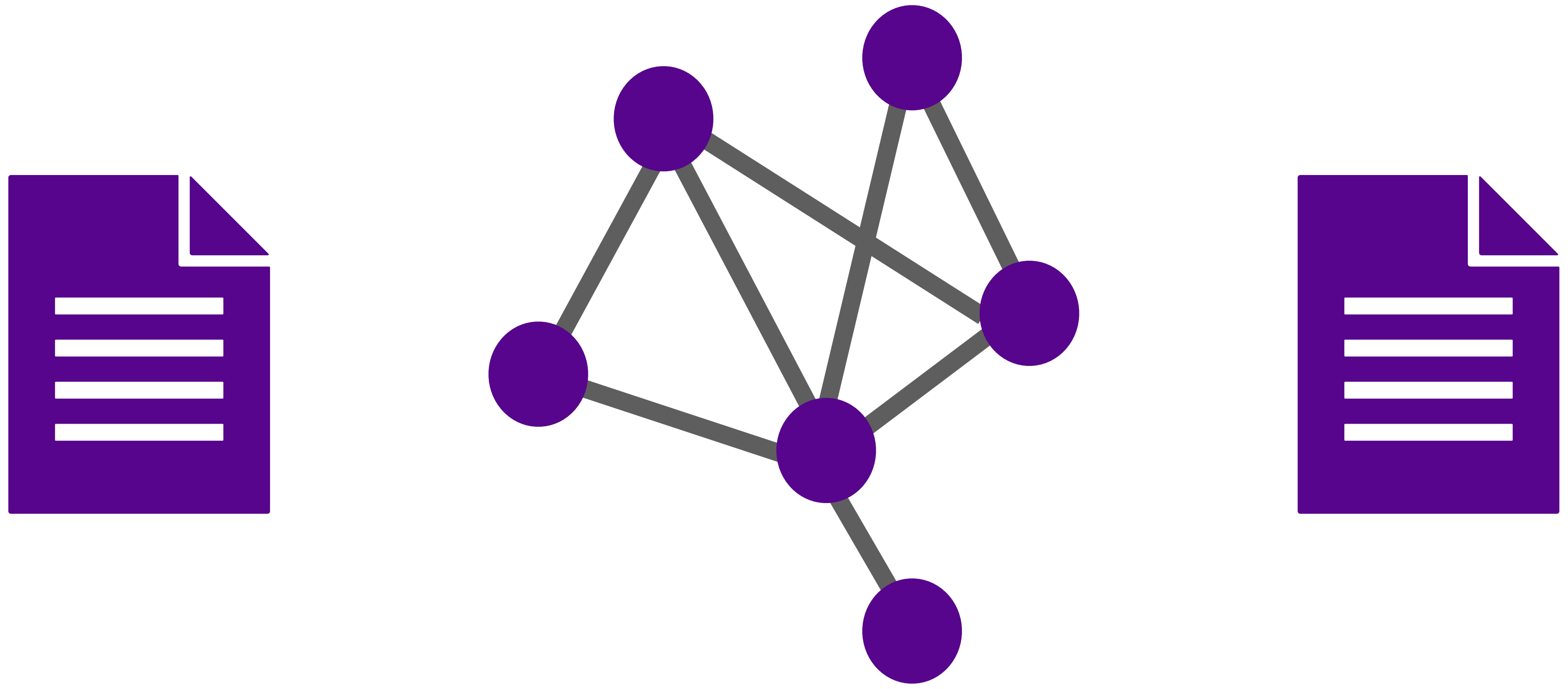
PolNet 2022 Workshop:

Methods at the Intersection of Network and Text Analysis



Sarah Shugars

NYU —> Rutgers

sarah.shugars@rutgers.edu

they/them

Workshop Outline

Part 0: Logistics

- Goals, expectations, & introductions

Part 1: Theory

- Conceptual approaches to working with text & networks

(Break)

Part 2: Practice

- Live coding (in Python)

Materials

All materials available at:
https://github.com/sshugars/PolNet2022

Goals

- Learn cool stuff
- Meet cool people
- Have fun
- Other goals?
    - ➡ Unmute or put them in the chat!

Expectations

- NO prior experience or training expected
  - ➡ I EXPECT you to ask questions!
  - ➡ Asking questions is how you learn

Please keep in touch!
sarah.shugars@rutgers.edu

- Being interdisciplinary means continually surrounding yourself with smart people who have expertise beyond your own
  - ➡ YOU are also incredibly smart and extremely capable!
  - ➡ Every one of you knows something the rest of us don't know
- Thank you for contributing to this community!

Introductions

- Too many people to do proper introductions...
- OPTIONAL: Put your info in spreadsheet – (shared only with workshop participants)

About Me

Faculty Fellow
New York University

Incoming Assistant Professor
Rutgers University

Sarah Shugars

Research

- Political talk & discourse
- Social Media
- Civic infrastructure

Methods

- Text-as-data / NLP
- Network analysis
- Computational social science

Personal & Contact

Pronouns: they/them
Twitter: @Shugars
Email: sarah.shugars@rutgers.edu

The Plan

Part 1: **Theory**

➡ Key concepts, methods, and approaches

Break

Part 2: **Practice**

➡ Live coding (in Python)

Part 1: **Theory**

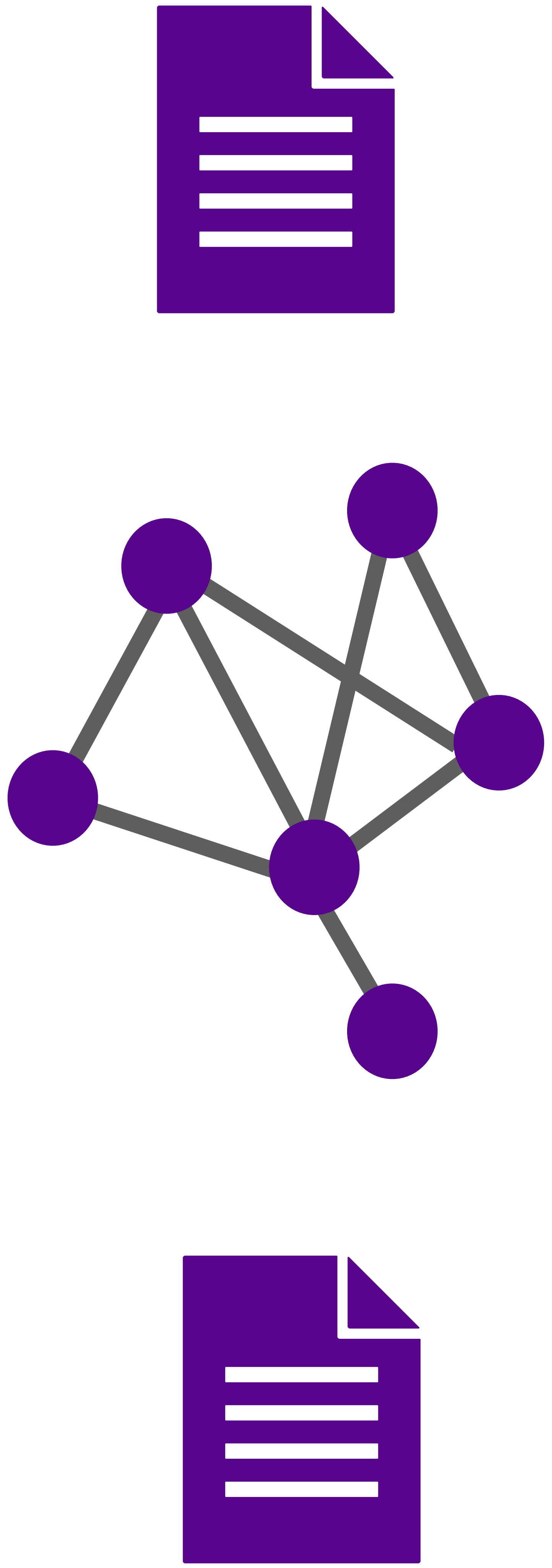
Why text?

Why networks?

What are these things?

What are we doing?

What is happening??



Computational Social Science Hot Take:

Computational execution is the easy part,
theory is the hard part!

- "Coding" is relatively easy
  - ➡ Computers are really good at calculating things

Once you learn the syntax, you can duplicate or look it up if needed.

- Making well-informed researcher decisions is hard
  - ➡ Requires theory: ***why*** are you doing something?

A life-long pursuit dependent on research question/context.

Computational Social Science Hot Take:

Computational execution is the easy part,
theory is the hard part!

- "Coding" is relatively easy
  - ➡ Computers are really good at calculating things

Learning to cook.

- Making well-informed researcher decisions is hard
  - ➡ Requires theory: ***why*** are you doing something?

Designing and constructing a kitchen.

Computational Social Science Hot Take:

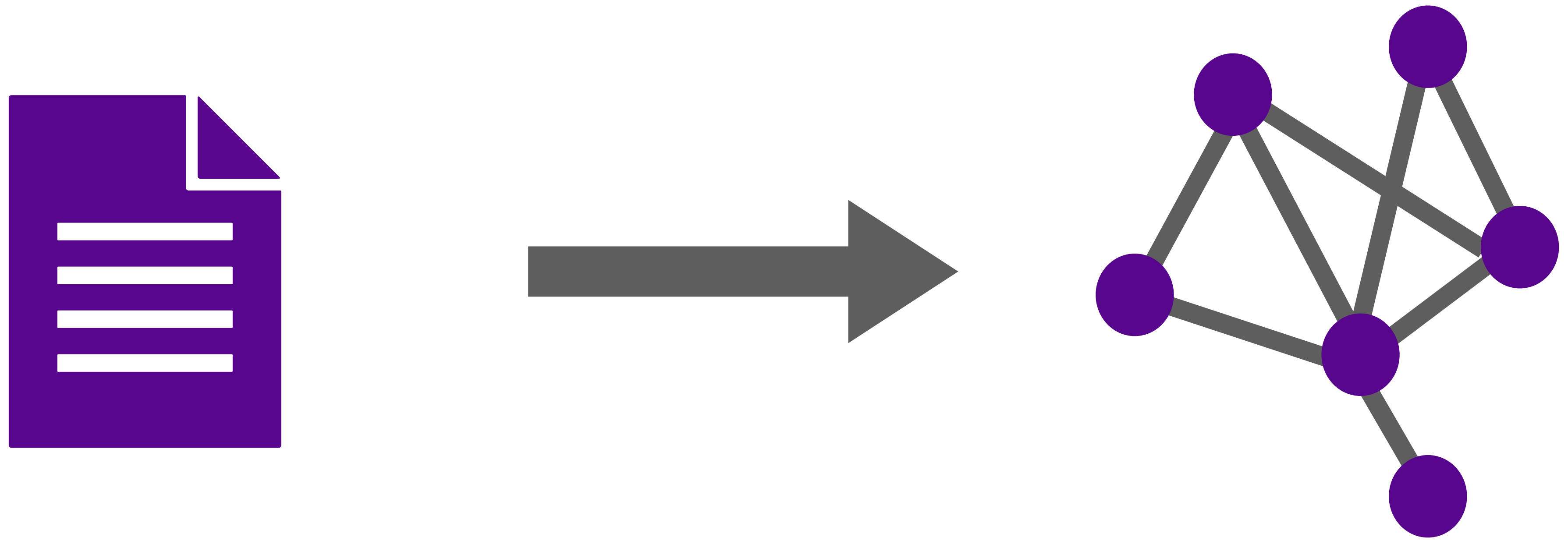
Computational execution is the easy part,
theory is the hard part!

This might seem scary or overwhelming at first...

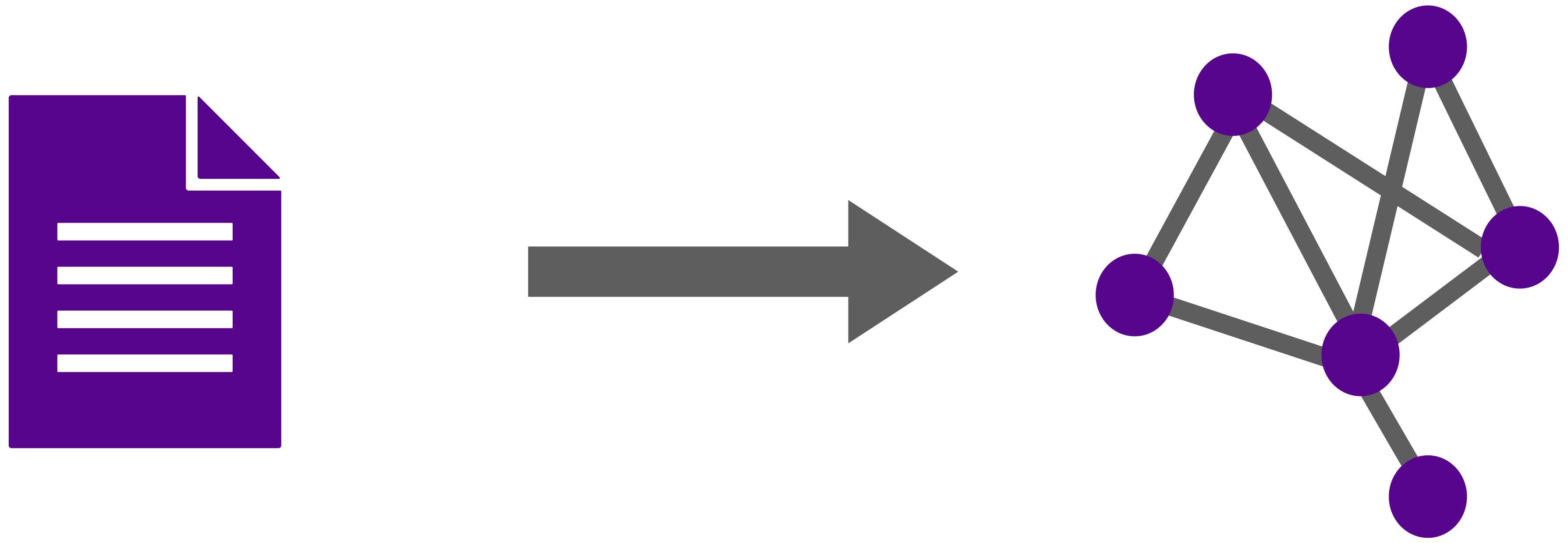
BUT...

Never forget that your PhD training will make you an expert in this.
And that training is invaluable.

Q: What is the best way to go from text(s) to network(s)?



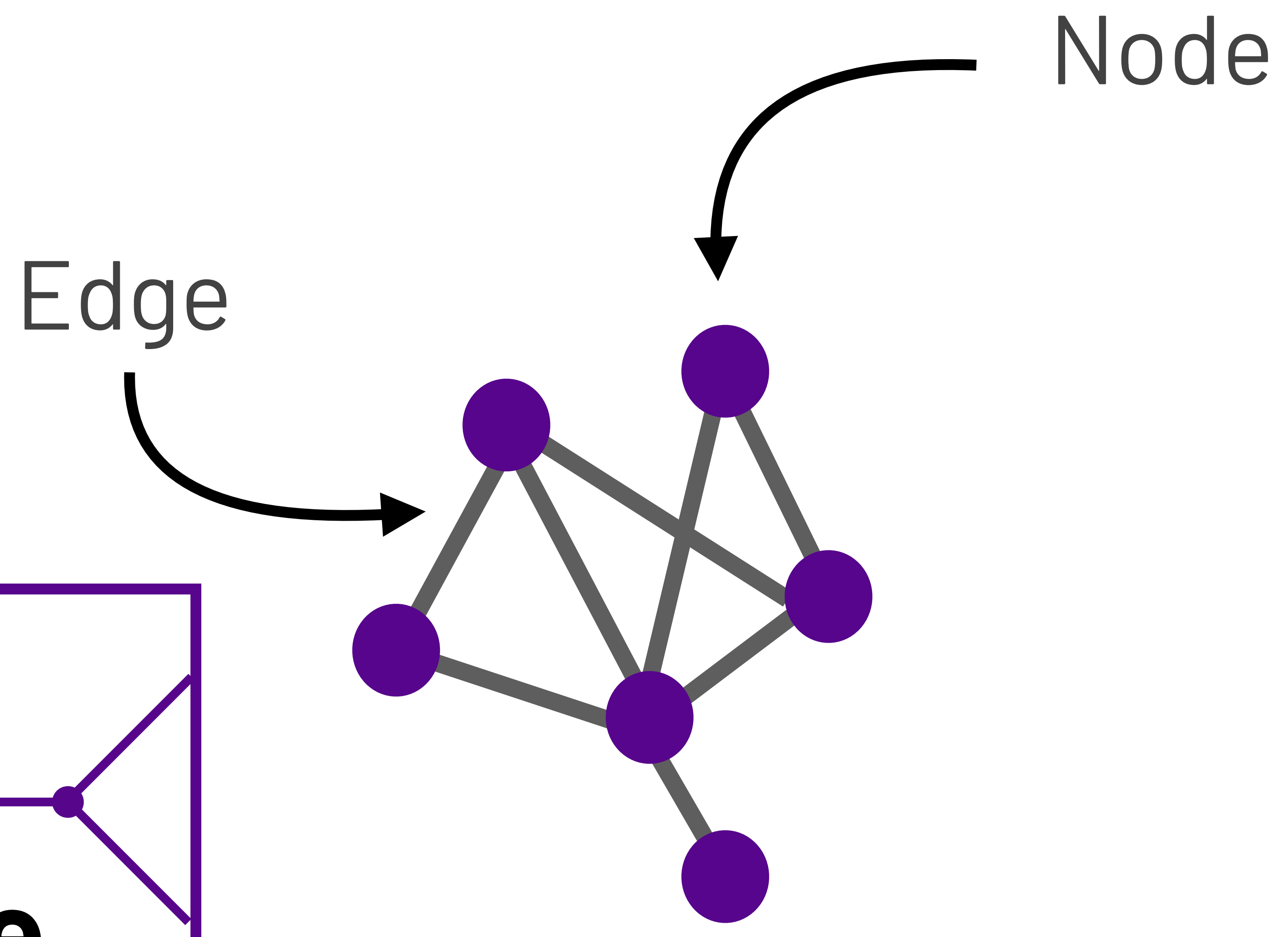
Q: What is the best way to go from text(s) to network(s)?



A: There is no singular best way.
It depends on your research question and requires theory-driven modeling choices.

Networks 101

- Networks are collections of things connected to other things
  - ➡ "Things" called nodes or vertices
  - ➡ "Connections" called edges



Pro Tip:

Network modeling is needed any time the connections between things are as or more important than the things themselves.

Example: Senario 1

- You survey a nationally representative, random sample of 2000 Americans

➡ **Q:** Can you treat their responses as **independent**?

➡ **A:** Yes. Very low probability respondents know / influence each other

Pro Tip:

Network modeling is needed any time the connections between things are as or more important than the things themselves.

Example: Senario 1

- You survey a nationally representative, random sample of 2000 Americans

➡ **Q:** Can you treat their responses as **connected**?

➡ **A:** Yes. For example, could look at shared media consumption.

Note: Whether this is helpful or not depends on your research question. Just because you can model something as a network, doesn't mean you need to.

Pro Tip:

Network modeling is needed any time the connections between things are as or more important than the things themselves.

Example: Senario 2

- You randomly assign members of a small, tight-knit community to treatment and control groups and examine adoption rates of a new technology

➡ **Q:** Can you assume the treatment and control outcomes are **independent**?

➡ **A:** No*. People will talk to each other and you might have spillover effects where treatment subjects influence control subjects

Pro Tip:

Network modeling is needed any time the connections between things are as or more important than the things themselves.

* Maybe, depending on research question and nature of treatment?

Example: Senario 2

- You randomly assign members of a small, tight-knit community to treatment and control groups and examine adoption rates of a new technology

➡ **Q:** Can you assume the treatment and control outcomes are **independent**?

➡ **A:** No*. People will talk to each other and you might have spillover effects where treatment subjects influence control subjects

Note: Sometimes considering the network is *essential* to your research question.

Pro Tip:

Network modeling is needed any time the connections between things are as or more important than the things themselves.

Text analysis 101

- Documents are collections of words*. They may:
  - ➡ Have structure (eg, grammatical rules)
  - ➡ Intend to convey something (meaning, information, emotion, etc)
- A single document is a "text" or "corpus"
- Multiple documents are "texts" or "corpora"

* What counts as a "word" can be very broad. 🎉👍💯

A "document" may be:

- A tweet
- A speech
- An article
- A book
- A chapter
- A paragraph
- Everything said by a given person in a given conversation
- Every article from a given journal

Thinking about texts and networks

Pro Tip:

The first step for any network analysis is figuring out: what are your nodes and what are your edges?

- A network is a **model** — you have to make choices about how you are modeling
  - ➡ Ideally you make good, theory-driven choices!
  - ➡ Generally no perfectly “right” way model
  - ➡ But, the more you do it, the better you will get at developing useful models

Modeling texts as networks

- What are your nodes?
  - ➡ Documents
  - ➡ Words
  - ➡ Concepts
  - ➡ Authors

- What are your edges?
  - ➡ Co-occurrence
  - ➡ "Similarity"
  - ➡ Grammatical structure

What is your unit of analysis? At what *scale* do you want to examine/compare?

What, conceptually, are the *connections* of interest?

Pro Tip:

The first step for any network analysis is figuring out: what are your nodes and what are your edges?

Modeling texts as networks

- What are your nodes?
  - ➡ Documents
  - ➡ Words
  - ➡ Concepts
  - ➡ Authors

Can have more than one type of node!

- What are your edges?
  - ➡ Co-occurrence
  - ➡ "Similarity"
  - ➡ Grammatical structure

There are essentially an infinite number of ways to model text(s) as network(s)

Pro Tip:

The first step for any network analysis is figuring out: what are your nodes and what are your edges?

Identifying nodes

- Often (but not always!) more conceptually clear as discrete units, ie:
  - A word
  - An author
  - A text with clear bounds
- BUT, can also be more complex:
  - A phrase of arbitrary length
  - A “concept” (what does that even mean??)

Personally, I would ***strongly*** recommend starting out in this territory.

A model that doesn’t recognize multi-word phrases may be slightly less nuanced, but for ***most*** research questions, it won’t actually make that much of a difference.

PSA: Language is complex. Text analysis is fundamentally about simplifying language.

Identifying edges

- Identifying "connections between textual elements" is typically conceptual harder
- But, there can still be some relatively simple ways of doing this:
  - Co-occurrence: two nodes are connected if they occur with the same ***span***
  - For example: occur within the same *k* words; within the same sentence, within the same paragraph, etc.

This is a sample document.

Example 1: Co-occurrence determined by window of k=2.

Edges (undirected):

- This – is
- This – a
- is – a
- is – sample
- a – document
- sample – document

Identifying edges

- Identifying "connections between textual elements" is typically conceptual harder
- But, there can still be some relatively simple ways of doing this:
  - Co-occurrence: two nodes are connected if they occur with the same ***span***
  - For example: occur within the same *k* words; within the same sentence, within the same paragraph, etc.

This is a sample document.

Example 2: Co-occurrence by being in the same sentence (or doc!)

Edges (undirected):

- This – is
- This – a
- This – sample
- This – document
- is – a
- is – sample
- is – document
- a – sample
- a – document
- sample – document

Identifying edges

- Identifying "connections between textual elements" is typically conceptual harder
- But, there can still be some relatively simple ways of doing this:
  - Co-occurrence: two nodes are connected if they occur with the same ***span***
- Can also have more complex definitions of an edge
  - For example, ***similar*** words are connected

Again, I would generally recommend starting with simpler models (ie, co-occurrence)

But! These ideas are also (excuse the pun) connected!

Text Analysis 102: Context and Embeddings

"You shall know a word by the company it keeps"
Firth (1954)

In other words:

Words which are similar occur in similar contexts

I drank a cup of **tea**.

I drank a cup of **coffee**.

Even if I don't know what "tea" and "coffee" mean, I know they are both "things I can drink a cup of"

Text Analysis 102: Context and Embeddings

"You shall know a word by the company it keeps"
Firth (1954)

In other words:

Words which are similar occur in similar contexts

I drank a cup of **tea**.

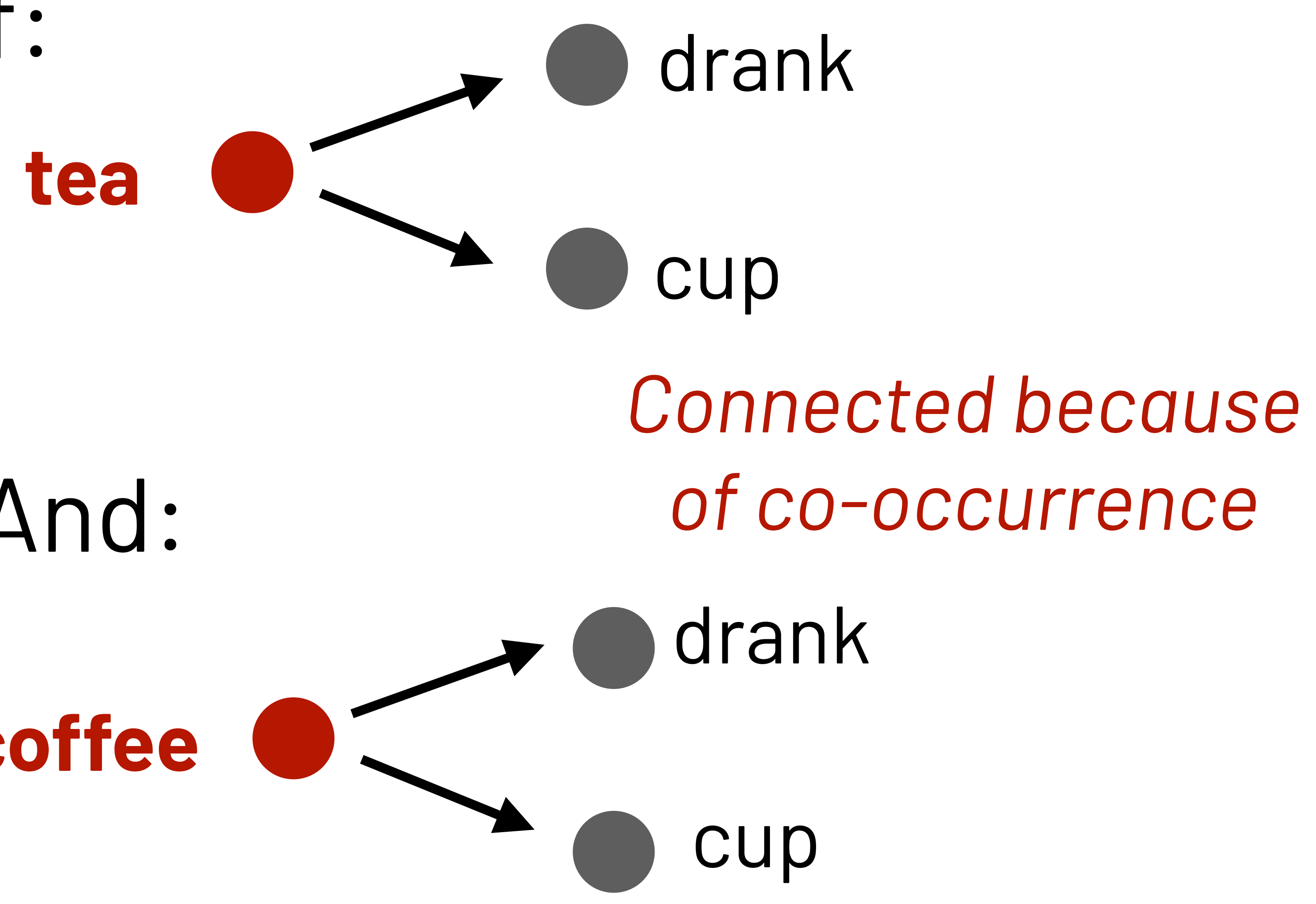
I drank a cup of **coffee**.

Even if I don't know what "tea" and "coffee" mean, I know they are both "things I can drink a cup of"

This is an inherently **networked** idea!!

If:

And:



Then:

"tea" is similar to **"coffee"**

Text Analysis 102: Context and Embeddings

More broadly, imagine I wanted to fill in the blank:

I drank a cup of _____ this morning.

Left context

Right context

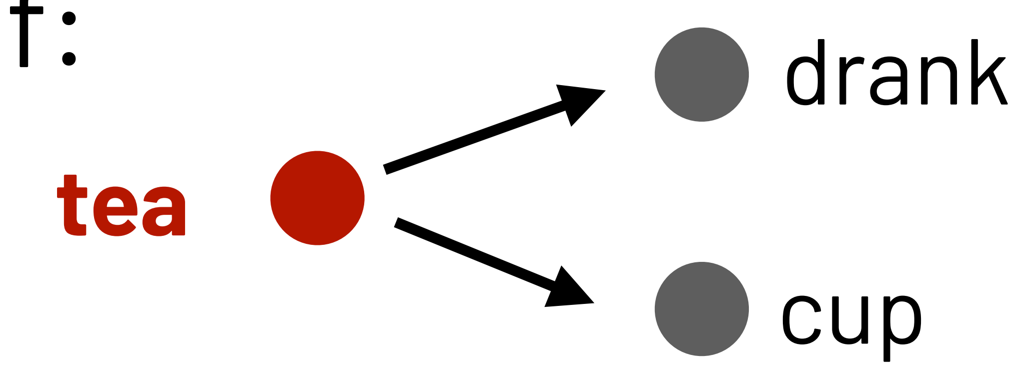
Target word, t

I'll predict my target word, t, is a word which frequently co-occurs with the observed context words:

- Coffee
- Tea
- Water? (Don't usually drink out of cups?)

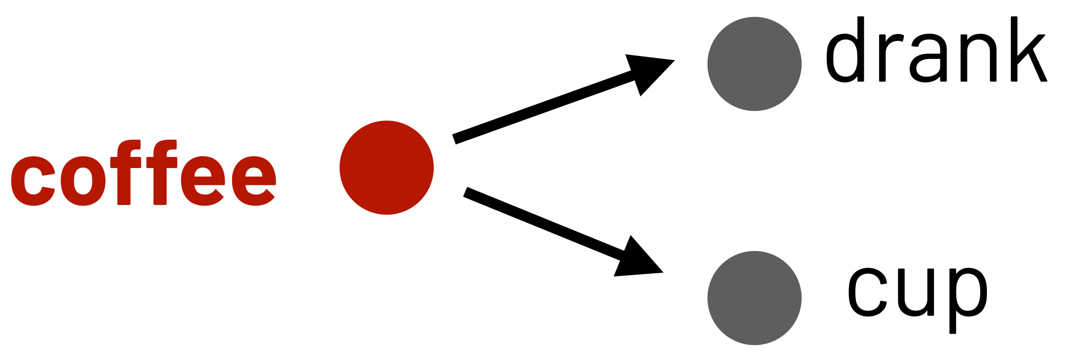
This is an inherently **networked** idea!!

If:



Connected because of co-occurrence

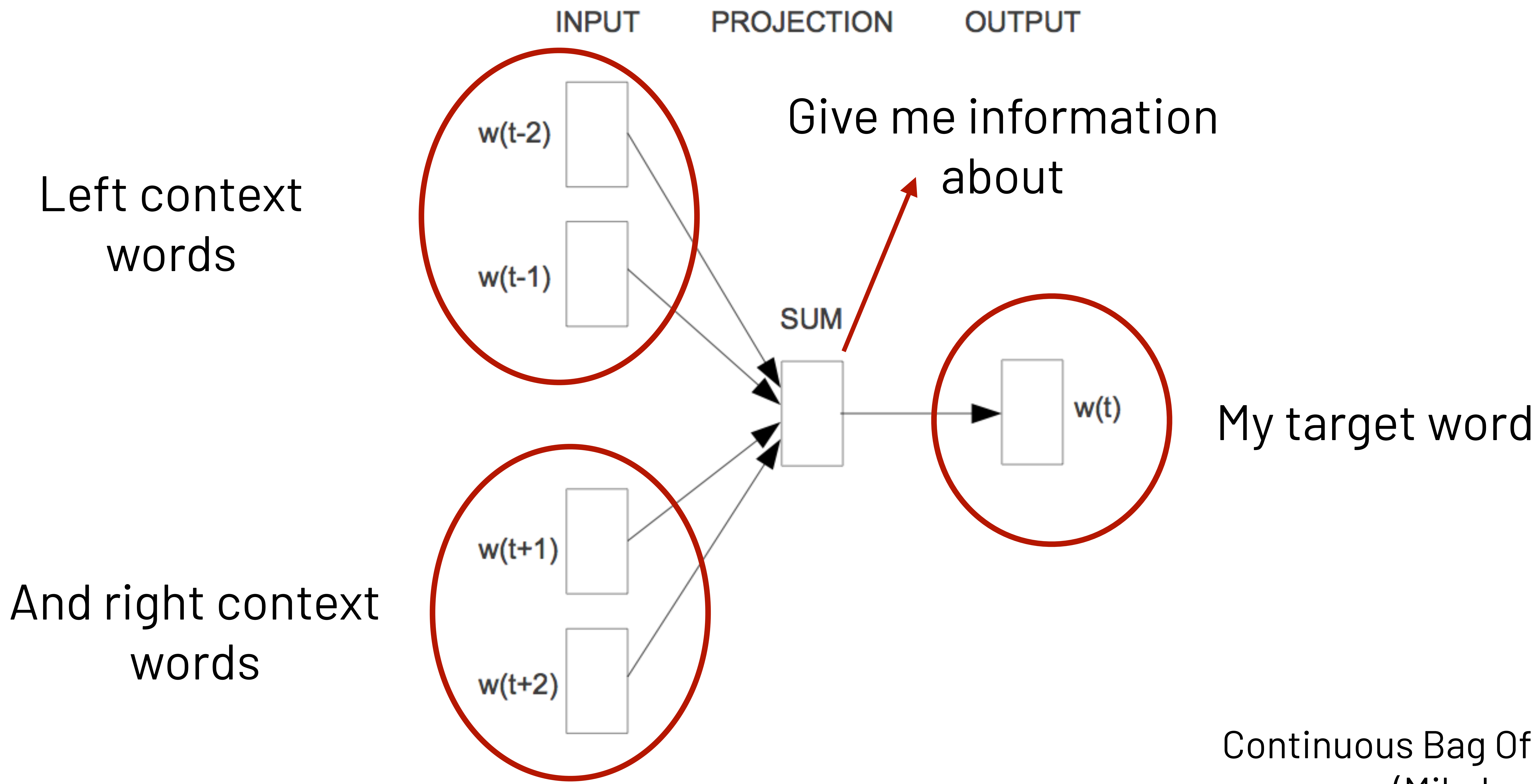
And:



Then:

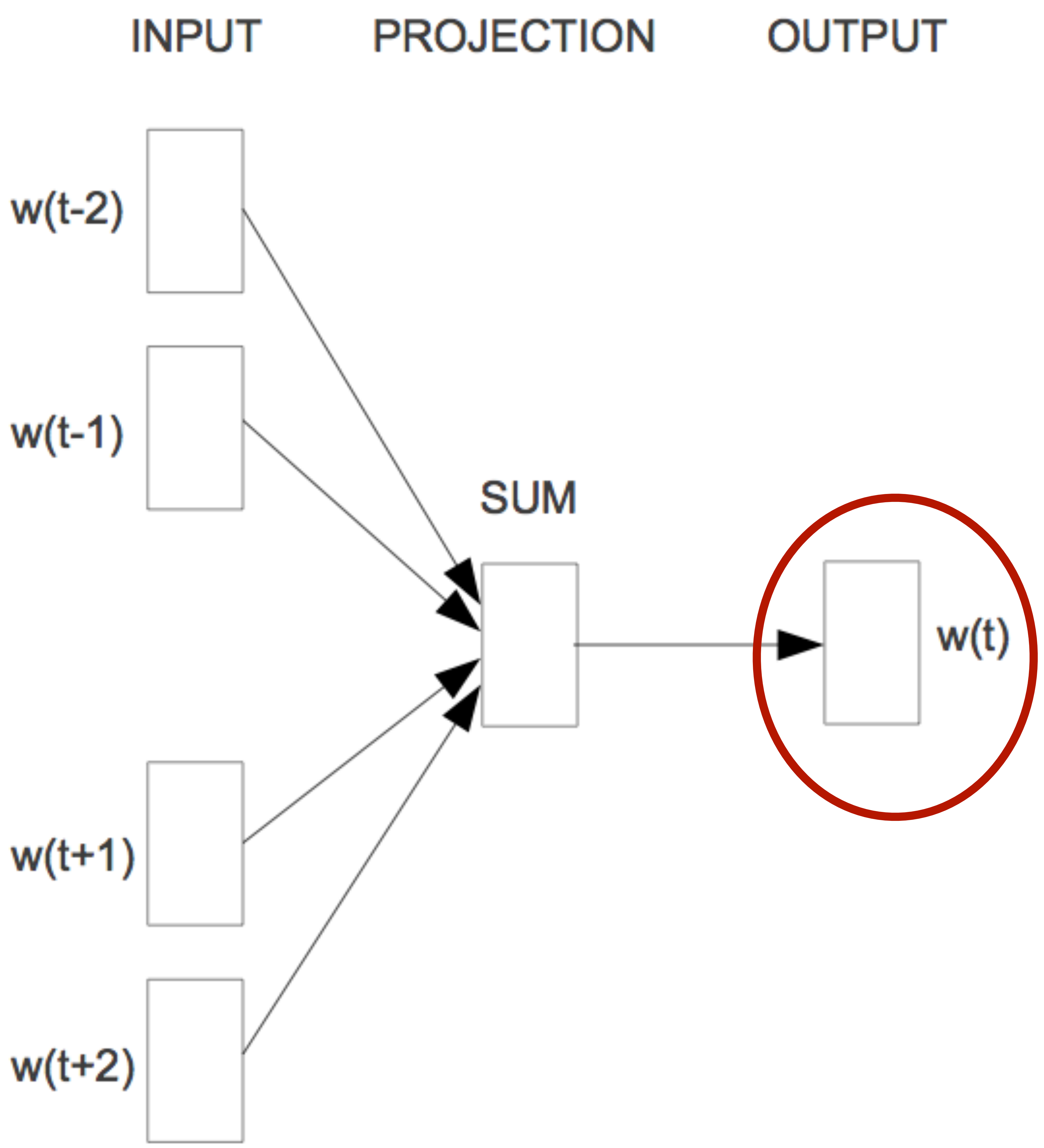
"tea" is similar to **"coffee"**

Text Analysis 102: Context and Embeddings



Continuous Bag Of Words (CBOW) model
(Mikolov et al., 2013)

Text Analysis 102: Context and Embeddings



But, I don't want to do this for just one target word...

I want to do this for all target words!

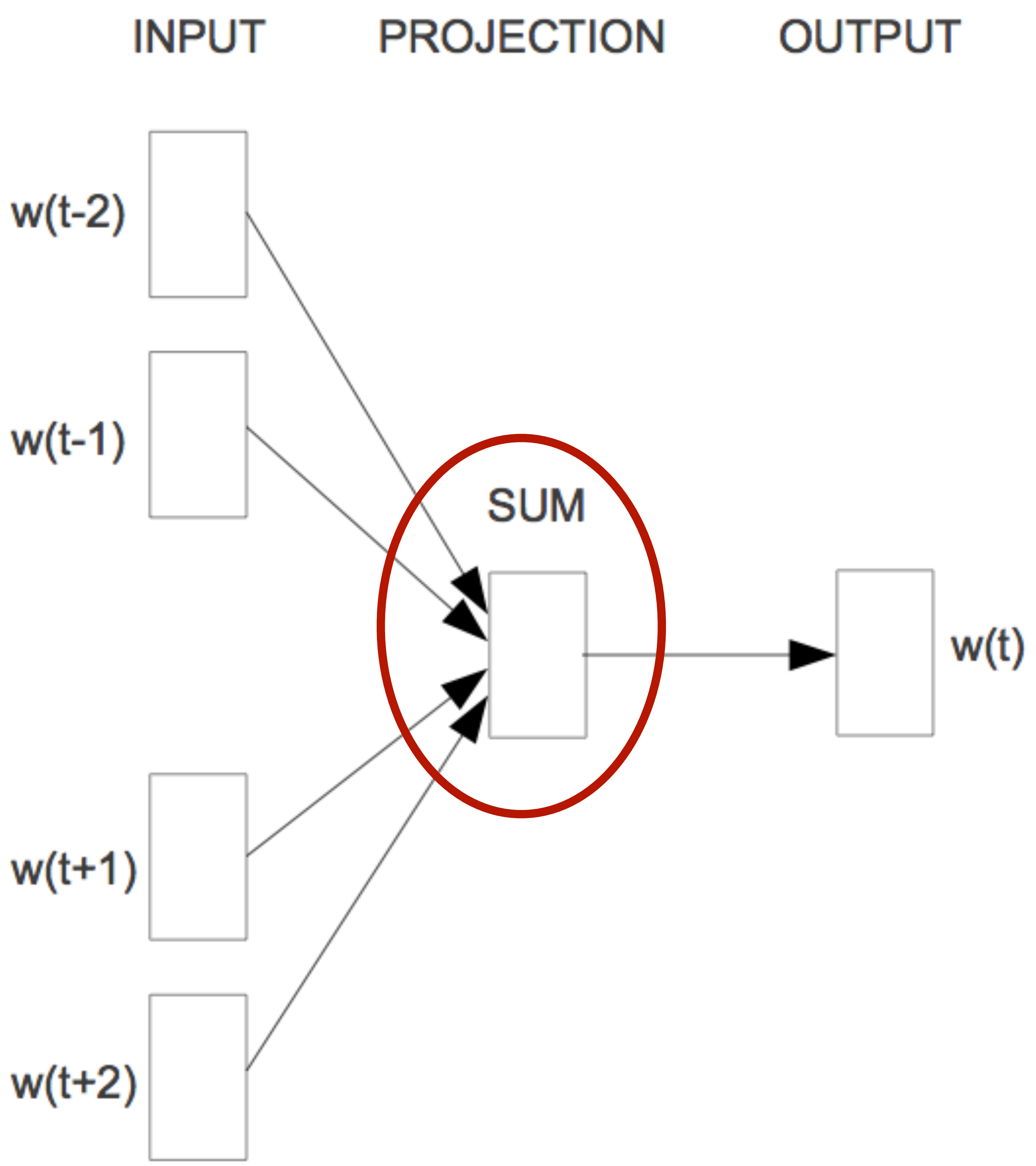
In other words, I want to embed my words in some high dimensional space...

And I want to do this in such a way that

The "embedding" for target word t (ie, the embedding representing that word)

Continuous Bag Of Words (CBOW) model
(Mikolov et al., 2013)

Text Analysis 102: Context and Embeddings



But, I don't want to do this for just one target word...

I want to do this for all target words!

In other words, I want to embed my words in some high dimensional space...

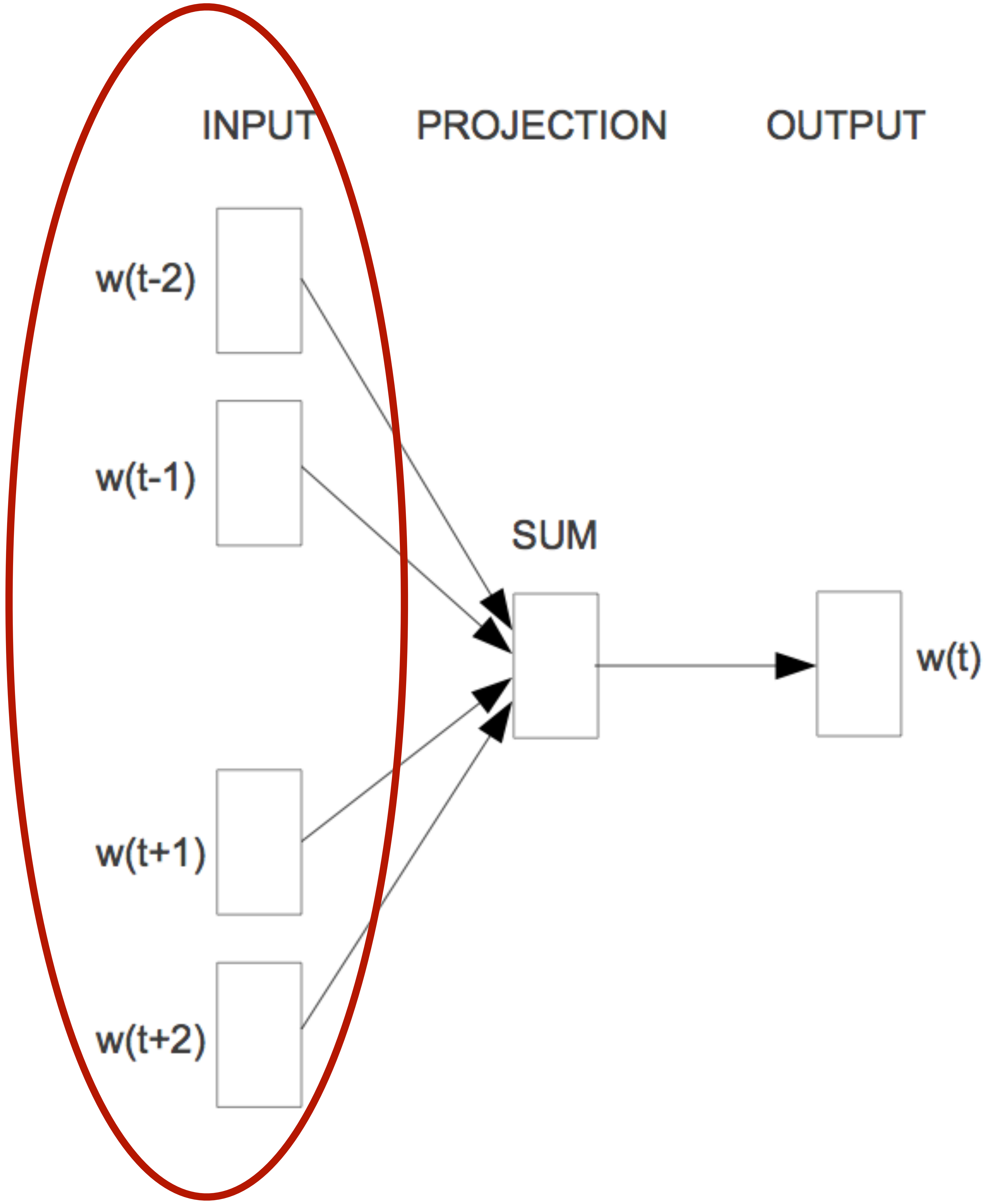
And I want to do this in such a way that

The "embedding" for target word t (ie, the embedding representing that word)

Is the sum of...

Continuous Bag Of Words (CBOW) model
(Mikolov et al., 2013)

Text Analysis 102: Context and Embeddings



But, I don't want to do this for just one target word...

I want to do this for all target words!

In other words, I want to embed my words in some high dimensional space...

And I want to do this in such a way that

The "embedding" for target word t (ie, the embedding representing that word)

Is the sum of...

Of all embeddings (vectors) of all words in context window

Continuous Bag Of Words (CBOW) model (Mikolov et al., 2013)

Text Analysis 102: Context and Embeddings

Specifically, I'll:

Choose length-n embeddings such that I maximize:

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T}}_{\text{Averaged over all words t}} \log \underbrace{p(w_t \mid w_{t-n}, \ldots w_{t-1}, w_{t+1}, \ldots, w_{t+n})}_{\text{Probability of seeing word t given context words}}$$

Averaged over all words t

Probability of seeing word t given context words

Don't be scared of math! It's just a concise way to summarize intuition. If it doesn't help you today, don't worry about it!

But, I don't want to do this for just one target word...

I want to do this for all target words!

In other words, I want to embed my words in some high dimensional space...

And I want to do this in such a way that

The "embedding" for target word t (ie, the embedding representing that word)

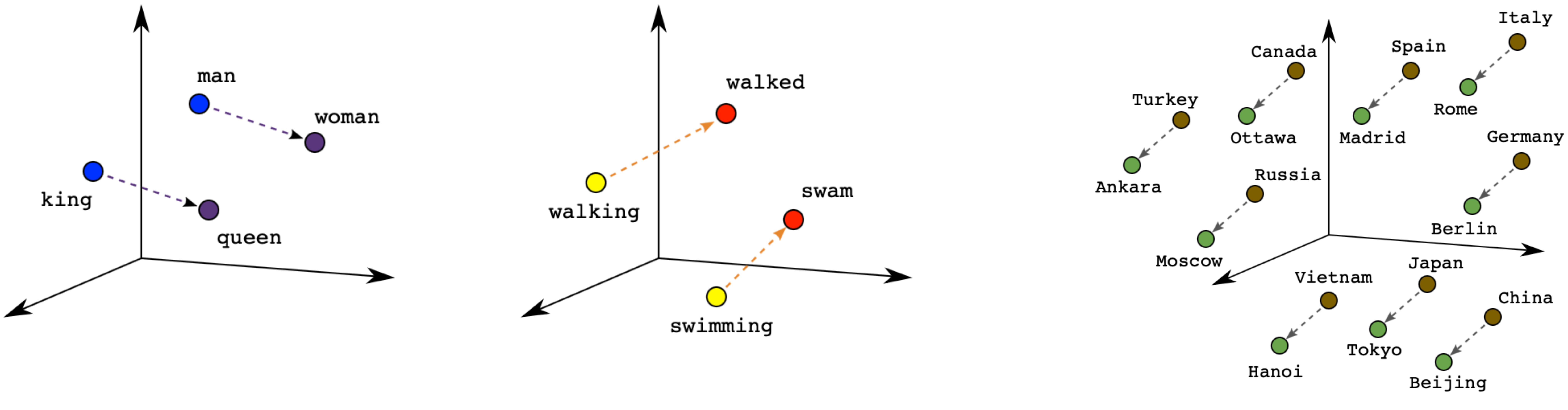
Is the sum of...

Of all embeddings (vectors) of all words in context window

Continuous Bag Of Words (CBOW) model
(Mikolov et al., 2013)

Text Analysis 102: Context and Embeddings

- Result is that words which appear in similar contexts will have similar embeddings
- Embeddings will reflect bias of training language



Male-Female

Verb Tense

Country-Capital

An example so classic I can't give you a proper citation

Text Analysis 102: Context and Embeddings

- Result of this model is that words which appear in similar contexts will have similar embeddings
- Unsupervised method – relies on word proximity having meaning
- Remarkably good at generating high dimensional representations of words
- Embeddings typically around 300 dimensions
- ***Can*** train your own models ***or*** use models that have been pre-trained on large corpora (eg, all news articles in Google)

Text Analysis 102: Context and Embeddings

Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research

Pedro L. Rodriguez, Vanderbilt University
Arthur Spirling, New York University

Word embeddings are becoming popular for political science research, yet we know little about their properties and performance. To help scholars seeking to use these techniques, we explore the effects of key parameter choices—including context window length, embedding vector dimensions, and pretrained versus locally fit variants—on the efficiency and quality of inferences possible with these models. Reassuringly we show that results are generally robust to such choices for political corpora of various sizes and in various languages. Beyond reporting extensive technical findings, we provide a novel crowdsourced "Turing test"–style method for examining the relative performance of any two models that produce substantive, text-based outputs. Our results are encouraging: popular, easily available pretrained embeddings perform at a level close to—or surpassing—both human coders and more complicated locally fit models. For completeness, we provide best practice advice for cases where local fitting is required.

Highly recommend:
Pedro Rodriguez and Arthur Spirling (2022)
Word Embeddings: What works, what doesn't, and how to tell the difference for applied research
Journal of Politics. https://doi.org/10.1086/715162

Moral: Using pre-trained models is just fine.

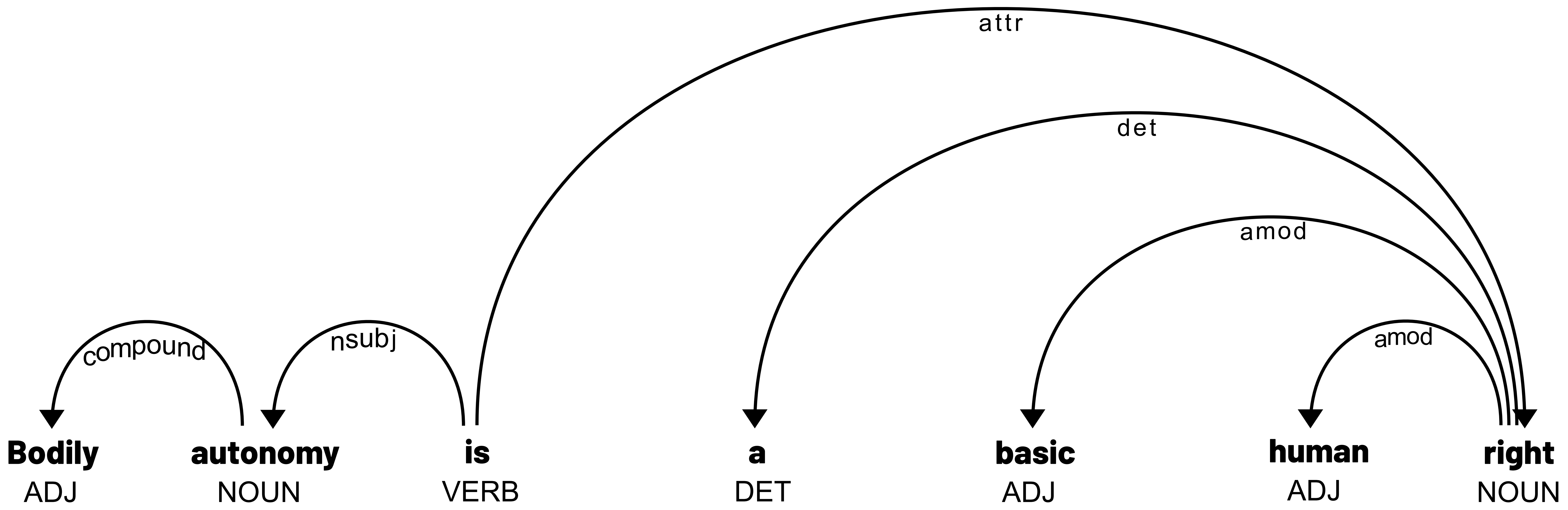
Identifying edges

- Identifying "connections between textual elements" is typically conceptual harder
- But, there can still be some relatively simple ways of doing this:
  - Co-occurrence: two nodes are connected if they occur with the same ***span***
- Can also have more complex definitions of an edge
  - For example, ***similar*** words are connected
  - Grammar determines connections

Again, I would generally recommend starting with simpler models (ie, co-occurrence)

But! These ideas are also (excuse the pun) connected!

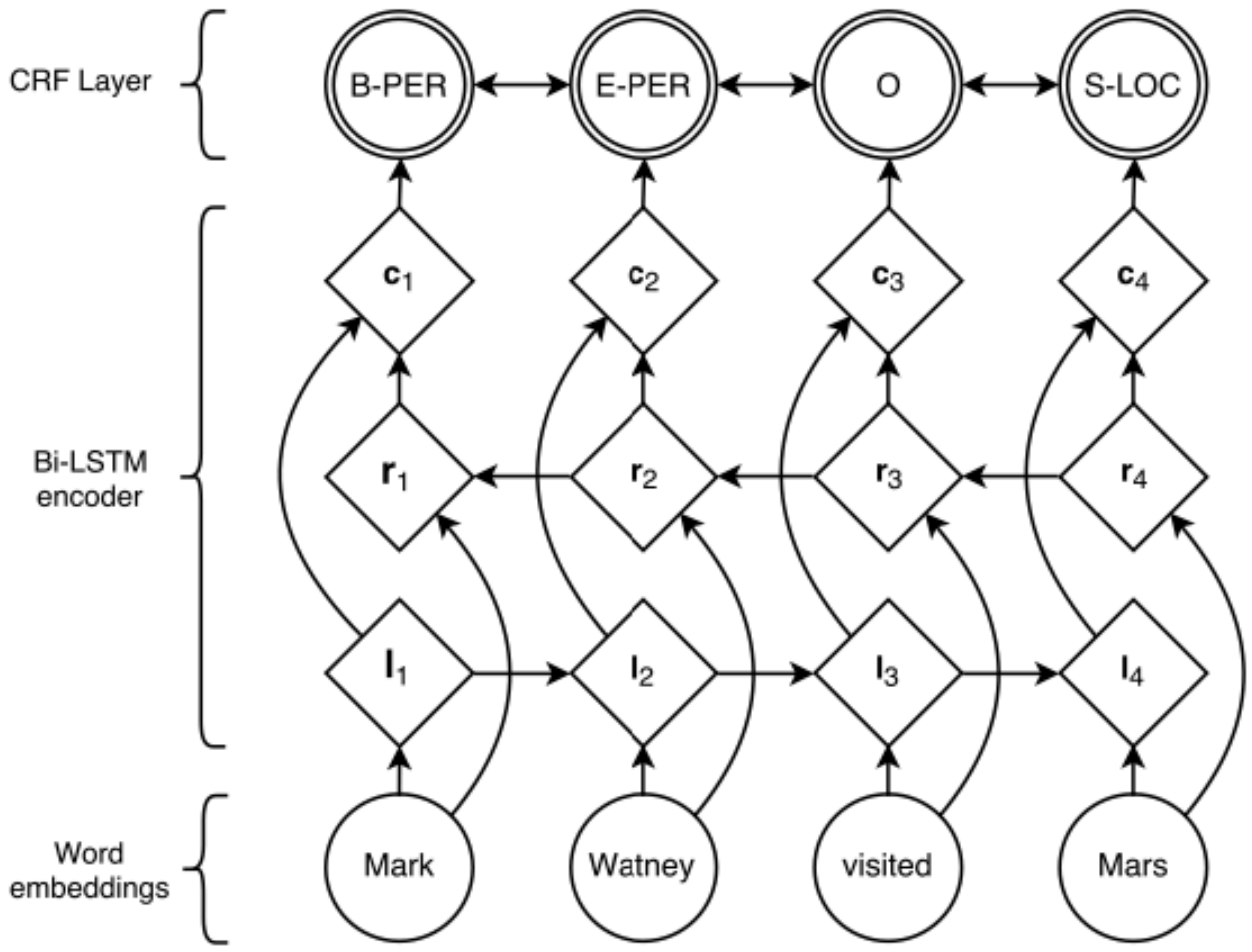
Semantic Parse Trees



Grammatical structure is inherently a networks structure!

Text Analysis 103: Predicting POS

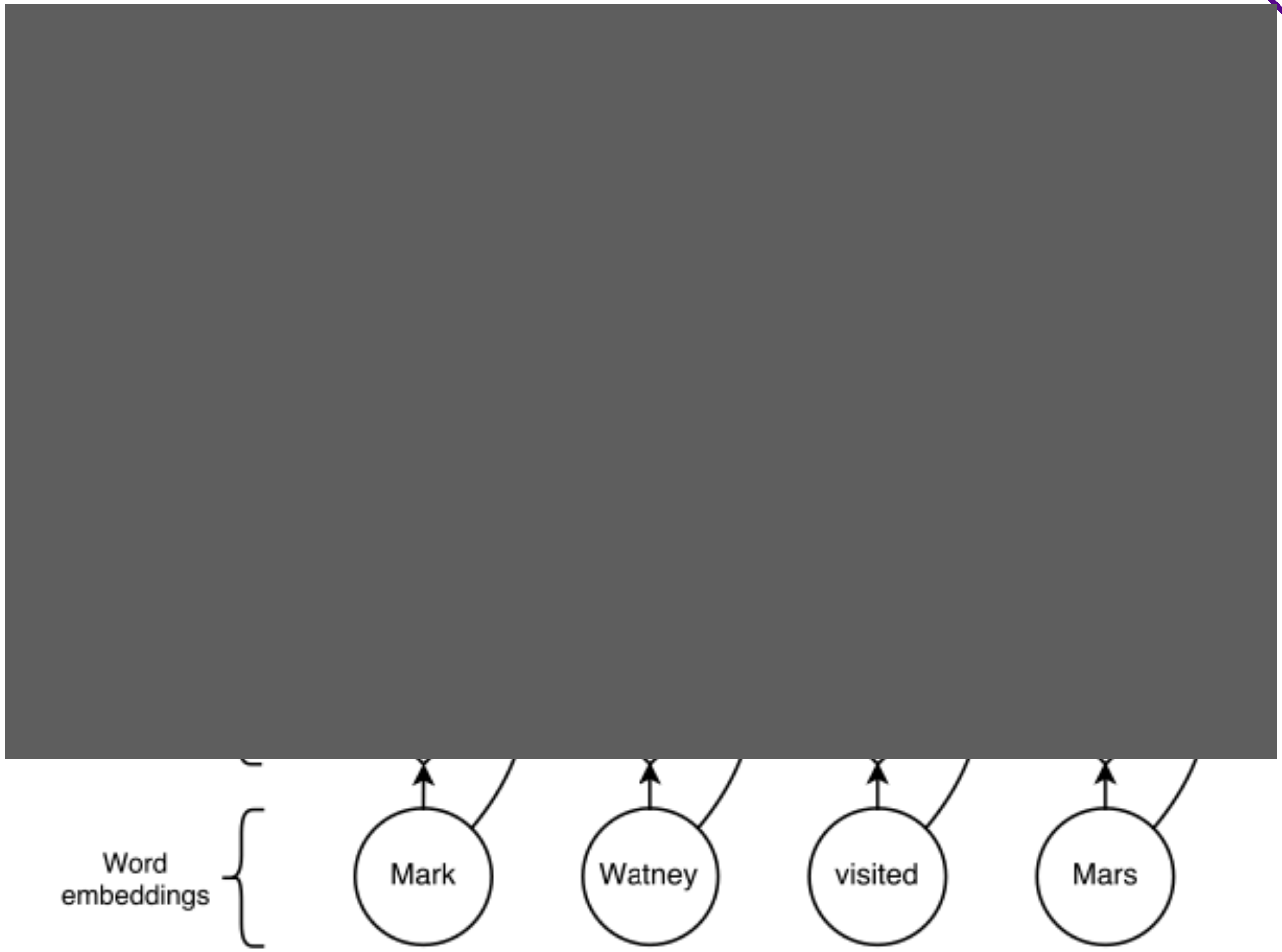
Identifying Part of Speech (POS) tags follows a similar intuition as we saw with word embeddings...



Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

Text Analysis 103: Predicting POS

Consider the input sentence:
"Mark Watney visited Mars"

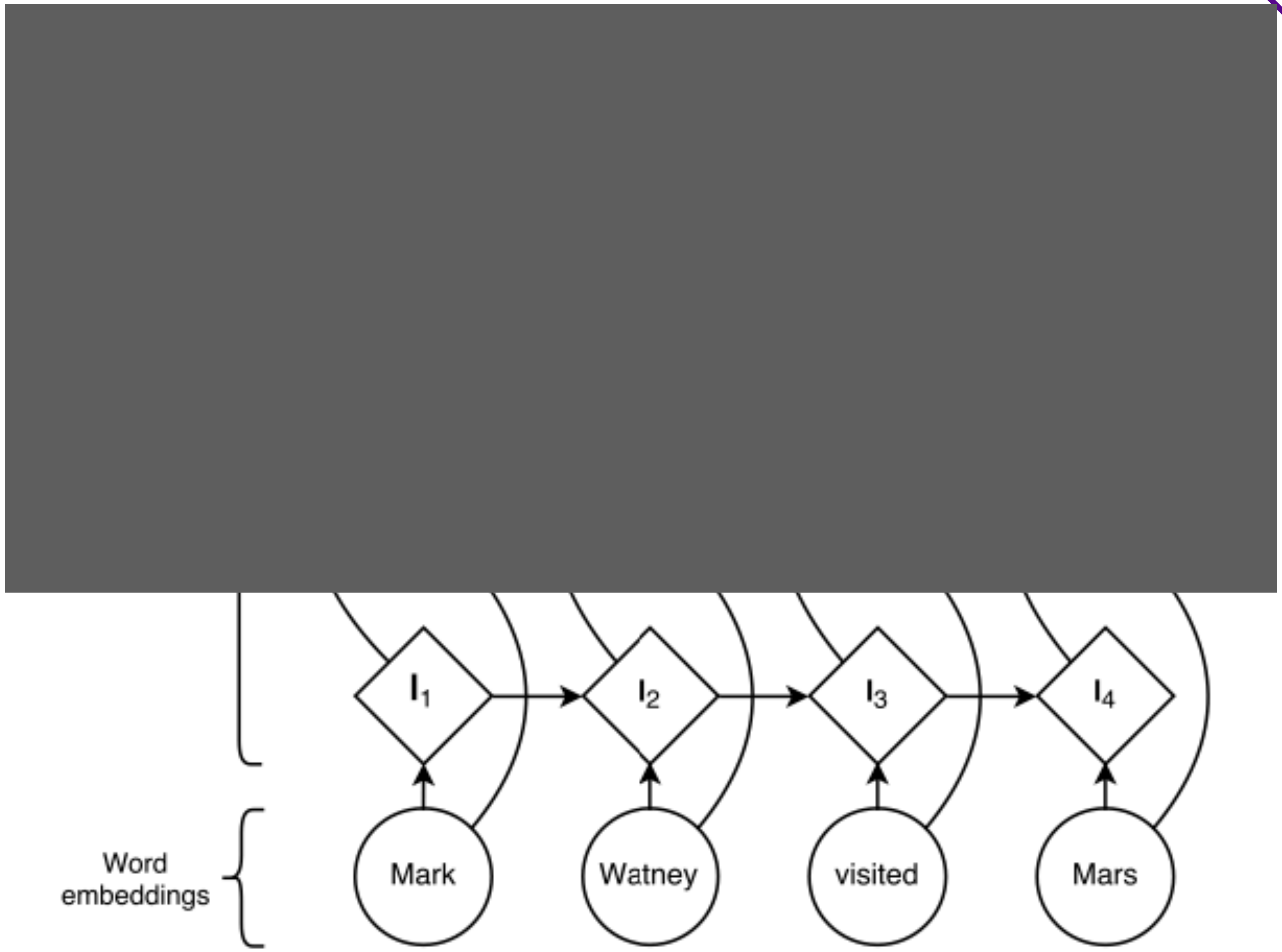


Input: N-Dimensional representation of each word
(More on what this means soon!)

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

Text Analysis 103: Predicting POS

Consider the input sentence:
"Mark Watney visited Mars"



Consider the "left context" (word(s) to the left)

Input: N-Dimensional representation of each word
(More on what this means soon!)

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

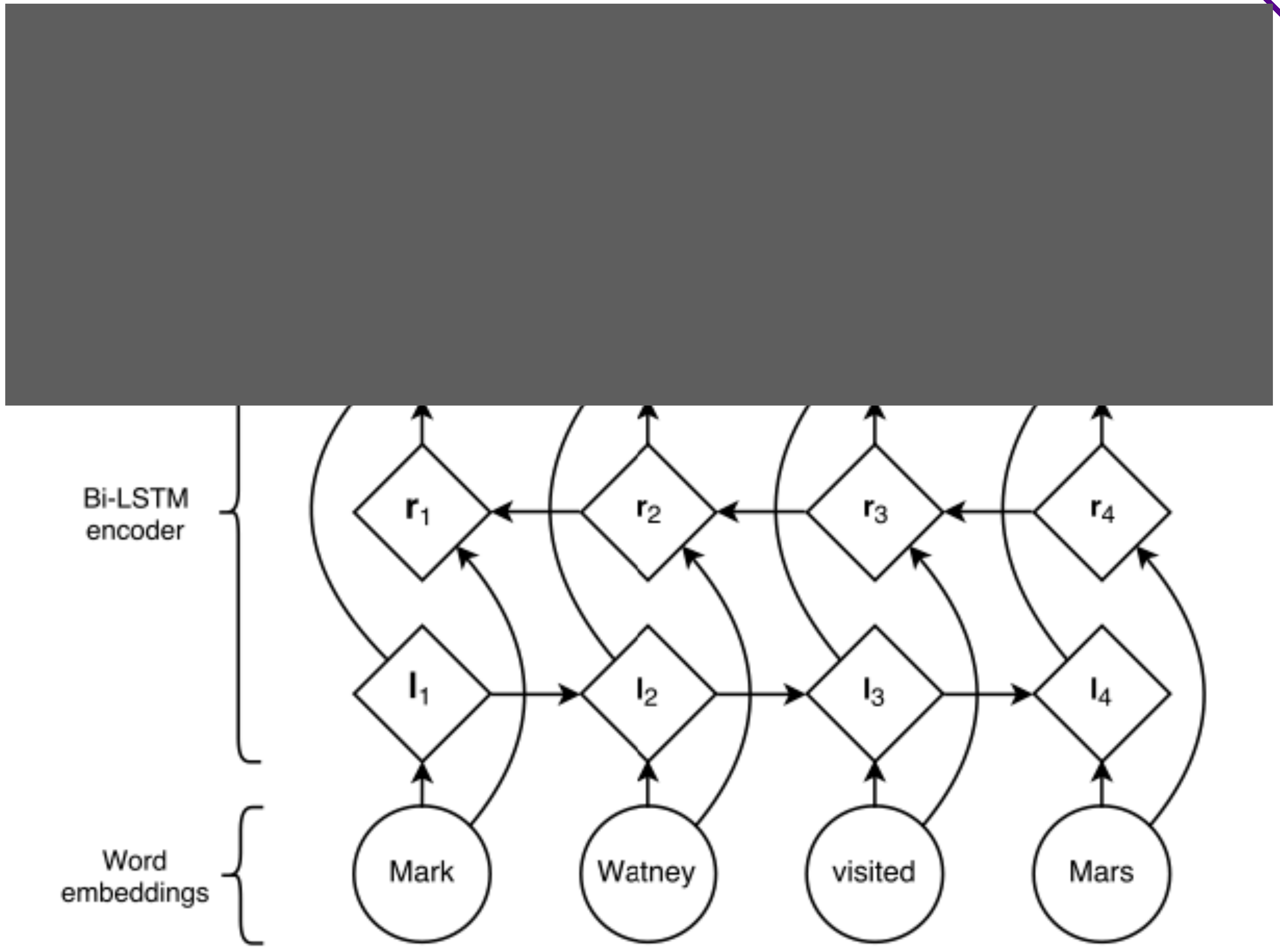
Text Analysis 103: Predicting POS

Consider the input sentence:
"Mark Watney visited Mars"

Consider the "right context" (word(s) to the right)

Consider the "left context" (word(s) to the left)

Input: N-Dimensional representation of each word
(More on what this means soon!)



Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

Text Analysis 103: Predicting POS

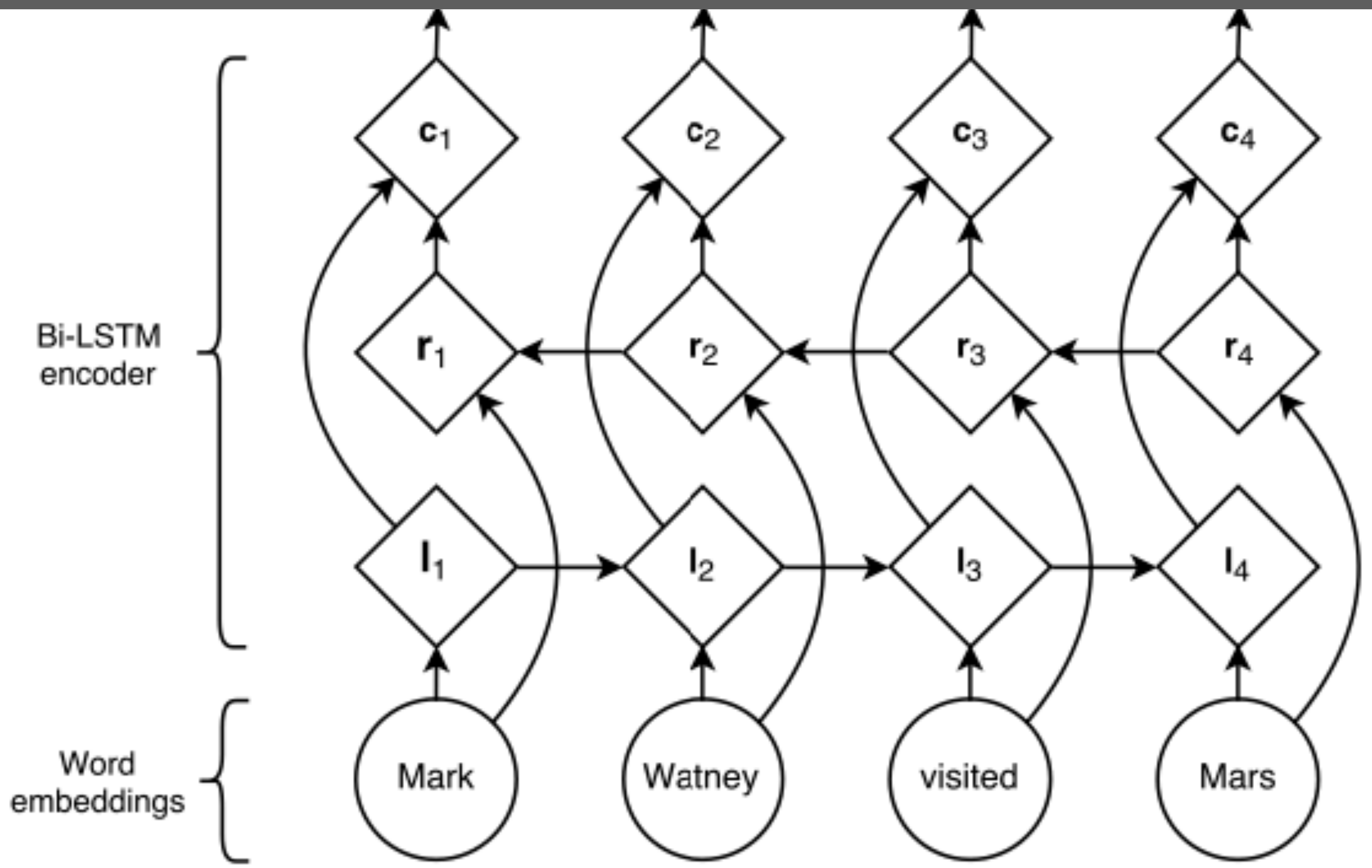
Consider the input sentence:
"Mark Watney visited Mars"

Concatenate L and R to consider entire context

Consider the "right context" (word(s) to the right)

Consider the "left context" (word(s) to the left)

Input: N-Dimensional representation of each word
(More on what this means soon!)



Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

Text Analysis 103: Predicting POS

Consider the input sentence:
"Mark Watney visited Mars"

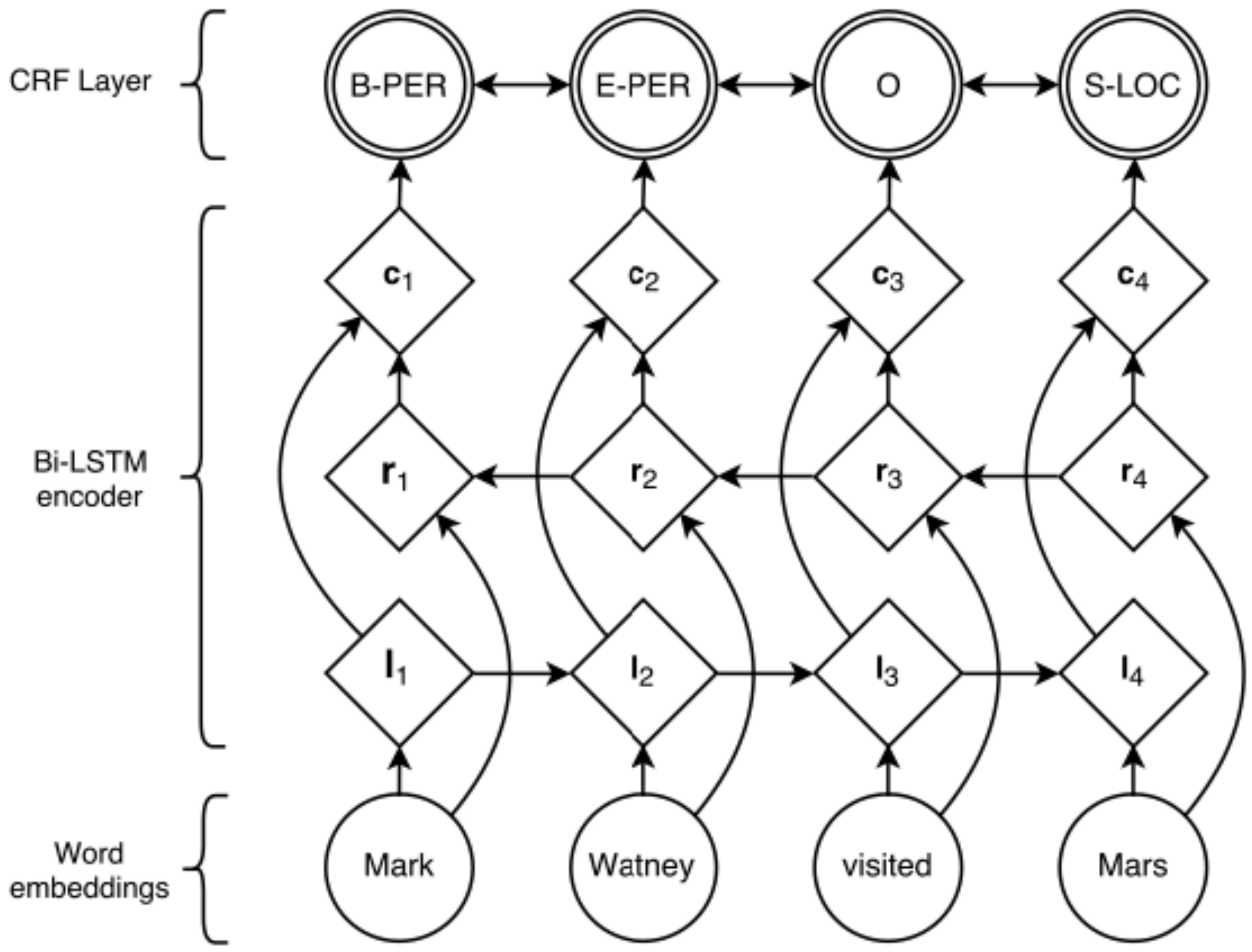
Jointly predict each word's label (or tag)

Concatenate L and R to consider entire context

Consider the "right context" (word(s) to the right)

Consider the "left context" (word(s) to the left)

Input: N-Dimensional representation of each word
(More on what this means soon!)



Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

Text and Network Methods

- Automated text parsing can be used for both:
  - Identifying edges (parse tree)
  - Identifying nodes (for example: named entities or phrases)

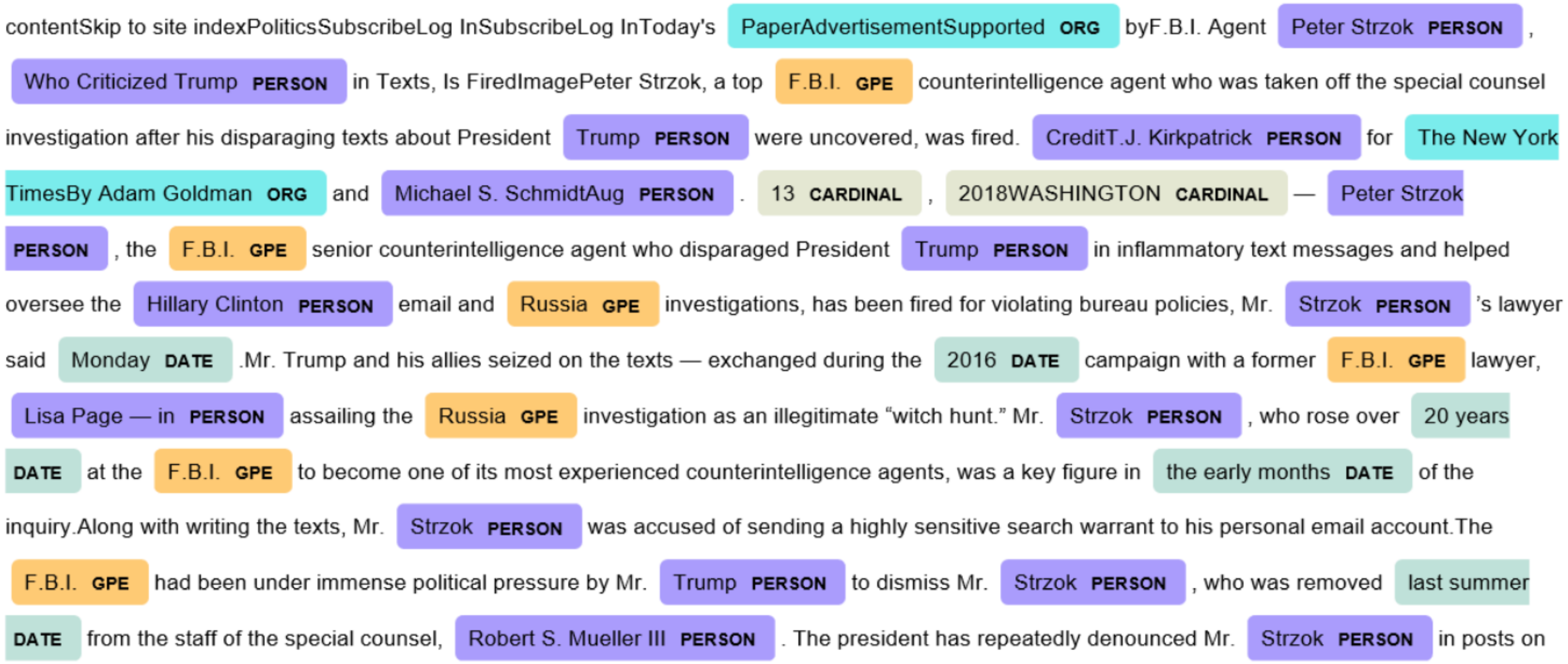


Image from https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2

Text and Network Methods

- Automated text parsing can be used for both:
  - Identifying edges (parse tree)
  - Identifying nodes (for example: named entities or phrases)

- Note that word similarity can ***also*** be used for both:
  - Identifying edges (similar words are connected)
  - Identifying nodes (similar words represent the same node)

A lot comes down to how you think about your specific data & research question!

Validating Your Approach

- You really have to look at your data!

- There are two basic ways to validate a model:
    1. Start with hand-labeled data and see how well your model does
    2. Hand-label a sample of model output to see how well it did

- No standard hand-coding procedure, typically want (at least) 2 independent coders per document

- Need to balance coding "as much data as possible" with time and effort required — depends on difficulty of task and how strong you want your findings to be. Ask around!

Validating Your Approach

- For example, can identify "named entities" by:
  - ➡ Hand coding and manually identifying entities. Gives the highest accuracy, but is not scalable to large corpora
  - ➡ Using trained neutral nets, typically ~85% accuracy

- "Using NLP" typically means using or adapting an implemented, pre-trained model
  - ➡ **Model training is data intensive** – should be trained on (for example) "the English language"
  - ➡ **VERY important to examine accuracy ON YOUR DATA**

Overview of Examples

1. Hashtag co-occurrence
   ➡ Easy NLP (Natural Language Processing) task
   ➡ Discussion of network modeling

2. Named entity co-occurrence
   ➡ Will use a pre-trained statistical model (using SpaCy)
   ➡ Discussion of validation/accuracy/challenges

3. "Concept" connections (if time)
   ➡ Getting creative!

Example 1: Hashtag Co-Occurrence

- Multiple documents (tweets)
- Documents contain key words (hashtags)
- Documents are connected if they share a word (hashtag)

- A single word is called a "unigram"
- Two words are a "bigram"
- Can also use "n-grams" of arbitrary length (named entities, specific phrases, etc)

Model Setup

- Nodes are words
- Edges indicate words co-occur

  (in document or within specified window)

Example 1: Hashtag Co-Occurrence

Document 1:

This is a tweet! #MyHashtag #What #NLProc

Document 2:

Another tweet. # NLProc

Document 3:

#MyHashtag is the best hashtag. #What

Example 1: Hashtag Co-Occurrence

Document 1:

This is a tweet! #MyHashtag #What #NLProc

Document 2:

Another tweet. #NLProc

Document 3:

#MyHashtag is the best hashtag. #What

Example 1: Hashtag Co-Occurrence

Document 1:

This is a tweet! #MyHashtag #What #NLProc

Document 2:

Another tweet. #NLProc

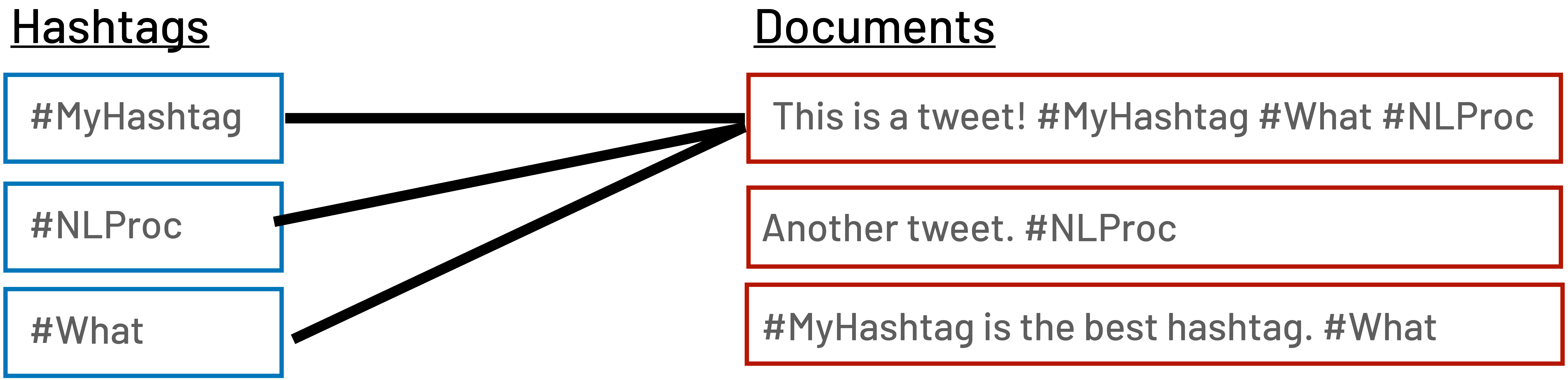
Document 3:

#MyHashtag is the best hashtag. #What

Example 1: Hashtag Co-Occurrence

Technically, this a **bipartite** network

- Two types of nodes
- Only connect to nodes of **other** type



Example 1: Hashtag Co-Occurrence

Technically, this a **bipartite** network

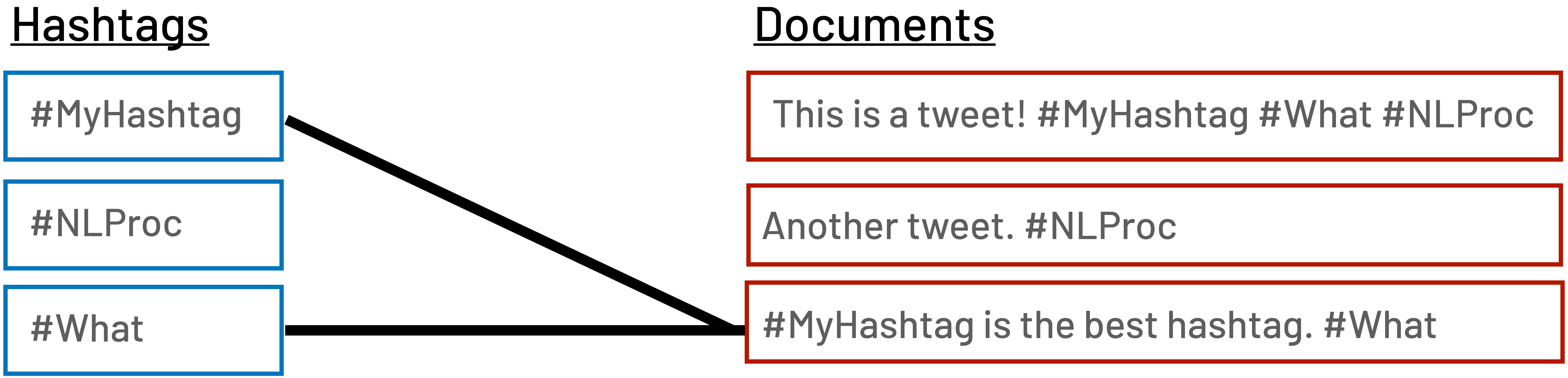
- Two types of nodes
- Only connect to nodes of **other** type

Hashtags	Documents
#MyHashtag	This is a tweet! #MyHashtag #What #NLProc
#NLProc	Another tweet. #NLProc
#What	#MyHashtag is the best hashtag. #What

Example 1: Hashtag Co-Occurrence

Technically, this a **bipartite** network

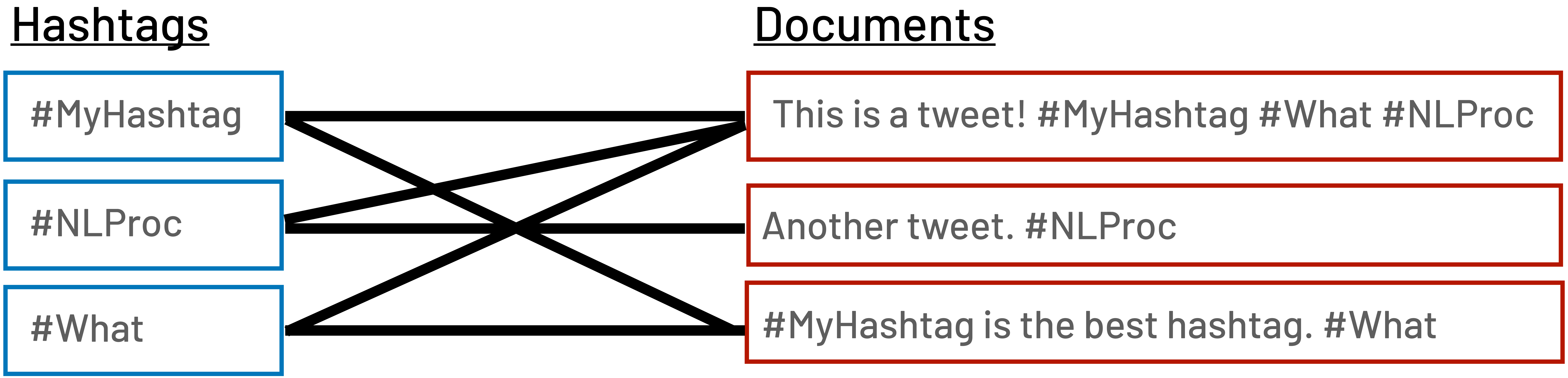
- Two types of nodes
- Only connect to nodes of **other** type



Example 1: Hashtag Co-Occurrence

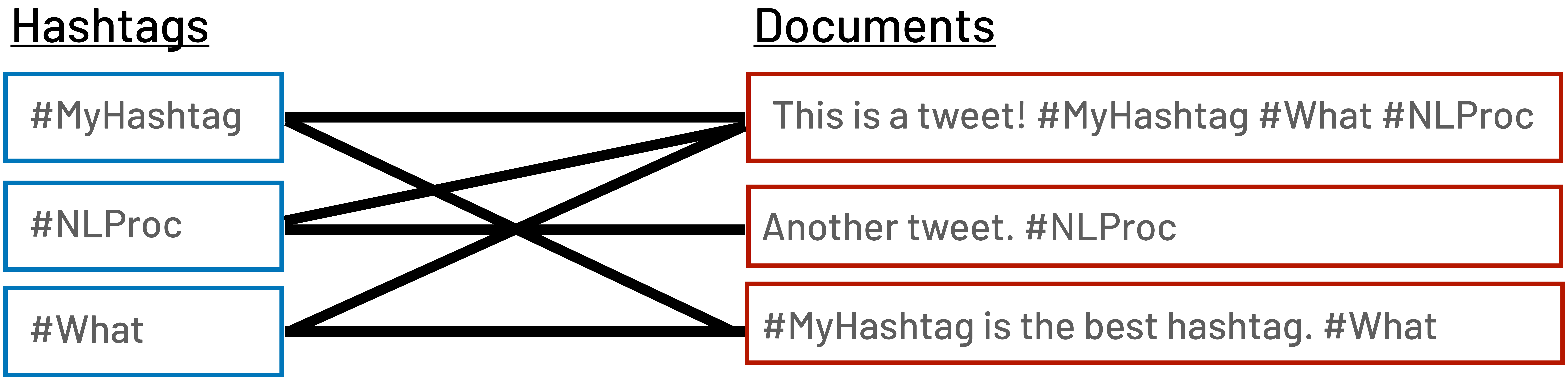
Technically, this a **bipartite** network

- Two types of nodes
- Only connect to nodes of **other** type



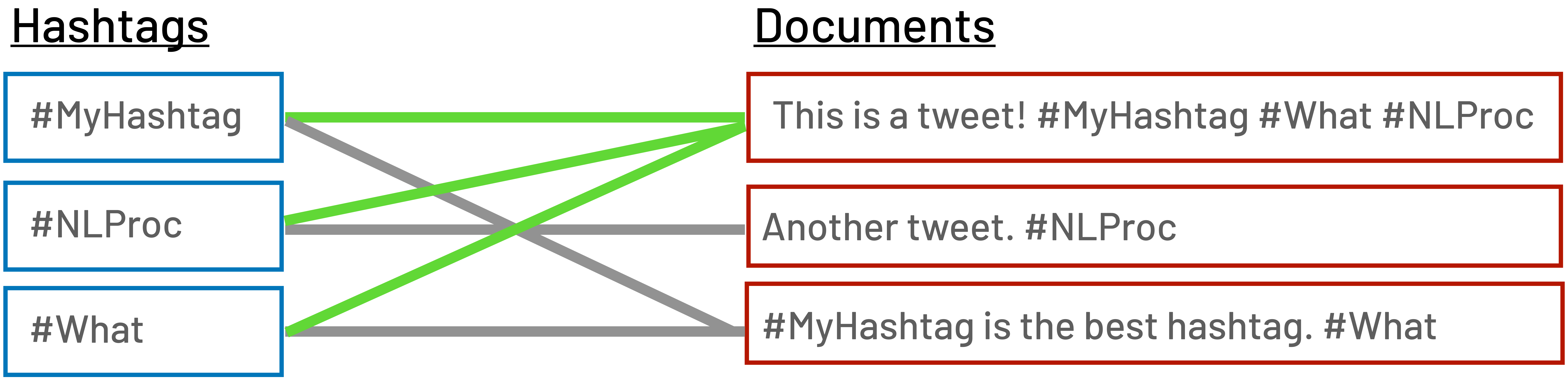
Example 1: Hashtag Co-Occurrence

If we take the **projection** of a bipartite network on one of the nodes sets, it gives us co-occurrence:



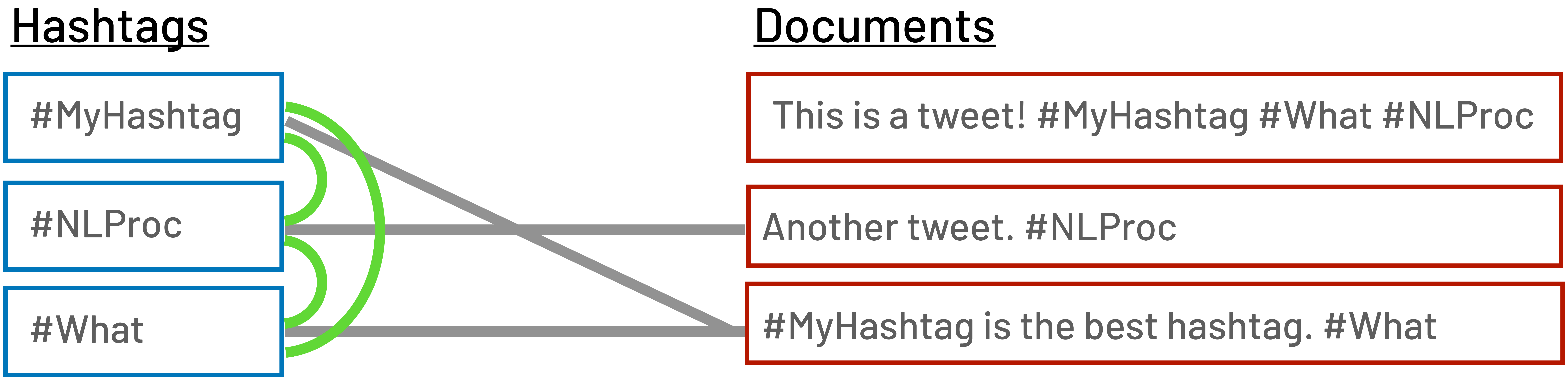
Example 1: Hashtag Co-Occurrence

If we take the **projection** of a bipartite network on one of the nodes sets, it gives us co-occurrence:



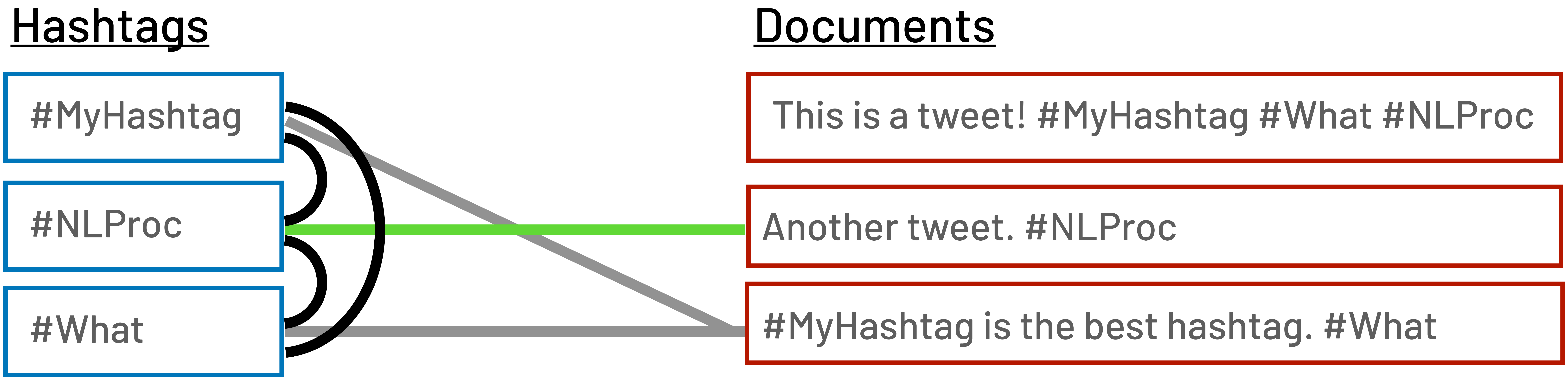
Example 1: Hashtag Co-Occurrence

If we take the **projection** of a bipartite network on one of the nodes sets, it gives us co-occurrence:



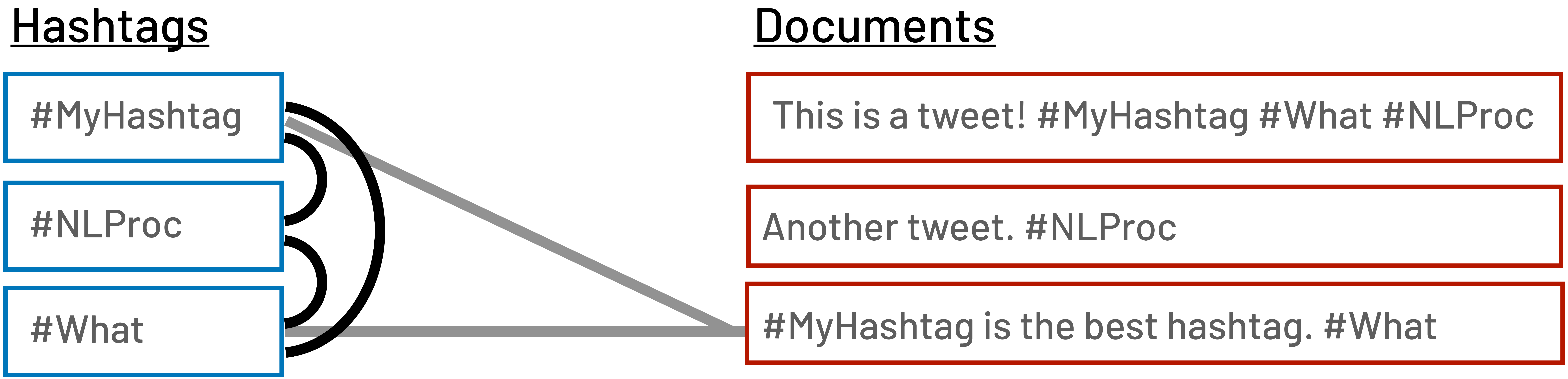
Example 1: Hashtag Co-Occurrence

If we take the **projection** of a bipartite network on one of the nodes sets, it gives us co-occurrence:



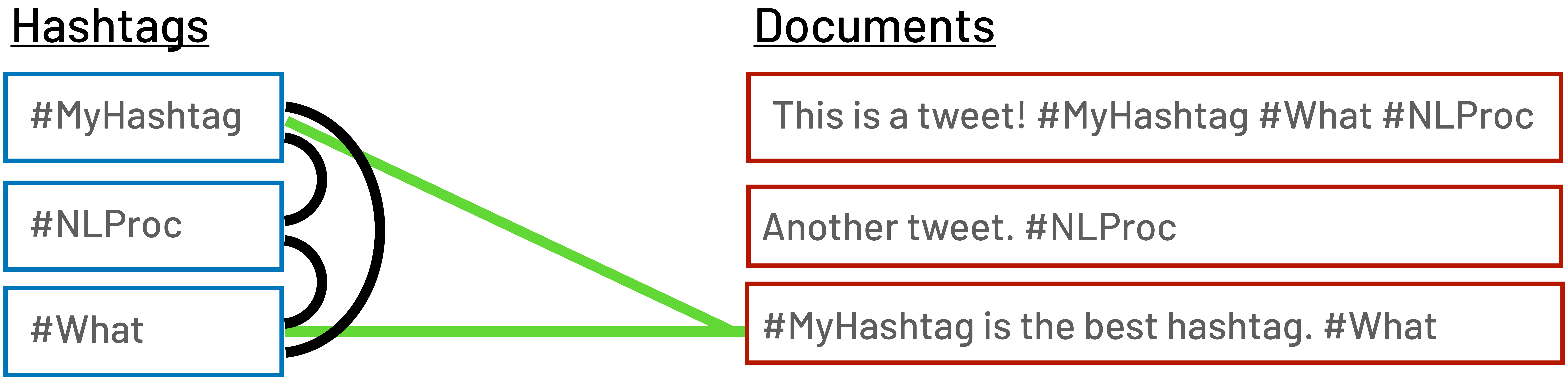
Example 1: Hashtag Co-Occurrence

If we take the **projection** of a bipartite network on one of the nodes sets, it gives us co-occurrence:



Example 1: Hashtag Co-Occurrence

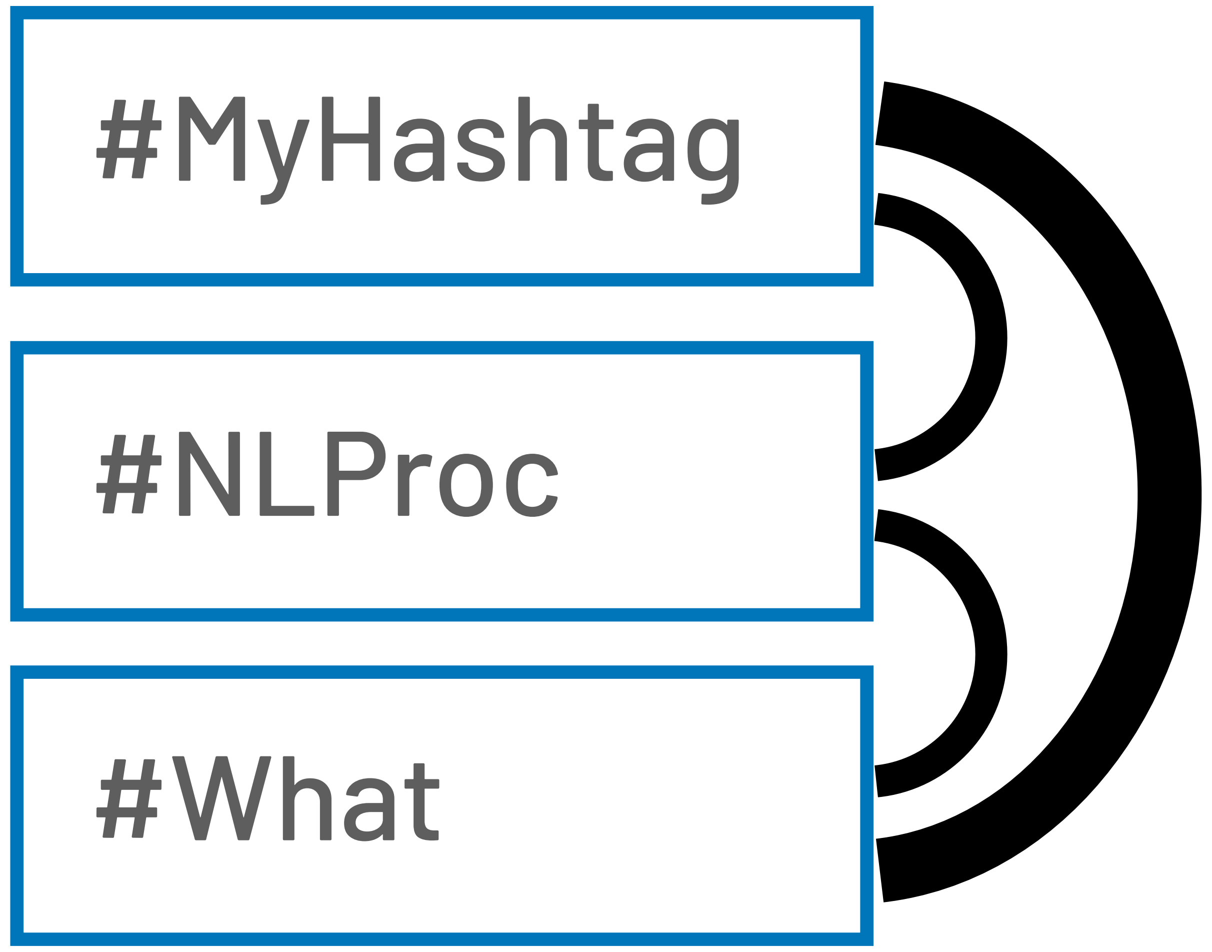
If we take the **projection** of a bipartite network on one of the nodes sets, it gives us co-occurrence:



Example 1: Hashtag Co-Occurrence

If we take the **projection** of a bipartite network on one of the nodes sets, it gives us co-occurrence:

Hashtags



Documents

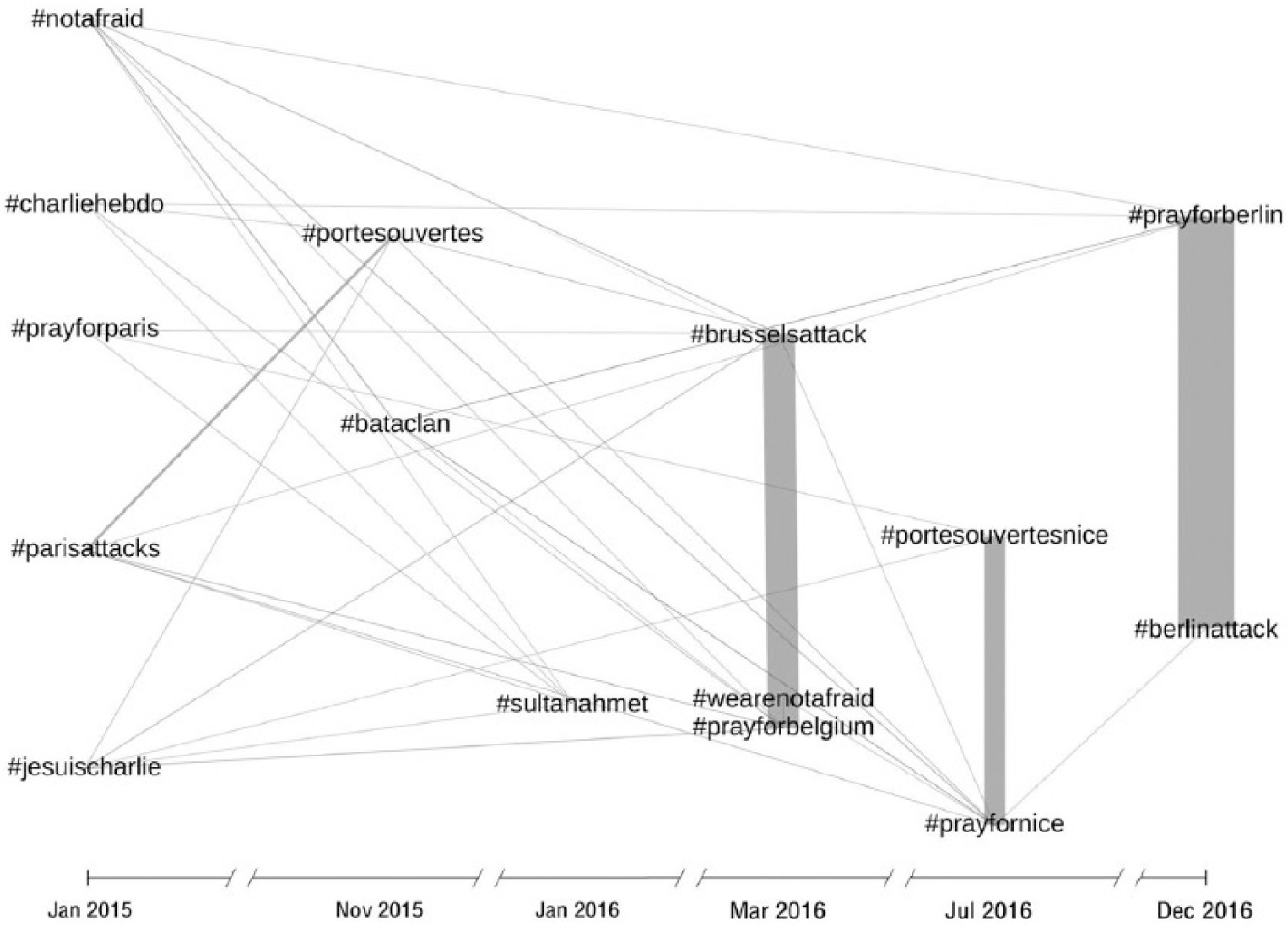
This is a tweet! #MyHashtag #What #NLProc

Another tweet. #NLProc

#MyHashtag is the best hashtag. #What

Example 1: Hashtag Co-Occurrence

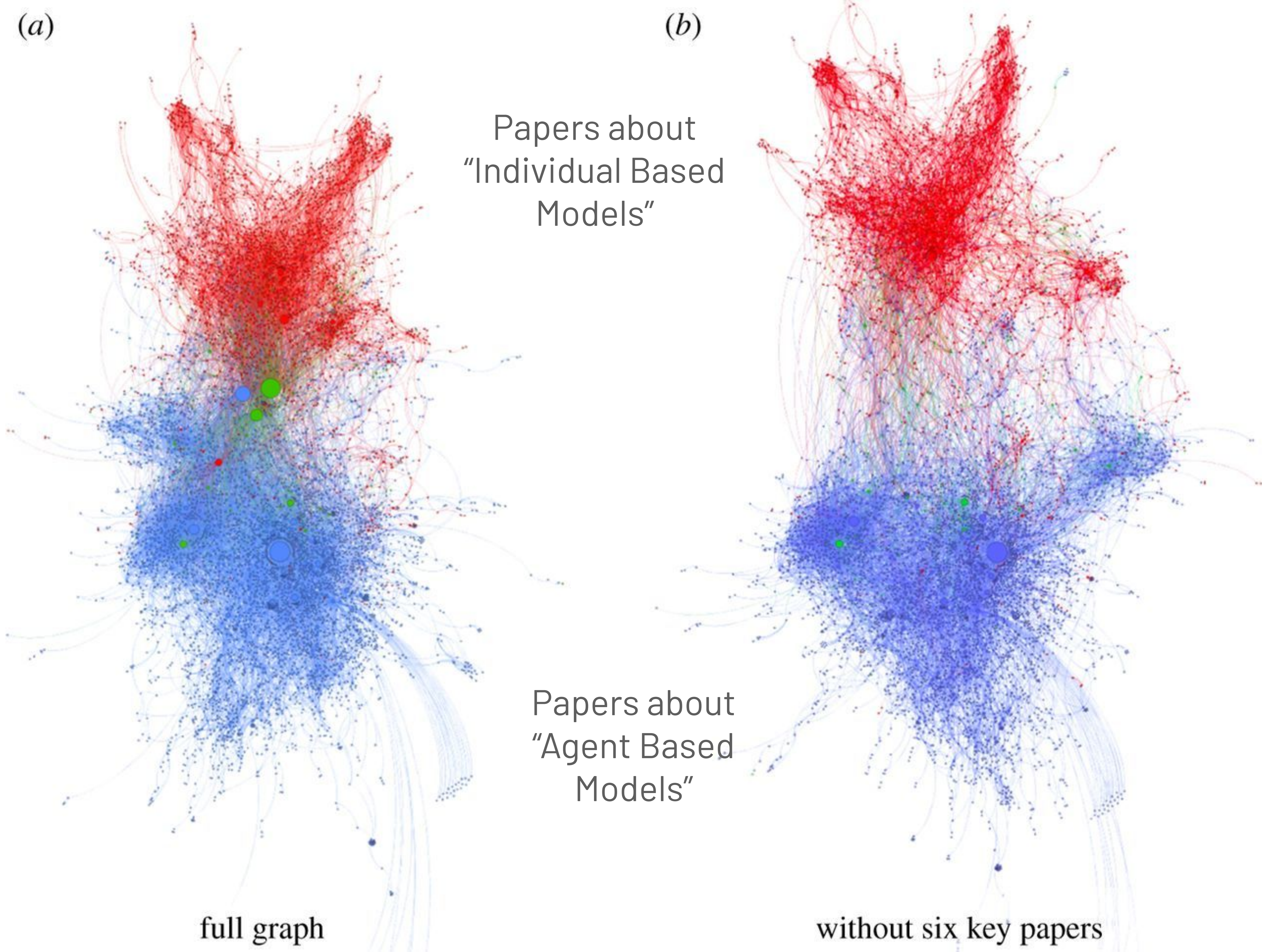
- The projection on hashtags (or words) tells us which words "go together"
  - ➡ the "discourse landscape"
  - ➡ What things are people talking about together?
  - ➡ What "discursive communities" arise?



Eriksson Krutrök, M., & Lindgren, S. (2018). Continued Contexts of Terror: Analyzing Temporal Patterns of Hashtag Co-Occurrence as Discursive Articulations. *Social Media + Society*. https://doi.org/10.1177/2056305118813649

Example 1: Hashtag Co-Occurrence

- The projection on tweets (documents) tells us which documents are "similar"
  - ➡ Which authors tend to talk similarly?
  - ➡ Useful for document recommendation



Vincenot, Christian E. 2018. How new concepts become universal scientific approaches: insights from citation network analysis of agent-based complex systems science. Proc. R. Soc. B. http://doi.org/10.1098/rspb.2017.2360

Example 1: Hashtag Co-Occurrence

Some notes:

- Any "co-occurrence" network is (probably?) a projection of a bipartite network

- In **code**, we can often skip the bipartite network and directly construct the co-occurrence network (we'll see this soon!)

- BUT, it's helpful to remember it's technically a projection!
  - ➡ Projections often have distinct structural features (eg, higher clustering)
  - ➡ The "other half" of the network may have useful metadata!

Example 2: Entity Co-Occurrence

- #Hashtags are easy to identify because they have an obvious textual symbol (eg, ‘#’)

- What if our “objects” of interest can have an arbitrary format?

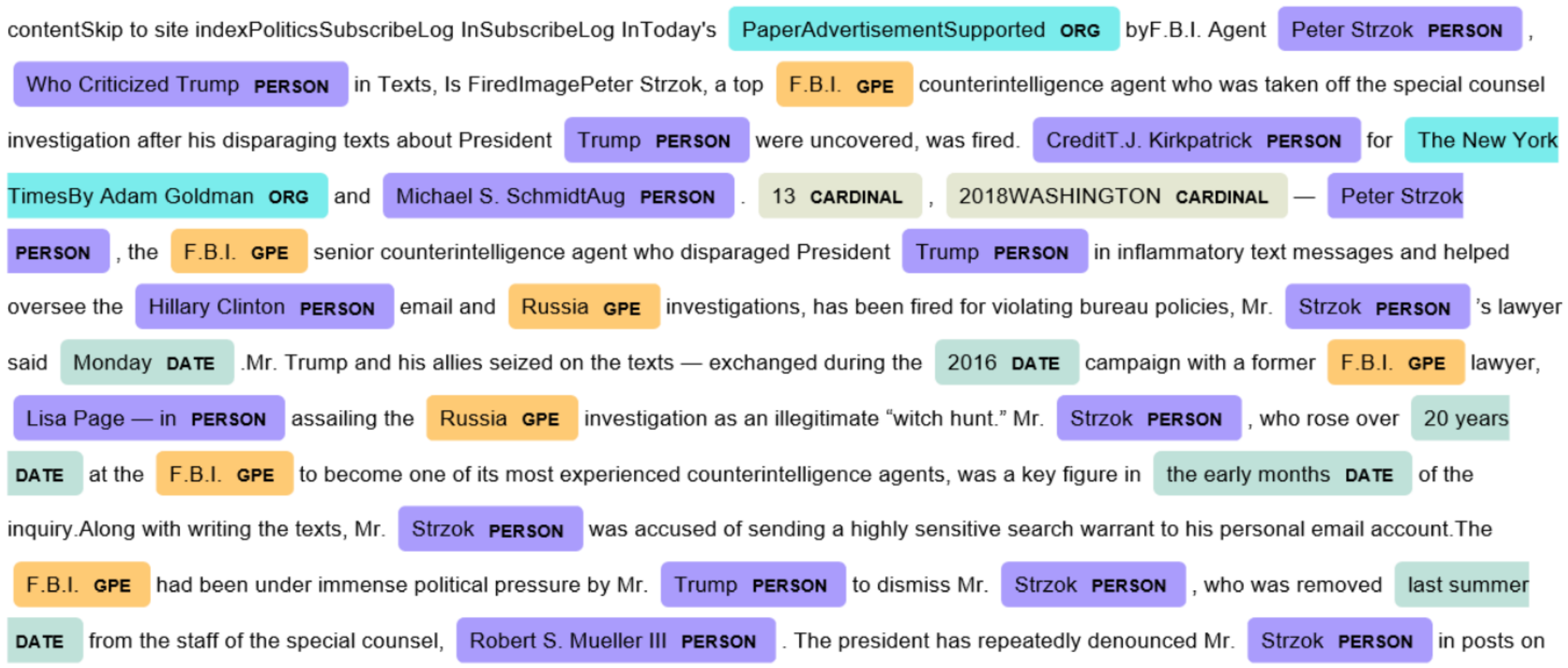

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate “witch hunt.” Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Image from https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2

Example 2: Entity Co-Occurrence

The network could be:

- Based on co-occurrence (same sentence/span)
- Between different types of entities (People —> Orgs)
- Based on other words (A [verbs] B)

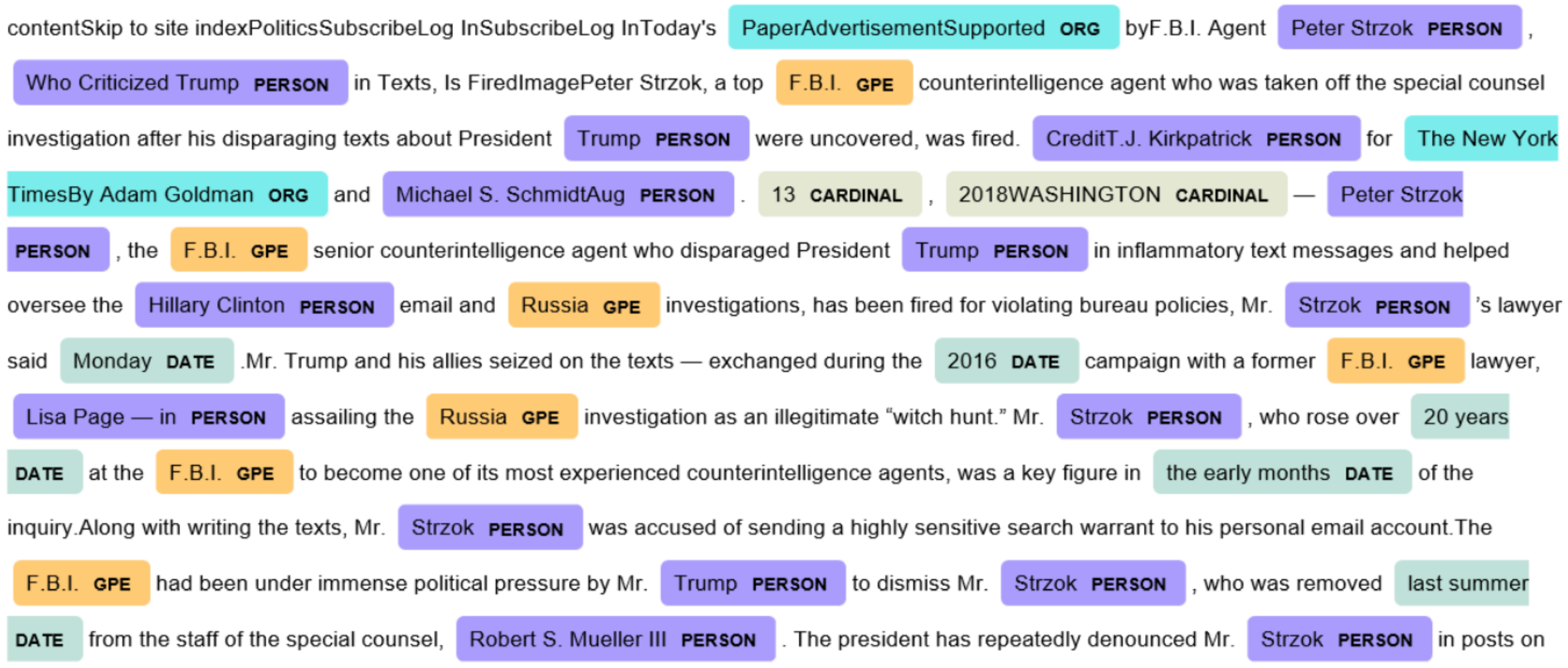
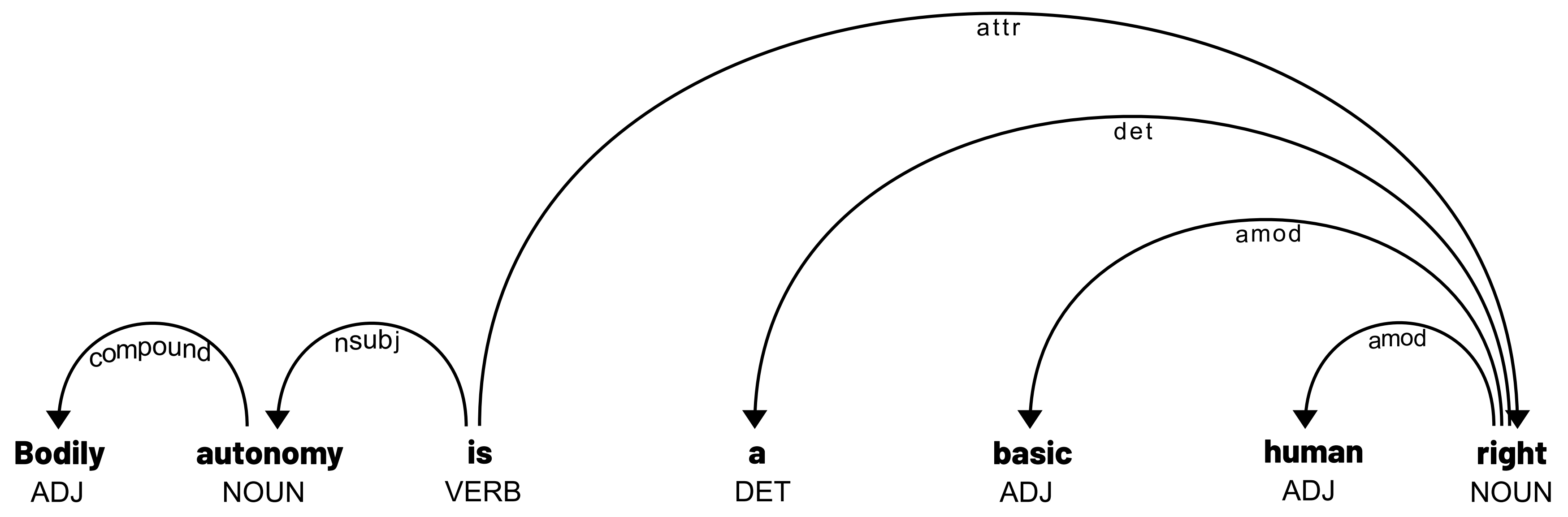

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Image from https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2

Example 3: Finding Connected Concepts

1. Use word embeddings to identify similar words

2. Use semantic parse tree to identify connections



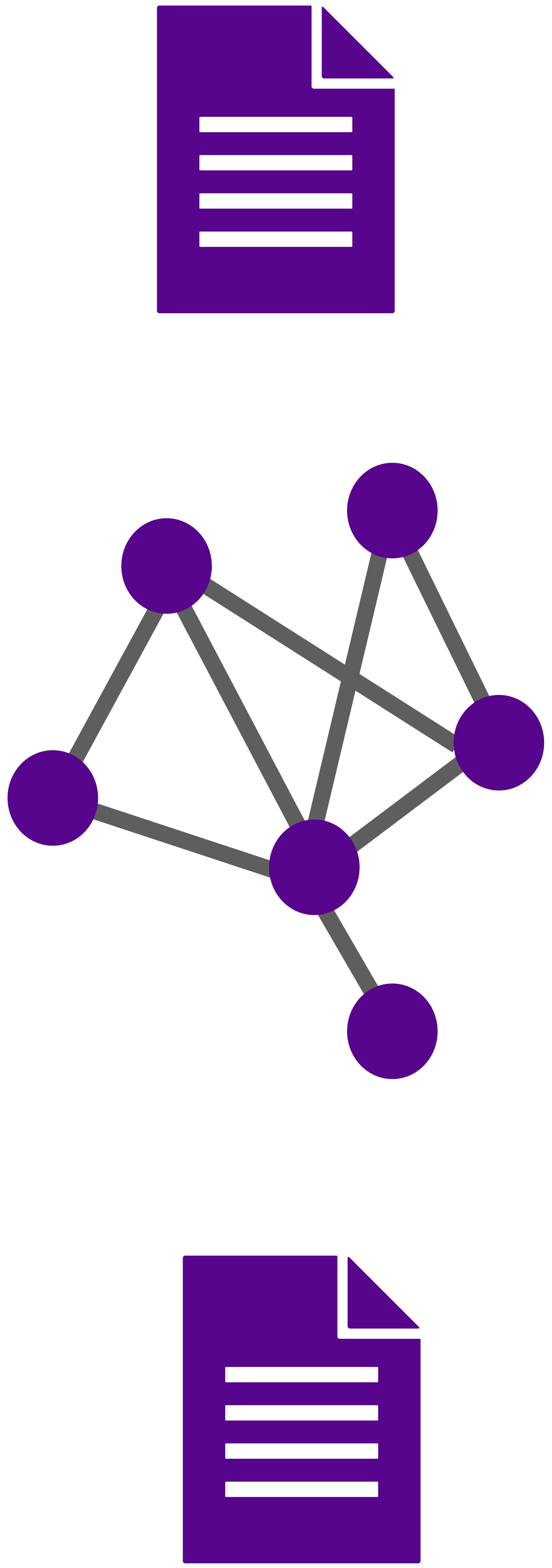
I've included some example code (that we may not have time for) but the high level take away is that there is LOTS of room for innovation and creativity!

But more researcher degrees of freedom means greater need for validation!!

Adapted from Shugars, S: The Structure of Reasoning: Measuring Justification and Preferences in Text
DOI: https://doi.org/10.17605/OSF.IO/PNWD8

Part 2: **Practice**

...But HOW???



Overview of Software

Network Analysis:

- We'll use Python + Networkx (https://networkx.org)
- R users: check out igraph (https://igraph.org/r/)
- Gephi great for visualizations, but also super buggy (https://gephi.org)

Overview of Software

Text analysis:

- We'll use Python + SpaCy (https://spacy.io)
- NLTK is another popular python package (https://www.nltk.org
  - NLTK has more options: eg, more data sets, more models, etc
  - SpaCy has fewer options but does what it does very well and is faster to integrate the newest NLP approaches
- R users: check out Quanteda (https://quanteda.io)
- Quanteda also has a SpaCy wrapper for R (https://spacyr.quanteda.io)

Overview of Software

This workshop will be in Python, but as you go forth in life, you should use whichever language you personally* feel most comfortable working in.

Neither is “better” than the other!

* Or maybe your collaborators

Now, we'll look at some code!