Information Extraction

CS6200

Information Retrieval

(and a sort of advertisement for NLP in the spring)

Information Extraction

- Automatically extract structure from text
  - annotate document using tags to identify extracted structure
- We've briefly mentioned one example
  - But part of speech tagging is so low-level it usually doesn't count as IE
- Named entity recognition
  - identify words that refer to something of interest in a particular application
  - e.g., people, companies, locations, dates, product names, prices, etc.

Named Entity Recognition

Fred Smith, who lives at 10 Water Street, Springfield, MA, is a long-time collector of **tropical fish.**

<PersonName><GivenName>Fred</GivenName> <Sn>Smith</Sn> </PersonName>, who lives at <address><Street >10 Water Street</Street>, <City>Springfield</City>, <State>MA</State></address>, is a long-time collector of tropical fish.

- Example showing semantic annotation of text using XML tags
- Information extraction also includes document structure and more complex features such as relationships and events

Named Entity Recognition

The Persian learned men say that the Phoenicians … came to our seas from the so-called Red Sea, and having settled in the country which they still occupy, at once began to make long voyages. Among other places to which they carried Egyptian and Assyrian merchandise, they came to Argos, which was at that time preeminent in every way among the people of what is now called Hellas. The Phoenicians came to Argos, and set out their cargo. On the fifth or sixth day after their arrival, when their wares were almost all sold, many women came to the shore and among them especially the daughter of the king, whose name was Io (according to Persians and Greeks alike), the daughter of Inachus. As these stood about the stern of the ship bargaining for the wares they liked, the Phoenicians incited one another to set upon them. Most of the women escaped: Io and others were seized and thrown into the ship, which then sailed away for Egypt.

Named Entity Recognition

The Persian learned men say that the Phoenicians … came to our seas from the so-called Red Sea, and having settled in the country which they still occupy, at once began to make long voyages. Among other places to which they carried Egyptian and Assyrian merchandise, they came to Argos, which was at that time preeminent in every way among the people of what is now called Hellas. The Phoenicians came to Argos, and set out their cargo. On the fifth or sixth day after their arrival, when their wares were almost all sold, many women came to the shore and among them especially the daughter of the king, whose name was Io (according to Persians and Greeks alike), the daughter of Inachus. As these stood about the stern of the ship bargaining for the wares they liked, the Phoenicians incited one another to set upon them. Most of the women escaped: Io and others were seized and thrown into the ship, which then sailed away for Egypt.

Person

Named Entity Recognition

The Persian learned men say that the Phoenicians … came to our seas from the so-called Red Sea, and having settled in the country which they still occupy, at once began to make long voyages. Among other places to which they carried Egyptian and Assyrian merchandise, they came to Argos, which was at that time preeminent in every way among the people of what is now called Hellas. The Phoenicians came to Argos, and set out their cargo. On the fifth or sixth day after their arrival, when their wares were almost all sold, many women came to the shore and among them especially the daughter of the king, whose name was Io (according to Persians and Greeks alike), the daughter of Inachus. As these stood about the stern of the ship bargaining for the wares they liked, the Phoenicians incited one another to set upon them. Most of the women escaped: Io and others were seized and thrown into the ship, which then sailed away for Egypt.

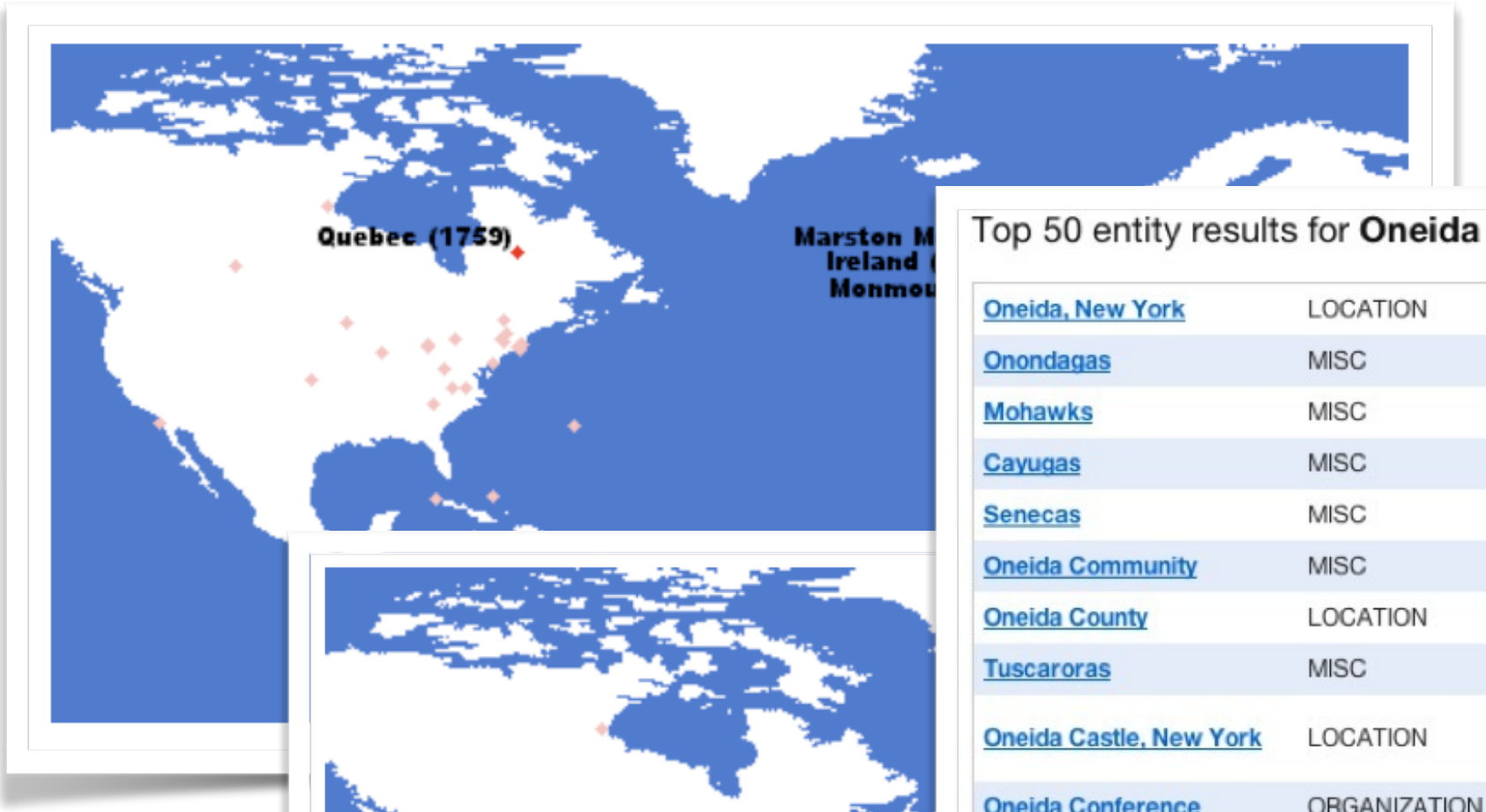
Person Location

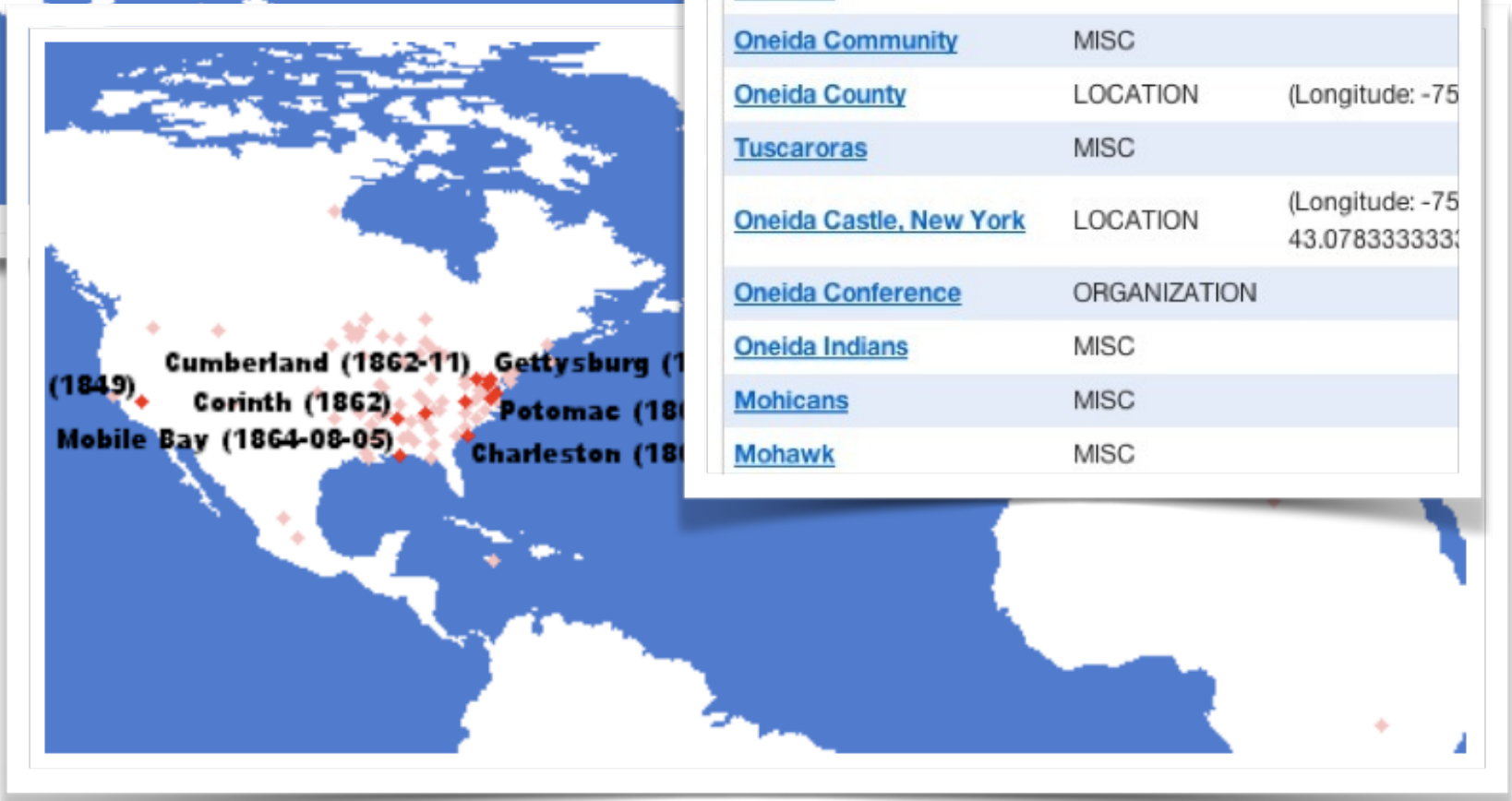
Named Entity Recognition

The Persian learned men say that the Phoenicians … came to our seas from the so-called Red Sea, and having settled in the country which they still occupy, at once began to make long voyages. Among other places to which they carried Egyptian and Assyrian merchandise, they came to Argos, which was at that time preeminent in every way among the people of what is now called Hellas. The Phoenicians came to Argos, and set out their cargo. On the fifth or sixth day after their arrival, when their wares were almost all sold, many women came to the shore and among them especially the daughter of the king, whose name was Io (according to Persians and Greeks alike), the daughter of Inachus. As these stood about the stern of the ship bargaining for the wares they liked, the Phoenicians incited one another to set upon them. Most of the women escaped: Io and others were seized and thrown into the ship, which then sailed away for Egypt.

Classes could also be, e.g., Wikipedia articles

Person Location Ethnic





Top 50 entity results for **Oneida**

Oneida, New York	LOCATION	(Longitude: -75
Onondagas	MISC	
Mohawks	MISC	
Cayugas	MISC	
Senecas	MISC	
Oneida Community	MISC	
Oneida County	LOCATION	(Longitude: -75
Tuscaroras	MISC	
Oneida Castle, New York	LOCATION	(Longitude: -75 43.078333333
Oneida Conference	ORGANIZATION	
Oneida Indians	MISC	
Mohicans	MISC	
Mohawk	MISC	

Named Entity Recognition

- *Rule-based*
  - Uses *lexicons* (lists of words and phrases) that categorize names
    - e.g., locations, peoples' names, organizations, etc.
  - Rules also used to verify or find new entity names
    - e.g., "<number> <word> street" for addresses
    - "<street address>, <city>" or "in <city>" to verify city names
    - "<street address>, <city>, <state>" to find new cities
    - "<title> <name>" to find new names

Named Entity Recognition

- Rules either developed manually by trial and error or using machine learning techniques
- *Statistical*
  - uses a probabilistic model of the words in and around an entity
  - probabilities estimated using *training data* (manually annotated text)
  - Hidden Markov Model (HMM) is one approach
    - Conditional Random Fields: similar structure, often higher accuracy, more expensive to train

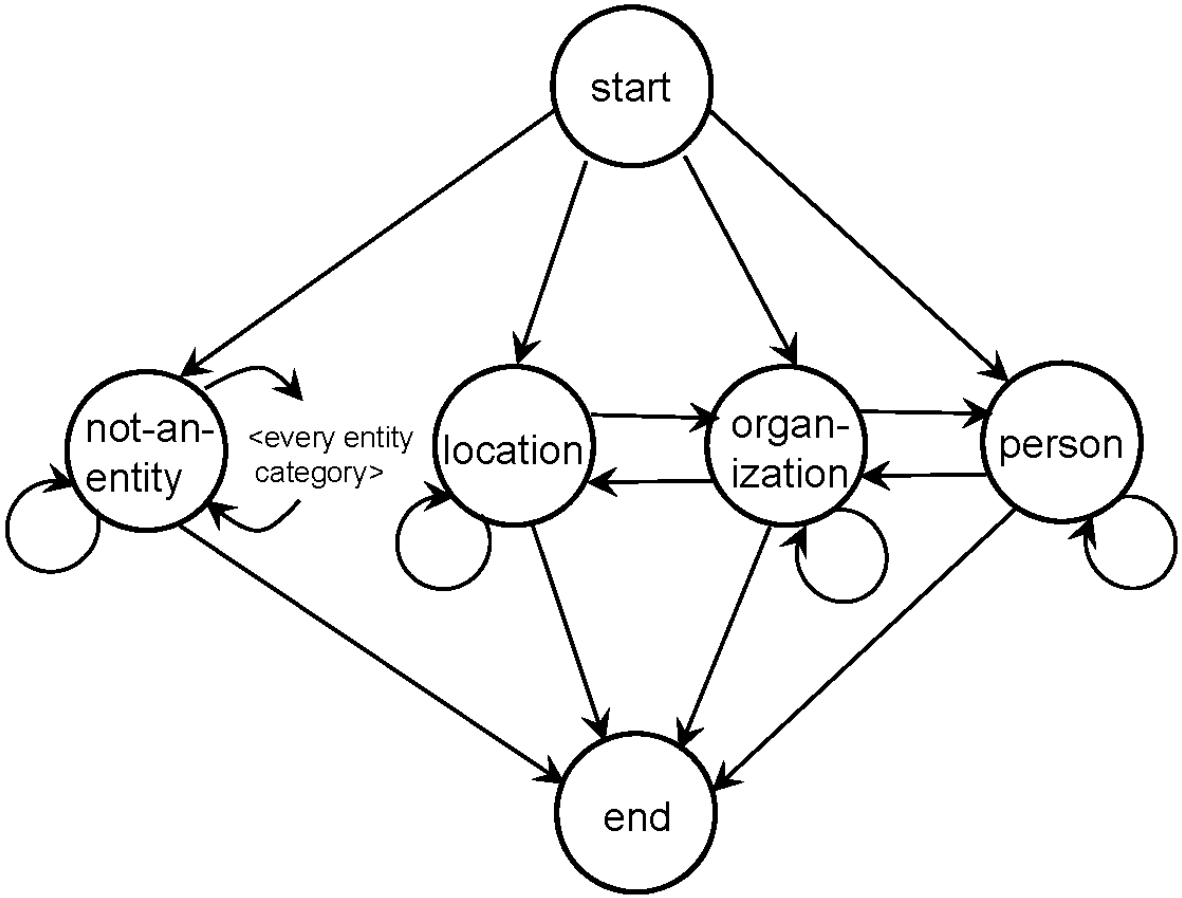
HMM for Extraction

- Resolve ambiguity in a word using *context*
  - e.g., "marathon" is a location or a sporting event, "boston marathon" is a specific sporting event
- Model context using a *generative* model of the sequence of words
  - *Markov property*: the next word in a sequence depends only on a small number of the previous words

HMM for Extraction

- *Markov Model* describes a process as a collection of states with transitions between them
  - each transition has a probability associated with it
  - next state depends only on current state and transition probabilities
- *Hidden Markov Model*
  - each state has a set of possible outputs
  - outputs have probabilities

HMM Sentence Model

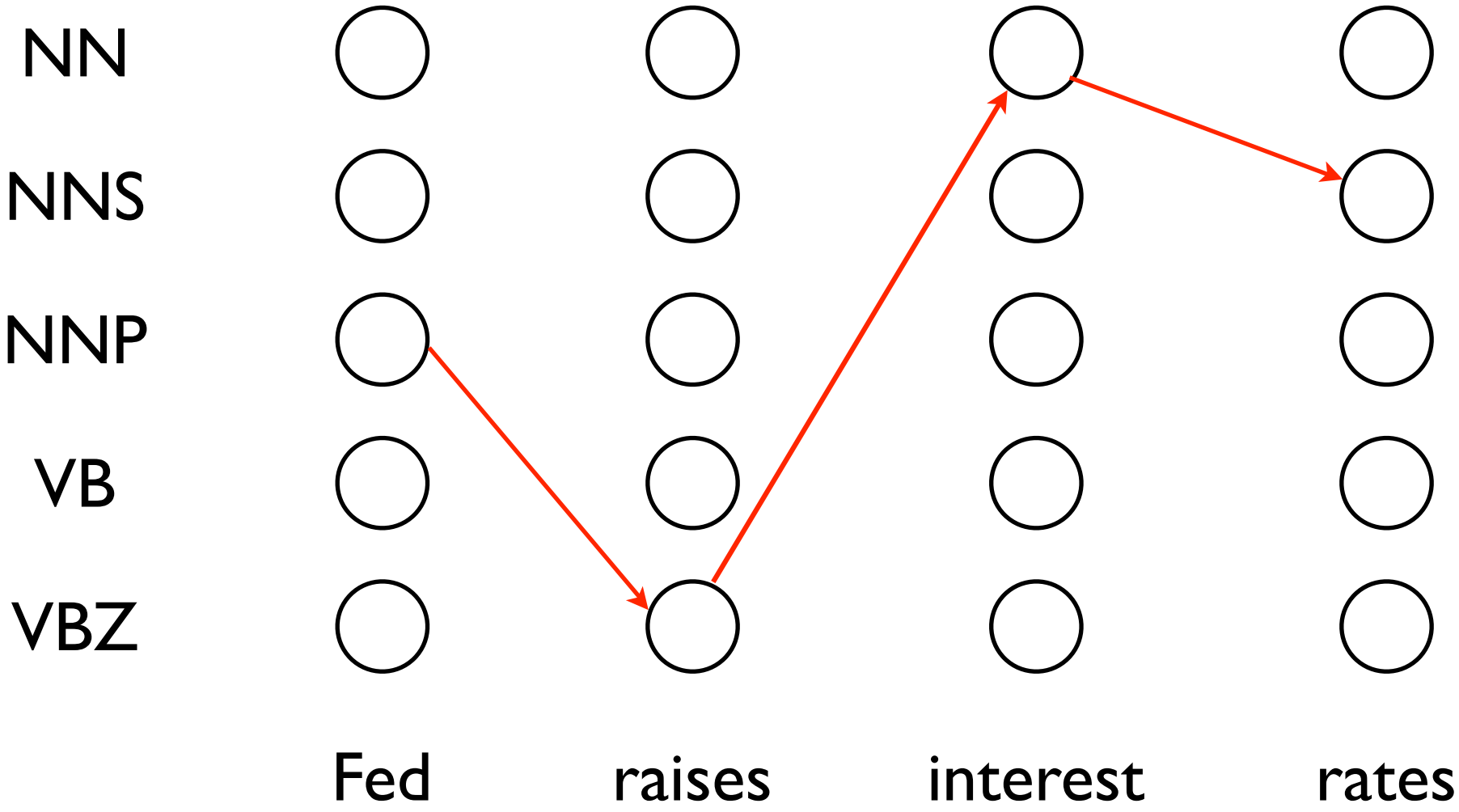


- Each state is associated with a probability distribution over words (the output)

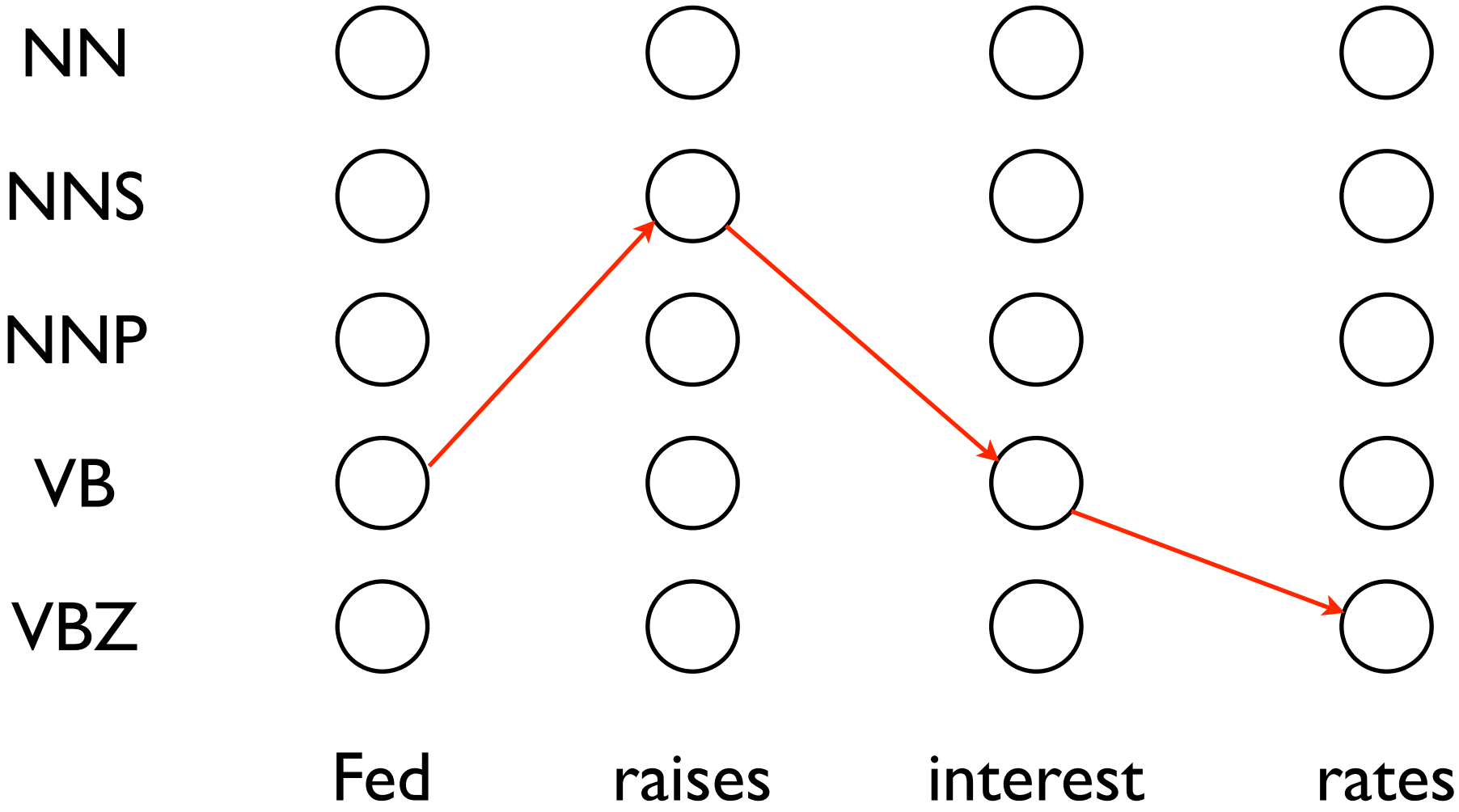
NER as Sequence Tagging

O	B-E	O	O	O	B-L	I-L
The	Phoenicians	came	from	the	Red	Sea

Sequence Tagging

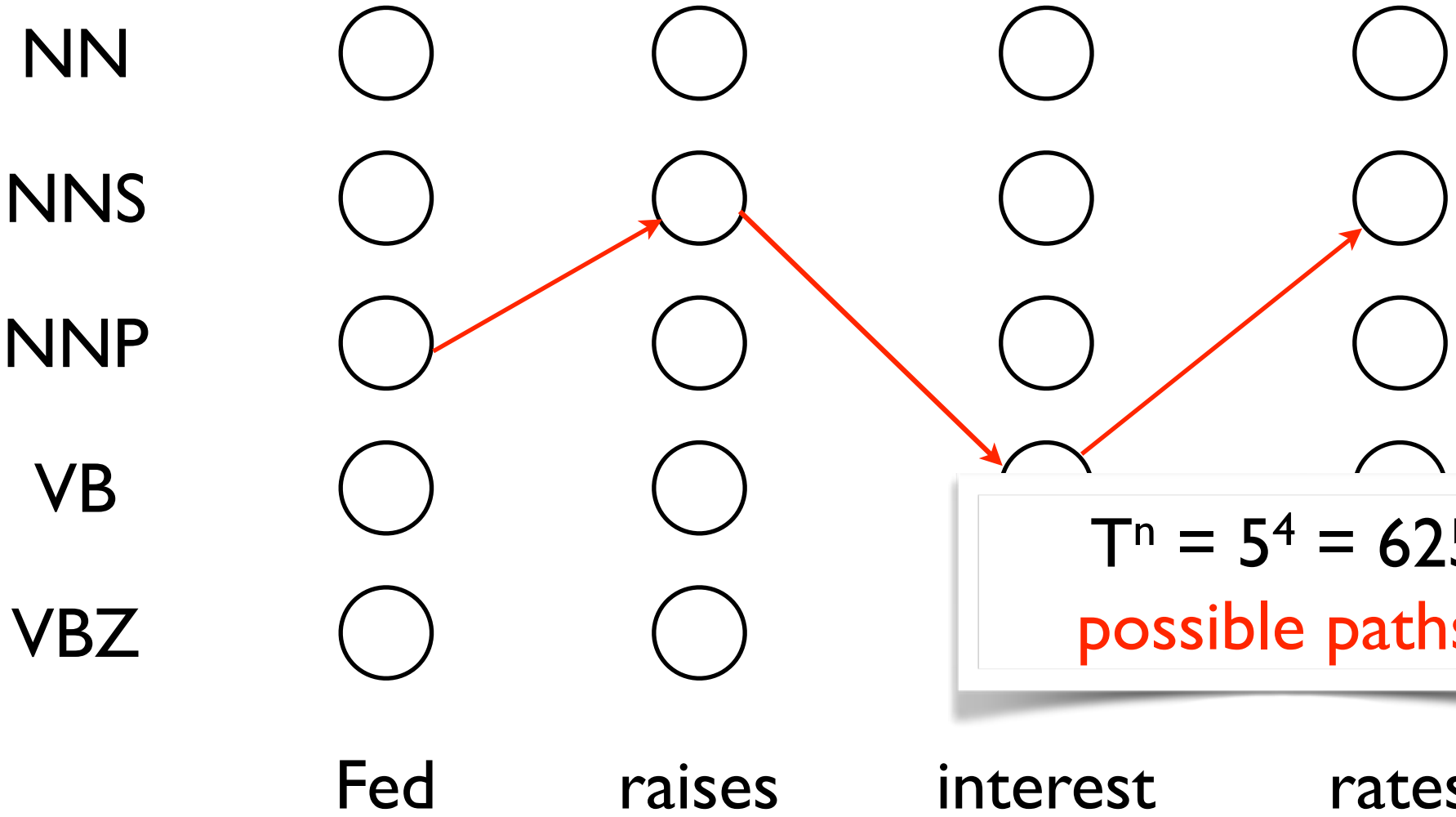


Sequence Tagging



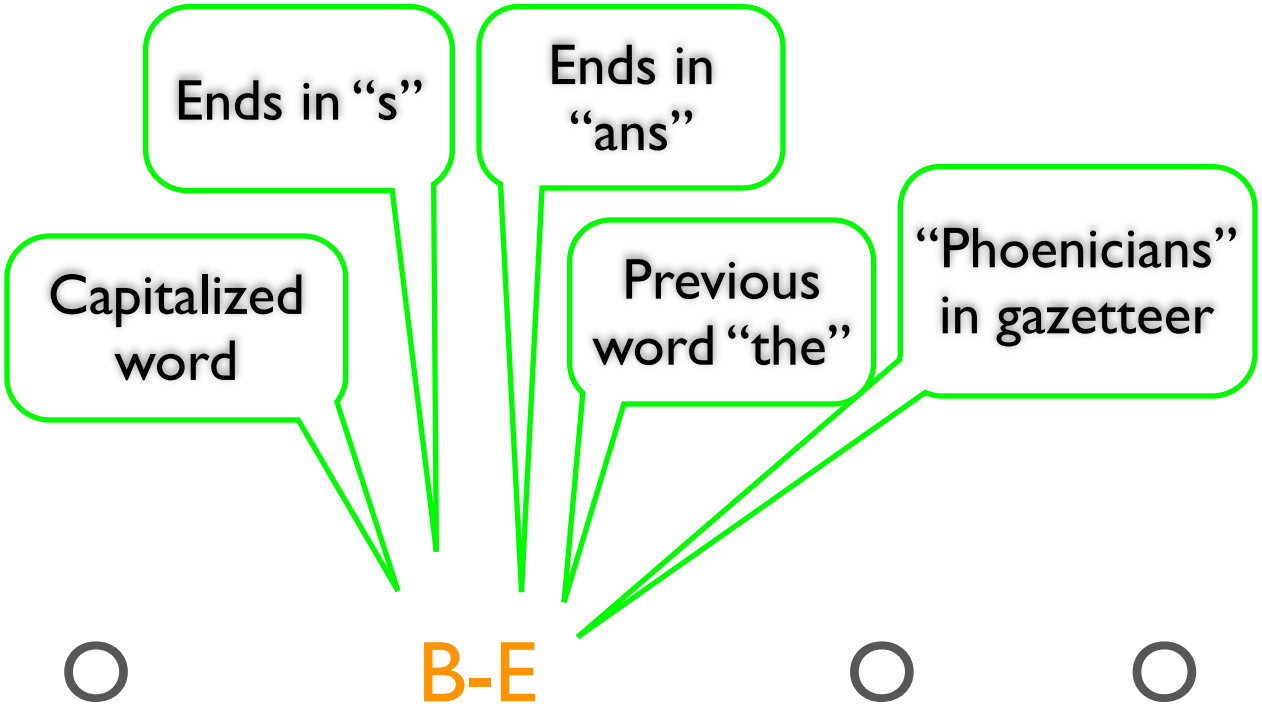
Sequence Tagging

Efficient (linear time) Shortest path = Viterbi algorithm

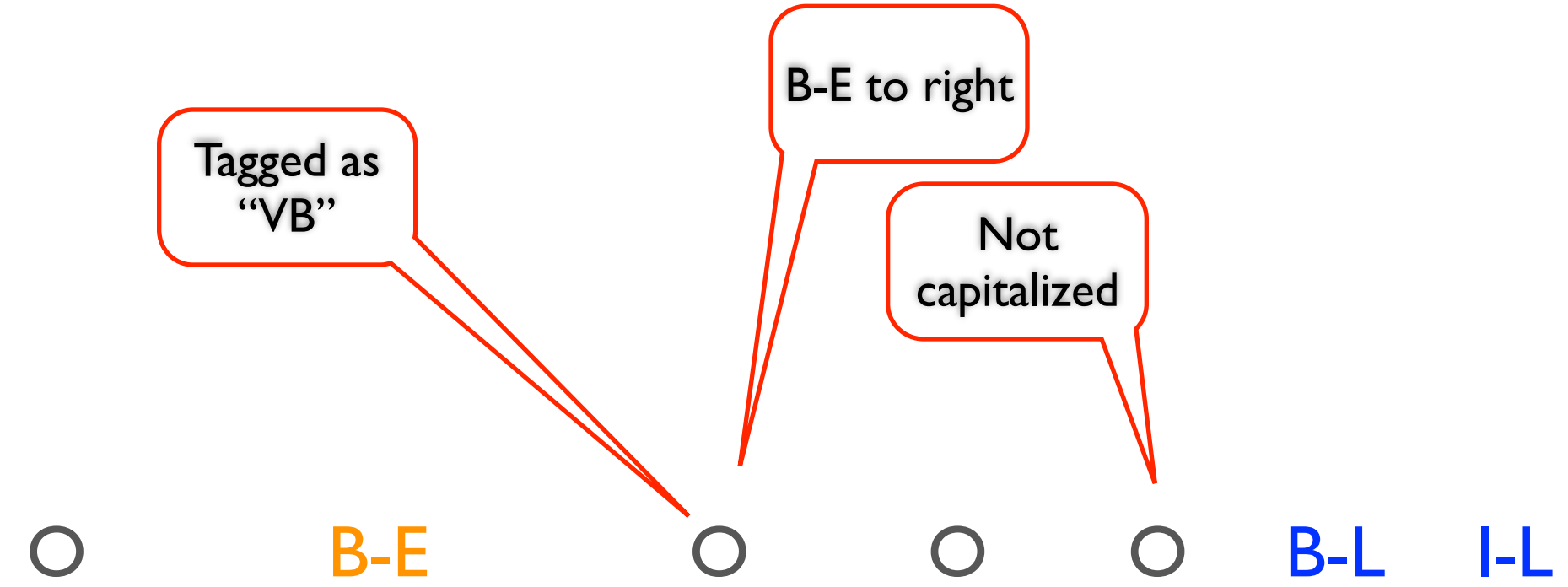


Can we specify that "Fed" always has the same tag in this document?

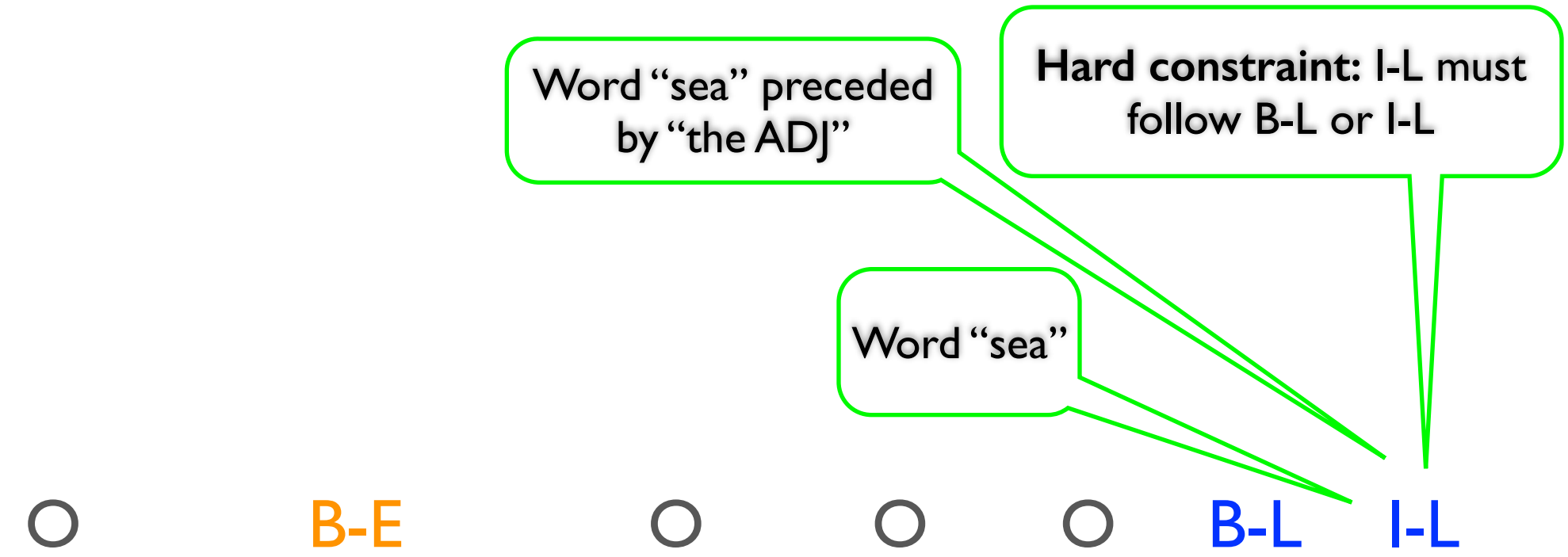
NER as Sequence Tagging



NER as Sequence Tagging

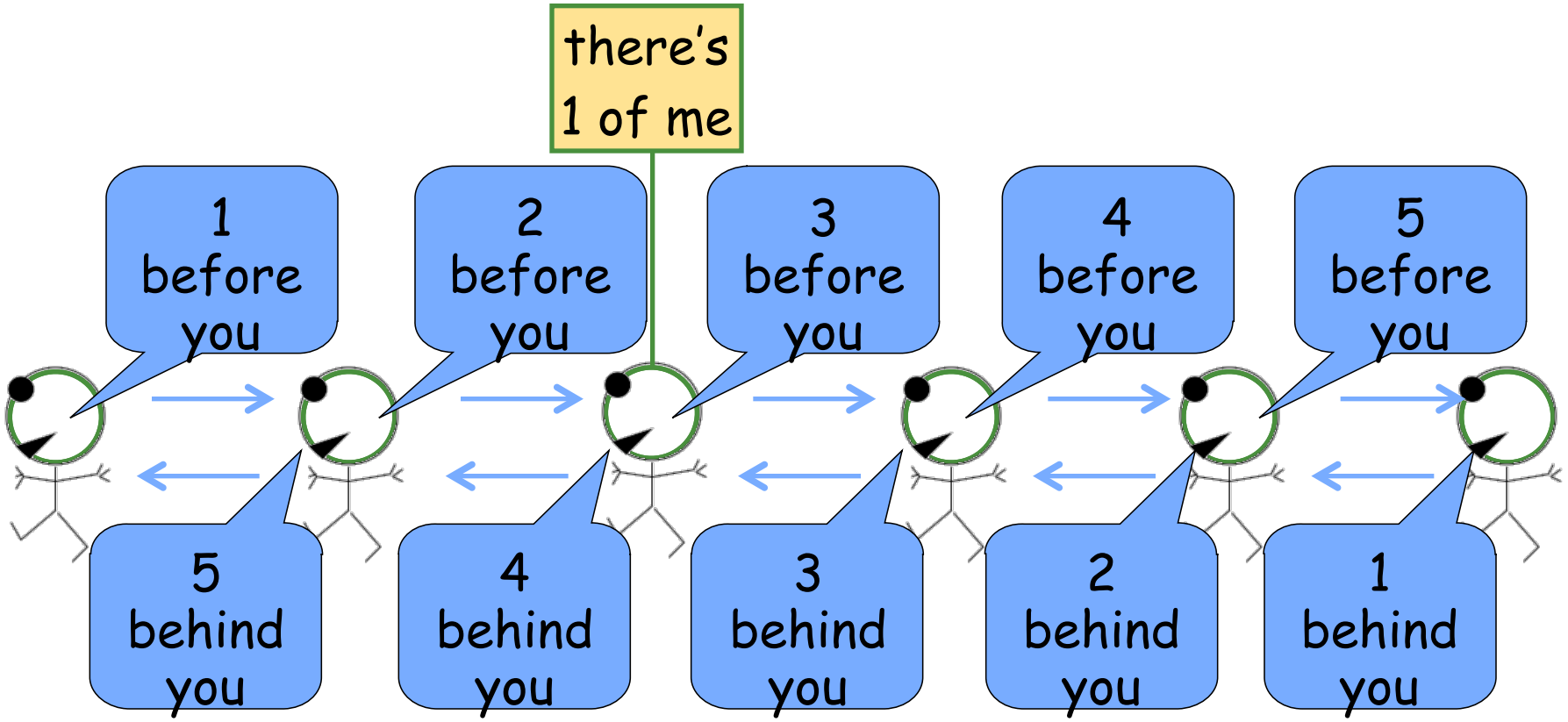


NER as Sequence Tagging



Great Ideas in ML: Message Passing

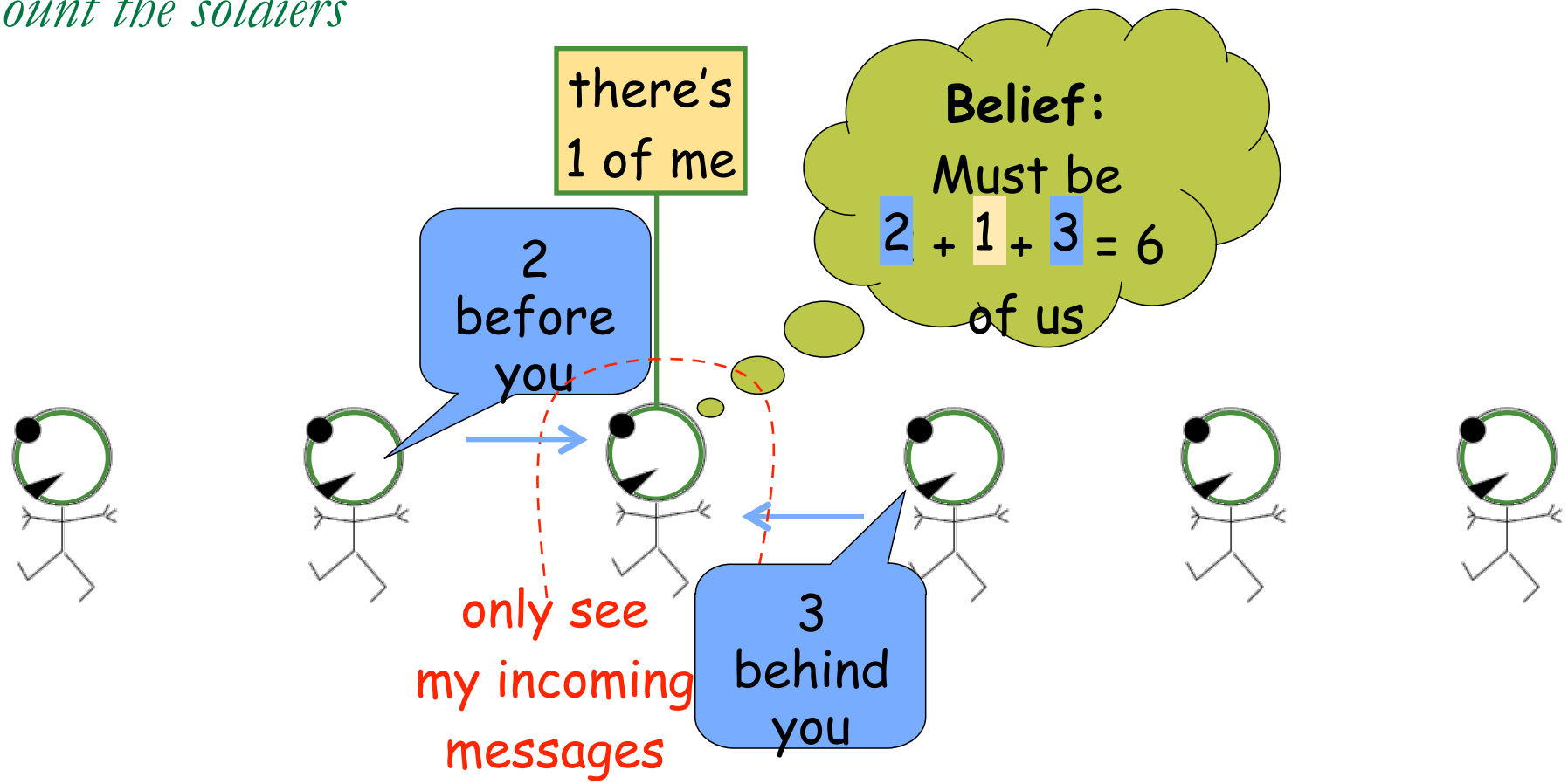
Count the soldiers



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

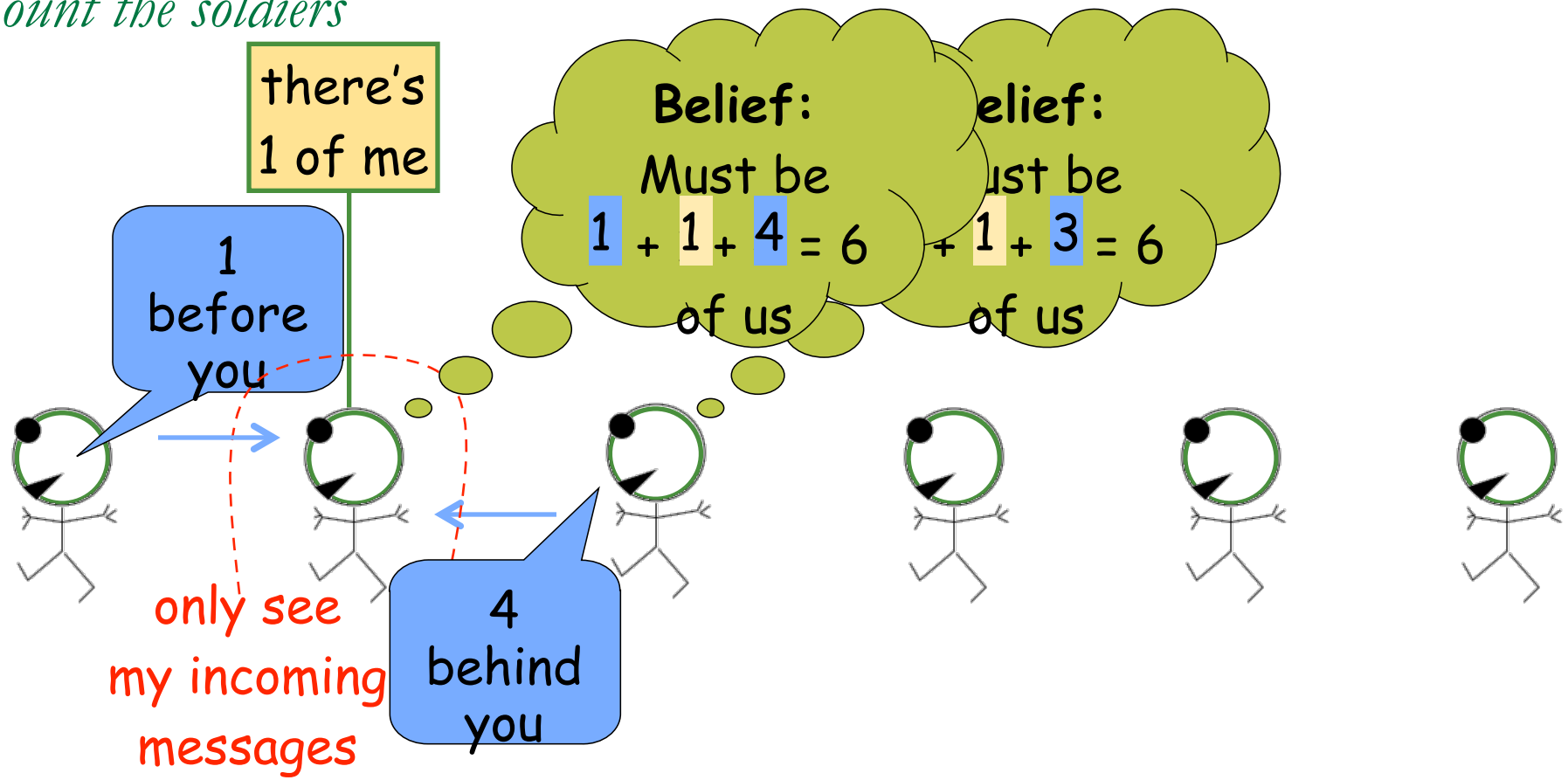
Count the soldiers



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

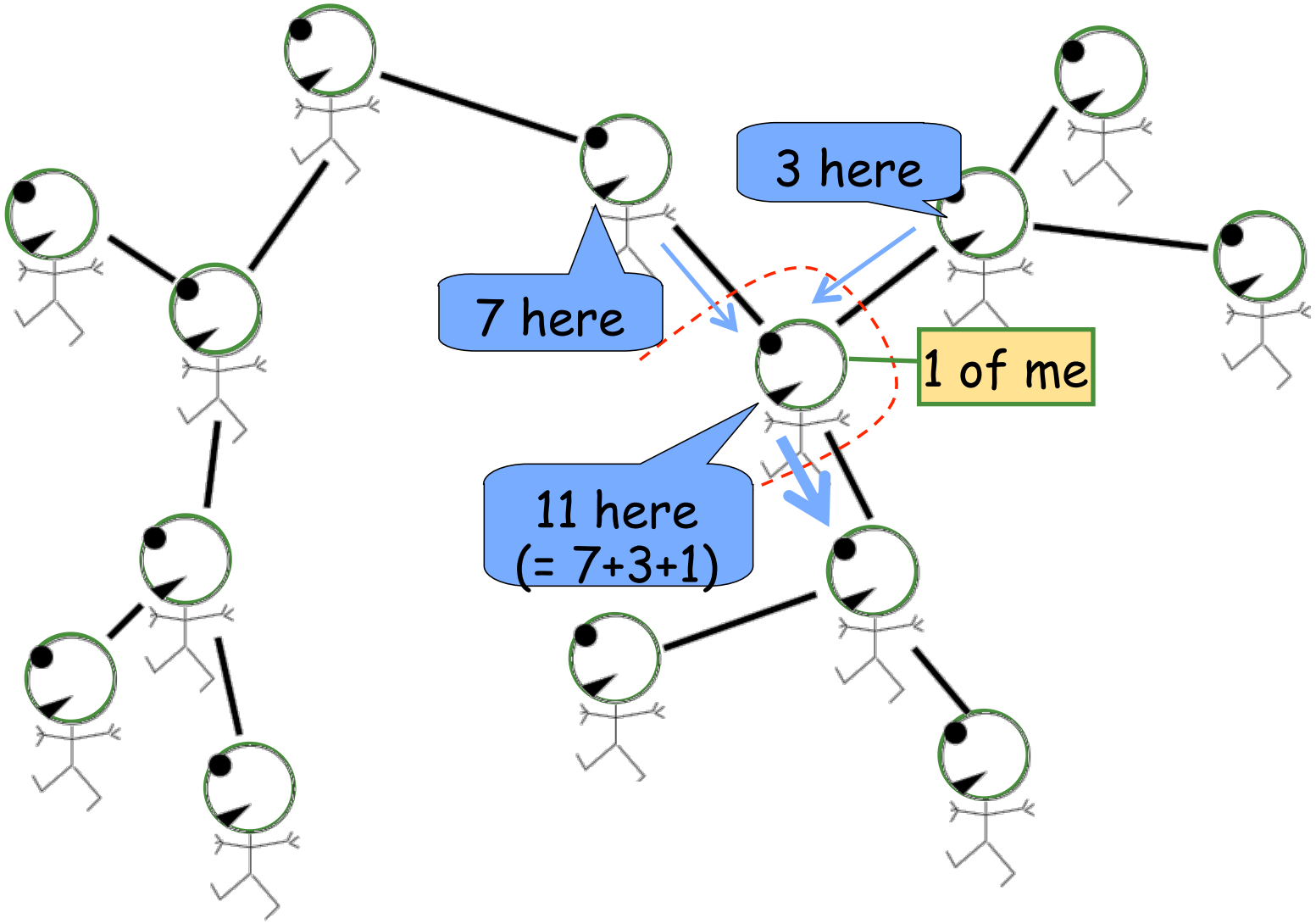
Count the soldiers



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

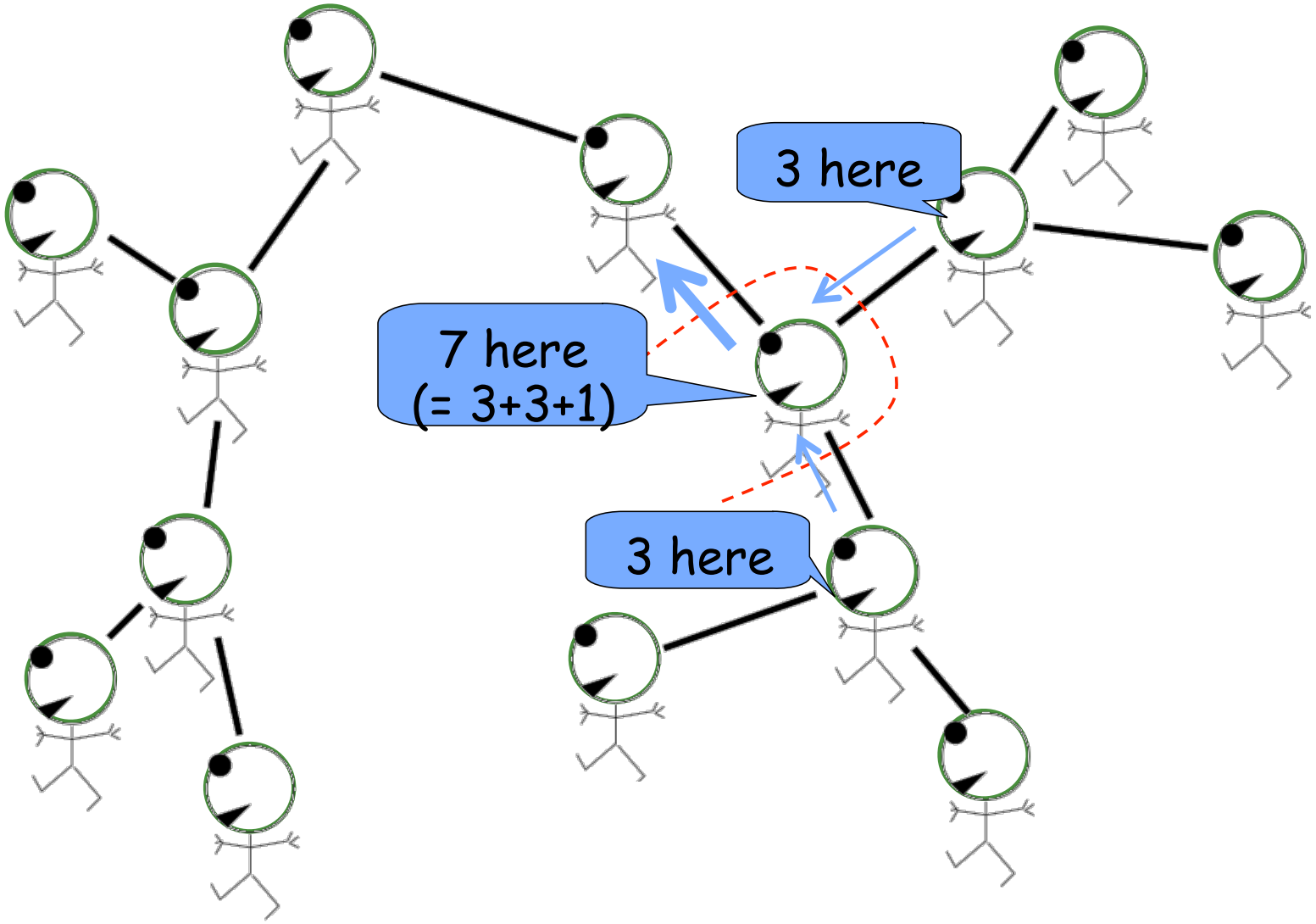
Each soldier receives reports from all branches of tree



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

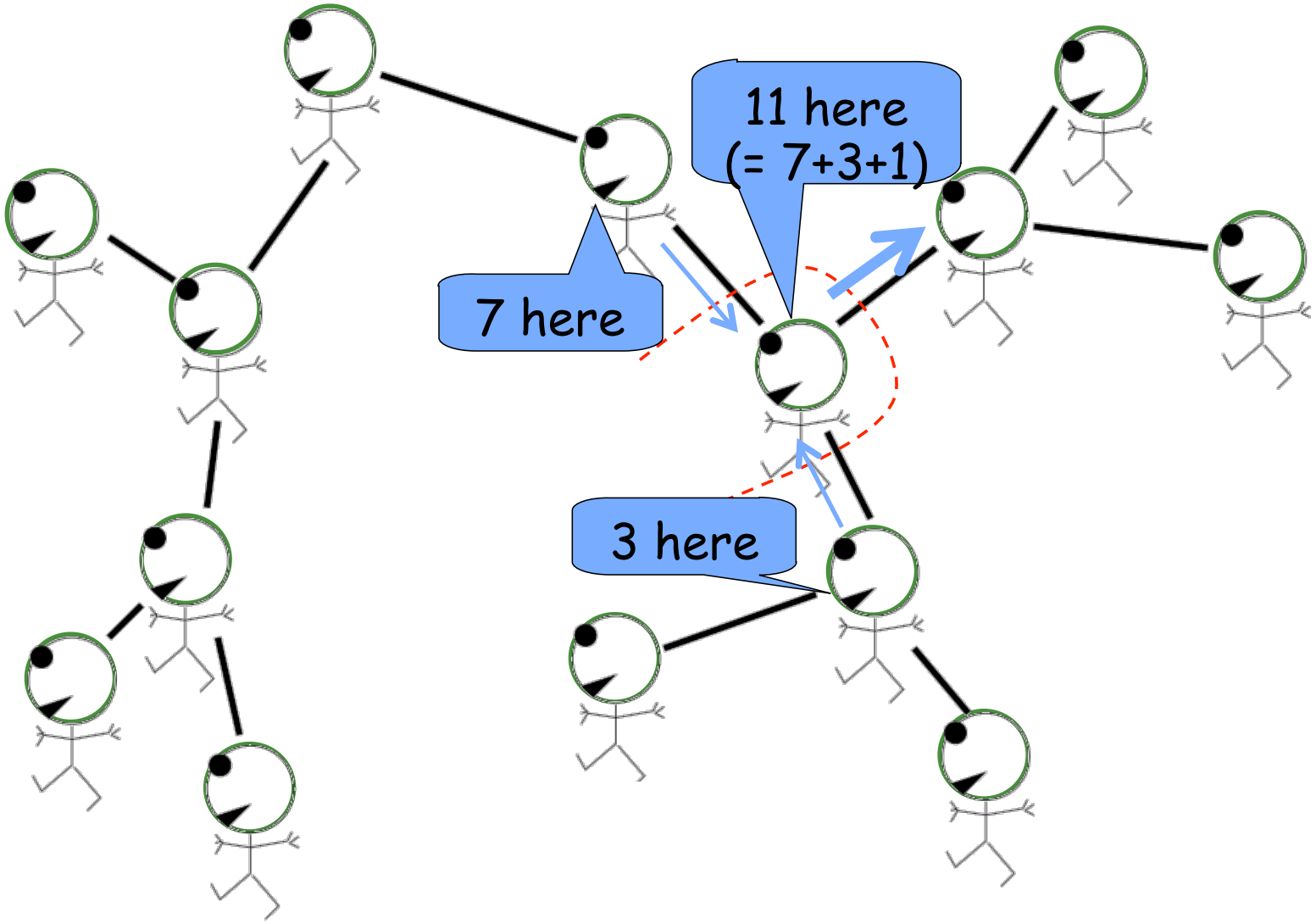
Each soldier receives reports from all branches of tree



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

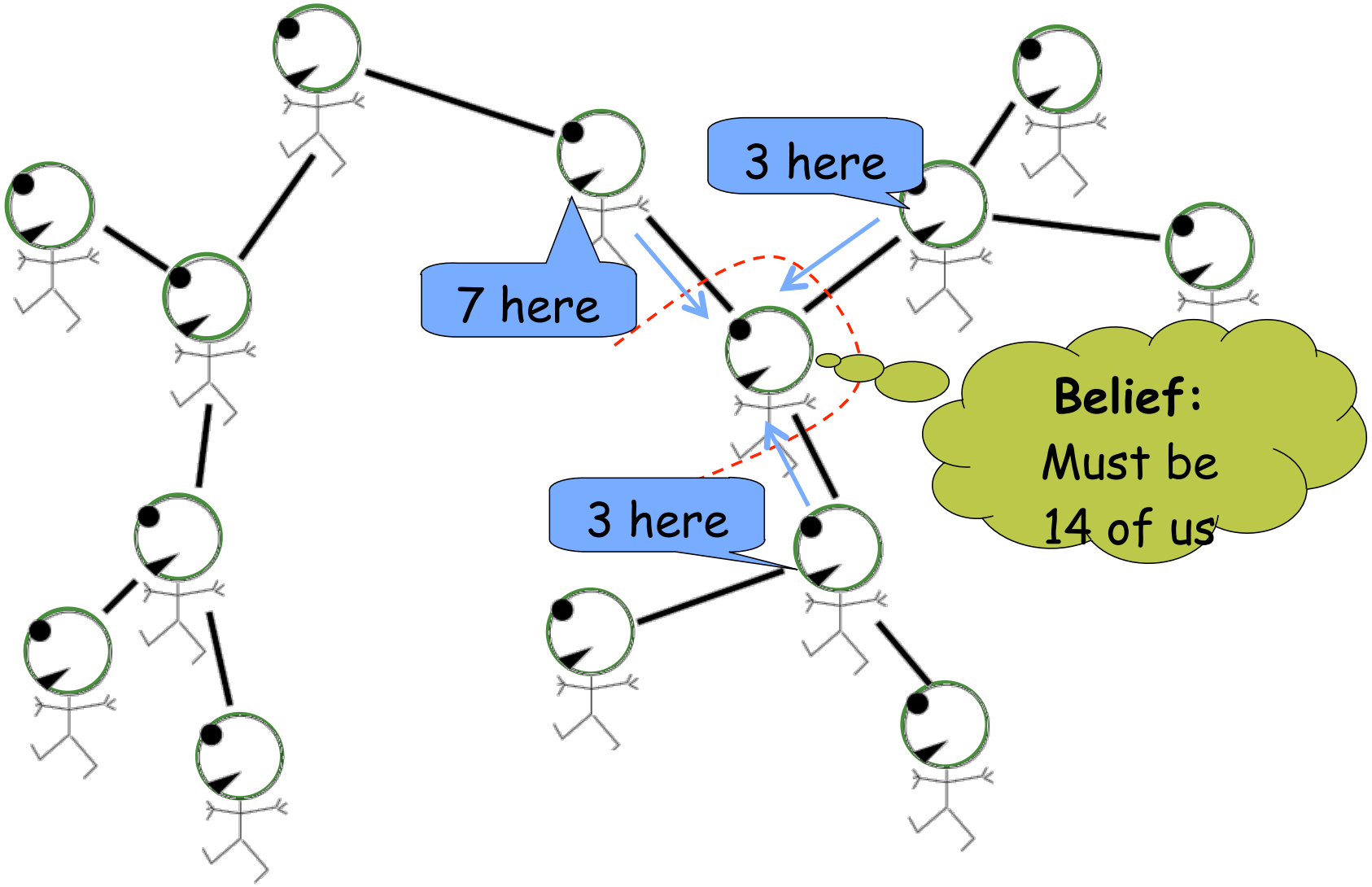
Each soldier receives reports from all branches of tree



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

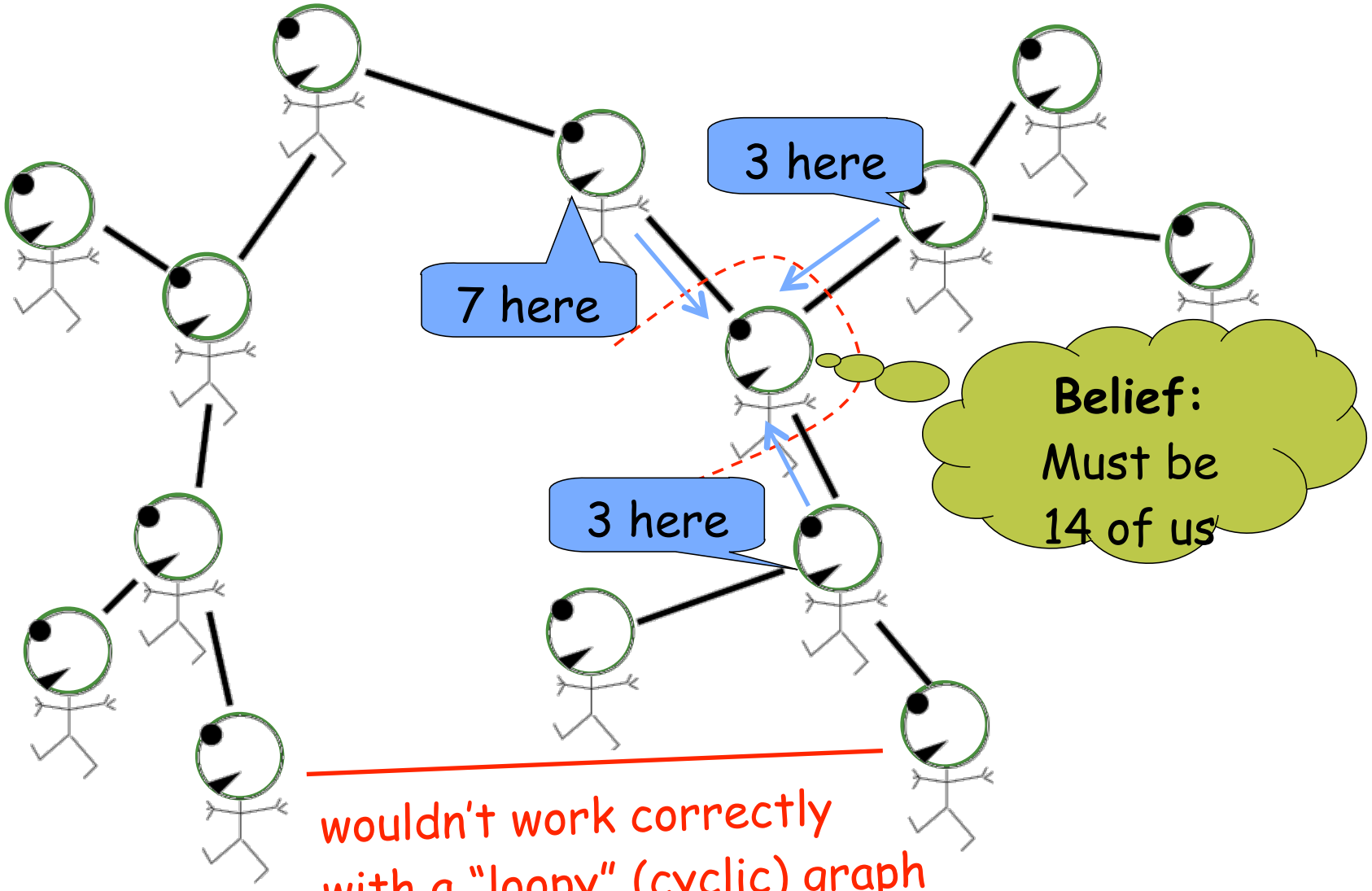
Each soldier receives reports from all branches of tree



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

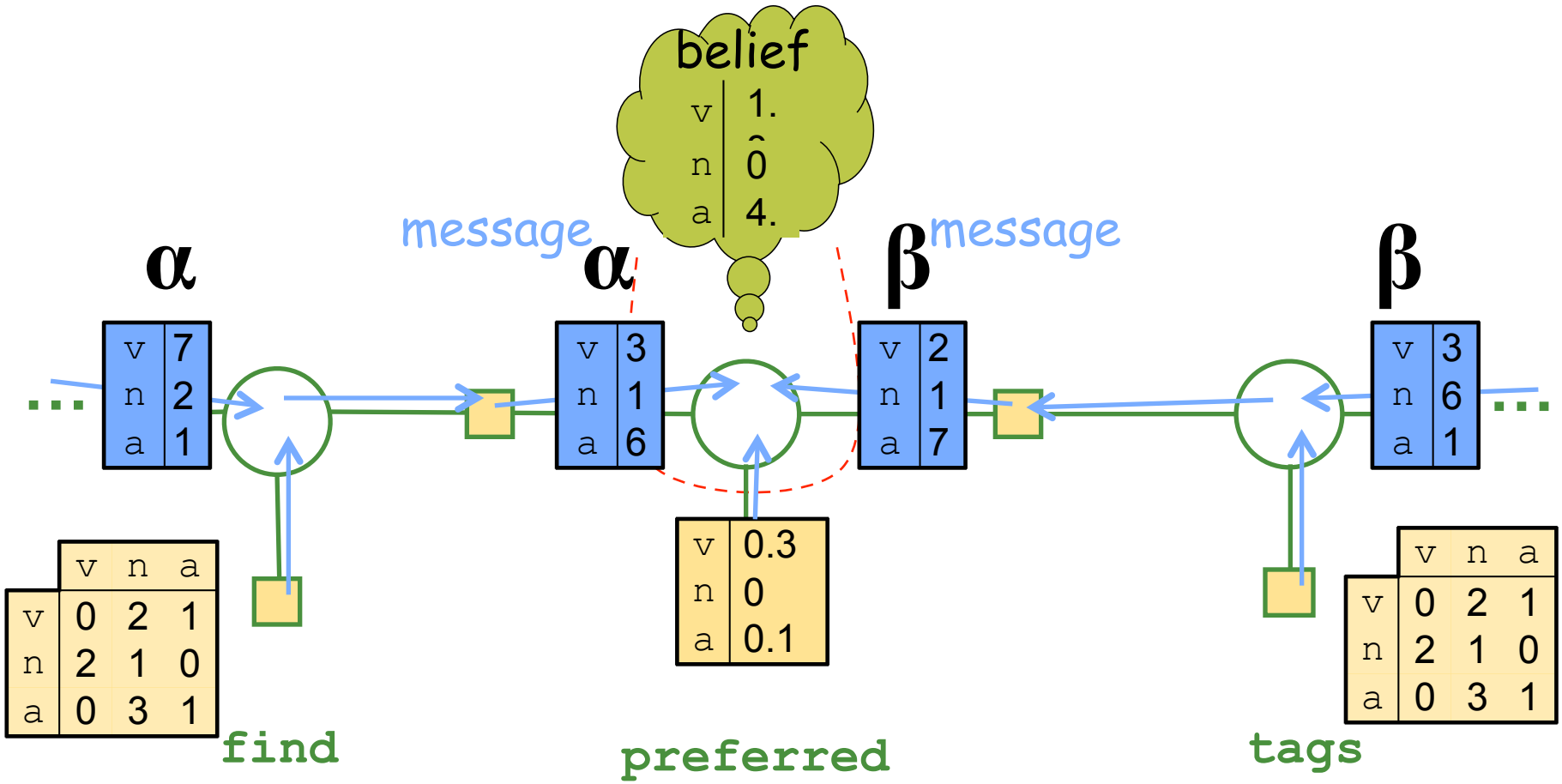
Each soldier receives reports from all branches of tree



adapted from MacKay (2003) textbook

Great ideas in ML: Forward-Backward

- In the CRF, message passing = forward-backward



Named Entity Recognition

- Accurate recognition requires about 1M words of training data (1,500 news stories)
  - –may be more expensive than developing rules for some applications
- Both rule-based and statistical can achieve about 90% effectiveness for categories such as names, locations, organizations

Internationalization

- 2/3 of the Web is in English
- About 50% of Web users do not use English as their primary language
- Many (maybe most) search applications have to deal with multiple languages
  - monolingual search: search in one language, but with many possible languages
  - cross-language search: search in multiple languages at the same time

Internationalization

- Many aspects of search engines are language-neutral
- Major differences:
  - Text encoding (converting to Unicode)
  - Tokenizing (many languages have no word separators)
  - Stemming
- Cultural differences may also impact interface design and features provided

Chinese “Tokenizing”

1. Original text

旱灾在中国造成的影响

(the impact of droughts in China)

2. Word segmentation

旱灾　在　中国　造成　的　影响

drought　at　china　make　　impact

3. Bigrams

旱灾　灾在　在中　中国　国造

造成　成的　的影　影响