

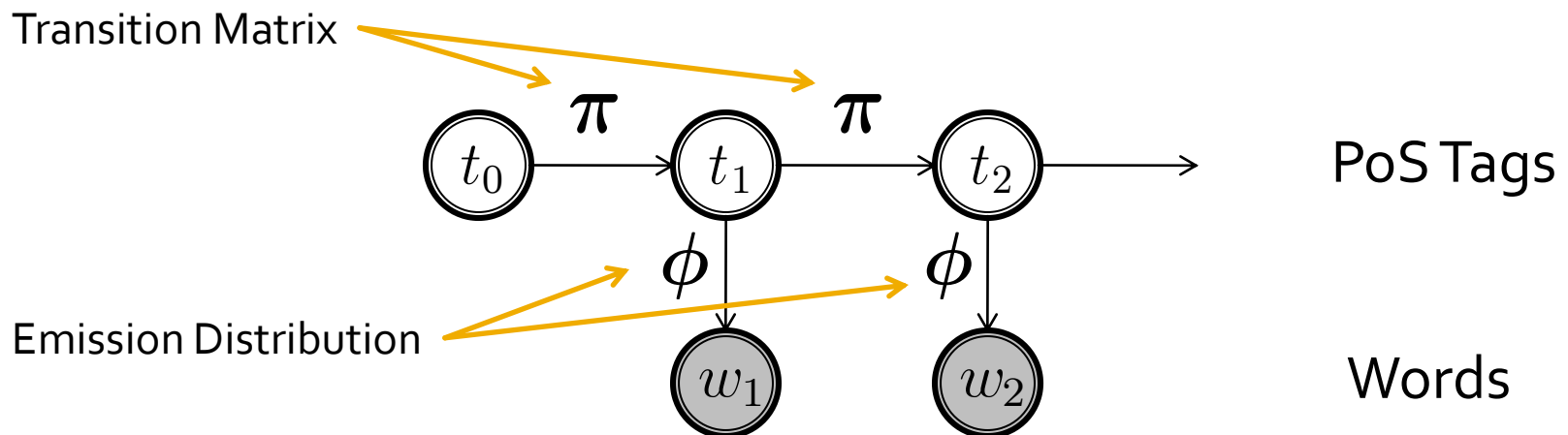
Jurgen Van Gael, Andreas Vlachos, Zoubin Ghahramani – University of Cambridge

# The Infinite HMM for POS Tagging

# Part of Speech Tagging

The	representative	put	chairs	on	the	table.
AT	NN	VBD	NNS	IN	AT	NN

- Goal: given text produce part-of-speech tags
- Approach: 1<sup>ST</sup> order Hidden Markov Model



# Big Picture

## Supervised

- Choose particular tag set (many choices)
- Data = Words + Gold PoS tags
- [Church 1988, Kupiec 1992, Merialdo 1994,...]

## Unsupervised

- Choose a particular HMM size
- Data = Words
- [..., Johnson 2007, Gao & Johnson 2008]

## Semi-Supervised

- [Goldwater & Griffiths 2007, ...]
- Future work ...

Introduction

**Questions**

Infinite HMM

Evaluation

Conclusion

# Question 1: Number of States?

On Wall Street Journal of Penn Treebank

- Johnson 2007 uses 50 – 40 – 25 – 10

Method	Variation of Information
EM with 50 states	4.46
EM with 40 states	4.37
EM with 25 states	4.23
EM with 10 states	4.32

- Gao & Johnson use 50 – 17
- Existing tag set sizes such as 45 – 61 – 87

***Question 1: Let data decide number of states?***

# Question 2: How to use States?

Toy sample from unsupervised HMM:

	NN	DT	VB
HMM State 1	2	0	0
HMM State 2	2	1	0
HMM State 3	0	1	4

- Think of HMM states as clusters ...
- Can we identify HMM clusters with PoS tags?
  - No, unless we make simplifications
  - Evaluate using clustering measures.
  - [Johnson, Gao & Johnson, Goldwater & Griffiths]

***Question 2: Can we use HMM clusters as features?***

Introduction

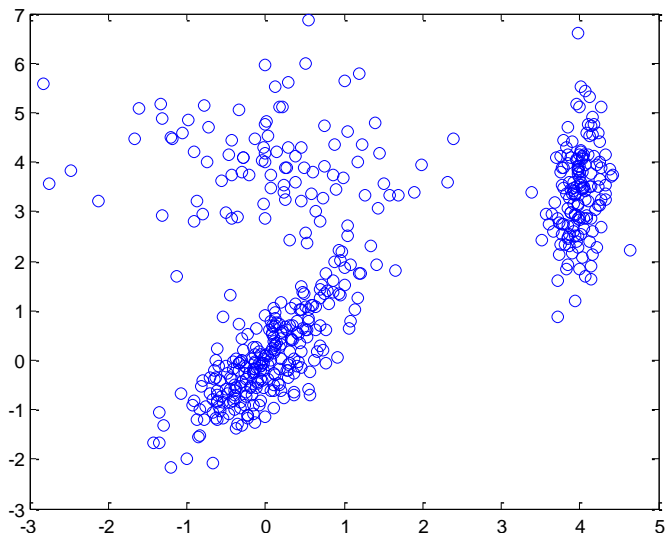
Questions

**Infinite HMM**

Evaluation

Conclusion

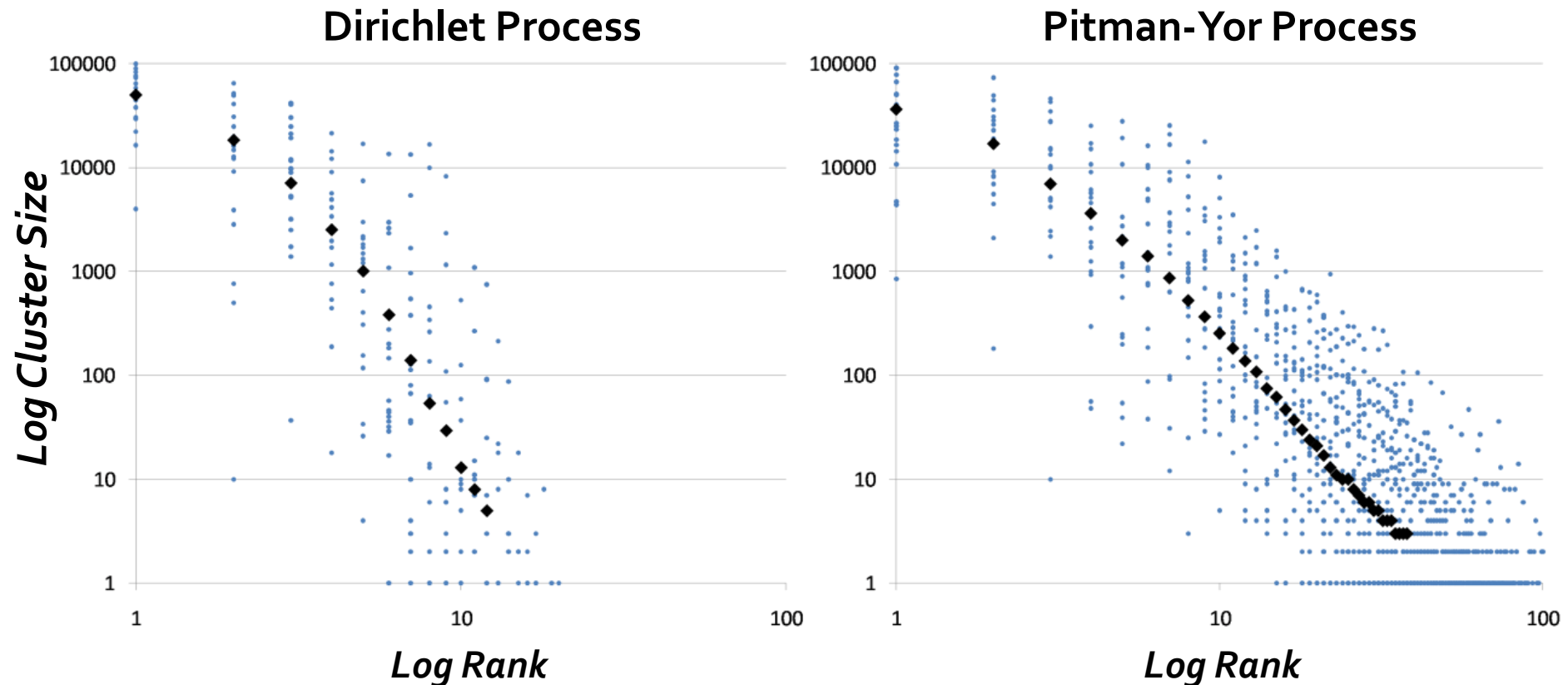
# Nonparametric Bayes



- How many clusters on the left?
- Parametric Model + Model Selection
- Nonparametric Bayes
  - Unbounded pool of clusters
  - Datapoint can belong to
    - Existing cluster
    - New cluster
  - One learning algorithm

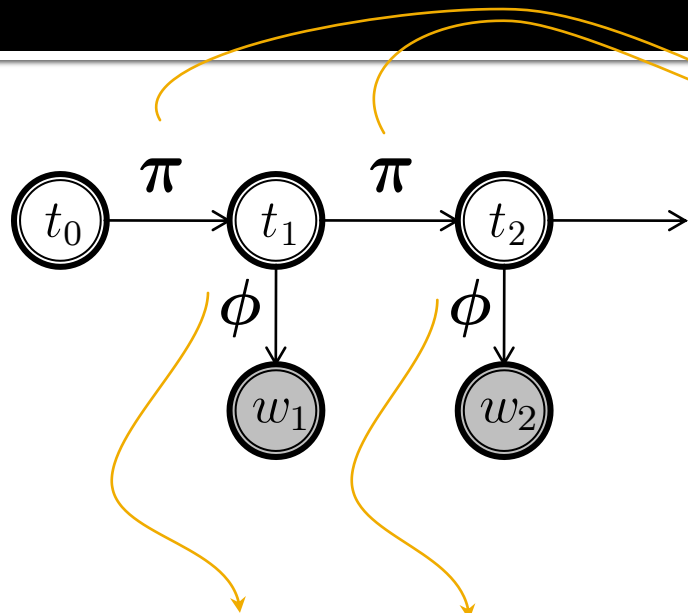
- Example: Language Modelling, [Teh 2006, Goldwater & Griffiths 2006], Kneser-Ney is special case of a NPBayes model
- ***Answer 1: use NPBayes for unsupervised PoS Tagging.***

# Nonparametric Bayes Priors

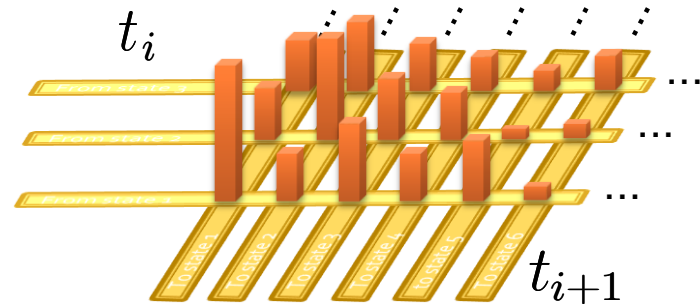


Pitman-Yor Process follows Zipf's Law

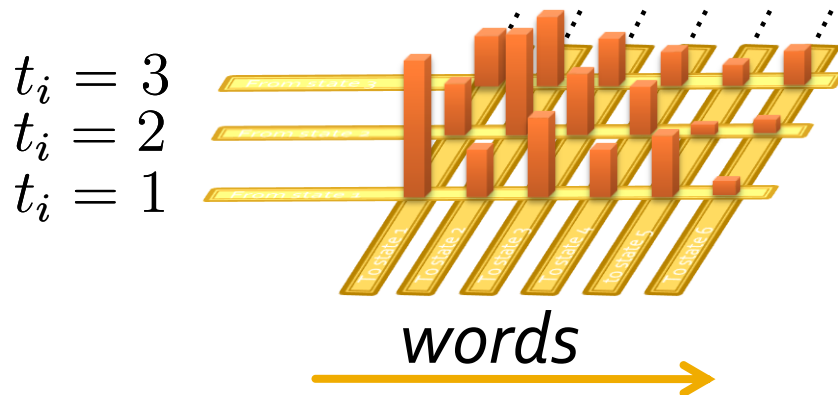
# The Infinite HMM [Beal et al. 2002]



Infinitely Large Transition Matrix



Infinitely Many Output Distributions



- Similar to Parametric HMM
- Two hyper parameters:
  - $\alpha$  controls sparsity of transition matrix
  - $\gamma$  controls the a priori expected number of states
- Can use DP or PY prior for transitions

# iHMM Inference (or Learning)

- Inference is the problem of computing posterior distributions
  - Generate samples
- Like in HMM we want to use dynamic program
  - Beam Sampling [Van Gael et al. 2008]
- In this work: parallelization using Map-Reduce

Introduction

Questions

Infinite HMM

**Evaluation**

Conclusion

# Experimental Setup





- Wall Street Journal (1.000.000 tokens)
- Whole corpus for learning
- 3 iHMM runs
  - DP-fixed: Dirichlet Process with fixed hypers
  - DP-learnt( $\alpha, \gamma$ ): Dirichlet Process with learned hypers
  - PY-fixed: Pitman-Yor Process with fixed hypers

# Prototypical iHMM Samples

State (count)	1 (115051)	2 (108665)	3 (105359)	8 (37250)	18 (21436)
	of (30840)	market (1321)	the (56250)	is (8944)	oil(374)
	in (20328)	quarter (1053)	a (21989)	was (4989)	computer (342)
	for (10610)	trading (754)	its (4883)	are (4854)	futures (271)
	to (9917)	bonds (731)	an (3700)	has (4301)	securities (242)
	on (6558)	business (645)	their (2300)	were (2659)	business (234)

# Evaluation I: Clustering Measures

	NN	DT	VB
HMM State 1	2	0	0
HMM State 2	2	1	0
HMM State 3	0	1	4

- Two desirable properties for clustering measure
  - *Homogeneity*: tokens assigned to the same state have the same PoS tag –  $H(Y|T)$
  - *Completeness*: tokens with the same PoS tag are assigned to the same state –  $H(T|Y)$
- More states:  completeness,  homogeneity
- Fewer states:  completeness,  homogeneity

# Evaluation I: Clustering Measures

Variation of Information = Homogeneity + Completeness

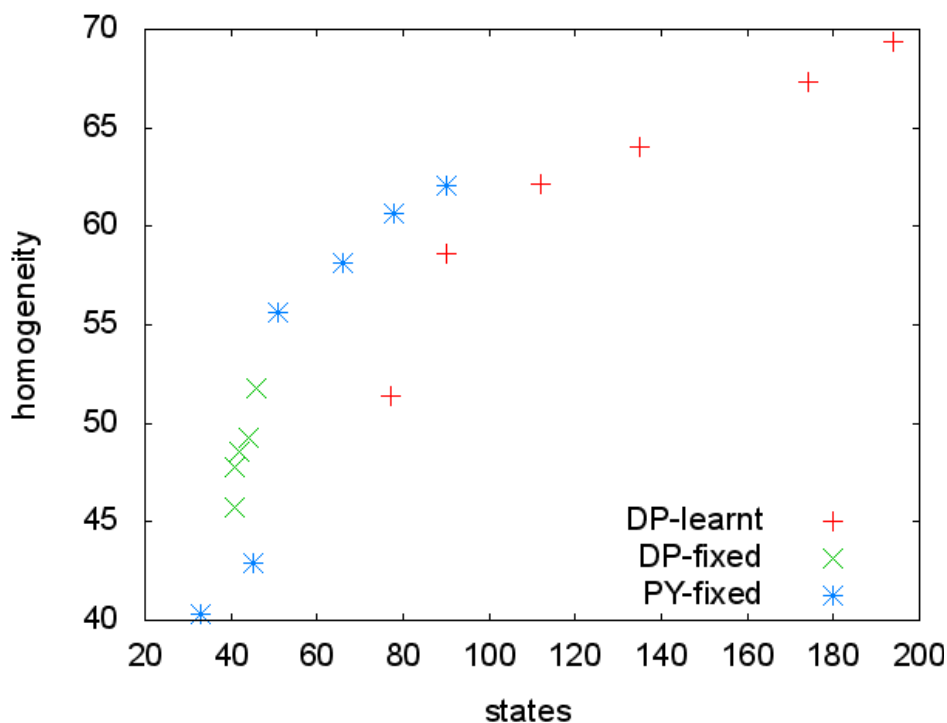
$$VI(Y,T) = H(Y|T) + H(T|Y)$$

Method	Variation of Information
EM (50) [G&J]	4.47555
VB (50) [G&J]	4.27911
GS (50) [G&J]	4.03886
iHMM (DP-F)	3.93
iHMM (DP-L)	4.32
iHMM (PY)	3.73

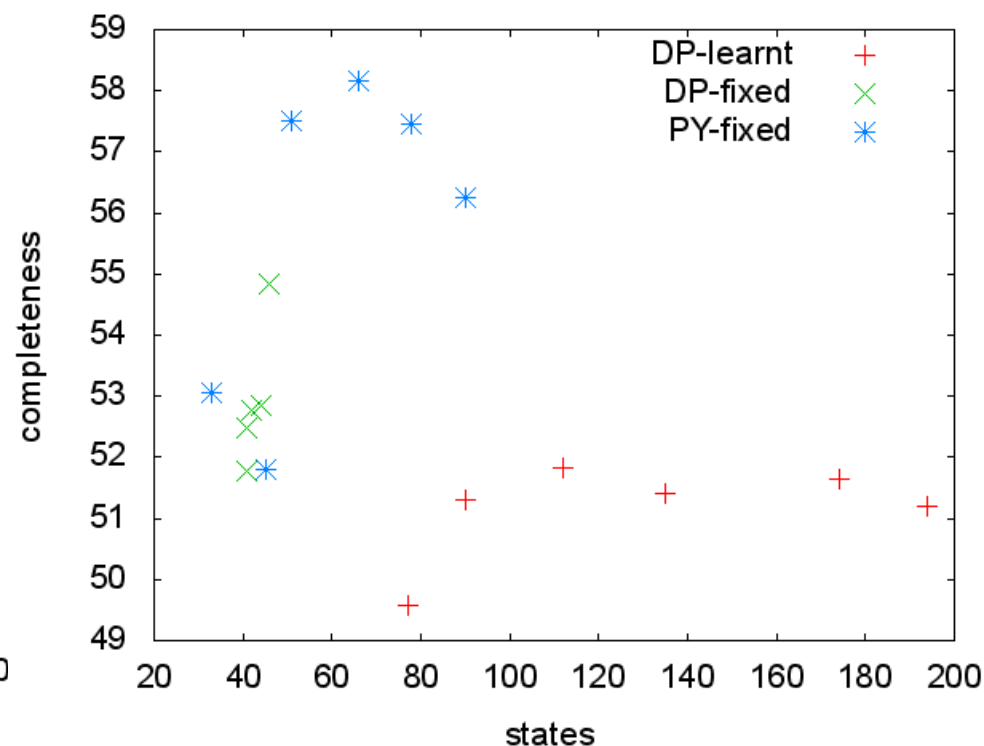
Variation of Information is comparable with Gao & Johnson

# Evaluation I: Clustering Measures

## Homogeneity wants a lot of states



## Completeness wants fewer states



*Maybe use states in other ways than interpreting them as PoS tags?*

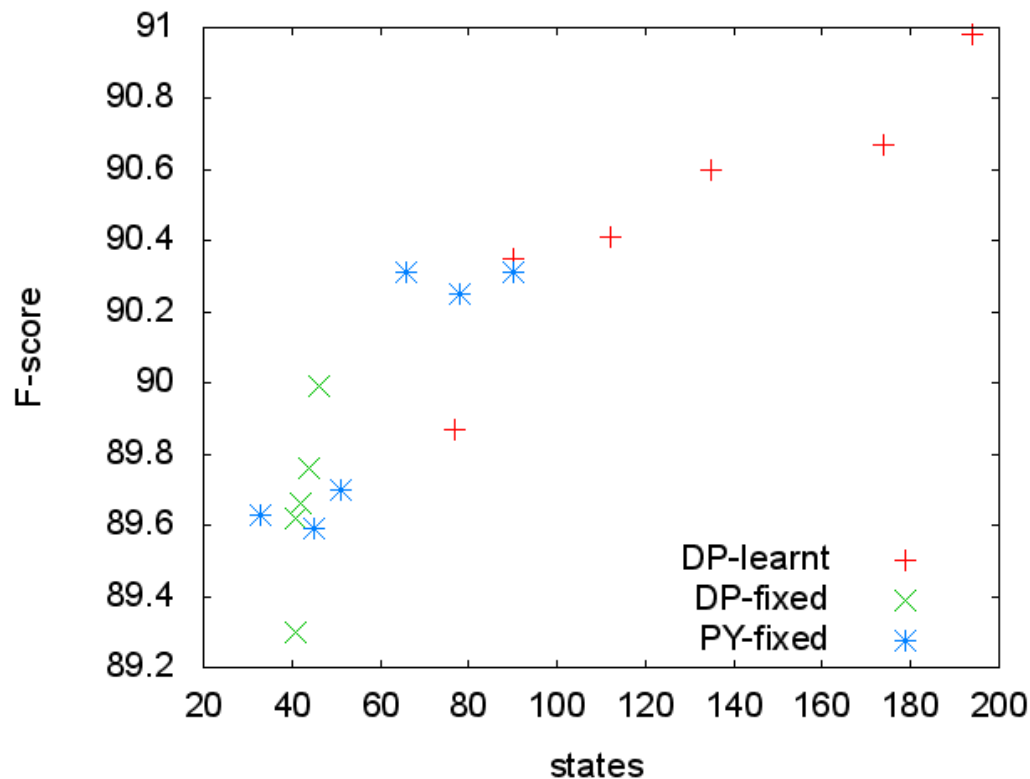
# Evaluation II: Shallow parsing

Rockwell	said	the	agreement	calls	for	it	to	supply
NNP	VBD	DT	NN	VBZ	IN	PRP	TO	VB
B-NP	B-VP	B-NP	I-NP	B-VP	B-SBAR	B-NP	B-VP	I-VP

- Can we replace the PoS tags with iHMM states as features?
- In shallow parsing we identify constituents but not their structure/role.
- CoNLL 2000 shared task used the (supervised) Brill tagger for PoS
- Sections 15-18 and 20 (CoNLL shared task training and test data) for extrinsic evaluation with CRF tagger

***Answer 2: Use CRF for extrinsic evaluation.***

# Results: Shallow Parsing



Comparison with baseline	
CRF Data	F-Score
Words + Brill	93.81
Words + iHMM	90.98
Words	88.58

DP-learnt provides the best features for shallow parsing

Introduction

Questions

Infinite HMM

Evaluation

**Conclusion**

# Conclusion + Future Work

We show that:

- the PY distribution can be integrated into the iHMM framework
- we can parallelize learning of the iHMM using Map/Reduce
- the iHMM automatically discovers putative states and their number
- adding iHMM states improves shallow parsing

Future work:

- Can we combine labeled and unlabeled data?
- Can we use unlabeled datasets 1 or 2 orders of magnitude larger?
- Semi-Supervised Nonparametric Part-of-Speech tagging using Wikipedia & Penn Treebank parallelized via Hadoop