Semantics

Tom Lippincott, Computer Science
(many slides credit to Matt Post)

Center for Digital Humanities
Johns Hopkins University

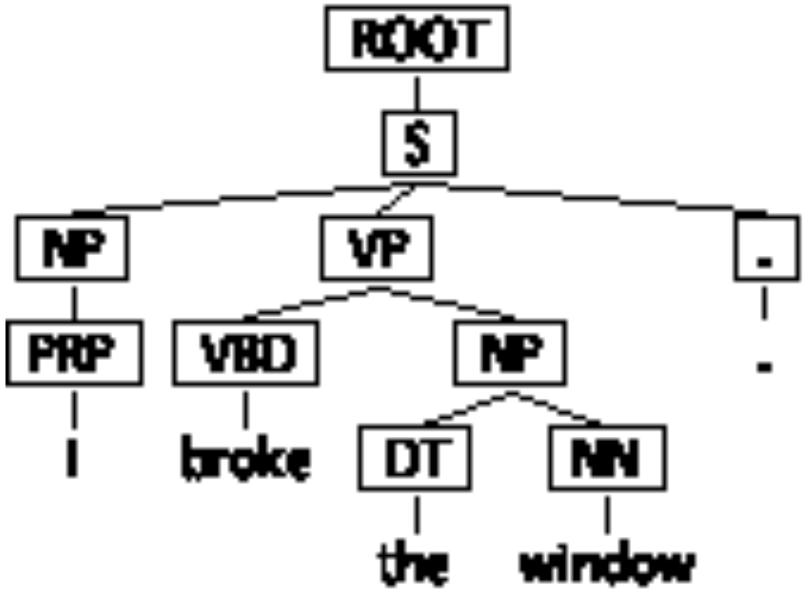
Knowledge of meaning

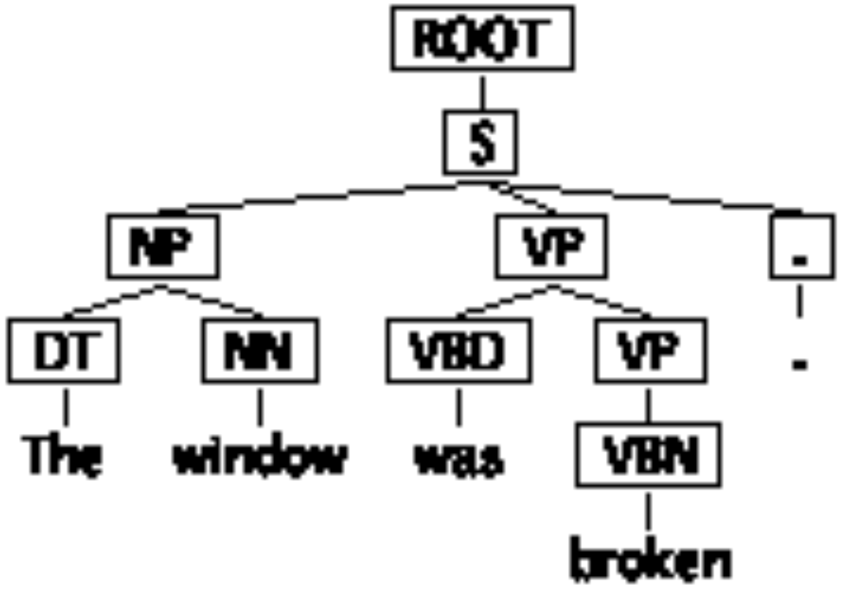
Aristotle taught Alexander
*Alexander taught Aristotle

He the almighty hurled
The almighty he hurled

Semantic Roles

- Syntax describes the grammatical relationships between words and phrases
  - But there are many different ways to express a particular meaning





- These variations miss an important generalization

- Structure is important, but one way it is important is as a “scaffolding for meaning”
- What we want to know is

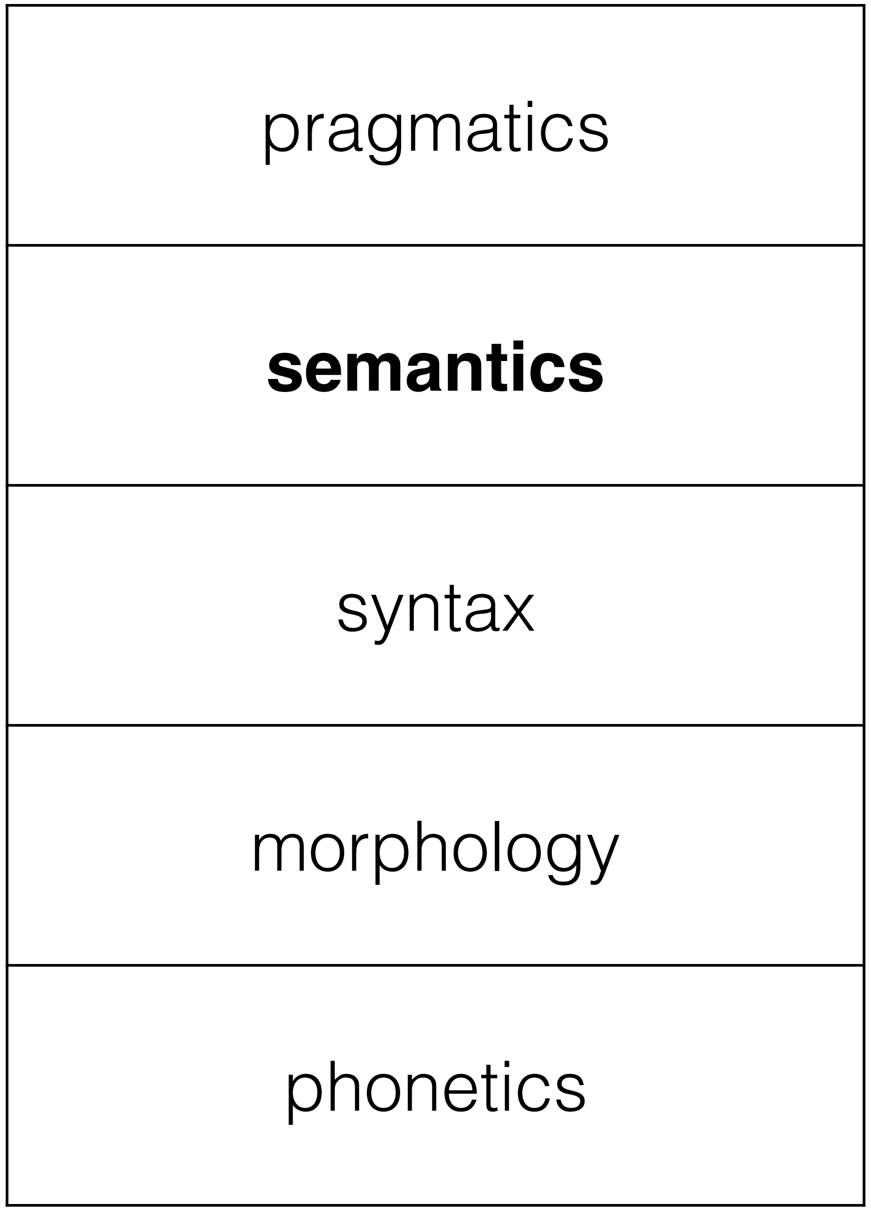
  who *did* ***what*** *to* ***whom***

  (and ***when****)*

  (and ***where****)*

  (and ***how****)*

A linguistic hierarchy



pragmatics

semantics

syntax

morphology

phonetics

how can we represent knowledge?

how do we do so in pursuit of solving some task?

Goal

- Given a sentence
  - answer the question “**who** did **what** to **whom** etc”
  - store answer in a machine-usable way

Goal

- Given a sentence
  - answer the question “**who** did **what** to **whom** etc”
  - store answer in a machine-usable way
- This requires
  - specifying some representation for meaning
  - specifying a representation for word relationships
  - mapping the words to these representations

Goal

- Given a sentence
  - answer the question “**who** did **what** to **whom** etc”
  - store answer in a machine-usable way
- This requires
  - specifying some representation for meaning
  - specifying a representation for word relationships
  - mapping the words to these representations
- **How do we represent meaning?**

Semantics

UNTIL RECENTLY → NOW

UNTIL RECENTLY

- Explicit representations
- Backed by human-constructed databases and ontologies
- Feature-based models

NOW

- End-to-end
- Backed by very large collections of unstructured human text
- Neural models

lexical semantics

“Word Senses and WordNet”
https://web.stanford.edu/~jurafsky/slp3/19.pdf

Words have many meanings

- Example
  - She pays 3% **interest** on the loan.
  - He showed a lot of **interest** in the painting.
  - Microsoft purchased a controlling **interest** in Google.
  - It is in the national **interest** to invade the Bahamas.
  - I only have your best **interest** in mind.
  - Playing chess is one of my **interests**.
  - Business **interests** lobbied for the legislation.

Words overlap in meaning

- What is the relationship among these words?
  - *{organization, team, group, association, conglomeration, institution, establishment, consortium, federation, agency, coalition, alliance, league, club, confederacy, syndicate, society, corporation}*
  - organisation?

Word senses can be organized

- **Synset**: a group of words with a shared meaning
  - This generalizes the notion of a word
  - Nowadays we'd think of this as a cluster in some high-dimensional space
- We can then define relationships between these sets of words

Relationships

- Many-many relationship between form and meaning

Relationships

- Many-many relationship between form and meaning
- Same forms

Relationships

- Many-many relationship between form and meaning
- Same forms
  - **polysemy** many related meanings

Relationships

- Many-many relationship between form and meaning
- Same forms
  - **polysemy** many related meanings
  - **homonymy** different meanings

Relationships

- Many-many relationship between form and meaning
- Same forms
  - **polysemy** many related meanings
  - **homonymy** different meanings
- Different forms

Relationships

- Many-many relationship between form and meaning
- Same forms
  - **polysemy** many related meanings
  - **homonymy** different meanings
- Different forms
  - **synonymy** same / similar meanings

Relationships

- Many-many relationship between form and meaning
- Same forms
  - **polysemy** many related meanings
  - **homonymy** different meanings
- Different forms
  - **synonymy** same / similar meanings
  - **antonymy** opposite or contrary meaning

More relationships

- Hypernym / hyponym
  - IS-A(animal, cat)
  - cat → feline → carnivore → placental mammal → mammal → vertebrate → …
- Meronymy (part / whole)
  - HAS-PART(cat, paw)
  - IS-PART-OF(paw, cat)
- Membership
  - IS-MEMBER-OF(professor, faculty)
  - HAS-MEMBER(faculty, professor)

WordNet

- English WordNet: https://wordnet.princeton.edu/

WordNet

- English WordNet: https://wordnet.princeton.edu/
  - nouns, verbs adjectives

WordNet

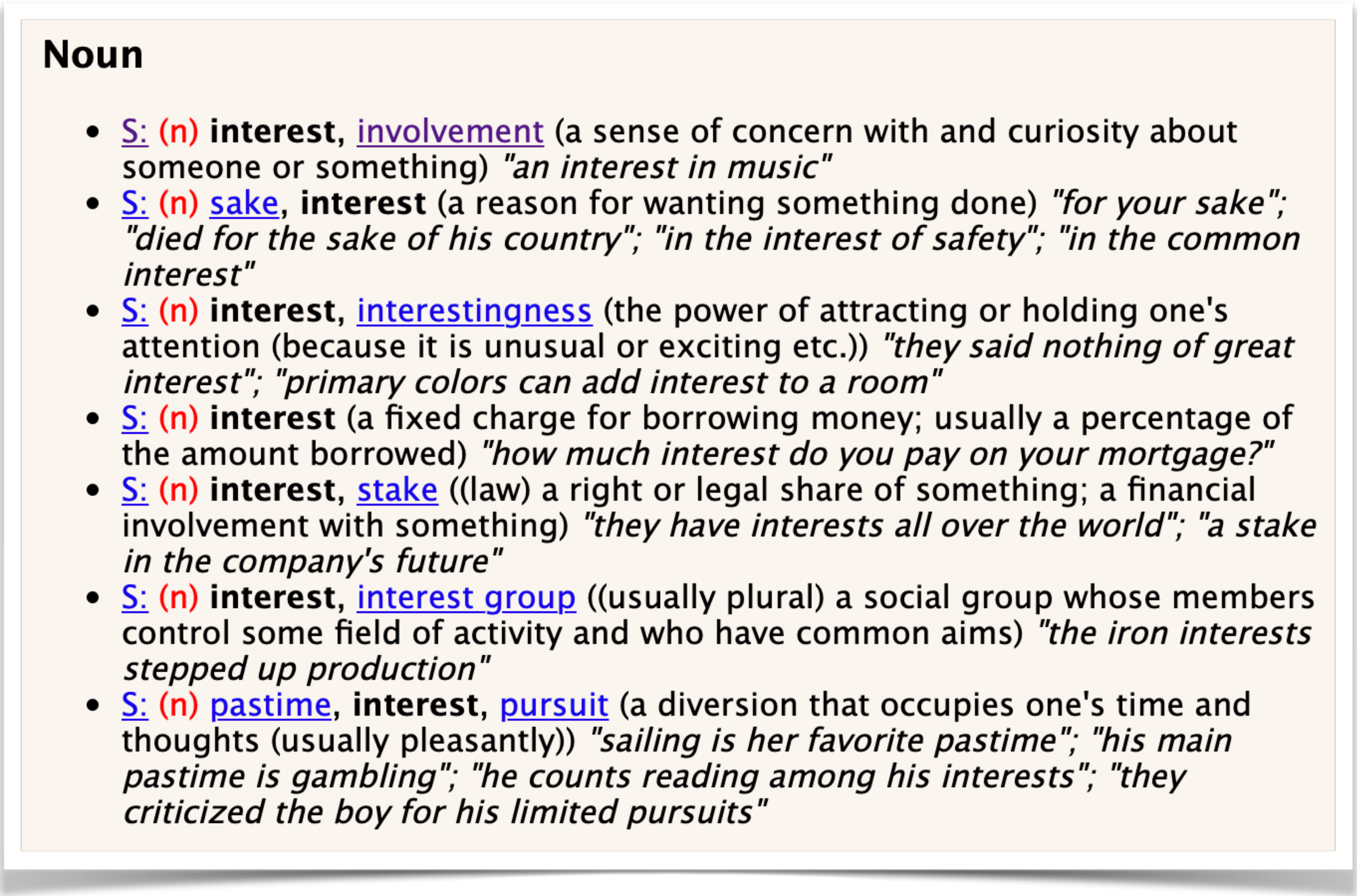
- English WordNet: https://wordnet.princeton.edu/
  - nouns, verbs adjectives
- Multilingual WordNet: http://compling.hss.ntu.edu.sg/omw/

WordNet

- English WordNet: https://wordnet.princeton.edu/
  - nouns, verbs adjectives
- Multilingual WordNet: http://compling.hss.ntu.edu.sg/omw/
- Examples: interest, tiger

Example (interest)

WordNet link



Noun

- S: (n) **interest**, involvement (a sense of concern with and curiosity about someone or something) *"an interest in music"*
- S: (n) sake, **interest** (a reason for wanting something done) *"for your sake"; "died for the sake of his country"; "in the interest of safety"; "in the common interest"*
- S: (n) **interest**, interestingness (the power of attracting or holding one's attention (because it is unusual or exciting etc.)) *"they said nothing of great interest"; "primary colors can add interest to a room"*
- S: (n) **interest** (a fixed charge for borrowing money; usually a percentage of the amount borrowed) *"how much interest do you pay on your mortgage?"*
- S: (n) **interest**, stake ((law) a right or legal share of something; a financial involvement with something) *"they have interests all over the world"; "a stake in the company's future"*
- S: (n) **interest**, interest group ((usually plural) a social group whose members control some field of activity and who have common aims) *"the iron interests stepped up production"*
- S: (n) pastime, **interest**, pursuit (a diversion that occupies one's time and thoughts (usually pleasantly)) *"sailing is her favorite pastime"; "his main pastime is gambling"; "he counts reading among his interests"; "they criticized the boy for his limited pursuits"*

Example (synset)

(a person who is gullible and easy to take advantage of)

S: (n) chump, fool, gull, mark, patsy, fall guy, sucker, soft touch, mug (a person who is gullible and easy to take advantage of)

Jurafsky & Martin, 3rd Ed., Ch 19. p. 6

Word Sense Disambiguation

- How can we map word (tokens) to the correct sense?

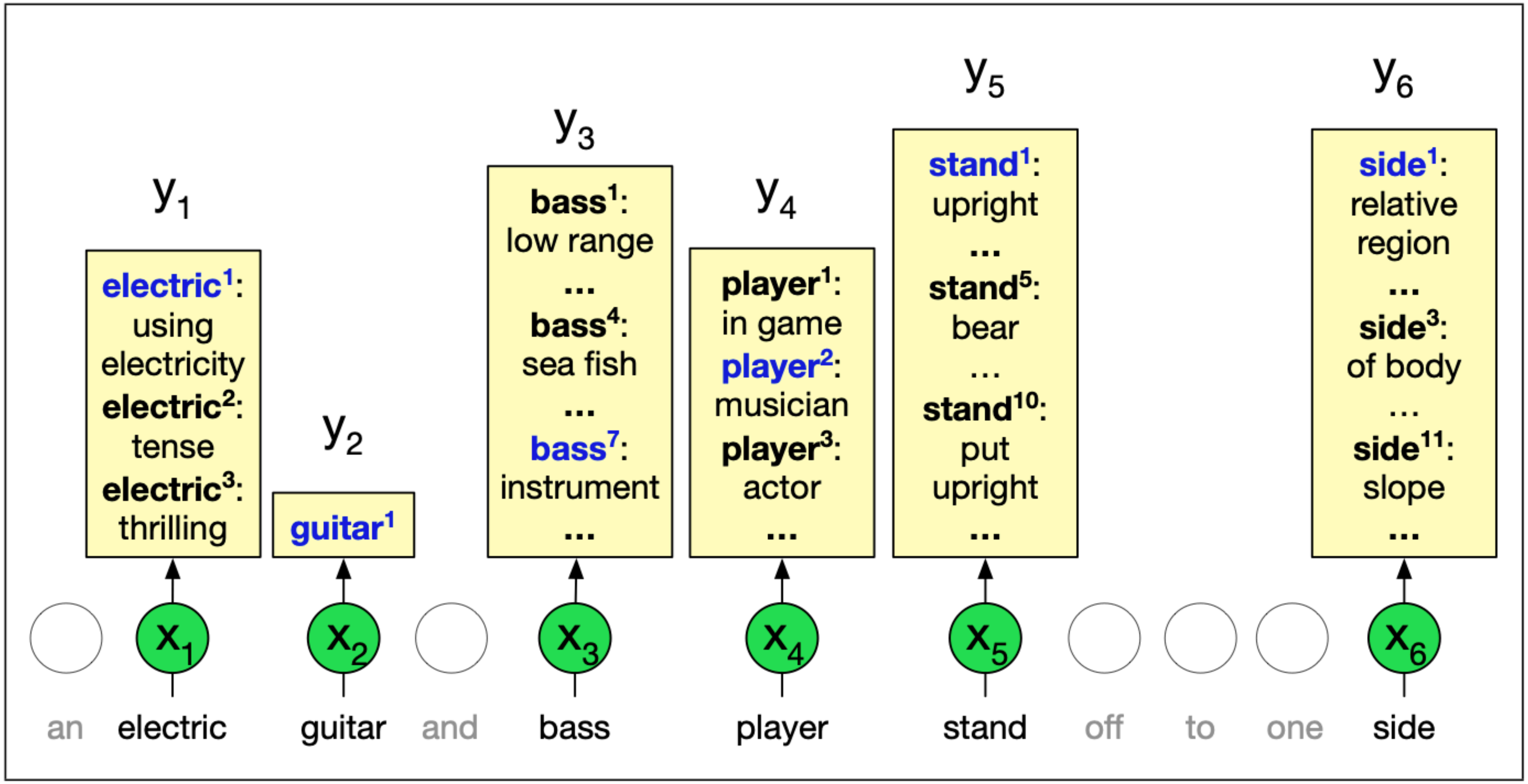


Figure 19.8 The all-words WSD task, mapping from input words (x) to WordNet senses (y). Only nouns, verbs, adjectives, and adverbs are mapped, and note that some words (like *guitar* in the example) only have one sense in WordNet. Figure inspired by Chaplot and Salakhutdinov (2018).

Supervised WSD

- Supervised approach
  - Take a corpus tagged with senses
  - Train a model on these tags
  - Apply to new data at test time
  -

Feature-based models

- Define features that are predictive of senses
  - window of words around the word
  - POS tags of window words
  - parse tree features
  - …you get the picture
- Learn a model using standard ML techniques, typically
  - P(sense | word, features)
  - e.g., maxent, naive Bayes, CRF

Contextual Embeddings

- The modern approach
- Compute contextual embeddings using (say) BERT or ELMo over a labeled dataset
  - produce a cluster by averaging the embeddings over the whole (labeled) training data
  - this produces a cluster for every sense of a word
  - at test time, again compute the contextual embedding, then assign by nearest-neighbors

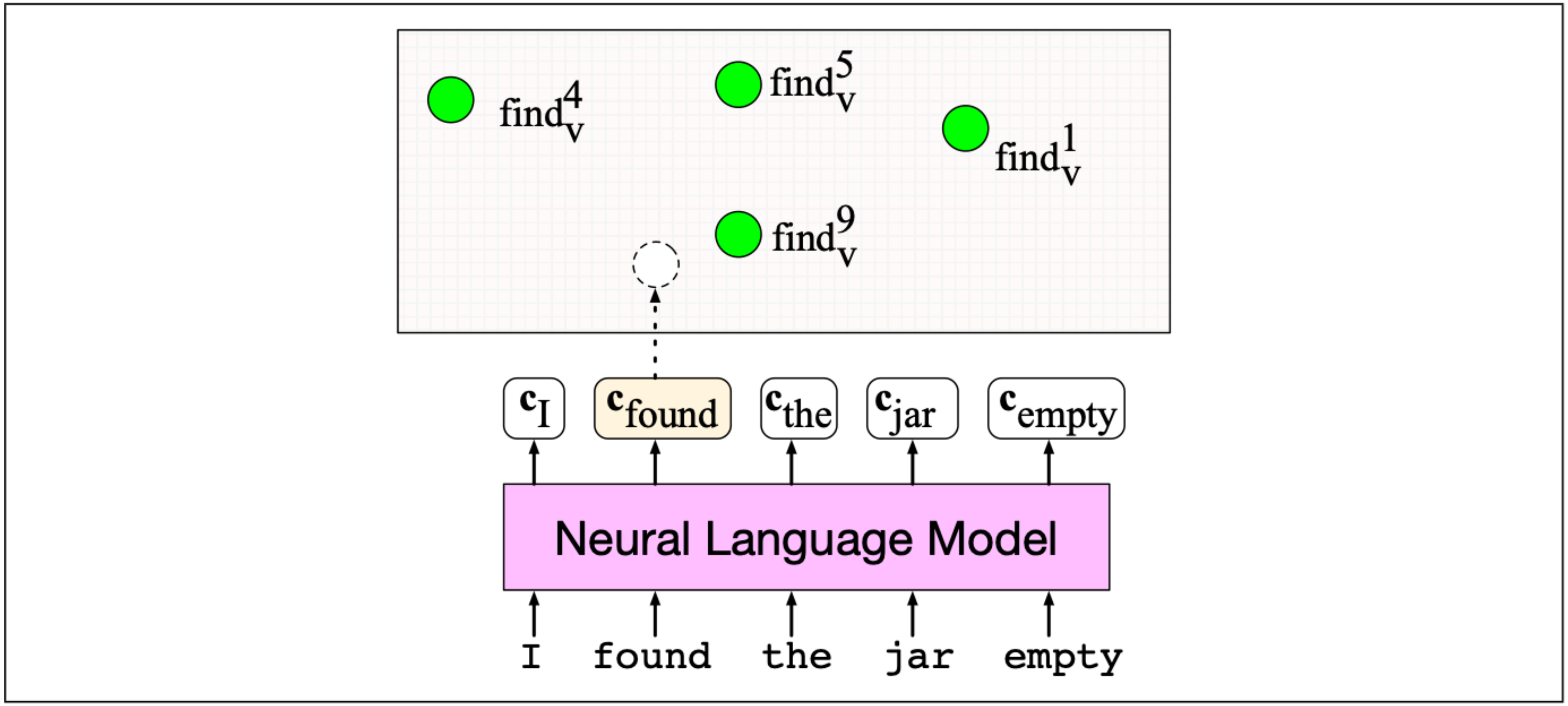


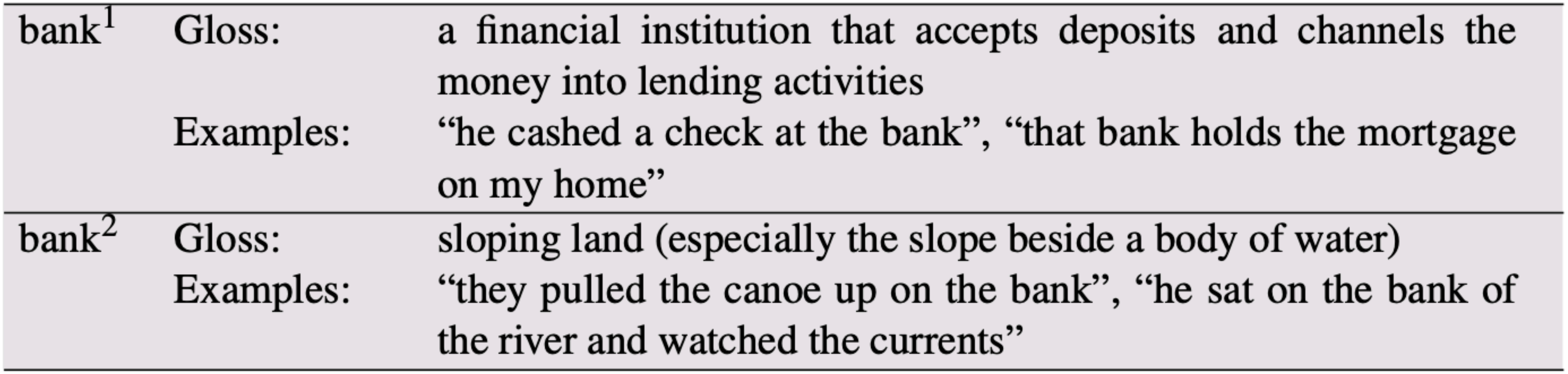
Figure 19.9 The nearest-neighbor algorithm for WSD. In green are the contextual embeddings precomputed for each sense of each word; here we just show a few of the senses for *find*. A contextual embedding is computed for the target word *found*, and the and then the nearest neighbor sense (in this case $\mathbf{find}_n^9$) would be chosen. Figure inspired by Loureiro and Jorge (2019).

Unsupervised WSD

- Consider:
  - we have Wordnet
    - which has groups of word forms, along with a gloss or definition
      - organized hierarchically
- What if you don’t have labeled data to choose from? How might you assign the correct word sense?

Lesk Algorithm

- (19.19) “The bank can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.”



bank^1	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	“he cashed a check at the bank”, “that bank holds the mortgage on my home”
bank^2	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	“they pulled the canoe up on the bank”, “he sat on the bank of the river and watched the currents”

  - which is the correct assignment?

Other approaches

- WordNet is huge, complicated, expensive to build
- Clustering
  - Instead of mapping words to predefined senses, using clustering algorithms to induce unlabeled clusters
  - Compute cluster centroids
  - At test time, assign words to clusters based on the nearest centroid
  - This has obvious connections to **word embeddings**

Summary

- Some takeaways:
  - Words can be grouped according to their overlapping senses called **synsets**
  - These groups can then be organized into an ontology with relationships
  - WordNet is a large database of these synsets, primarily for English
- Further reading:
  - Jurafsky & Martin, 3rd Ed., Chapter 19 https://web.stanford.edu/~jurafsky/slp3/19.pdf

semantic role labeling

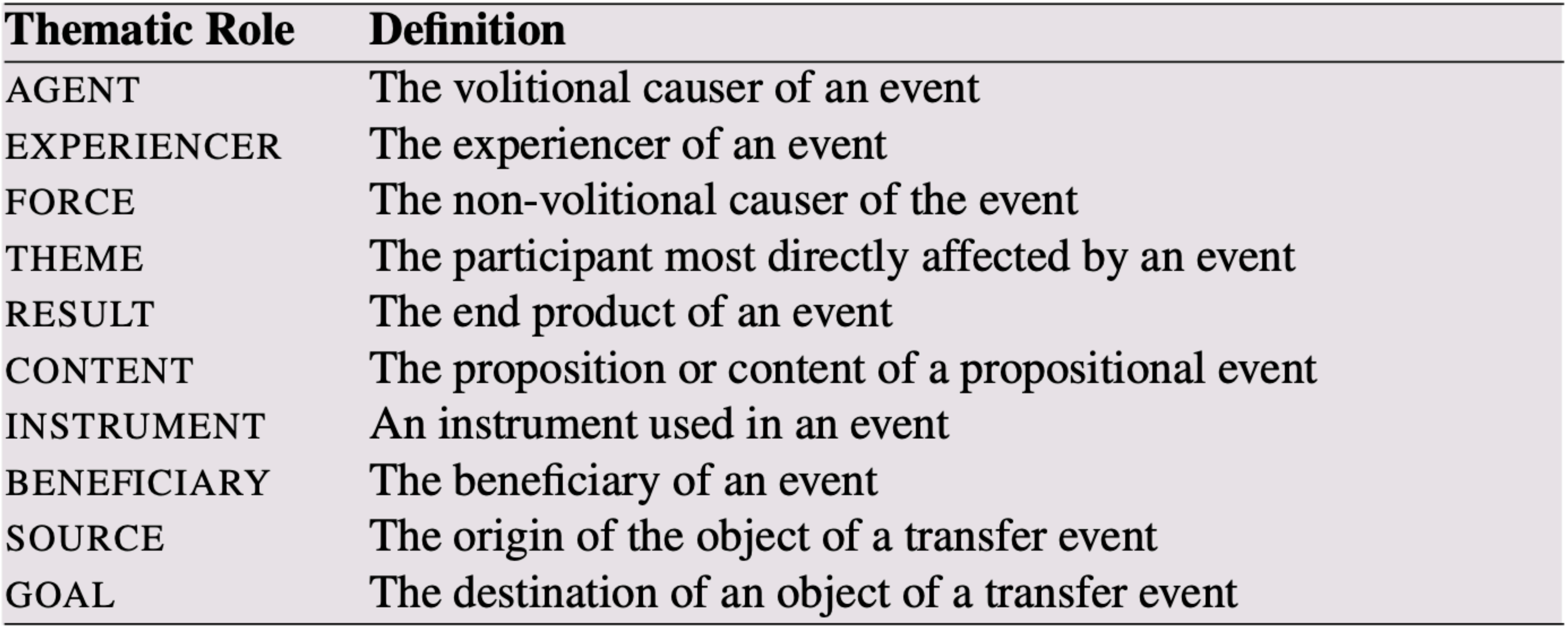
Semantic Role Labeling

- Assuming we can disambiguate a word, can we get back to the core question of identifying word relationships?
- Example sentence pair from before
  - *I broke the window*
  - *The window was broken by me*
- There is a generalization here involving the types of participants

Much of the structure here follows Chapter 20 of Jurafsky & Martin, 3rd Ed.

https://web.stanford.edu/~jurafsky/slp3/20.pdf

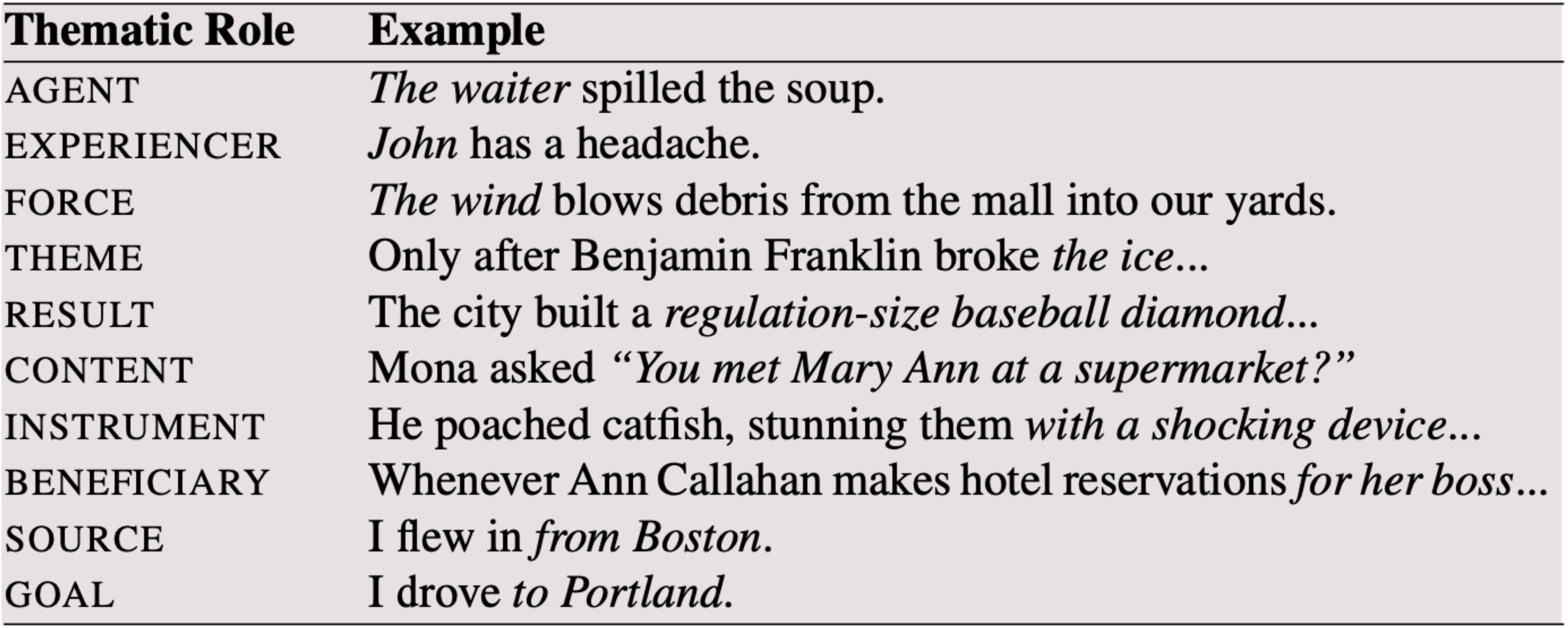
Thematic Roles



Thematic Role	Definition
AGENT	The volitional causer of an event
EXPERIENCER	The experiencer of an event
FORCE	The non-volitional causer of the event
THEME	The participant most directly affected by an event
RESULT	The end product of an event
CONTENT	The proposition or content of a propositional event
INSTRUMENT	An instrument used in an event
BENEFICIARY	The beneficiary of an event
SOURCE	The origin of the object of a transfer event
GOAL	The destination of an object of a transfer event

Figure 20.1 Some commonly used thematic roles with their definitions.

Thematic Roles



Thematic Role	**Example**
AGENT	*The waiter* spilled the soup.
EXPERIENCER	*John* has a headache.
FORCE	*The wind* blows debris from the mall into our yards.
THEME	Only after Benjamin Franklin broke *the ice*...
RESULT	The city built a *regulation-size baseball diamond*...
CONTENT	Mona asked *"You met Mary Ann at a supermarket?"*
INSTRUMENT	He poached catfish, stunning them *with a shocking device*...
BENEFICIARY	Whenever Ann Callahan makes hotel reservations *for her boss*...
SOURCE	I flew in *from Boston*.
GOAL	I drove *to Portland*.

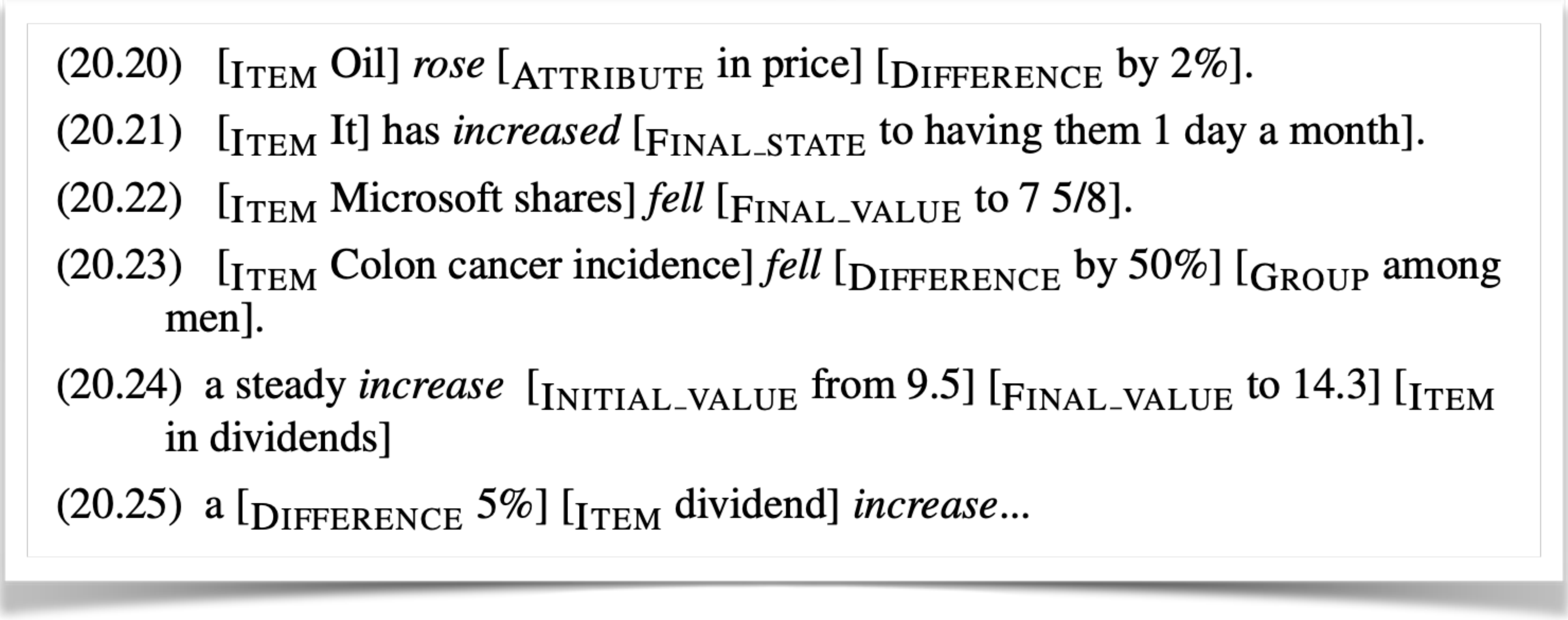
Figure 20.2 Some prototypical examples of various thematic roles.

FrameNet

- **frame**: the general background information relating to an event that is invoked and filled by the sentence
  - established idea in cognitive science and semantics
  - related to the idea of **scripts** (story patterns that underly an event or report)

Example

- Consider these sentences



(20.20) [$_{\text{ITEM}}$ Oil] *rose* [$_{\text{ATTRIBUTE}}$ in price] [$_{\text{DIFFERENCE}}$ by 2%].

(20.21) [$_{\text{ITEM}}$ It] has *increased* [$_{\text{FINAL\_STATE}}$ to having them 1 day a month].

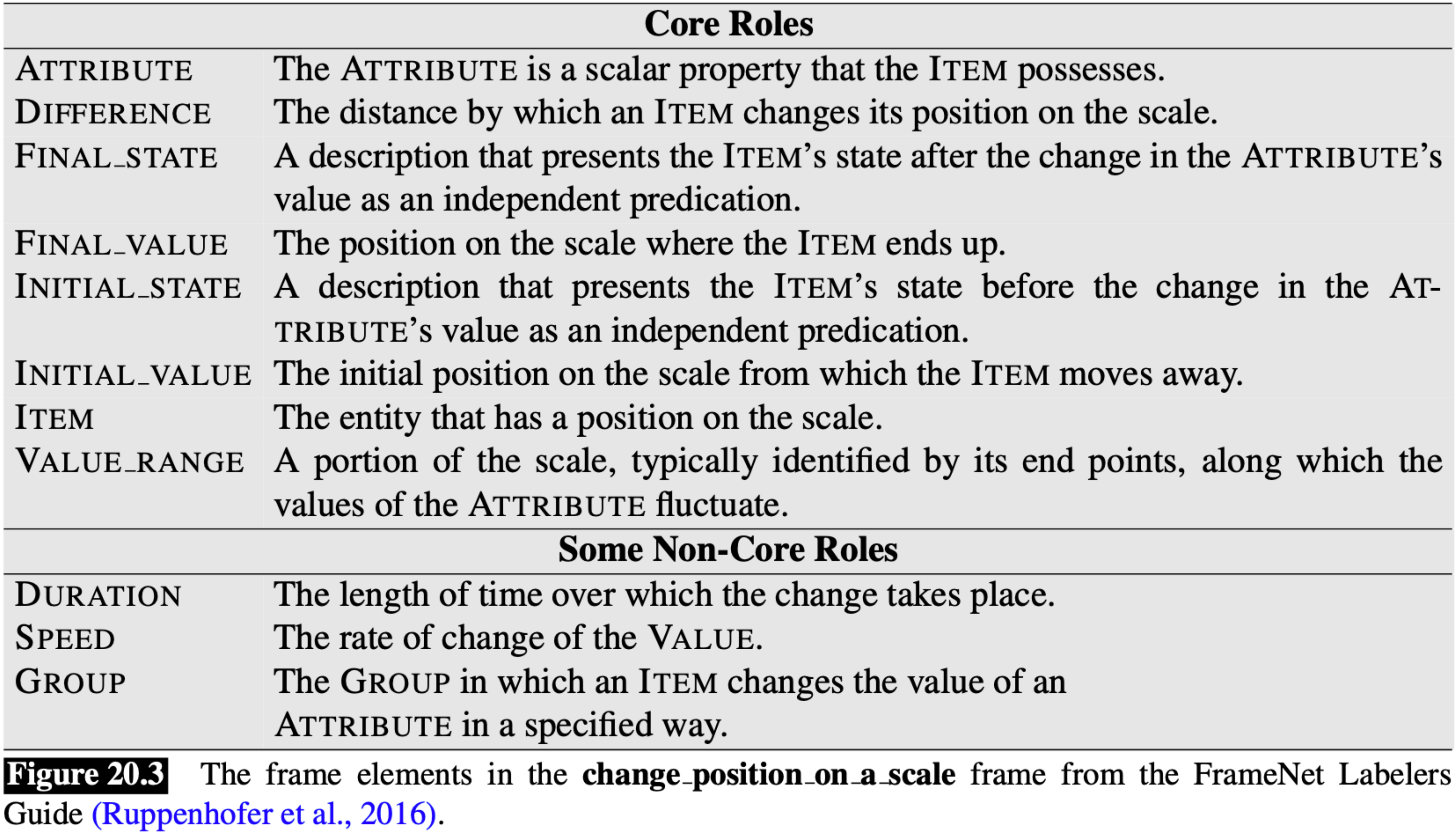
(20.22) [$_{\text{ITEM}}$ Microsoft shares] *fell* [$_{\text{FINAL\_VALUE}}$ to 7 5/8].

(20.23) [$_{\text{ITEM}}$ Colon cancer incidence] *fell* [$_{\text{DIFFERENCE}}$ by 50%] [$_{\text{GROUP}}$ among men].

(20.24) a steady *increase* [$_{\text{INITIAL\_VALUE}}$ from 9.5] [$_{\text{FINAL\_VALUE}}$ to 14.3] [$_{\text{ITEM}}$ in dividends]

(20.25) a [$_{\text{DIFFERENCE}}$ 5%] [$_{\text{ITEM}}$ dividend] *increase*...

these can be thought of as invoking the following frame



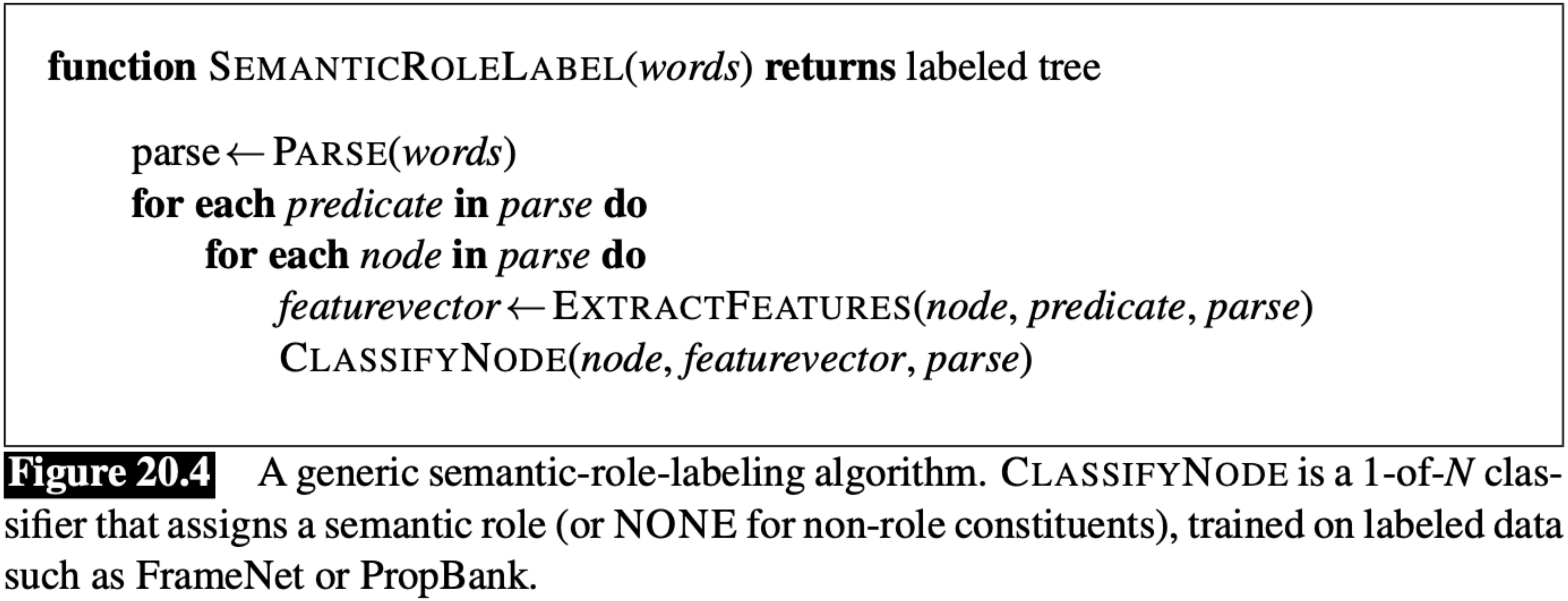
Core Roles	
ATTRIBUTE	The ATTRIBUTE is a scalar property that the ITEM possesses.
DIFFERENCE	The distance by which an ITEM changes its position on the scale.
FINAL_STATE	A description that presents the ITEM's state after the change in the ATTRIBUTE's value as an independent predication.
FINAL_VALUE	The position on the scale where the ITEM ends up.
INITIAL_STATE	A description that presents the ITEM's state before the change in the ATTRIBUTE's value as an independent predication.
INITIAL_VALUE	The initial position on the scale from which the ITEM moves away.
ITEM	The entity that has a position on the scale.
VALUE_RANGE	A portion of the scale, typically identified by its end points, along which the values of the ATTRIBUTE fluctuate.
Some Non-Core Roles	
DURATION	The length of time over which the change takes place.
SPEED	The rate of change of the VALUE.
GROUP	The GROUP in which an ITEM changes the value of an ATTRIBUTE in a specified way.

Figure 20.3 The frame elements in the **change_position_on_a_scale** frame from the FrameNet Labelers Guide (Ruppenhofer et al., 2016).

Semantic Role Labeling: the task

- Determine semantic roles of words in a sentence
  - Input: *You can't blame the program for being unable to identify it.*
  - Output: [You]COGNIZER can't [blame]TARGET [the program]EVALUEE [for being unable to identify it]REASON

The algorithm



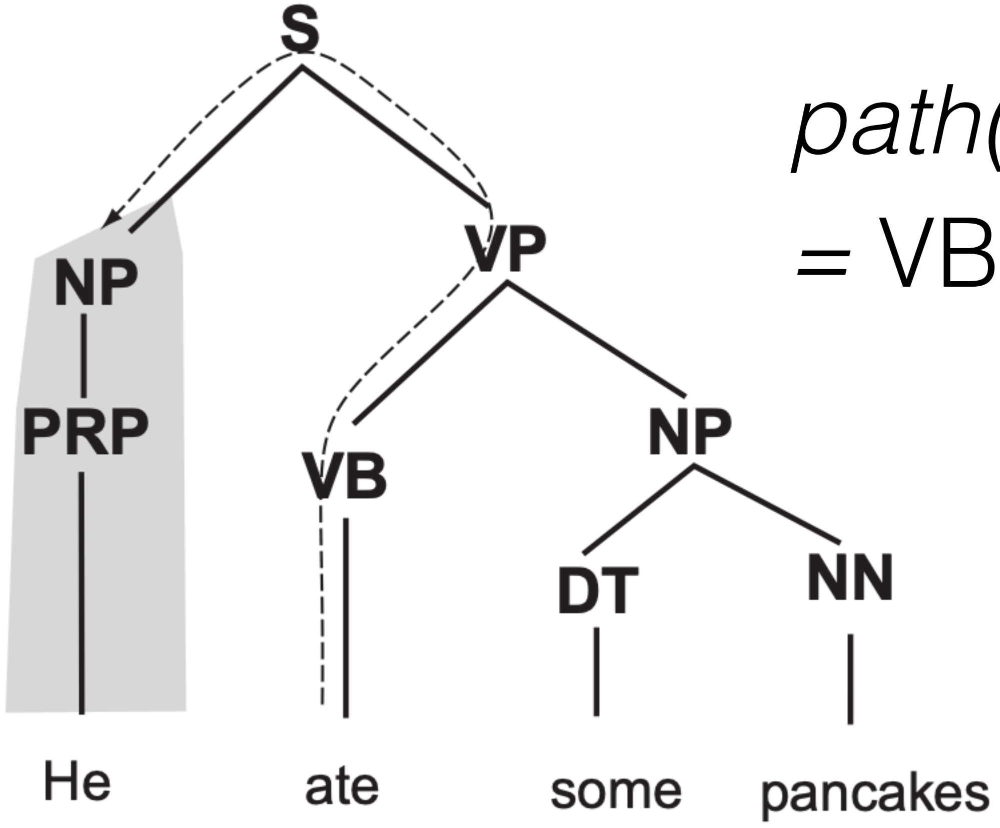
```
function SEMANTICROLELABEL(words) returns labeled tree

    parse ← PARSE(words)
    for each predicate in parse do
        for each node in parse do
            featurevector ← EXTRACTFEATURES(node, predicate, parse)
            CLASSIFYNODE(node, featurevector, parse)
```

Figure 20.4 A generic semantic-role-labeling algorithm. CLASSIFYNODE is a 1-of-*N* classifier that assigns a semantic role (or NONE for non-role constituents), trained on labeled data such as FrameNet or PropBank.

Features

- Nonterminal label (“NP”)
- Governing category (“S” or “VP” = subject or object)
- Parse tree path
- Position (before or after predicate)
- Head word
- Many, many more



path(ate→He)

= VB ↑ VP ↑ S ↓ NP

- Trained with discriminative ML algorithms (SVM, MaxEnt)

Bringing it together

- This can finally bring us to the point where we have tuples, say of the form (action, agent, patient, [theme])
  - e.g., (saw, man, bird, telescope)
- How can we use these?

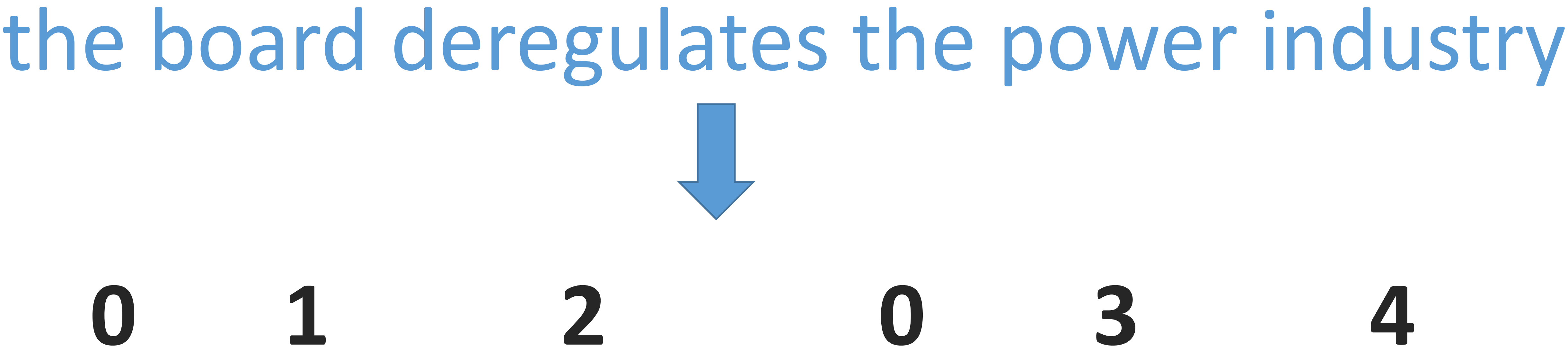
Bringing it together

- This can finally bring us to the point where we have tuples, say of the form (action, agent, patient, [theme])
  - e.g., (saw, man, bird, telescope)
- How can we use these?
- Maybe question answering:
  - build large database of tuples
  - for a new question:
    - map it to a tuple
    - match it against the database, fill in the slot

Language data: sequences of symbols?

the board deregulates the power industry

Language data: sequences of symbols?



the board deregulates the power industry

0 1 2 0 3 4

Language data: sequences of symbols?

the board deregulates the power industry

↓

0 1 2 0 3 4

Identity

deregulates != John revises tax and deregulate

Words aren’t atomic

deregulates

What’s in a word?

deregulates

What’s in a word?

deregulates

What’s in a word?

deregulates

Old-school AI approach

Create some formalism

Manually add entries to generate a “lexicon”

Rules for composing entries, inflecting words, etc

WordNet

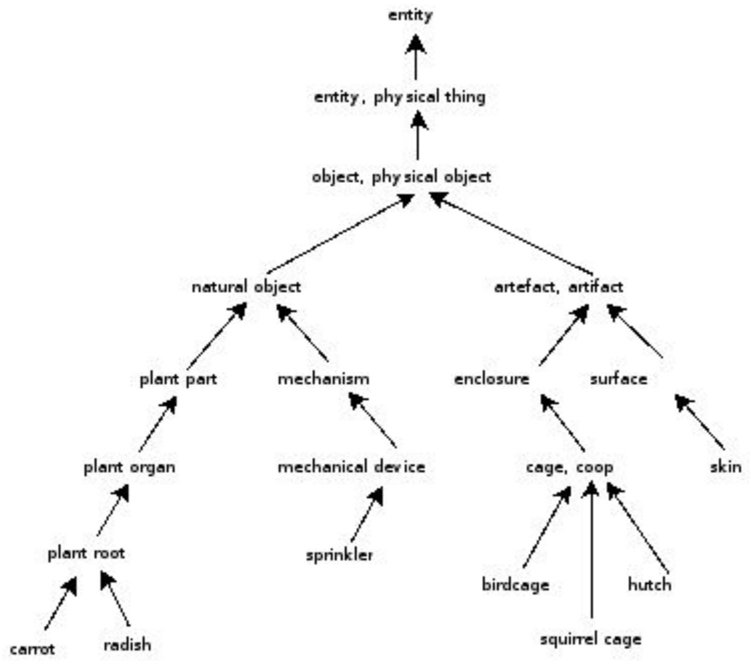


Figure 1. "is a" relation ex ample

Lots of problems

Tremendous amount of effort involved

Inconsistent coverage

Not data-driven

Word Embeddings

Represent a word as a ***vector of real numbers***

Word Embeddings

Represent a word as a ***vector of real numbers***

deregulates =

0.5
0.8
-0.6
1.3
0.0

Word Embeddings

Represent a word as a ***vector of real numbers***

deregulates =

0.5	0.4
0.8	0.0
-0.6	0.8
1.3	0.0
0.0	1.1

=

tariff

Word Embeddings

Represent a word as a ***vector of real numbers***

deregulates =

0.5	0.4	**Trade-relatedness**
0.8	0.0	**Present-tense-ish**
-0.6	0.8	**Polarity**
1.3	0.0	**Verb-like**
0.0	1.1	**Noun-like**

=

tariff

Word Embeddings

Many useful properties

Word similarity via cosine-similarity

Document similarity, e.g. compare average vectors

Composition: Brother – Man + Woman = Sister, ran – run = “past tense”

“You will know a word by the company it keeps” –Firth, 1957

Representing a word’s “neighborhood”

Documents	cat	milk	vet	farm	cow	dog
1	0	2	3	12	10	2
2	12	1	11	0	0	0
3	1	11	0	1	14	0
4	0	0	13	1	0	10

We want to learn “topics”

Documents	cat	milk	vet	farm	cow	dog
1	0	2	3	12	10	2
2	12	1	11	0	0	0
3	1	11	0	1	14	0
4	0	0	13	1	0	10

We want to learn “topics”

Documents	cat	milk	vet	farm	cow	dog
1	0	2	3	12	10	2
2	12	1	11	0	0	0
3	1	11	0	1	14	0
4	0	0	13	1	0	10

We want to learn “topics”

Documents	cat	milk	vet	farm	cow	dog
1	0	2	3	12	10	2
2	12	1	11	0	0	0
3	1	11	0	1	14	0
4	0	0	13	1	0	10

Beyond equality: how similar are words/docs?

Word-by-word	dog	cat	cow	milk
dog	Max	High	Low	Low
cat	High	Max	Low	Low
cow	Low	Low	Max	High
milk	Low	Low	High	Max

Doc-by-doc	1	2	3	4
1	Max	Low	High	Low
2	Low	Max	Low	High
3	High	Low	Max	Low
4	Low	High	Low	Max

Pets!

Word-by-word	dog	cat	cow	milk
dog	Max	High	Low	Low
cat	High	Max	Low	Low
cow	Low	Low	Max	High
milk	Low	Low	High	Max

Doc-by-doc	1	2	3	4
1	Max	Low	High	Low
2	Low	Max	Low	High
3	High	Low	Max	Low
4	Low	High	Low	Max

Livestock!

Word-by-word	dog	cat	cow	milk
dog	Max	High	Low	Low
cat	High	Max	Low	Low
cow	Low	Low	Max	High
milk	Low	Low	High	Max

Doc-by-doc	1	2	3	4
1	Max	Low	High	Low
2	Low	Max	Low	High
3	High	Low	Max	Low
4	Low	High	Low	Max

One approach: matrix factorization

Low-rank matrix approximation

Total-least-squares

Latent Semantic Analysis/Indexing

Principle Components Analysis

Singular Value Decomposition

One approach: matrix factorization

Low-rank matrix approximation

Total-least-squares

Latent Semantic Analysis/Indexing

Principle Components Analysis

Singular Value Decomposition

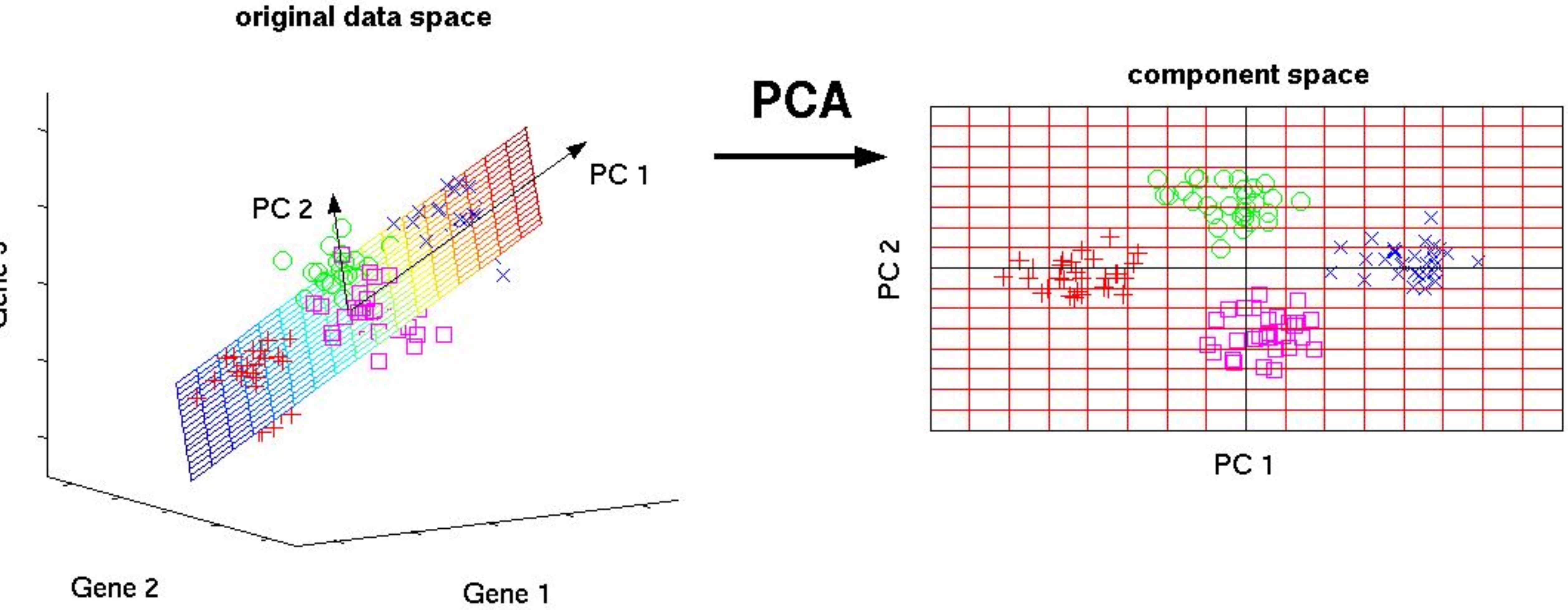
Express matrix as a product of (simpler) matrices

Involves de-correlating various components

E.g. the simpler matrices are orthonormal

Usually some calculation of eigenvalues/vectors

PCA: Principle component analysis



SVD: Singular value decomposition

$$M = U\Sigma \mathrm{V}^{\mathrm{T}}$$

U and V are orthonormal

Σ is diagonal

SVD: Singular value decomposition

$$M = U\Sigma V^*$$

$$\mathbf{U} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{V}^* = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{0.2} & 0 & 0 & 0 & \sqrt{0.8} \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix}$$

LSA: Latent Semantic Analysis

$$M = U\Sigma V^*$$

$$\mathbf{U} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{V}^* = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{0.2} & 0 & 0 & 0 & \sqrt{0.8} \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix}$$

LSA: Latent Semantic Analysis

$$M = U\Sigma V^*$$

Dimensions of Σ correspond to the number of topics

A row in U corresponds to the document's "topic proportions"

A column in V* corresponds to a topic's "word mixture"

LSA: Latent Semantic Analysis

Topic 1 = [1.3 * dog, 2.4 * cat, -.3 * horse, …]

Document 1 = [.98 * topic 1, -1.8 * topic 2, …]

We Want a Story!

Connect well-formed probability distributions

This “generative story” explains how the data came to be

Describe this via a Bayesian Network (BN)

Example: Car Alarm

Example: Car Alarm

E

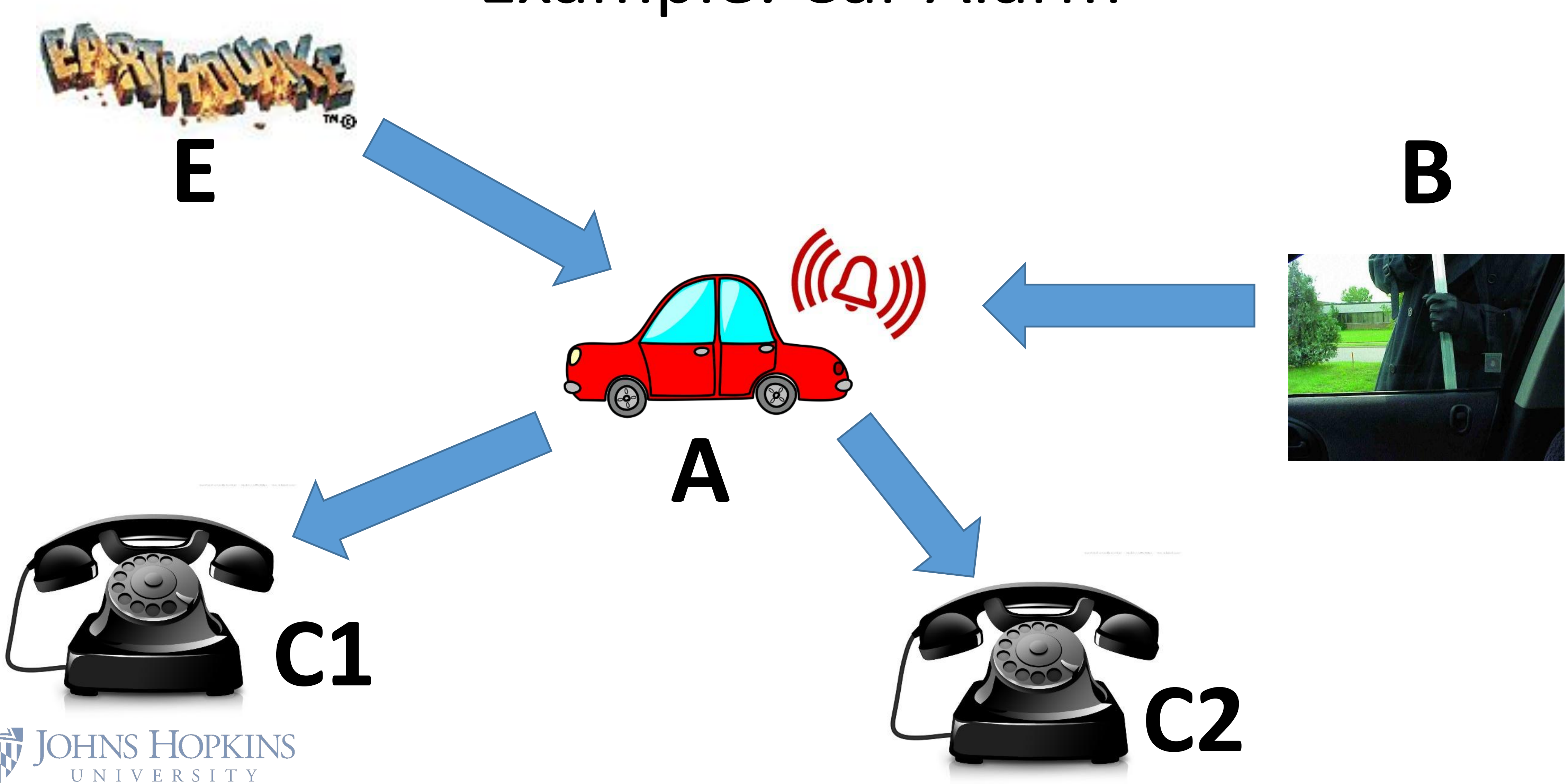
B

A

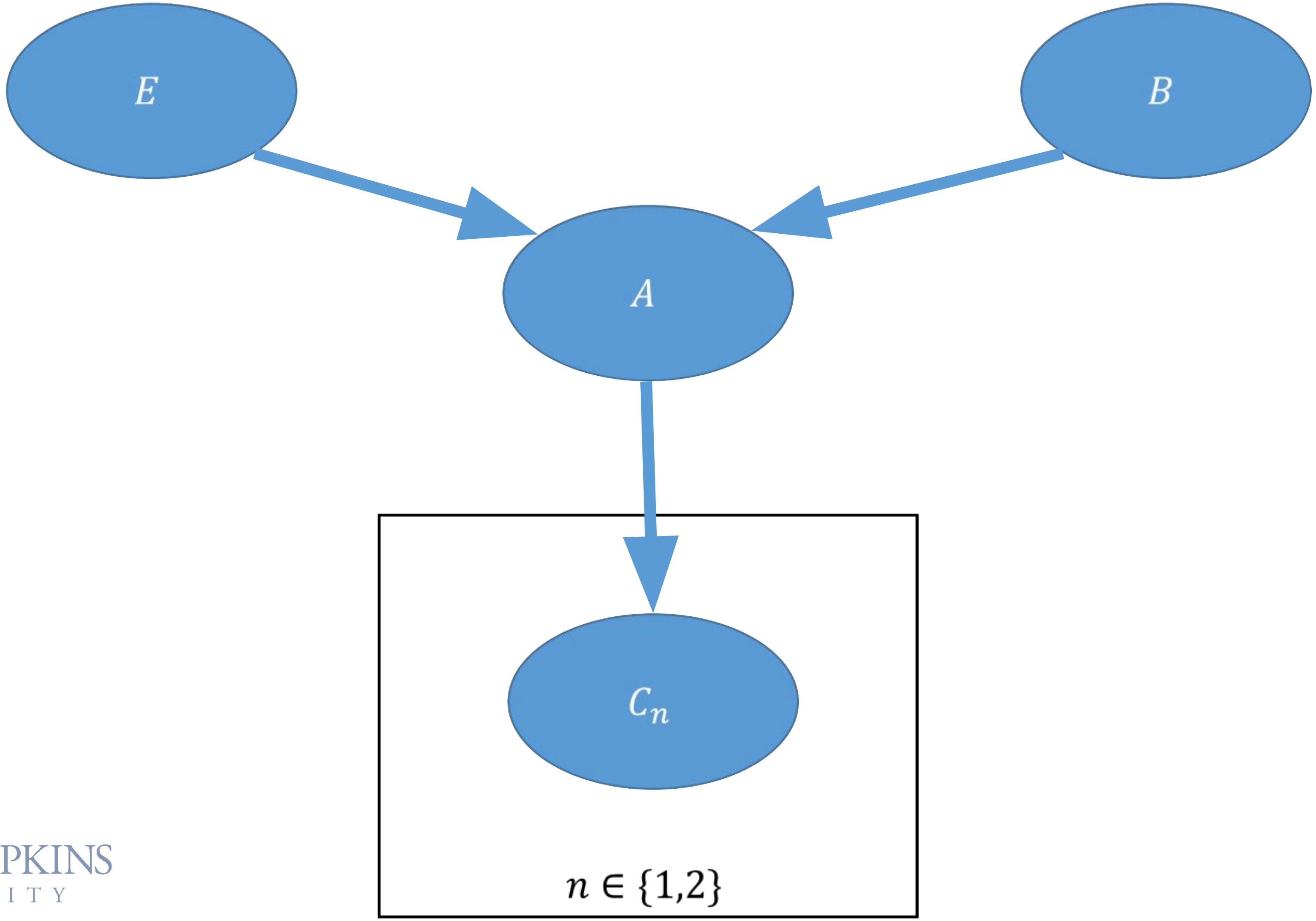
C1

C2

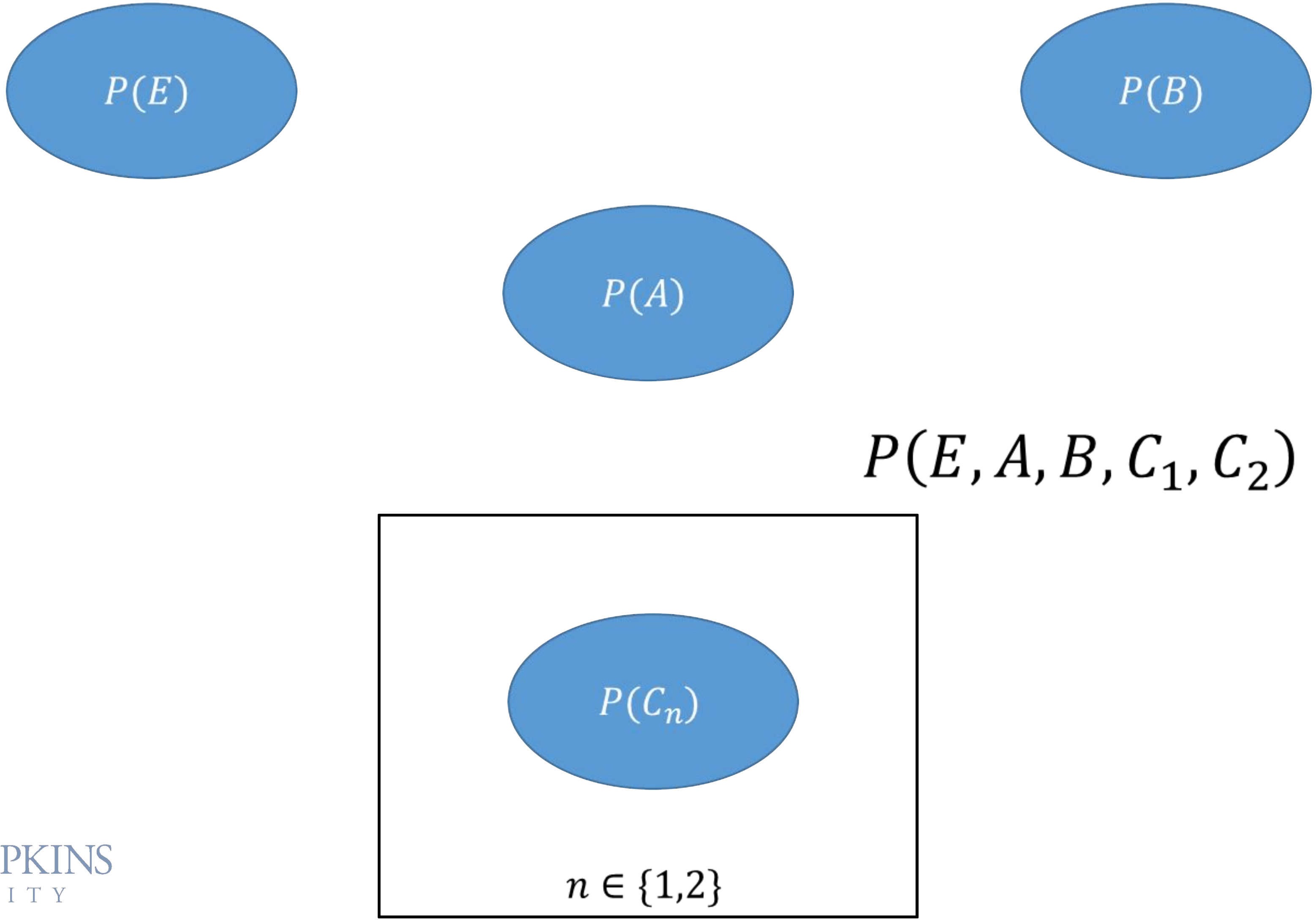
Example: Car Alarm



Bayesian Networks

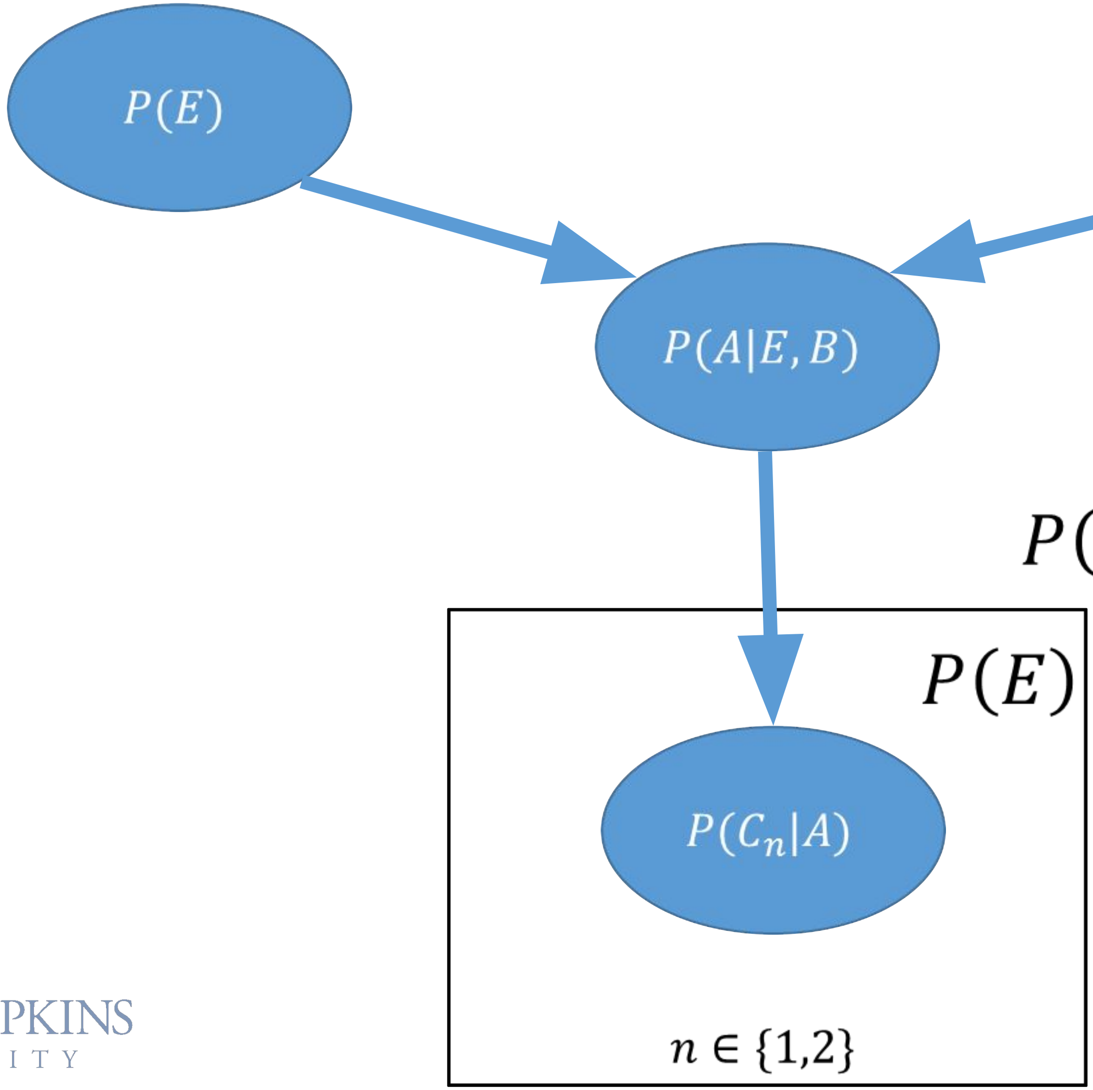


Bayesian Networks



$P(E, A, B, C_1, C_2)$

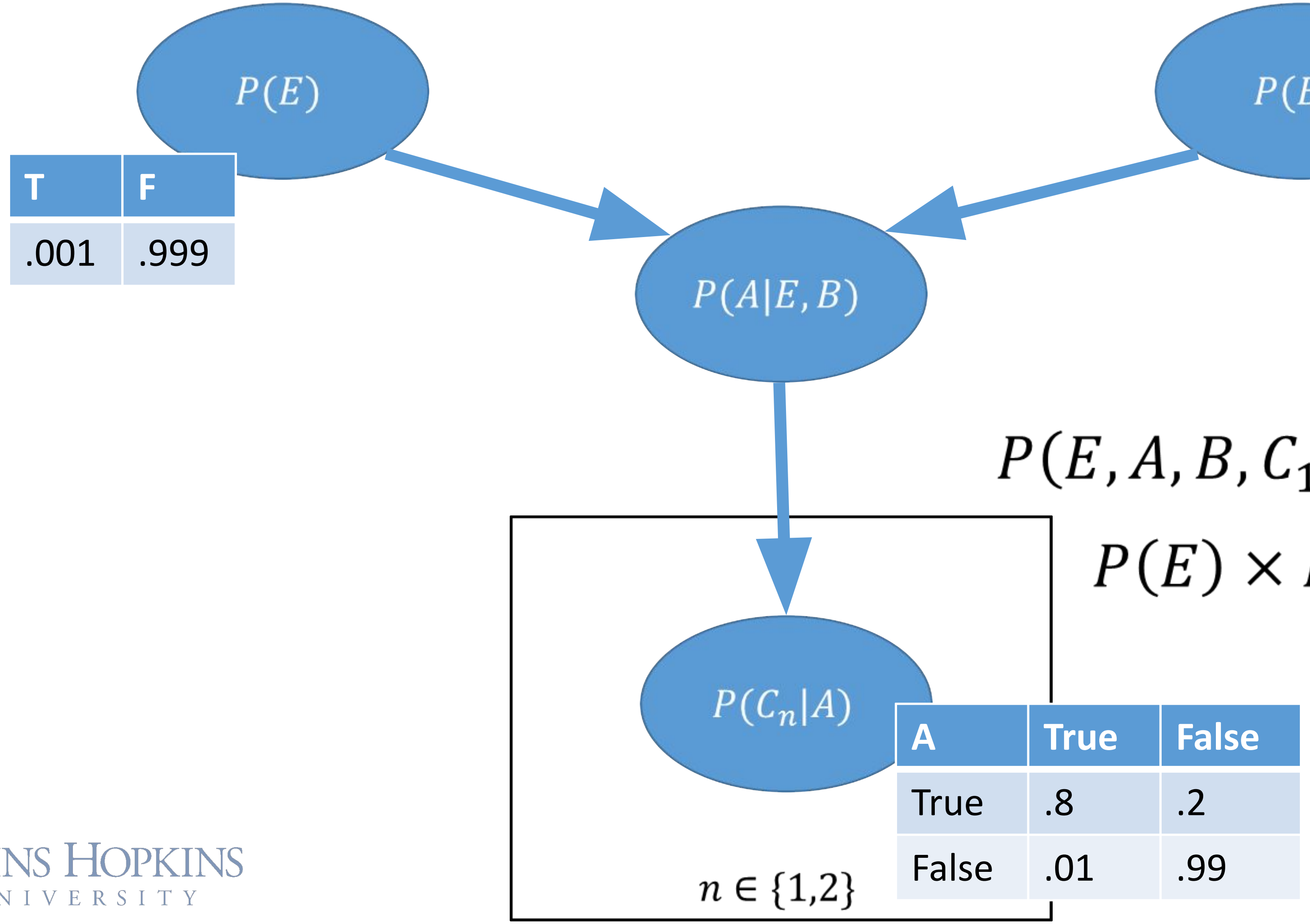
Bayesian Networks



$$P(E, A, B, C_1, C_2) =$$

$$P(E) \times P(B) \times P(A|E,B) \times \prod P(C_n|A)$$

Bayesian Networks



T	F
.001	.999

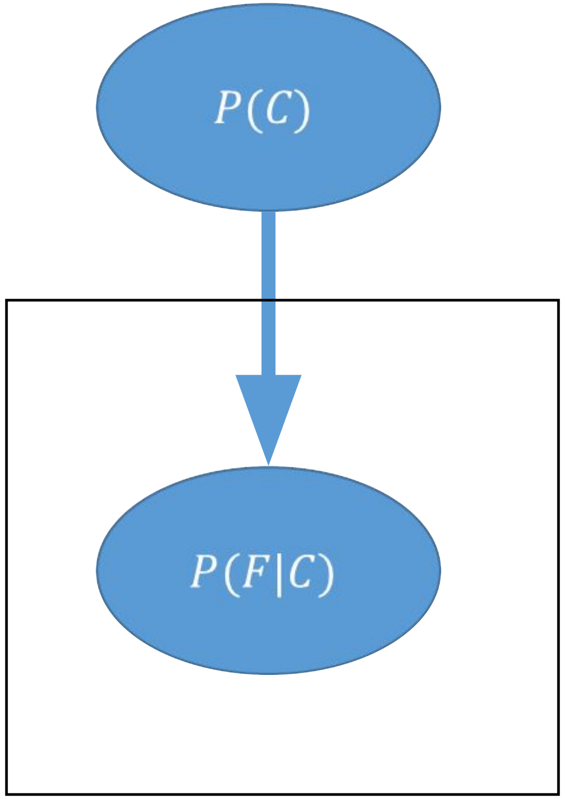
T	F
.1	.9

A	True	False
True	.8	.2
False	.01	.99

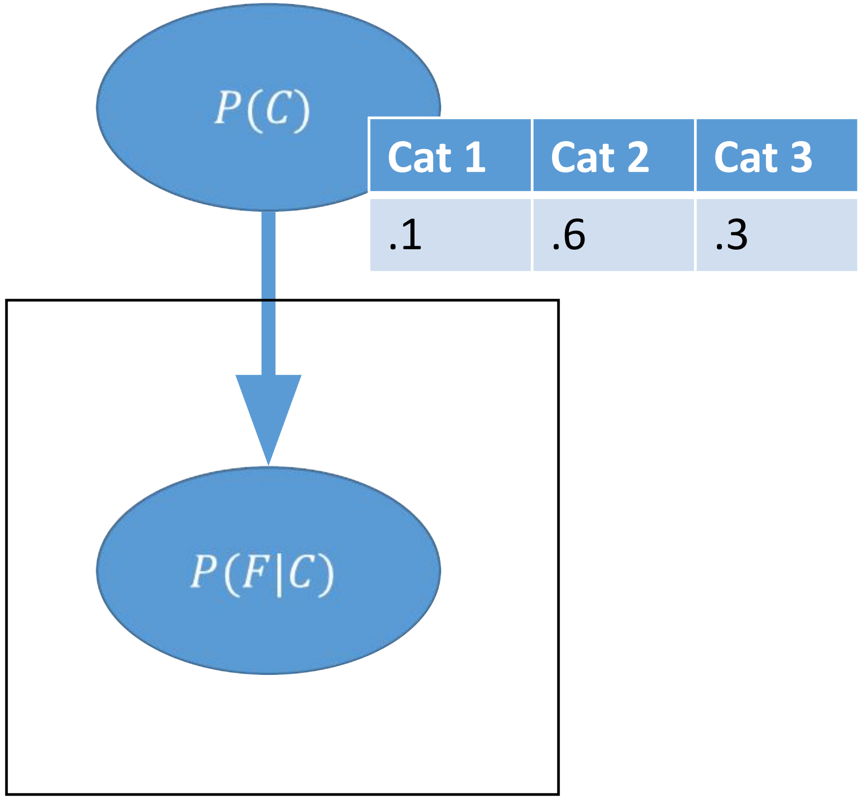
$$P(E, A, B, C_1, C_2) =$$

$$P(E) \times P(A) \times P(B) \times \prod P(C_n)$$

Naïve Bayes is a simple story

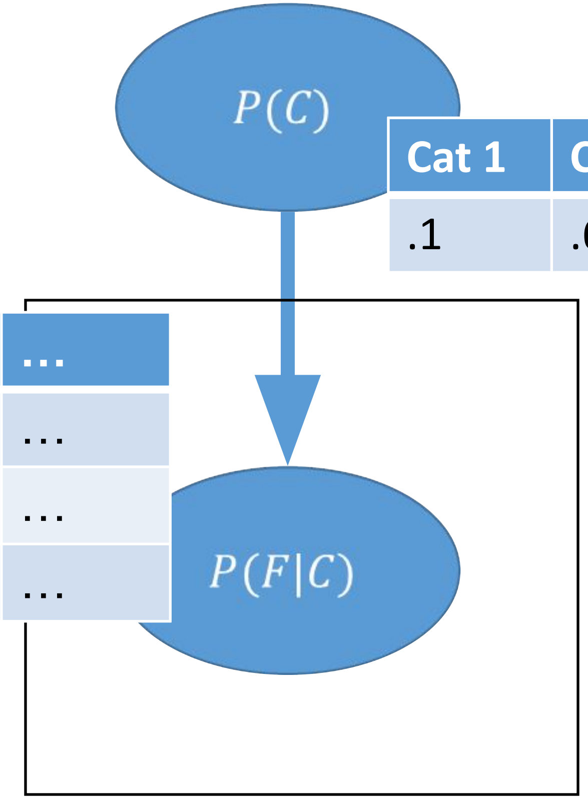


Naïve Bayes is a simple story



Cat 1	Cat 2	Cat 3
.1	.6	.3

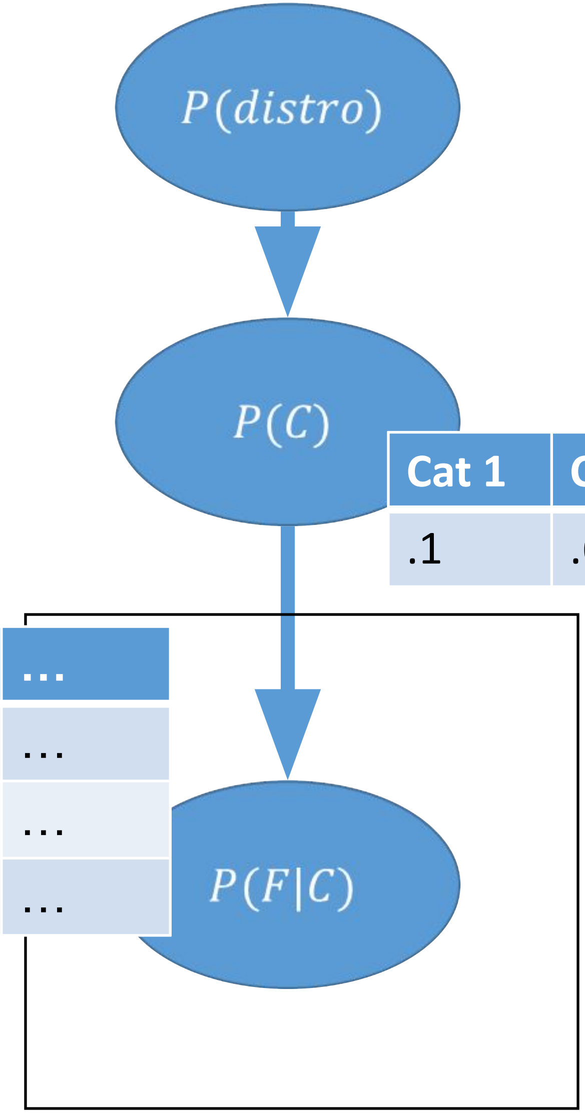
Naïve Bayes is a simple story



Cat 1	Cat 2	Cat 3
.1	.6	.3

Cat	house	frog	…
1	.01	.003	…
2	.5	.001	…
3	.2	.02	…

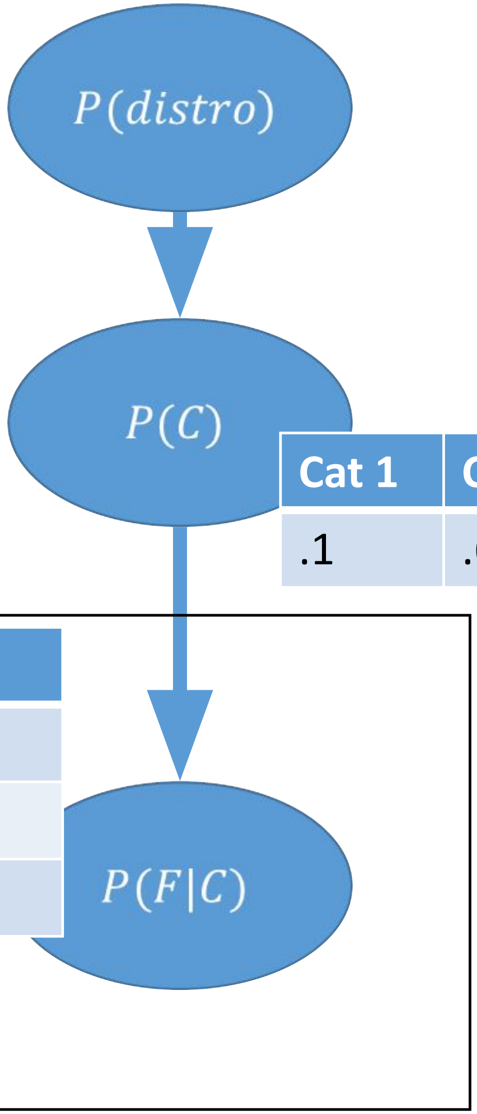
Naïve Bayes is a simple story



Cat 1	Cat 2	Cat 3
.1	.6	.3

Cat	house	frog	…
1	.01	.003	…
2	.5	.001	…
3	.2	.02	…

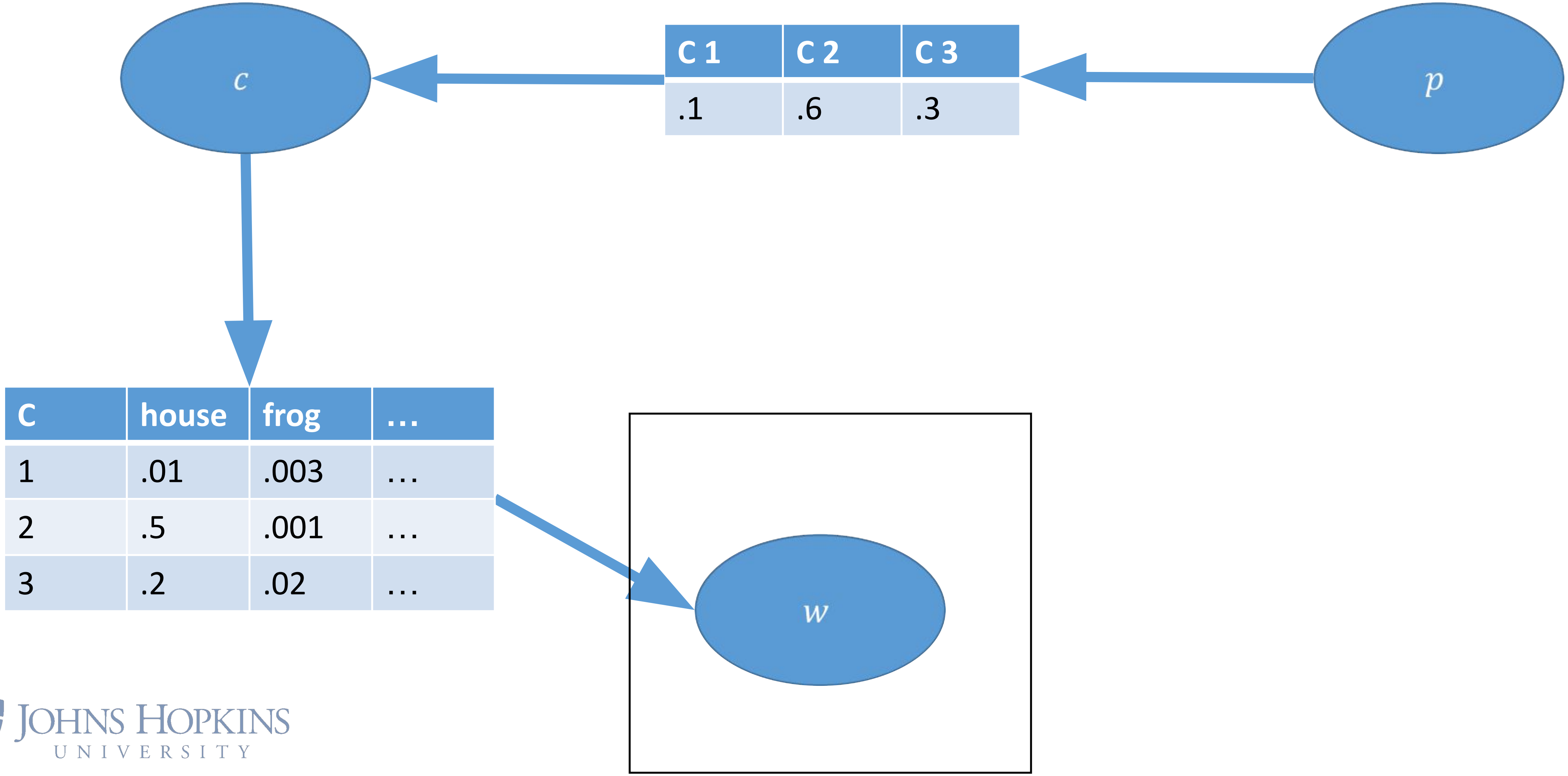
Side note: BNs versus Factor Graphs



Cat 1	Cat 2	Cat 3
.1	.6	.3

Cat	house	frog	…
1	.01	.003	…
2	.5	.001	…
3	.2	.02	…

Side note: BNs versus Factor Graphs



C 1	C 2	C 3
.1	.6	.3

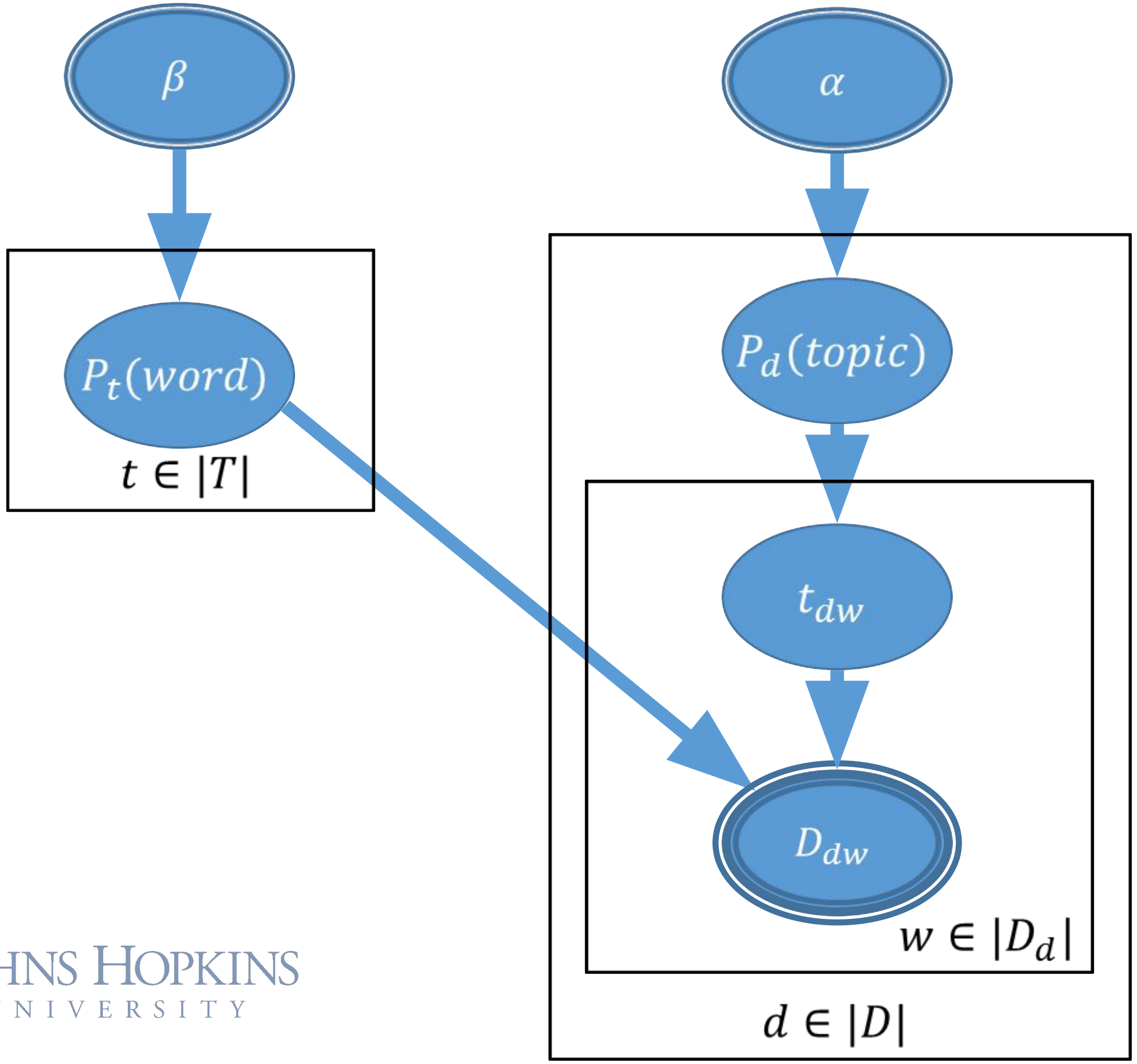
C	house	frog	…
1	.01	.003	…
2	.5	.001	…
3	.2	.02	…

LDA: Latent Dirichlet Analysis (err…. “allocation”)

“Latent” because it focuses on unobserved “topic” variables

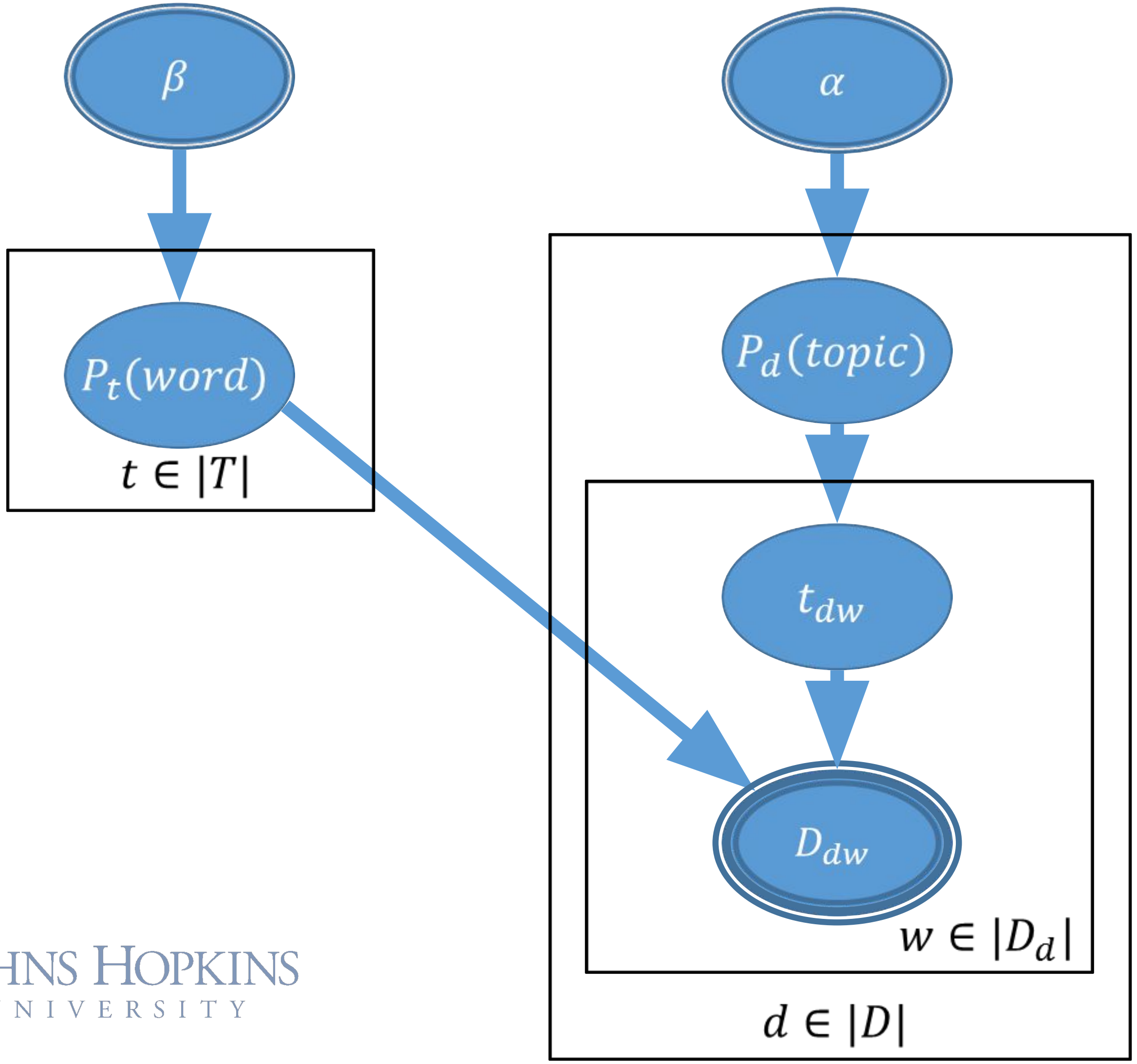
“Dirichlet” because this distribution is used as the prior to capture intuitions and improve inference

LDA: data, model, and story



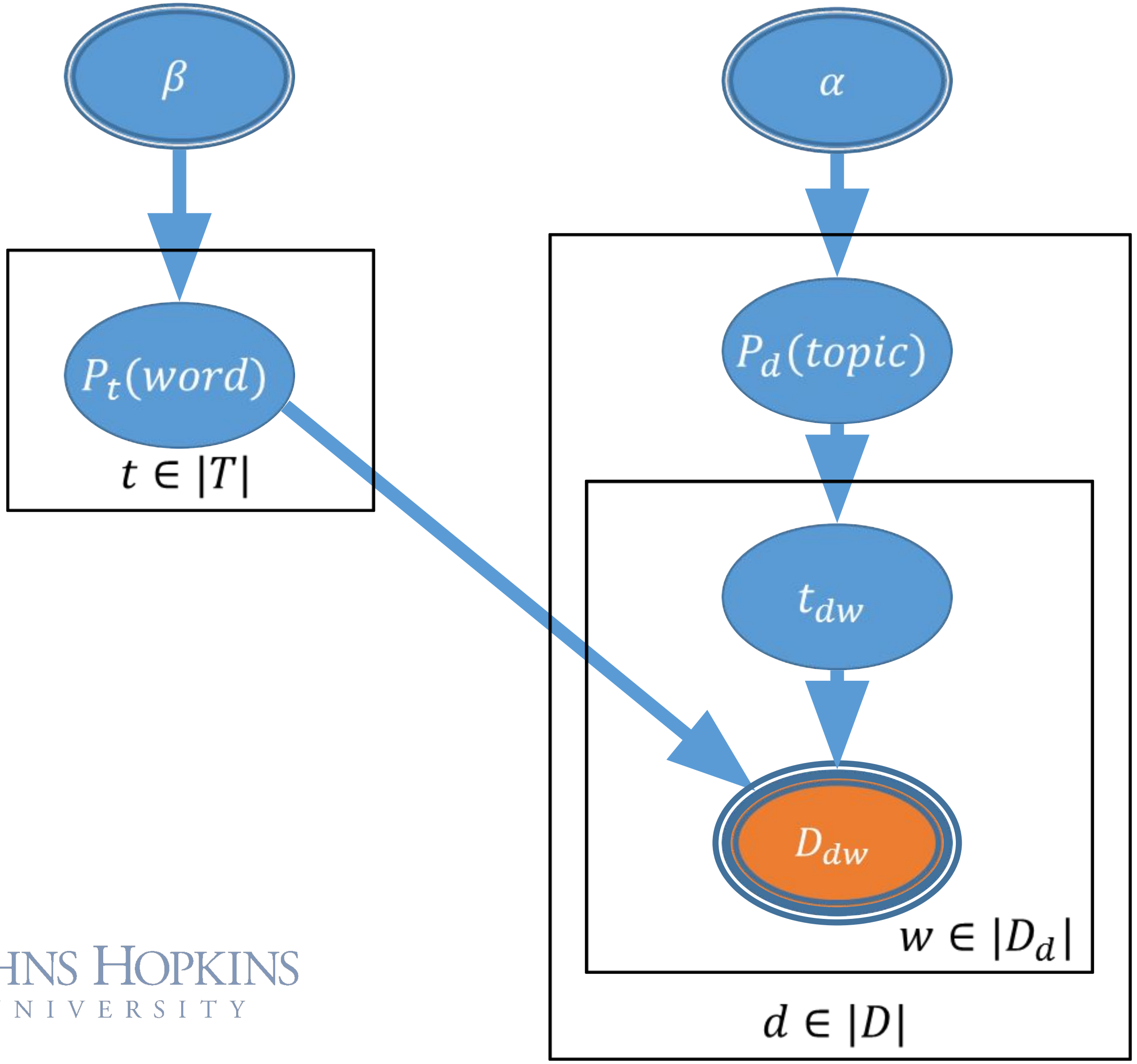
	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

Ragged data matrix (could be rectangular…)



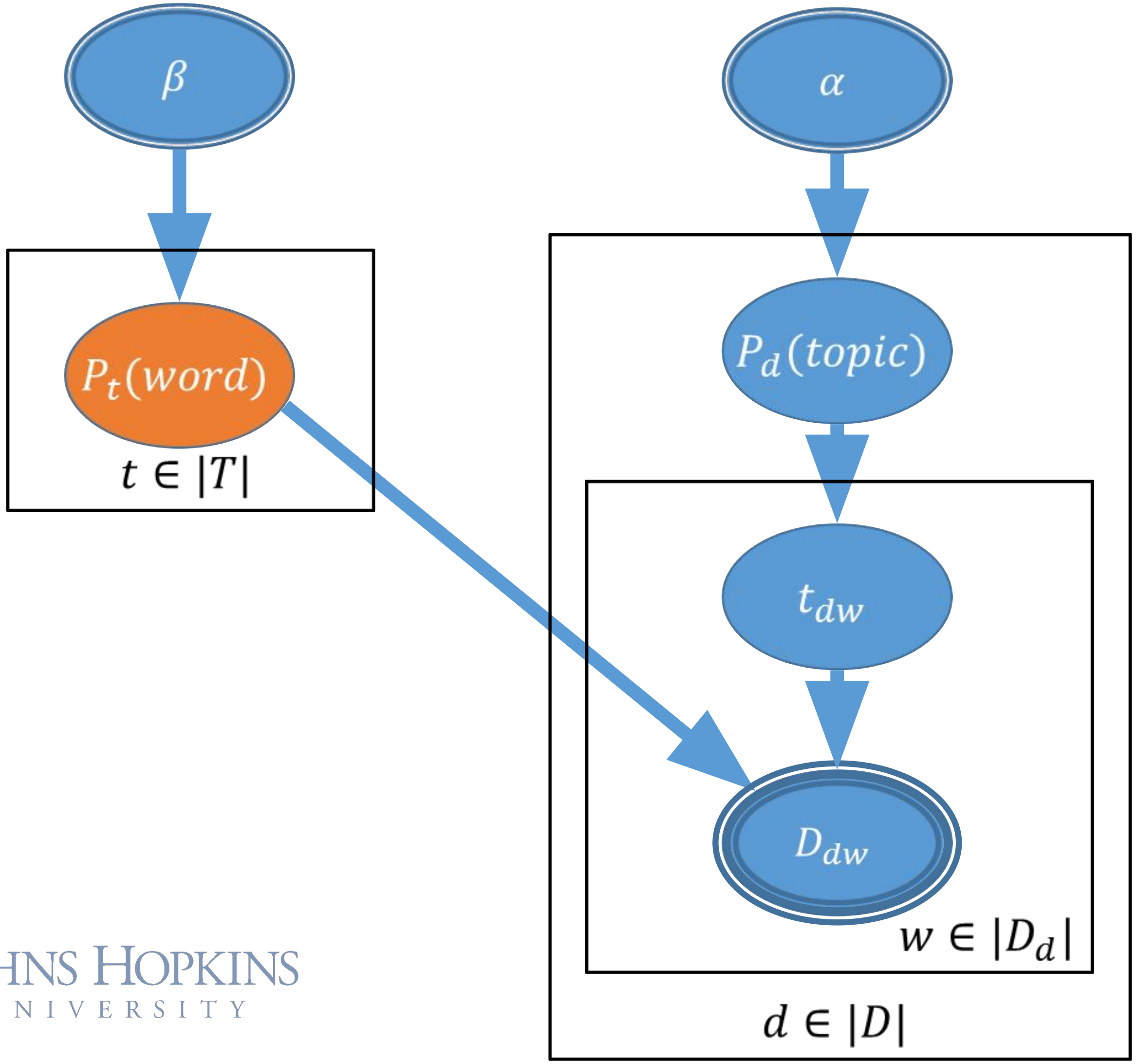
	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

Corresponds to the observed variable



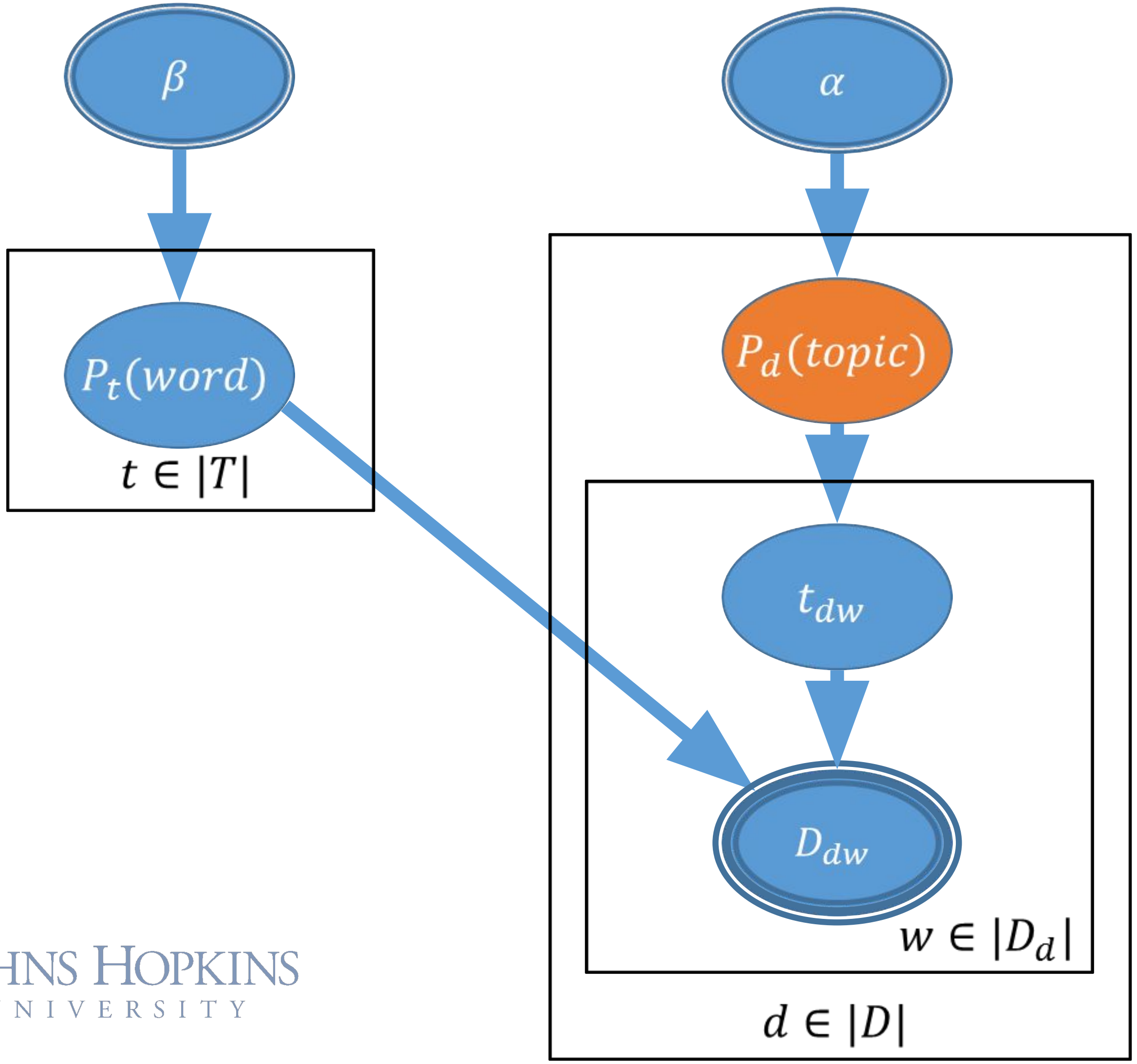
	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

We think there are “topics”



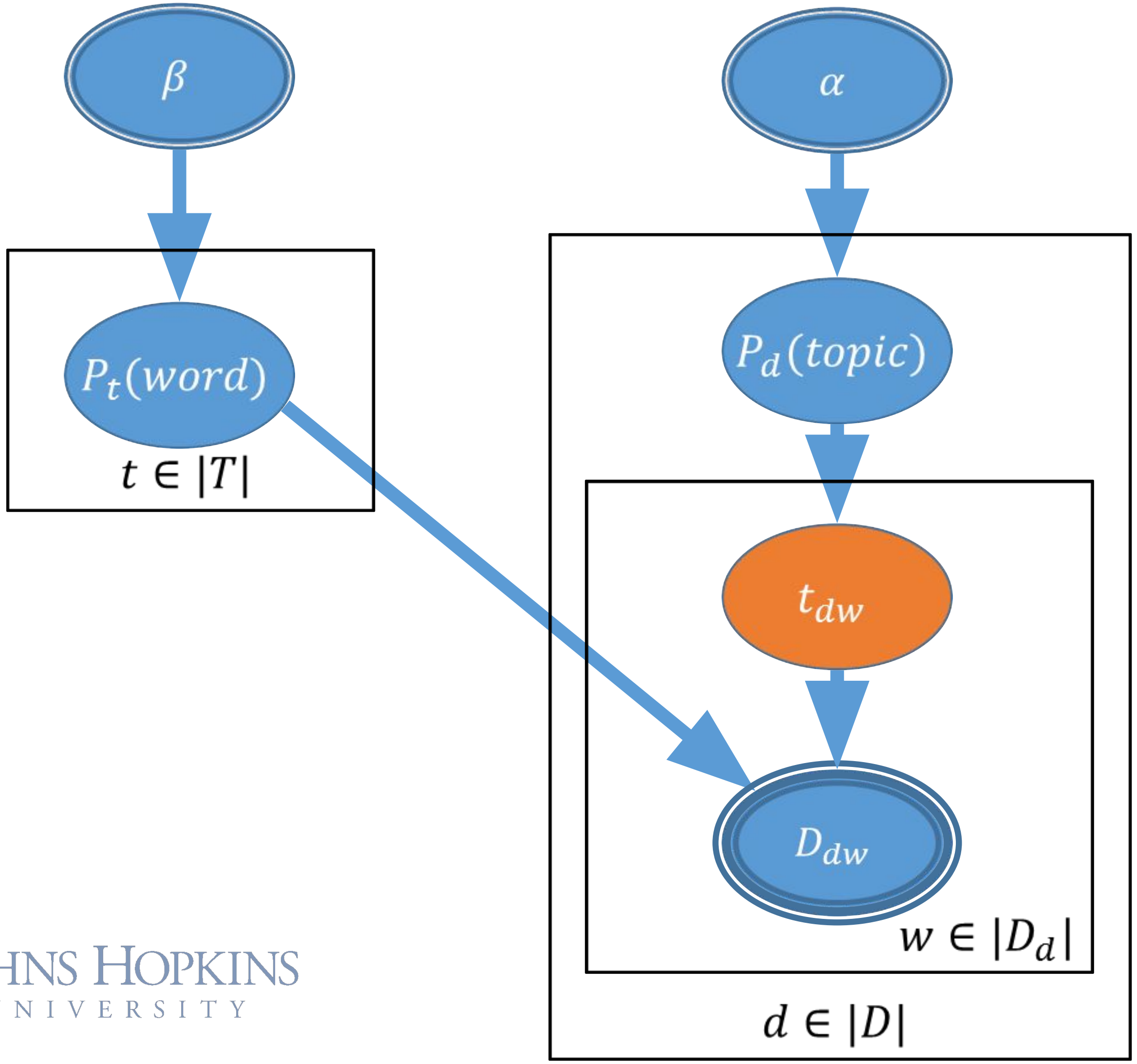
	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

Each document is some mixture of topics



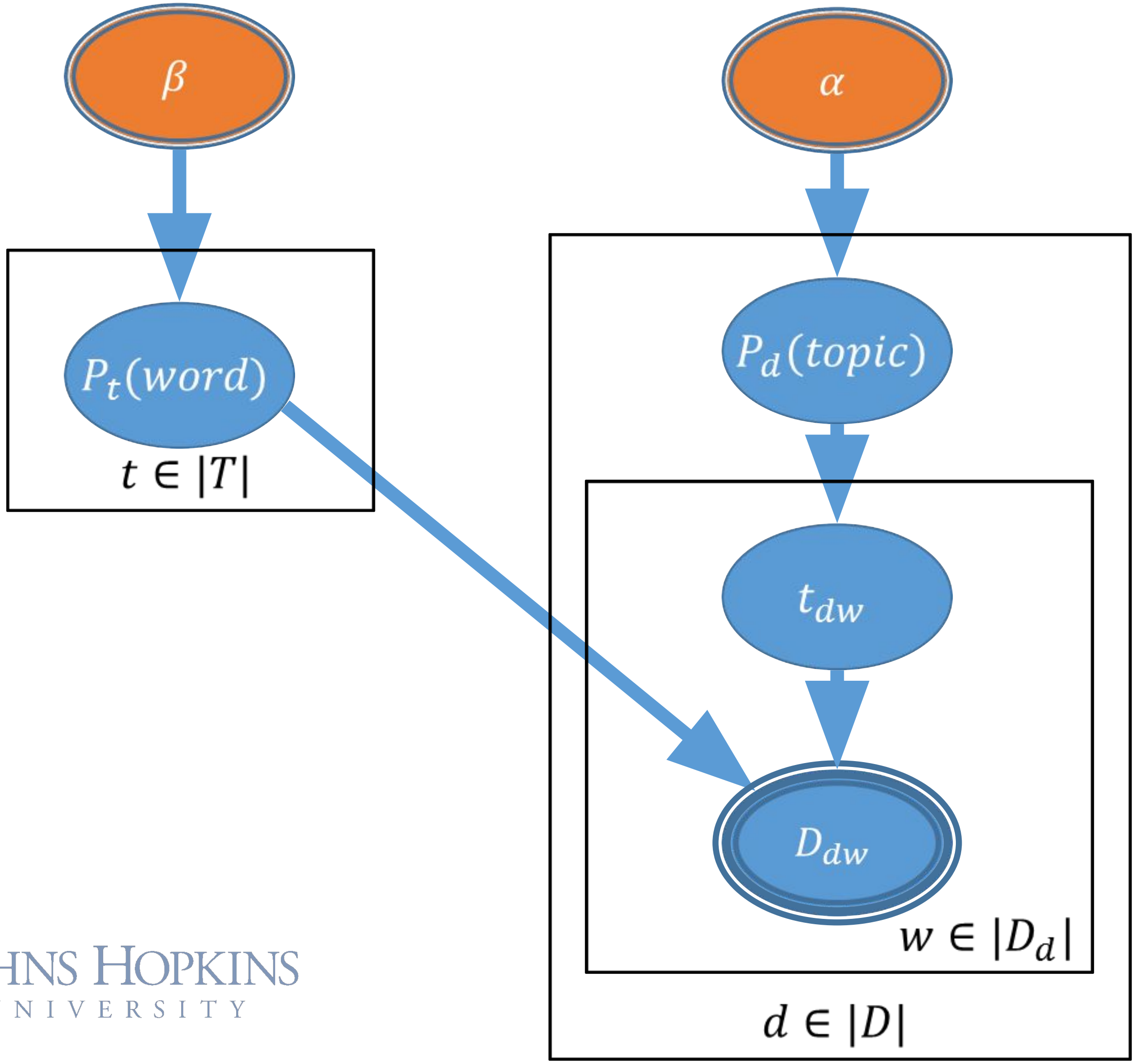
	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

Each word is associated with a topic



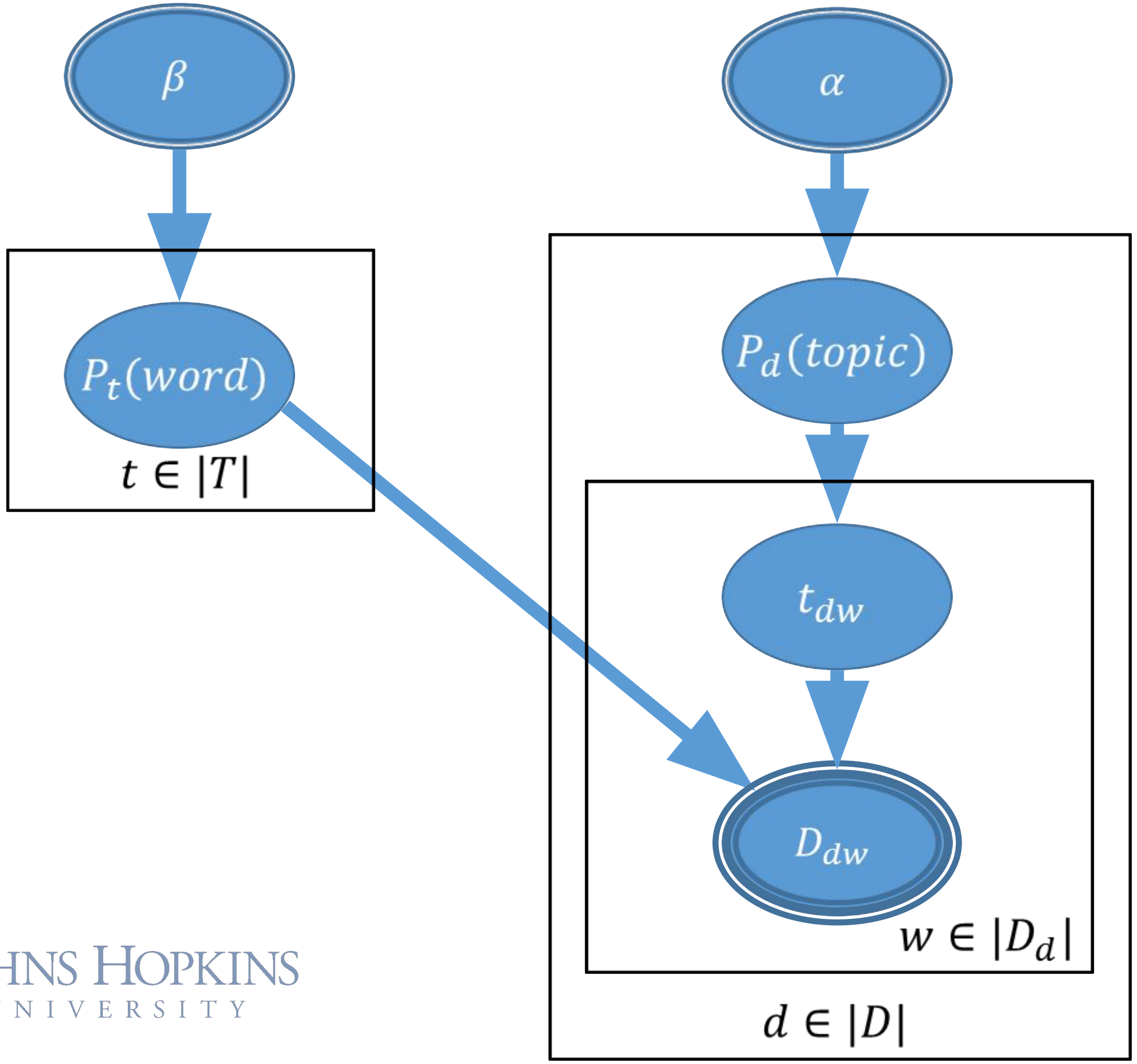
	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

α and β are "hyper-parameters"



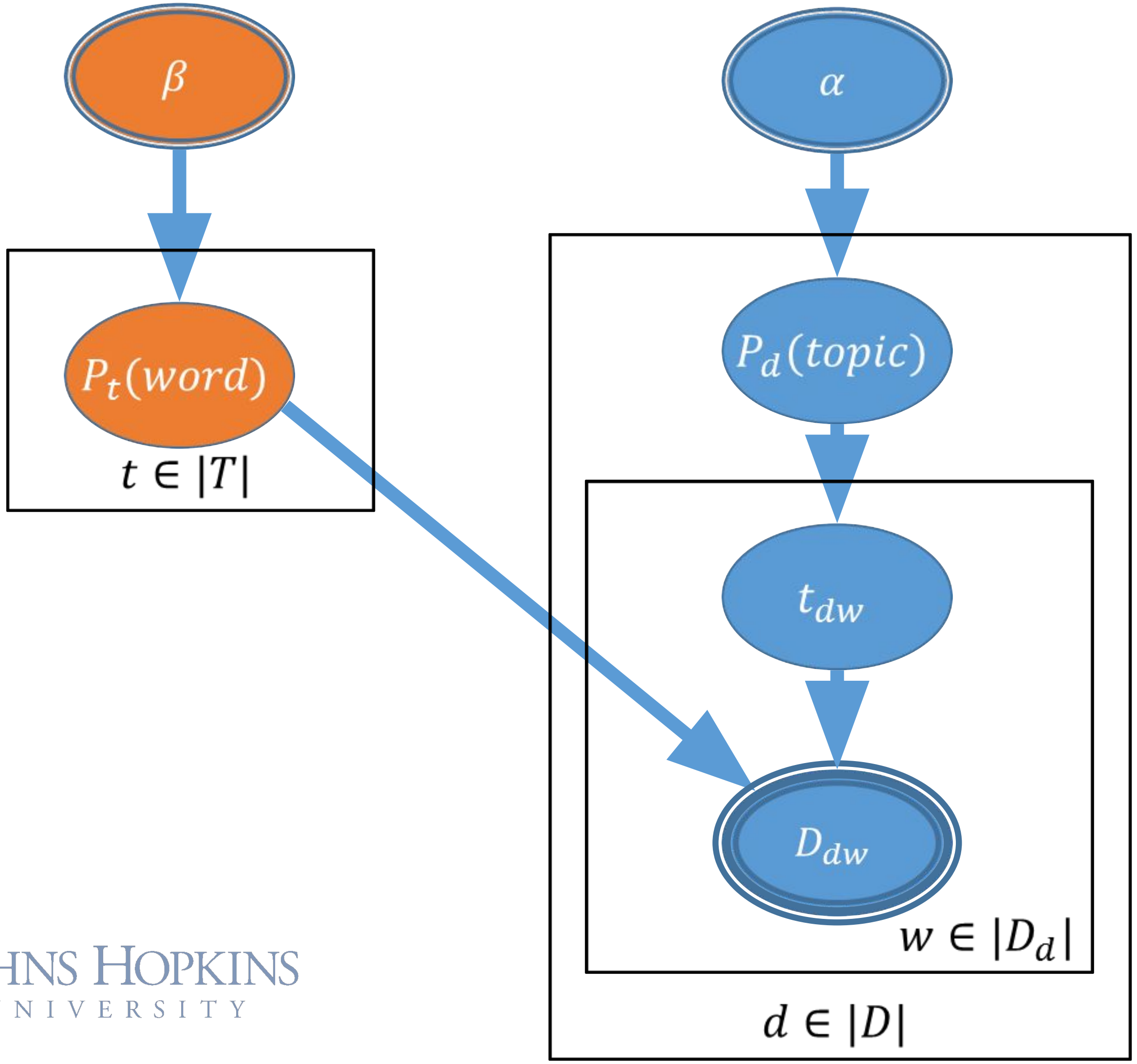
	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

Here’s “how we generate documents”:



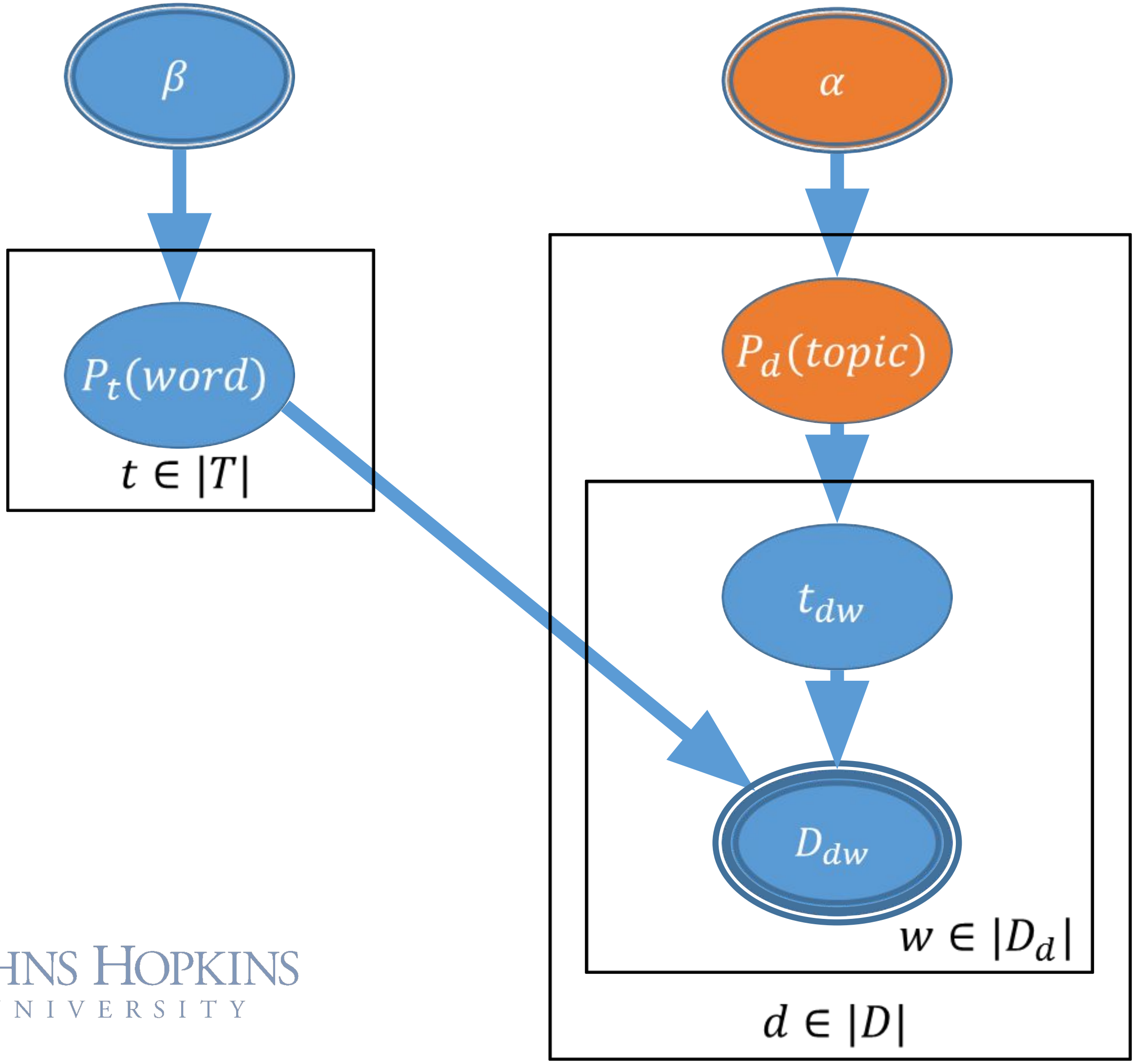
	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

Topic distributions are chosen “first”



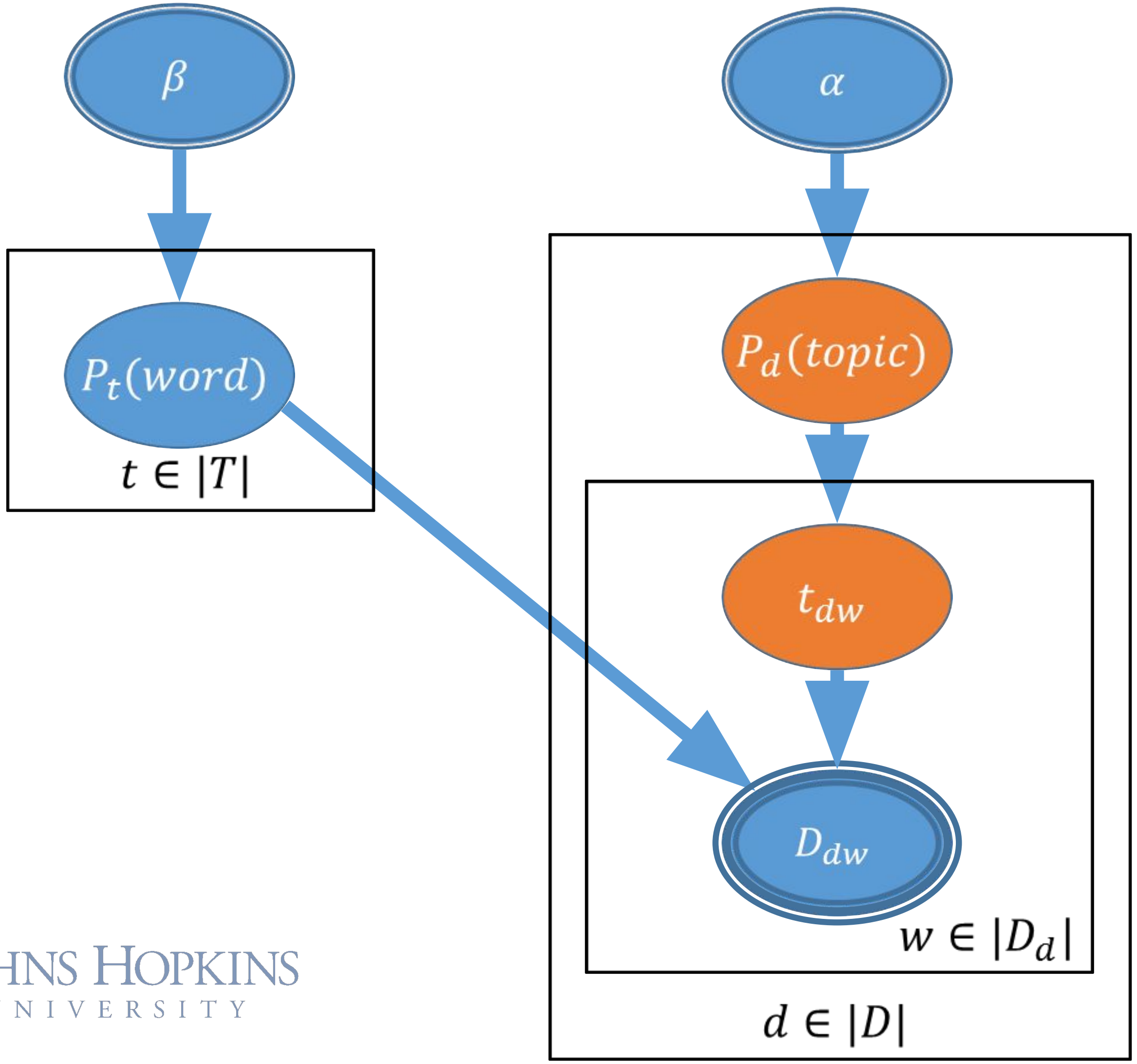
	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

New document: pick topic proportions



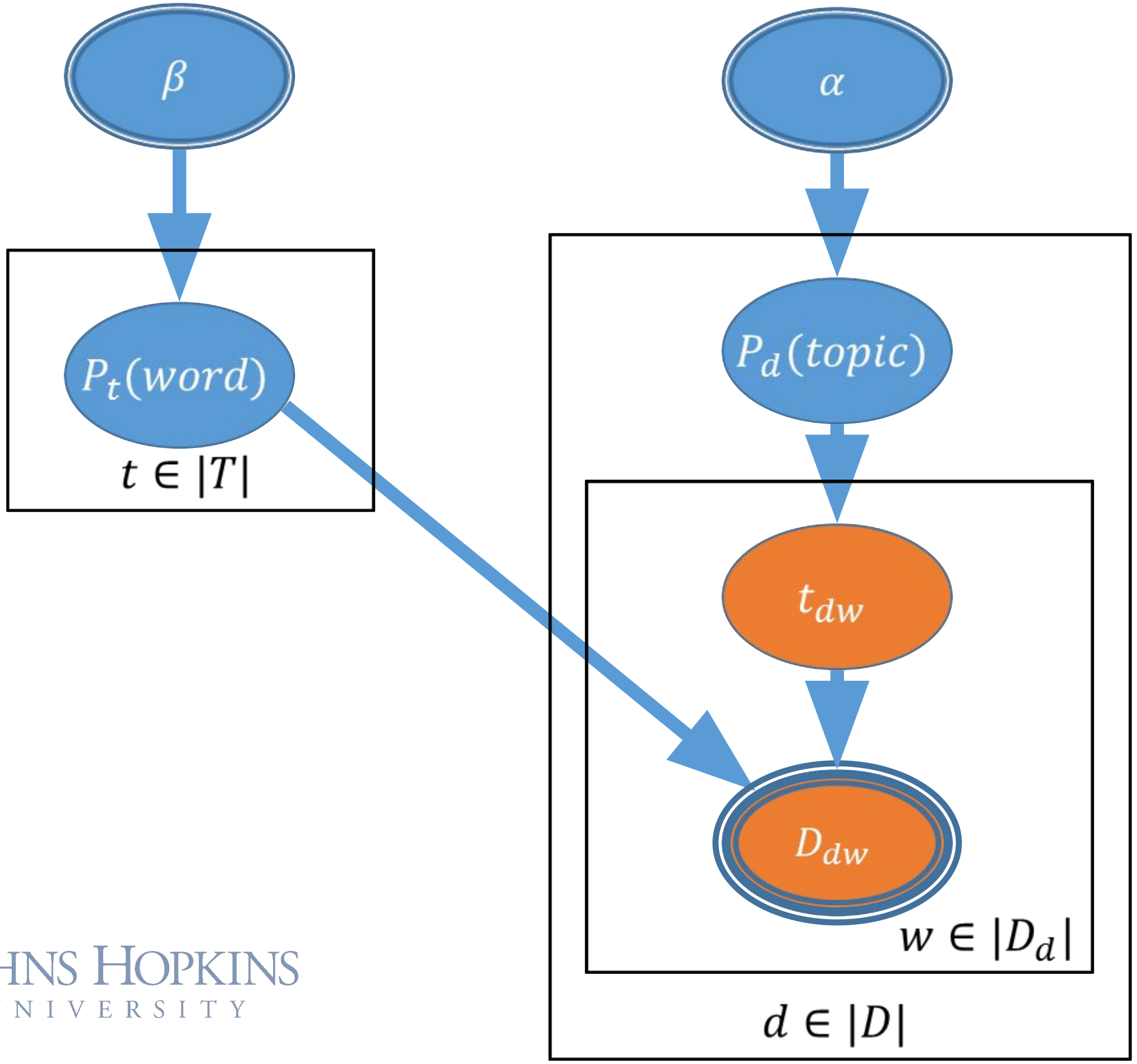
	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

Each word: pick a topic from the document's distribution



	1	2	3	4
1	This	is	a	sentence
2	So	is	this	
…				

Pick a word from that topic’s distribution



	1	2	3	4
1	1 This	1 is	1 a	2 sentence
2	3 So	1 is	2 this	
…				

Markov Chain Monte Carlo

Markov chain: a “state” is a complete assignment of hidden variables

Monte Carlo: the next state is chosen at random (“sampled”)

Gibbs Sampling

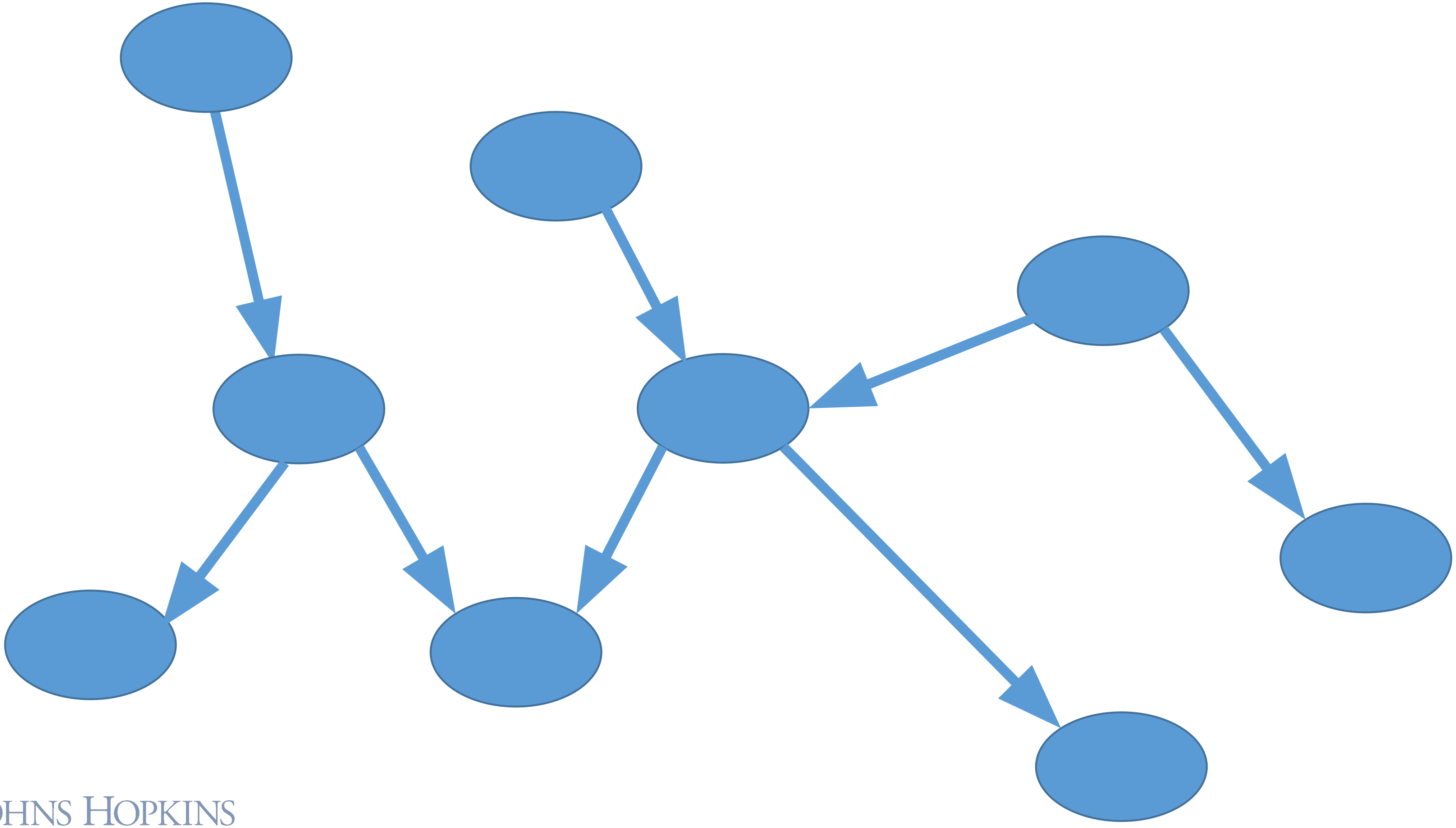
Initialize all variables to values in their range

Calculate probability of this configuration

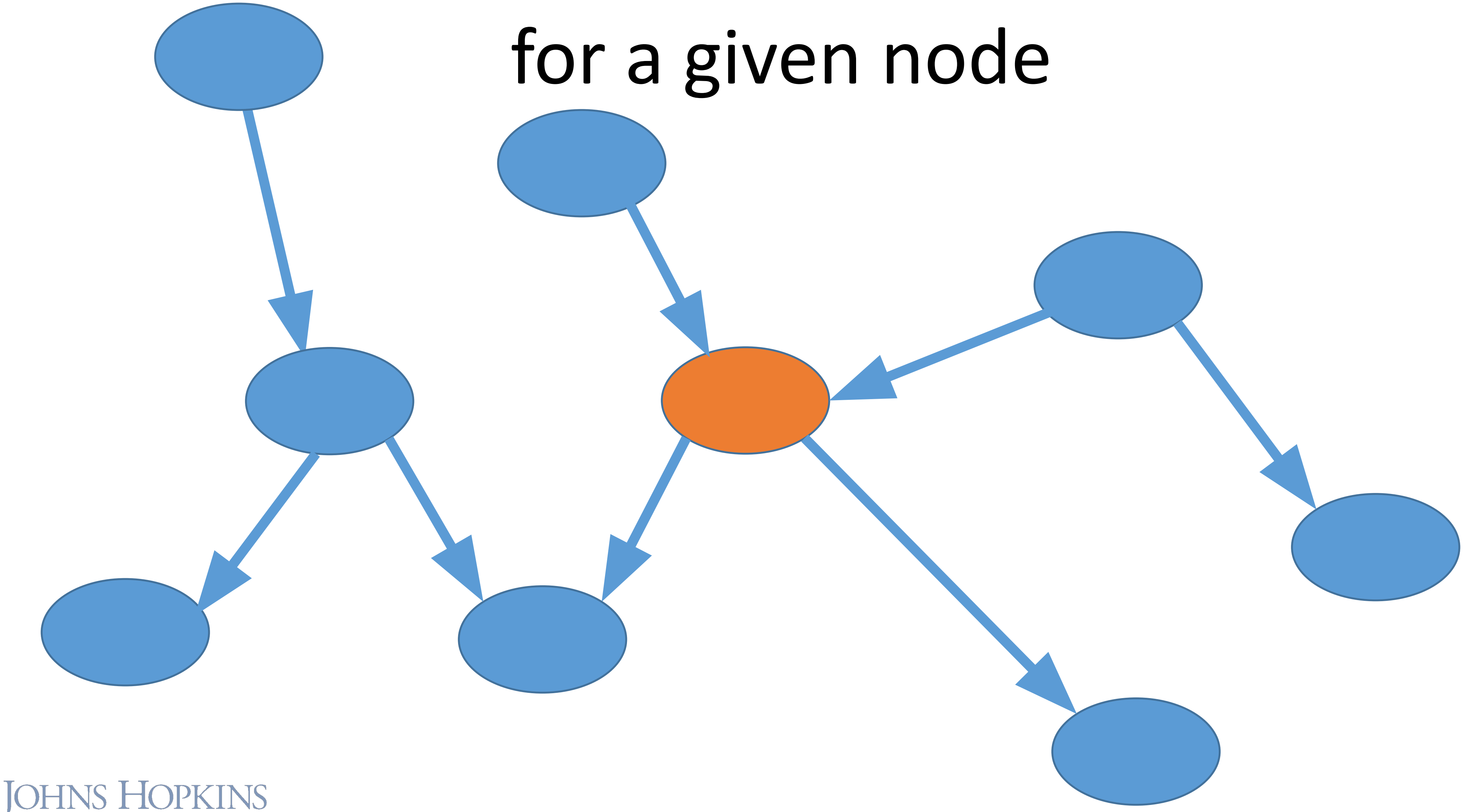
Loop:

- Pick variable at random
- Remove its value from consideration
- Choose a new value for it based on its “Markov blanket”

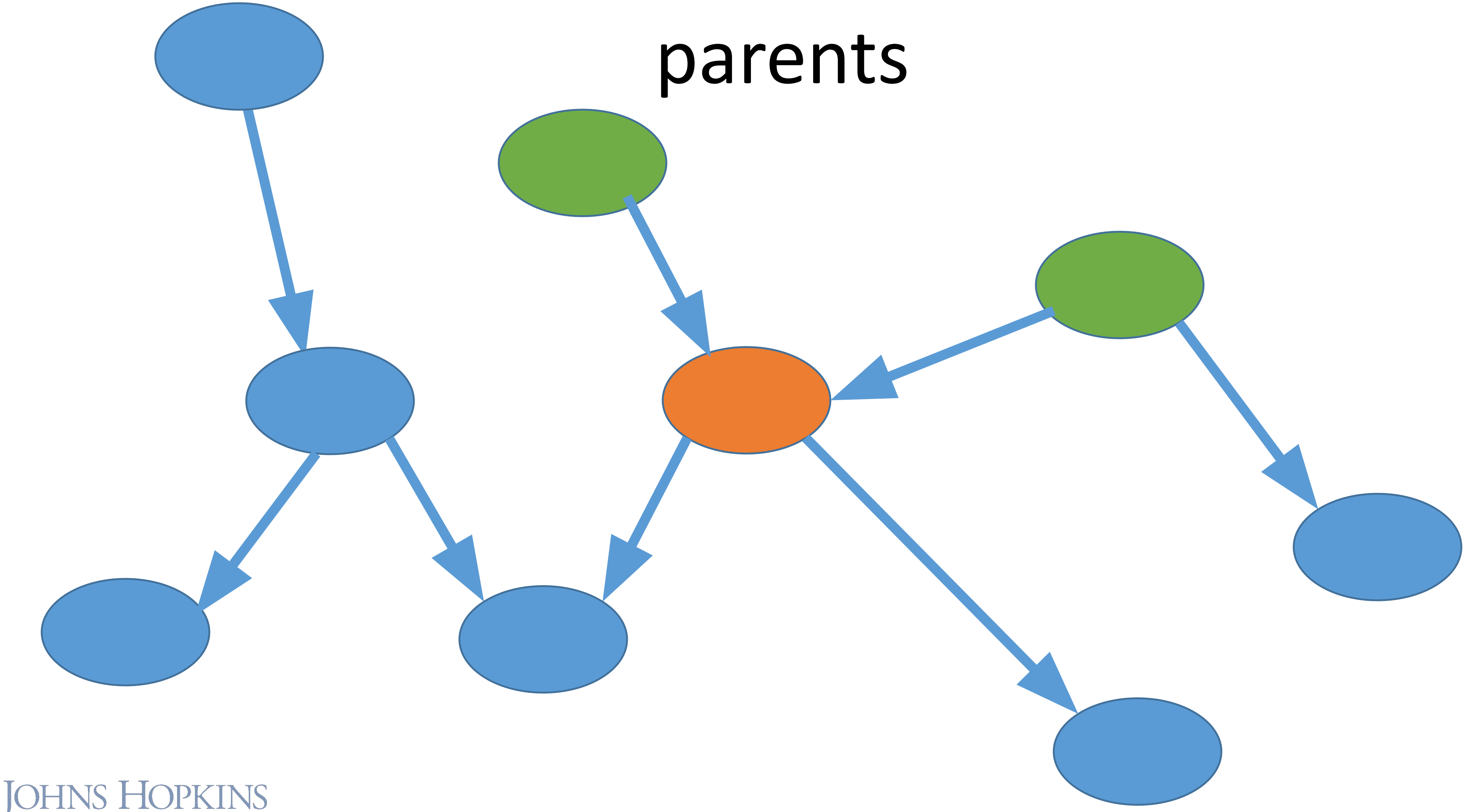
Markov Blanket



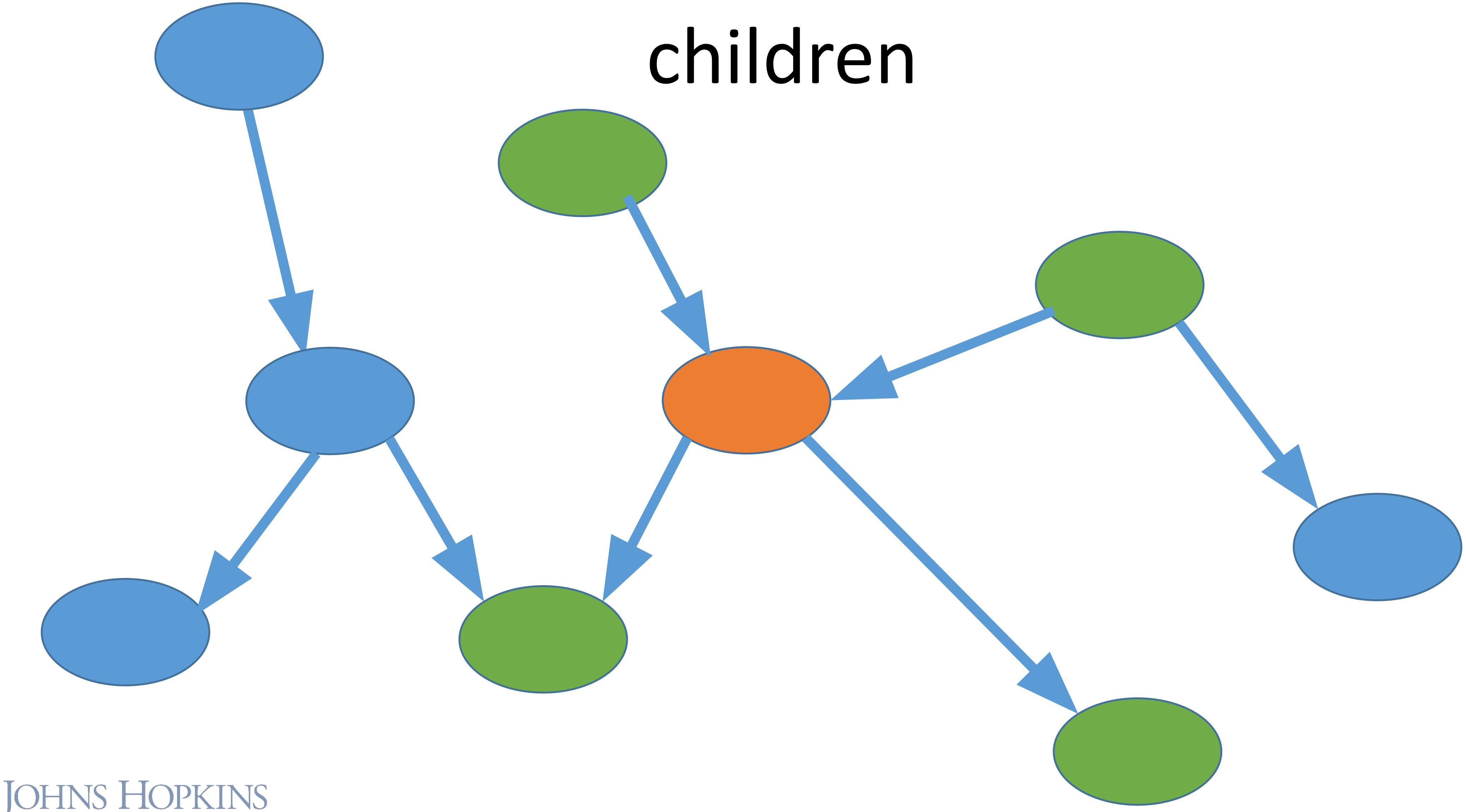
Markov Blanket for a given node



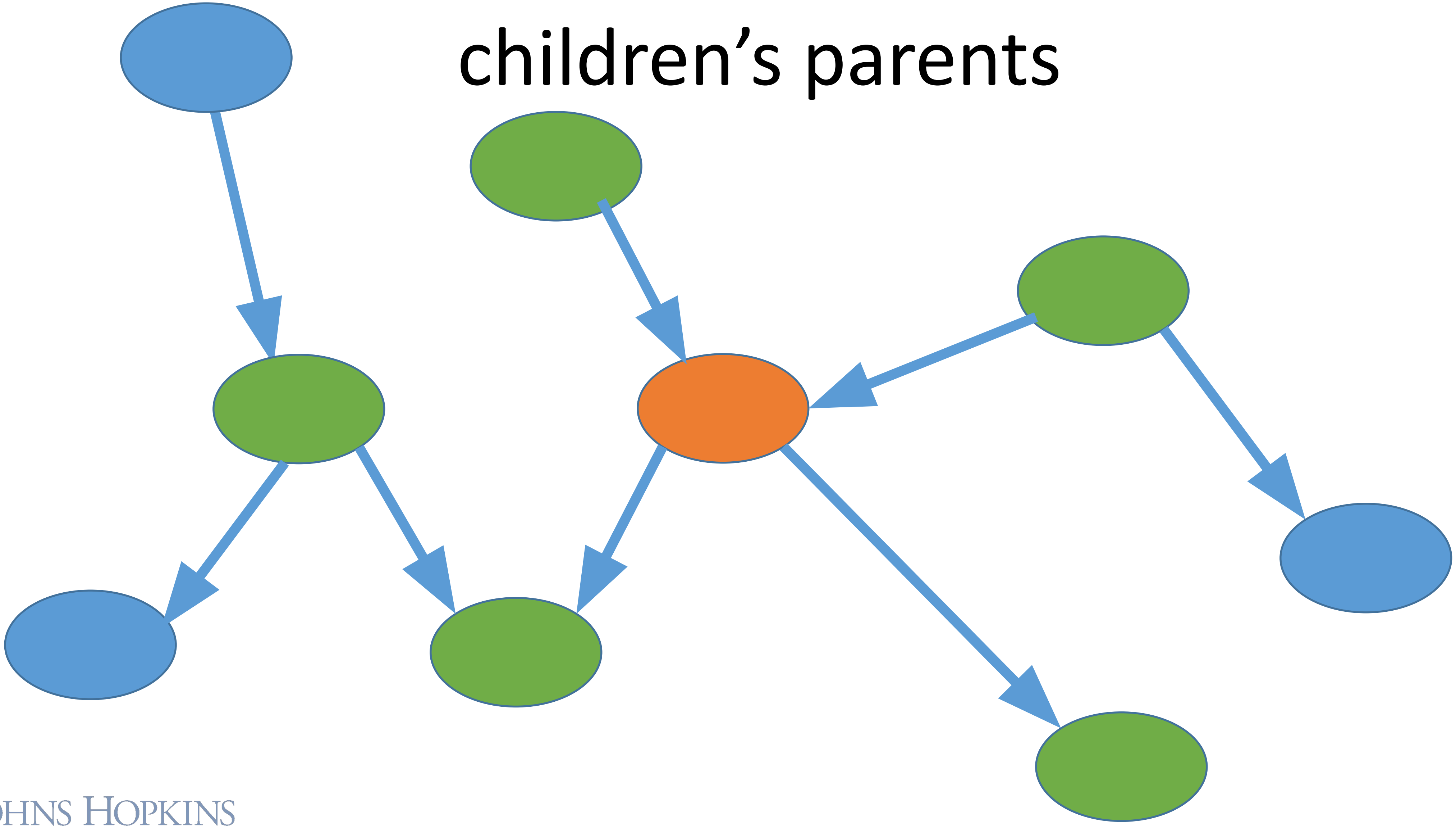
Markov Blanket
parents



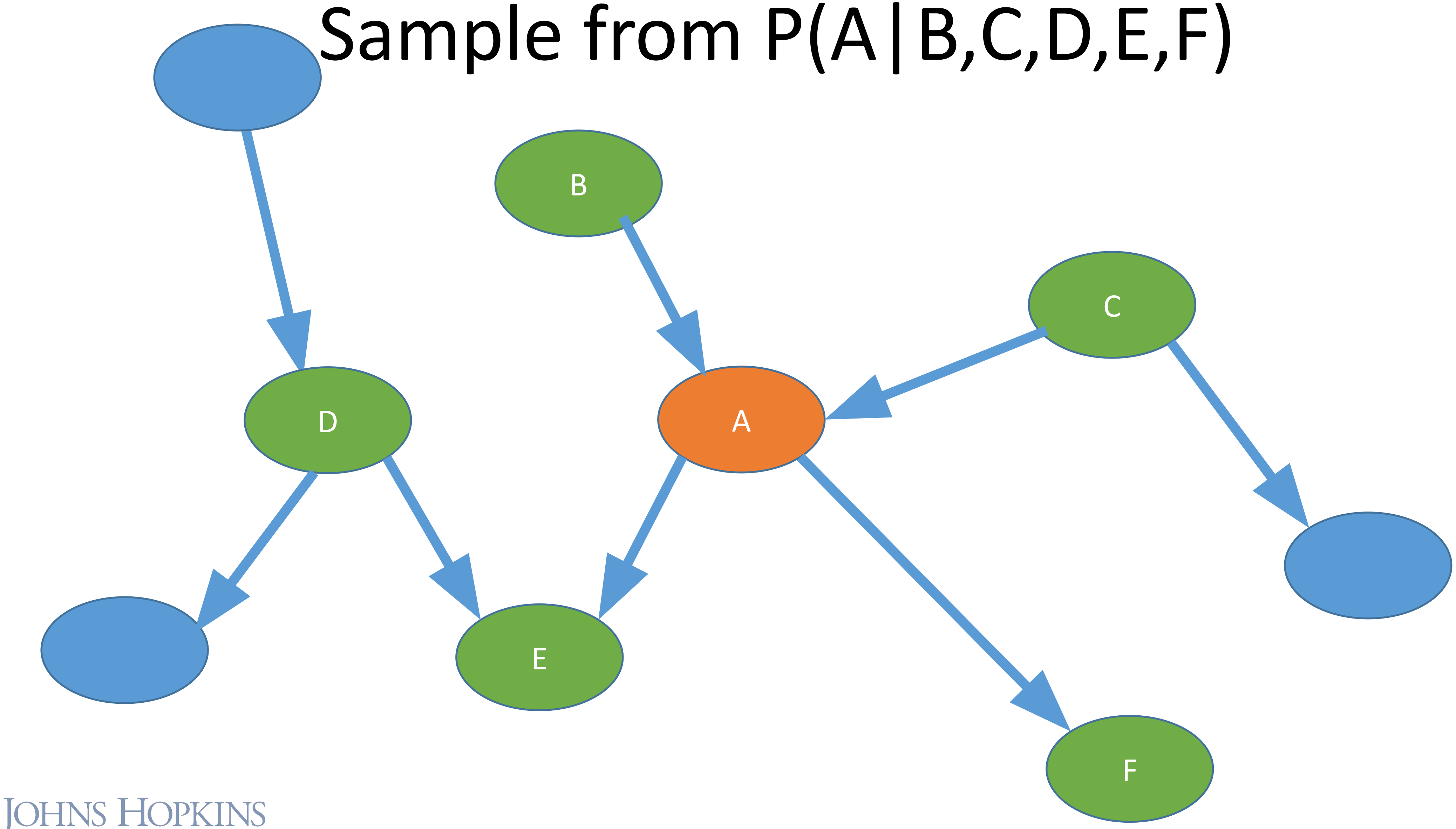
Markov Blanket children



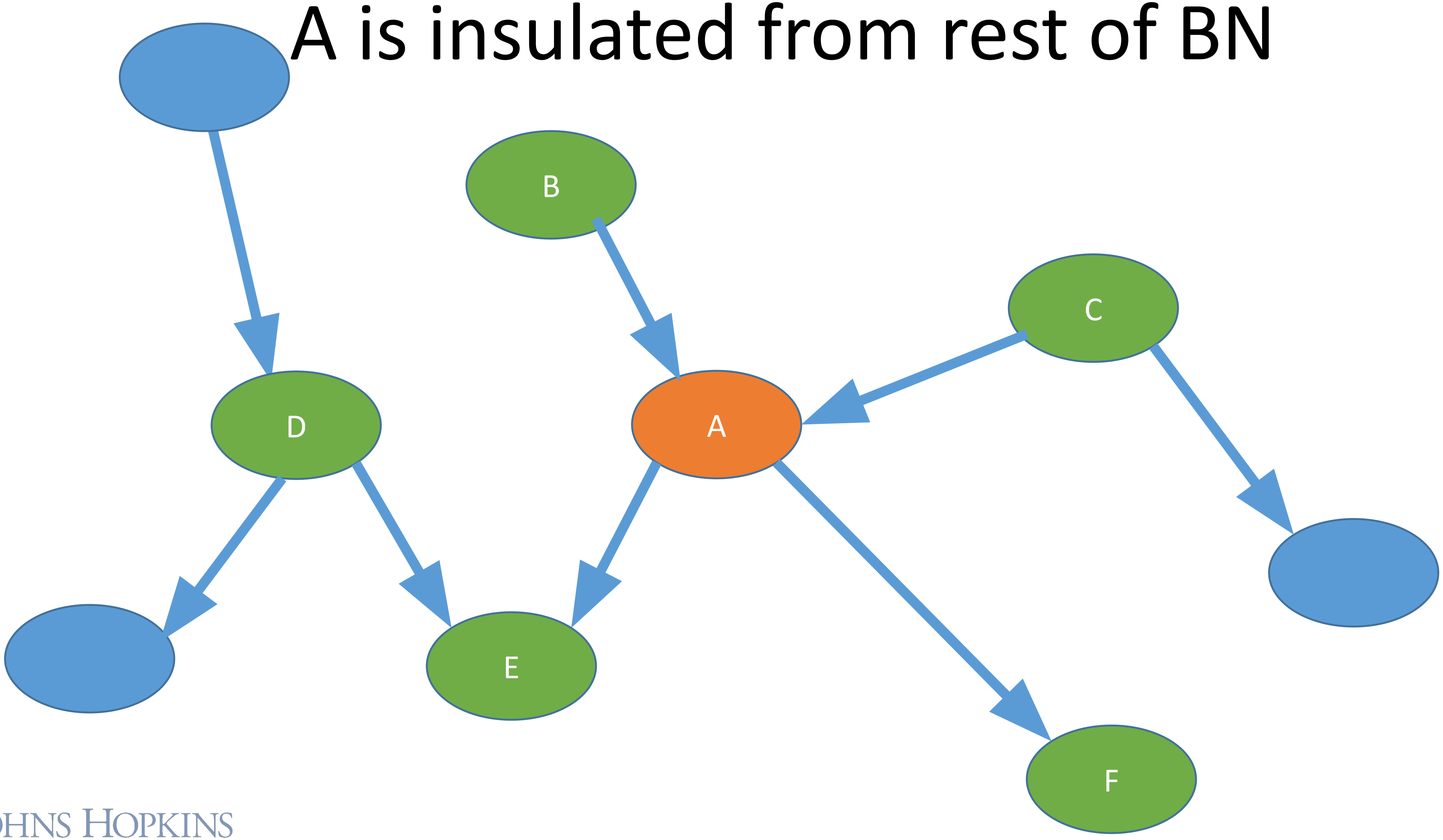
Markov Blanket
children's parents



Sample from P(A|B,C,D,E,F)



A is insulated from rest of BN



Sampling and tracking “sufficient statistics”

For LDA, need to track:

n_{dt}: Number of times topic t occurs in document d

n_{tw}: number of times word w is drawn from topic t

n_t: total occurrences of topic t

For random document d, word w, with current topic t:

Decrement $n_{dt}[d, t], n_{tw}[t, w], n_t[t]$

Choose new topic with probability proportional to true distribution

Conjugacy and “Collapsed” Sampling

Two distributions: Dirichlet and Categorical

Categorically-distributed values

Basic approach:

$$TopicWordProportions \sim \beta$$
$$DocumentTopicProportions \sim \alpha$$
$$Topic \sim DocumentTopicProportions$$
$$Word \sim TopicWordProportions[Topic]$$

Conjugacy and "Collapsed" Sampling

Ultimately though, we care about the *discrete values*

$$P(w|\beta) = \int_c P(w|c)P(c|\beta)$$

The choice of the Dirichlet distribution has a great property:

$$P(w|\beta) \propto \mathrm{n_w} + \beta_w$$

Fiddling with hyper-parameters

α and β encode *prior probabilities* for words and topics

They also encode the sparsity of the distributions

Dirichlet prior
Which items/categories are likely?

	Run	Jump	walk	talk
	.1	.1	.2	.1

	run	jump	walk	talk
	100	100	200	100

Dirichlet prior
Which items/categories are likely?

	run	jump	walk	talk
	.1	.1	.2	.1

	run	jump	walk	talk
	100	100	200	100

Draw 10k samples

	run	jump	walk	talk
Average	.2	.2	.4	.2

Dirichlet prior
How “concentrated” are they?

	Run	Jump	Walk	talk
	.1	.1	.2	.1
	.016	.003	.279	.557
	.00000000001	.591	.407	.0009
	.102	.00000000001	.891	.006

	run	jump	walk	Talk
	100	100	200	100
	.197	.187	.401	.217
	.229	.206	.356	.208
	.184	.188	.438	.190

Minka’s Fixed-Point Optimization

Iterative way to estimate α and β

Symmetric and asymmetric variants

Rerun every X iterations of Gibbs sampling

Properties of Gibbs Sampling

Ergodic implies guaranteed to converge to the global optimum!

But we don’t know when, or even how to check…

Can get an idea of how things are going, e.g. *Perplexity:*

$$2^{-\sum p(x)\log(p(x))}$$

How “surprising” is the data given the current parameters

Alternative approach: Variational Methods

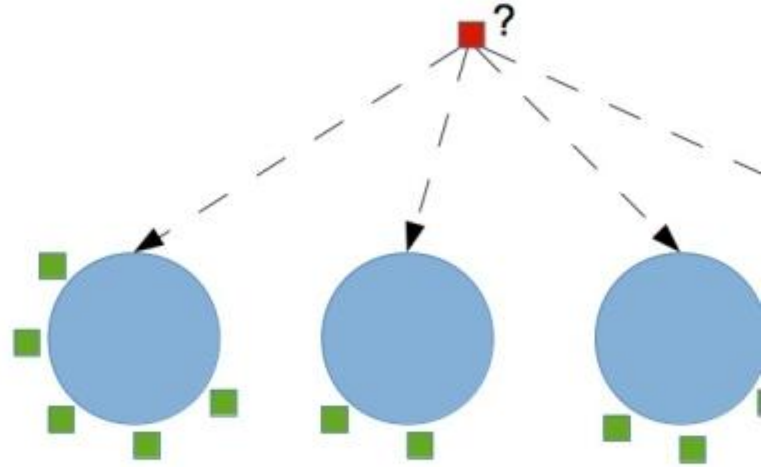
Pretend there are more independencies in the graph

Gives a simpler function that can be directly optimized

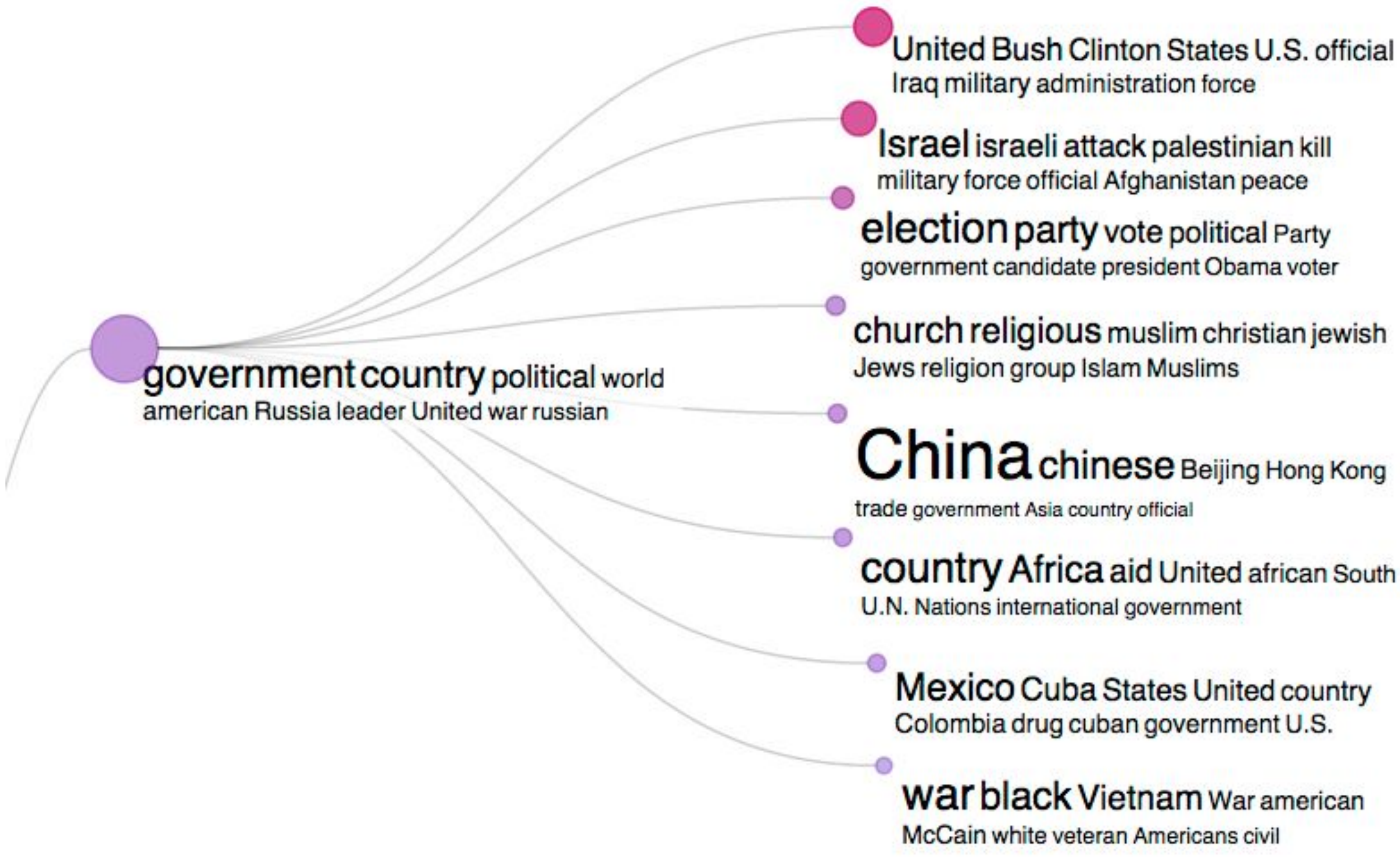
Only get a local optimum…

But we get it fast, and know when we get it!

Non-parametrics: “Chinese Restaurant Process”



Hierarchical topic models


government country political world
american Russia leader United war russian
United Bush Clinton States U.S. official
Iraq military administration force
Israel israeli attack palestinian kill
military force official Afghanistan peace
election party vote political Party
government candidate president Obama voter
church religious muslim christian jewish
Jews religion group Islam Muslims
China chinese Beijing Hong Kong
trade government Asia country official
country Africa aid United african South
U.N. Nations international government
Mexico Cuba States United country
Colombia drug cuban government U.S.
war black Vietnam War american
McCain white veteran Americans civil

Uses

Query expansion:

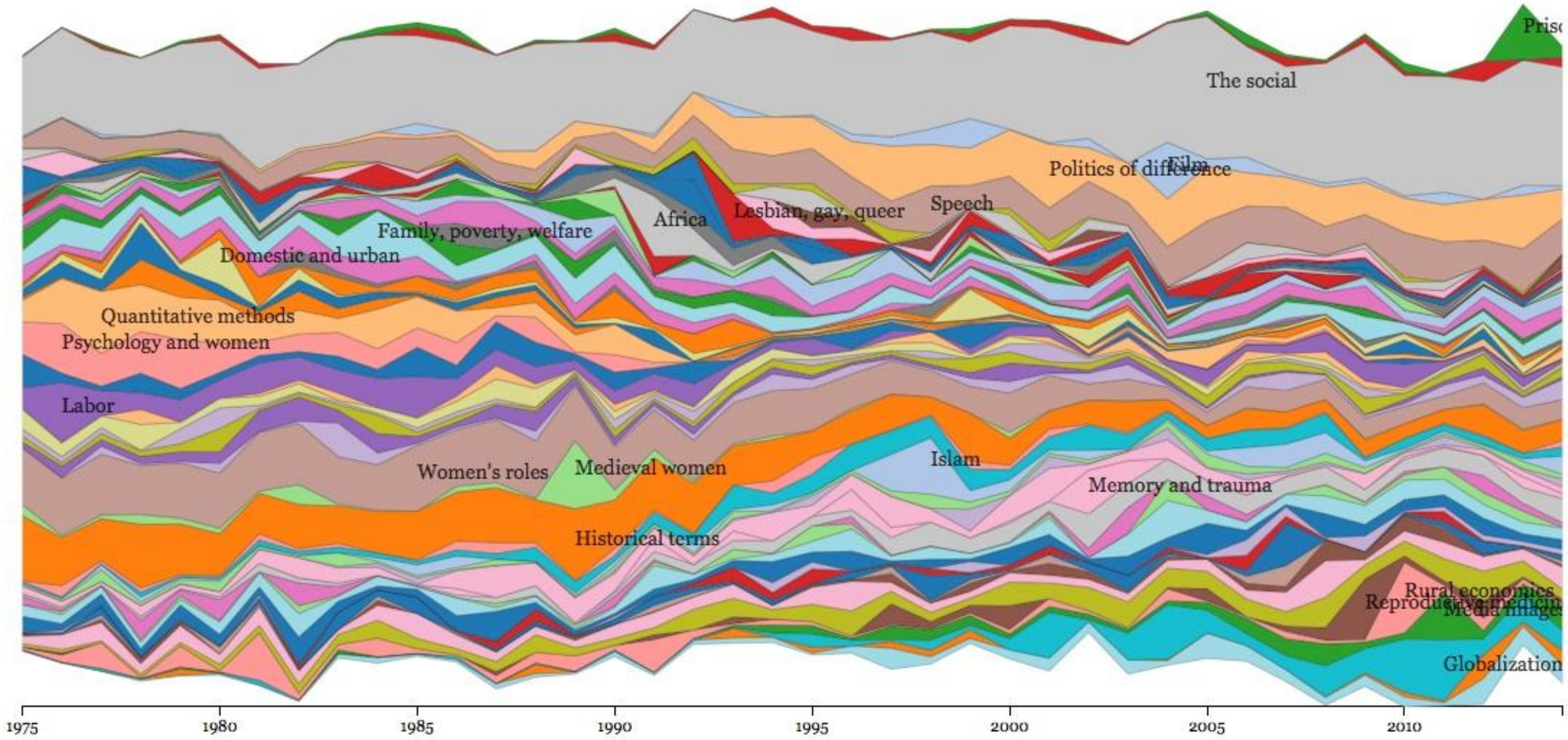
Initial query: “peaceful dog types”

Useful document: “labs are a gentle breed”

Classification features

Recommender systems etc.

Topic distributions over time and place


Prisc
The social
Politics of difference
Film
Speech
Lesbian, gay, queer
Africa
Family, poverty, welfare
Domestic and urban
Quantitative methods
Psychology and women
Labor
Women's roles
Medieval women
Islam
Memory and trauma
Historical terms
Rural economics
Reproductive medicin
Media images
Globalization