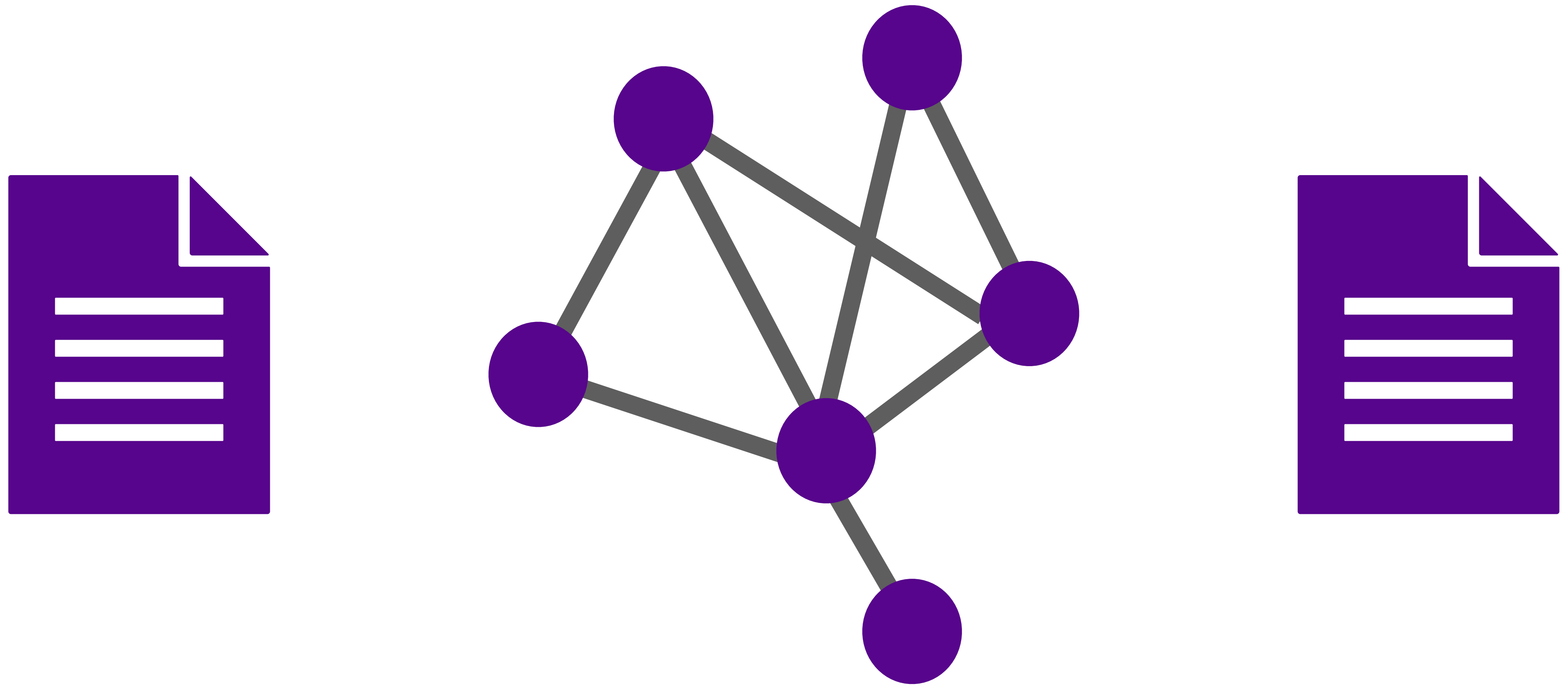
PolNet 2021 Workshop:

Methods at the Intersection of Network and Text Analysis



Sarah Shugars

Faculty Fellow, Center for Data Science

sarah.shugars@nyu.edu

they/she

Workshop Outline

Part 0: Logistics

- Goals, expectations, & introductions

Part 1: Theory

- Conceptual approaches to working with text & networks

Part 2: Practice

- Live coding (in Python)

Materials

All materials available at: https://github.com/sshugars/PolNet2021

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
- 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

Sarah Shugars

On the market for TT positions starting Fall 2022!

Research

- Political talk & "reasoning"
- Social Media
- Civic infrastructure

Methods

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

Personal & Contact

Pronouns: they/she
Twitter: @Shugars
Email: sarah.shugars@nyu.edu

Part 1: **Theory**

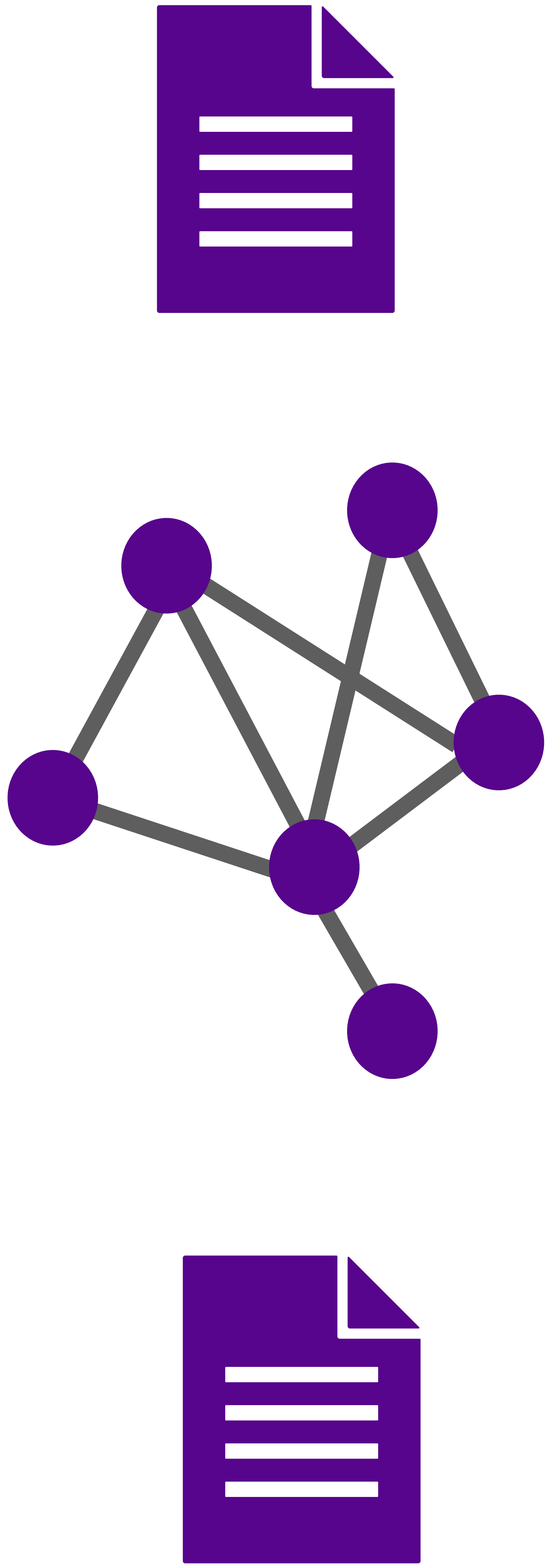
Why text?

Why networks?

What are these things?

What are we doing?

What is happening??



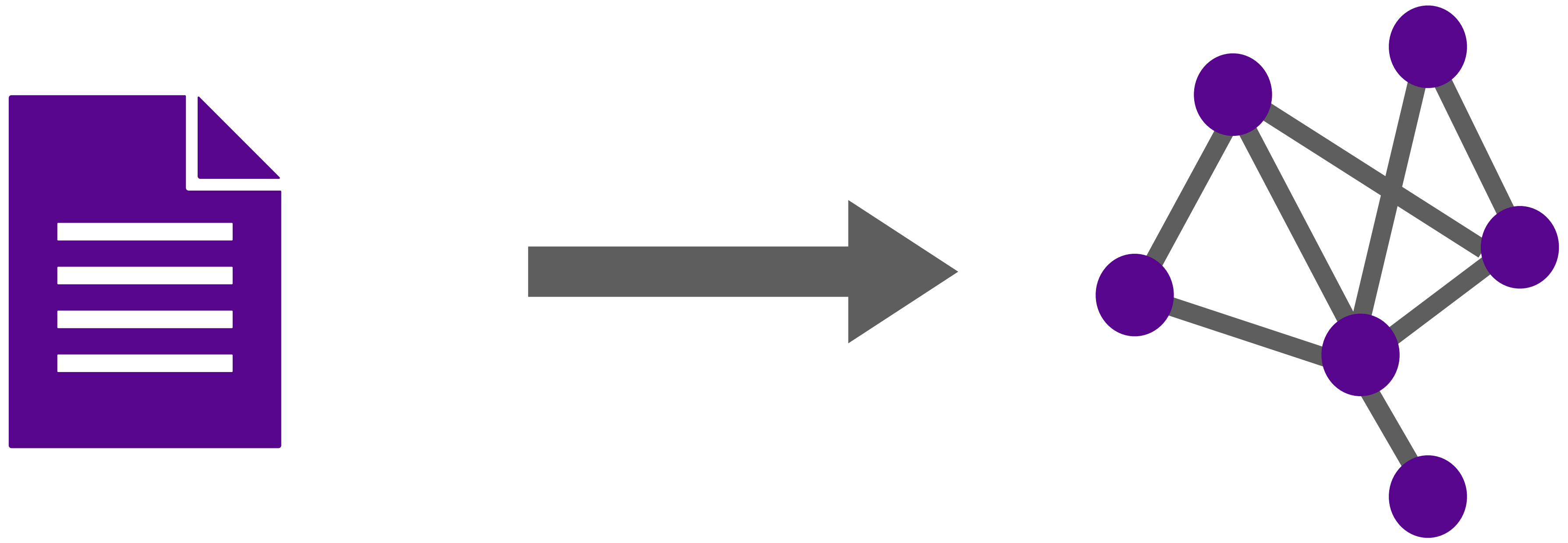
Computational Social Science Hot Take:

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

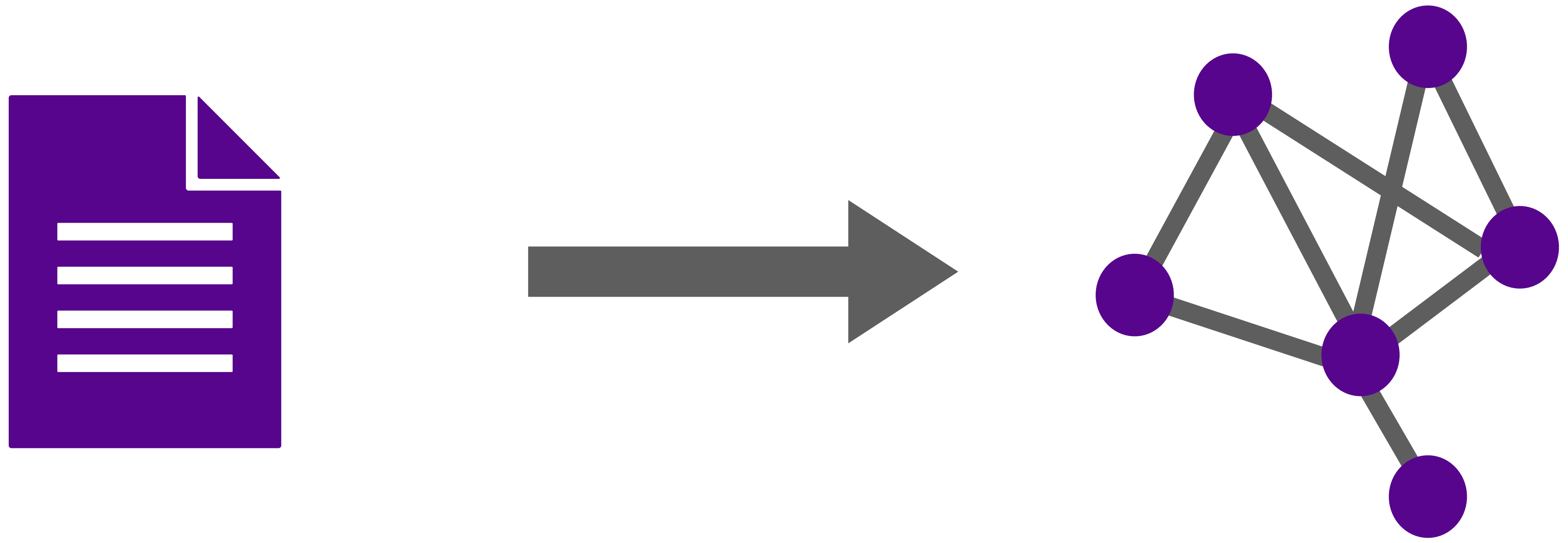
- "Coding" is relatively easy
  - ➡ Computers are really good at calculating things

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

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



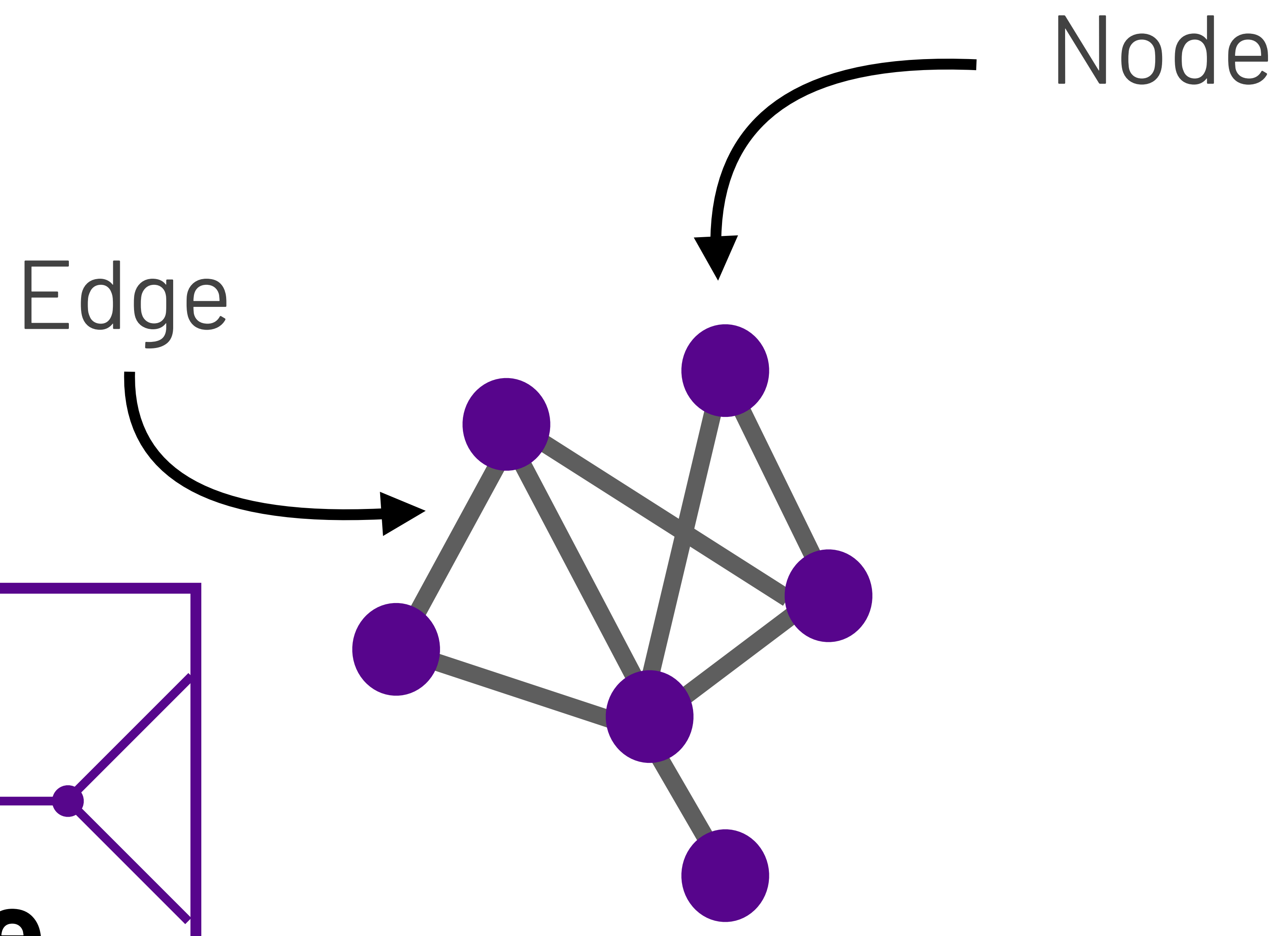
Q: What is the best way to go from text(s) to network(s)?



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.

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?

Text analysis 101

- Documents are collection of words*. 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. 🎉👍💯

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!

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?

Overview of Examples

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

2. Named entity co-occurence
   - ➡ 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

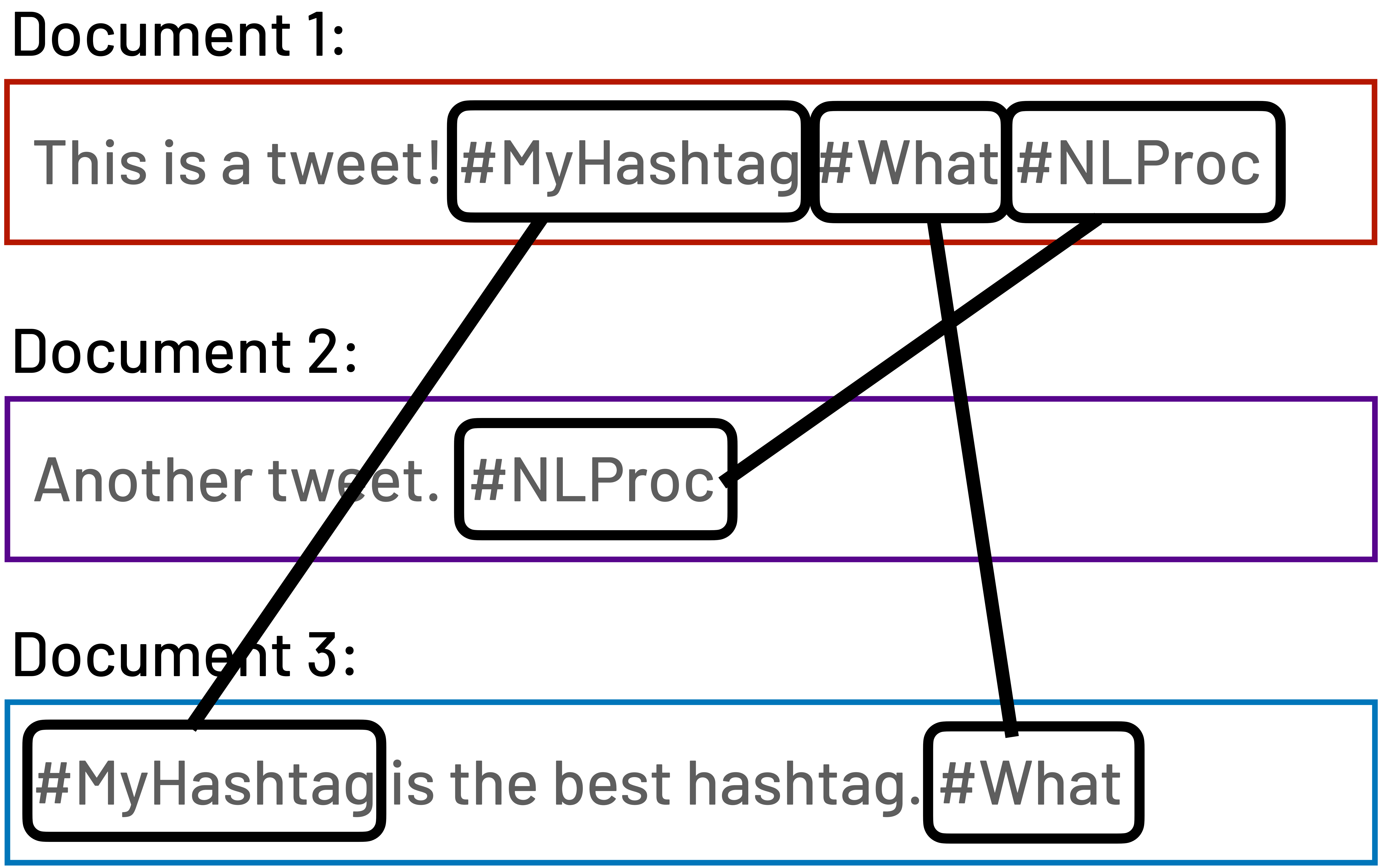
Document 2:

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

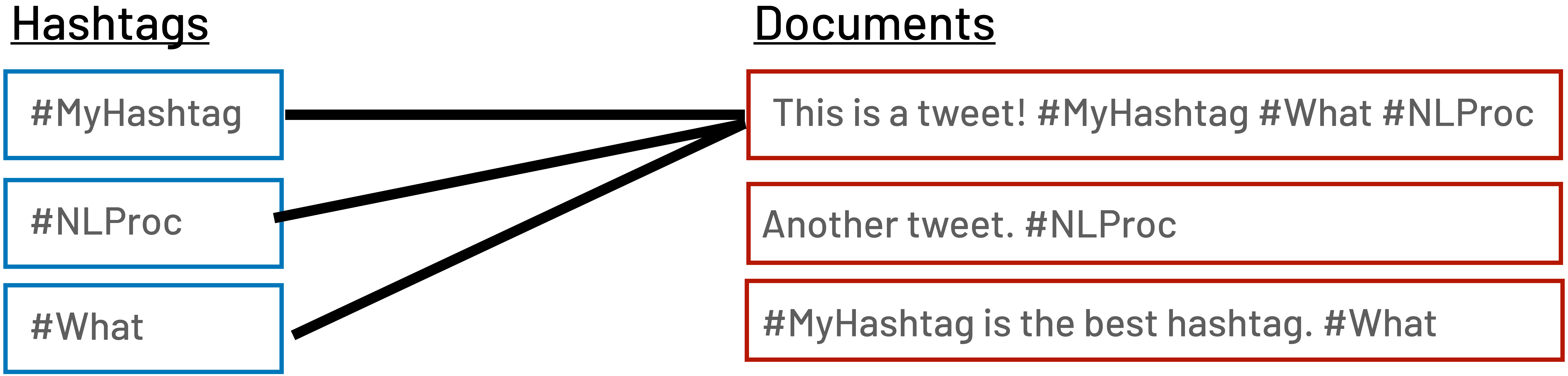
Document 3:

#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



Hashtags

#MyHashtag

#NLProc

#What

Documents

This is a tweet! #MyHashtag #What #NLProc

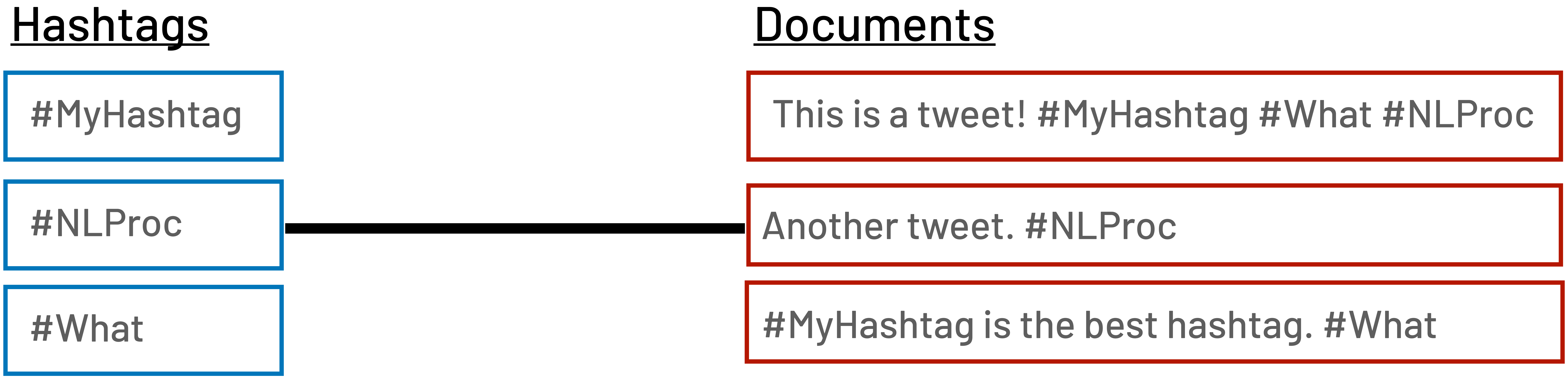
Another tweet. #NLProc

#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



Hashtags

#MyHashtag

#NLProc

#What

Documents

This is a tweet! #MyHashtag #What #NLProc

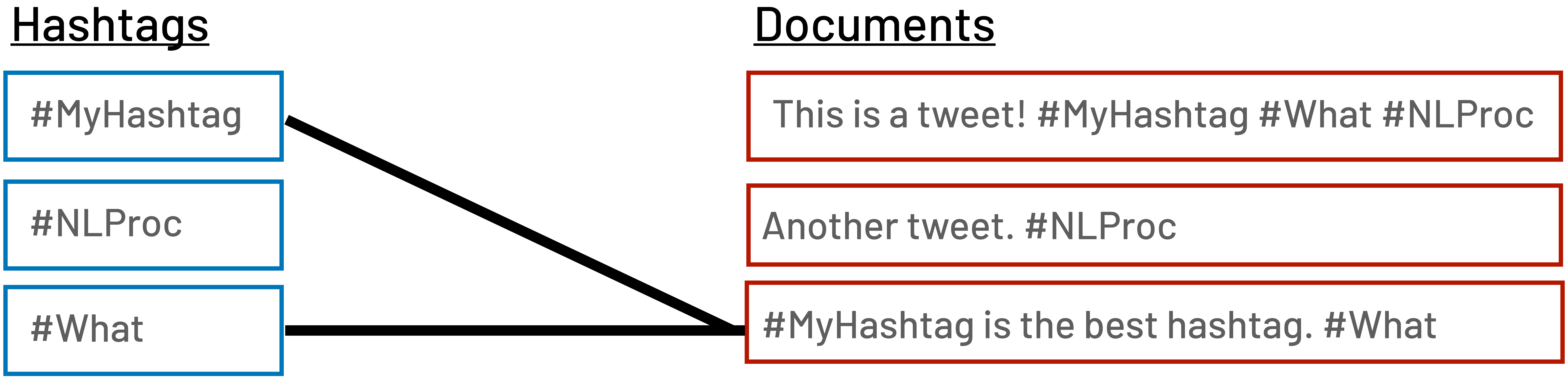
Another tweet. #NLProc

#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



Hashtags

#MyHashtag

#NLProc

#What

Documents

This is a tweet! #MyHashtag #What #NLProc

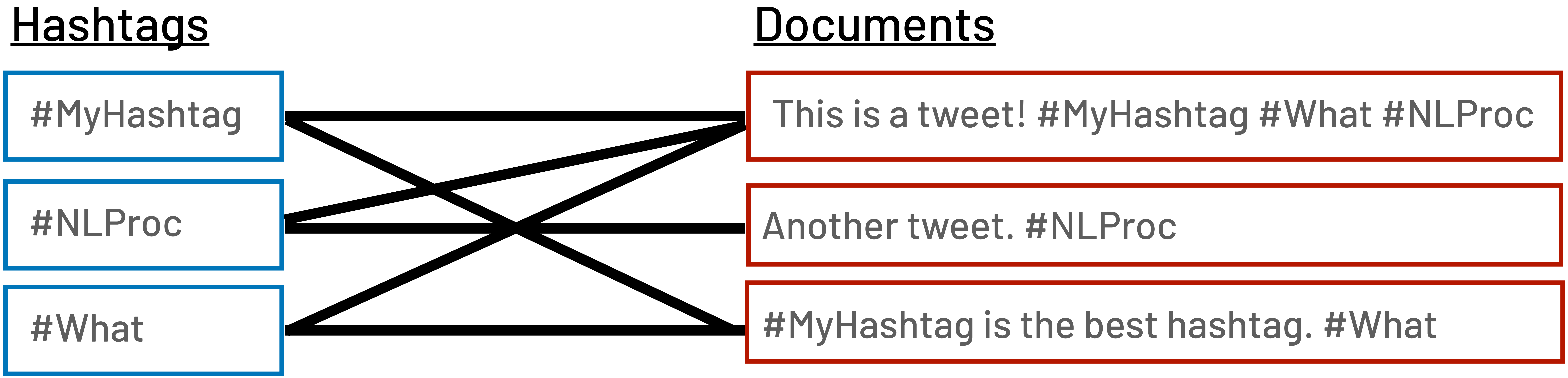
Another tweet. #NLProc

#MyHashtag is the best hashtag. #What

Example 1: Hashtag Co-Occurrence

Technically, this a **bipartite** network

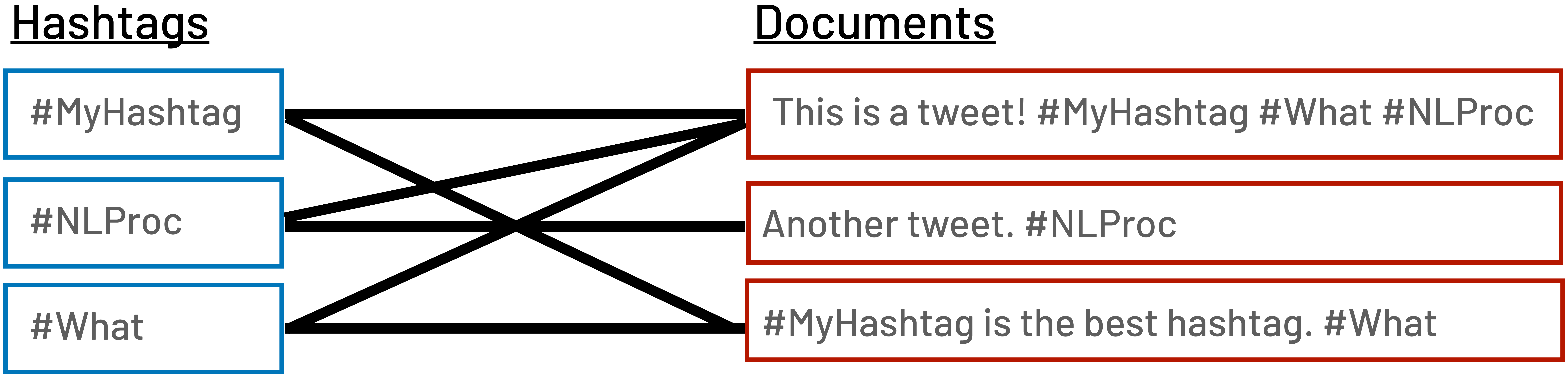
- Two types of nodes
- Only connect to nodes of **other** type



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

Example 1: Hashtag Co-Occurrence

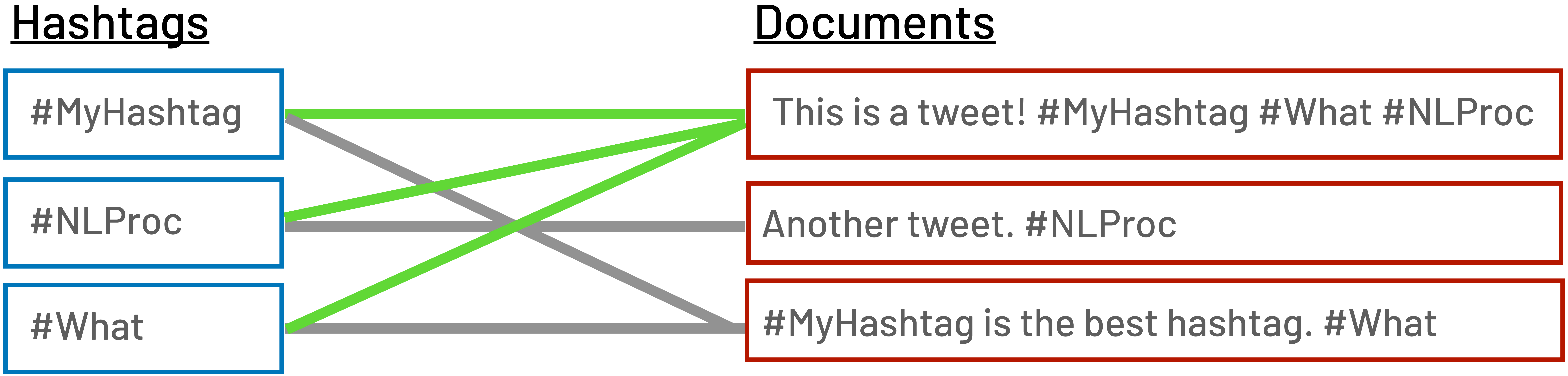
If we take the **projection** of a bipartite network on one of the nodes sets, it gives us co-occurence:



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

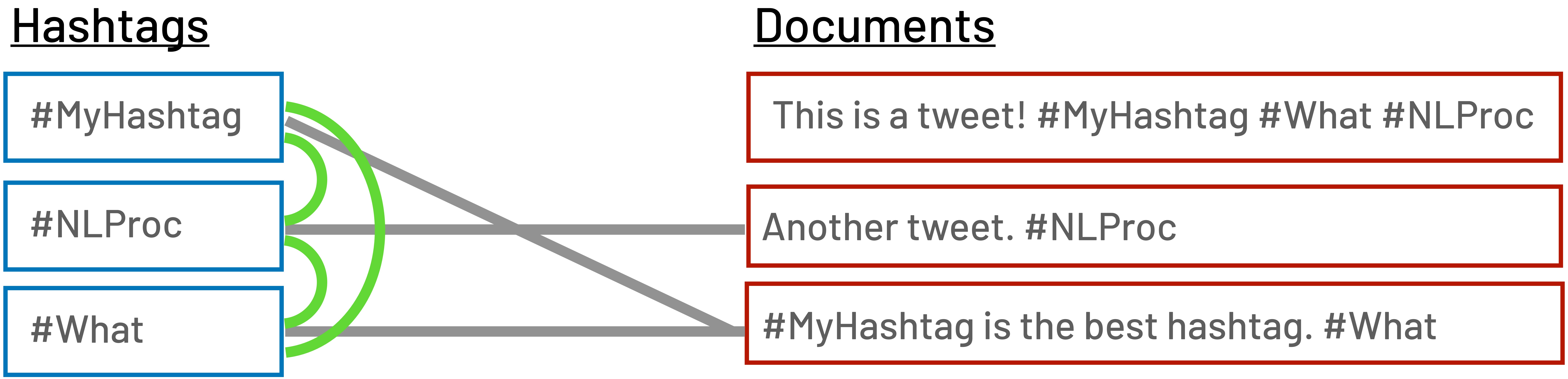
Example 1: Hashtag Co-Occurrence

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


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

Example 1: Hashtag Co-Occurrence

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



Hashtags

#MyHashtag

#NLProc

#What

Documents

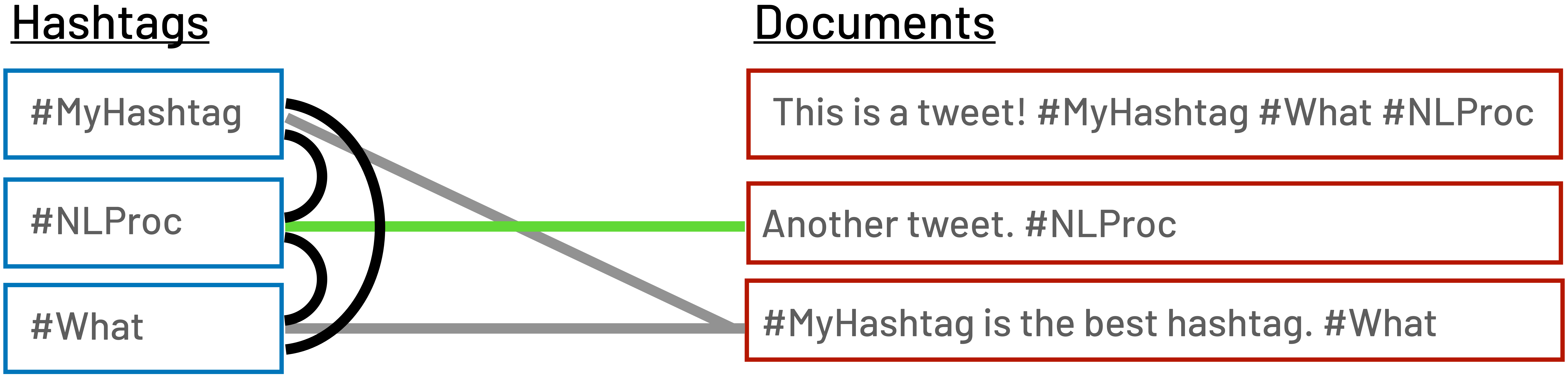
This is a tweet! #MyHashtag #What #NLProc

Another tweet. #NLProc

#MyHashtag is the best hashtag. #What

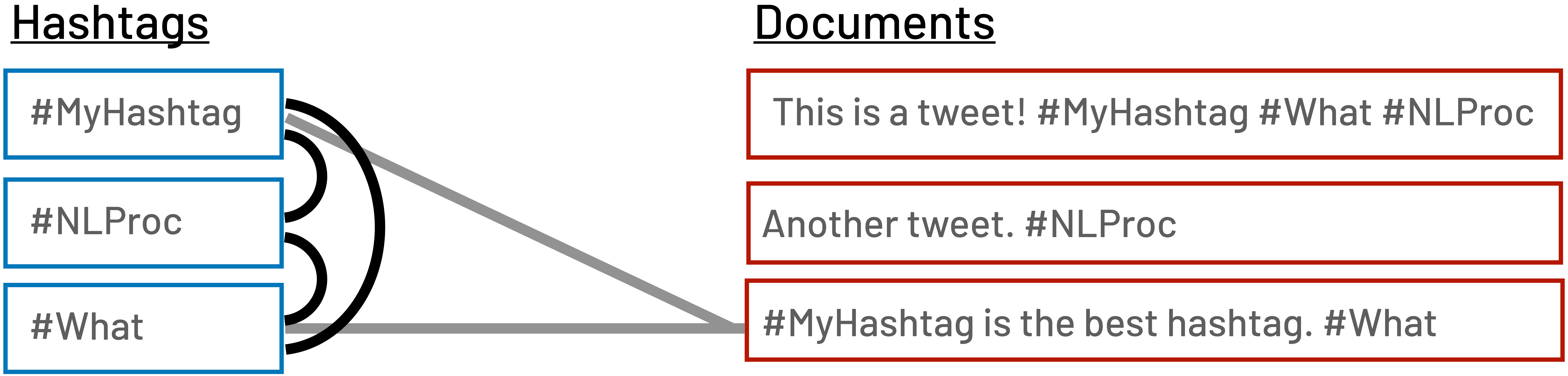
Example 1: Hashtag Co-Occurrence

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


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

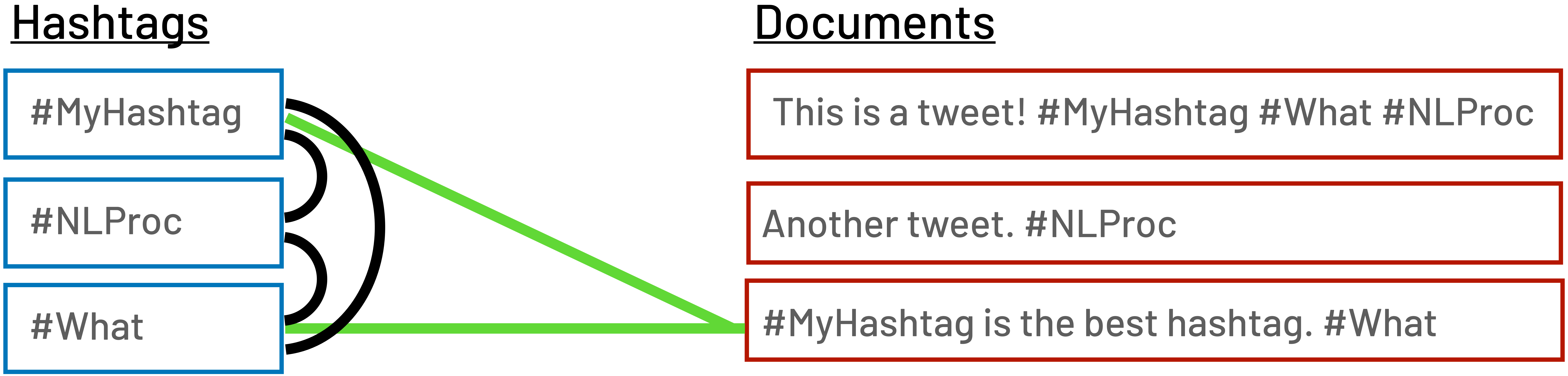
Example 1: Hashtag Co-Occurrence

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


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

Example 1: Hashtag Co-Occurrence

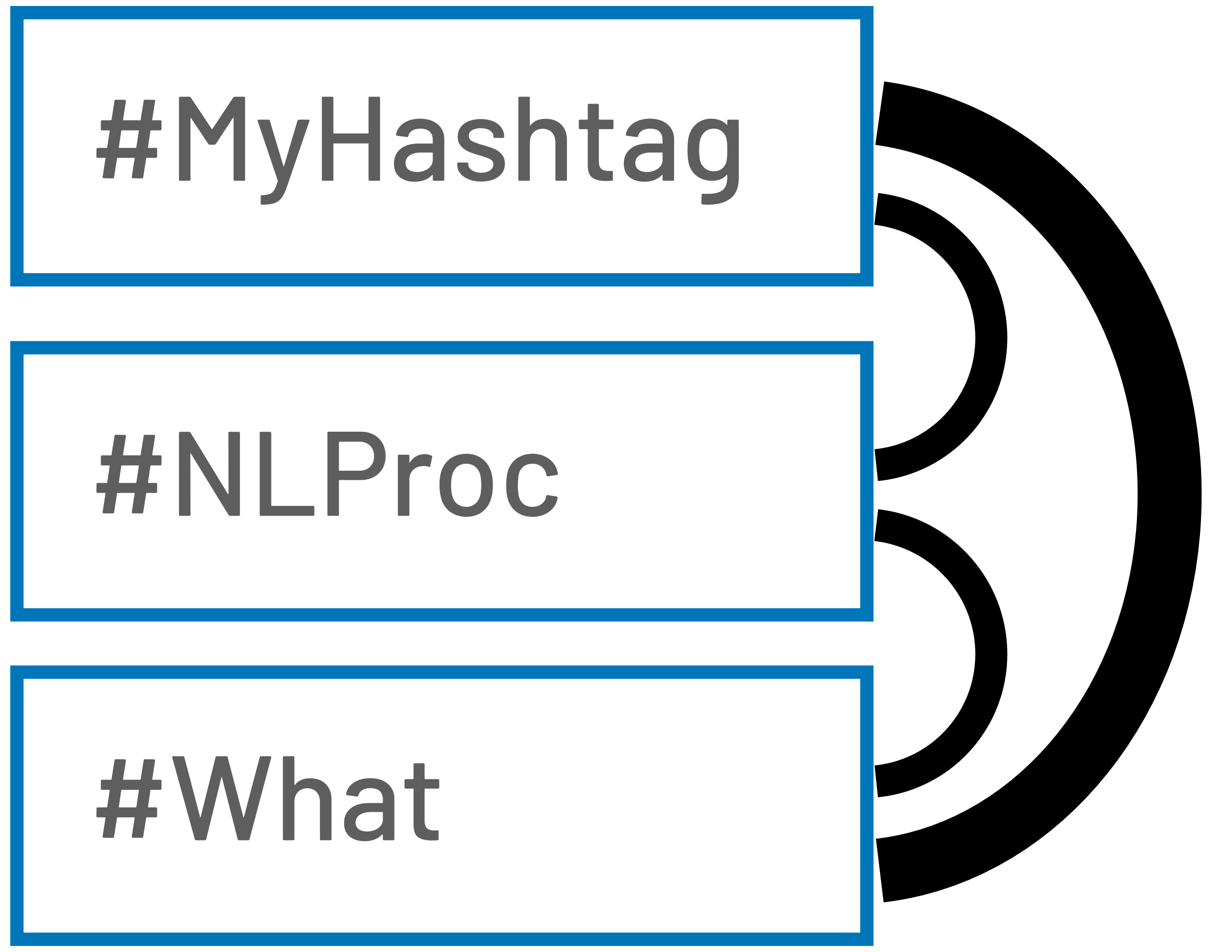
If we take the **projection** of a bipartite network on one of the nodes sets, it gives us co-occurance:


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

Example 1: Hashtag Co-Occurrence

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

<u>Hashtags</u>



#MyHashtag

#NLProc

#What

<u>Documents</u>

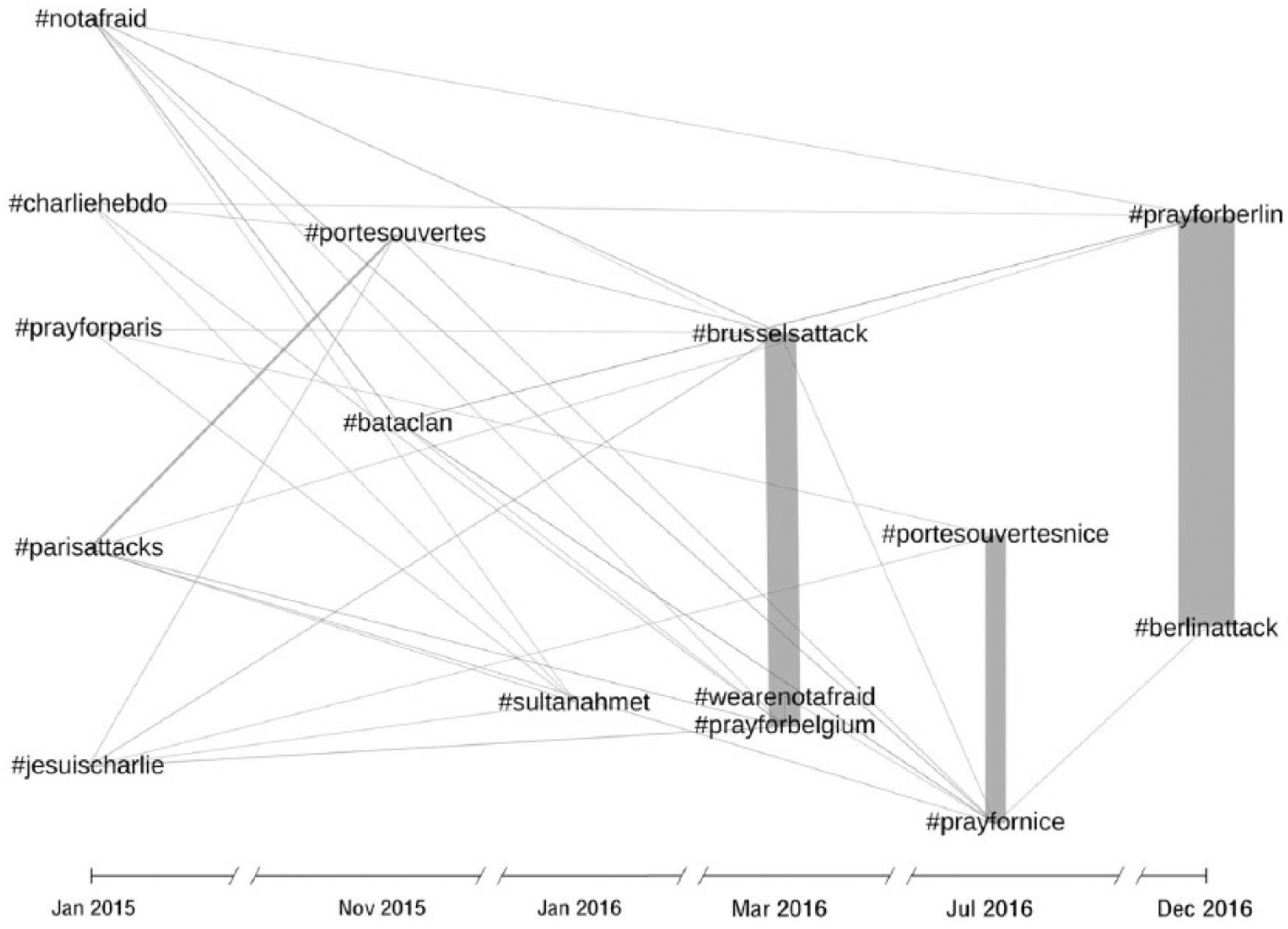
This is a tweet! #MyHashtag #What #NLProc

Another tweet. #NLProc

#MyHashtag is the best hashtag. #What

Example 1: Hashtag Co-Occurrence

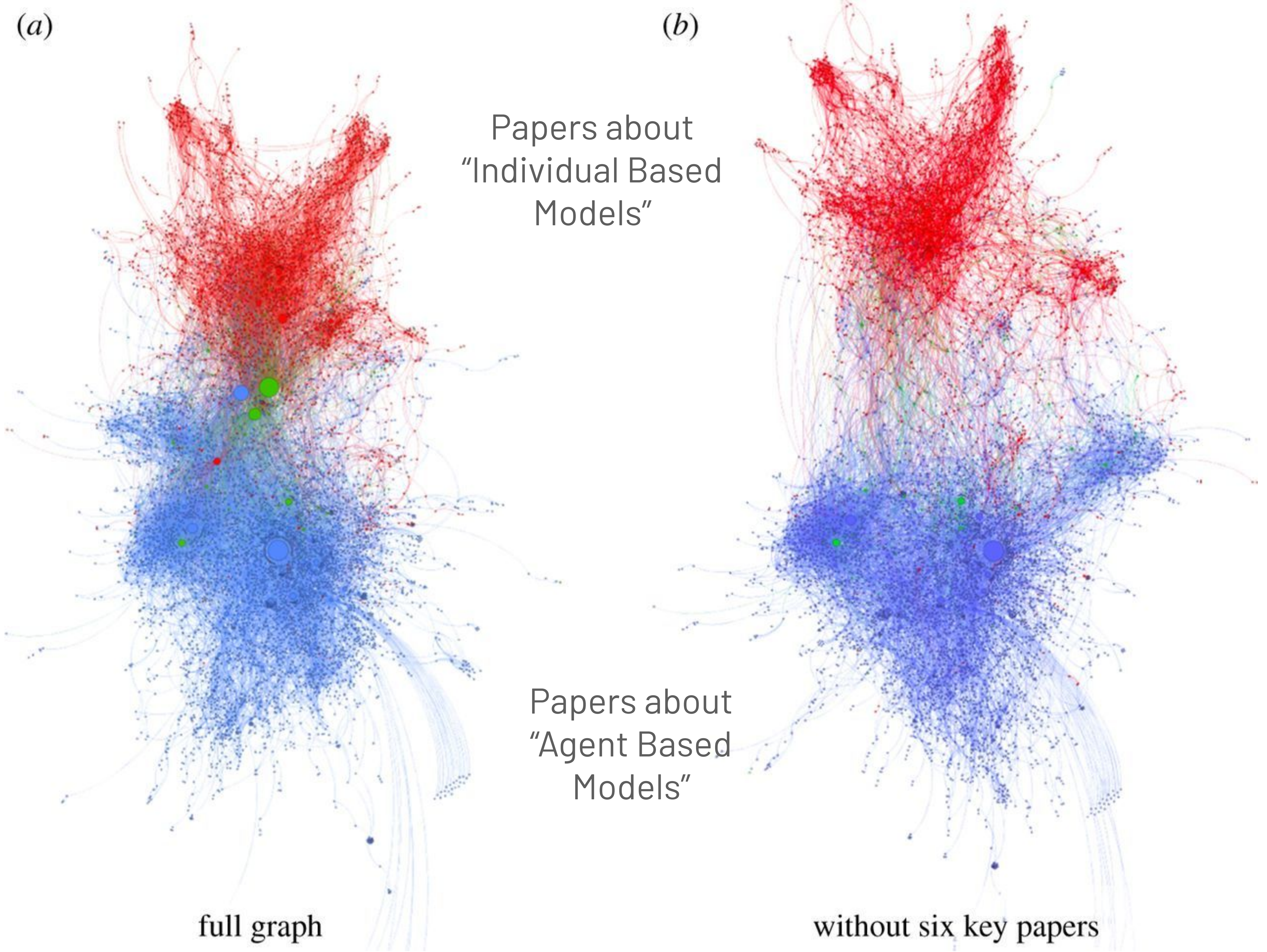
- The projection on hashtags (or words) tells us which words "go together"
  - ➡ the "discourse landscape"
  - ➡ What things are people talking about together?
  - ➡ What "discursive communities" arise?



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



Papers about “Individual Based Models”

Papers about “Agent Based Models”

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-occurence” network is (probably?) a projection of a bipartite network

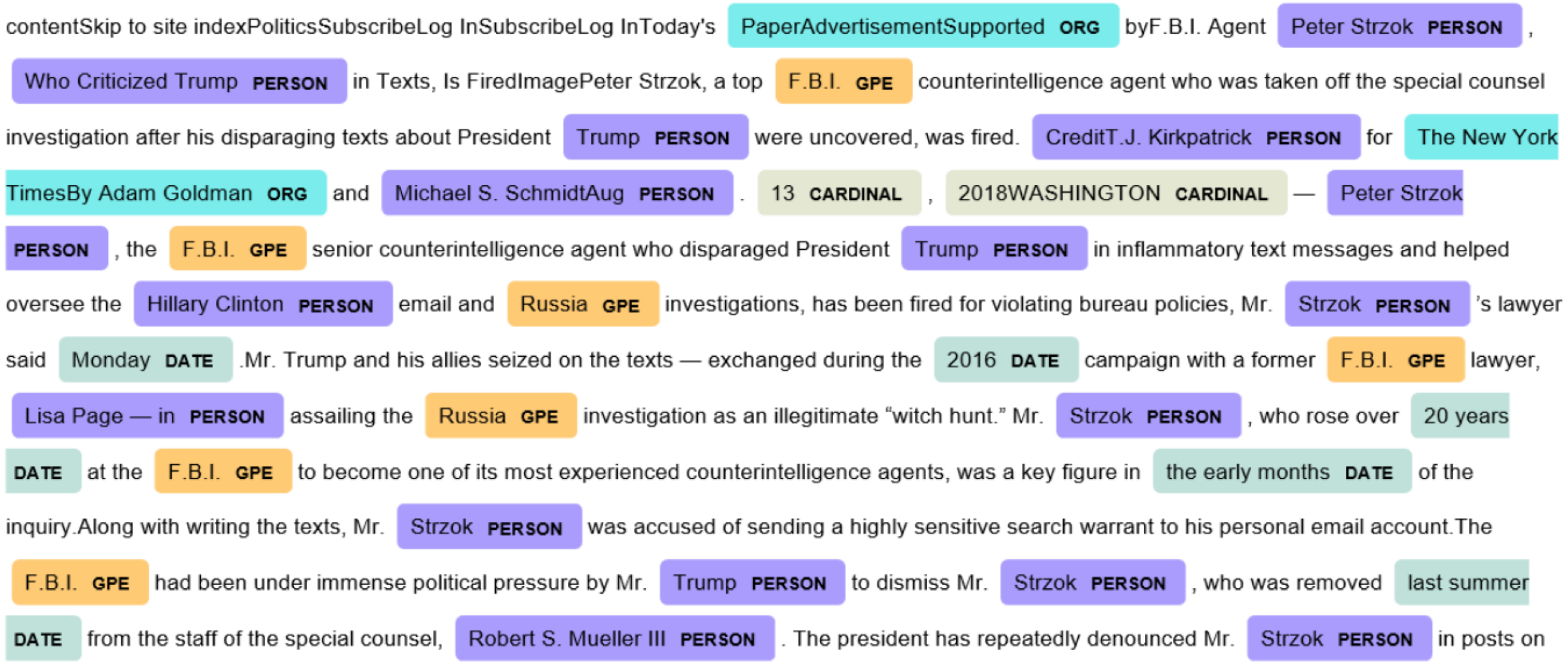
- In **code**, we can often skip the bipartite network and directly construct the co-occurence network (we’ll see this soon!)

- 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

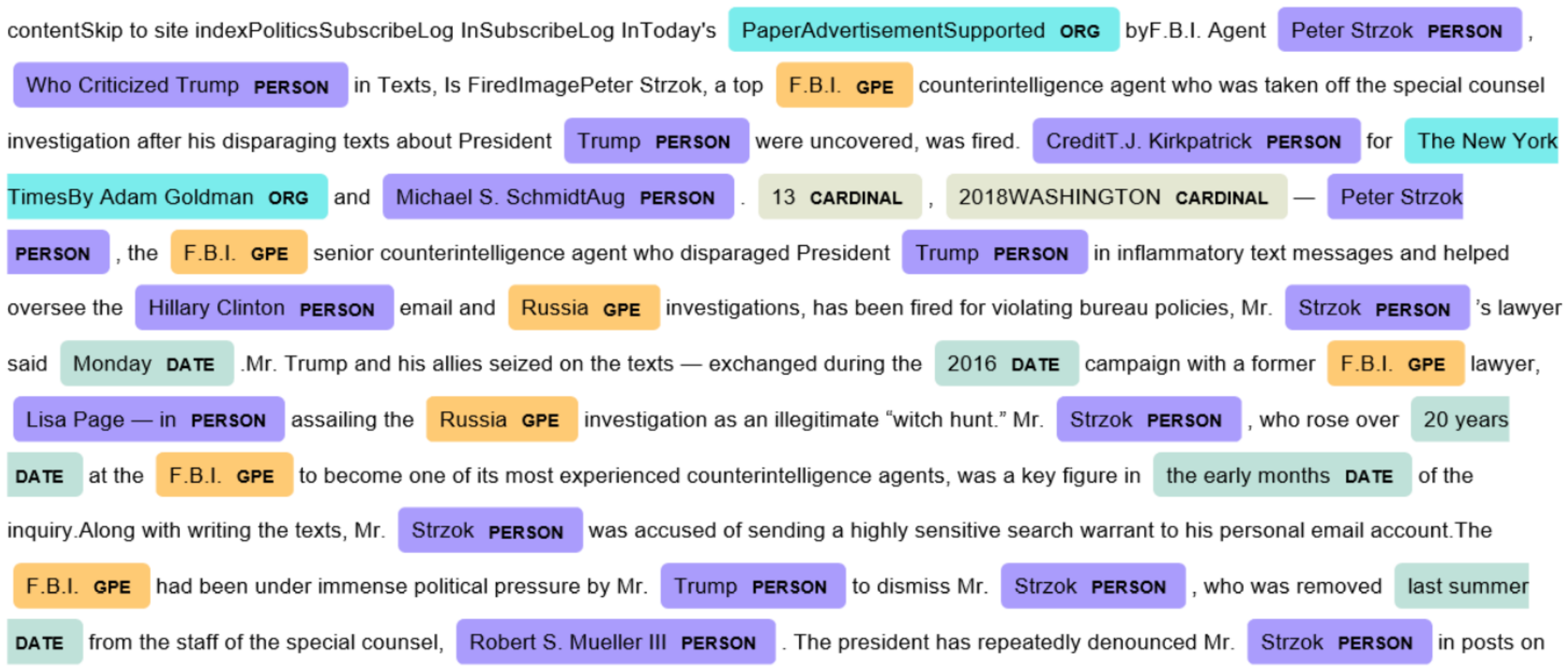
Image from https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2

Example 2: Entity Co-Occurrence

- Before we get to the NLP challenge, let’s assume that we have annotated text

The network could be:

- Based on co-occurence (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 2: Entity Co-Occurrence

- How do we identify those entities?
  - ➡ Hand coding by 2 or more coders gives the highest accuracy, but is not scalable to large corpora
  - ➡ Using trained neutral nets, NLP has gotten pretty good (~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**

Validation Approaches

- 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

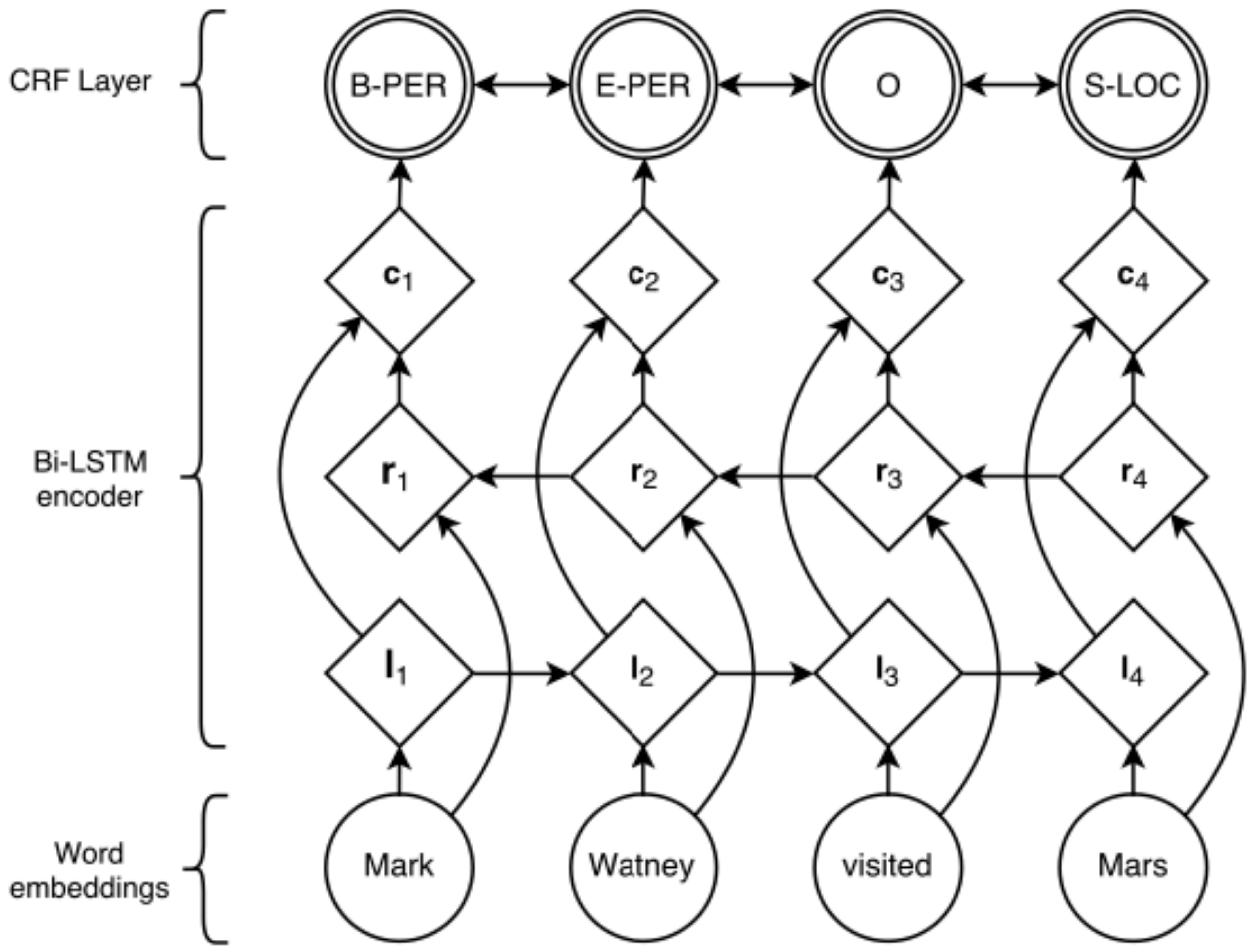
- The moral: You really have to look at your data

- 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!

Text Analysis 102: Predicting Labels

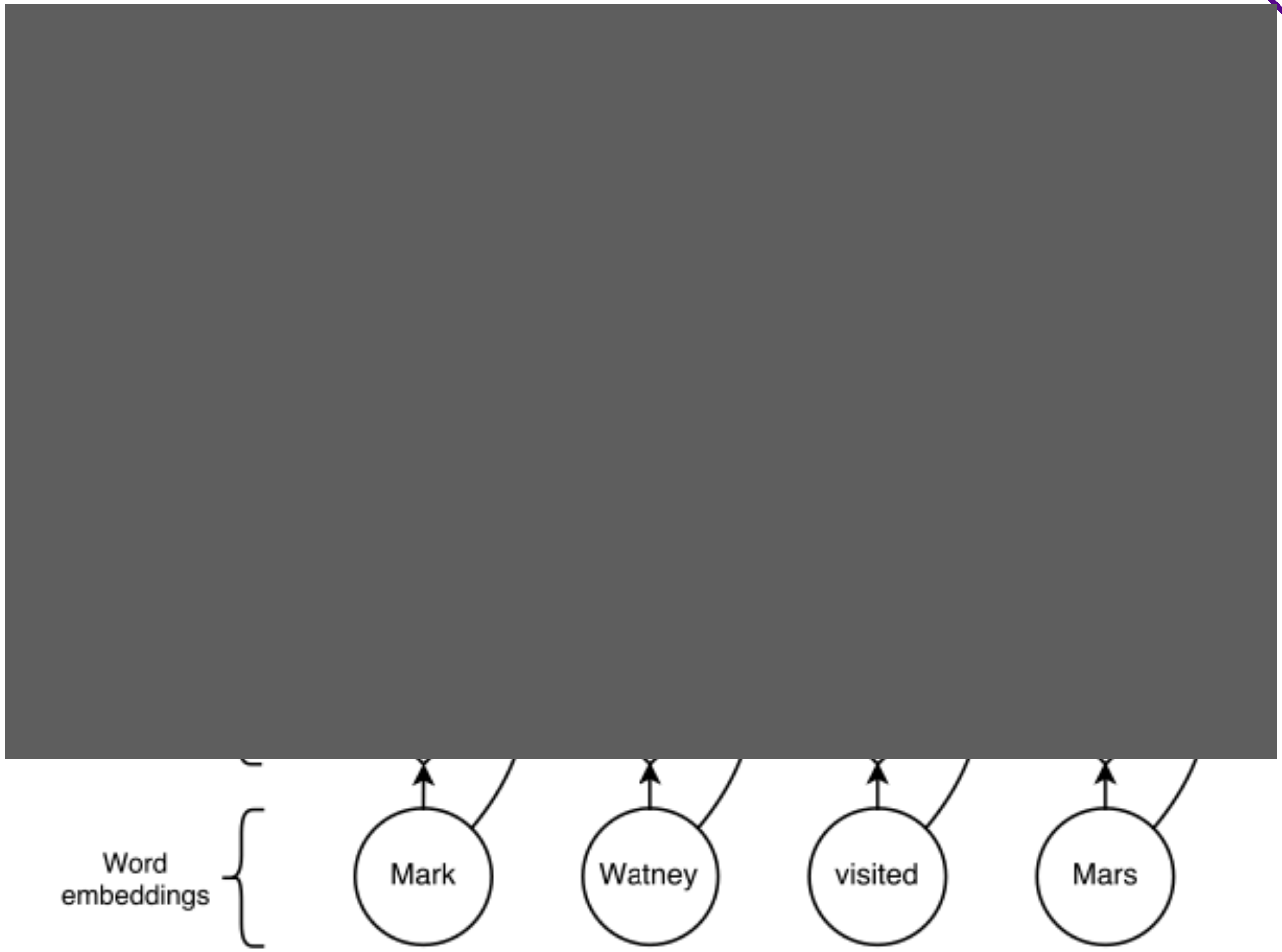
Named entity recognition in the sentence:
“Mark Watney visited Mars”



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

Text Analysis 102: Predicting Labels

Named entity recognition in the sentence:
“Mark Watney visited Mars”

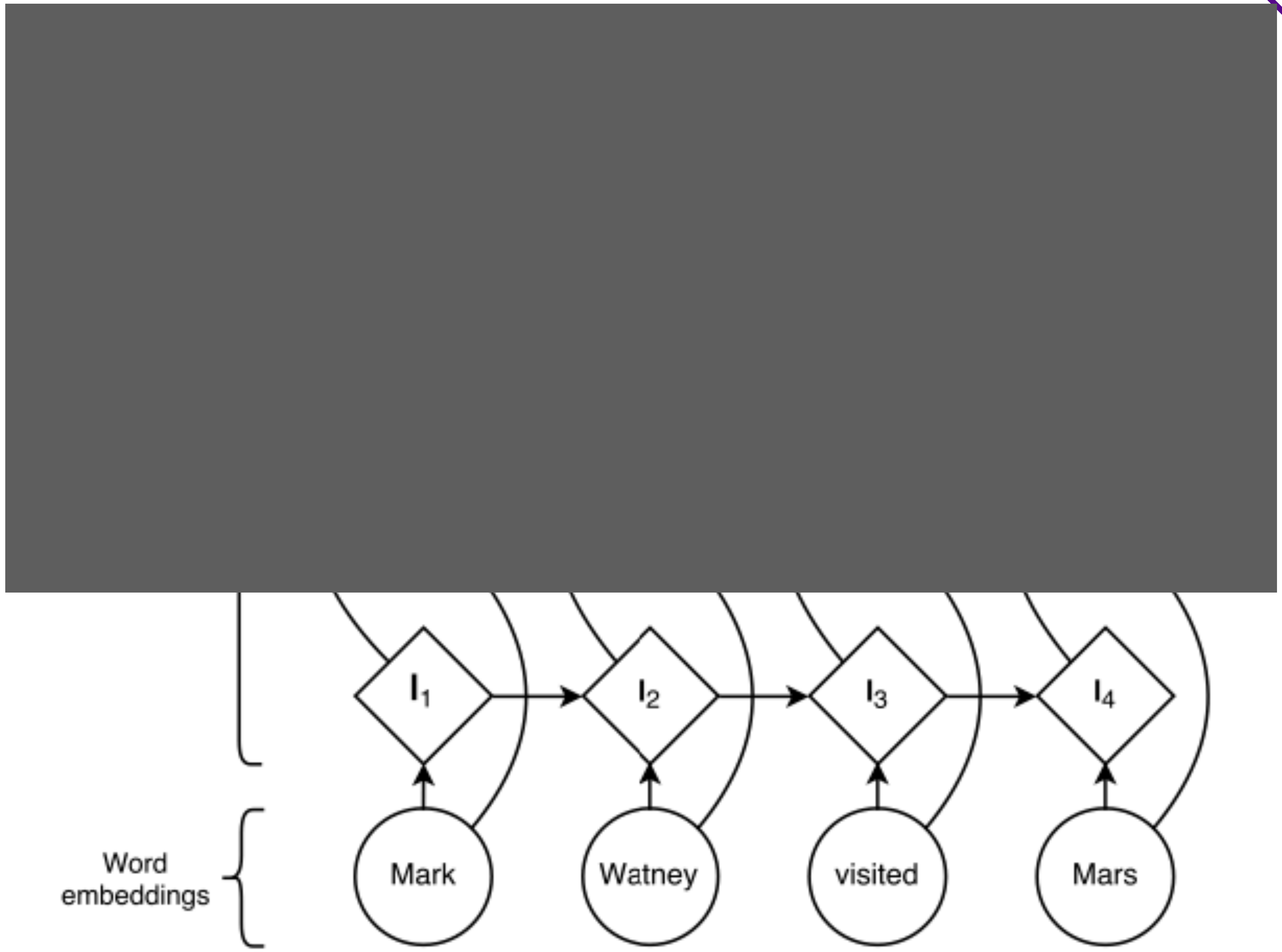


Input: N-Dimensional representation of each word
(More on what this means soon!)

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

Text Analysis 102: Predicting Labels

Named entity recognition in the 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.

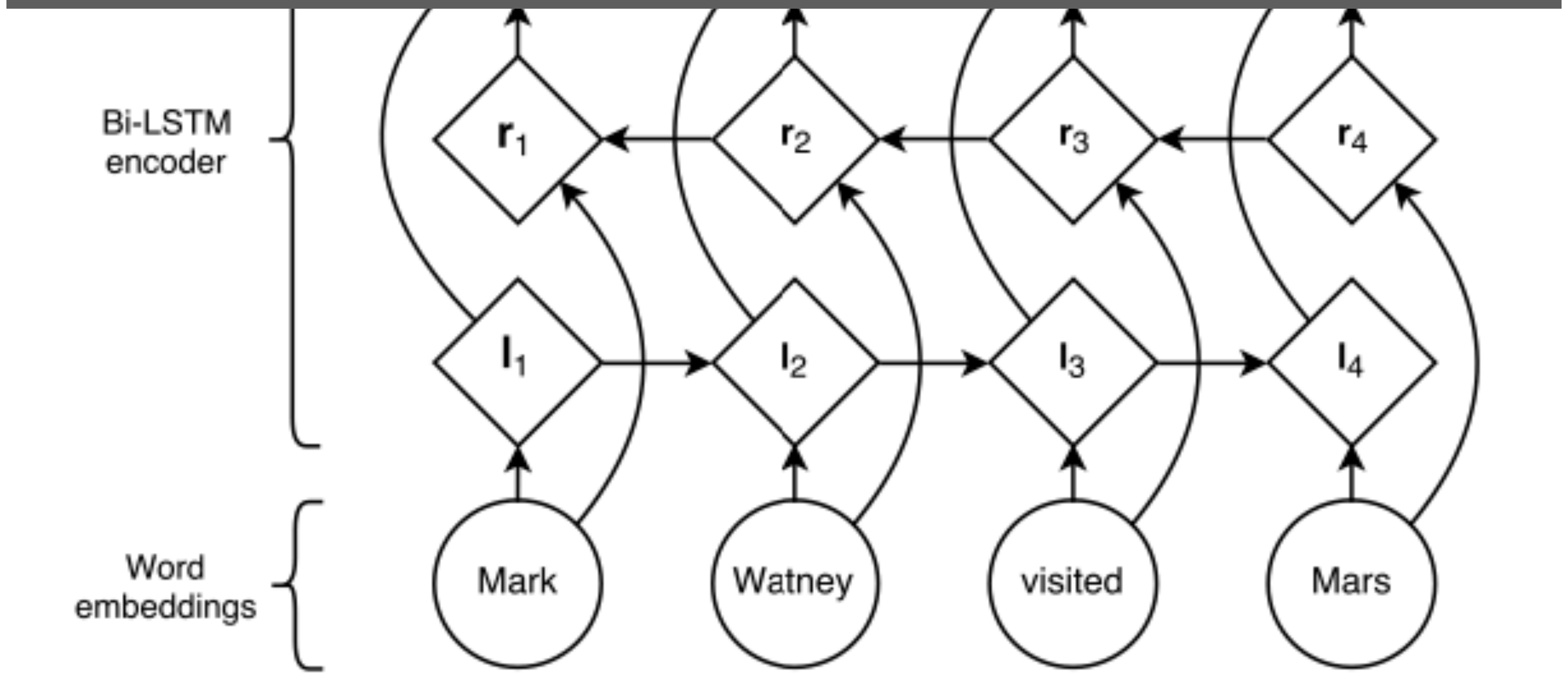
Text Analysis 102: Predicting Labels

Named entity recognition in the 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 102: Predicting Labels

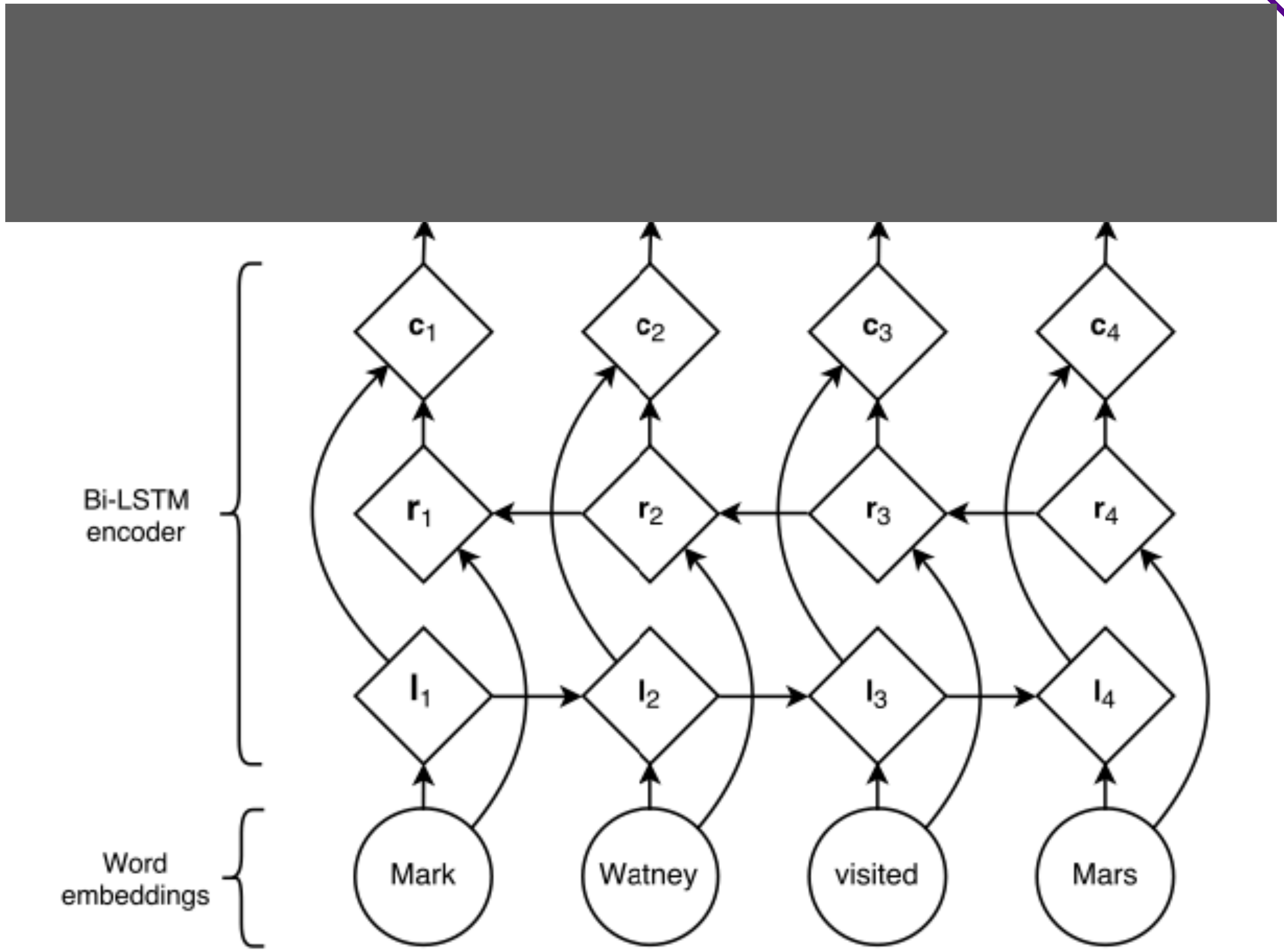
Named entity recognition in the sentence:
"Mark Watney visited Mars"

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 102: Predicting Labels

Named entity recognition in the sentence:
“Mark Watney visited Mars”

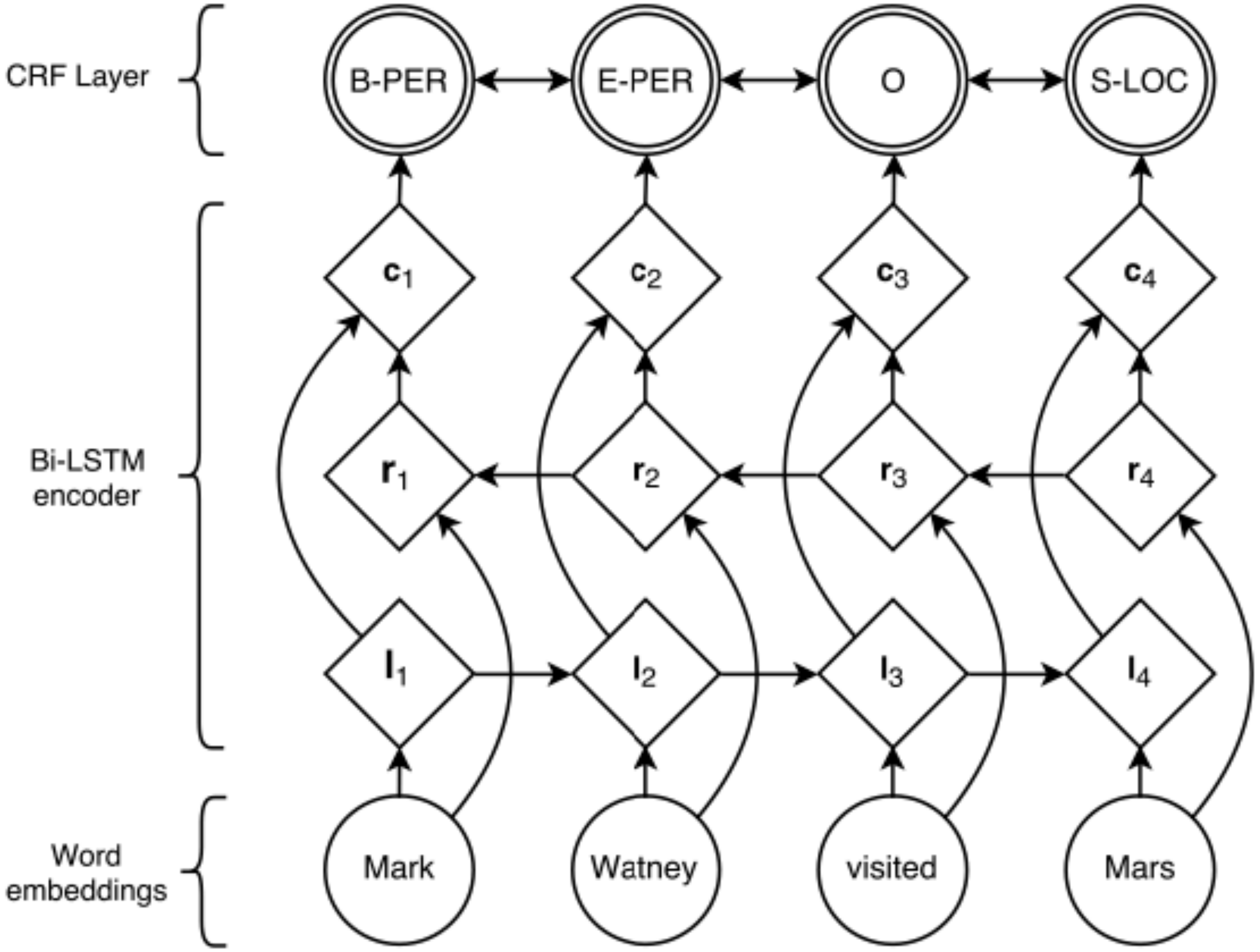
Jointly predict each word’s label (or tag)

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 Analysis 102: Context and Embeddings

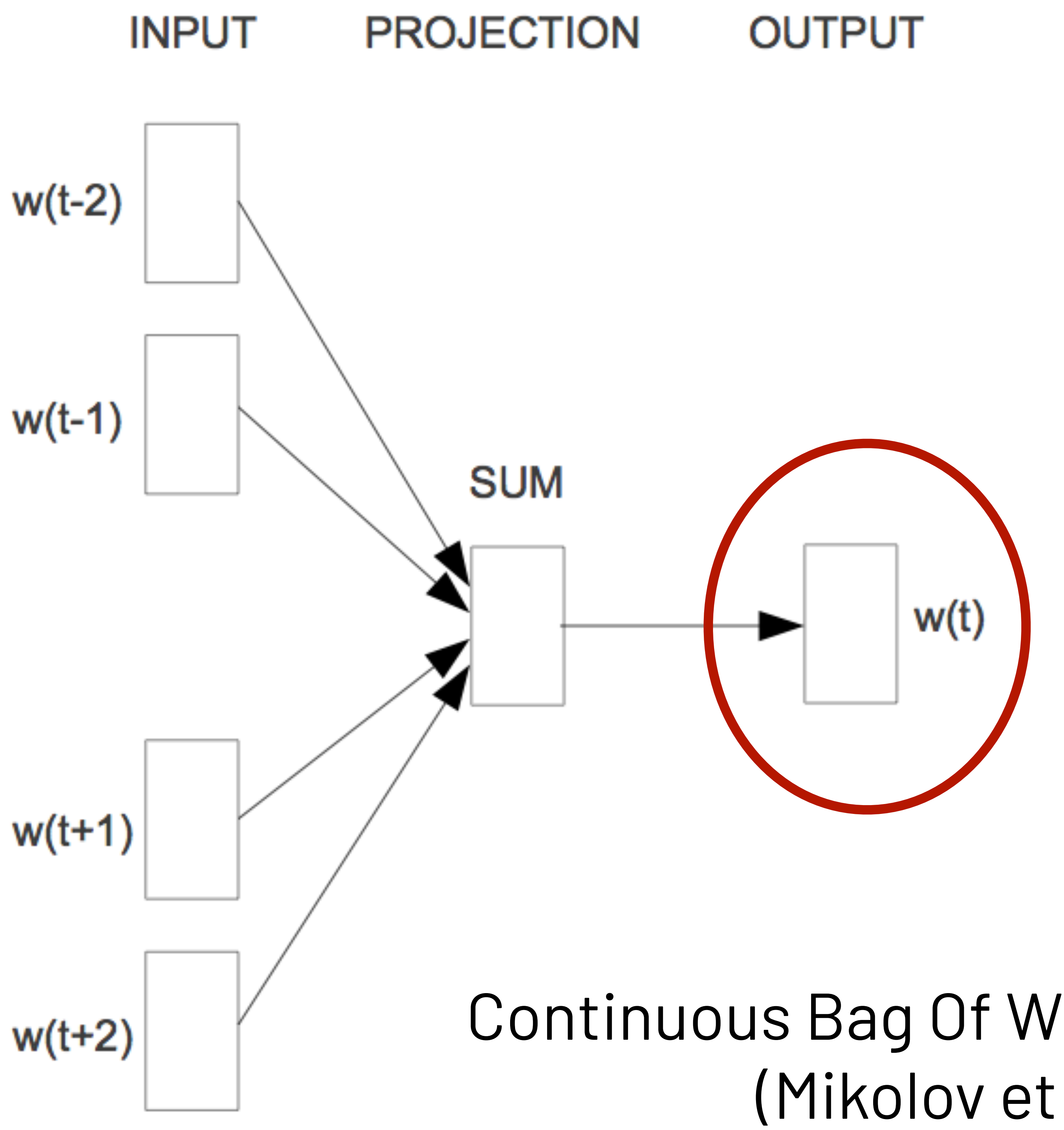
"You shall know a word by the company it keeps"
Firth (1954)

Highly recommend:
Pedro Rodriguez and Arthur Spirling (Forthcoming)
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

Text Analysis 102: Context and Embeddings

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

Highly recommend:
Pedro Rodriguez and Arthur Spirling (Forthcoming)
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

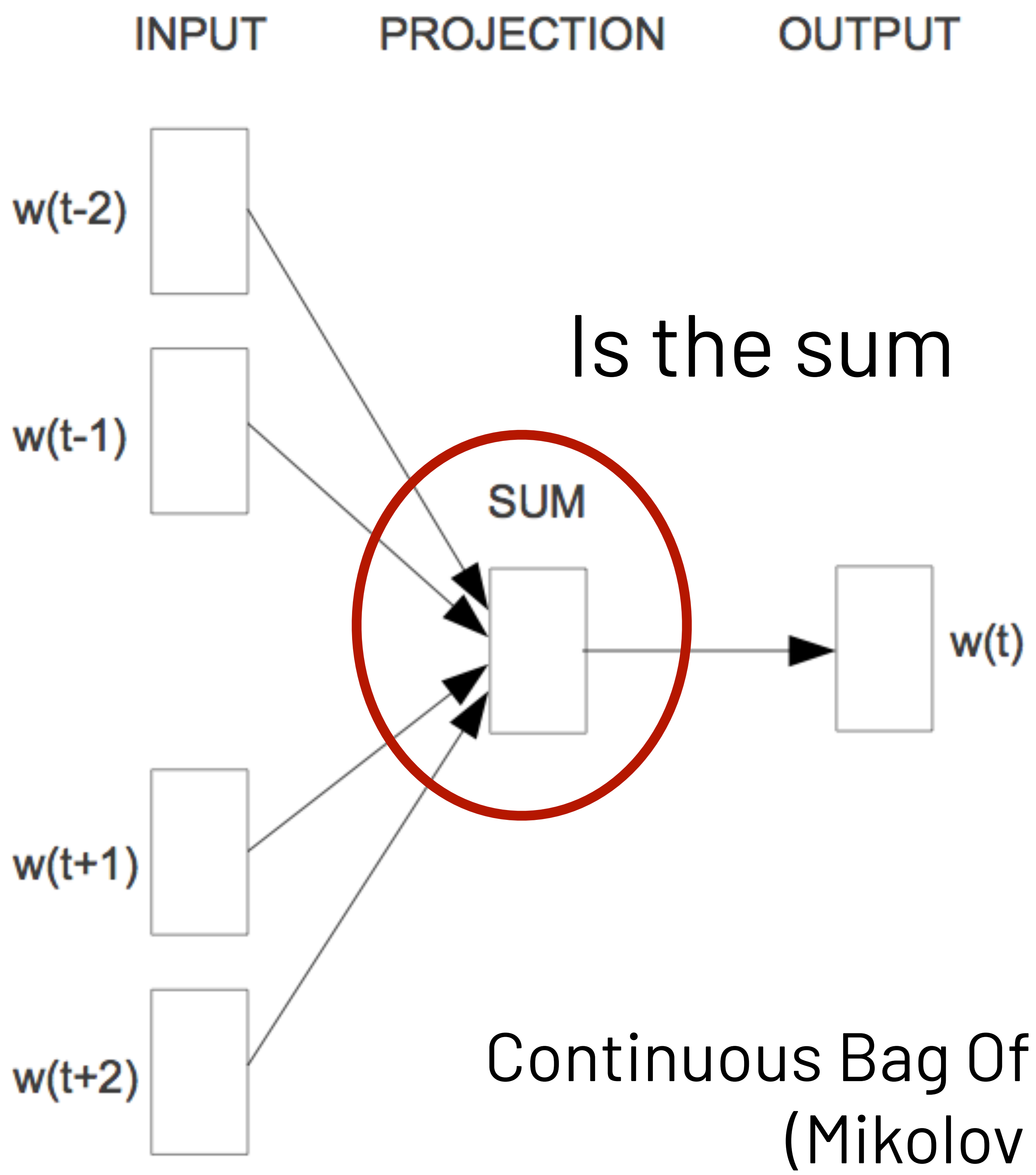


“Embedding” (vector) for target word t

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

Text Analysis 102: Context and Embeddings

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



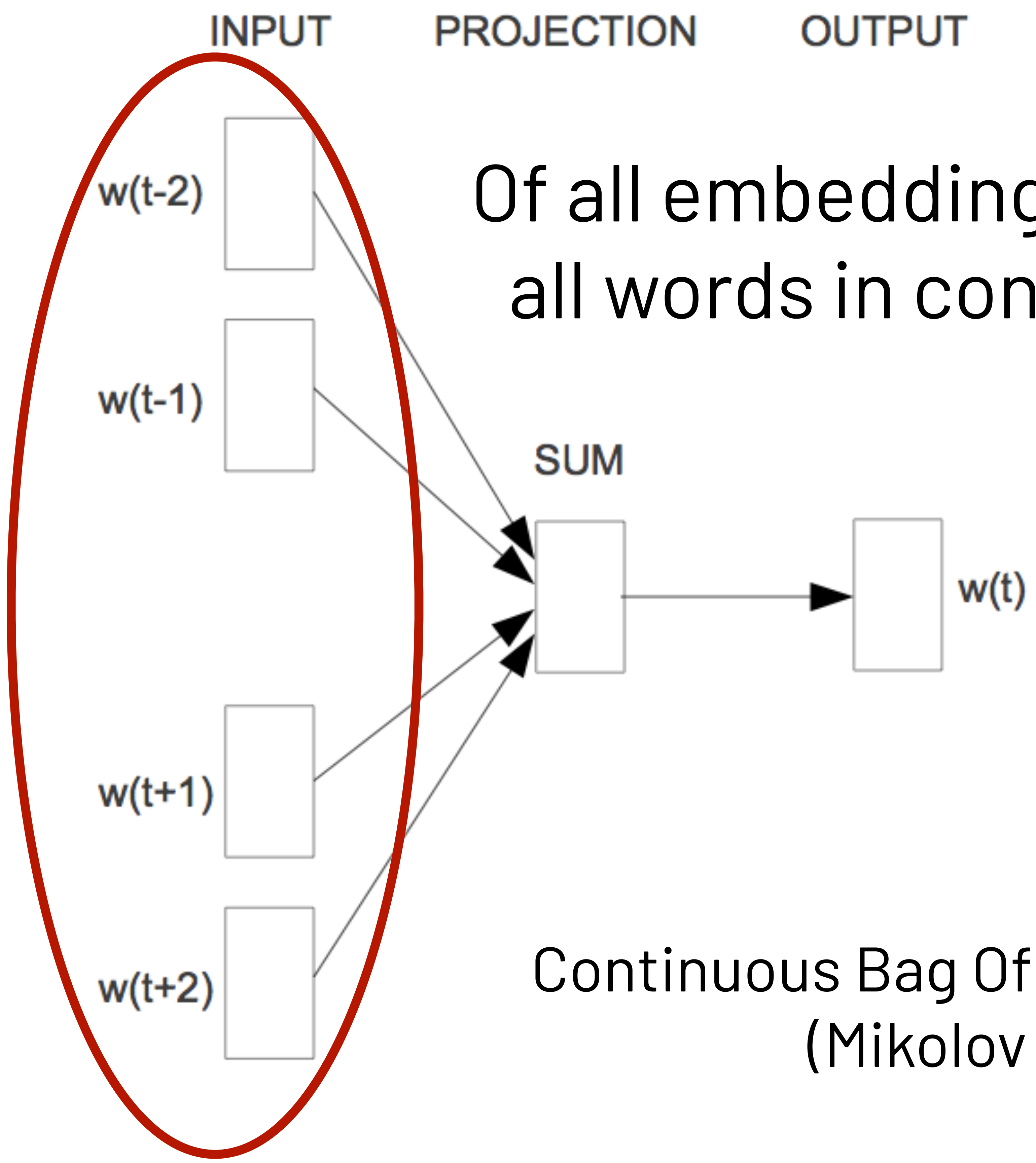
Is the sum

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

Highly recommend:
Pedro Rodriguez and Arthur Spirling (Forthcoming)
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

Text Analysis 102: Context and Embeddings

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



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

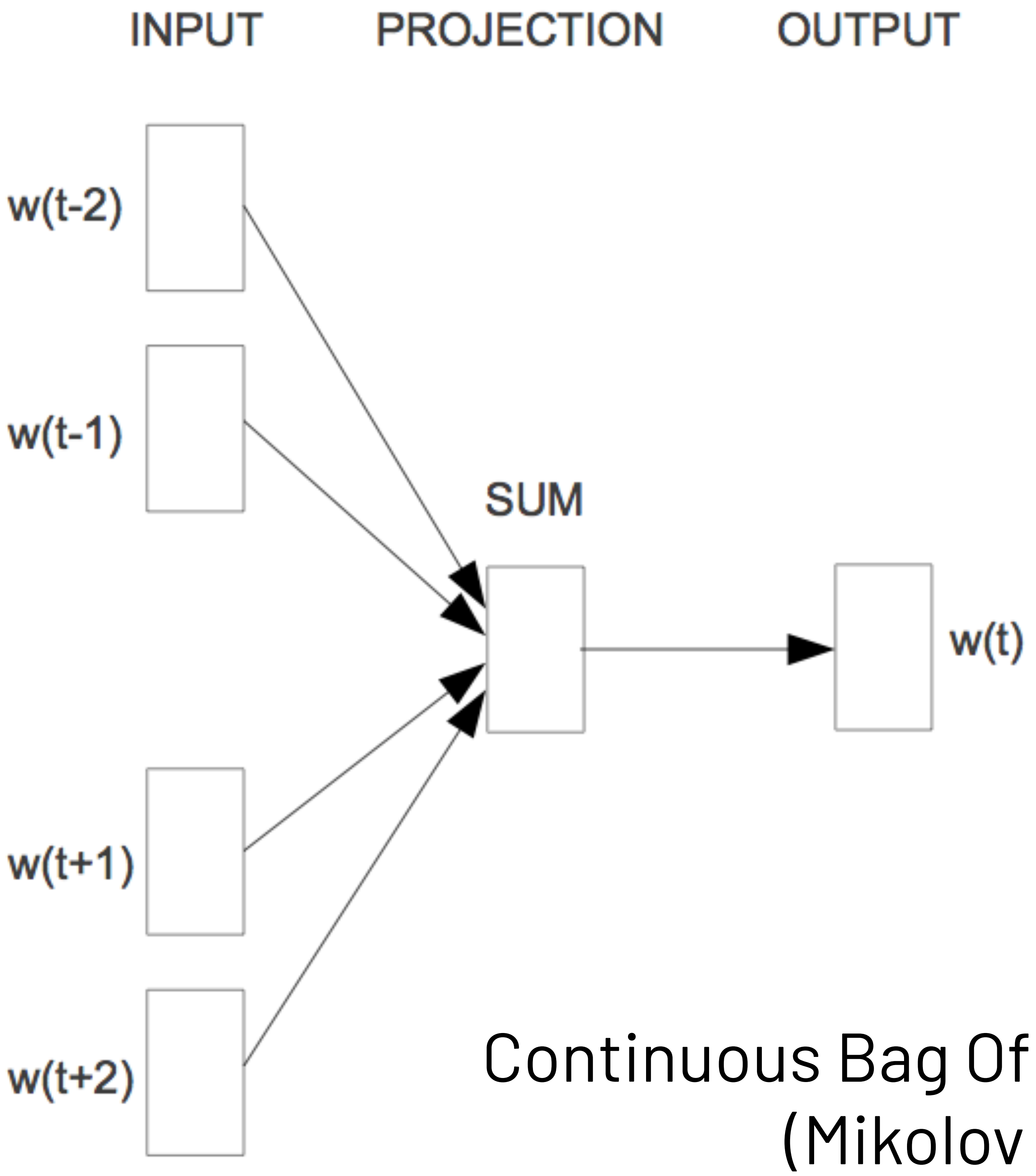
Continuous Bag Of Words (CBOW) model
(Mikolov et al., 2013)

Highly recommend:
Pedro Rodriguez and Arthur Spirling (Forthcoming)
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

$$J_\theta = \frac{1}{T} \sum_{t=1}^{T} \log p(w_t \,|\, w_{t-n}, \ldots w_{t-1}, w_{t+1}, \ldots, w_{t+n})$$

Text Analysis 102: Context and Embeddings

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



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

Highly recommend:
Pedro Rodriguez and Arthur Spirling (Forthcoming)
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

Choose n-length embeddings such that:

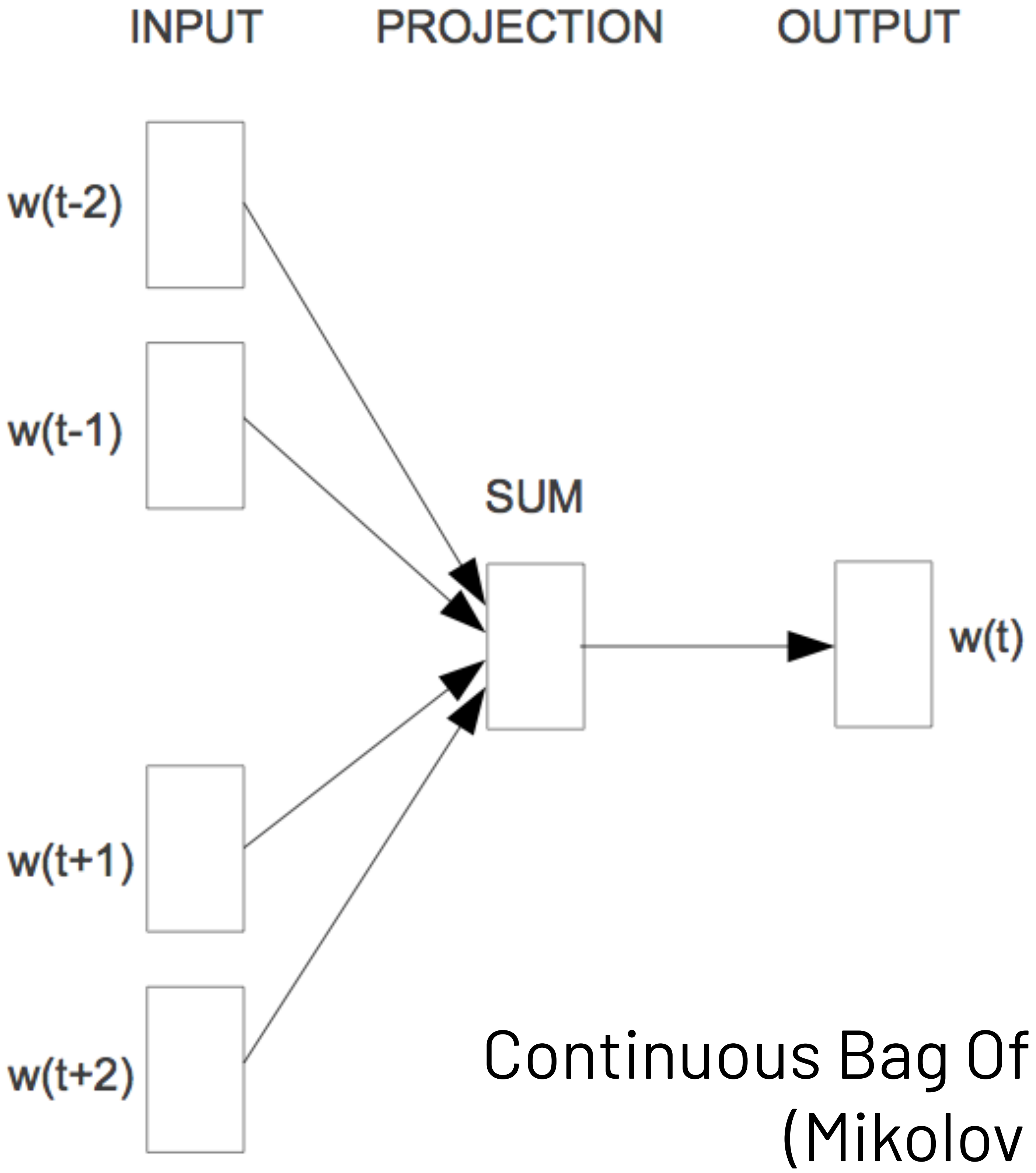
$$J_\theta = \frac{1}{T} \sum_{t=1}^{T} \log p(w_t \,|\, w_{t-n}, \ldots w_{t-1}, w_{t+1}, \ldots, w_{t+n})$$

Probability of seeing word t
given context words

Text Analysis 102: Context and Embeddings

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

Highly recommend:
Pedro Rodriguez and Arthur Spirling (Forthcoming) 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



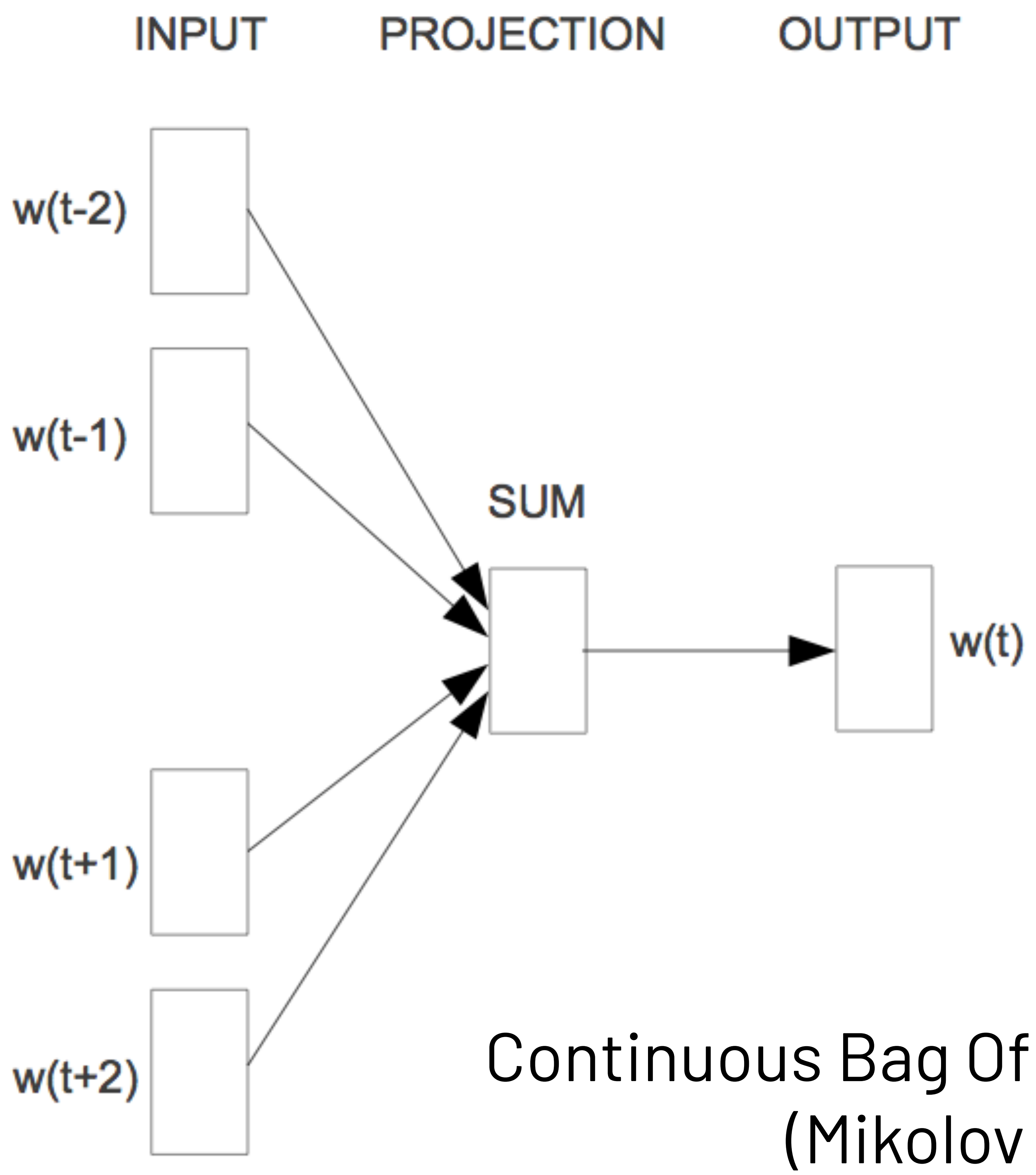
Continuous Bag Of Words (CBOW) model
(Mikolov et al., 2013)

Choose n-length embeddings such that:

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

Text Analysis 102: Context and Embeddings

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



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

Highly recommend:
Pedro Rodriguez and Arthur Spirling (Forthcoming)
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

Choose n-length embeddings such that:

$$J_\theta = \underbrace{\frac{1}{T} \sum_{t=1}^{T}}_{\text{Averaged over all words T}} \log p(w_t \mid w_{t-n}, \ldots w_{t-1}, w_{t+1}, \ldots, w_{t+n})$$

A Note on Fancy Models

- Almost always optimized to (roughly) maximize the probability of seeing words together

- How do we know the probability of seeing words together?

- This is what happens in the pre-training: the model “looks at” a bunch of labeled data and (essentially) counts how frequently different types of things co-occur. This is how the model “learns” that a verb commonly follows a noun (in English).

- The moral: Most machine learning is fancy counting

Text Analysis 102: Context and Embeddings

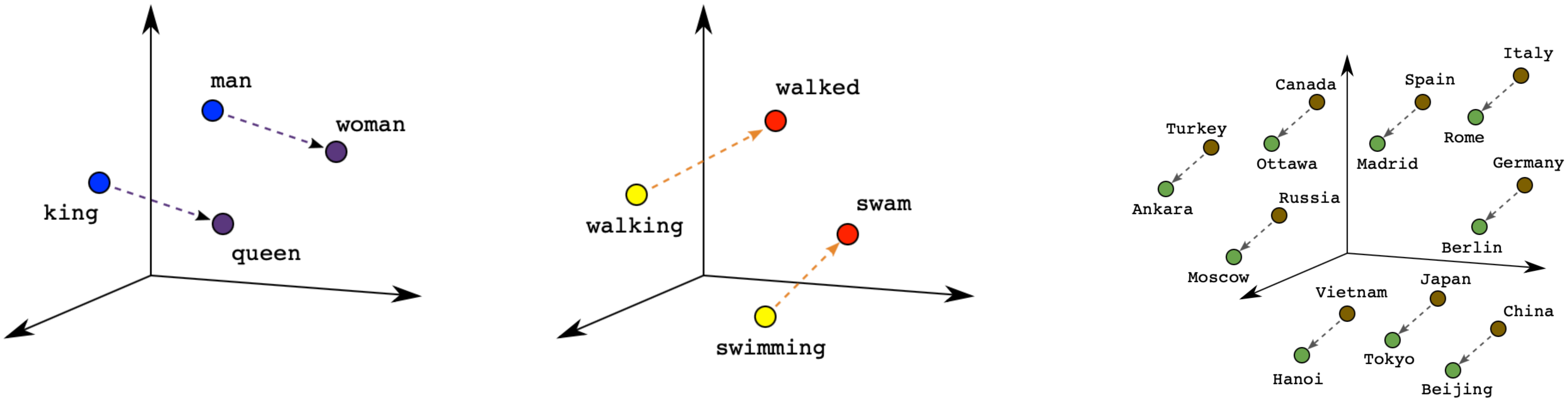
$$J_\theta = \frac{1}{T} \sum_{t=1}^{T} \log p(w_t \mid w_{t-n}, \ldots w_{t-1}, w_{t+1}, \ldots, w_{t+n})$$

Probability of seeing word t
given context words,
averaged over all words T

- 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

Text Analysis 102: Context and Embeddings

- Embeddings will reflect bias of training language



Country-Capital

Male-Female

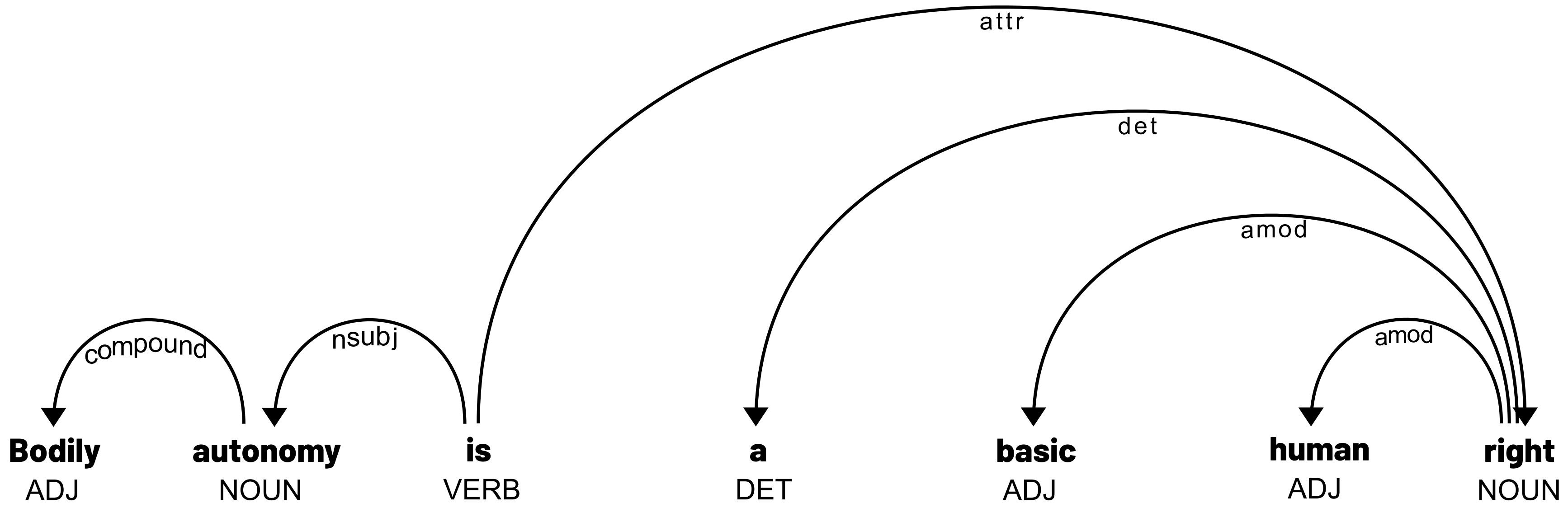
Verb Tense

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

Example 3: Finding Connected Concepts

1. Use word embeddings to identify similar words

2. Use semantic parse tree to identify connections



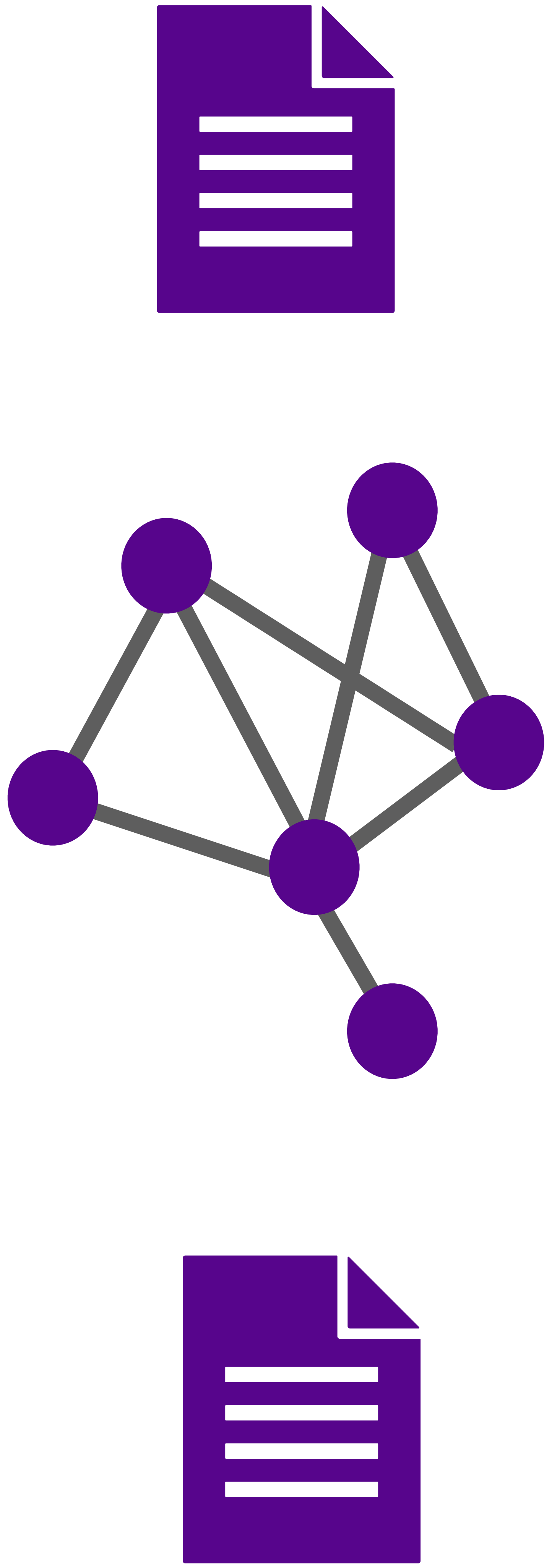
I've included some example code (that we may not have time for) but the high level take away is that there is LOTS of room for innovation and creativity!

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!