verbs!**)

- **WSJ2** — **It is.** (**pronouns!**)
  Darkness fell.

Idea I: Baby Steps ... as Graduated Learning

- **WSJ1** — **Atone** (**verbs!**)

- **WSJ2** — **It is.** (**pronouns!**)
  Darkness fell.

- **WSJ3** — **Become a Lobbyist**
  But many have. (**determiners!**)
  They didn't.

Idea I: Baby Steps ... and Related Notions

Idea I: Baby Steps ... and Related Notions

- **shaping** **(Skinner, 1938)**

Idea I: Baby Steps ... and Related Notions

- **shaping** Psychology / Cognitive Science (Skinner, 1938)
- **less is more** (Kail, 1984; Newport, 1988; 1990)
- **starting small** (Elman, 1993)
  - **scaffold on model complexity** [restrict memory]
  - **scaffold on data complexity** [restrict input]

Idea I: Baby Steps ... and Related Notions

- **shaping** Psychology / Cognitive Science (Skinner, 1938)
- **less is more** (Kail, 1984; Newport, 1988; 1990)
- **starting small** (Elman, 1993)
  - **scaffold on model complexity** [restrict memory]
  - **scaffold on data complexity** [restrict input]
- **stepping stones** (Brown et al., 1993)
- **coarse-to-fine** NLP / AI (Charniak and Johnson, 2005)
- **curriculum learning** (Bengio et al., 2009)

Idea I: Baby Steps ... and Related Notions

- **shaping** Psychology / Cognitive Science (Skinner, 1938)
- **less is more** (Kail, 1984; Newport, 1988; 1990)
- **starting small** (Elman, 1993)
  - **scaffold on model complexity** [restrict memory]
  - **scaffold on data complexity** [restrict input]
- **stepping stones** (Brown et al., 1993)
- **coarse-to-fine** NLP / AI (Charniak and Johnson, 2005)
- **curriculum learning** (Bengio et al., 2009)

- **continuation methods** (Allgower and Georg, 1990)
- **deterministic annealing** Math / OR (Rose, 1998)

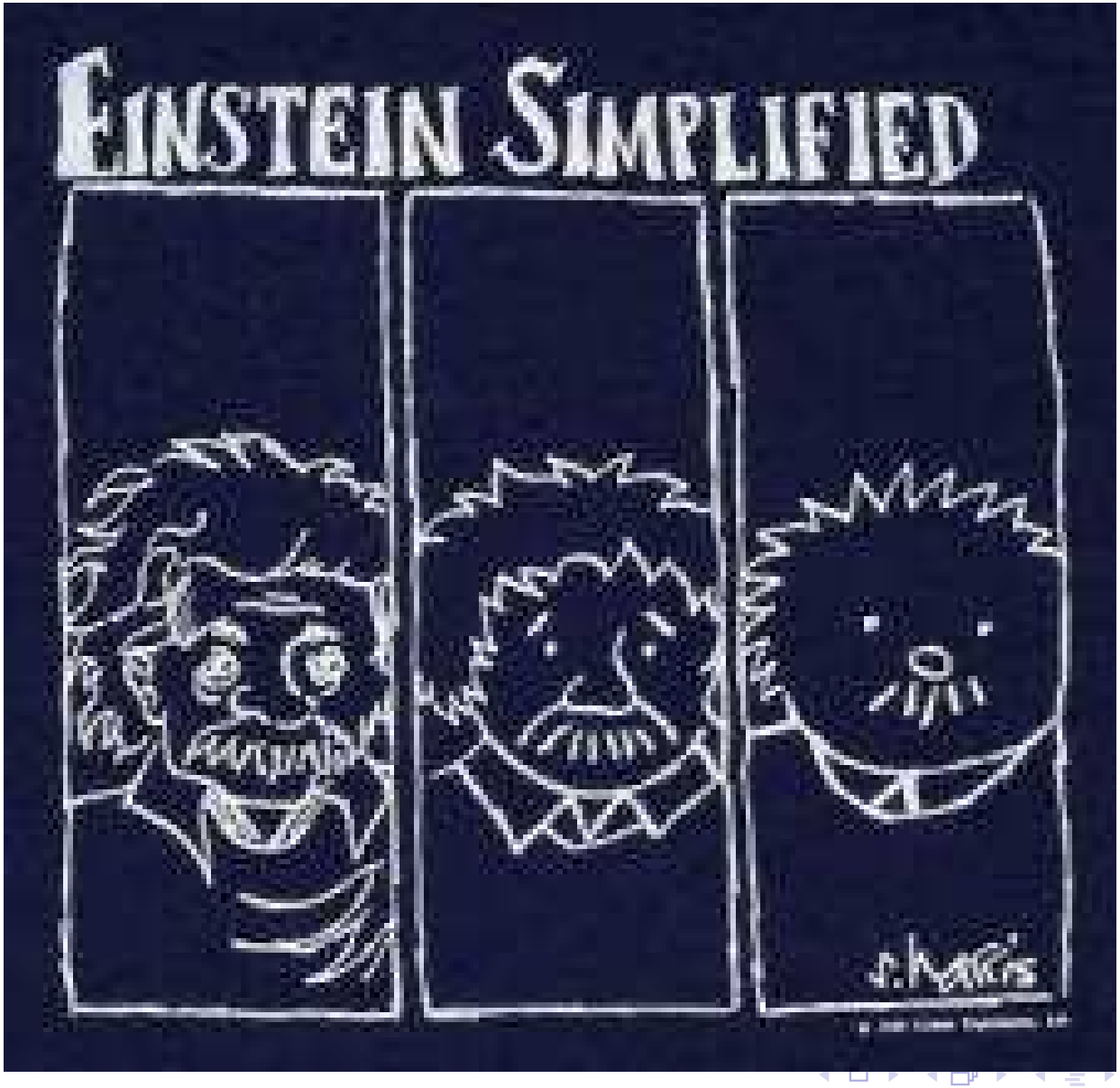
Idea I: Baby Steps ... and Related Notions

- **shaping** Psychology / Cognitive Science (Skinner, 1938)
- **less is more** (Kail, 1984; Newport, 1988; 1990)
- **starting small** (Elman, 1993)
  - **scaffold on model complexity** [restrict memory]
  - **scaffold on data complexity** [restrict input]
- **stepping stones** (Brown et al., 1993)
- **coarse-to-fine** NLP / AI (Charniak and Johnson, 2005)
- **curriculum learning** (Bengio et al., 2009)

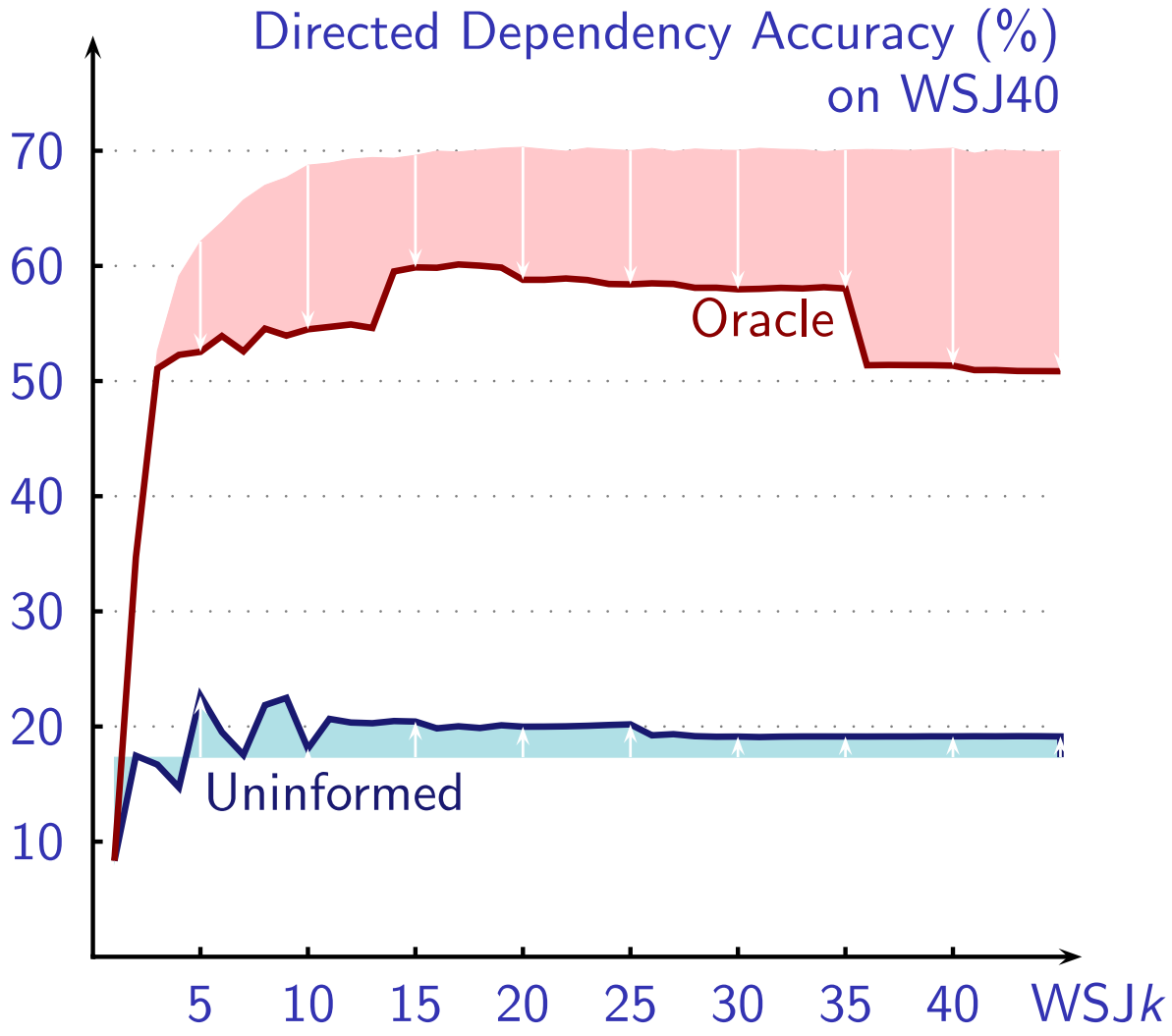
- **continuation methods** (Allgower and Georg, 1990)
- **deterministic annealing** Math / OR (Rose, 1998)

successive approximations!

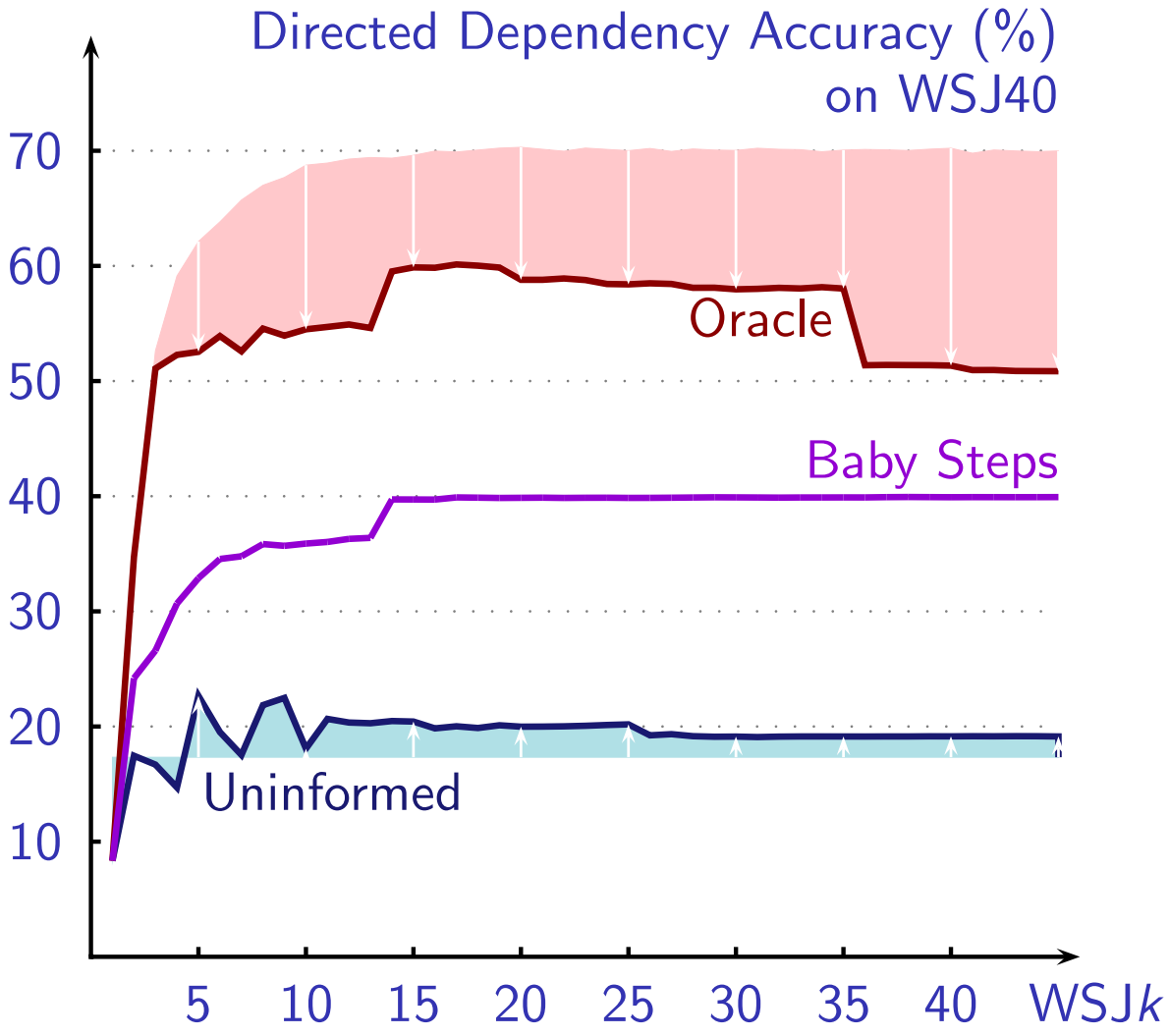
Aside I: Speaking of successive approximations...



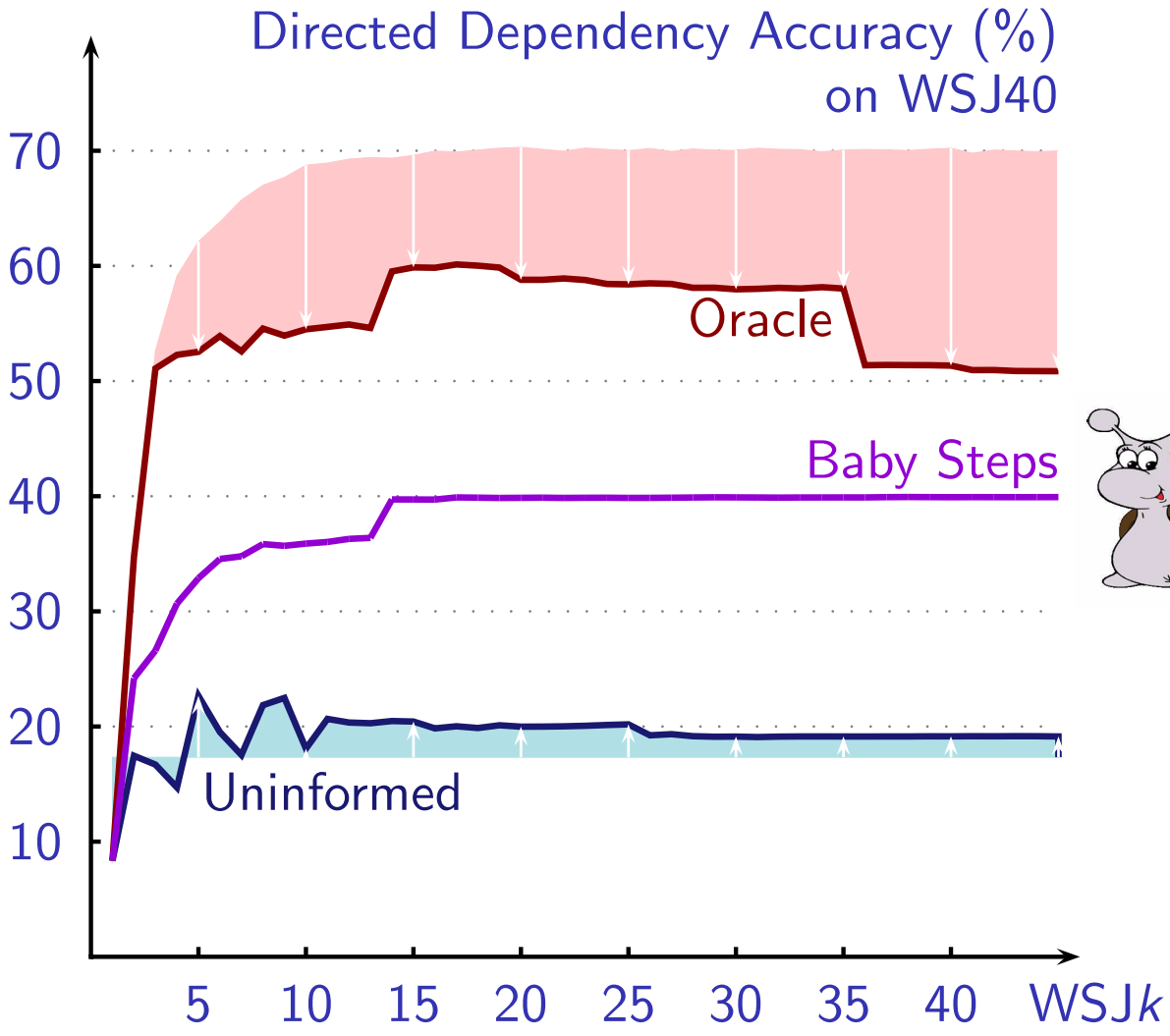
Idea I: Baby Steps ... Observations & Results!

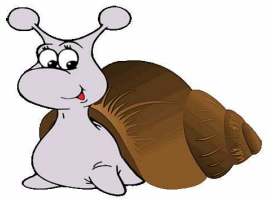


Idea I: Baby Steps ... Observations & Results!

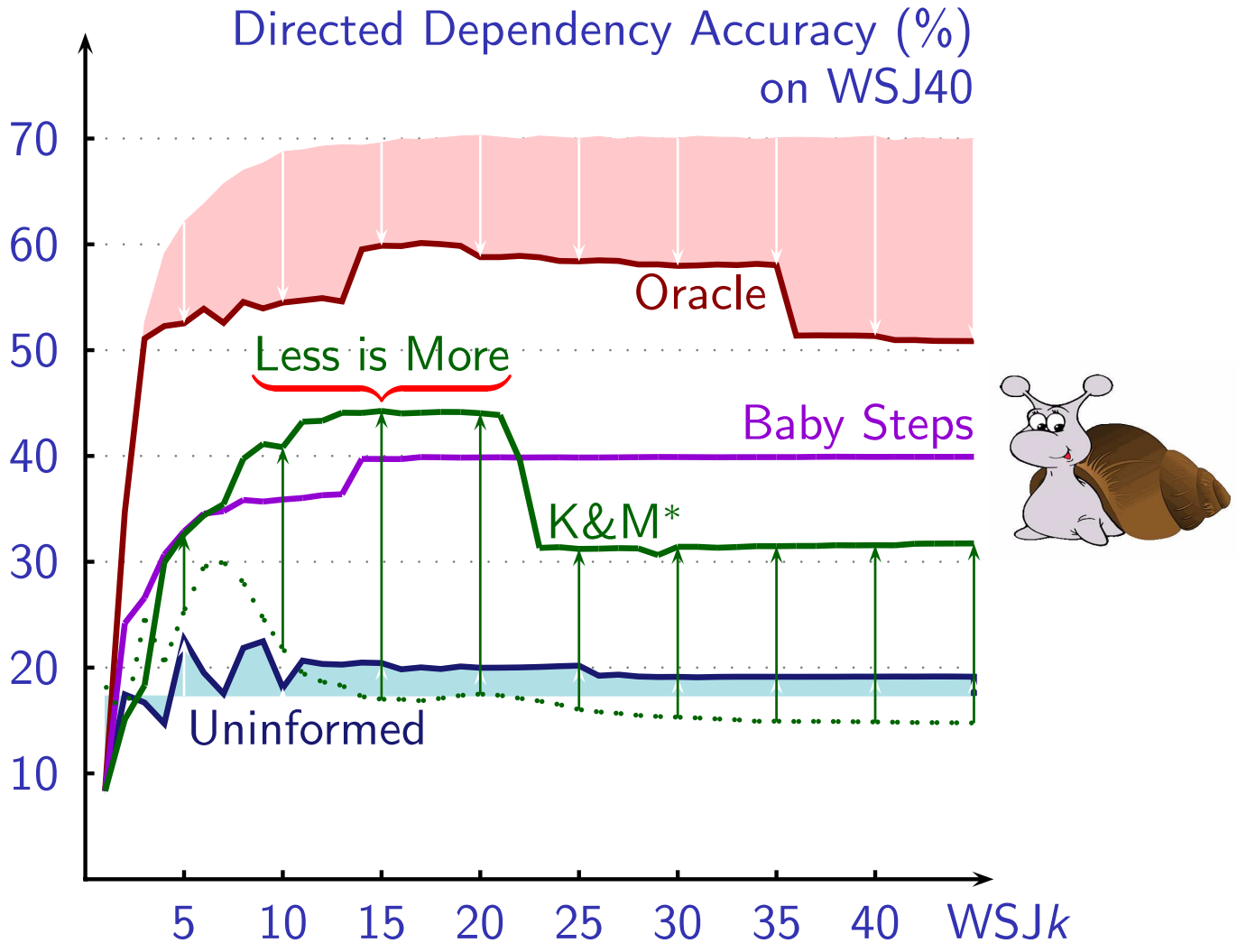


Idea I: Baby Steps ... Observations & Results!

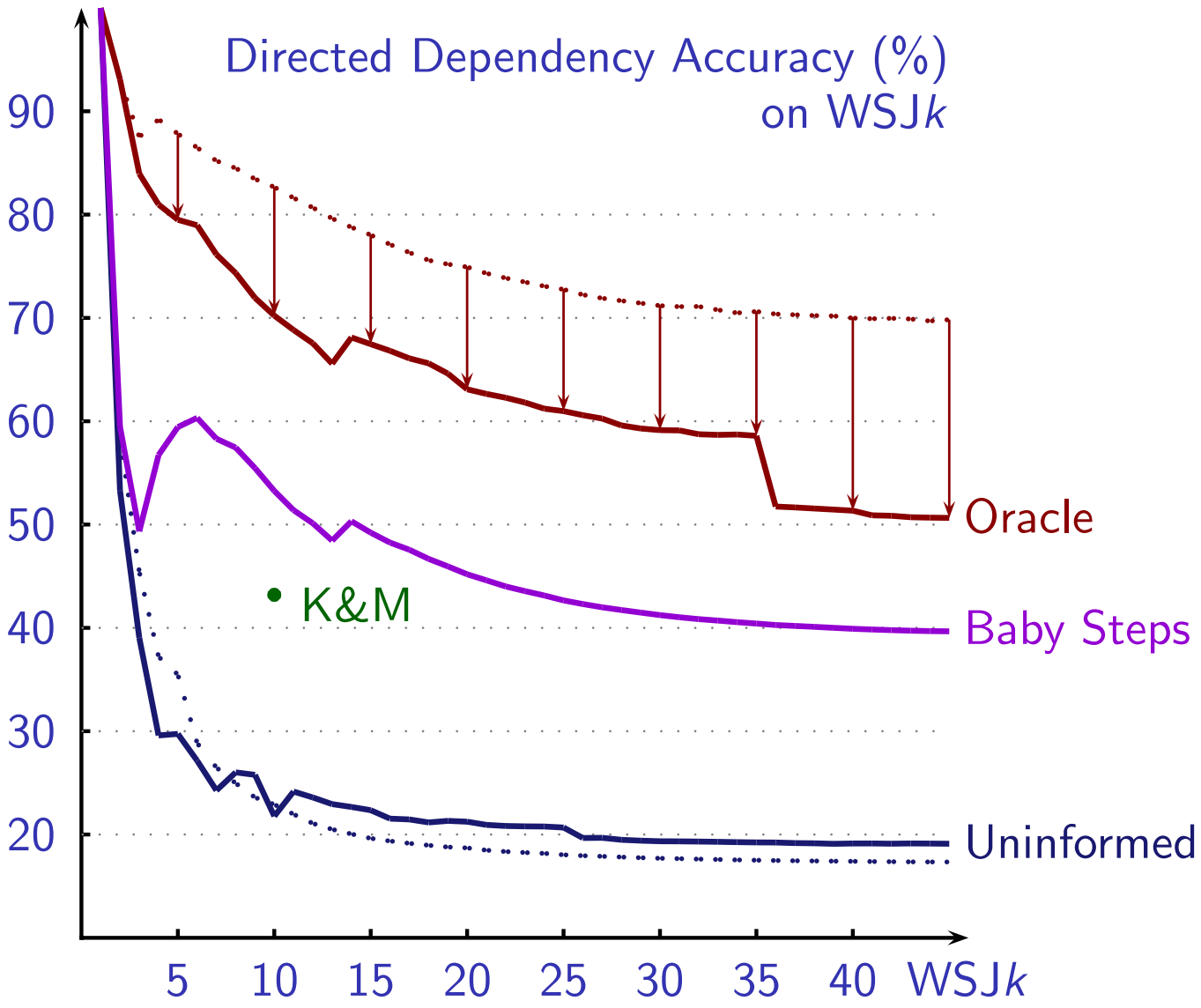




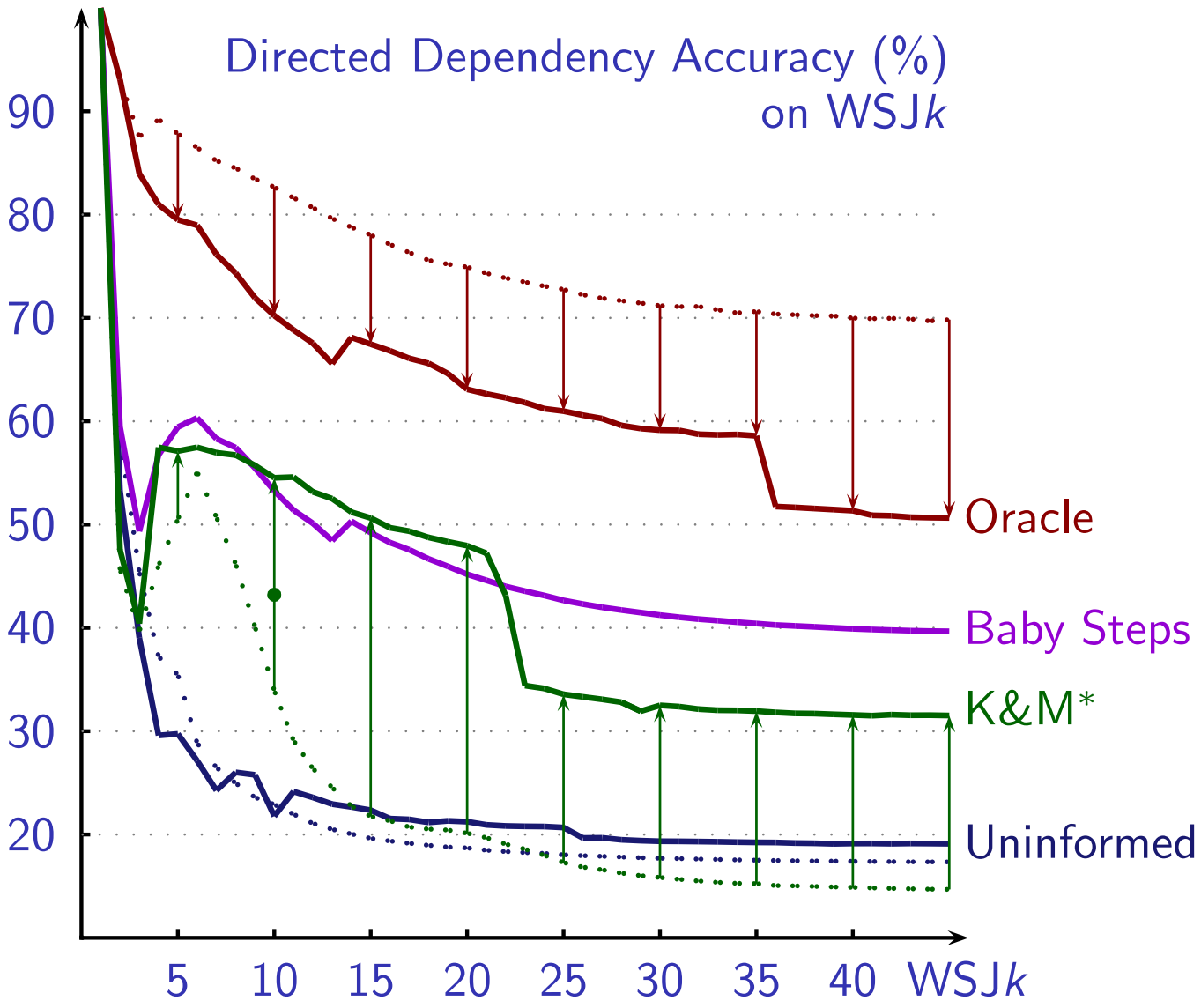
Idea I: Baby Steps ... Observations & Results!



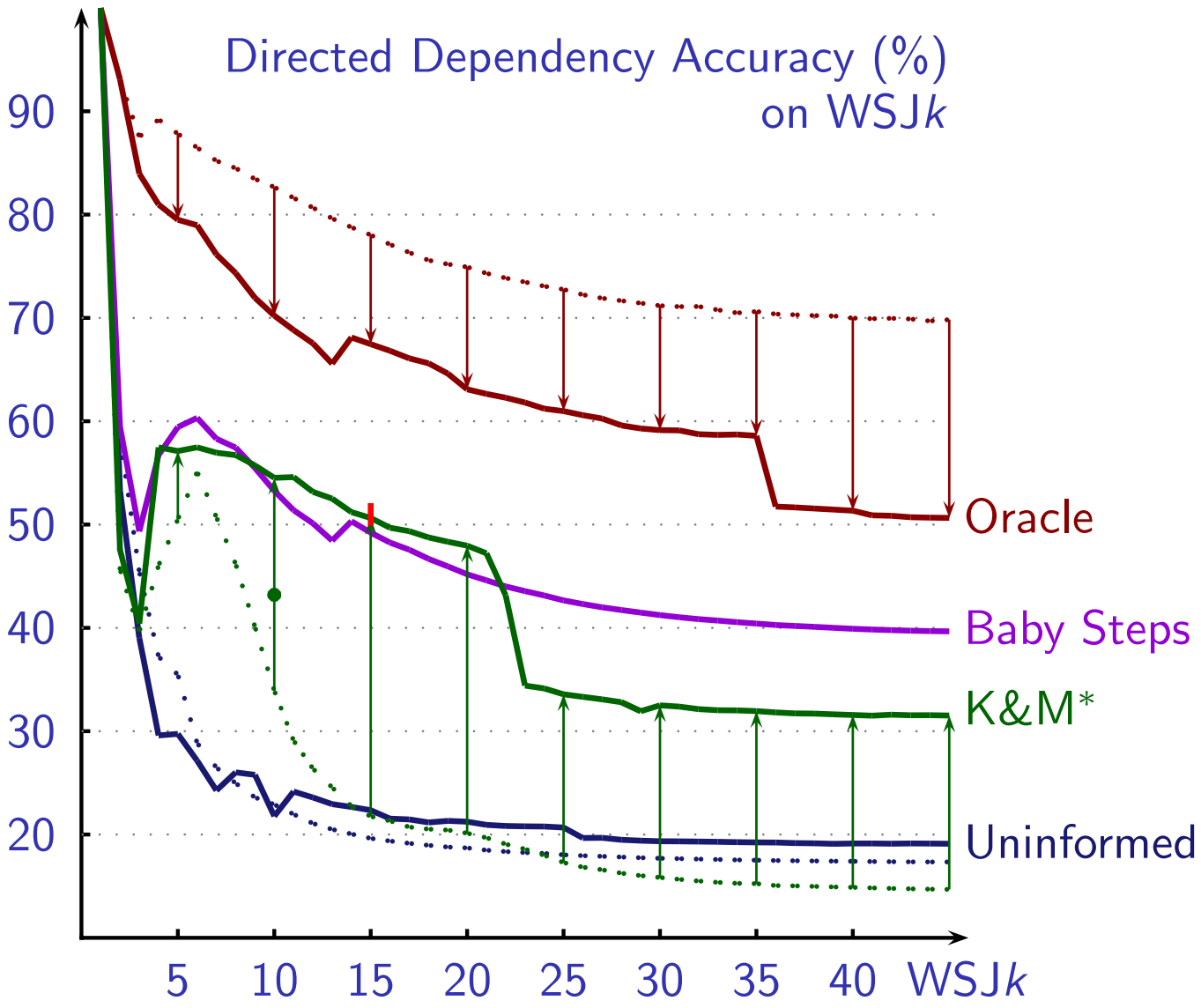
Idea II: Less is More & Leapfrog ... a Hack!



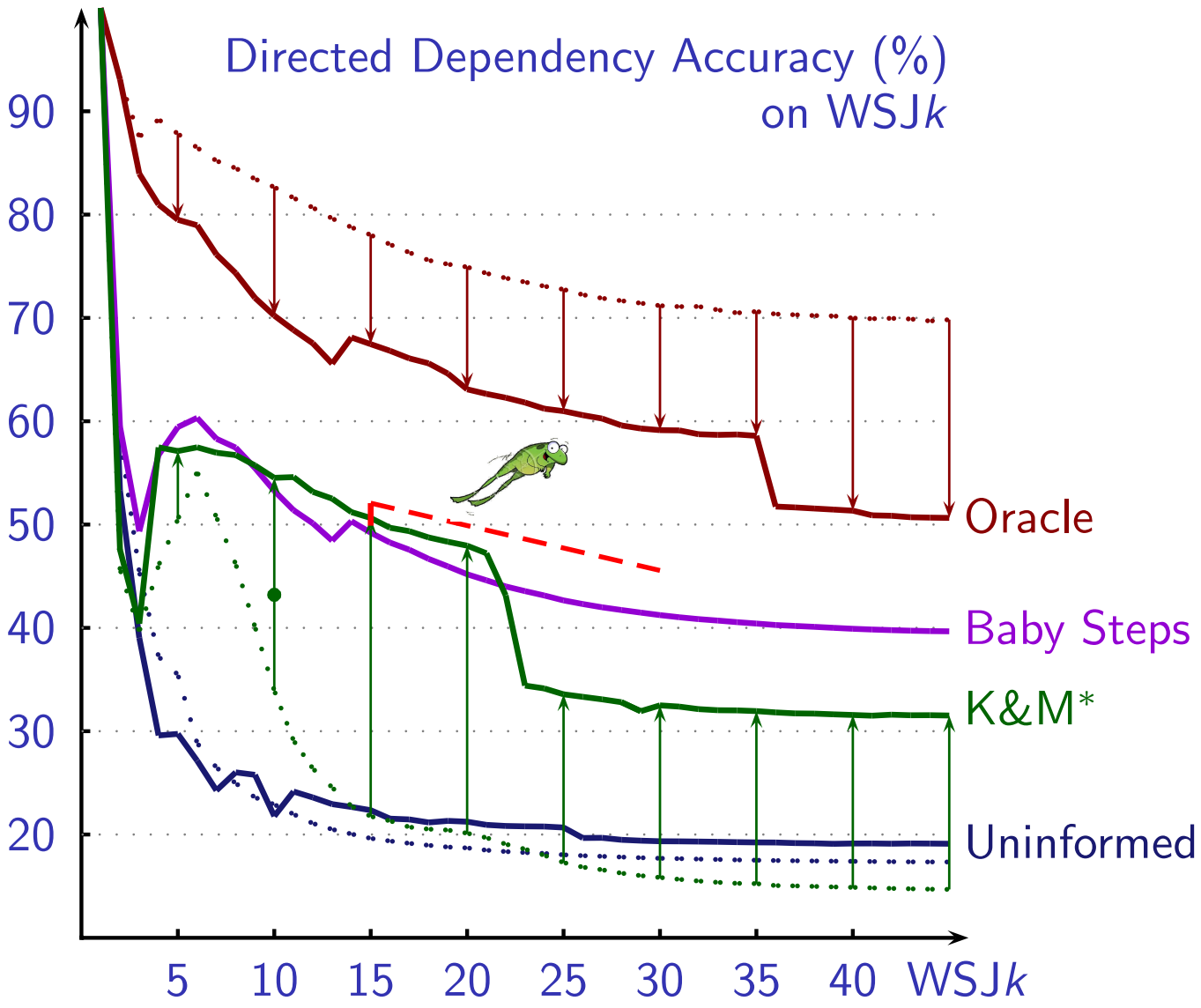
Idea II: Less is More & Leapfrog ... a Hack!



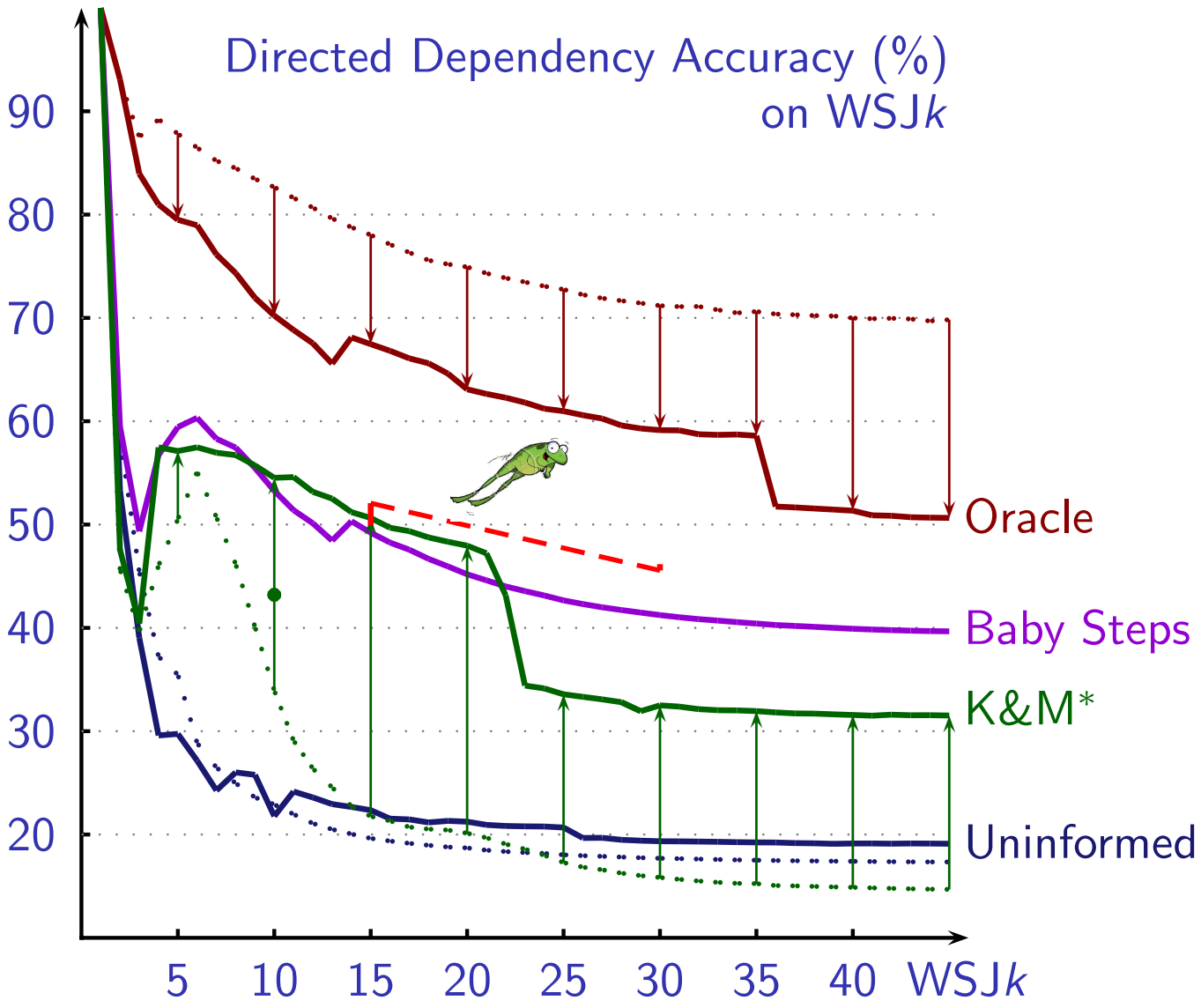
Idea II: Less is More & Leapfrog ... a Hack!



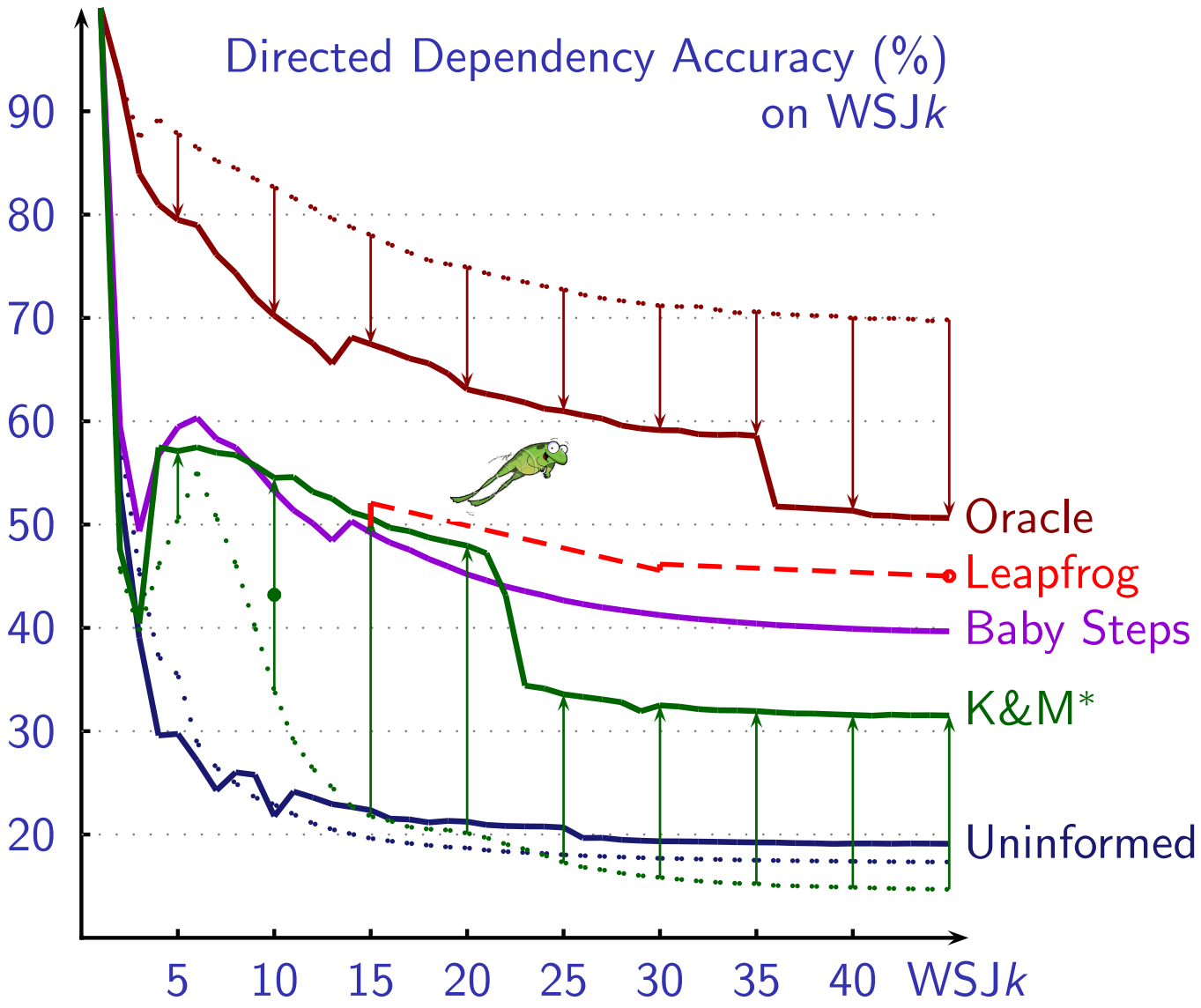
Idea II: Less is More & Leapfrog ... a Hack!



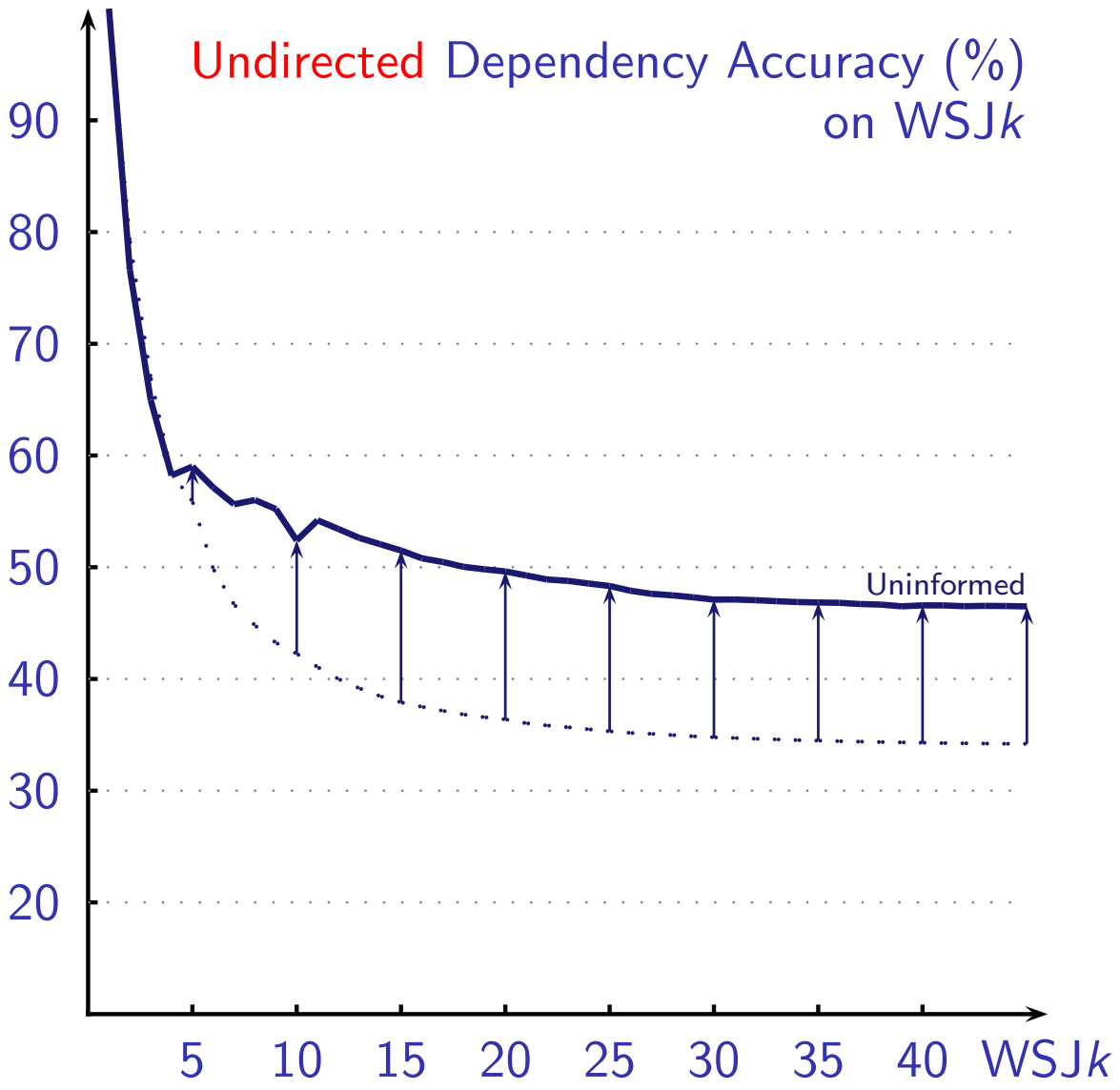
Idea II: Less is More & Leapfrog ... a Hack!



Idea II: Less is More & Leapfrog ... a Hack!



Aside II: Also, to be fair...

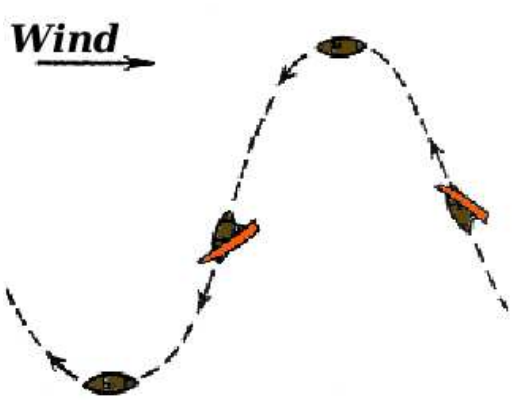


Lateen EM: Intuition

- **Use both objectives (a primary and a secondary).**

Lateen EM: Intuition

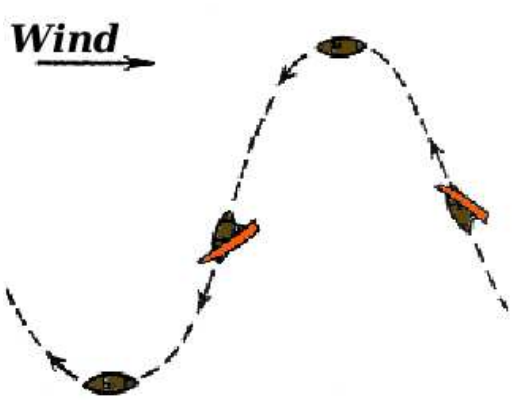
- **Use both objectives (a primary and a secondary).**



As a captain can't count on favorable winds, so an unsupervised learner can't rely on co-operative gradients.

Lateen EM: Intuition

- **Use both objectives (a primary and a secondary).**



As a captain can't count on favorable winds, so an unsupervised learner can't rely on co-operative gradients.

Lateen strategies de-emphasize fixed points, e.g., by tacking around local attractors, in a zig-zag fashion.

Lateen EM: Simple Algorithm

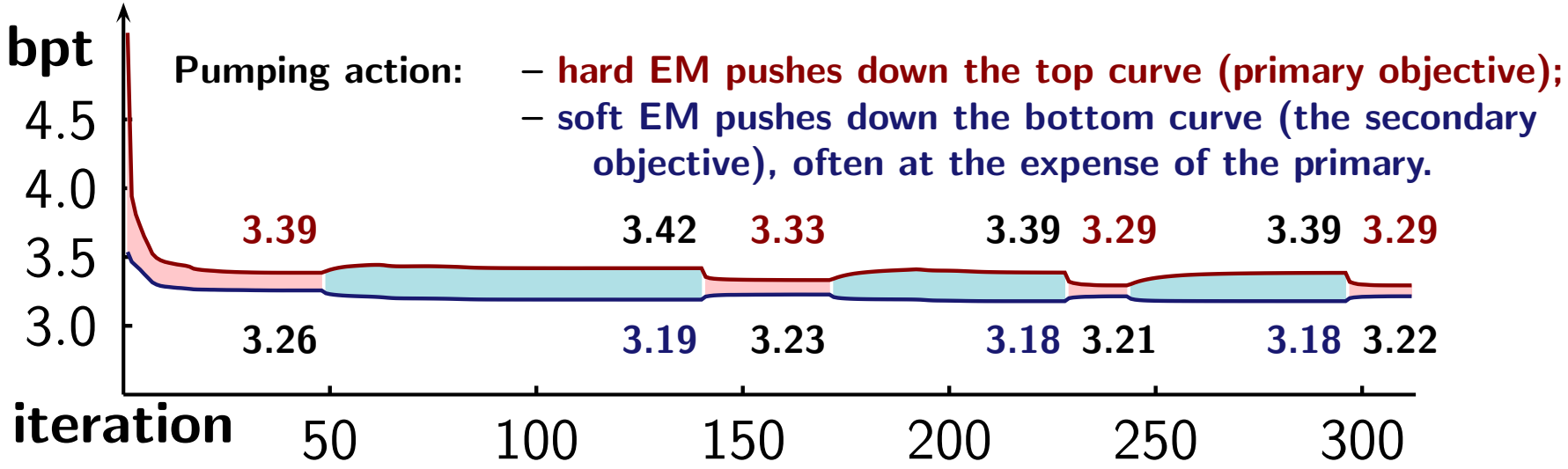
Lateen EM: Simple Algorithm

- **Alternate ordinary soft and hard EM algorithms: switching when stuck helps escape local optima.**

Lateen EM: Simple Algorithm

- **Alternate ordinary soft and hard EM algorithms: switching when stuck helps escape local optima.**

E.g., Italian grammar induction improves from 41.8% to 56.2% after three lateen alternations:



Lateen EM: Early-Stopping

Lateen EM: Early-Stopping

- **Use one objective to validate moves proposed by the other: stop if the secondary objective gets worse.**

Lateen EM: Early-Stopping

- **Use one objective to validate moves proposed by the other: stop if the secondary objective gets worse.**
- **30% faster, on average, than either standard EM.**

Optimization: Summary

Optimization: Summary

- **Take guesswork out of tuning EM:**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change;**
  - **Viterbi EM and Baby Steps don't require initializers.**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change;**
  - **Viterbi EM and Baby Steps don't require initializers.**
- **Exploit multiple views and iterate (Blum and Mitchell, 1998):**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change;**
  - **Viterbi EM and Baby Steps don't require initializers.**
- **Exploit multiple views and iterate (Blum and Mitchell, 1998):**
  - **Lateen EM optimizes for Viterbi parse trees *and* forests;**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change;**
  - **Viterbi EM and Baby Steps don't require initializers.**
- **Exploit multiple views and iterate (Blum and Mitchell, 1998):**
  - **Lateen EM optimizes for Viterbi parse trees *and* forests;**
  - **Baby Steps focuses on simpler sentences in the data.**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change;**
  - **Viterbi EM and Baby Steps don't require initializers.**
- **Exploit multiple views and iterate (Blum and Mitchell, 1998):**
  - **Lateen EM optimizes for Viterbi parse trees *and* forests;**
  - **Baby Steps focuses on simpler sentences in the data.**
- **Capitalize on EM's advantages:**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change;**
  - **Viterbi EM and Baby Steps don't require initializers.**
- **Exploit multiple views and iterate (Blum and Mitchell, 1998):**
  - **Lateen EM optimizes for Viterbi parse trees *and* forests;**
  - **Baby Steps focuses on simpler sentences in the data.**
- **Capitalize on EM's advantages:**
  - **guarantee to not harm a primary likelihood**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change;**
  - **Viterbi EM and Baby Steps don't require initializers.**
- **Exploit multiple views and iterate (Blum and Mitchell, 1998):**
  - **Lateen EM optimizes for Viterbi parse trees *and* forests;**
  - **Baby Steps focuses on simpler sentences in the data.**
- **Capitalize on EM's advantages:**
  - **guarantee to not harm a primary likelihood;**
  - **begin with large steps in a parameter space.**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change;**
  - **Viterbi EM and Baby Steps don't require initializers.**
- **Exploit multiple views and iterate (Blum and Mitchell, 1998):**
  - **Lateen EM optimizes for Viterbi parse trees *and* forests;**
  - **Baby Steps focuses on simpler sentences in the data.**
- **Capitalize on EM's advantages:**
  - **guarantee to not harm a primary likelihood;**
  - **begin with large steps in a parameter space.**
- **Ameliorate EM's disadvantages:**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change;**
  - **Viterbi EM and Baby Steps don't require initializers.**
- **Exploit multiple views and iterate (Blum and Mitchell, 1998):**
  - **Lateen EM optimizes for Viterbi parse trees *and* forests;**
  - **Baby Steps focuses on simpler sentences in the data.**
- **Capitalize on EM's advantages:**
  - **guarantee to not harm a primary likelihood;**
  - **begin with large steps in a parameter space.**
- **Ameliorate EM's disadvantages:**
  - **avoid taking disproportionately many (and ever-smaller) steps to approach a likelihood's fixed point**

Optimization: Summary

- **Take guesswork out of tuning EM:**
  - **Lateen EM ties termination to a sign change;**
  - **Viterbi EM and Baby Steps don't require initializers.**
- **Exploit multiple views and iterate (Blum and Mitchell, 1998):**
  - **Lateen EM optimizes for Viterbi parse trees *and* forests;**
  - **Baby Steps focuses on simpler sentences in the data.**
- **Capitalize on EM's advantages:**
  - **guarantee to not harm a primary likelihood;**
  - **begin with large steps in a parameter space.**
- **Ameliorate EM's disadvantages:**
  - **avoid taking disproportionately many (and ever-smaller) steps to approach a likelihood's fixed point;**
  - **escape by changing objectives and mixing solutions!**

Constraints

Web Markup

Punctuation

Capitalization

Constraints: Supervised and Unsupervised

Constraints: Supervised and Unsupervised

- **compact summaries of high-level insights into a domain**

Constraints: Supervised and Unsupervised

- **compact summaries of high-level insights into a domain — can significantly reduce the search space**

Constraints: Supervised and Unsupervised

- **compact summaries of high-level insights into a domain**
  — can significantly reduce the search space
  — often easier to enforce than to model

Constraints: Supervised and Unsupervised

- **compact summaries of high-level insights into a domain**
  — can significantly reduce the search space
  — often easier to enforce than to model

- **relevant to unsupervised learning (less rope to hang self)**

Constraints: Supervised and Unsupervised

- **compact summaries of high-level insights into a domain**
  — can significantly reduce the search space
  — often easier to enforce than to model

- **relevant to unsupervised learning (less rope to hang self)**
  — in general, steer at the "right" regularities in data

Constraints: Supervised and Unsupervised

- **compact summaries of high-level insights into a domain**
  — can significantly reduce the search space
  — often easier to enforce than to model

- **relevant to unsupervised learning (less rope to hang self)**
  — in general, steer at the "right" regularities in data
  — linguistic structure underdetermined by raw text

Constraints: Supervised and Unsupervised

- **compact summaries of high-level insights into a domain**
  — can significantly reduce the search space
  — often easier to enforce than to model

- **relevant to unsupervised learning (less rope to hang self)**
  — in general, steer at the "right" regularities in data
  — linguistic structure underdetermined by raw text

- **partial bracketings (Pereira and Schabes, 1992)**

Constraints: Supervised and Unsupervised

- compact summaries of high-level insights into a domain
  — can significantly reduce the search space
  — often easier to enforce than to model

- relevant to unsupervised learning (less rope to hang self)
  — in general, steer at the "right" regularities in data
  — linguistic structure underdetermined by raw text

- partial bracketings (Pereira and Schabes, 1992)

- synchronous grammars (Alshawi and Douglas, 2000)
- linear-time parsing, skewness of trees, etc. (Seginer, 2007)
- sparse posterior regularization (Ganchev et al., 2009)

Syntax of Markup: Common Constituents

..., but [$_S$ [$_{NP}$ the <a> *Toronto Star*][$_{VP}$ reports [$_{NP}$ this] [$_{PP}$ in the softest possible way],[$_S$ stating ...]]]

Syntax of Markup: Common Constituents

..., but [$_S$ [$_{NP}$ the <a>*Toronto Star*][$_{VP}$ reports [$_{NP}$ this] [$_{PP}$ in the softest possible way],[$_S$ stating ...]]]

$$S \rightarrow \underline{NP\ \ VP}$$

Syntax of Markup: Common Constituents

..., but [$_S$ [$_{NP}$ the <a>*Toronto Star*][$_{VP}$ reports [$_{NP}$ this] [$_{PP}$ in the softest possible way],[$_S$ stating ...]]]

S → NP VP → DT NNP NNP VBZ NP PP S

Syntax of Markup: Constituent Productions

	%
NP → NNP NNP	9.6
NP → NNP	4.6
NP → NP PP	3.4
NP → NNP NNP NNP	2.4
NP → DT NNP NNP	2.1
NP → NN	1.8
NP → DT NNP NNP NNP	1.7
NP → DT NN	1.7
NP → DT NNP NNP	1.6
S → NP VP	1.4
NP → DT NNP NNP NNP	1.2
NP → DT JJ NN	1.1
NP → NNS	1.0
NP → JJ NN	0.8
NP → NP NP	0.8
	35.3

Syntax of Markup: Constituent Productions

	%
NP → NNP NNP	9.6
NP → NNP	4.6
NP → NP PP	3.4
NP → NNP NNP NNP	2.4
NP → DT NNP NNP	2.1
NP → NN	1.8
NP → DT NNP NNP NNP	1.7
NP → DT NN	1.7
NP → DT NNP NNP	1.6
S → NP VP	1.4
NP → DT NNP NNP NNP	1.2
NP → DT JJ NN	1.1
NP → NNS	1.0
NP → JJ NN	0.8
NP → NP NP	0.8
	35.3

Syntax of Markup: Constituent Productions

	%
NP → NNP NNP	9.6
NP → NNP	4.6
NP → NP PP	3.4
NP → NNP NNP NNP	2.4
NP → DT NNP NNP	2.1
NP → NN	1.8
NP → DT NNP NNP NNP	1.7
NP → DT NN	1.7
NP → DT NNP NNP	1.6
S → NP VP	1.4
NP → DT NNP NNP NNP	1.2
NP → DT JJ NN	1.1
NP → NNS	1.0
NP → JJ NN	0.8
NP → NP NP	0.8
	35.3

Syntax of Markup: Common Dependencies

..., but [$_S$ [$_{NP}$ the <a>*Toronto Star*][$_{VP}$ reports [$_{NP}$ this] [$_{PP}$ in the softest possible way],[$_S$ stating ...]]]

Syntax of Markup: Common Dependencies

..., but [$_S$ [$_{NP}$ the <a>*Toronto Star*][$_{VP}$ reports [$_{NP}$ this] [$_{PP}$ in the softest possible way],[$_S$ stating ...]]]

Syntax of Markup: Common Dependencies

..., but [$_S$ [$_{NP}$ the <a>*Toronto Star*][$_{VP}$ reports [$_{NP}$ this] [$_{PP}$ in the softest possible way],[$_S$ stating ...]]]

"the <a>*Star* reports"

Syntax of Markup: Head-Outward Spawns

	%
NNP	24.4
NN	8.1
DT NNP	6.1
DT NN	5.9
NNS	4.5
NNPS	1.4
VBG	1.3
NNP NNP NN	1.2
VBD	1.0
IN	1.0
VBN	1.0
DT JJ NN	0.9
VBZ	0.9
POS NNP	0.9
JJ	0.8
	59.4

Syntax of Markup: Head-Outward Spawns

	%
NNP	24.4
NN	8.1
DT NNP	6.1
DT NN	5.9
NNS	4.5
NNPS	1.4
VBG	1.3
NNP NNP NN	1.2
VBD	1.0
IN	1.0
VBN	1.0
DT JJ NN	0.9
VBZ	0.9
POS NNP	0.9
JJ	0.8
	59.4

Syntax of Markup: Head-Outward Spawns

	%
NNP	24.4
NN	8.1
DT NNP	6.1
DT NN	5.9
NNS	4.5
NNPS	1.4
VBG	1.3
NNP NNP NN	1.2
VBD	1.0
IN	1.0
VBN	1.0
DT JJ NN	0.9
VBZ	0.9
POS NNP	0.9
JJ	0.8
	59.4

Formatting:

- **strong connection between markup and syntax**

Formatting:

- **strong connection between markup and syntax**
- **yields a suite of accurate parsing constraints**

Formatting:

- strong connection between markup and syntax
- yields a suite of accurate parsing constraints
- improves grammar induction over web data

Formatting:

- **strong connection between markup and syntax**
- **yields a suite of accurate parsing constraints**
- **improves grammar induction over web data, but...**

Formatting:

- strong connection between markup and syntax
- yields a suite of accurate parsing constraints
- improves grammar induction over web data, but...
  - works better with more natural language-resources!

Formatting:

- strong connection between markup and syntax
- yields a suite of accurate parsing constraints
- improves grammar induction over web data, but...
  - works better with more natural language-resources!

- missing structural cues:

Formatting:

- **strong connection between markup and syntax**
- **yields a suite of accurate parsing constraints**
- **improves grammar induction over web data, but...**
  - **works better with more natural language-resources!**

- **missing structural cues:**
  — e.g., punctuation **(will make heavy use)**

Formatting:

- strong connection between markup and syntax
- yields a suite of accurate parsing constraints
- improves grammar induction over web data, but...
  - works better with more natural language-resources!

- missing structural cues:
  — e.g., punctuation (will make heavy use)
  and capitalization (will not discuss today)

Formatting:

- strong connection between markup and syntax
- yields a suite of accurate parsing constraints
- improves grammar induction over web data, but...
  - works better with more natural language-resources!

- missing structural cues:
  — e.g., punctuation (will make heavy use)
  and capitalization (will not discuss today)

- raw word streams often difficult even for humans
  — e.g., transcribed utterances (Kim and Woodland, 2002)

Example: Raw Word Stream

ALTHOUGH IT PROBABLY HAS REDUCED THE LEVEL OF EXPENDITURES FOR SOME PURCHASERS UTILIZATION MANAGEMENT LIKE MOST OTHER COST CONTAINMENT STRATEGIES DOESN'T APPEAR TO HAVE ALTERED THE LONG-TERM RATE OF INCREASE IN HEALTH-CARE COSTS THE INSTITUTE OF MEDICINE AN AFFILIATE OF THE NATIONAL ACADEMY OF SCIENCES CONCLUDED AFTER A TWO-YEAR STUDY

Example:

Example:

[$_{\text{SBAR}}$ **Although it probably has reduced the level of expenditures for some purchasers**],

Example:

[$_{\text{SBAR}}$ **Although it probably has reduced the level of expenditures for some purchasers**], [$_{\text{NP}}$ **utilization management**] —

Example:

[$_{\text{SBAR}}$ Although it probably has reduced the level of expenditures for some purchasers], [$_{\text{NP}}$ utilization management] — [$_{\text{PP}}$ like most other cost containment strategies] —

Example:

[$_{\text{SBAR}}$ Although it probably has reduced the level of expenditures for some purchasers], [$_{\text{NP}}$ utilization management] — [$_{\text{PP}}$ like most other cost containment strategies] — [$_{\text{VP}}$ doesn't appear to have altered the long-term rate of increase in health-care costs],

Example:

[$_{\text{SBAR}}$ Although it probably has reduced the level of expenditures for some purchasers], [$_{\text{NP}}$ utilization management] — [$_{\text{PP}}$ like most other cost containment strategies] — [$_{\text{VP}}$ doesn't appear to have altered the long-term rate of increase in health-care costs], [$_{\text{NP}}$ the Institute of Medicine],

Example:

[$_{\text{SBAR}}$ **Although it probably has reduced the level of expenditures for some purchasers**], [$_{\text{NP}}$ **utilization management**] — [$_{\text{PP}}$ **like most other cost containment strategies**] — [$_{\text{VP}}$ **doesn't appear to have altered the long-term rate of increase in health-care costs**], [$_{\text{NP}}$ **the Institute of Medicine**], [$_{\text{NP}}$ **an affiliate of the National Academy of Sciences**],

Example:

[$_{\text{SBAR}}$ Although it probably has reduced the level of expenditures for some purchasers], [$_{\text{NP}}$ utilization management] — [$_{\text{PP}}$ like most other cost containment strategies] — [$_{\text{VP}}$ doesn't appear to have altered the long-term rate of increase in health-care costs], [$_{\text{NP}}$ the Institute of Medicine], [$_{\text{NP}}$ an affiliate of the National Academy of Sciences], [$_{\text{VP}}$ concluded after a two-year study].

Intuition:

- **strong constraint**

Intuition:

- **strong constraint: (head $\leftarrow$ head) in training**

Intuition:

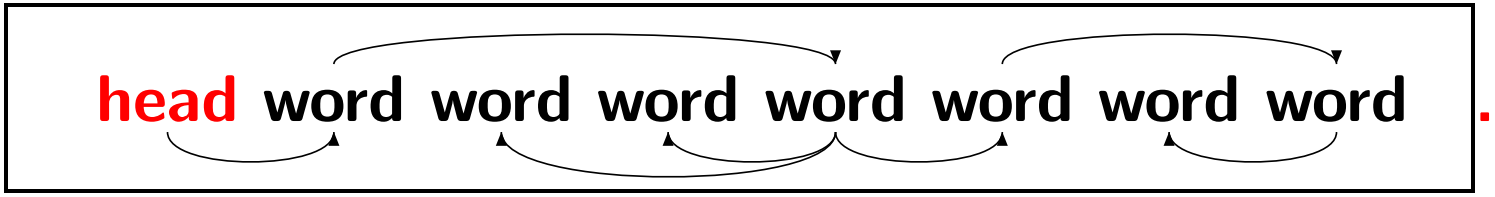
- **strong constraint: (head ← head) in training**

word head	,	head word word	,

head word word word word word word word	.

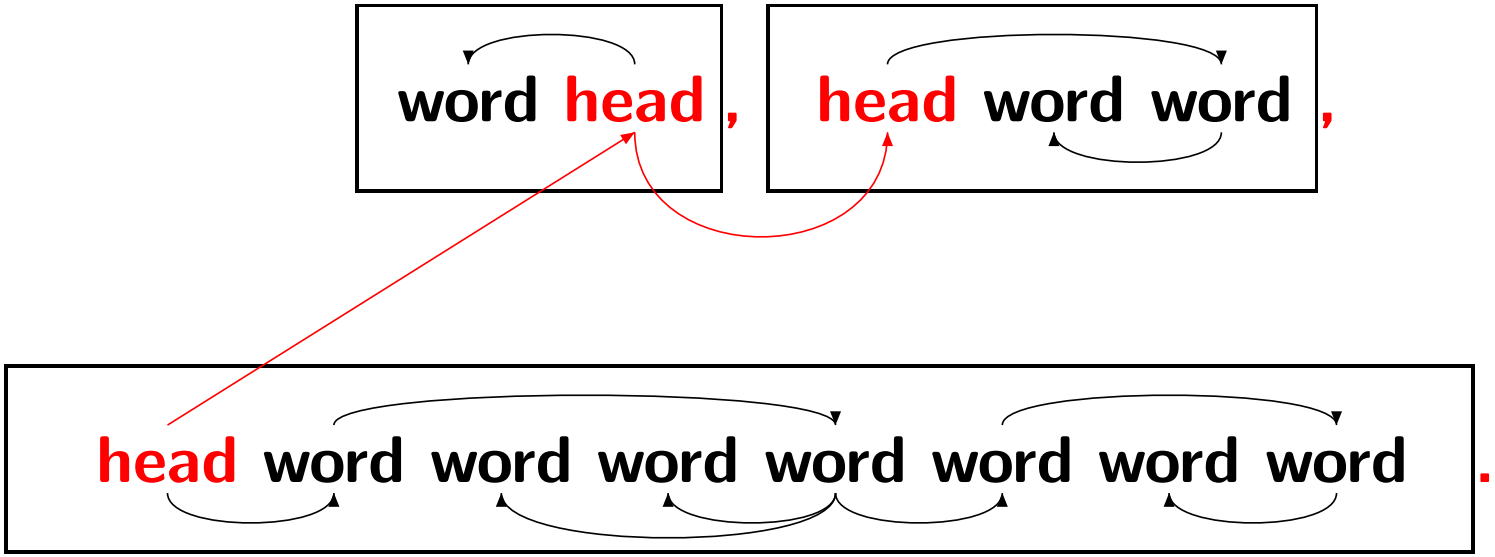
Intuition:

- **strong constraint: (head ← head) in training**



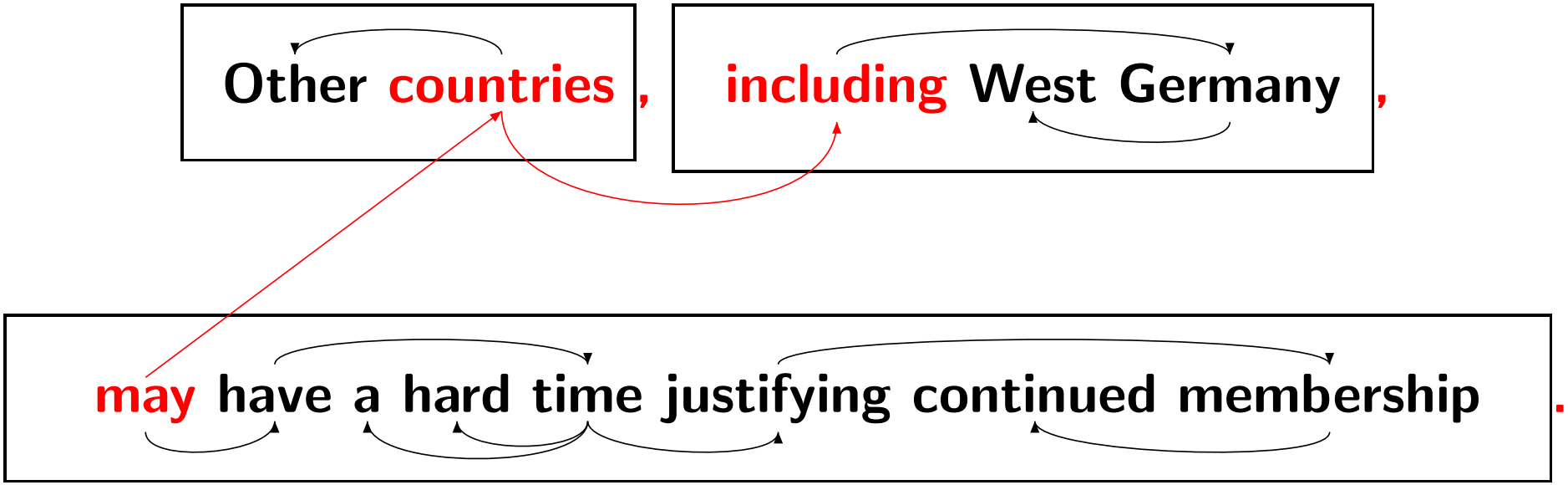
Intuition:

- **strong constraint: (head ← head) in training**



Intuition:

- **strong constraint: (head ← head) in training**



Intuition:

- **weak constraint**

Intuition:

- **weak constraint: (head $\leftarrow$ external word) in inference**

Intuition:

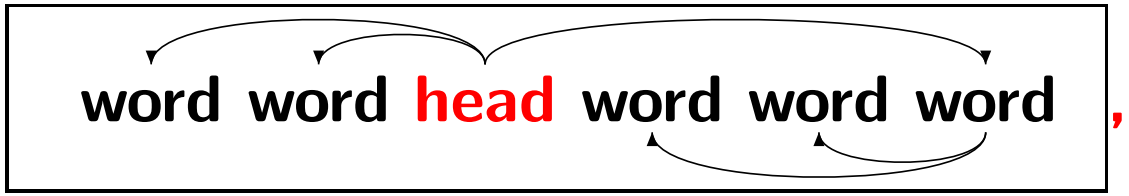
- **weak constraint: (head ← external word) in inference**

word word head word word word ,

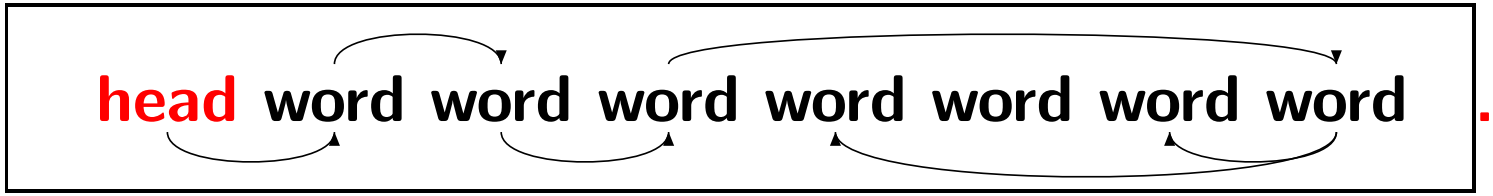
head word word word word word word word .

Intuition:

- **weak constraint: (head ← external word) in inference**



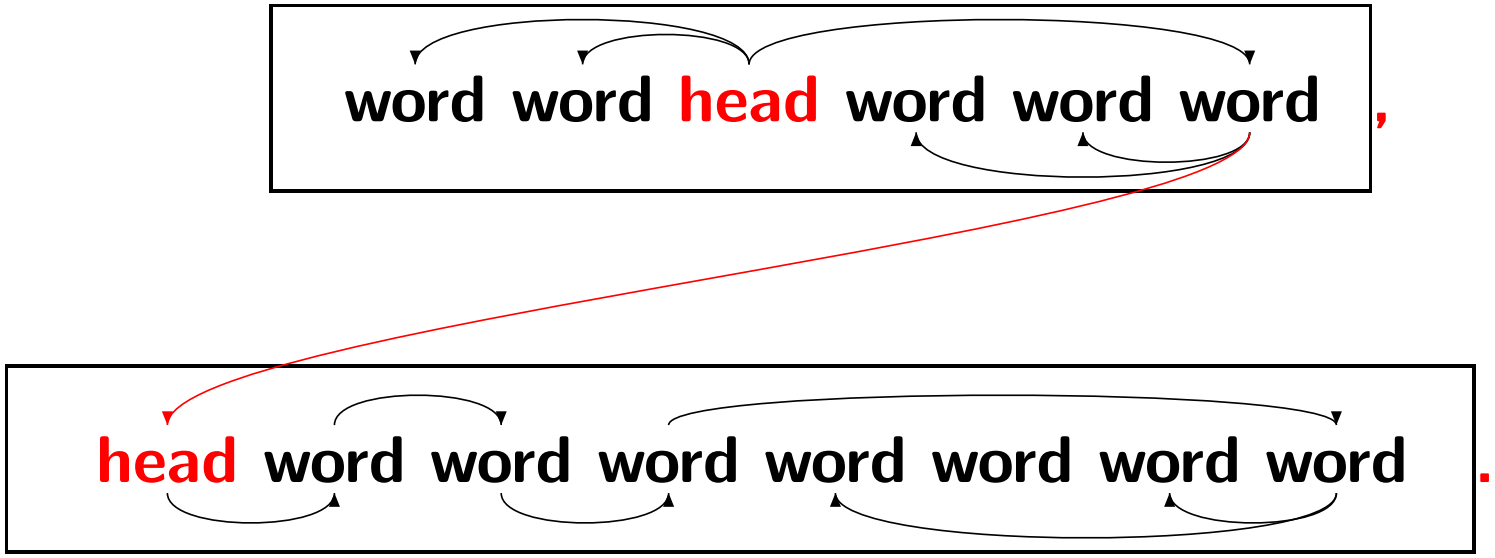
,



.

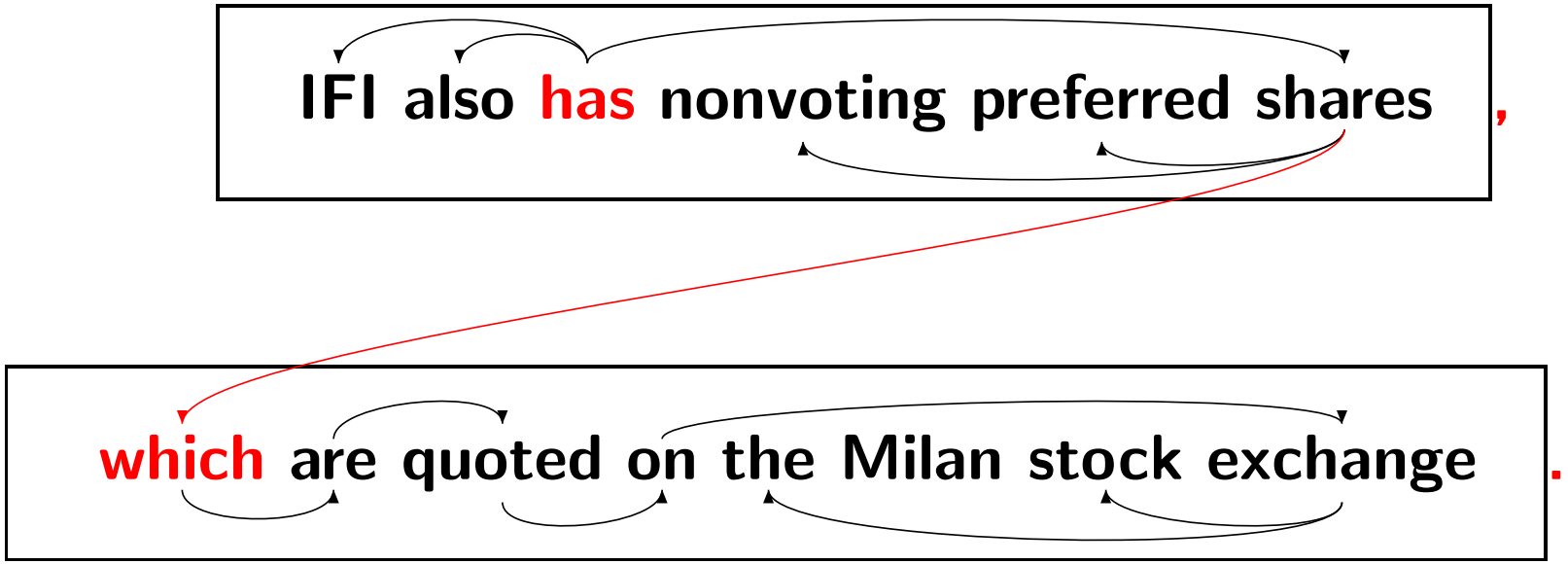
Intuition:

- **weak constraint: (head ← external word) in inference**



Intuition:

- **weak constraint: (head ← external word) in inference**



Linguistic Analysis:

- **punctuation and syntax are related (Nunberg, 1990; Briscoe, 1994; Jones 1994; Doran, 1998,** *inter alia***)**

Linguistic Analysis:

- **punctuation and syntax are related (Nunberg, 1990; Briscoe, 1994; Jones 1994; Doran, 1998, *inter alia*)**
- **49.4% of inter-punctuation fragments are constituents**

Linguistic Analysis:

- **punctuation and syntax are related (Nunberg, 1990; Briscoe, 1994; Jones 1994; Doran, 1998, *inter alia*)**
- **49.4% of inter-punctuation fragments are constituents**
- **many fragments derive precisely themselves:**

	%
IN	9.6
NN	7.2
NNP	6.3
CD	3.8
VBD	3.2
VBZ	3.0
RB	2.8
VBG	2.2
VBP	1.9
NNS	1.8
WDT	1.6
MD	1.1
VBN	1.1
IN VBD	1.0
JJ	0.7
	52.8

Training:

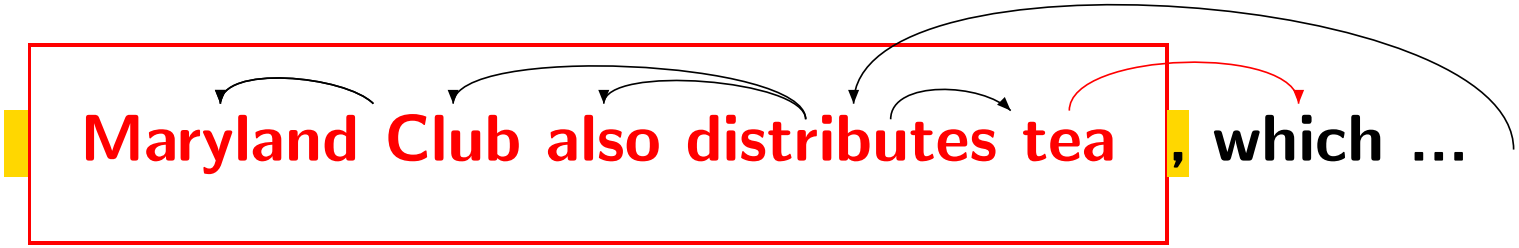
- **strong constraint, e.g.,**

... arrests followed a "*Snake Day*" at Utrecht ...

— already 74.0% agreement with head-percolated trees.

Inference:

- **weak constraint, e.g.,**



— now 92.9% agreement with head-percolated trees!

Modeling

Context-Sensitive Unsupervsed Tags
Dependency-and-Boundary Models
Reduced Models of Grammar

Example: Actually, state-of-the-art uses gold tags!

IN PRP RB VBZ VBN DT NN IN NNS IN DT
NNS NN NN IN RBS JJ NN NN NNS VBZ RB
VB TO VB VBN DT JJ NN IN NN IN JJ NNS
DT NNP IN NNP DT NN IN DT NNP NNP IN
NNPS VBD IN DT JJ NN

Word Categories: Two benefits of tag-sets...

- **Grouping: pooling statistics of words that play similar syntactic roles improves generalization (reduces sparsity):**

Word Categories: Two benefits of tag-sets...

- **Grouping: pooling statistics of words that play similar syntactic roles improves generalization (reduces sparsity):**
  - **indeed, working directly with words does not do well.**

Word Categories: Two benefits of tag-sets...

- **Grouping: pooling statistics of words that play similar syntactic roles improves generalization (reduces sparsity):**
  - **indeed, working directly with words does not do well.**
- **Disambiguation: for words that take on multiple parts of speech, knowing gold tags limits the parsing search space:**

Word Categories: Two benefits of tag-sets...

- **Grouping: pooling statistics of words that play similar syntactic roles improves generalization (reduces sparsity):**
  - **indeed, working directly with words does not do well.**
- **Disambiguation: for words that take on multiple parts of speech, knowing gold tags limits the parsing search space:**
  - **reducing manual tags to monosemous clusterings lowers performance below that of the unsupervised categories constructed by Finkel and Manning (2009) / Clark (2003).**

Word Categories: Two benefits of tag-sets...

- **Grouping: pooling statistics of words that play similar syntactic roles improves generalization (reduces sparsity):**
  - **indeed, working directly with words does not do well.**
- **Disambiguation: for words that take on multiple parts of speech, knowing gold tags limits the parsing search space:**
  - **reducing manual tags to monosemous clusterings lowers performance below that of the unsupervised categories constructed by Finkel and Manning (2009) / Clark (2003).**

Can improve with context-sensitive unsupervised clusters!

Word Categories: Two benefits of tag-sets...

- **Grouping: pooling statistics of words that play similar syntactic roles improves generalization (reduces sparsity):**
  - **indeed, working directly with words does not do well.**
- **Disambiguation: for words that take on multiple parts of speech, knowing gold tags limits the parsing search space:**
  - **reducing manual tags to monosemous clusterings lowers performance below that of the unsupervised categories constructed by Finkel and Manning (2009) / Clark (2003).**

Can improve with context-sensitive unsupervised clusters!

1. **start with a hard assignment;** **(standard word clustering)**

Word Categories: Two benefits of tag-sets...

- **Grouping**: pooling statistics of words that play similar syntactic roles improves generalization (reduces sparsity):
  - indeed, working directly with words does not do well.
- **Disambiguation**: for words that take on multiple parts of speech, knowing gold tags limits the parsing search space:
  - reducing manual tags to monosemous clusterings lowers performance below that of the unsupervised categories constructed by Finkel and Manning (2009) / Clark (2003).

Can improve with **context-sensitive** unsupervised clusters!

1. start with a hard assignment; (standard word clustering)
2. inject context-colored noise; (get out of the local optimum)

Word Categories: Two benefits of tag-sets...

- **Grouping**: pooling statistics of words that play similar syntactic roles improves generalization (reduces sparsity):
  - indeed, working directly with words does not do well.
- **Disambiguation**: for words that take on multiple parts of speech, knowing gold tags limits the parsing search space:
  - reducing manual tags to monosemous clusterings lowers performance below that of the unsupervised categories constructed by Finkel and Manning (2009) / Clark (2003).

Can improve with **context-sensitive** unsupervised clusters!

1. start with a hard assignment; (standard word clustering)
2. inject context-colored noise; (get out of the local optimum)
3. Viterbi-train a bitag HMM. (the unTagger)

Word Categories: Breaking the barrier...

- **Swapped out gold tags for unsupervised word categories:**

System Description	*Accuracy*
"punctuation" with gold tags	**58.4**

Word Categories: Breaking the barrier...

- **Swapped out gold tags for unsupervised word categories:**

System Description	*Accuracy*
"punctuation" with **gold** tags	**58.4**
"punctuation" with **monosemous** induced tags	**58.2 (−0.2)**

Word Categories: Breaking the barrier...

- **Swapped out gold tags for unsupervised word categories:**

System Description	*Accuracy*	
"punctuation" with **gold** tags	**58.4**	
"punctuation" with **monosemous** induced tags	**58.2**	**(−0.2)**
"punctuation" with **context-sensitive** induced tags	**59.1**	**(+0.7)**

Word Categories: Breaking the barrier...

- **Swapped out gold tags for unsupervised word categories:**

System Description	*Accuracy*
"punctuation" with **gold** tags	**58.4**
"punctuation" with **monosemous** induced tags	**58.2 (−0.2)**
"punctuation" with **context-sensitive** induced tags	**59.1 (+0.7)**

  - **only a small drop from switching to (monosemous) unsupervised clusters — previous systems lost** ~ 5**pts**

Word Categories: Breaking the barrier...

- **Swapped out gold tags for unsupervised word categories:**

System Description	*Accuracy*	
"punctuation" with **gold** tags	**58.4**	
"punctuation" with **monosemous** induced tags	**58.2**	**(−0.2)**
"punctuation" with **context-sensitive** induced tags	**59.1**	**(+0.7)**

  - **only a small drop from switching to (monosemous) unsupervised clusters — previous systems lost** $\sim$ 5**pts;**
  - **first state-of-the-art system to improve with (context-sensitive) unsupervised clusters!**

Dependency-and-Boundary Models (DBMs):

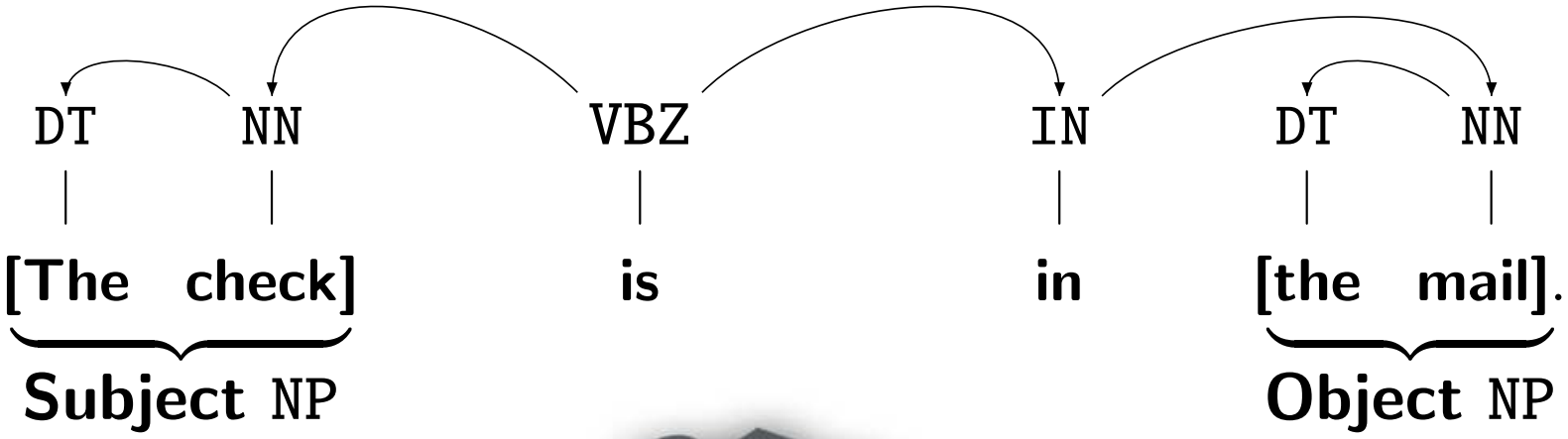
- **Use boundary cues in head-driven dependency grammars.**

Dependency-and-Boundary Models (DBMs):

- **Use boundary cues in head-driven dependency grammars.**
- **E.g., induce structure by working inwards from edges**

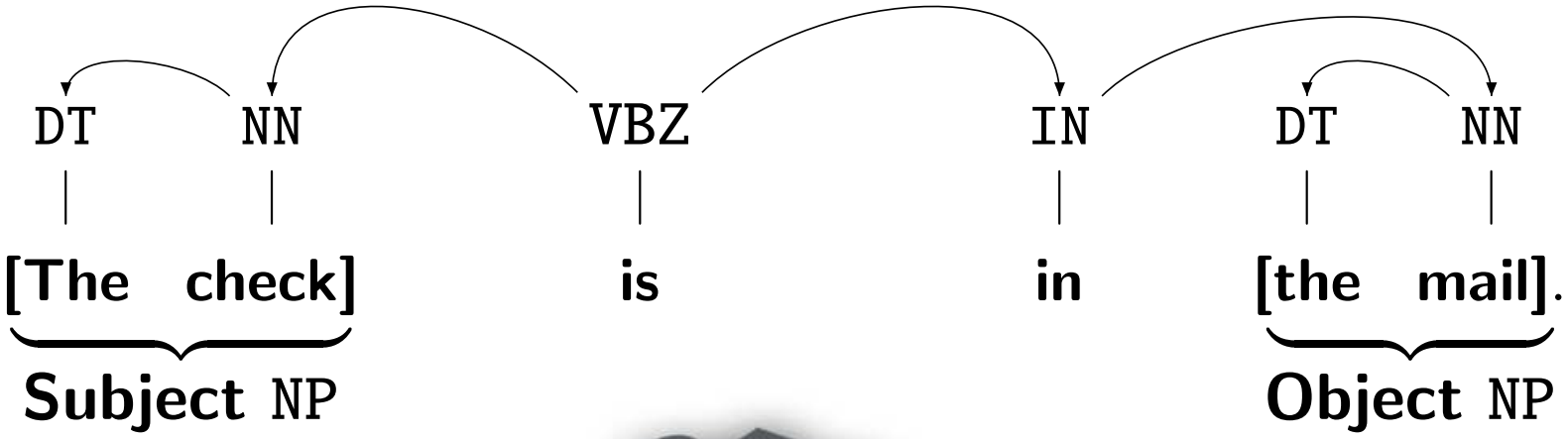
Dependency-and-Boundary Models (DBMs):

- **Use boundary cues in head-driven dependency grammars.**
- **E.g., induce structure by working inwards from edges:**



Dependency-and-Boundary Models (DBMs):

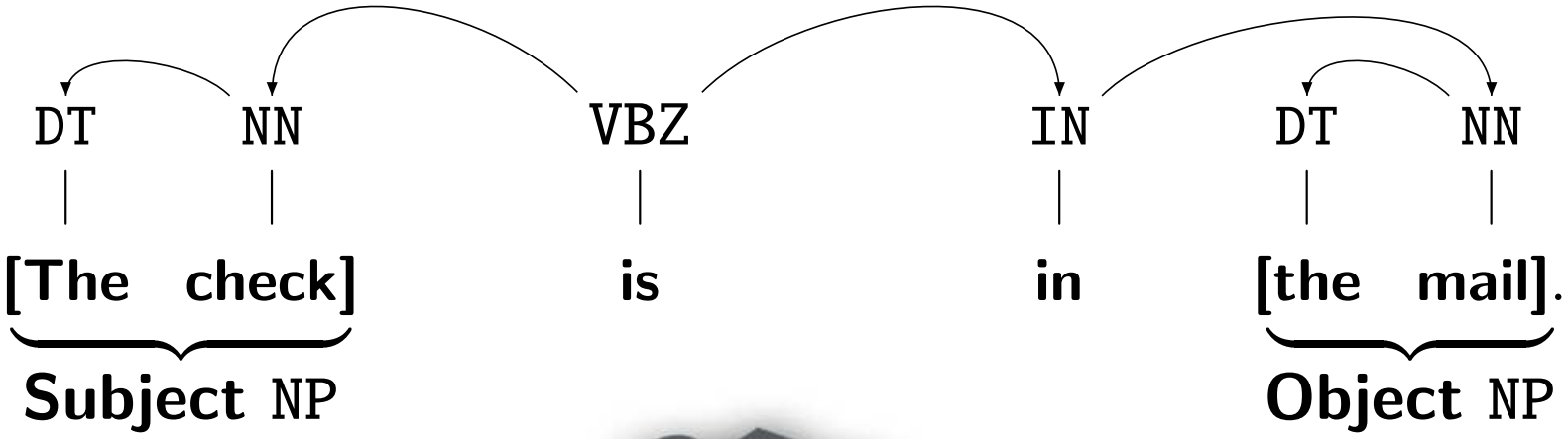
- **Use boundary cues in head-driven dependency grammars.**
- **E.g., induce structure by working inwards from edges:**



- **learn from left fringe (determiner DT) to parse object NP**

Dependency-and-Boundary Models (DBMs):

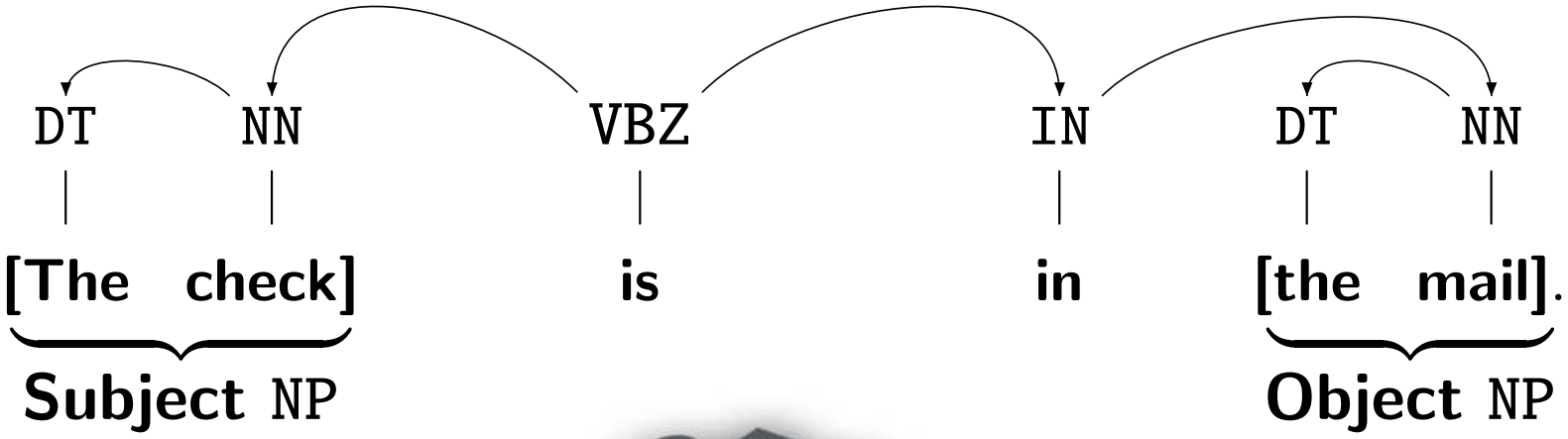
- **Use boundary cues in head-driven dependency grammars.**
- **E.g., induce structure by working inwards from edges:**



- **learn from left fringe (determiner DT) to parse object NP**
- **based on right fringe (noun NN), correctly parse subject NP**

Dependency-and-Boundary Models (DBMs):

- **Use boundary cues in head-driven dependency grammars.**
- **E.g., induce structure by working inwards from edges:**



- **learn from left fringe (determiner DT) to parse object NP**
- **based on right fringe (noun NN), correctly parse subject NP**
- **between them, glean make-up of larger phrases (e.g., VP)**

Dependency-and-Boundary Models: Highlights

Dependency-and-Boundary Models: Highlights

- **DBM-1: Use words at the fringes.**

Dependency-and-Boundary Models: Highlights

- **DBM-1: Use words at the fringes.**
  - **truly head-outward model** **(Alshawi, 1996)**

Dependency-and-Boundary Models: Highlights

- **DBM-1: Use words at the fringes.**
  - **truly head-outward model** **(Alshawi, 1996)**
  - **conditions on what can more often be seen!**

Dependency-and-Boundary Models: Highlights

- **DBM-1: Use words at the fringes.**
  - **truly head-outward model** **(Alshawi, 1996)**
  - **conditions on what can more often be seen!**
- **DBM-2: Fragments differ from complete sentences.**

Dependency-and-Boundary Models: Highlights

- **DBM-1: Use words at the fringes.**
  - **truly head-outward model** **(Alshawi, 1996)**
  - **conditions on what can more often be seen!**
- **DBM-2: Fragments differ from complete sentences.**
  - **incomplete fragments are uncharacteristically short**

Dependency-and-Boundary Models: Highlights

- **DBM-1: Use words at the fringes.**
  - **truly head-outward model** **(Alshawi, 1996)**
  - **conditions on what can more often be seen!**
- **DBM-2: Fragments differ from complete sentences.**
  - **incomplete fragments are uncharacteristically short**
  - **roots of fragments are generally not verbs or modals**

Dependency-and-Boundary Models: Highlights

- **DBM-1: Use words at the fringes.**
  - **truly head-outward model** **(Alshawi, 1996)**
  - **conditions on what can more often be seen!**
- **DBM-2: Fragments differ from complete sentences.**
  - **incomplete fragments are uncharacteristically short**
  - **roots of fragments are generally not verbs or modals**
- **DBM-3: Learn to stitch together fragments.**

Class-based, head-outward generation (Alshawi, 1996)

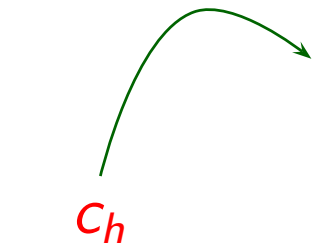
Class-based, head-outward generation (Alshawi, 1996)

c_h

c_e

$dir = \mathrm{R}$ $\qquad$ $adj = \mathrm{T}$

Class-based, head-outward generation (Alshawi, 1996)

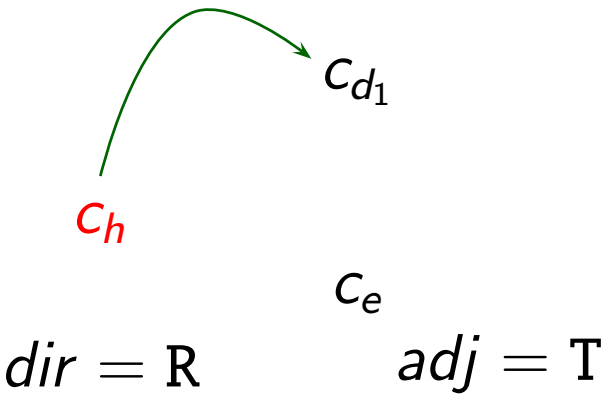


c_h

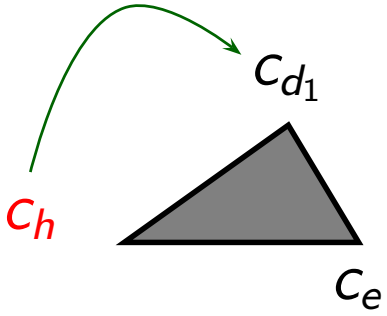
c_e

$dir = \mathrm{R}$ $\quad$ $adj = \mathrm{T}$

Class-based, head-outward generation (Alshawi, 1996)

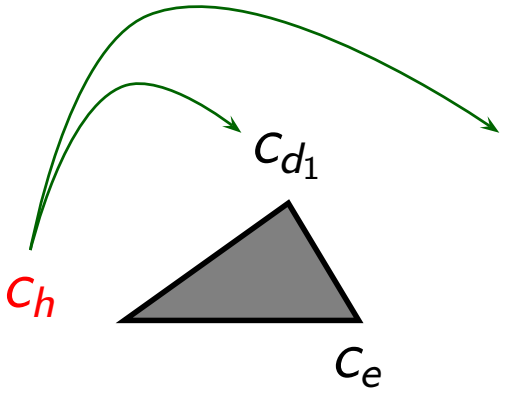


Class-based, head-outward generation (Alshawi, 1996)



$dir = \mathrm{R}$ $adj = \mathrm{F}$

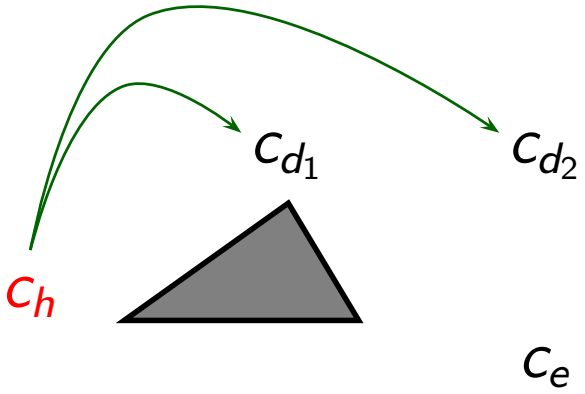
Class-based, head-outward generation (Alshawi, 1996)



$dir = \mathrm{R}$ $adj = \mathrm{F}$

Class-based, head-outward generation

(Alshawi, 1996)

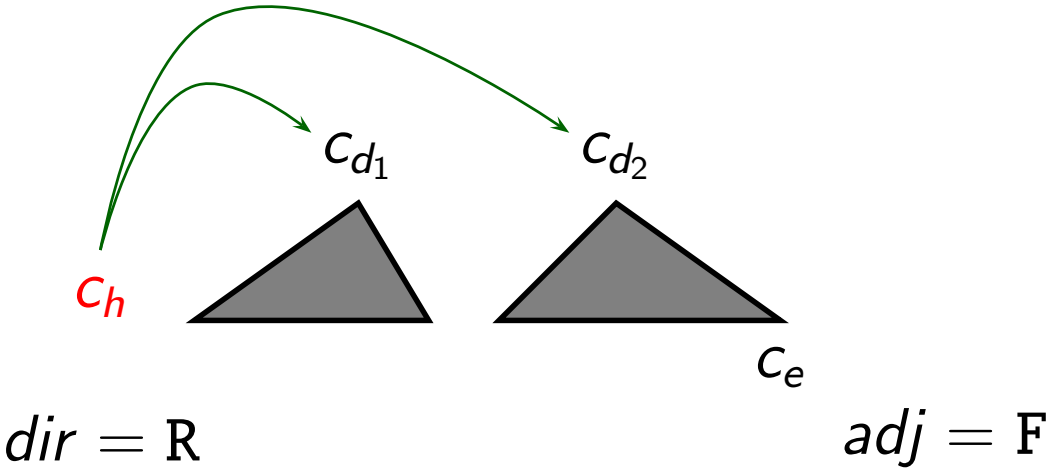


$dir = \mathrm{R}$

$adj = \mathrm{F}$

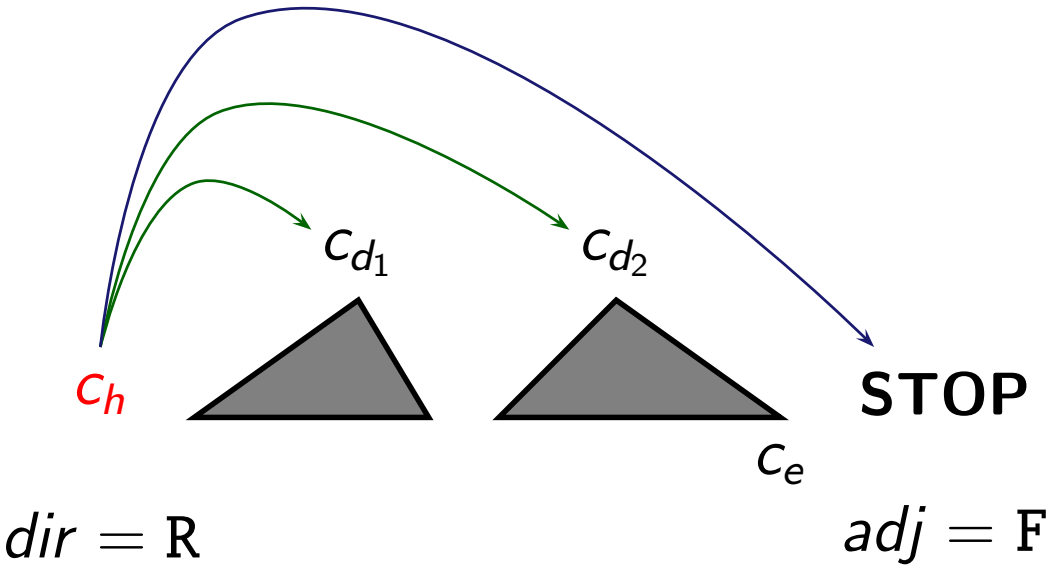
Class-based, head-outward generation

(Alshawi, 1996)

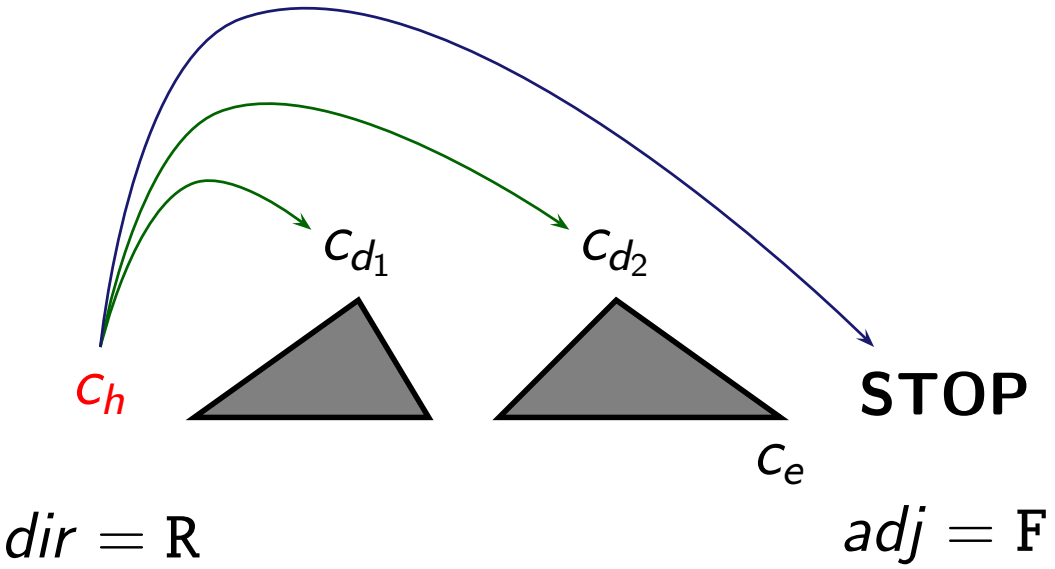


Class-based, head-outward generation

(Alshawi, 1996)

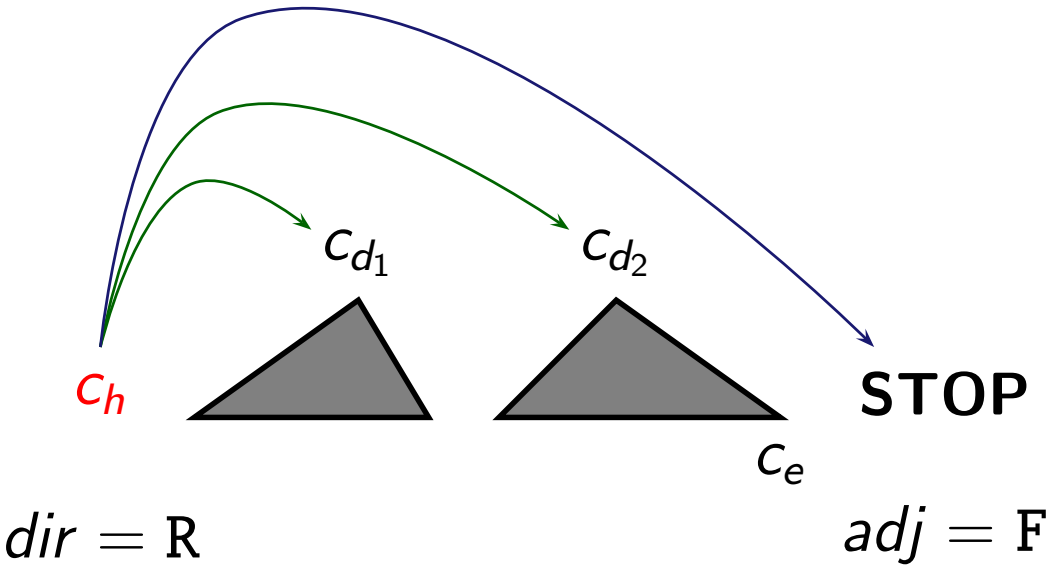


Class-based, head-outward generation (Alshawi, 1996)



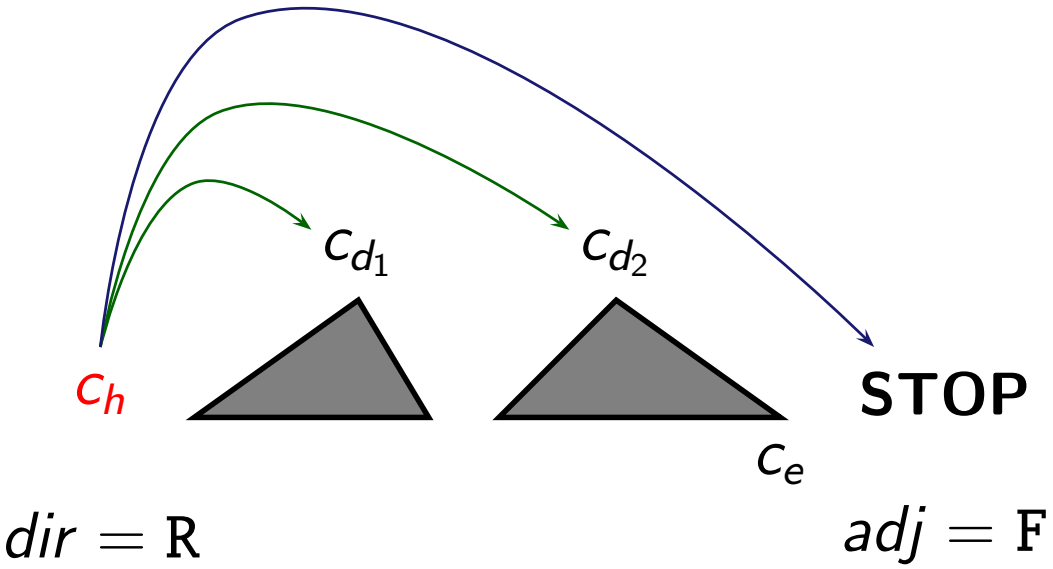
$\mathbb{P}_{\mathrm{ROOT}}(c_h \mid comp)$

Class-based, head-outward generation (Alshawi, 1996)



$$\mathbb{P}_{\text{ROOT}}(c_h \mid comp)\ \mathbb{P}_{\text{ATTACH}}(c_d \mid c_h, dir, cross)$$

Class-based, head-outward generation (Alshawi, 1996)



$$\mathbb{P}_{\text{ROOT}}(c_h \mid comp)\ \mathbb{P}_{\text{ATTACH}}(c_d \mid c_h, dir, cross)\ \mathbb{P}_{\text{STOP}}(\mid dir, adj, c_e, comp)$$

How to better exploit more data better?

- **more text in long sentences, but those can be hard...**

How to better exploit more data better?

- **more text in long sentences, but those can be hard...**
  - **require Viterbi training, punctuation constraints, etc.**

How to better exploit more data better?

- **more text in long sentences, but those can be hard...**
  - **require Viterbi training, punctuation constraints, etc.**
- **could we "start small"** *and still* **use more data?**

How to better exploit more data better?

- **more text in long sentences, but those can be hard...**
  - **require Viterbi training, punctuation constraints, etc.**
- **could we "start small"** *and still* **use more data?**
  - **... and wouldn't it be nice if we could just split things up!**

What if we chopped up input at punctuation?

- **impact on *quantity* of data (with a 15-token threshold):**
  - **more and simpler word sequences incorporated earlier**

What if we chopped up input at punctuation?

- **impact on *quantity* of data (with a 15-token threshold):**
  - **more and simpler word sequences incorporated earlier**
  - **much *more dense* coverage of available data:**
    - **number of training inputs goes up 2.2x**
    - **number of tokens increases 4.3x**

What if we chopped up input at punctuation?

- **impact on *quantity* of data (with a 15-token threshold):**
  - **more and simpler word sequences incorporated earlier**
  - **much *more dense* coverage of available data:**
    - **number of training inputs goes up 2.2x**
    - **number of tokens increases 4.3x**

- **but, also impact on *quality* of data**

What if we chopped up input at punctuation?

- **impact on *quantity* of data (with a 15-token threshold):**
  - **more and simpler word sequences incorporated earlier**
  - **much *more dense* coverage of available data:**
    - **number of training inputs goes up 2.2x**
    - **number of tokens increases 4.3x**

- **but, also impact on *quality* of data:**
  - **many fewer complete sentences exhibiting full structure**

What if we chopped up input at punctuation?

- **impact on *quantity* of data (with a 15-token threshold):**
  - **more and simpler word sequences incorporated earlier**
  - **much *more dense* coverage of available data:**
    - **number of training inputs goes up 2.2x**
    - **number of tokens increases 4.3x**

- **but, also impact on *quality* of data:**
  - **many fewer complete sentences exhibiting full structure**
  - **even *less representative* than short inputs:**
    - **but mostly phrases and clauses...**

What if we chopped up input at punctuation?

- **impact on *quantity* of data (with a 15-token threshold):**
  - **more and simpler word sequences incorporated earlier**
  - **much *more dense* coverage of available data:**
    - **number of training inputs goes up 2.2x**
    - **number of tokens increases 4.3x**
- **but, also impact on *quality* of data:**
  - **many fewer complete sentences exhibiting full structure**
  - **even *less representative* than short inputs:**
    - **but mostly phrases and clauses...**
- **... however, we have an appropriate model family!**

Example (cont'd)

Example (cont'd)

	length & type		*left & right*	
complete	**51**	**S**	**IN**	**NN**
incomplete	**12**	**SBAR**	**IN**	**NNS**
	2	**NP**	**NN**	**NN**
	6	**PP**	**IN**	**NNS**
	14	**VP**	**VBZ**	**NNS**
	4	**NP**	**DT**	**NNP**
	8	**NP**	**DT**	**NNPS**
	5	**VP**	**VBD**	**NN**

Example (cont'd)

DBM-2

	length & type		*left & right*	
complete	**51**	**S**	**IN**	**NN**
incomplete	**12**	**SBAR**	**IN**	**NNS**
	2	**NP**	**NN**	**NN**
	6	**PP**	**IN**	**NNS**
	14	**VP**	**VBZ**	**NNS**
	4	**NP**	**DT**	**NNP**
	8	**NP**	**DT**	**NNPS**
	5	**VP**	**VBD**	**NN**

Example (cont'd)

	length & type		**DBM-1** *left & right*	
complete	**51**	**S**	**IN**	**NN**
incomplete	**12**	**SBAR**	**IN**	**NNS**
	2	**NP**	**NN**	**NN**
	6	**PP**	**IN**	**NNS**
	14	**VP**	**VBZ**	**NNS**
	4	**NP**	**DT**	**NNP**
	8	**NP**	**DT**	**NNPS**
	5	**VP**	**VBD**	**NN**

Example (cont'd)

	length & type		*left & right*	
complete	**51**	**S**	**IN**	**NN**
incomplete	**12**	**SBAR**	**IN**	**NNS**
	2	**NP**	**NN**	**NN**
	6	**PP**	**IN**	**NNS**
	14	**VP**	**VBZ**	**NNS**
	4	**NP**	**DT**	**NNP**
	8	**NP**	**DT**	**NNPS**
DBM-3	**5**	**VP**	**VBD**	**NN**

partial parse forests

"easy-first" (Goldberg and Elhadad, 2010)

Example (cont'd)

	length & type		*left & right*	
complete	**51**	**S**	**IN**	**NN**
incomplete	**12**	**SBAR**	**IN**	**NNS**
	2	**NP**	**NN**	**NN**
	6	**PP**	**IN**	**NNS**
reduced model	**14**	**VP**	**VBZ**	**NNS**
DBM-0	**4**	**NP**	**DT**	**NNP**
	8	**NP**	**DT**	**NNPS**
	5	**VP**	**VBD**	**NN**

partial parse forests **"easy-first" (Goldberg and Elhadad, 2010)**

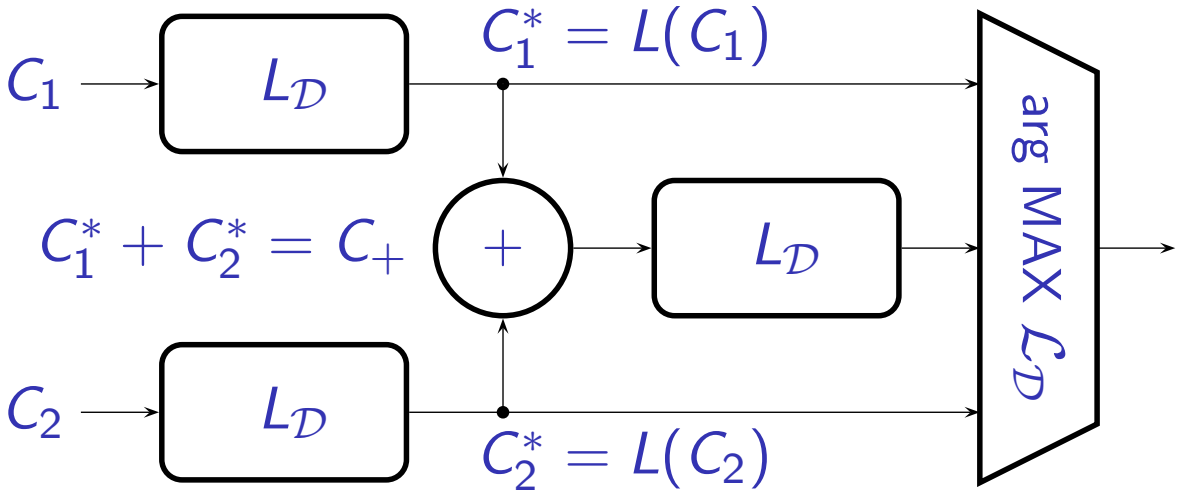
Integrated System

System: Combiners...

(from Leapfrog)

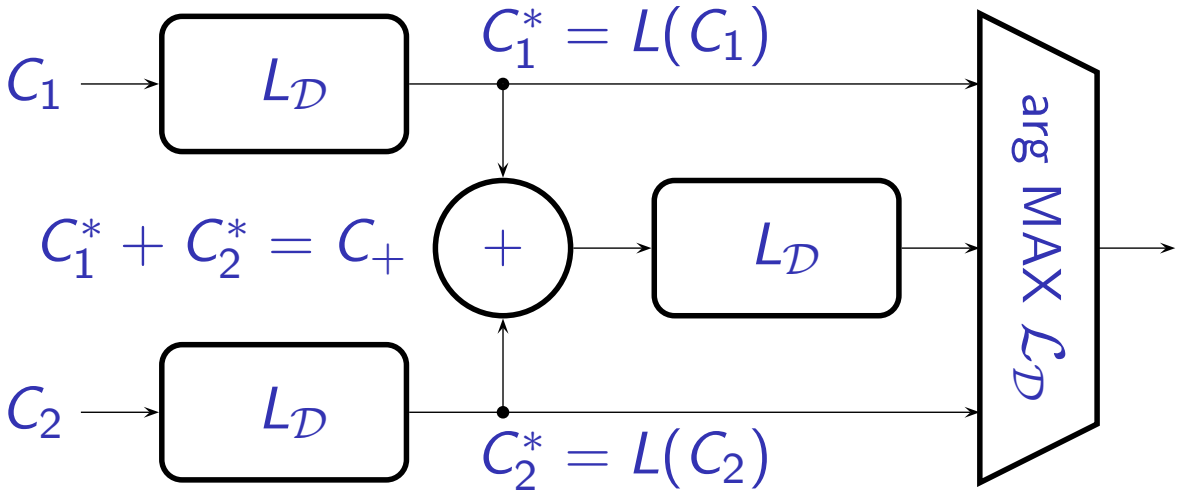
System: Combiners...

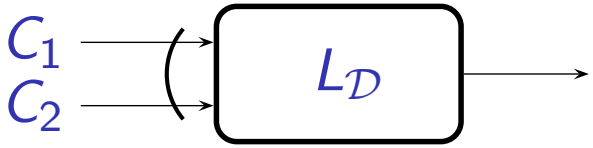
(from Leapfrog)



System: Combiners...

(from Leapfrog)



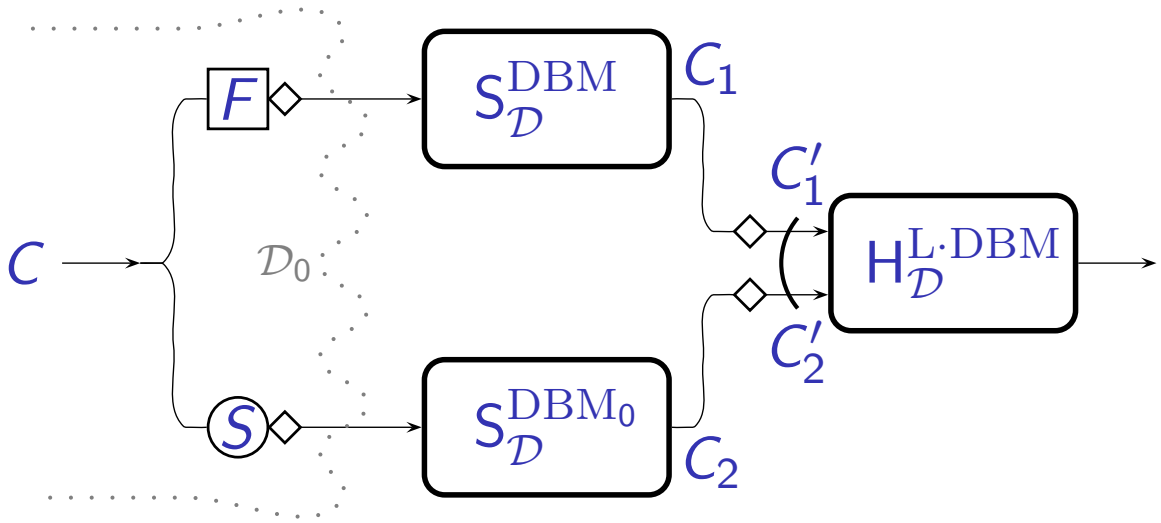


System: Inductors...

(from Baby Steps and Lateen EM)

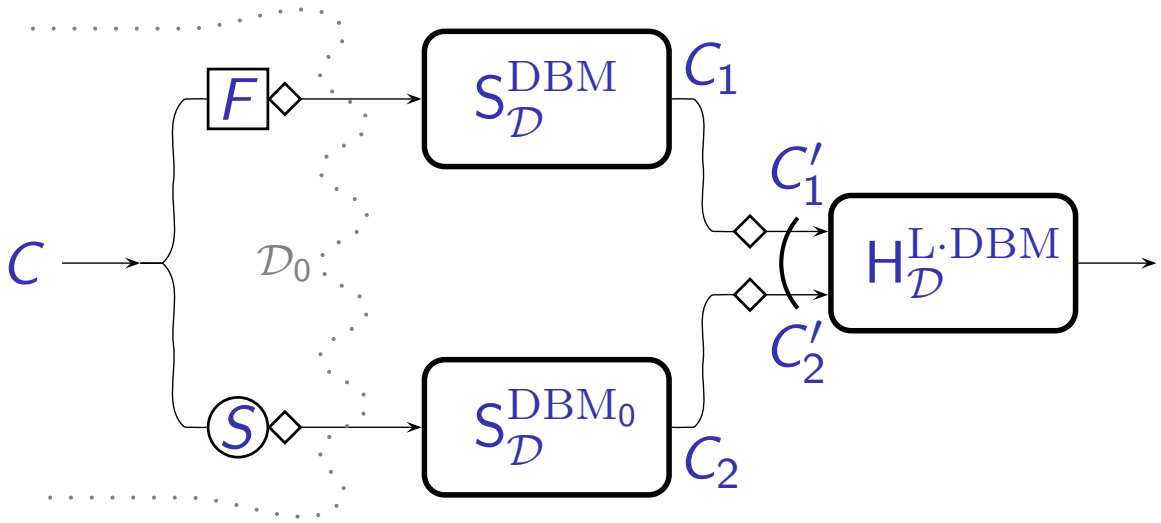
System: Inductors...

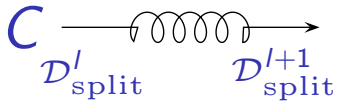
(from Baby Steps and Lateen EM)



System: Inductors...

(from Baby Steps and Lateen EM)



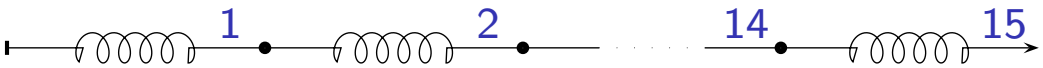


System: Iterating...

(from Baby Steps, Less is More, etc.)

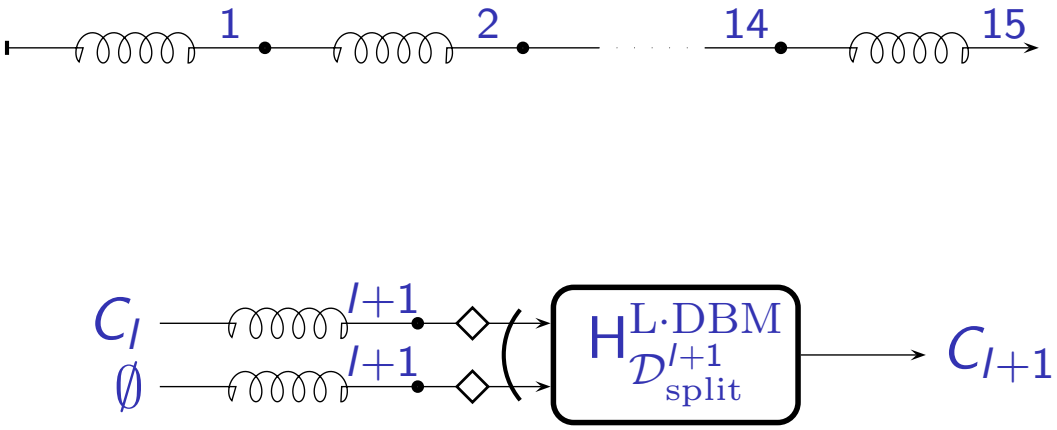
System: Iterating...

(from Baby Steps, Less is More, etc.)



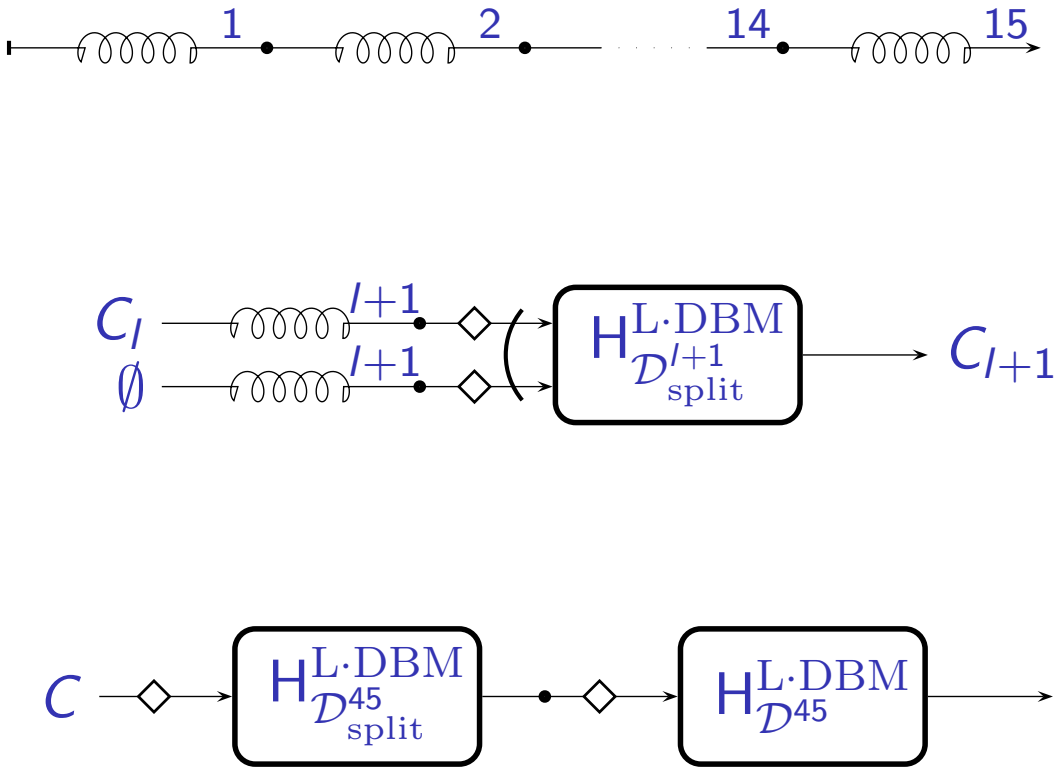
System: Iterating...

(from Baby Steps, Less is More, etc.)



System: Iterating...

(from Baby Steps, Less is More, etc.)



System: Results ... on Section 23 of WSJ

System: Results ... on Section 23 of WSJ

Right-Branching **(Klein and Manning, 2004)** **31.7%**

System: Results ... on Section 23 of WSJ

Right-Branching	**(Klein and Manning, 2004)**	**31.7%**
DMV	**@10**	**34.2%**

System: Results ... on Section 23 of WSJ

Right-Branching	(Klein and Manning, 2004)	31.7%
DMV	@10	34.2%
Baby Steps	@15	39.2%
Baby Steps	@45	39.4%

System: Results ... on Section 23 of WSJ

Right-Branching	(Klein and Manning, 2004)	31.7%
DMV	@10	34.2%
Baby Steps	@15	39.2%
Baby Steps	@45	39.4%
Soft Parameter Tying	(Cohen and Smith, 2009)	42.2%

System: Results ... on Section 23 of WSJ

Right-Branching	(Klein and Manning, 2004)	31.7%
DMV	@10	34.2%
Baby Steps	@15	39.2%
Baby Steps	@45	39.4%
Soft Parameter Tying	(Cohen and Smith, 2009)	42.2%
Less is More	@15	44.1%

System: Results ... on Section 23 of WSJ

Right-Branching	(Klein and Manning, 2004)	31.7%
DMV	@10	34.2%
Baby Steps	@15	39.2%
Baby Steps	@45	39.4%
Soft Parameter Tying	(Cohen and Smith, 2009)	42.2%
Less is More	@15	44.1%
Leapfrog	@45	45.0%

System: Results ... on Section 23 of WSJ

Right-Branching	(Klein and Manning, 2004)	31.7%
DMV	@10	34.2%
Baby Steps	@15	39.2%
Baby Steps	@45	39.4%
Soft Parameter Tying	(Cohen and Smith, 2009)	42.2%
Less is More	@15	44.1%
Leapfrog	@45	45.0%
	(Gimpel and Smith, 2012)	53.1%
	(Gillenwater et al., 2010)	53.3%
	(Bisk and Hockenmaier, 2012)	53.3%
	(Blunsom and Cohn, 2010)	55.7%
	(Tu and Honavar, 2012)	57.0%

System: Results ... on Section 23 of WSJ

Right-Branching	(Klein and Manning, 2004)	31.7%
DMV	@10	34.2%
Baby Steps	@15	39.2%
Baby Steps	@45	39.4%
Soft Parameter Tying	(Cohen and Smith, 2009)	42.2%
Less is More	@15	44.1%
Leapfrog	@45	45.0%
	(Gimpel and Smith, 2012)	53.1%
	(Gillenwater et al., 2010)	53.3%
	(Bisk and Hockenmaier, 2012)	53.3%
	(Blunsom and Cohn, 2010)	55.7%
	(Tu and Honavar, 2012)	57.0%
Integrated System	— no initializers or gold tags —	64.4%

System: Results ... on Section 23 of WSJ

Right-Branching	(Klein and Manning, 2004)	**31.7%**
DMV	@10	**34.2%**
Baby Steps	@15	**39.2%**
Baby Steps	@45	**39.4%**
Soft Parameter Tying	(Cohen and Smith, 2009)	**42.2%**
Less is More	@15	**44.1%**
Leapfrog	@45	**45.0%**
	(Gimpel and Smith, 2012)	**53.1%**
	(Gillenwater et al., 2010)	**53.3%**
	(Bisk and Hockenmaier, 2012)	**53.3%**
	(Blunsom and Cohn, 2010)	**55.7%**
	(Tu and Honavar, 2012)	**57.0%**
Integrated System	— no initializers or gold tags —	**64.4%**
Supervised DBM		*76.3%*

System: Results ... on Section 23 of WSJ

Right-Branching	**(Klein and Manning, 2004)**	**31.7%**
DMV	**@10**	**34.2%**
Baby Steps	**@15**	**39.2%**
Baby Steps	**@45**	**39.4%**
Soft Parameter Tying	**(Cohen and Smith, 2009)**	**42.2%**
Less is More	**@15**	**44.1%**
Leapfrog	**@45**	**45.0%**

During the same time, *supervised (constituency)* **parsing advanced from 91.8F_1 (Petrov, 2010) to 92.4F_1 (Shindo et al., 2012).**

Integrated System	**— no initializers or gold tags —**	**64.4%**
Supervised DBM		*76.3%*

System: Results ... on 23 CoNLL sets

- **Also state-of-the-art for multi-lingual evaluation, across 19 languages from disparate families:**

English	+	**Arabic**	**Greek**
		Basque	**Hungarian**
		Bulgarian	**Italian**
		Catalan	**Japanese**
		Chinese	**Portuguese**
		Czech	**Slovenian**
		Danish	**Spanish**
		Dutch	**Swedish**
		German	**Turkish**

Thanks!

Omri Abend, Eneko Agirre, Hiyan Alshawi, Serafim Batzoglou, John Bauer, Thorsten Brants, Dan Cer, Nate Chambers, Angel Chang, Pi-Chuan Chang, Wanxiang Che, Johnny Chen, Jenny Finkel, Andy Golding, Spence Green, Sonal Gupta, David Hall, Dan Jurafsky, Eisar Lipkovitz, Ting Liu, Chris Manning, Marie de Marneffe, David McClosky, Ryan McDonald, Liz Morin, Andrew Ng, Peter Norvig, Art Owen, Marius Pașca, Fernando Pereira, Slav Petrov, Daniel Pipes, Agnieszka Purves, Daniel Ramage, Marta Recasens, Roi Reichart, Roy Schwartz, Richard Socher, Mihai Surdeanu, Julie Tibshirani, Mengqiu Wang, Eric Yeh,

annoymous reviewers, and many others of Stanford NLP, Fannie & John Hertz Foundation, and Google Inc.

Thanks!

Omri Abend, Eneko Agirre, Hiyan Alshawi, Serafim Batzoglou, John Bauer, Thorsten Brants, Dan Cer, Nate Chambers, Angel Chang, Pi-Chuan Chang, Wanxiang Che, Johnny Chen, Jenny Finkel, Andy Golding, Spence Green, Sonal Gupta, David Hall, Dan Jurafsky, Eisar Lipkovitz, Ting Liu, Chris Manning, Marie de Marneffe, David McClosky, Ryan McDonald, Liz Morin, Andrew Ng, Peter Norvig, Art Owen, Marius Pașca, Fernando Pereira, Slav Petrov, Daniel Pipes, Agnieszka Purves, Daniel Ramage, Marta Recasens, Roi Reichart, Roy Schwartz, Richard Socher, Mihai Surdeanu, Julie Tibshirani, Mengqiu Wang, Eric Yeh,

annoymous reviewers, and many others of Stanford NLP, Fannie & John Hertz Foundation, and Google Inc.

Thanks!

Omri Abend, Eneko Agirre, Hiyan Alshawi, Serafim Batzoglou, John Bauer, Thorsten Brants, Dan Cer, Nate Chambers, Angel Chang, Pi-Chuan Chang, Wanxiang Che, Johnny Chen, Jenny Finkel, Andy Golding, Spence Green, Sonal Gupta, David Hall, Dan Jurafsky, Eisar Lipkovitz, Ting Liu, Chris Manning, Marie de Marneffe, David McClosky, Ryan McDonald, Liz Morin, Andrew Ng, Peter Norvig, Art Owen, Marius Pașca, Fernando Pereira, Slav Petrov, Daniel Pipes, Agnieszka Purves, Daniel Ramage, Marta Recasens, Roi Reichart, Roy Schwartz, Richard Socher, Mihai Surdeanu, Julie Tibshirani, Mengqiu Wang, Eric Yeh,

annoymous reviewers, and many others of Stanford NLP, Fannie & John Hertz Foundation, and Google Inc.

Thanks!

Omri Abend, Eneko Agirre, Hiyan Alshawi, Serafim Batzoglou, John Bauer, Thorsten Brants, Dan Cer, Nate Chambers, Angel Chang, Pi-Chuan Chang, Wanxiang Che, Johnny Chen, Jenny Finkel, Andy Golding, Spence Green, Sonal Gupta, David Hall, Dan Jurafsky, Eisar Lipkovitz, Ting Liu, Chris Manning, Marie de Marneffe, David McClosky, Ryan McDonald, Liz Morin, Andrew Ng, Peter Norvig, Art Owen, Marius Pașca, Fernando Pereira, Slav Petrov, Daniel Pipes, Agnieszka Purves, Daniel Ramage, Marta Recasens, Roi Reichart, Roy Schwartz, Richard Socher, Mihai Surdeanu, Julie Tibshirani, Mengqiu Wang, Eric Yeh,

annoymous reviewers, and many others of Stanford NLP, Fannie & John Hertz Foundation, and Google Inc.

Thanks!

Omri Abend, Eneko Agirre, Hiyan Alshawi, Serafim Batzoglou, John Bauer, Thorsten Brants, Dan Cer, Nate Chambers, Angel Chang, Pi-Chuan Chang, Wanxiang Che, Johnny Chen, Jenny Finkel, Andy Golding, Spence Green, Sonal Gupta, David Hall, Dan Jurafsky, Eisar Lipkovitz, Ting Liu, Chris Manning, Marie de Marneffe, David McClosky, Ryan McDonald, Liz Morin, Andrew Ng, Peter Norvig, Art Owen, Marius Pașca, Fernando Pereira, Slav Petrov, Daniel Pipes, Agnieszka Purves, Daniel Ramage, Marta Recasens, Roi Reichart, Roy Schwartz, Richard Socher, Mihai Surdeanu, Julie Tibshirani, Mengqiu Wang, Eric Yeh,

annoymous reviewers, and many others of Stanford NLP, Fannie & John Hertz Foundation, and Google Inc.

Dramatization: http://www.youtube.com/embed/ncFCdCjBqcE?start=4&end=66.5

"... the real challenge is to make simple things look beautiful."

— Glenn Corteza, Tanguero Argentino

Thanks again!

Questions?