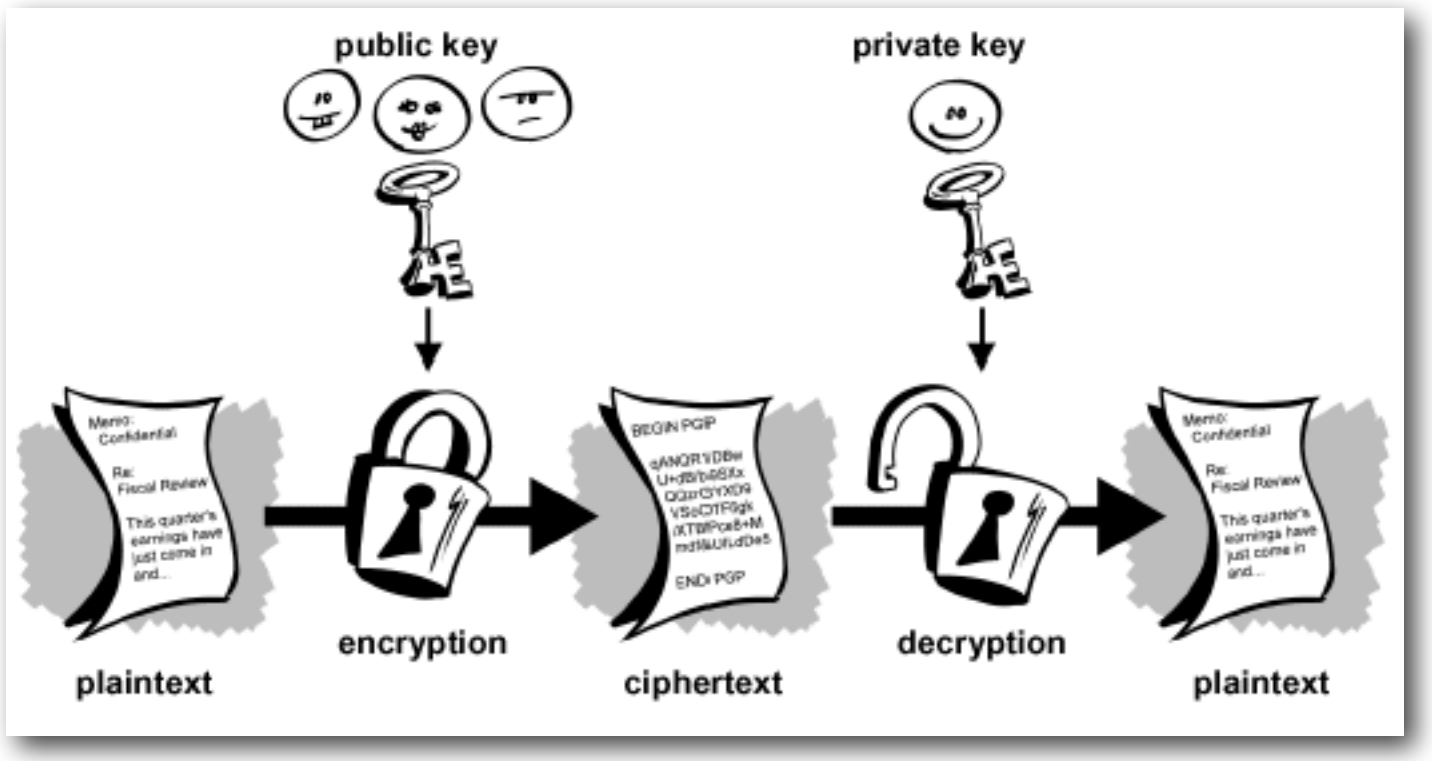
Get the data!

- 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

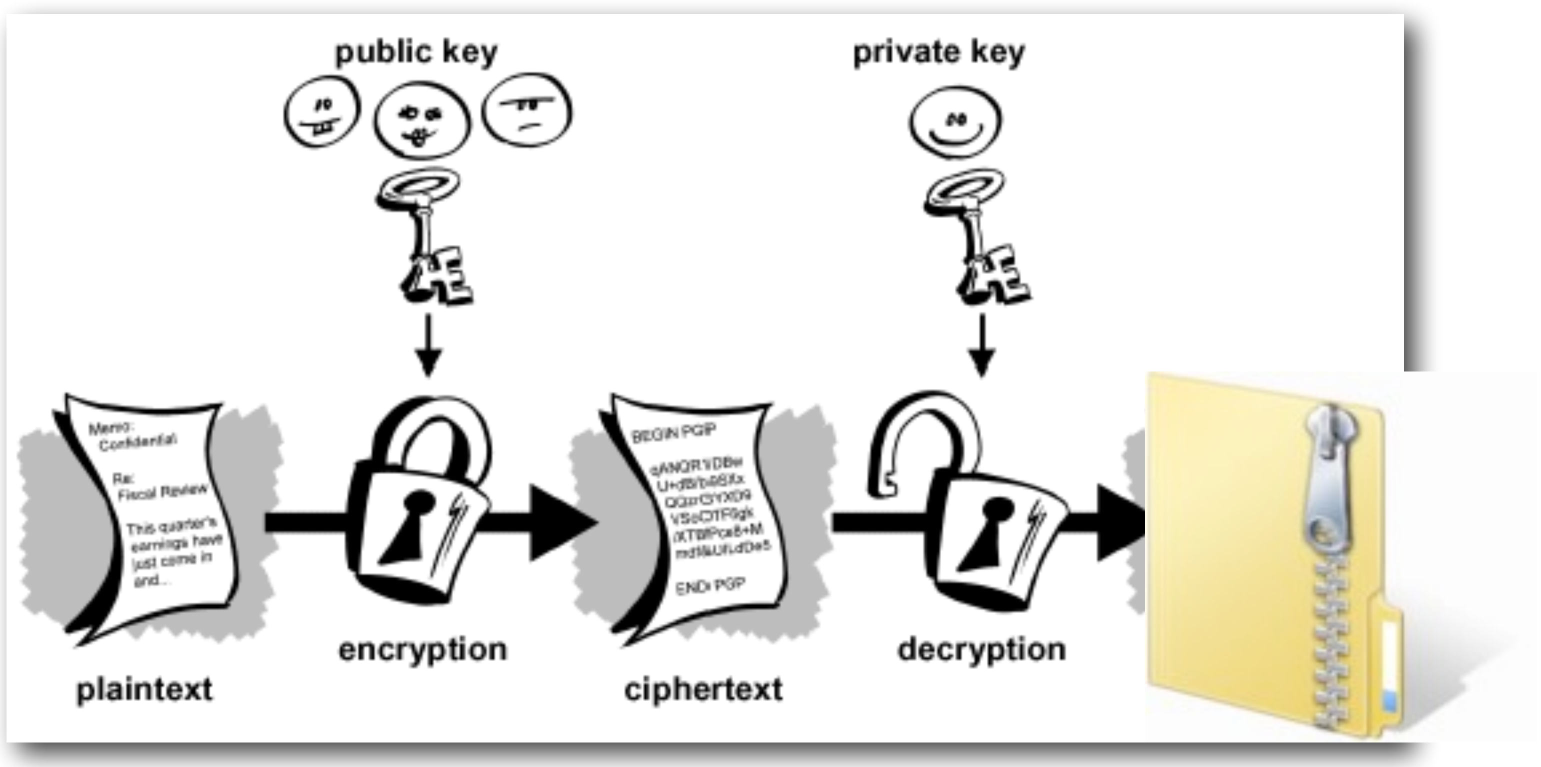
Get the data!

- 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



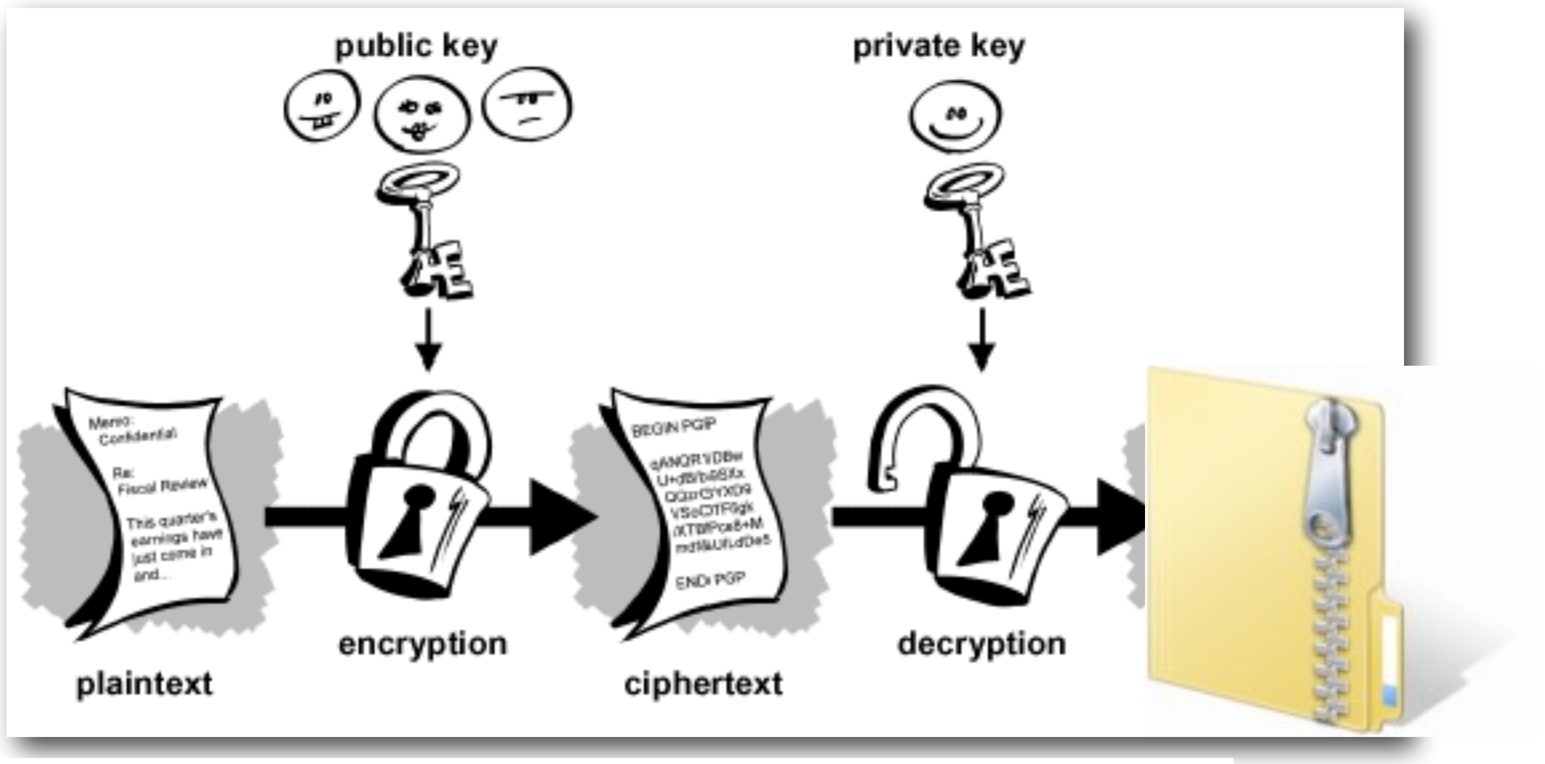
Get the data!

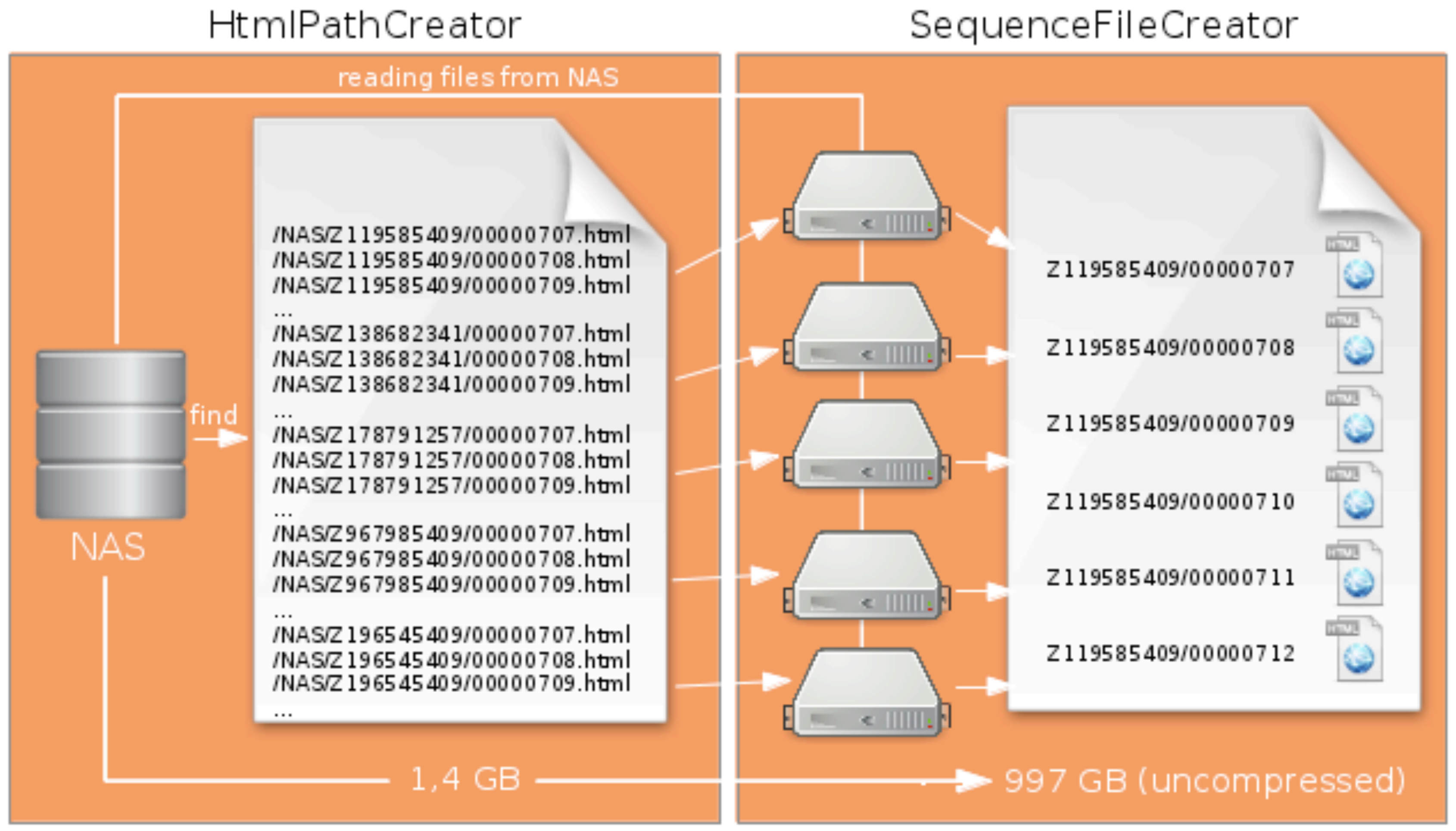
- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV



Get the data!

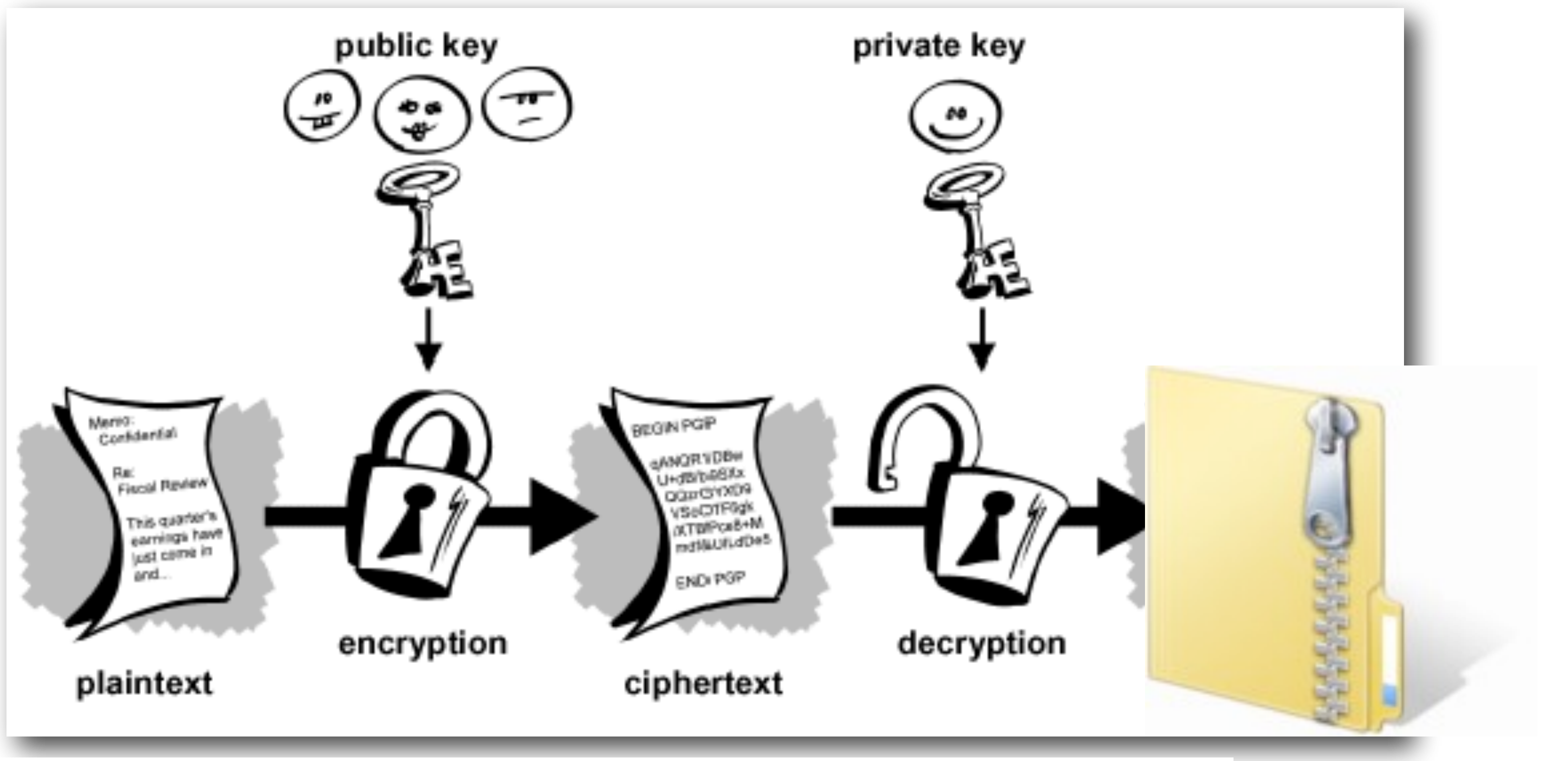
- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV





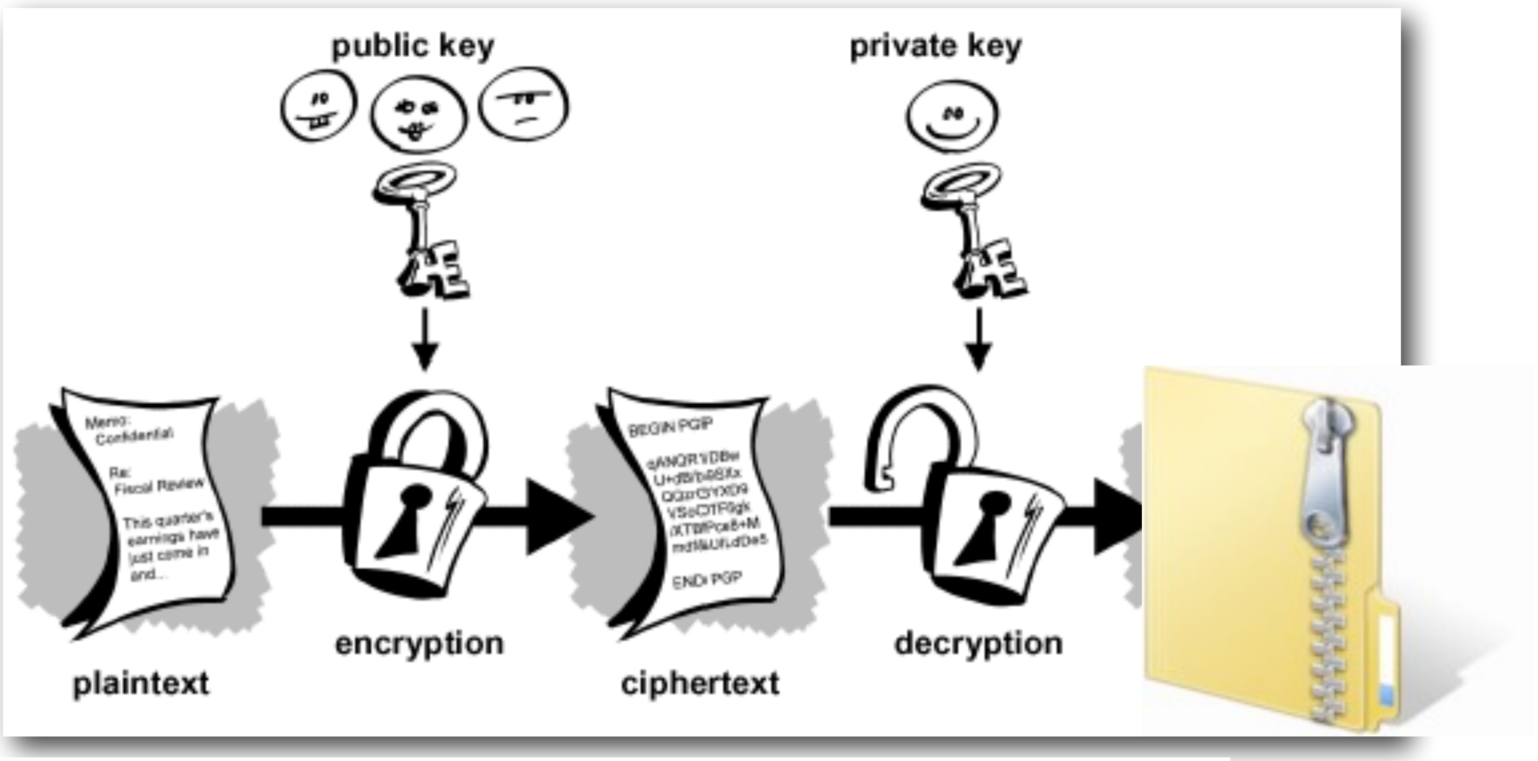
Get the data!

- 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



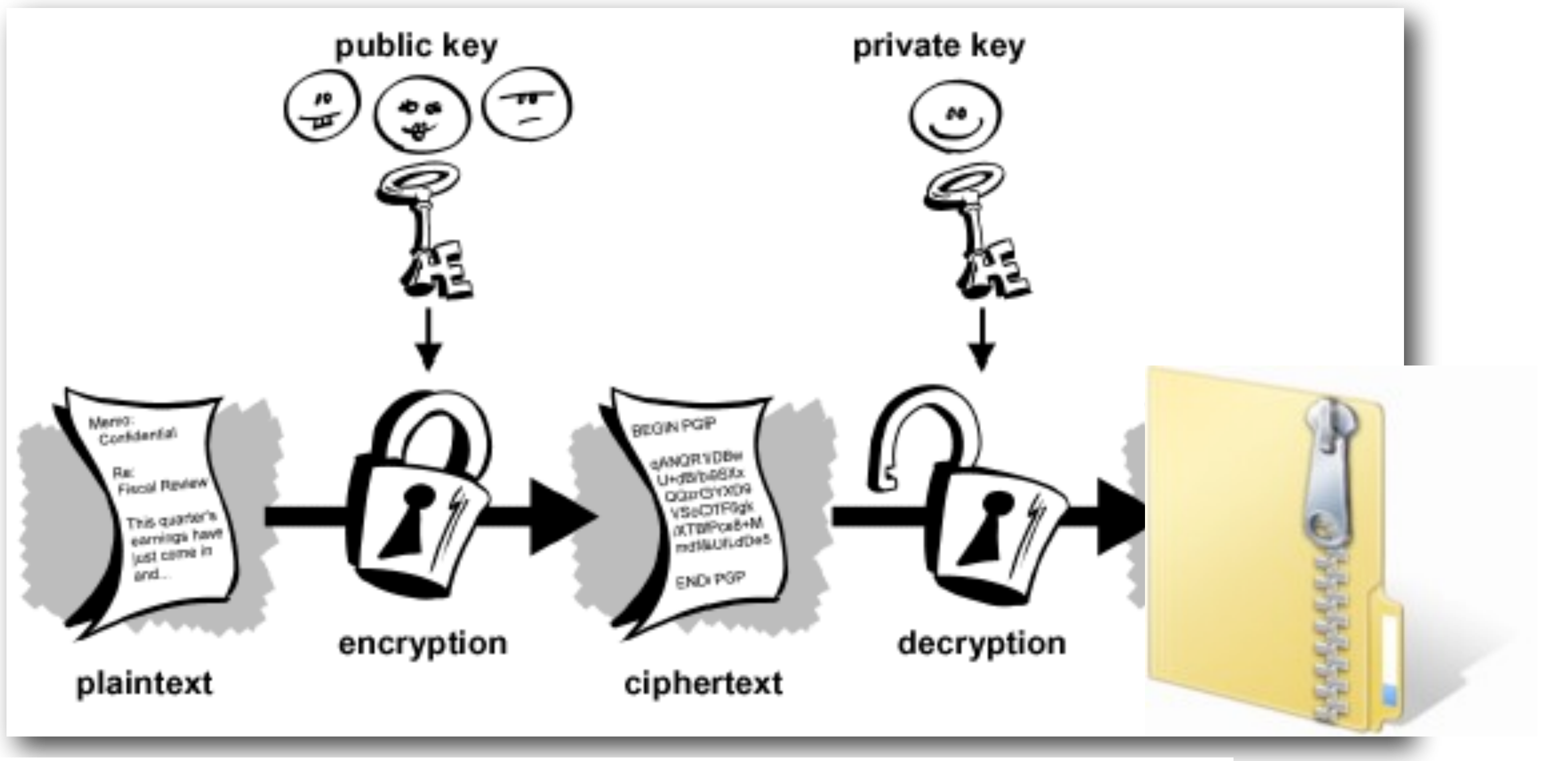
Get the data!

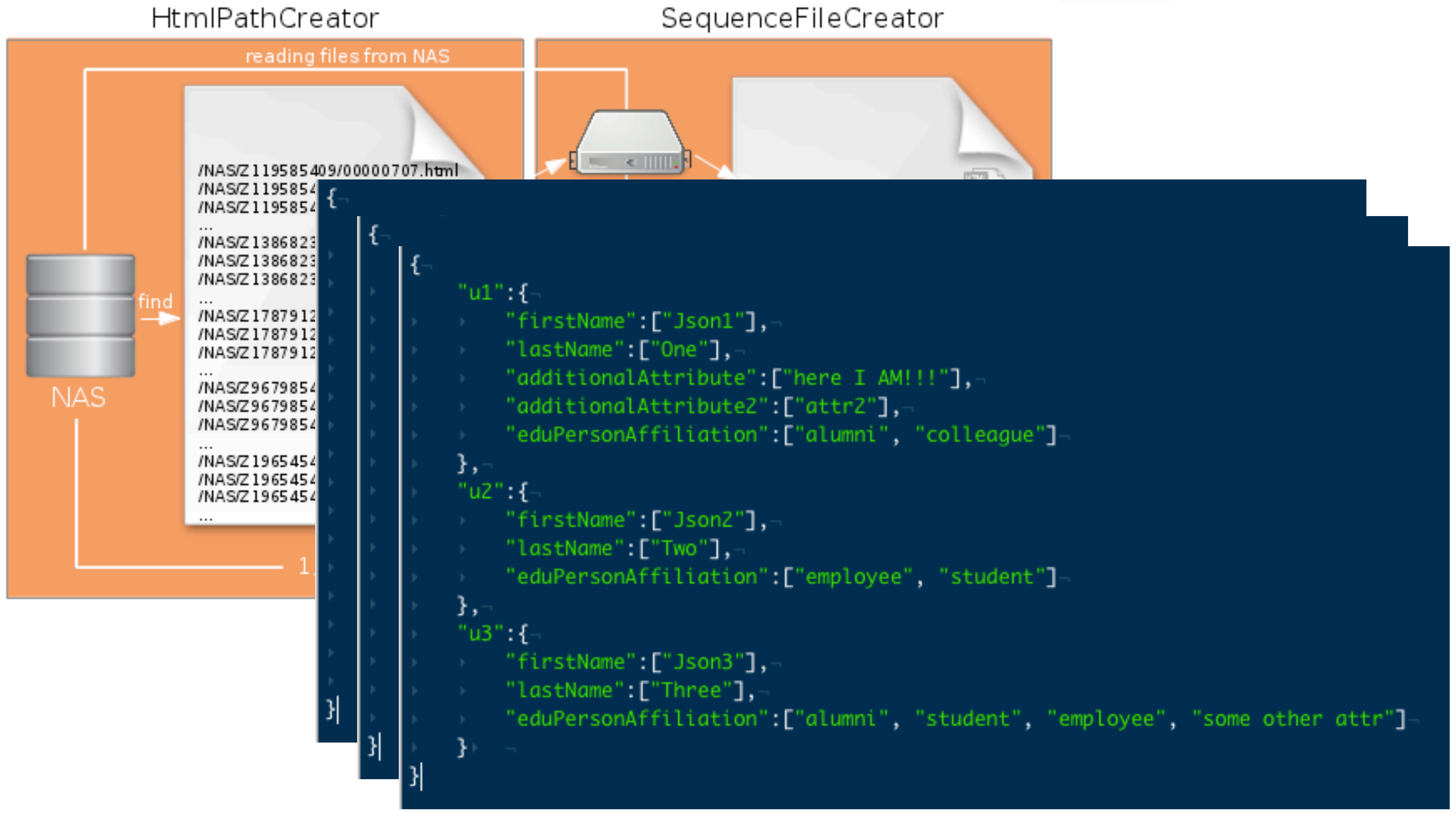
- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV



Get the data!

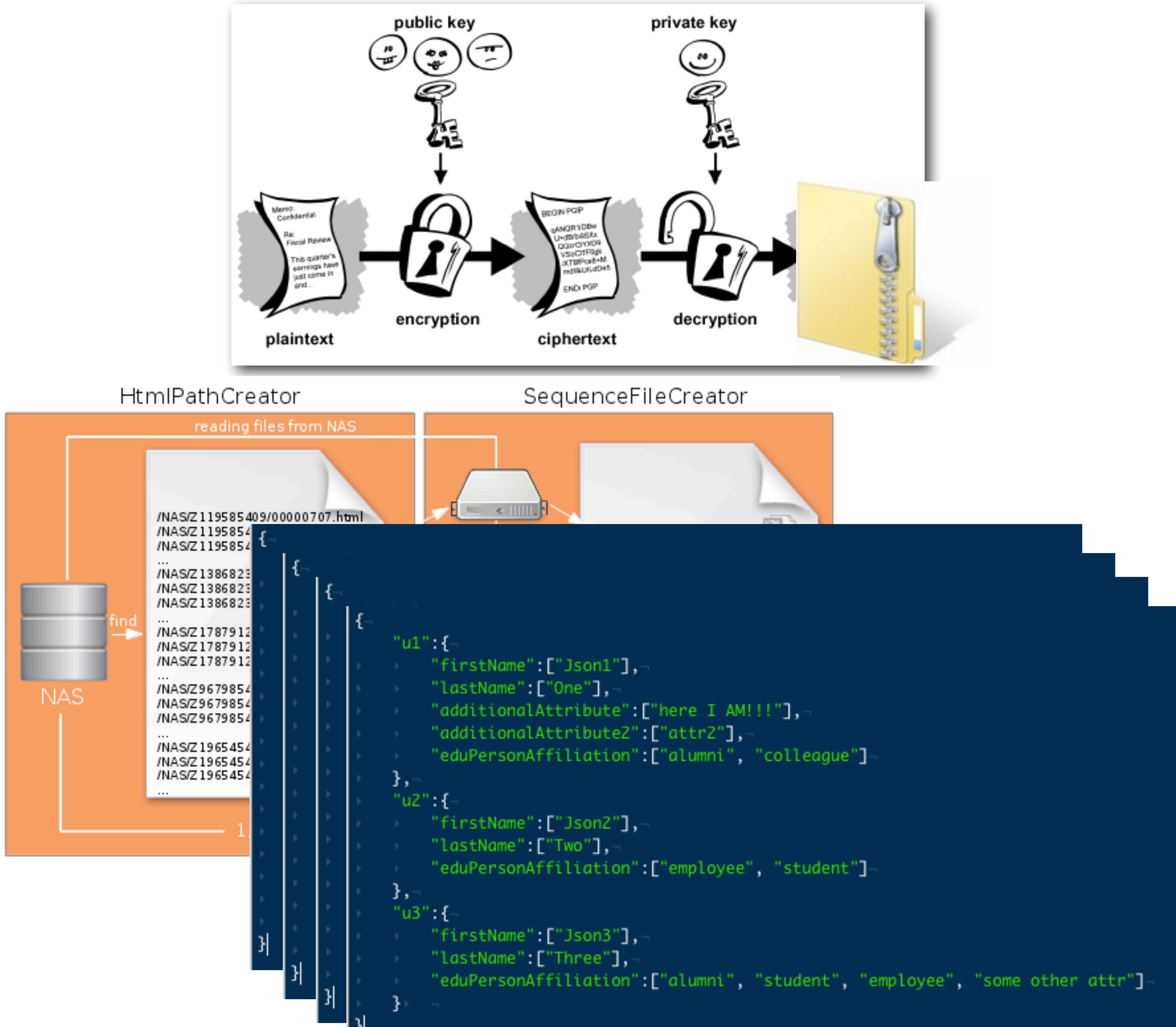
- Data sets are typically compressed in large batches of files.
- The files are:
  - Encrypted with gpg
  - Compressed .gz/.bz/.xz files
  - Hadoop sequence .sc files
  - Thrift/JSON/XML/CSV





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

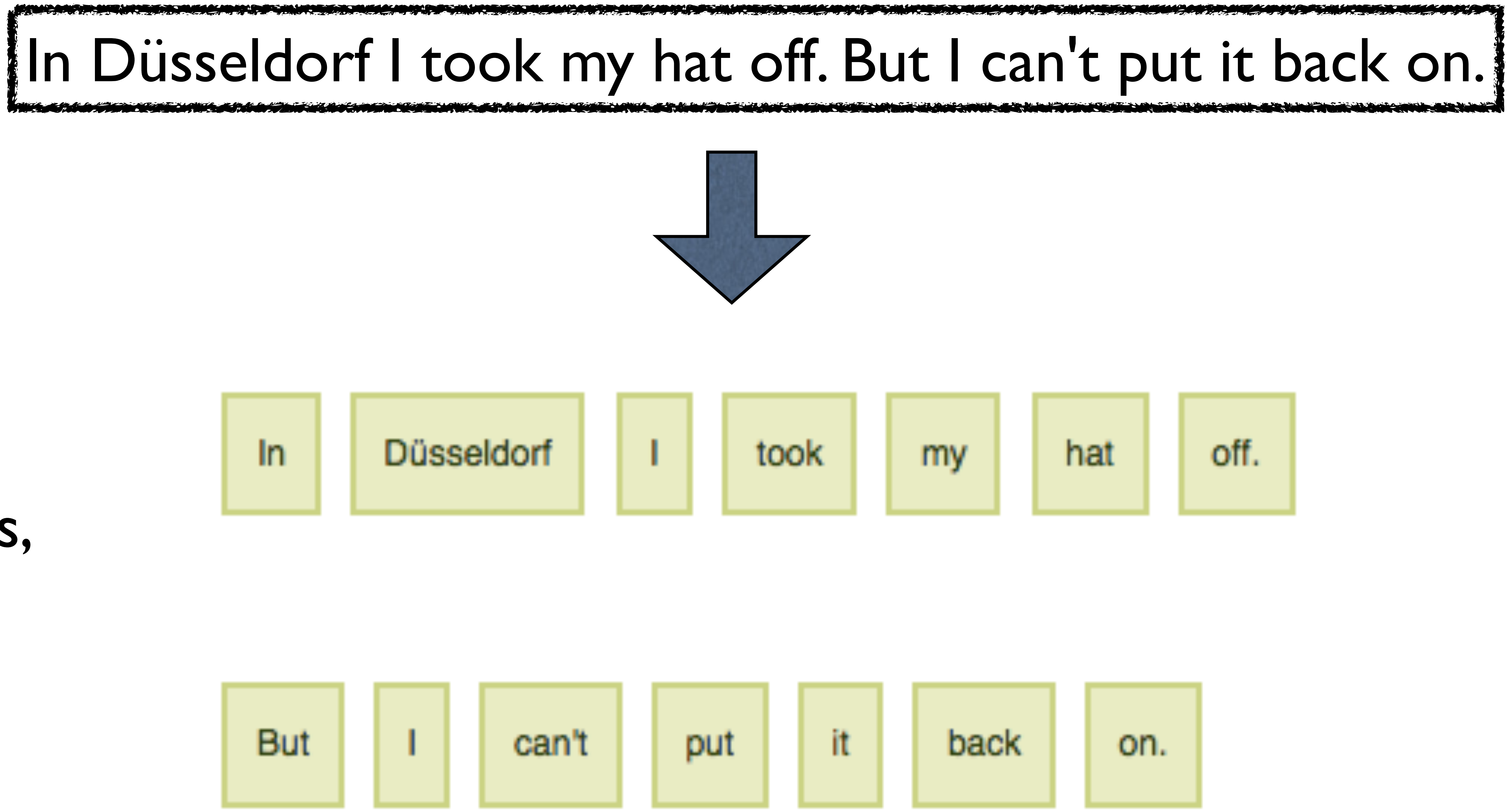
- Split a textual document into sentences.
  - Was that an abbreviation?
  - Was that inside a quote?

```
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

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



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	*to*
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential *there*	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	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-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 *there*	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	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-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

PRP	MD	RB	VB	TO	VB	NN	IN	JJ	NN	IN	VBG	VB	NNS	IN	NNP	NNP	.
I	can	still	manage	to	have	fun	in	graduate	school	by	taking	Text	Analytics	with	Dr.	Grant	.

Named Entity Recognition (NER)

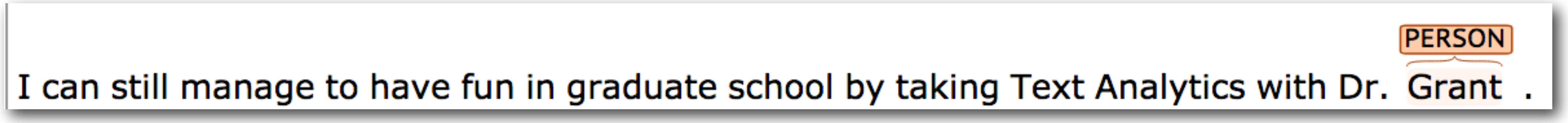
- Identify the tokens in a sentence that correspond to a Entity.

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
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.

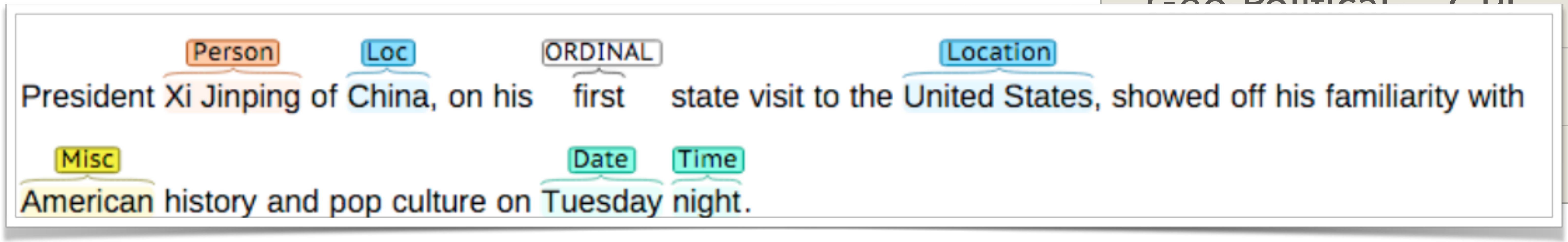
Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes seas
Geo-Political	GPE	Countries states, provinces, counties
Facility	FAC	Bridges,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

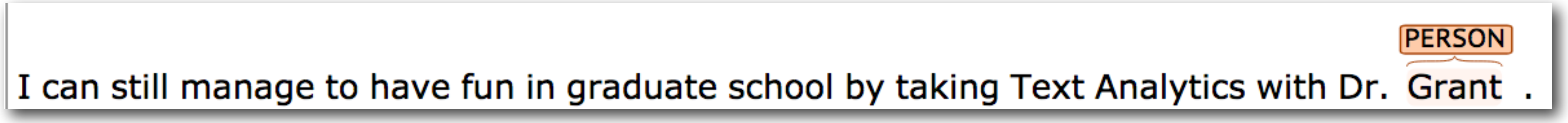


Named Entity Recognition (NER)

- 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



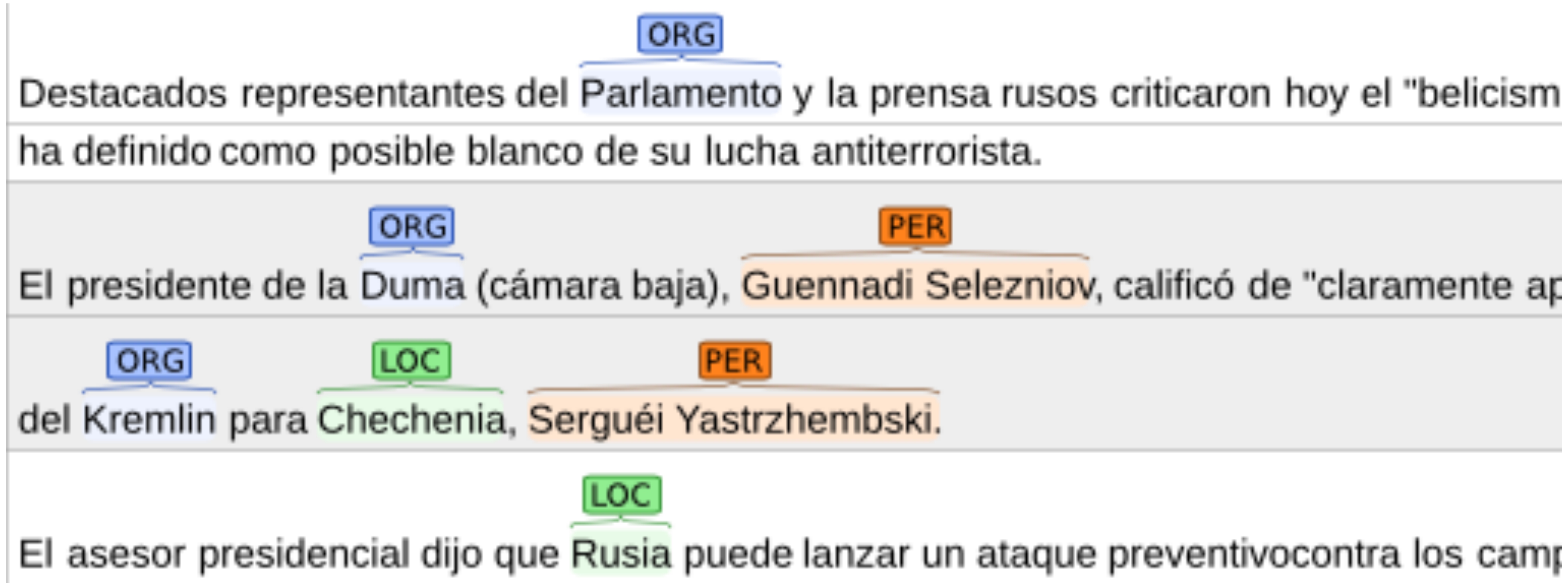


Chunking

- Identify sequences of non-overlapping labels

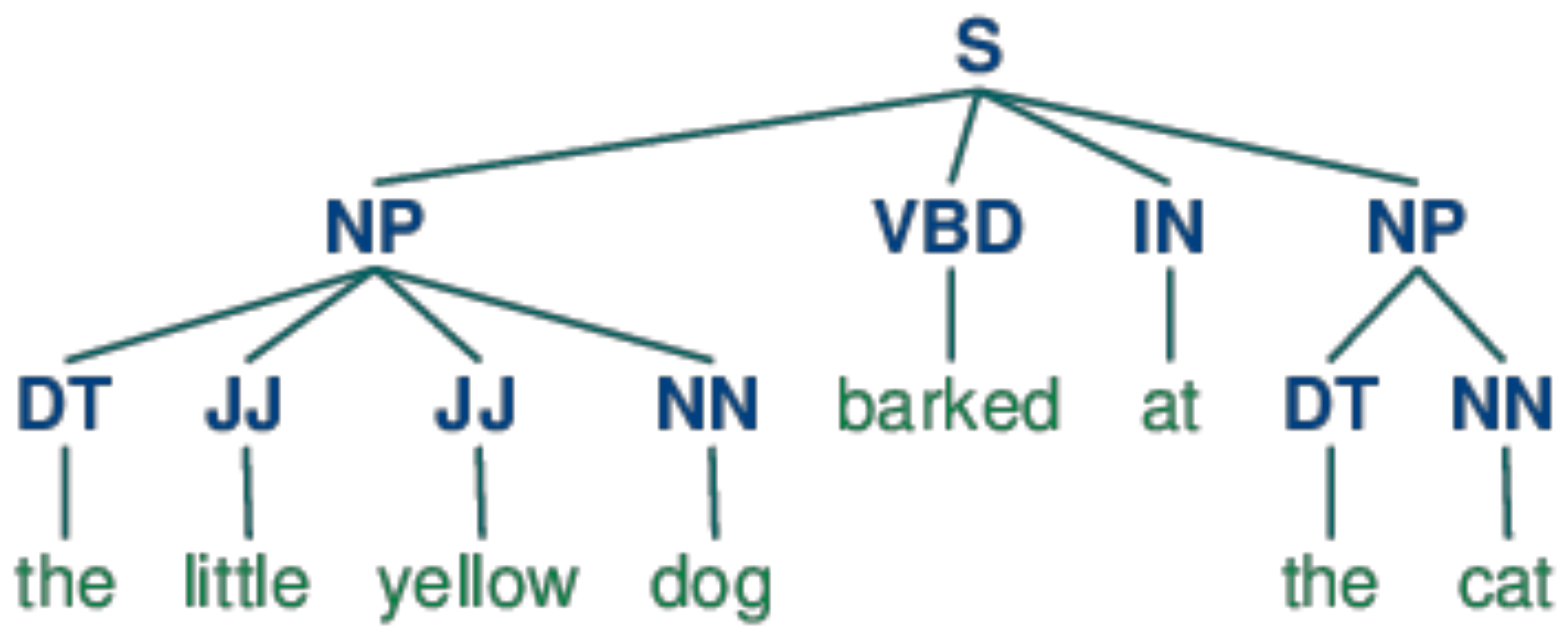
Chunking

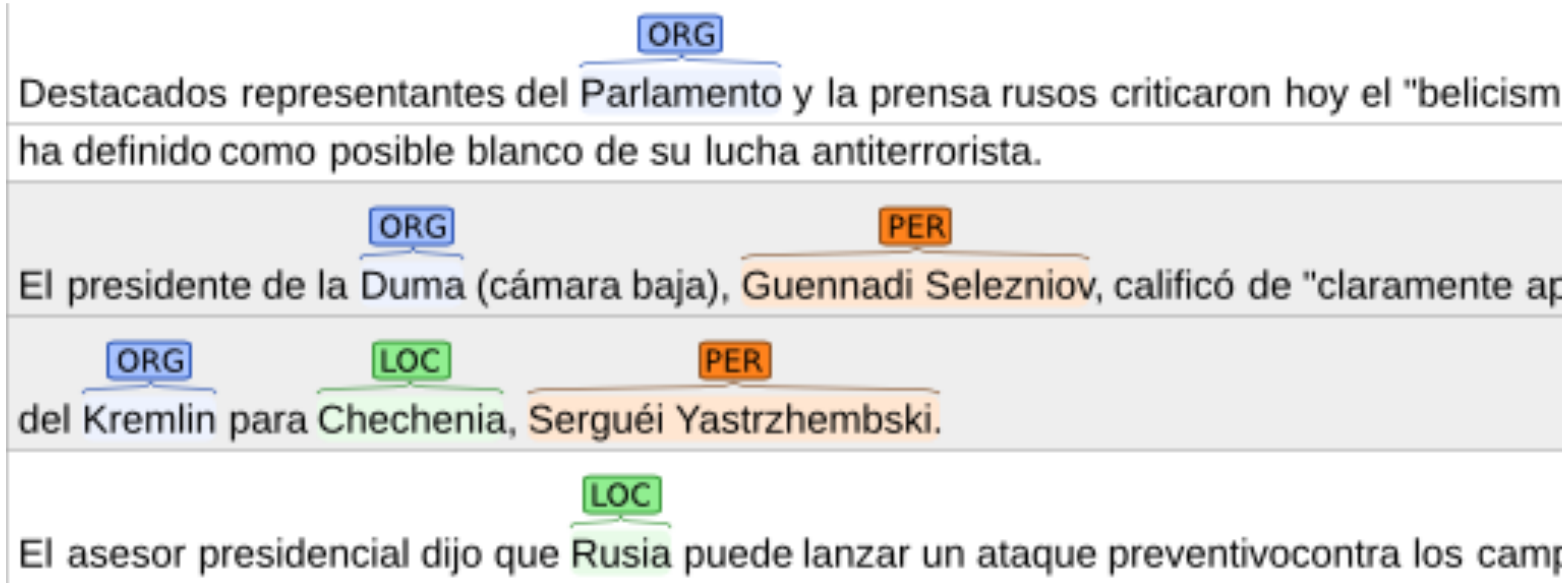
- Identify sequences of non-overlapping labels



Chunking

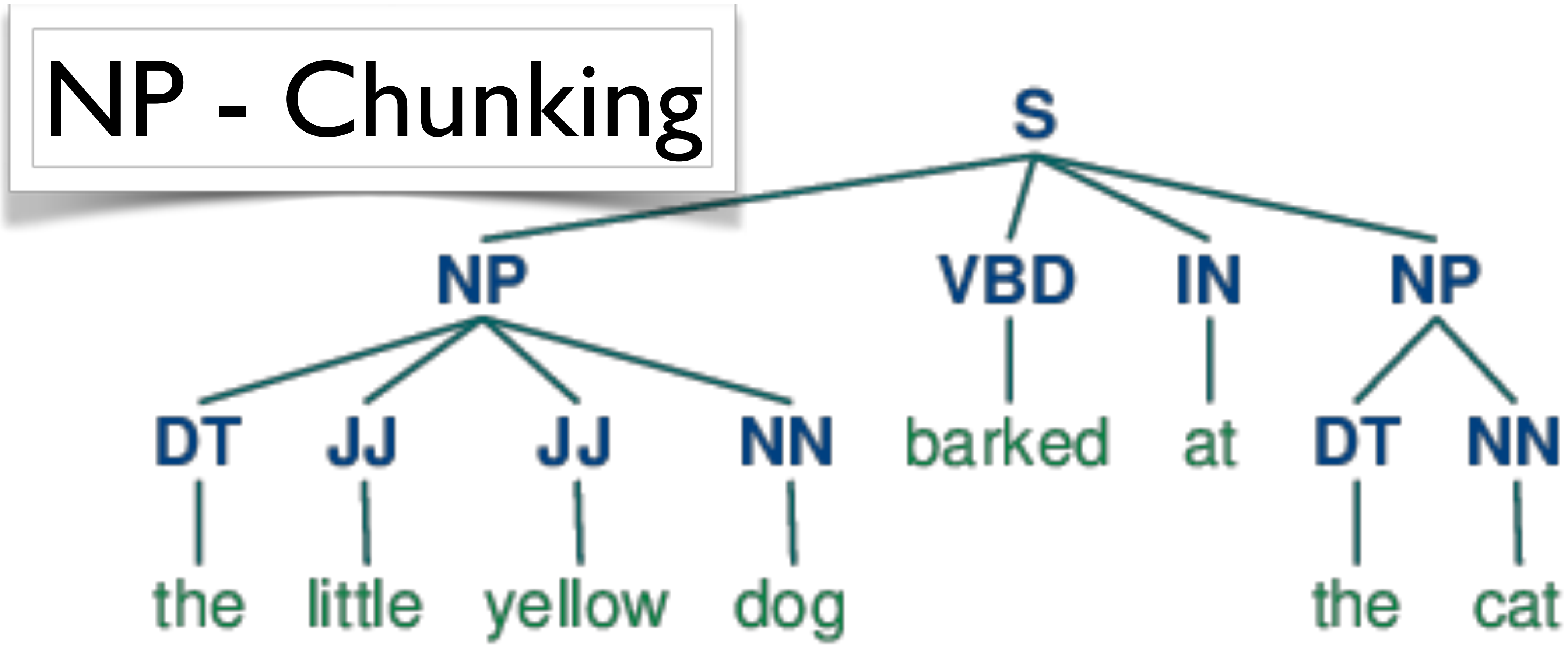
- Identify sequences of non-overlapping labels

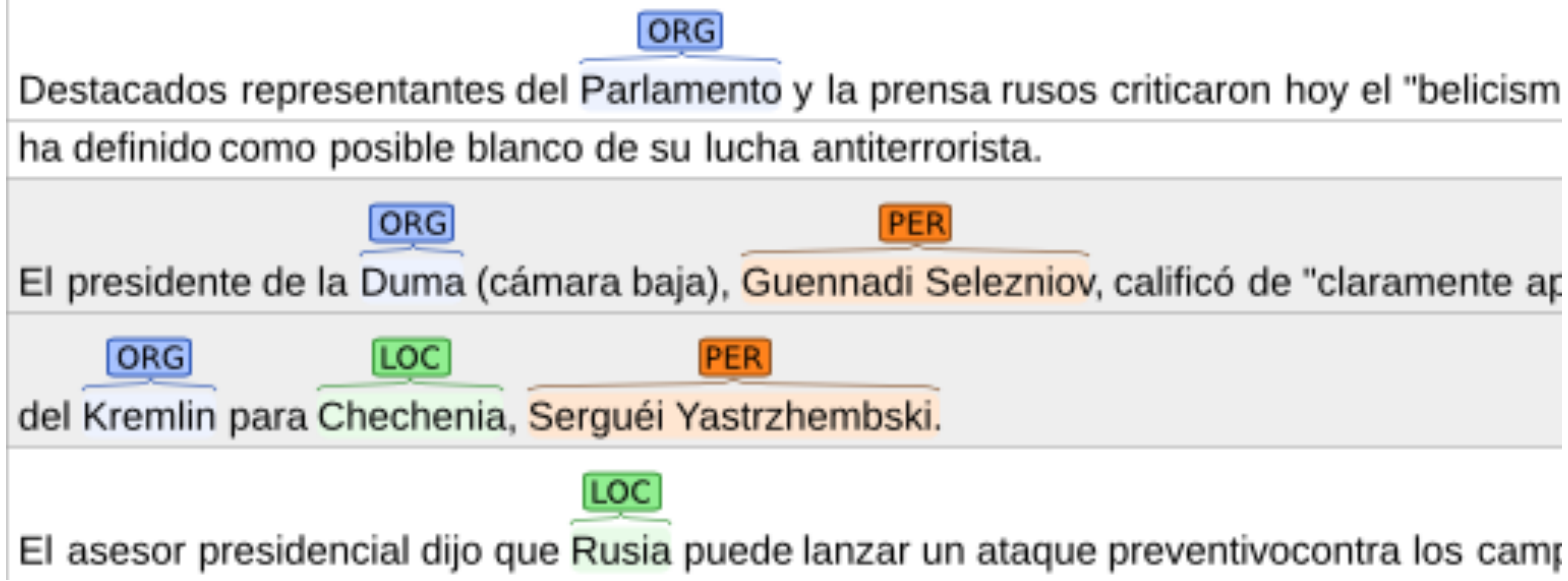




Chunking

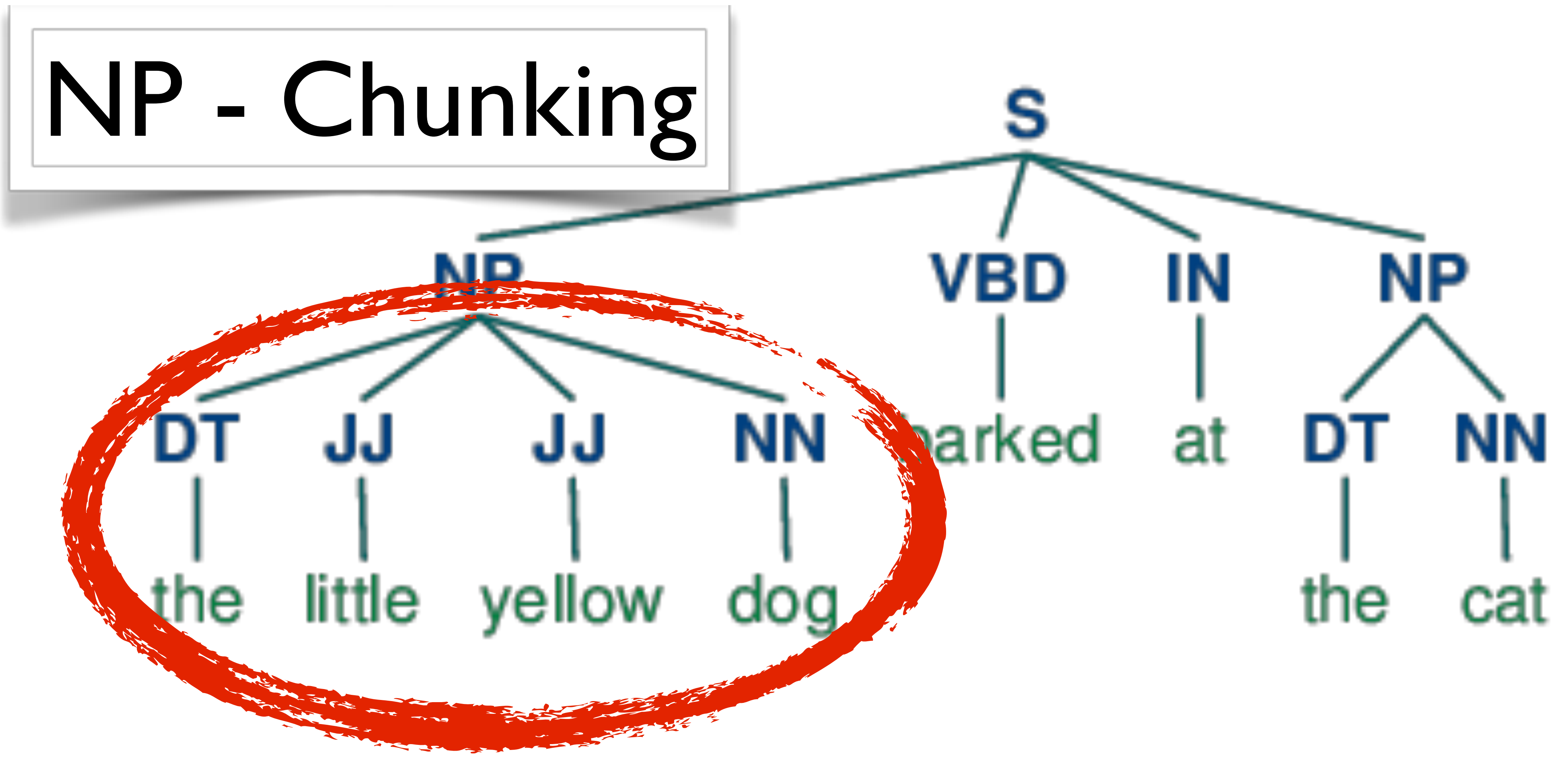
- Identify sequences of non-overlapping labels

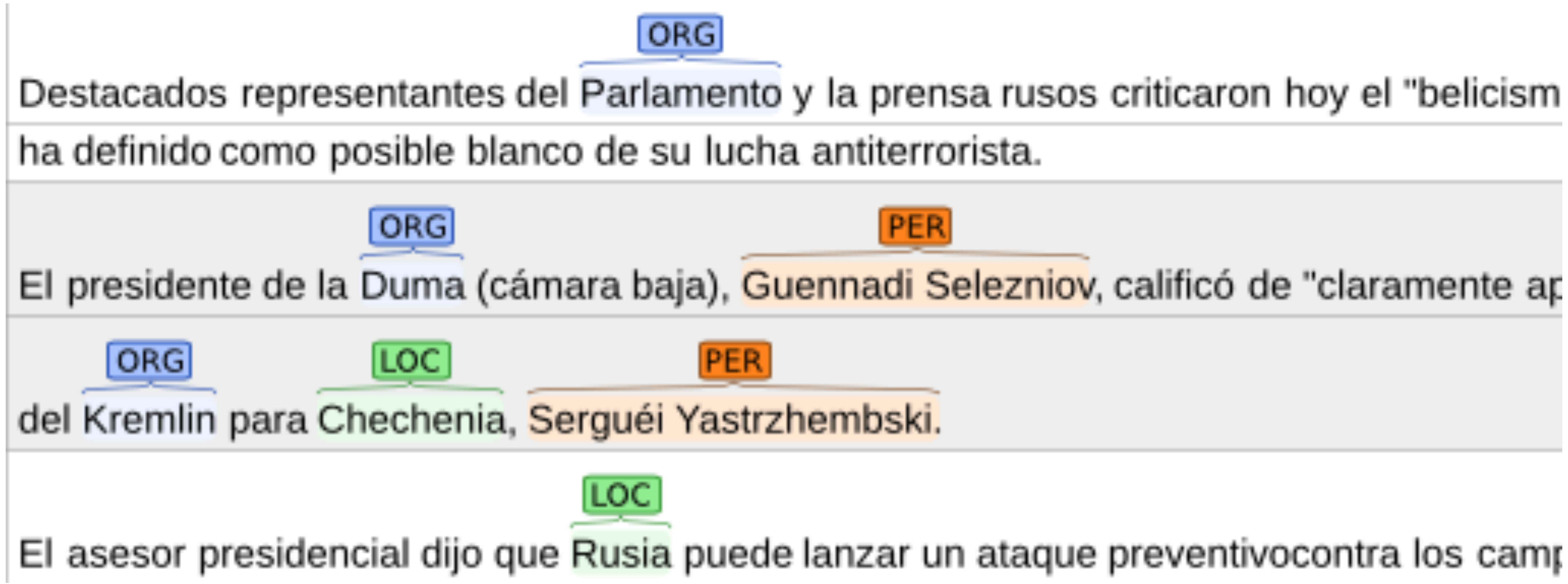




Chunking

- Identify sequences of non-overlapping labels

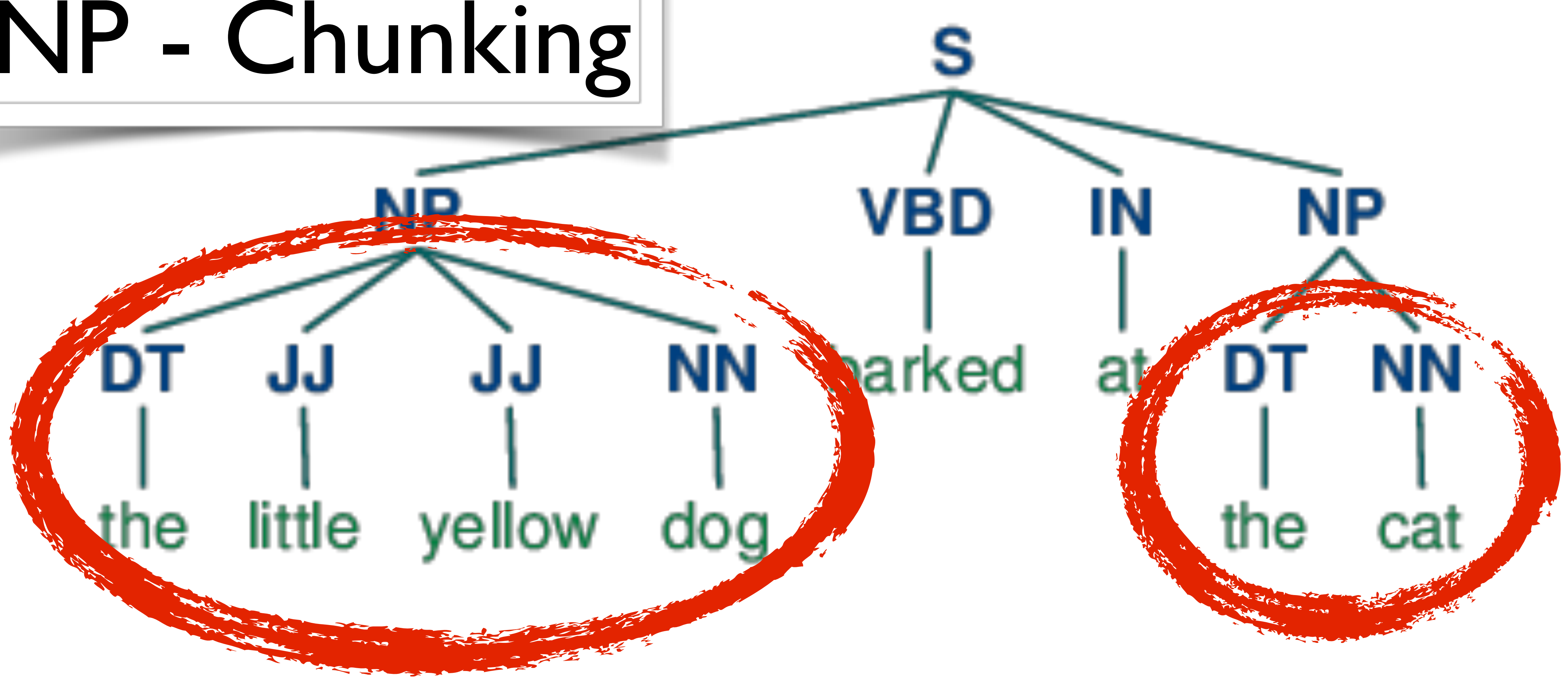


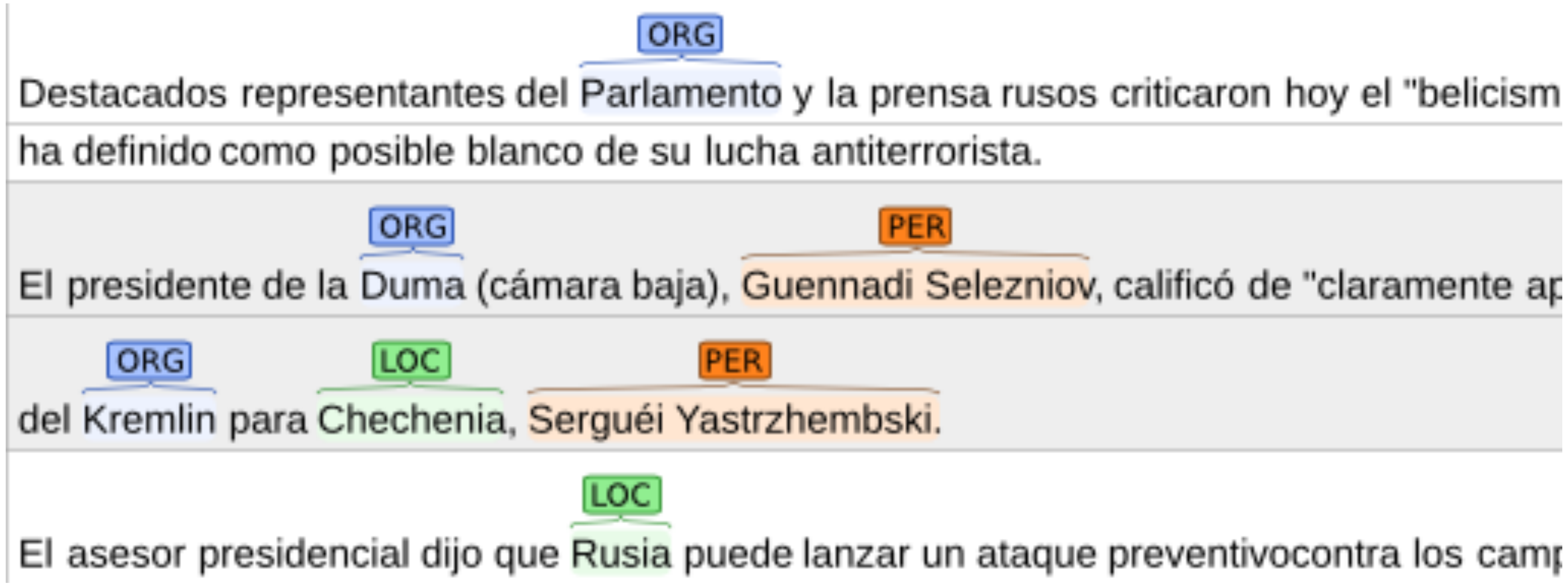


Chunking

- Identify sequences of non-overlapping labels

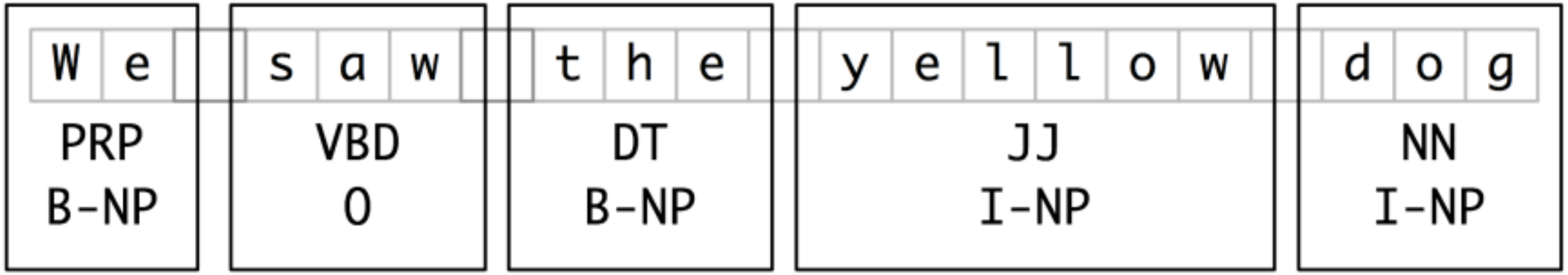
NP - Chunking





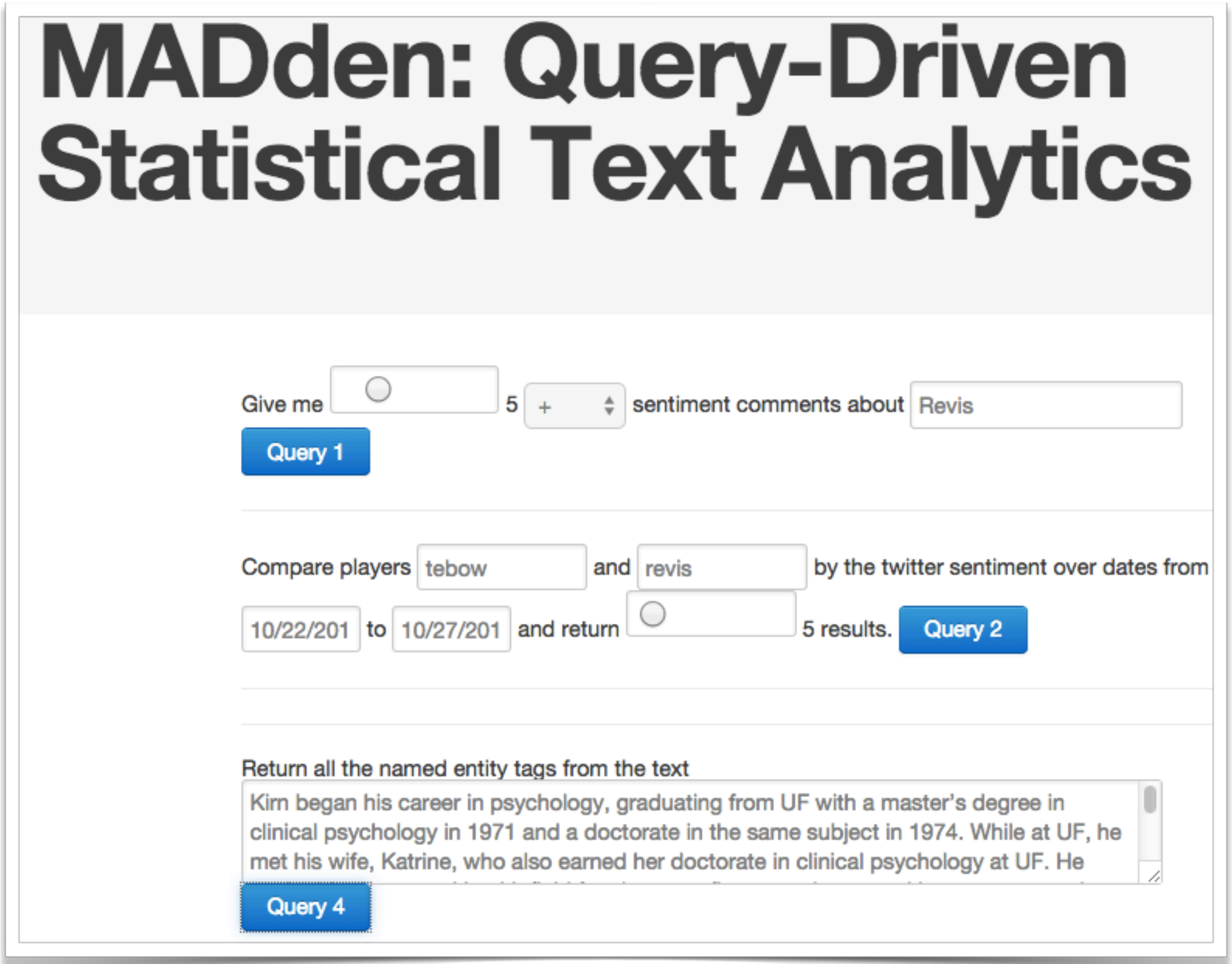
Chunking

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

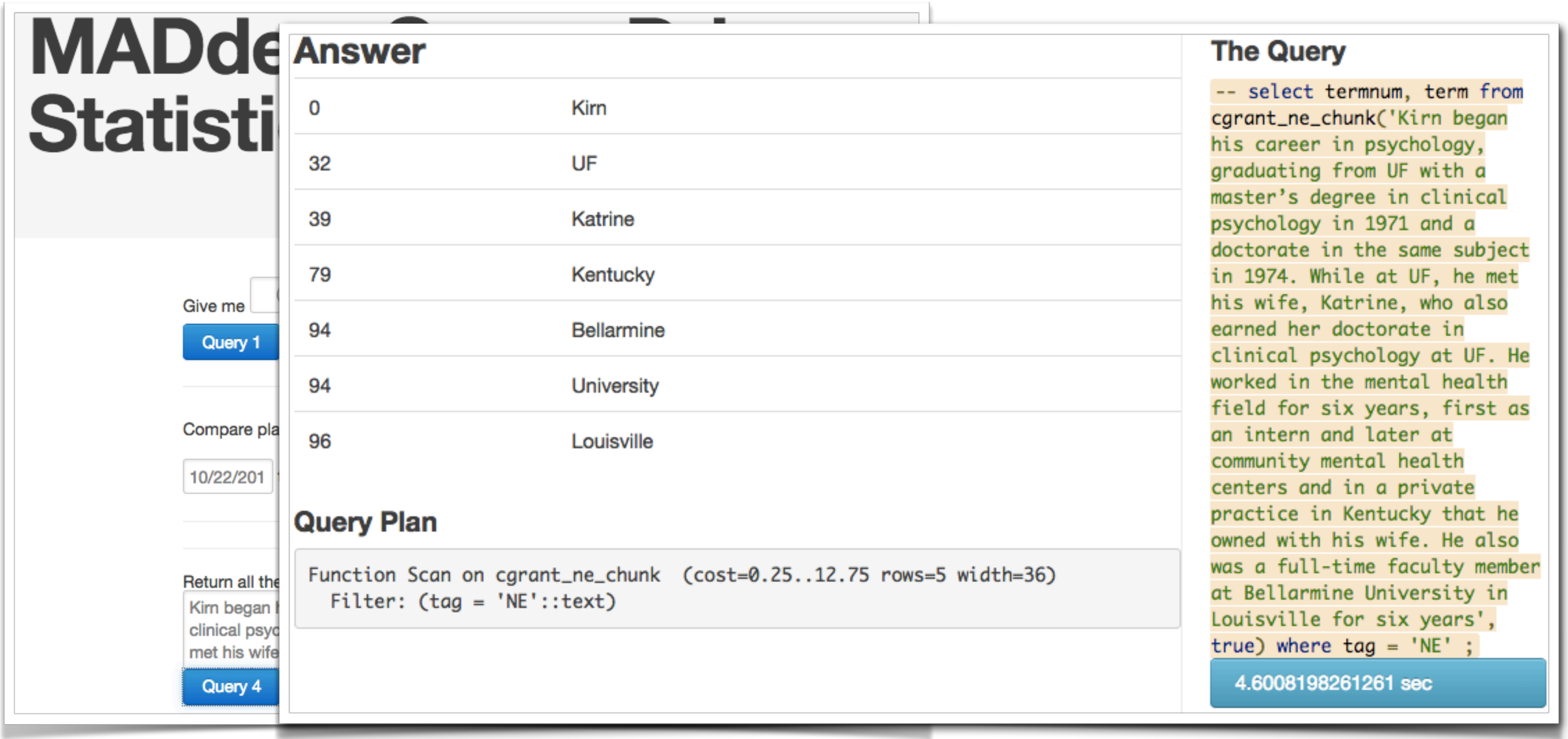


MADden

MADden


MADden: Query-Driven Statistical Text Analytics
Give me
5
+
sentiment comments about
Revis
Query 1
Compare players
tebow
and
revis
by the twitter sentiment over dates from
10/22/201
to
10/27/201
and return
5 results.
Query 2
Return all the named entity tags from the text
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
Query 4

MADden



MADde
Statisti
Give me
Query 1
Compare pla
10/22/201
Return all the
Kirn began
clinical psyc
met his wife
Query 4
Answer
0 Kirn
32 UF
39 Katrine
79 Kentucky
94 Bellarmine
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

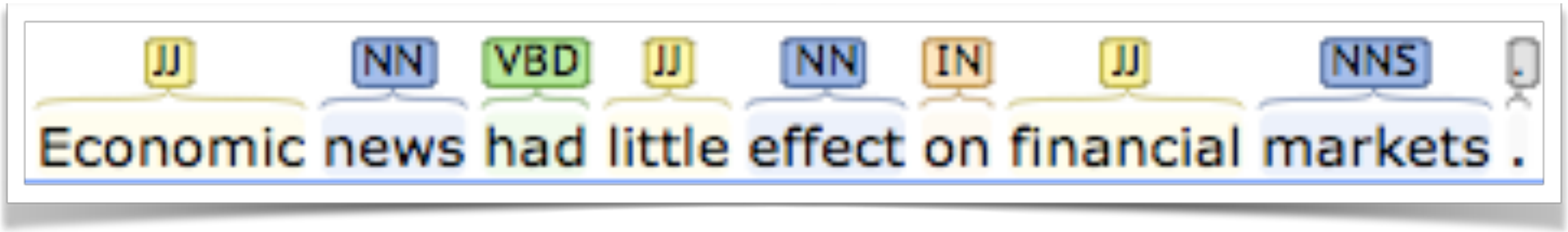
Dependency Parsing

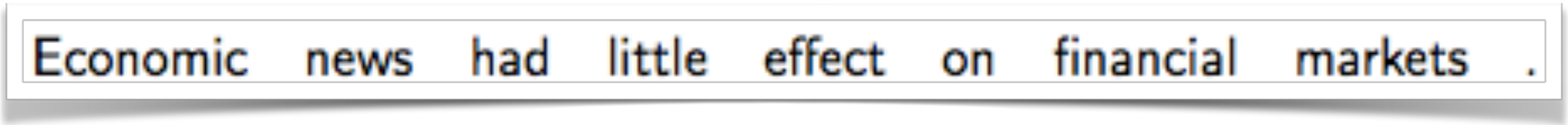
- A graph depicting the relationship between a word (head) and its dependents.
  - Starts with a verb and finds the related subject and object.
  - Useful in understanding phrases
  - Similar to chunking
  - Very close to **semantic relationships**
  - Link grammar is the most notable implementation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

Dependency Parsing

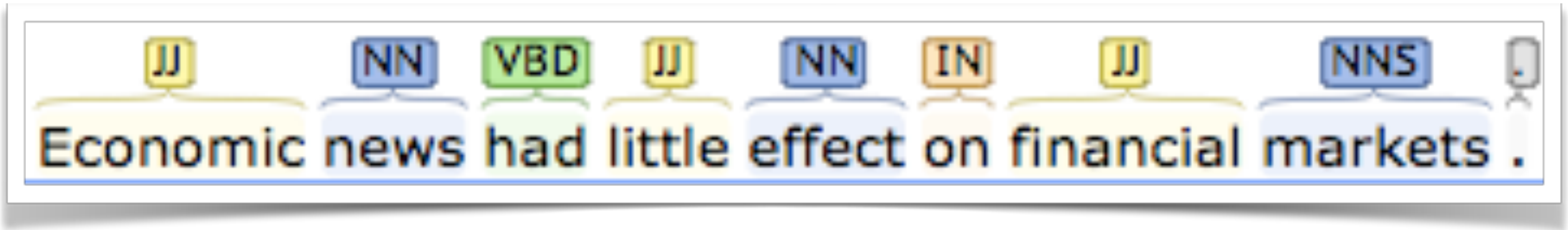
Economic news had little effect on financial markets .

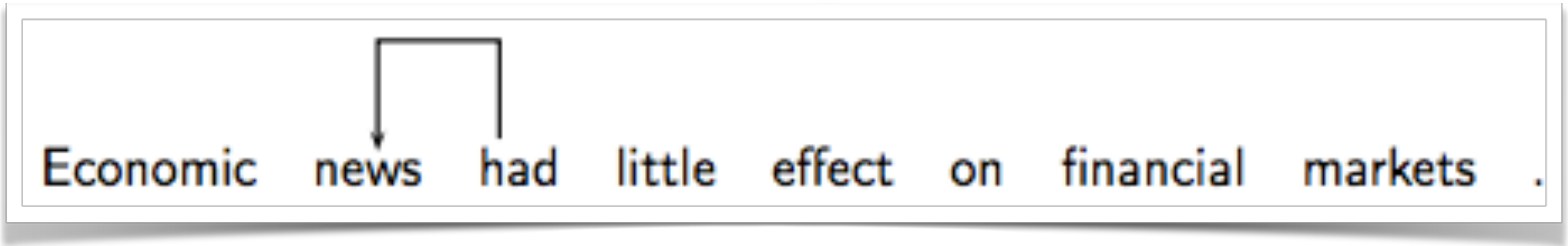
Dependency Parsing



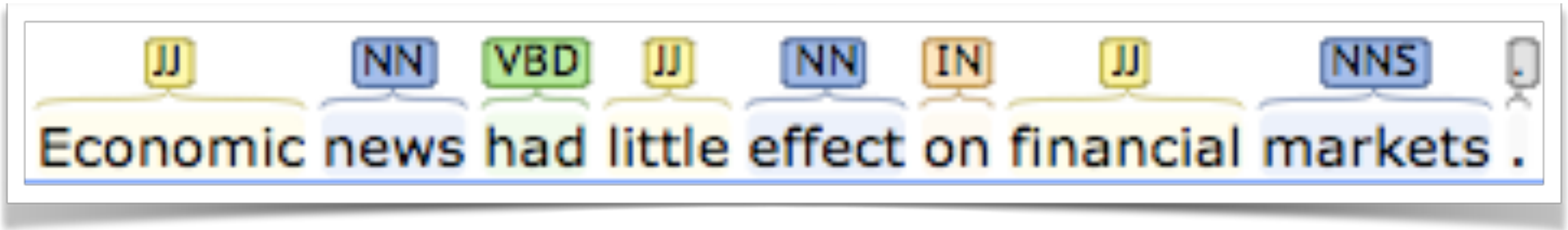


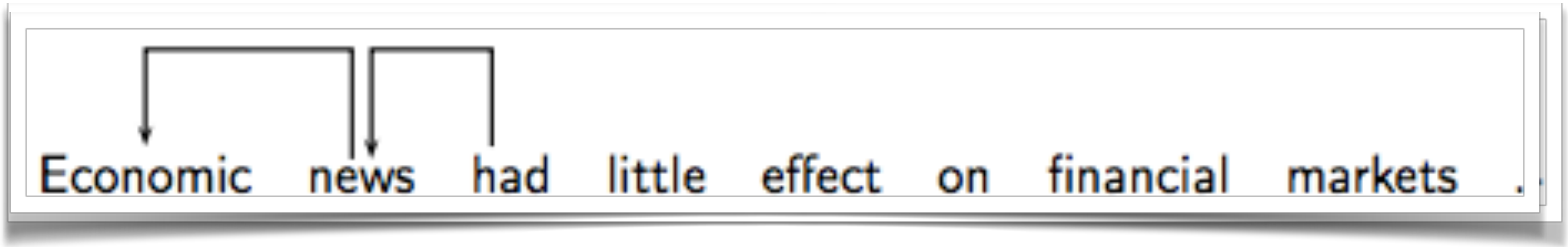
Dependency Parsing



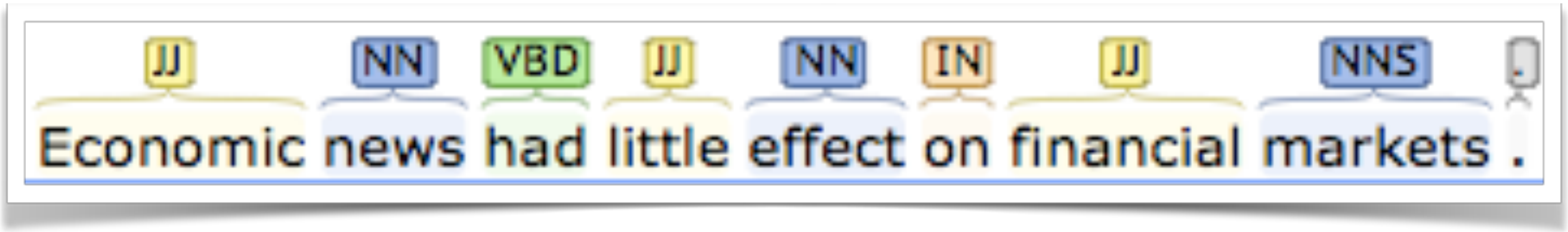


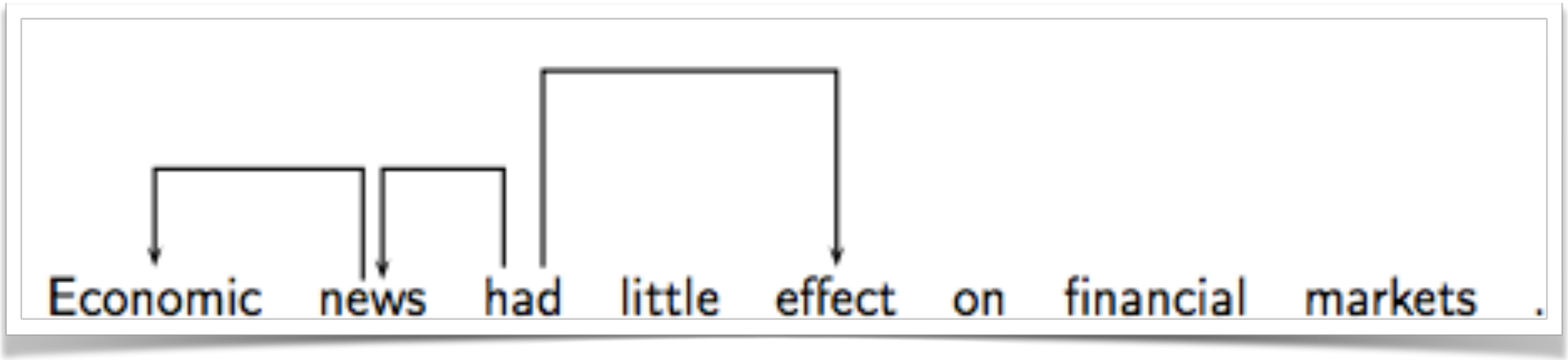
Dependency Parsing



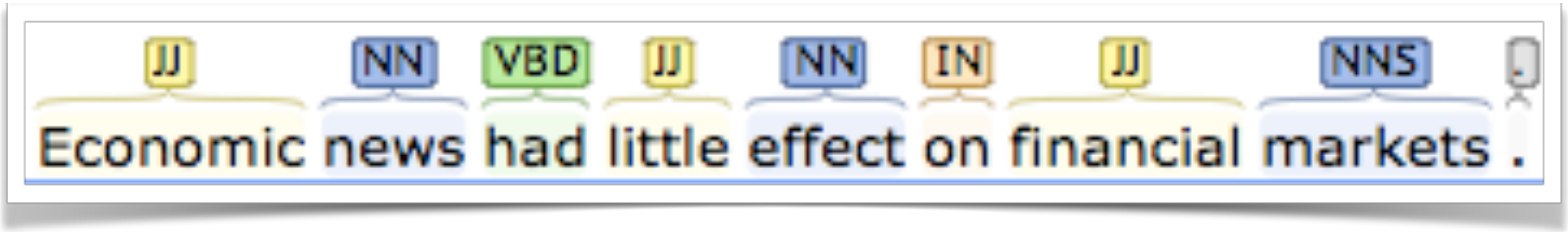


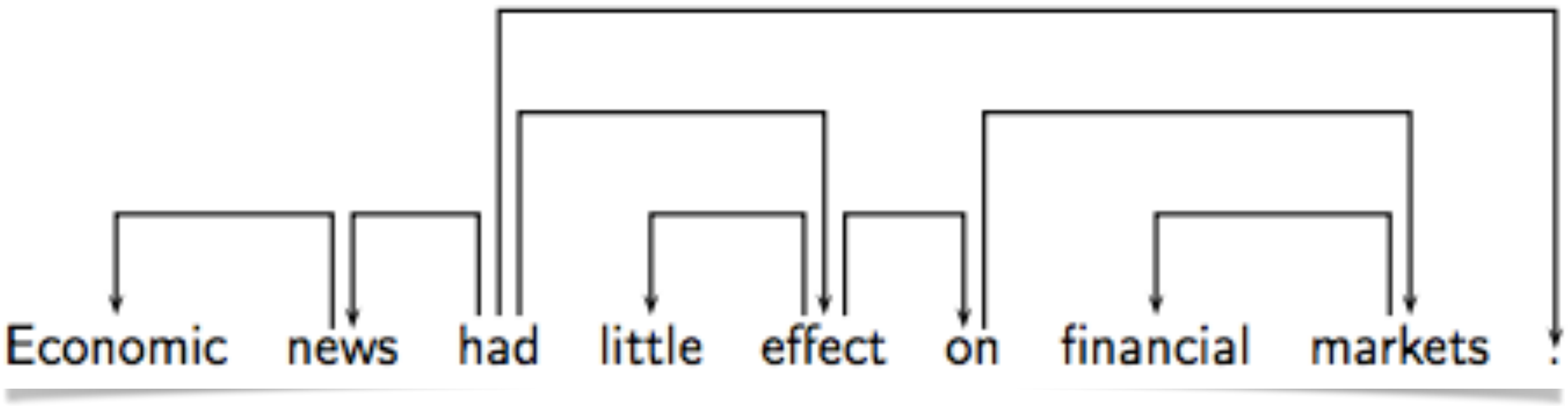
Dependency Parsing



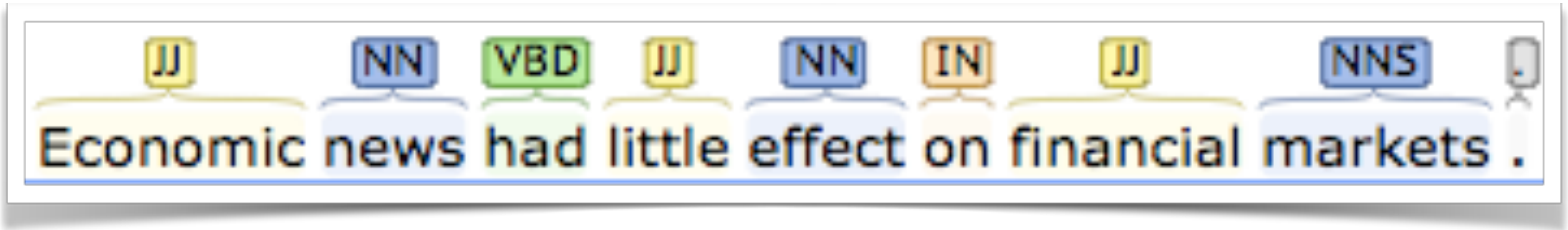


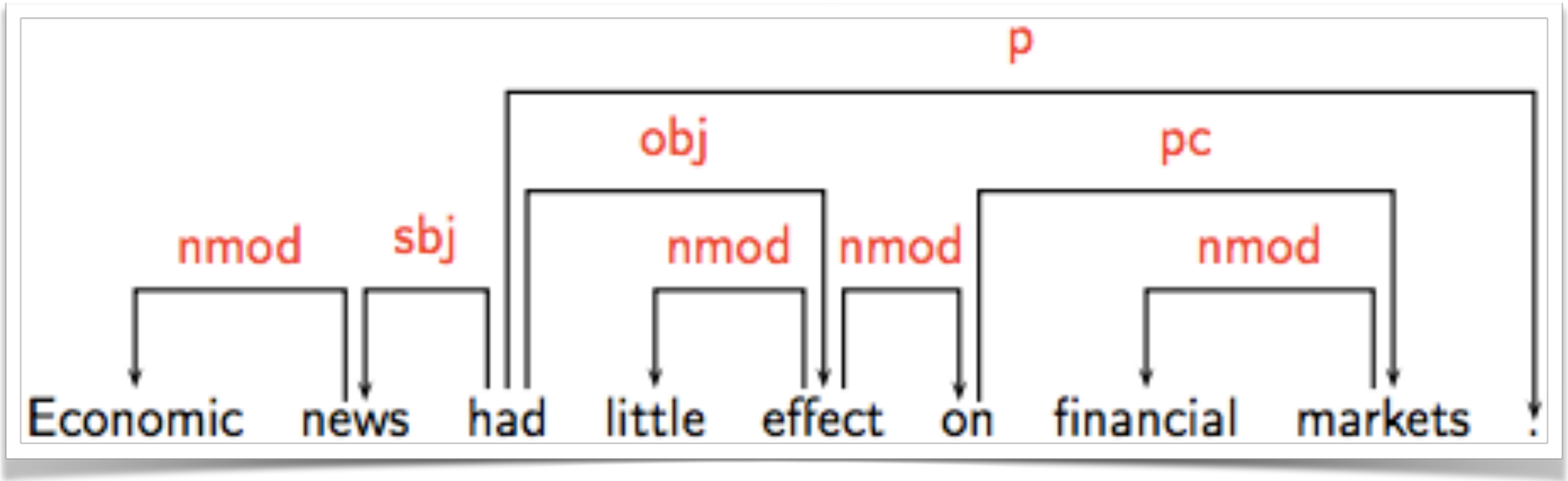
Dependency Parsing





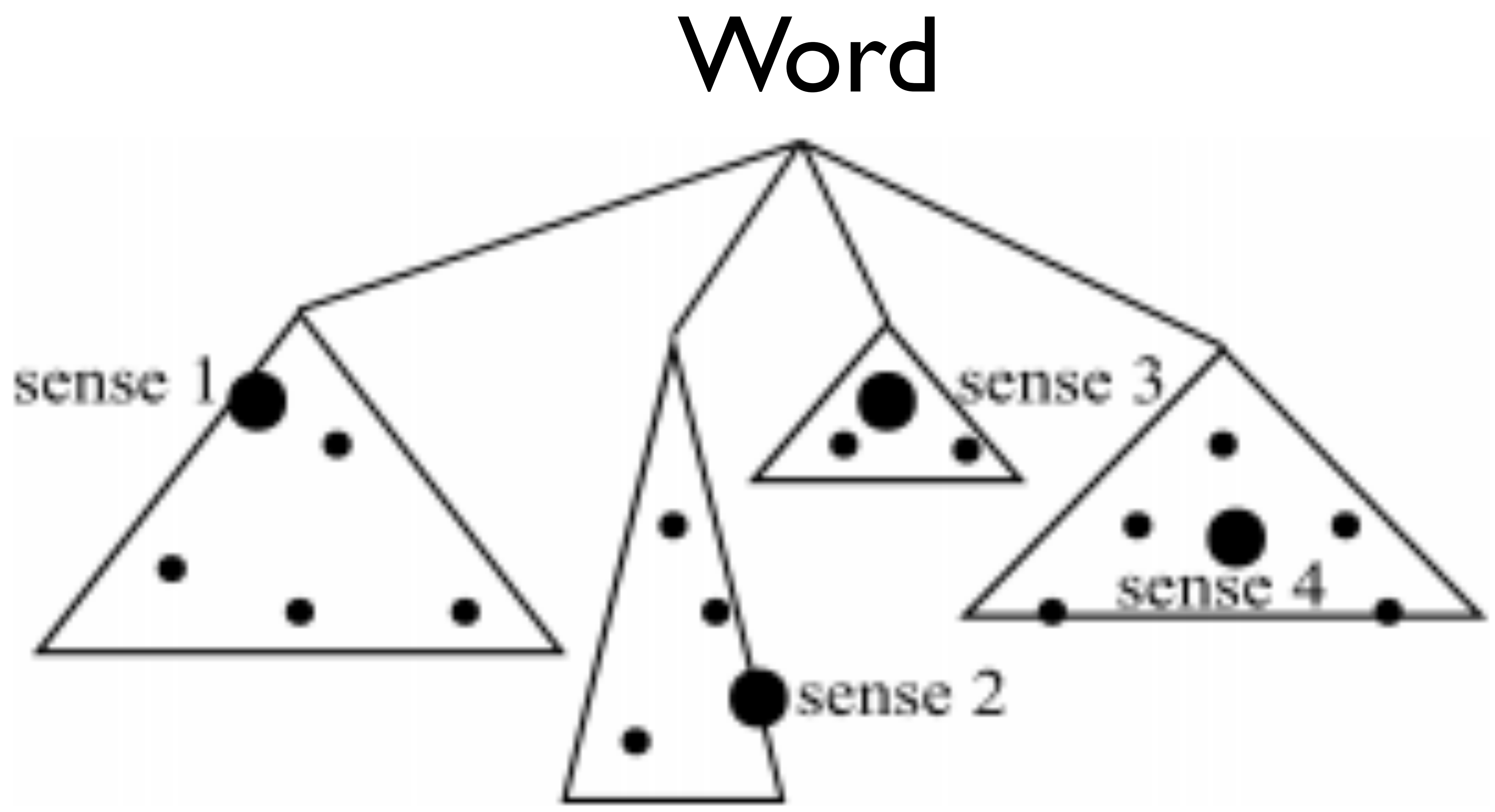
Dependency Parsing





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.

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.

Co-Reference Resolution

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

Entity/Co-reference Chains

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

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.

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**