

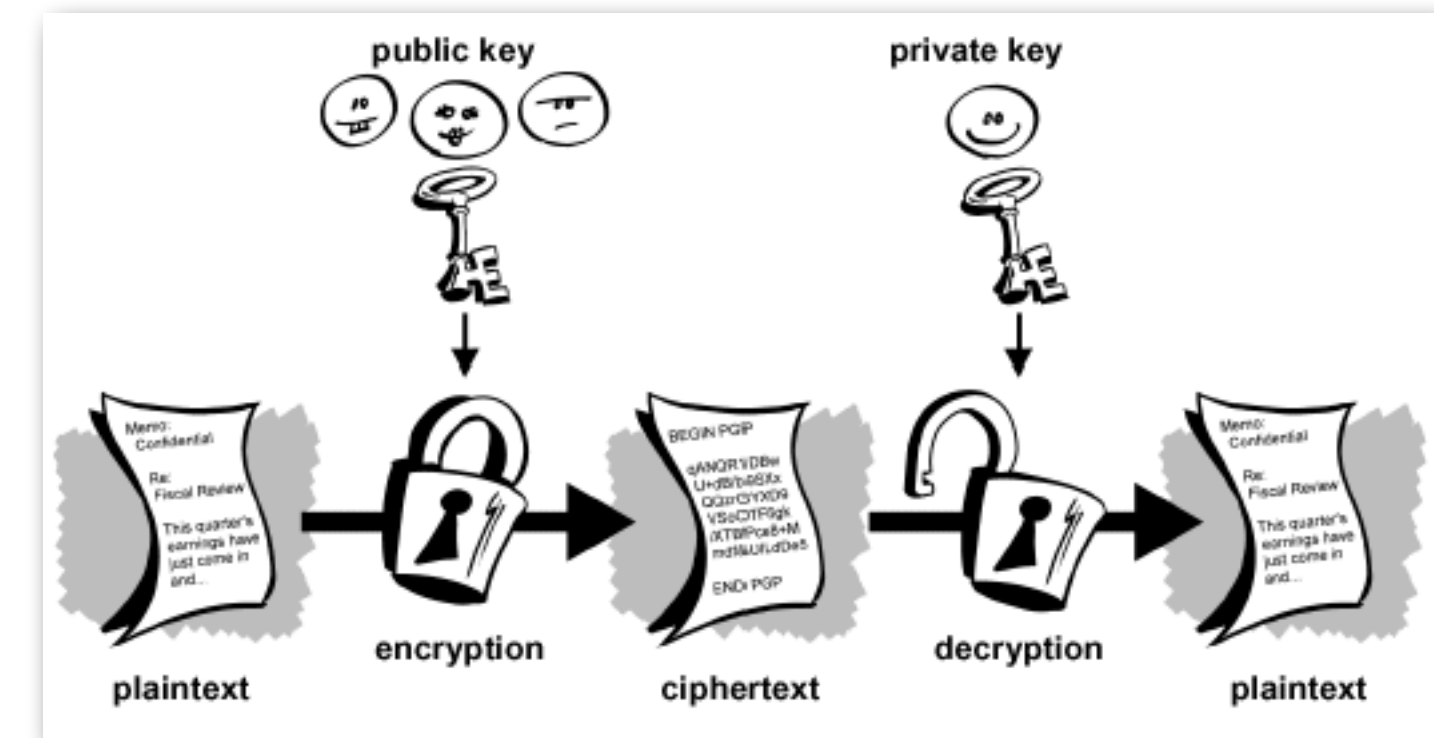


# Get the data!

- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV

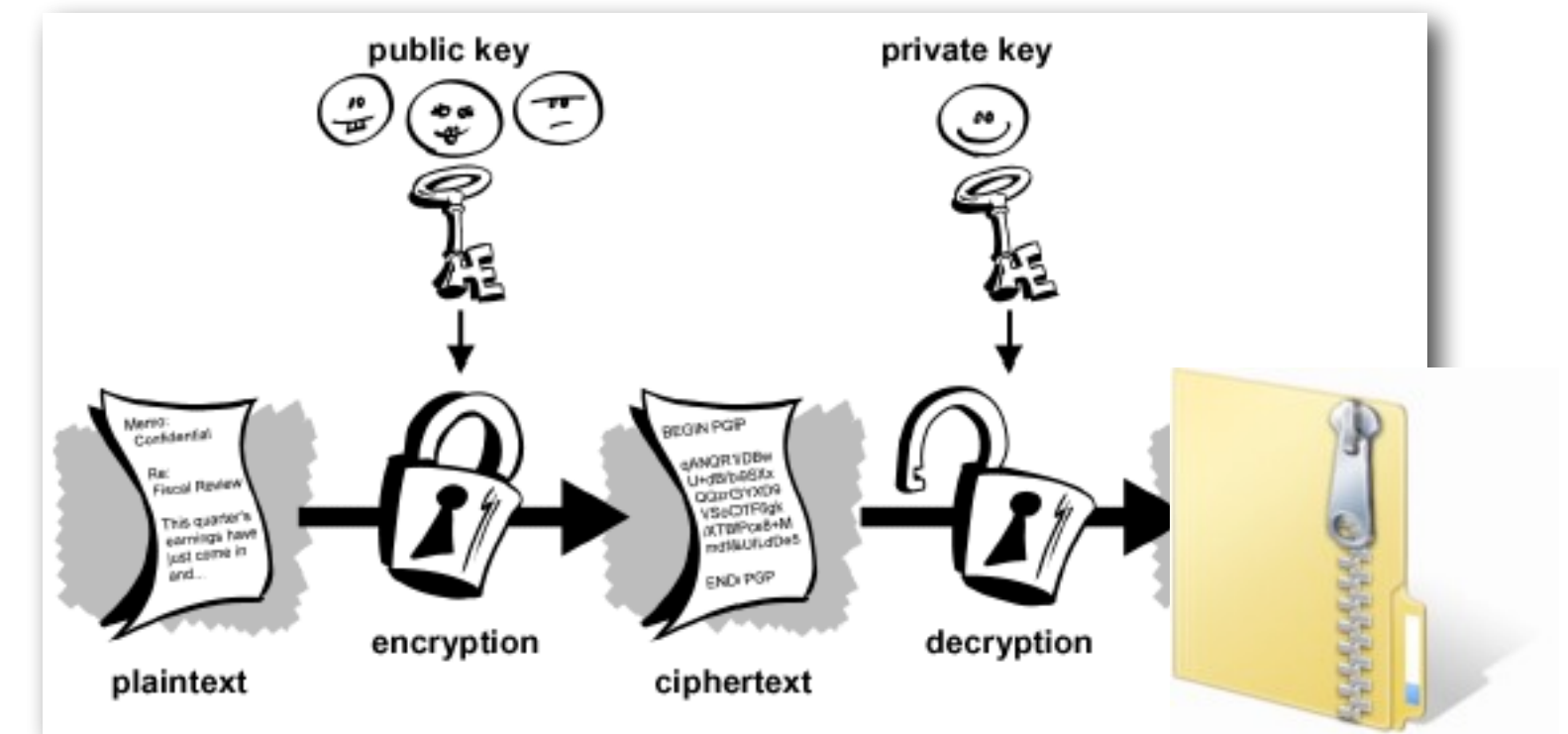
# Get the data!

- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV



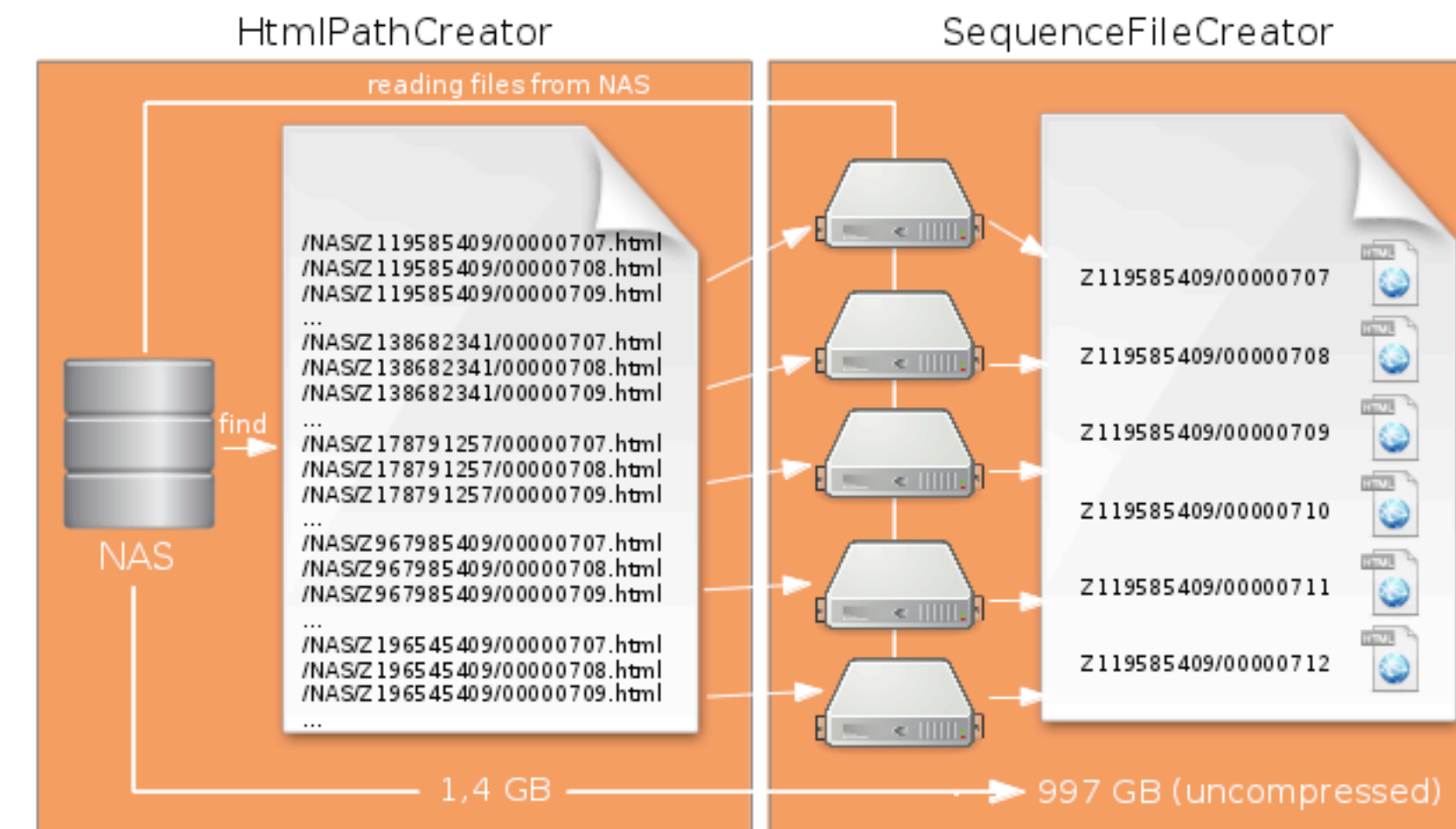
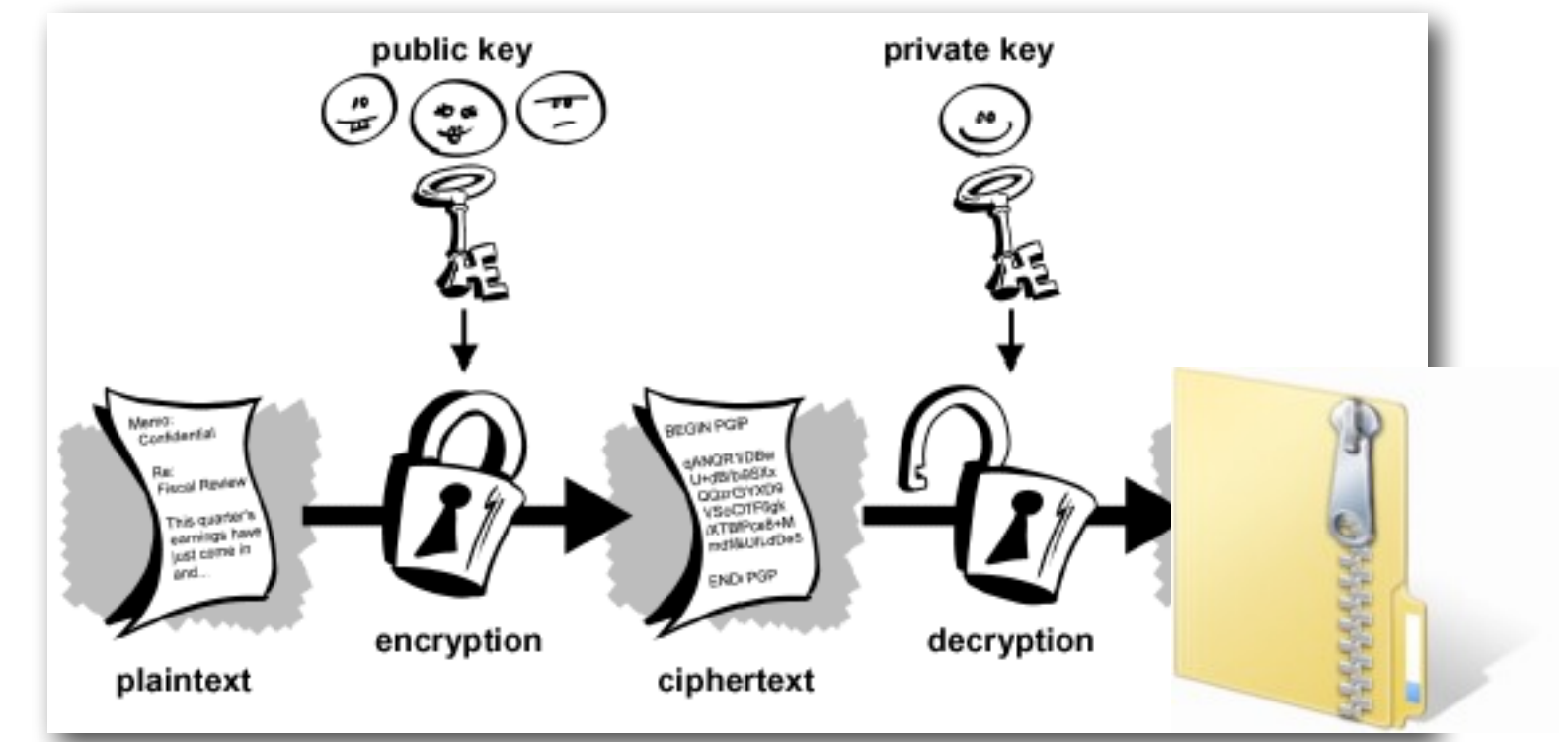
# Get the data!

- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV



# Get the data!

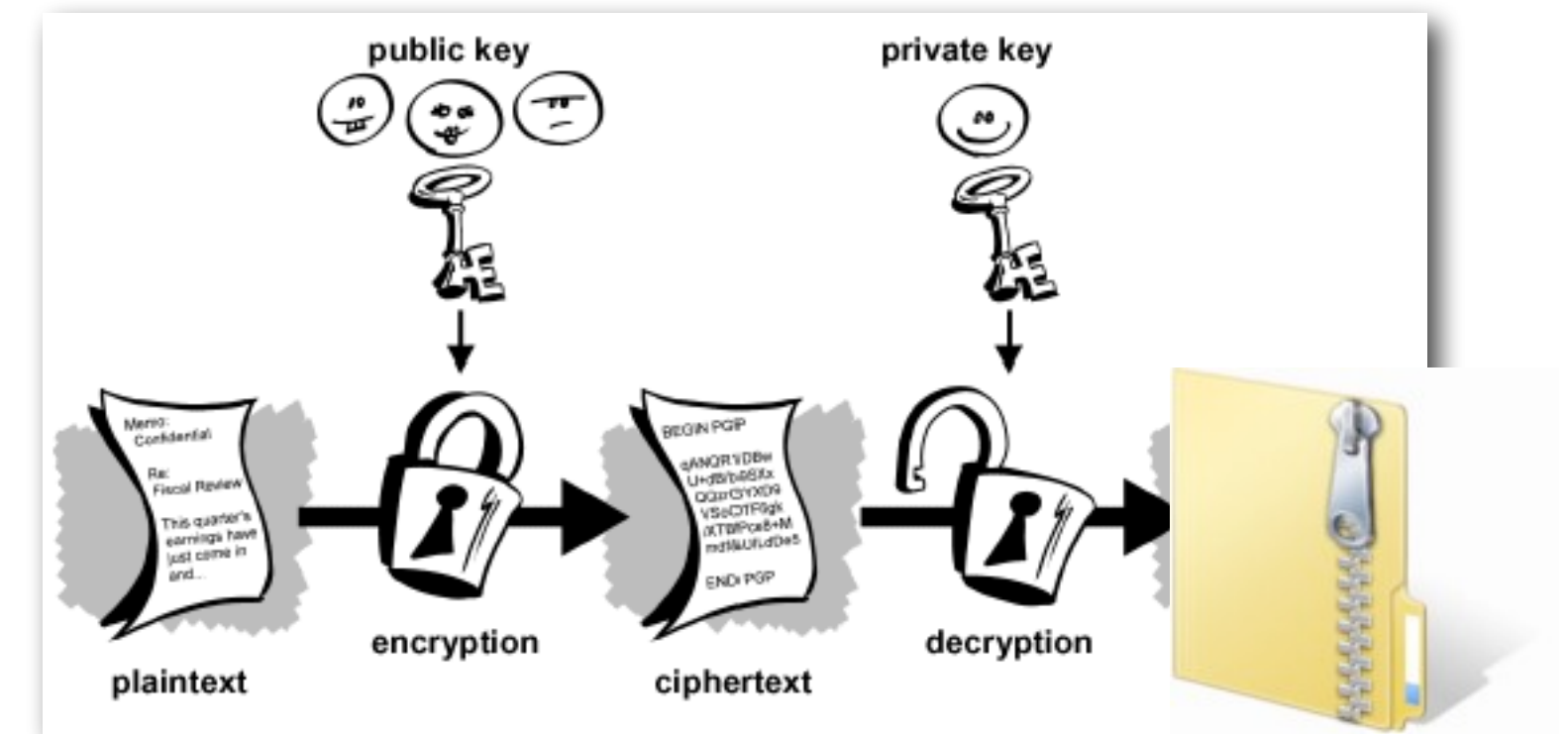
- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV





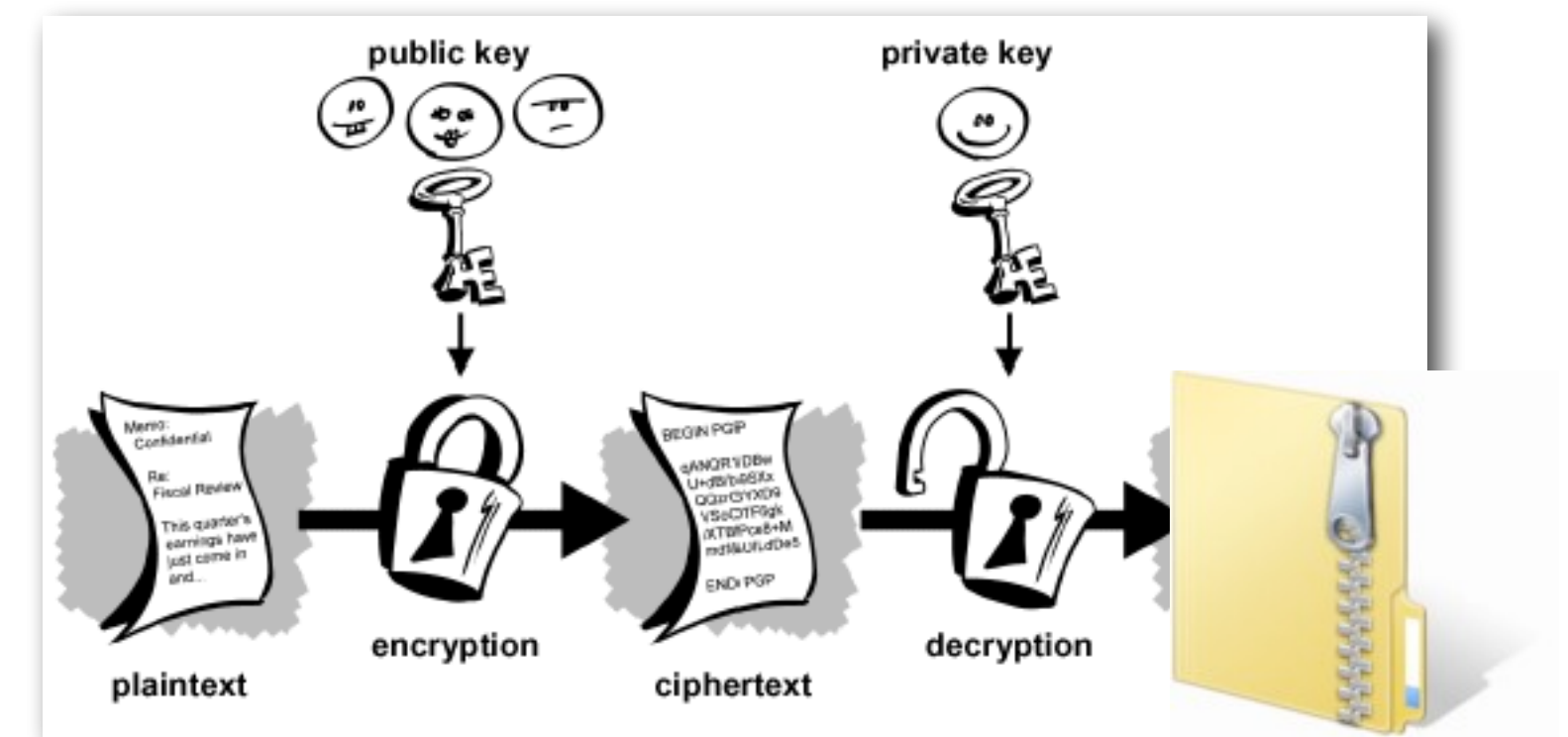
# Get the data!

- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV



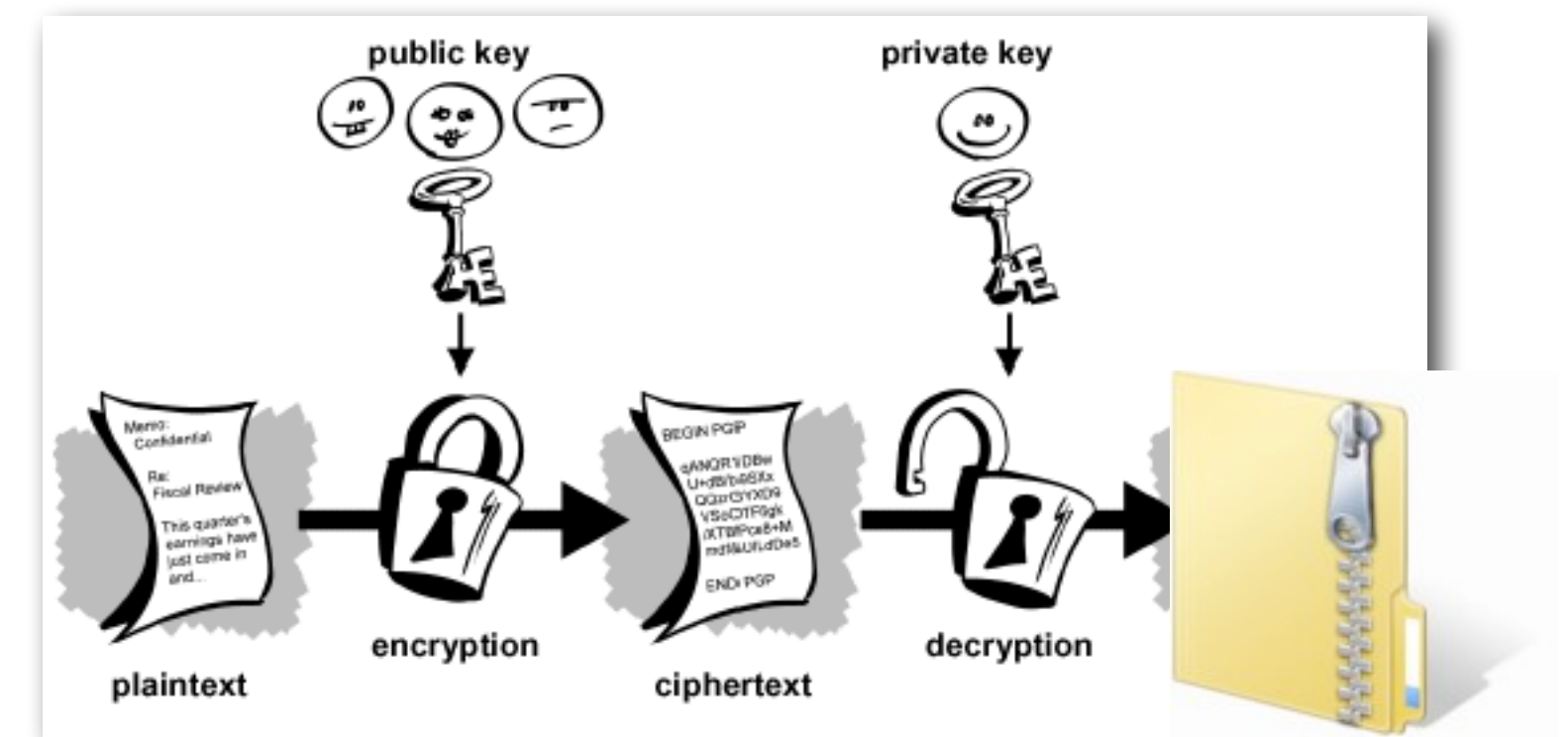
# Get the data!

- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV



# Get the data!

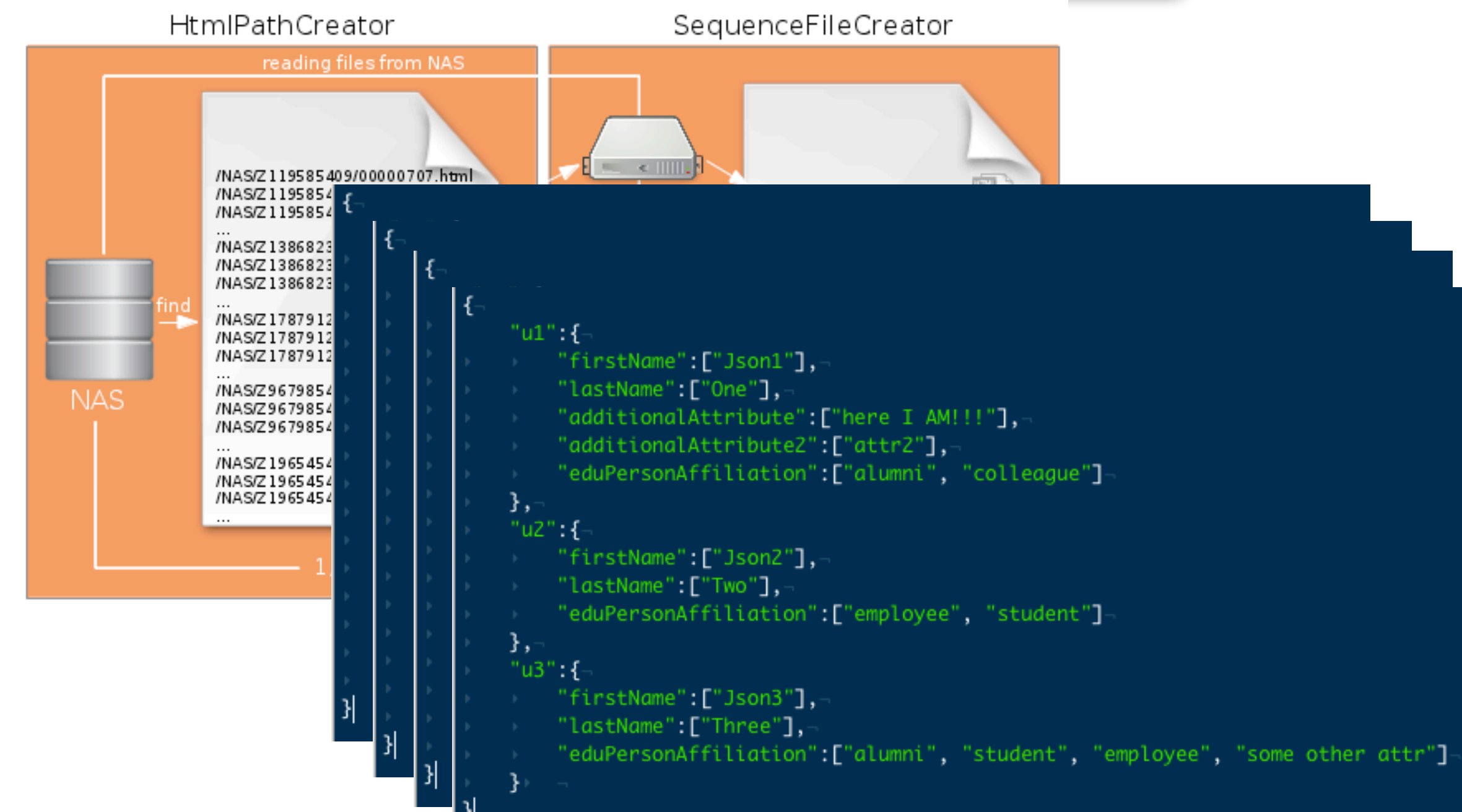
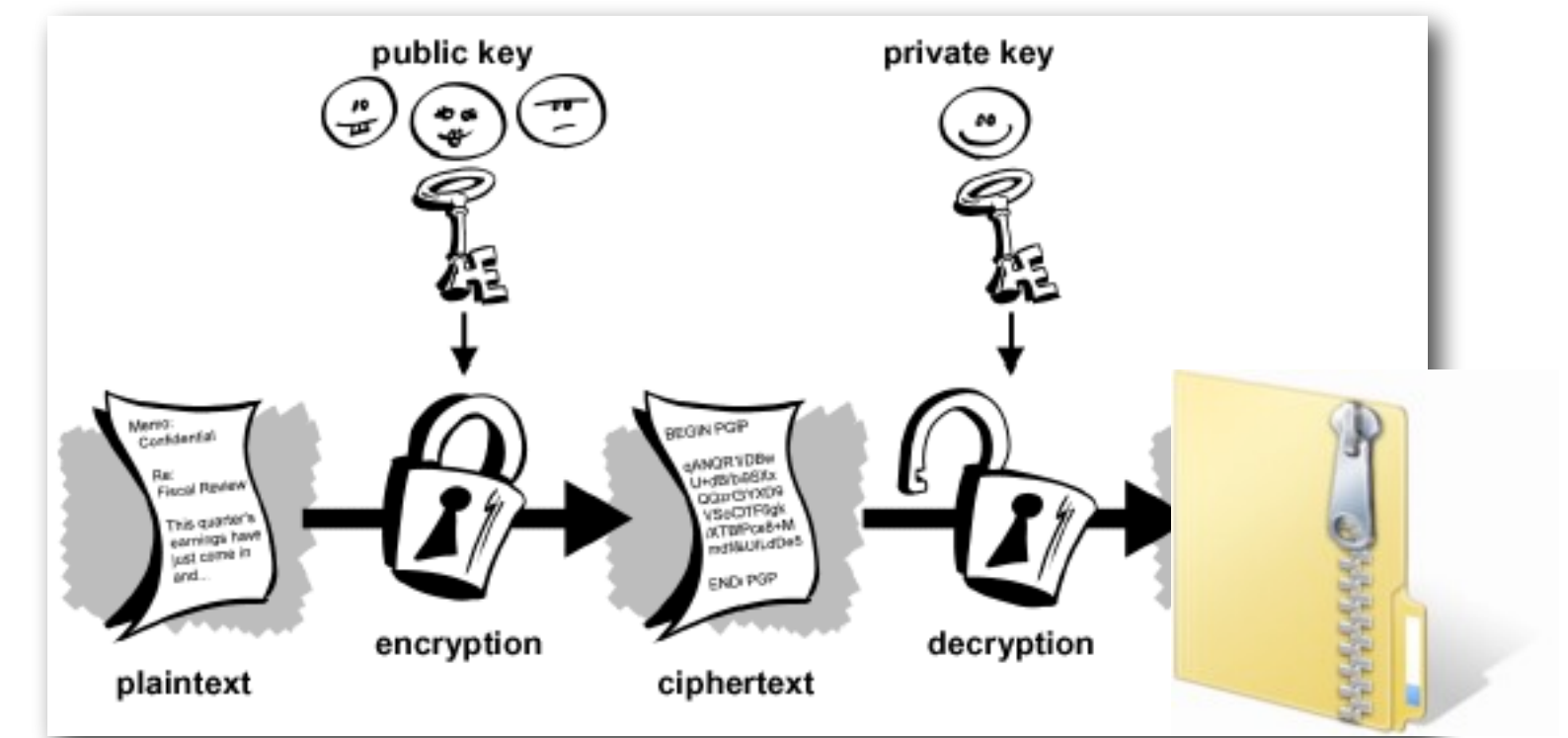
- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV





# Get the data!

- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV







# Sentence Segmentation

- Split a textual document into sentences.
- Was that an abbreviation?
- Was that inside a quote?

```
She stopped.  She said, "Hello there," and then went on.
^              ^
He's vanished!  What will we do?  It's up to us.
^              ^              ^              ^
Please add 1.5 liters to the tank.
^
```

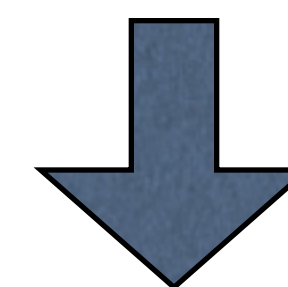
```
sentence =
tokenize.sent_tokenize(text)
```



# Word Tokenization

In Düsseldorf I took my hat off. But I can't put it back on.

- Split a sentence into tokens
- `text.split(" ")` is not always enough
- What about apostrophe, abbreviations, misspellings, URIs, different languages?



In Düsseldorf I took my hat off.

But I can't put it back on.



# Part-of-Speech Tagging (POS)

- Classifying word tokens into parts of speech





# Part-of-Speech Tagging (POS)

- Classifying word tokens into parts of speech

The Penn Treebank POS tagset.

1. CC	Coordinating conjunction	25. TO	<i>to</i>
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential <i>there</i>	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	<i>wh</i> -determiner
10. LS	List item marker	34. WP	<i>wh</i> -pronoun
11. MD	Modal	35. WP\$	Possessive <i>wh</i> -pronoun
12. NN	Noun, singular or mass	36. WRB	<i>wh</i> -adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. \$	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP\$	Possessive pronoun	43. )	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. '	Left open single quote
22. RBS	Adverb, superlative	46. "	Left open double quote
23. RP	Particle	47. '	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. "	Right close double quote

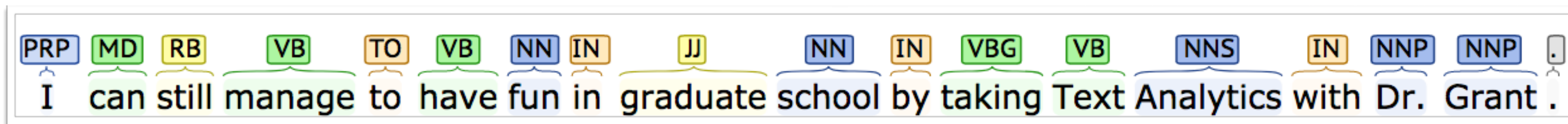


# Part-of-Speech Tagging (POS)

- Classifying word tokens into parts of speech

The Penn Treebank POS tagset.

1. CC	Coordinating conjunction	25. TO	to
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential <i>there</i>	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	<i>wh</i> -determiner
10. LS	List item marker	34. WP	<i>wh</i> -pronoun
11. MD	Modal	35. WP\$	Possessive <i>wh</i> -pronoun
12. NN	Noun, singular or mass	36. WRB	<i>wh</i> -adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. \$	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP\$	Possessive pronoun	43. )	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. '	Left open single quote
22. RBS	Adverb, superlative	46. "	Left open double quote
23. RP	Particle	47. '	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. "	Right close double quote





# Named Entity Recognition (NER)

- Identify the tokens in a sentence that correspond to a Entity.

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes seas
Geo-Political	GPE	Countries states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles



# Named Entity Recognition (NER)

- Identify the tokens in a sentence that correspond to a Entity.

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes seas
Geo-Political	GPE	Countries states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles

I can still manage to have fun in graduate school by taking Text Analytics with Dr. PERSON Grant .

# Named Entity Recognition (NER)

- Identify the tokens in a sentence that correspond to a Entity.

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes seas
Geo Political	GPE	Countries states, provinces, counties
		Bridges, buildings, airports
		Planes, trains, and automobiles

President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

I can still manage to have fun in graduate school by taking Text Analytics with Dr. Grant .

# Chunking

- Identify sequences of non-overlapping labels



# Chunking

- Identify sequences of non-overlapping labels

Destacados representantes del **ORG** Parlamento y la prensa rusos criticaron hoy el "belicismo" que ha definido como posible blanco de su lucha antiterrorista.

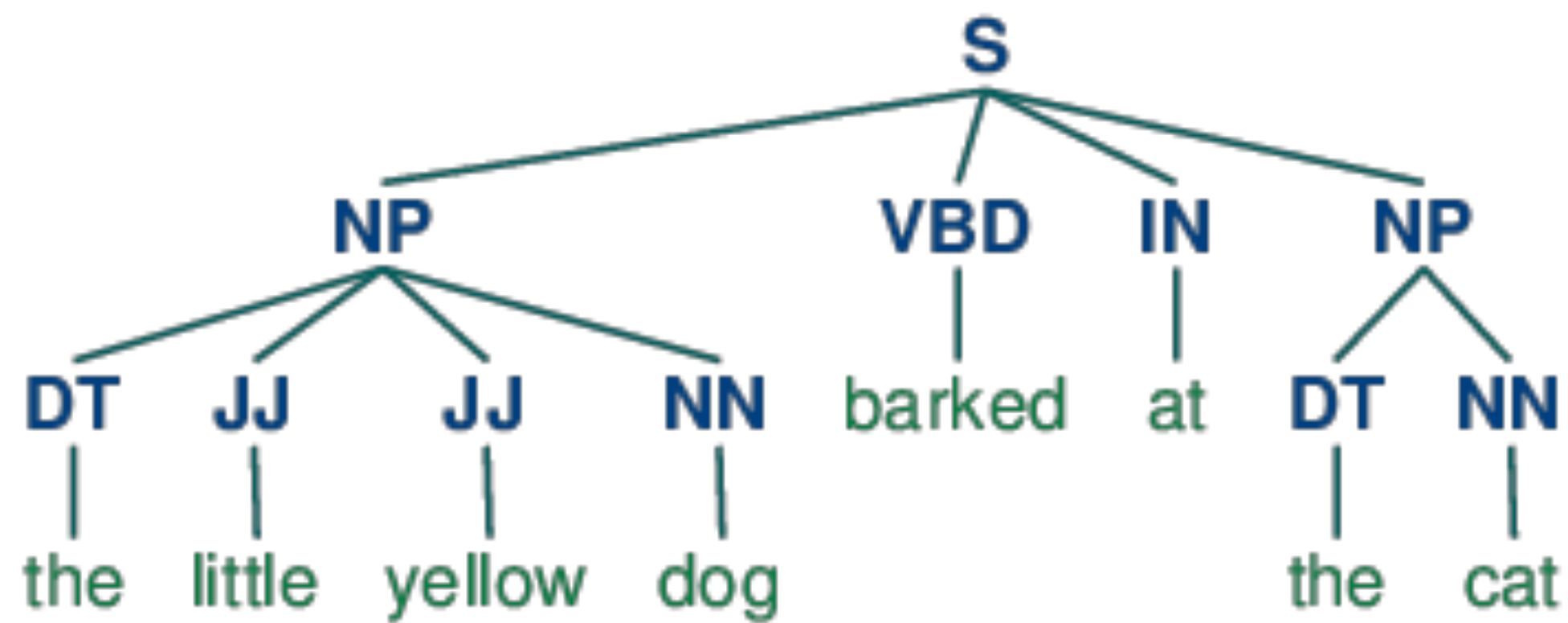
El presidente de la Duma (cámara baja), **ORG** Guennadi Selezniiov, calificó de "claramente aporofóbico" el discurso del **PER** presidente ruso Vladimir Putin.

del Kremlin para Chechenia, **LOC** Serguéi Yastrzhembski.

El asesor presidencial dijo que **LOC** Rusia puede lanzar un ataque preventivo contra los campos de entrenamiento de los terroristas.

# Chunking

- Identify sequences of non-overlapping labels



Destacados representantes del **ORG** Parlamento y la prensa rusos criticaron hoy el "belicismo" ha definido como posible blanco de su lucha antiterrorista.

El presidente de la Duma (cámara baja), **ORG** Guennadi Seleznirov, **PER** calificó de "claramente ap

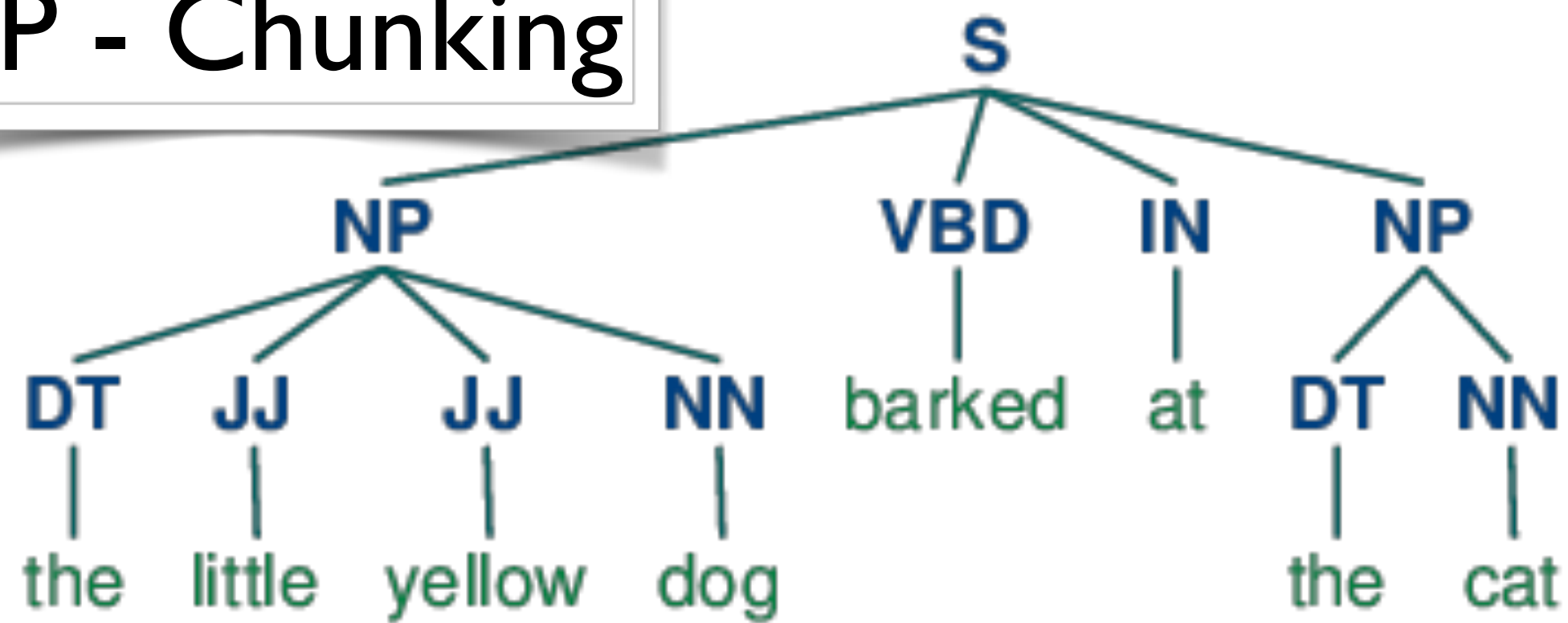
del **ORG** Kremlin para **LOC** Chechenia, **PER** Serguéi Yastrzhembski.

El asesor presidencial dijo que **LOC** Rusia puede lanzar un ataque preventivo contra los camp

# Chunking

- Identify sequences of non-overlapping labels

## NP - Chunking



Destacados representantes del **ORG** Parlamento y la prensa rusos criticaron hoy el "belicismo" ha definido como posible blanco de su lucha antiterrorista.

El presidente de la Duma (cámara baja), **ORG** Guennadi Selezniiov, **PER** calificó de "claramente ap

del **ORG** Kremlin para **LOC** Chechenia, **PER** Serguéi Yastrzhembski.

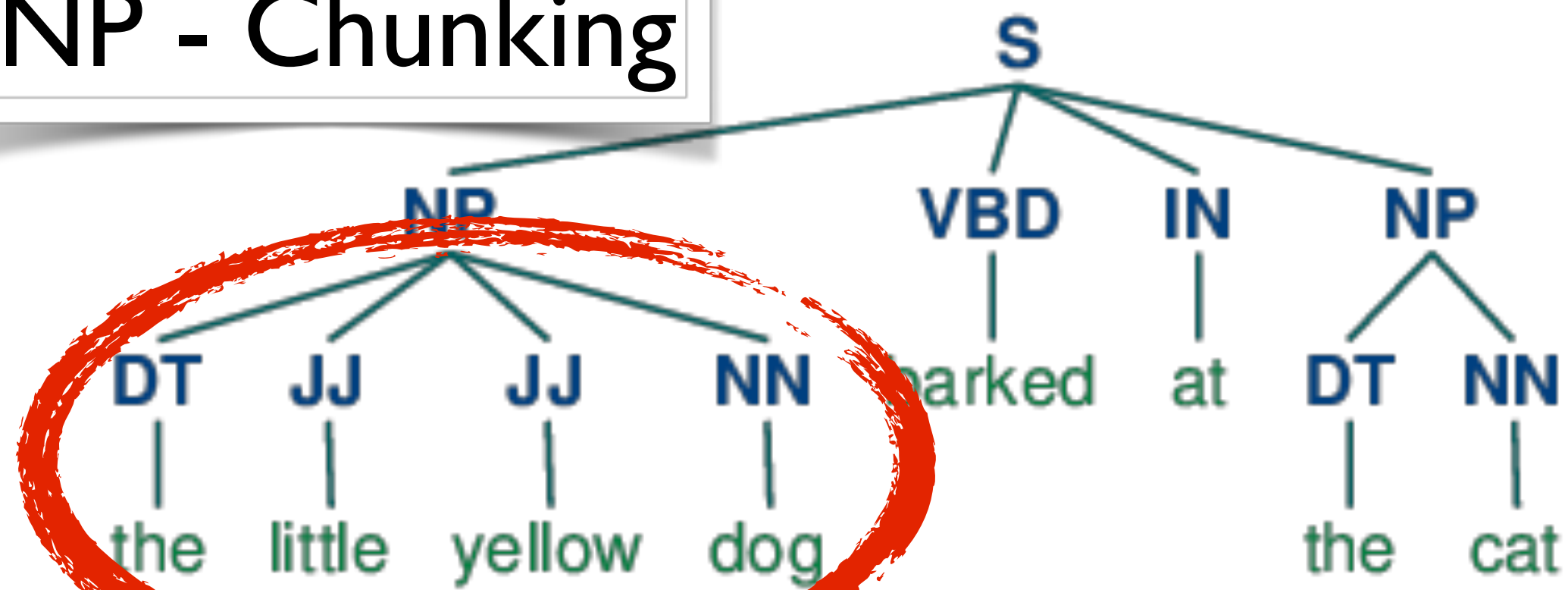
El asesor presidencial dijo que **LOC** Rusia puede lanzar un ataque preventivo contra los camp



# Chunking

- Identify sequences of non-overlapping labels

## NP - Chunking



Destacados representantes del **ORG** Parlamento y la prensa rusos criticaron hoy el "belicismo" ha definido como posible blanco de su lucha antiterrorista.

El presidente de la Duma (cámara baja), **ORG** Guennadi Selezniiov, **PER** calificó de "claramente ap

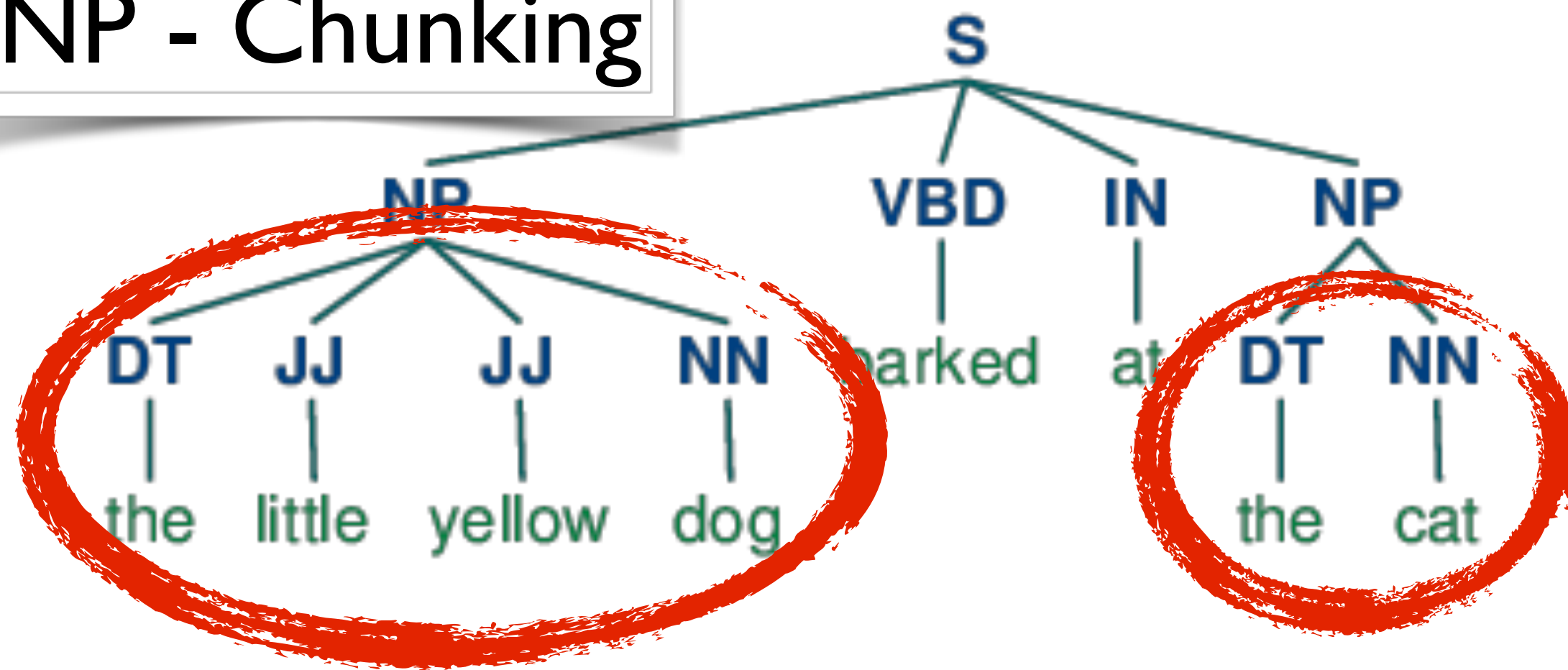
del **ORG** Kremlin para **LOC** Chechenia, **PER** Serguéi Yastrzhembski.

El asesor presidencial dijo que **LOC** Rusia puede lanzar un ataque preventivo contra los camp

# Chunking

- Identify sequences of non-overlapping labels

## NP - Chunking



Destacados representantes del **ORG** Parlamento y la prensa rusos criticaron hoy el "belicismo" ha definido como posible blanco de su lucha antiterrorista.

El presidente de la Duma (cámara baja), **ORG** Guennadi Seleznirov, **PER** calificó de "claramente ap

del **ORG** Kremlin para **LOC** Chechenia, **PER** Serguéi Yastrzhembski.

El asesor presidencial dijo que **LOC** Rusia puede lanzar un ataque preventivo contra los camp



# Chunking

- IOB Representation
  - Every token is **I**n a chunk or **O**ut of a chunk.
  - Distinguish the **B**eginnings of chunks.

W	e	s	a	w	t	h	e	y	e	l	l	o	w	d	o	g
PRP		VBD			DT			JJ						NN		
B-NP		O			B-NP			I-NP						I-NP		

MADden



# MADden

## MADden: Query-Driven Statistical Text Analytics

Give me  sentiment comments about

Query 1

Compare players  and  by the twitter sentiment over dates from  to  and return  results.

Query 2

Return all the named entity tags from the text

Kim began his career in psychology, graduating from UF with a master's degree in clinical psychology in 1971 and a doctorate in the same subject in 1974. While at UF, he met his wife, Katrine, who also earned her doctorate in clinical psychology at UF. He

Query 4

# MADden

## MADden Statistical

Give me

Query 1

Compare pla

10/22/201

Return all the

Kim began  
clinical psych  
met his wife

Query 4

### Answer

0	Kirn
32	UF
39	Katrine
79	Kentucky
94	Bellarmino
94	University
96	Louisville

### Query Plan

```
Function Scan on cgrant_ne_chunk (cost=0.25..12.75 rows=5 width=36)  
  Filter: (tag = 'NE'::text)
```

### The Query

```
-- select termnum, term from  
cgrant_ne_chunk('Kirn began  
his career in psychology,  
graduating from UF with a  
master's degree in clinical  
psychology in 1971 and a  
doctorate in the same subject  
in 1974. While at UF, he met  
his wife, Katrine, who also  
earned her doctorate in  
clinical psychology at UF. He  
worked in the mental health  
field for six years, first as  
an intern and later at  
community mental health  
centers and in a private  
practice in Kentucky that he  
owned with his wife. He also  
was a full-time faculty member  
at Bellarmine University in  
Louisville for six years',  
true) where tag = 'NE' ;
```

4.6008198261261 sec



# Dependency Parsing

- A graph depicting the relationship between a word (head) and its dependents.
- Starts with a verb and finds the related subject and object.
- Useful in understanding phrases
- Similar to chunking
- Very close to **semantic relationships**
- Link grammar is the most notable implementation (in AbiWord)

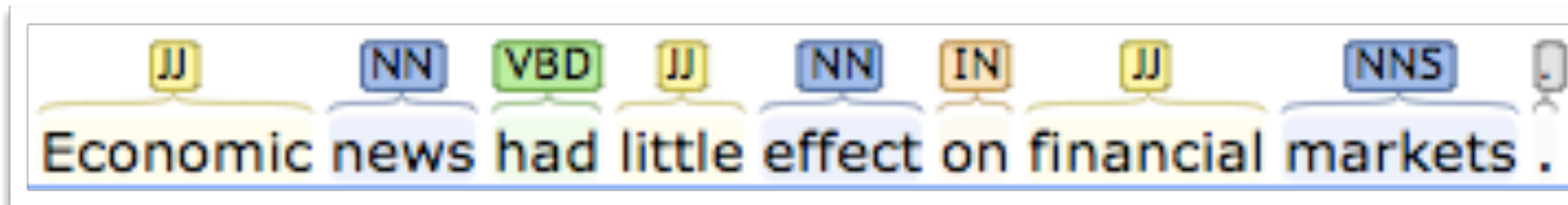


# Dependency Parsing

Economic news had little effect on financial markets .

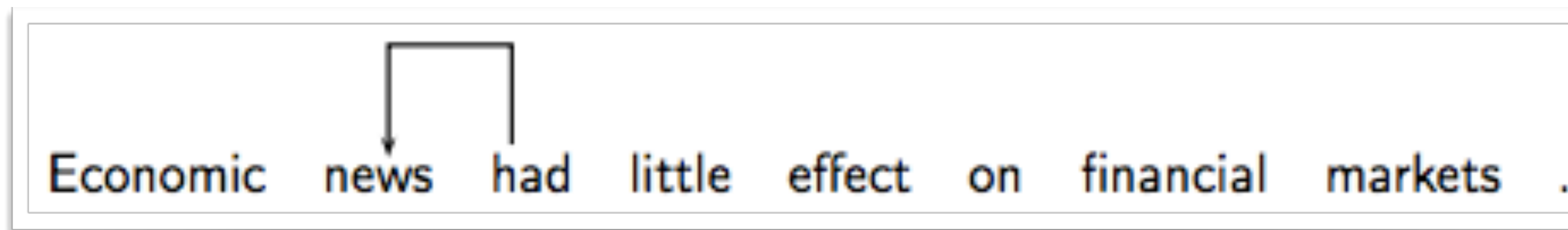
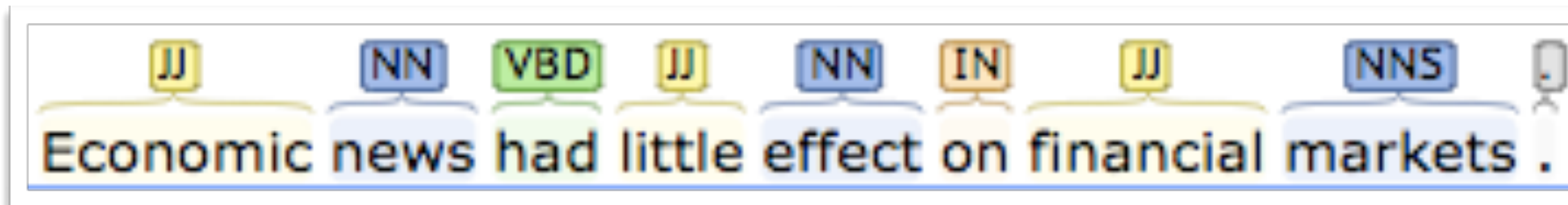


# Dependency Parsing

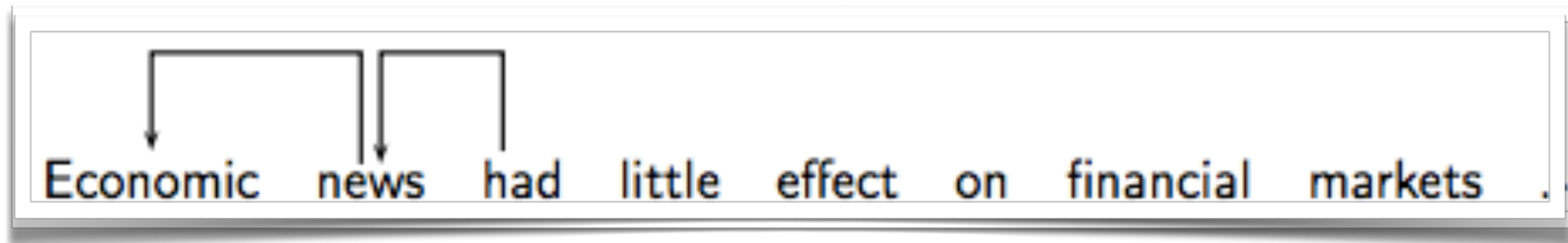
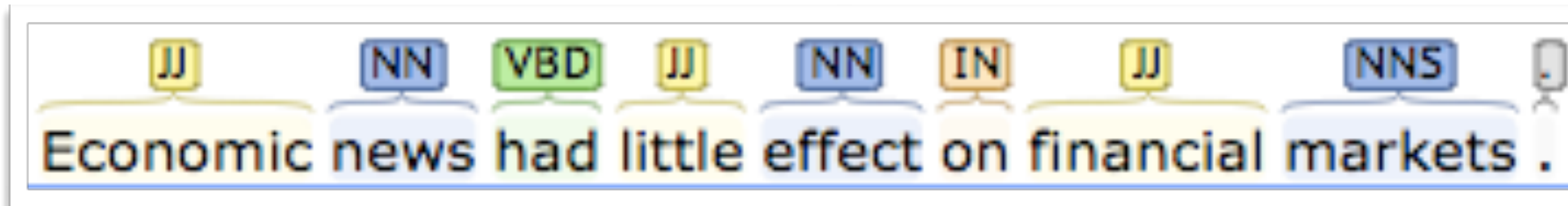


Economic news had little effect on financial markets .

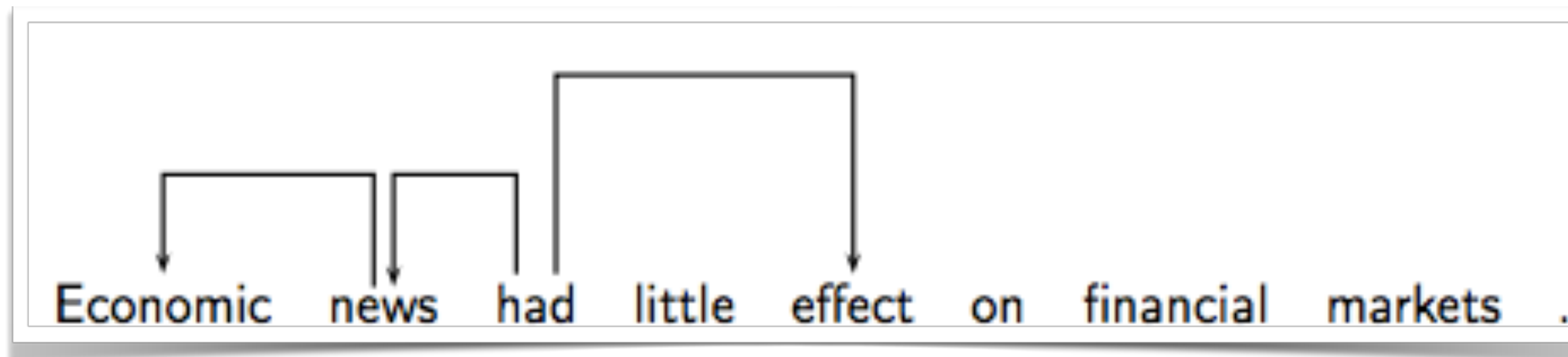
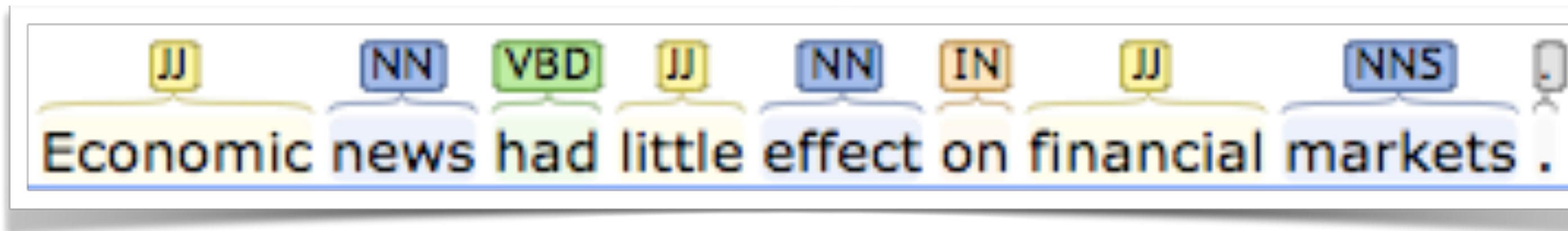
# Dependency Parsing



# Dependency Parsing

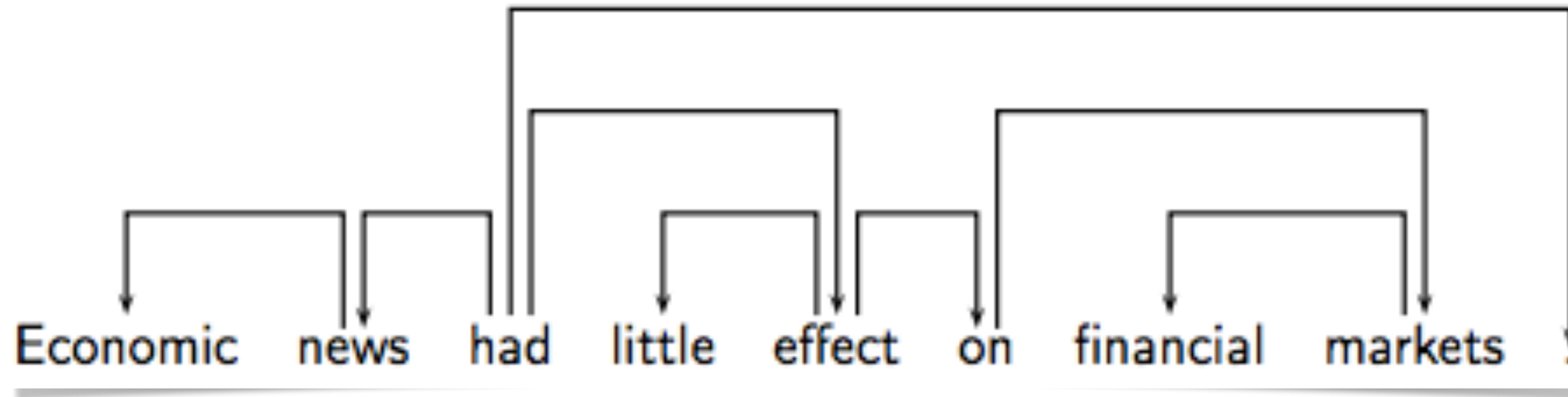
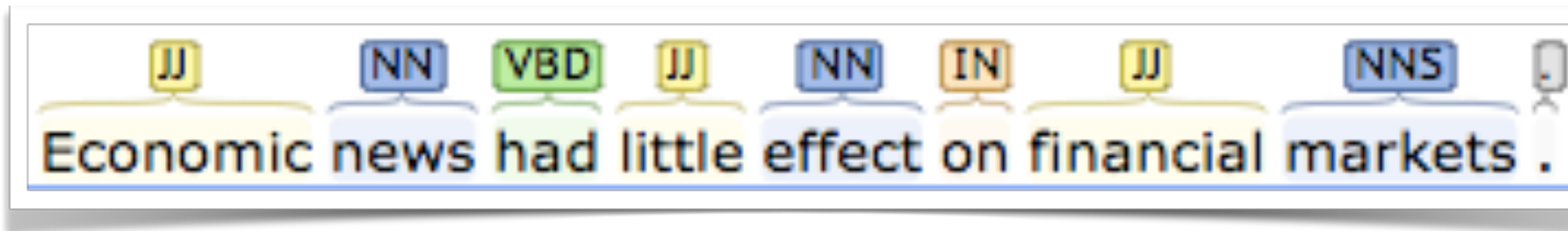


# Dependency Parsing

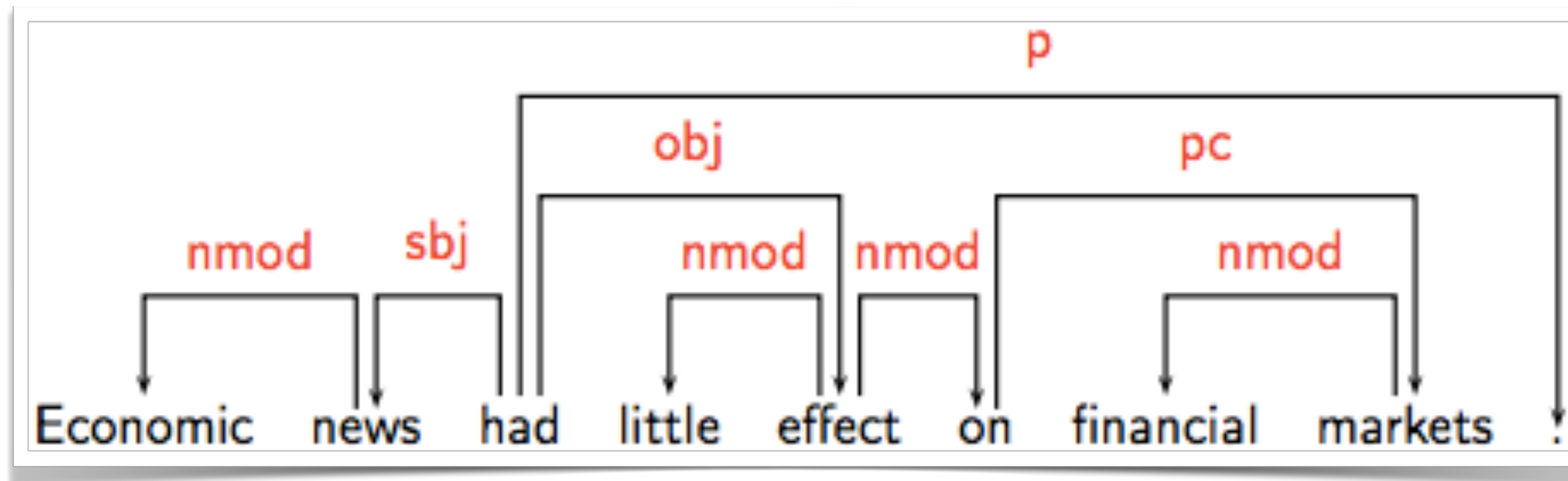
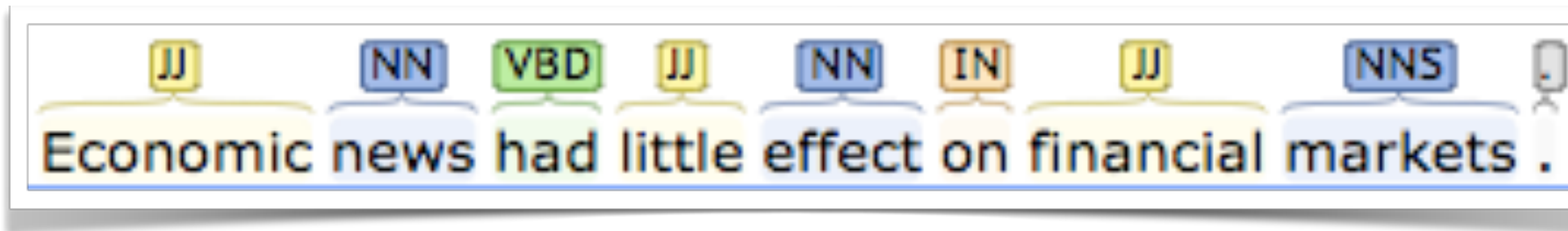




# Dependency Parsing

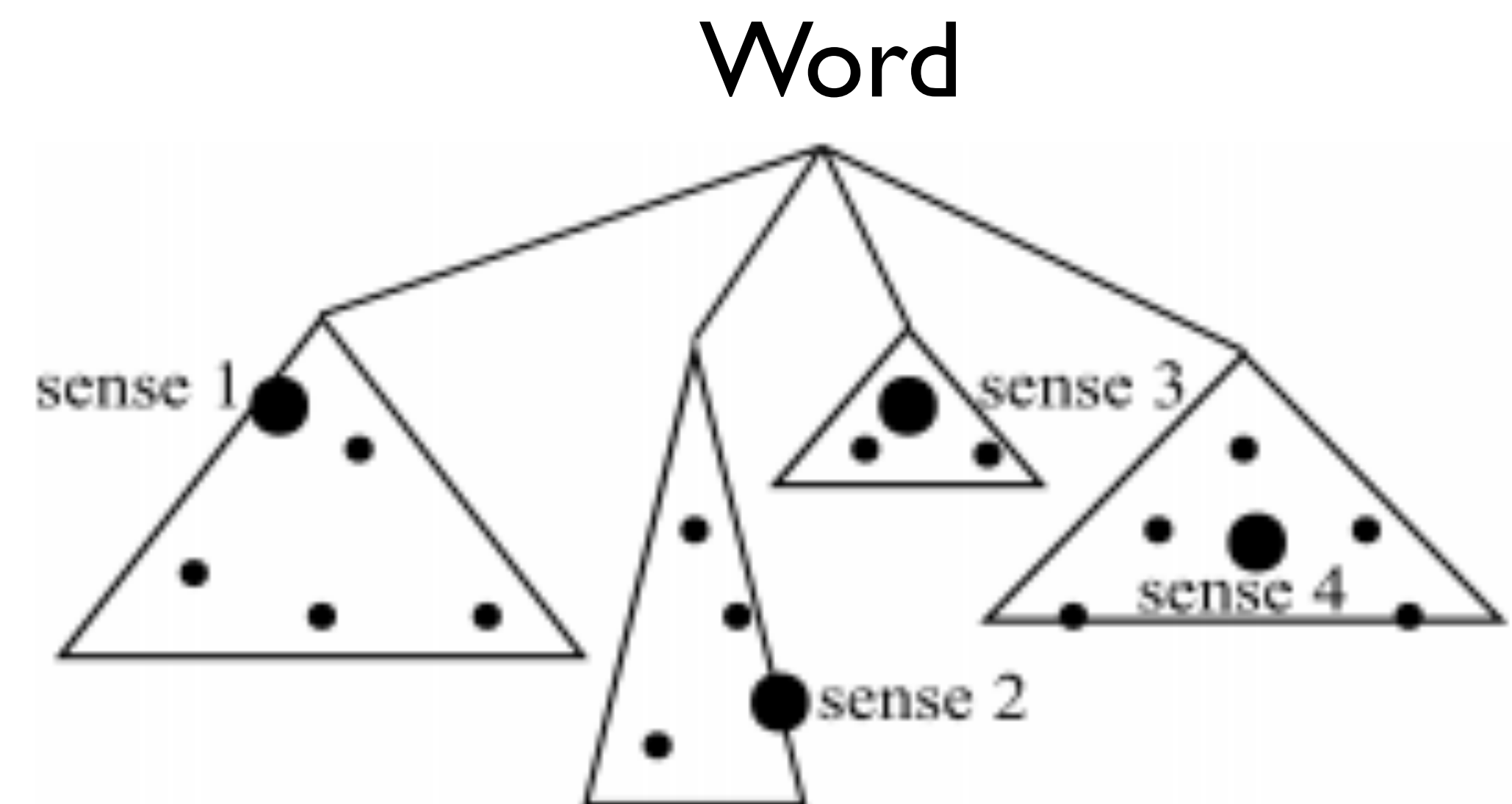


# Dependency Parsing



# Word-Sense Disambiguation

- Classifying the meaning of a word among many possible interpretations.
- Classification can be done in a myriad of ways.
- Still an open NLP problem
- I like bass!

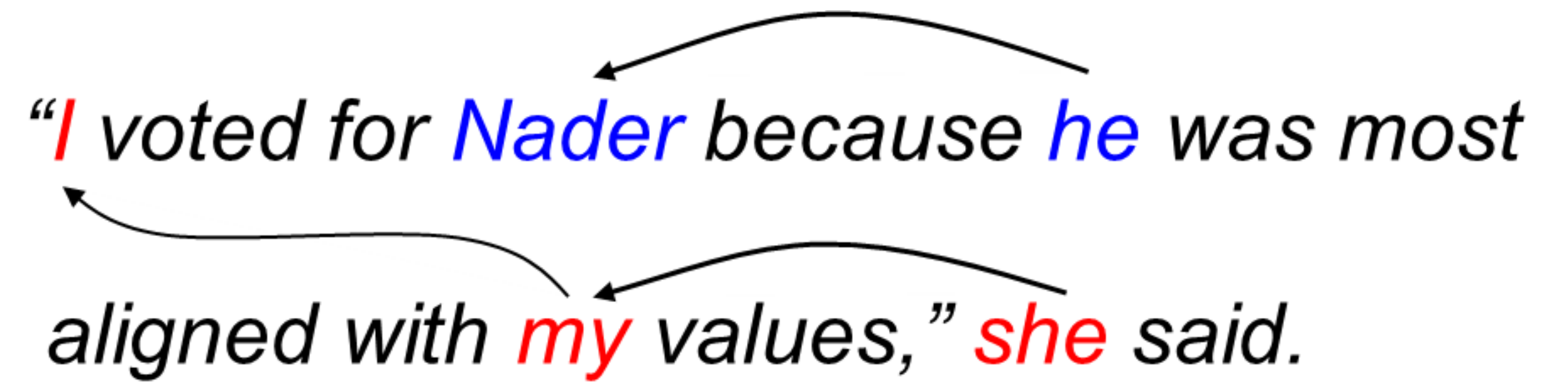




# Co-Reference Resolution

- Determining the mentions in a document that correspond to the **same** entity.

*“I voted for **Nader** because **he** was most aligned with **my** values,” **she** said.*



The diagram illustrates co-reference resolution in the sentence: "I voted for Nader because he was most aligned with my values," she said. Arrows indicate the following relationships: an arrow from "Nader" to "he", an arrow from "my" to "she", and an arrow from "I" to "she".



# Co-Reference Resolution

- Determining the mentions in a document that correspond to the **same** entity.

MMAX2 1.12 /home/yannick/tmp/MUC-MMAX/muc6/test/ws93\_022.0297.mmax [modified]  
File Settings Display Tools Plugins Info ShowNLPanel  
DOCID: wsj93\_022\_0297  
DOCNO: 930504-0023 .  
HL: IBM appoints Chrysler's York as finance chief --- computer maker's move signals strategy of cuts in costs, asset sales --- by Michael W. Miller and Douglas Lavin staff reporters of the wall street journal  
DD: 05/04/93  
wall street journal (j) Page 33 of IBM automobiles ( aut ), computers ( cpr )  
TXT: International Business Machines Corp. continued its executive makeover by hiring Jerome B. York, an architect of the turnaround at Chrysler Corp., to become chief financial officer.  
Mr. York, 54 years old, is a West Point graduate who helped transform Chrysler by slashing costs and selling billions of dollars in assets.  
His appointment is a strong sign that IBM's new chairman, Louis V. Gerstner Jr., plans a similar strategy at the wounded computer giant.  
Mr. Gerstner raced to hire Mr. York after meeting him for the first time just three weeks ago in IBM's Manhattan offices.  
In his first month, Mr. Gerstner has also brought in outsiders to run IBM's communications and disk-drive business, and is searching for a new head of personnel.  
Mr. York was executive vice president for finance and a board member at Chrysler, where he spent 12 years in financial posts and running several car and truck divisions.  
Chrysler did n't name a successor yesterday.

*"I voted for Nader because he was most aligned with my values," she said.*

# Co-Reference Resolution

- Determining the mentions in a document that correspond to the **same** entity.

The screenshot shows a document with several paragraphs. Red lines connect mentions of 'IBM' and 'Chrysler' across different paragraphs, illustrating co-reference chains. For example, 'IBM' is mentioned in the first paragraph, and 'the wounded computer giant' is mentioned in the second paragraph. A red line connects these two mentions, indicating they refer to the same entity. Other lines connect 'Chrysler' mentions to 'the wounded computer giant' mention.

*"I voted for Nader because he was most aligned with my values," she said.*

Entity/Co-reference Chains





# Cross-Document Entity Resolution

- Take coreference chains from *across documents* and match the ones that correspond to the same real world entity.
- A type of clustering problem.
- Use the features from the document.

**Thomas Cruise**



**Michael Jordan**



# Entity Resolution Model

- Entity Resolution/Coreference Resolution is an ubiquitous problem.



# Entity Resolution Model

- Entity Resolution/Coreference Resolution is an ubiquitous problem.
- Many models and many domains.

# Entity Resolution Model

- Entity Resolution/Coreference Resolution is an ubiquitous problem.
- Many models and many domains.
- We use the McCallum method (McCallum, Wellner 2004)

# Entity Resolution Model

- Entity Resolution/Coreference Resolution is an ubiquitous problem.
- Many models and many domains.
- We use the McCallum method (McCallum, Wellner 2004)
  - Statistically sound — Based on conditional random fields (CRF).

# Entity Resolution Model

- Entity Resolution/Coreference Resolution is an ubiquitous problem.
- Many models and many domains.
- We use the McCallum method (McCallum, Wellner 2004)
  - Statistically sound — Based on conditional random fields (CRF).
  - Relational — Does not assume independence.



# Entity Resolution Example

# Entity Resolution Example

We extract set of noun phrases from text documents using *named entity recognition*.

# Entity Resolution Example

We extract set of noun phrases from text documents using *named entity recognition*.

# Entity Resolution Example

We extract set of noun phrases from text documents using *named entity recognition*.

...Late-night host and comic **Jimmy Fallon** was born on...



# Entity Resolution Example

We extract set of noun phrases from text documents using *named entity recognition*.

...Late-night host and comic **Jimmy Fallon** was born on...

...the Kanye West appearance on **Jimmy Kimmel** Live last...

# Entity Resolution Example

We extract set of noun phrases from text documents using *named entity recognition*.

...Late-night host and comic **Jimmy Fallon** was born on...

...the Kanye West appearance on **Jimmy Kimmel** Live last...

...you work at **Jimmy John's** sammich shops, where...

# Entity Resolution Example

We extract set of noun phrases from text documents using *named entity recognition*.

...Late-night host and comic **Jimmy Fallon** was born on...

...the Kanye West appearance on **Jimmy Kimmel** Live last...

...you work at **Jimmy John's** sammich shops, where...

**Jimmy Kimmel** Shares His "Only Complaint" About **Jimmy Fallon**

# Entity Resolution Example

We extract set of noun phrases from text documents using *named entity recognition*.



...Late-night host and comic **Jimmy Fallon** was born on...

...the Kanye West appearance on **Jimmy Kimmel** Live last...

...you work at **Jimmy John's** sammich shops, where...

**Jimmy Kimmel** Shares His "Only Complaint" About **Jimmy Fallon**



# Entity Resolution Example

We extract set of noun phrases from text documents using *named entity recognition*.



...Late-night host and comic **Jimmy Fallon** was born on...

...the Kanye West appearance on **Jimmy Kimmel** Live last...

...you work at **Jimmy John's** sammich shops, where...

**Jimmy Kimmel** Shares His "Only Complaint" About **Jimmy Fallon**