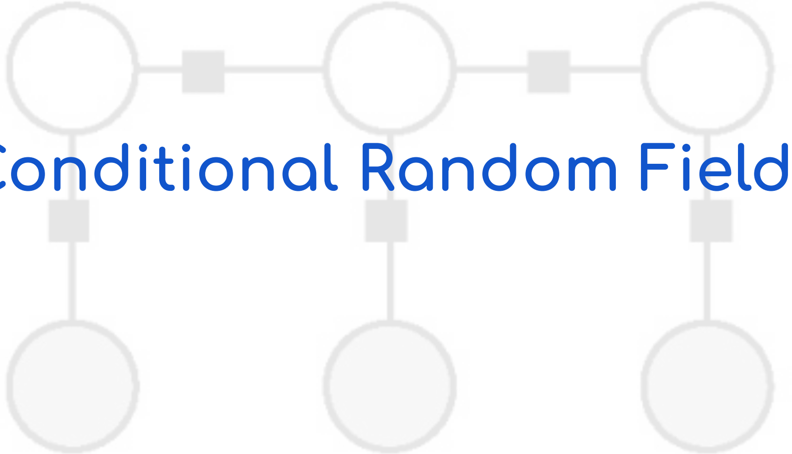


Conditional Random Fields

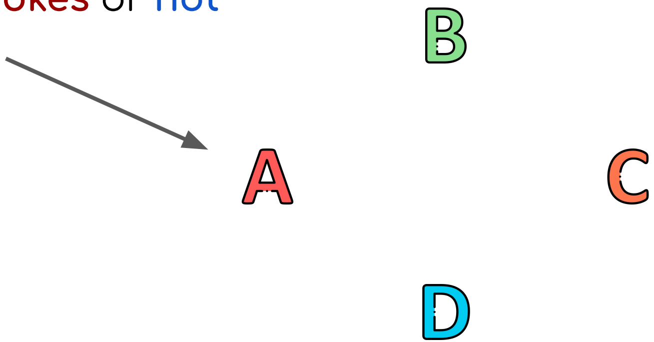
Taki Eddine MEKHALFA
Taki-Eddine.MEKHALFA@inria.fr

B

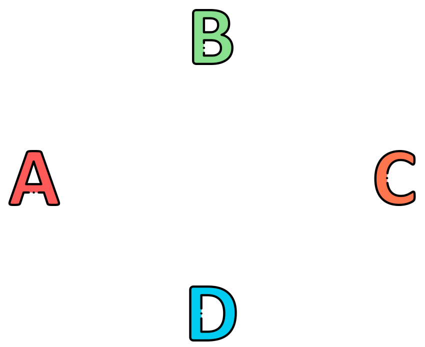
A C

D

Random variable: A smokes or not



Suppose you know that A and B are friends! **What does this tell you?**



Factor Graphs and Markov property

Suppose you know that A and B are friends! What does this tell you?

=> They might have similar smoking habits!

B

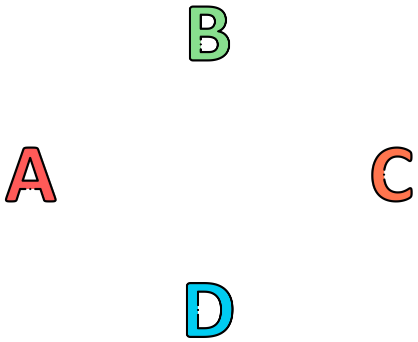
A C

D

Factor Graphs and Markov property

Suppose you know that A and B are friends! What does this tell you?

=> They might have similar smoking habits! Is this useful to model smoking behaviors?



Suppose you know that A and B are friends! What does this tell you?

=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

B

A C

D

Factor Graphs and Markov property

Suppose you know that A and B are friends! What does this tell you?

=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

B

A C

D

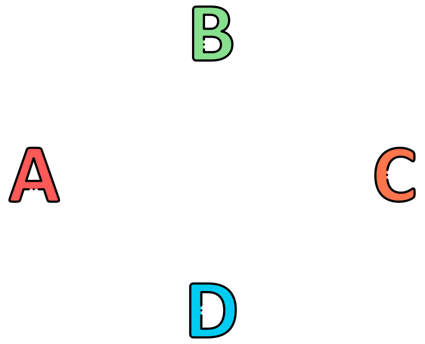
Suppose you know that A and B are friends! What does this tell you?

=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

=> Use Graphs!



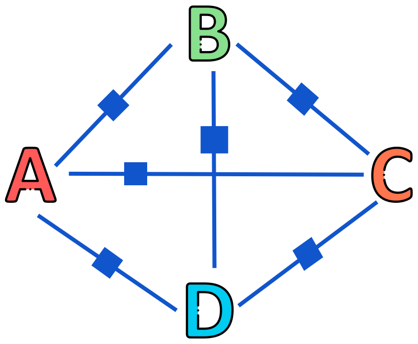
Suppose you know that A and B are friends! What does this tell you?

=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

=> Use Graphs!



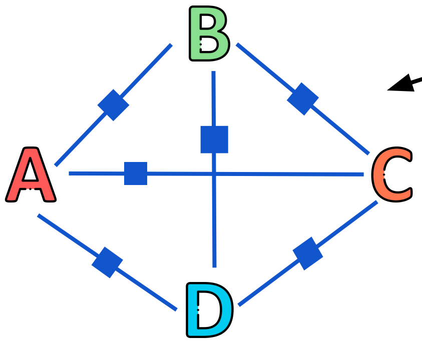
Suppose you know that A and B are friends! What does this tell you?

=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

=> Use Graphs!



What's the problem?

Factor Graphs and Markov property

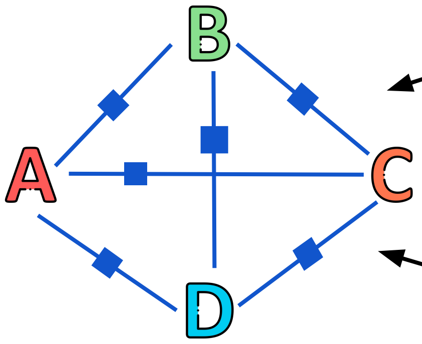
Suppose you know that A and B are friends! What does this tell you?

=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

=> Use Graphs!



What's the problem?

Every variable depends on every other variable!

Suppose you know that A and B are friends! What does this tell you?

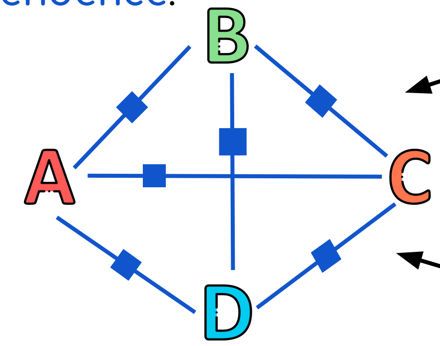
=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

=> Use Graphs!

=> Need a way to represent independence!



What's the problem?

Every variable depends on every other variable!

Suppose you know that A and B are friends! What does this tell you?

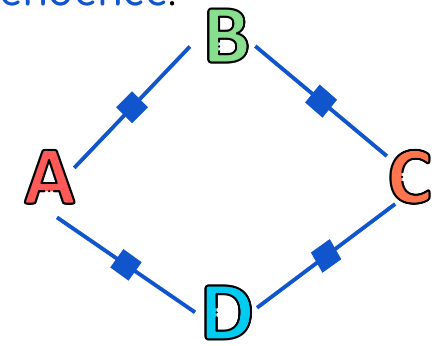
=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

=> Use Graphs!

=> Need a way to represent independence!



Suppose you know that A and B are friends! What does this tell you?

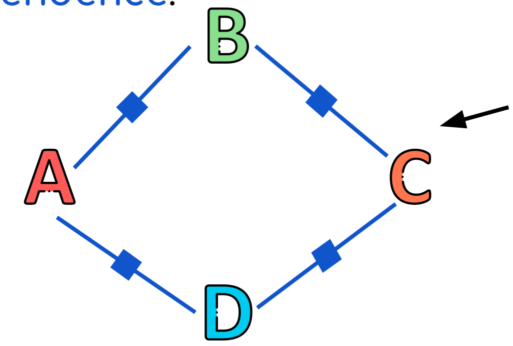
=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

=> Use Graphs!

=> Need a way to represent independence!



Does this mean B and D that are totally independent?

Factor Graphs and Markov property

Suppose you know that A and B are friends! What does this tell you?

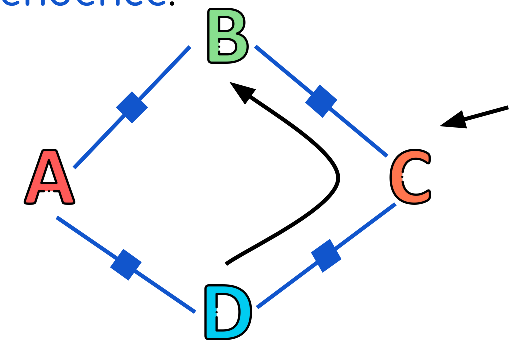
=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

=> Use Graphs!

=> Need a way to represent independence!



Does this mean B and D that are totally independent?

Suppose you know that A and B are friends! What does this tell you?

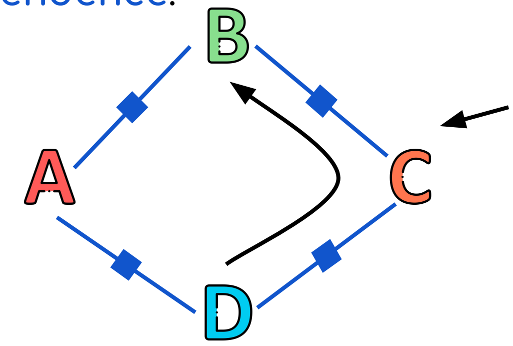
=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

=> Use Graphs!

=> Need a way to represent independence!



Does this mean B and D that are totally independent?

B Is independent from D only when we know A and C

Factor Graphs and Markov property

Suppose you know that A and B are friends! What does this tell you?

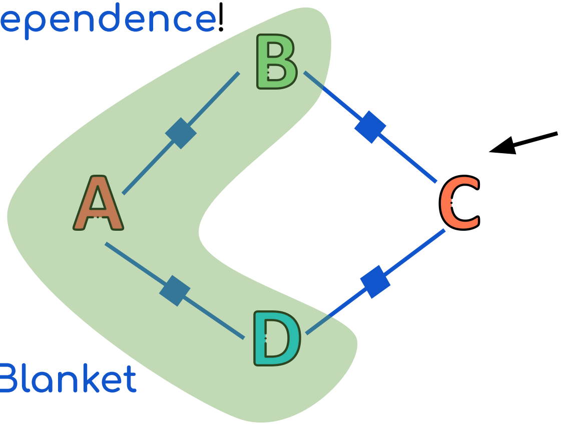
=> They might have similar smoking habits! Is this useful to model smoking behaviors?

=> Call it a feature then!

=> Need a way to represent interactions!

=> Use Graphs!

=> Need a way to represent independence!

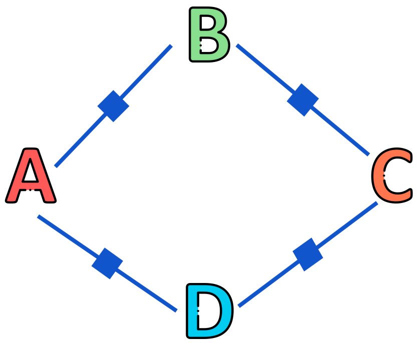


Called the Markovian Blanket

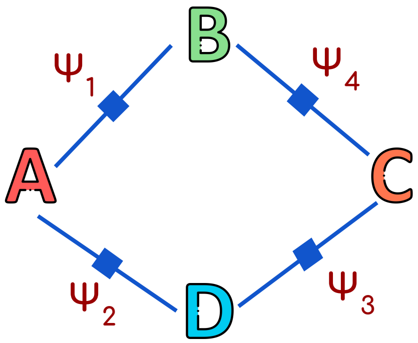
Does this mean B and D that are totally independent?

B Is independent from D only when we know A and C

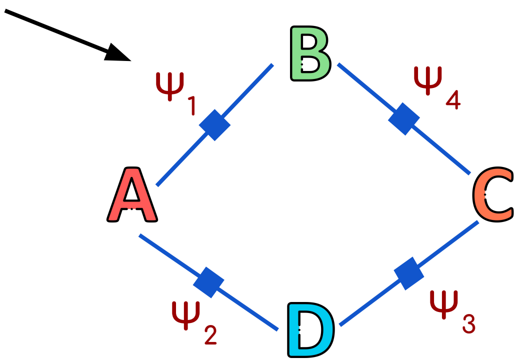
Factor Graphs and Compatibility



Factor Graphs and Compatibility

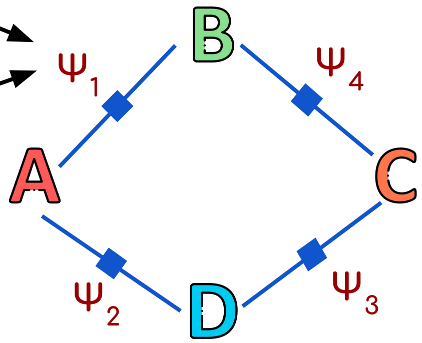


Call it a FACTOR



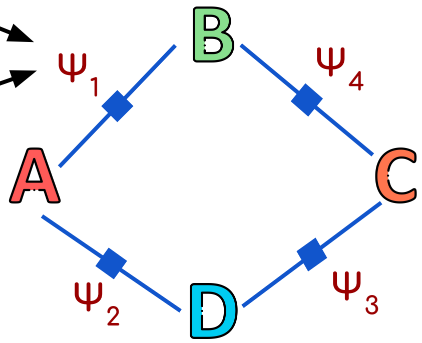
Call it a FACTOR

Compatibility between A and B



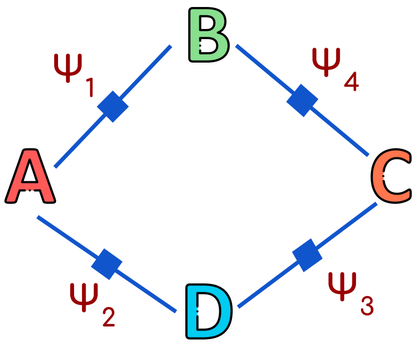
Call it a FACTOR

Compatibility between A and B



ASSUMPTION: A given state of variables (we say a possible world) is more likely iff the overall compatibility is higher

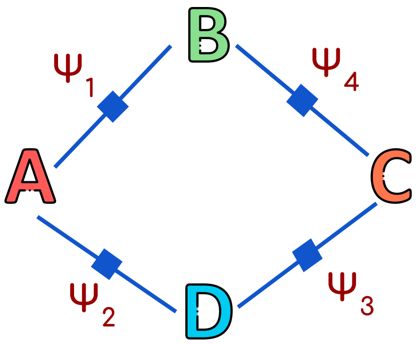
How to define compatibility in our case?



How to define compatibility in our case?

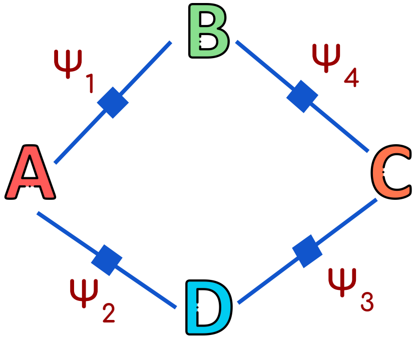
$\psi_1(A, B) =$

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



ψ_1(A , B) =

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



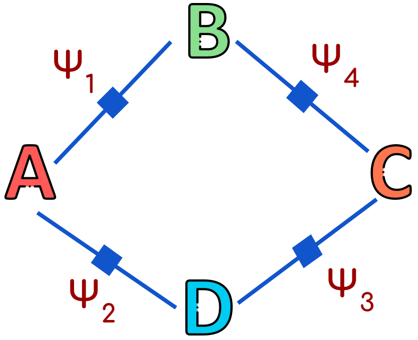
Suppose that:

A Is friends with B. C is friends with D

What world is most likely between the following two?:

ψ_1(A , B) =

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



Suppose that:

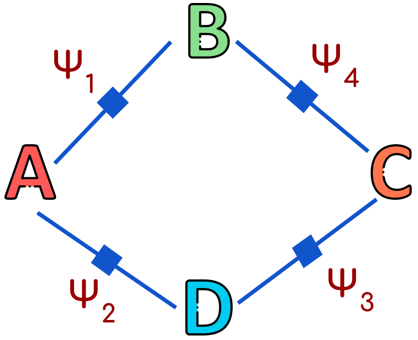
A Is friends with B. C is friends with D

What world is most likely between the following two?:

1. Smokes{A , B } and Does not Smoke {C , D }
2. Smokes{ A , C } and Does not Smoke {B , D }

ψ_1(A , B) =

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



Suspose that:

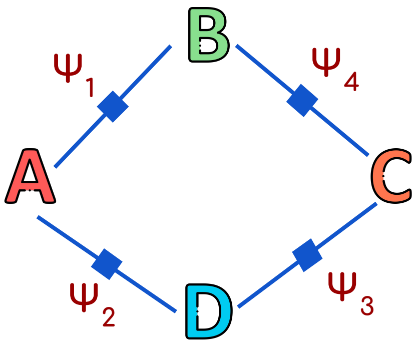
A Is friends with B. C is friends with D

What world is most likely between the following two?:

1. Smokes{A , B } and Does not Smoke {C , D }
2. Smokes{ A , C } and Does not Smoke {B , D }

ψ_1(A , B) =

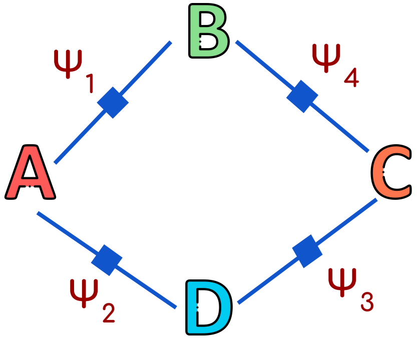
- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



P(Smokes{ A, B} and Does not Smoke { C, D})

Ψ_1(A , B) =

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits

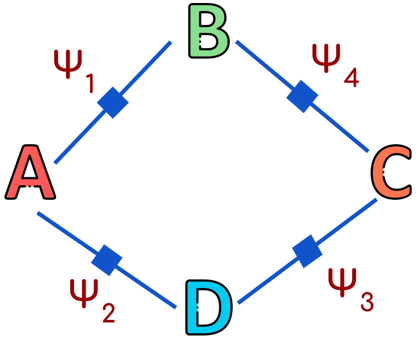


P(Smokes{ A, B} and Does not Smoke { C, D})

$\propto\Psi$(Smokes{A , B } and Does not Smoke {C , D })

$\Psi_1(A, B) =$

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



P(Smokes{ A, B} and Does not Smoke { C, D})

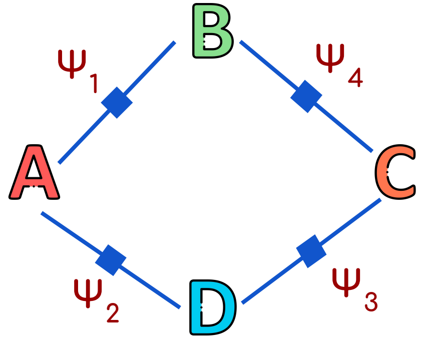
$\propto \Psi$(Smokes{A, B } and Does not Smoke {C, D })

$= \exp(\Psi_1(A, B)) \times \exp(\Psi_2(A, D))\ \exp(\Psi_3(B, C))\ \exp(\Psi_4(C, D))$

=?

$\Psi_1(A, B) =$

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



P(Smokes{ A, B} and Does not Smoke { C, D})

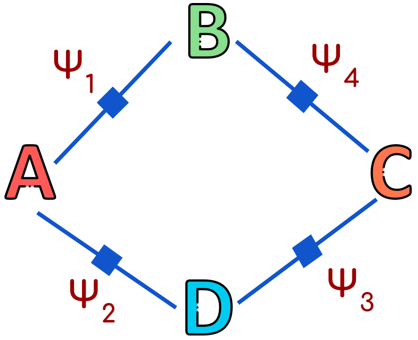
$\propto \Psi$(Smokes{A , B } and Does not Smoke {C , D })

$= \exp(\Psi_1(A, B)) \times \exp(\Psi_2(A, D))\ \exp(\Psi_3(B, C))\ \exp(\Psi_4(C, D))$

$= \exp(1) \times \exp(0.5) \times \exp(0.5) \times \exp(1) = \exp(3)$

Ψ_1(A , B) =

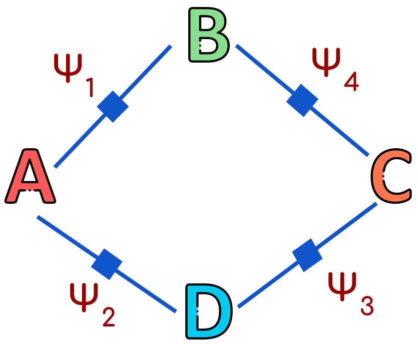
- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



P(Smokes{ A, C } and Does not Smoke {B , D})

Ψ_1(A , B) =

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



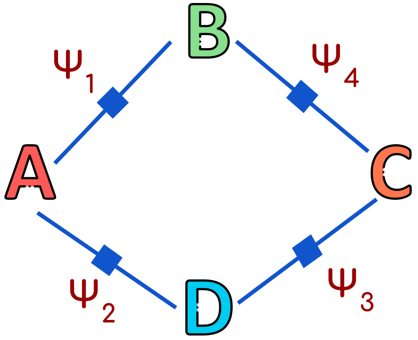
P(Smokes{ A, C } and Does not Smoke {B , D})

$\propto \Psi$(Smokes{A ,C } and Does not Smoke {B , D })

Factor Graphs and Compatibility

$\psi_1(A, B) =$

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



P(Smokes{ A, C } and Does not Smoke {B, D})

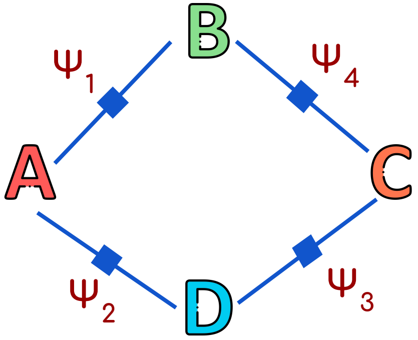
$\propto \psi$(Smokes{A, C} and Does not Smoke {B, D})

$= \exp(\psi_1(A, B)) \times \exp(\psi_2(A, D))\ \exp(\psi_3(B, C))\ \exp(\psi_4(C, D))$

= ?

Ψ_1(A , B) =

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



P(Smokes{ A, C } and Does not Smoke {B , D})

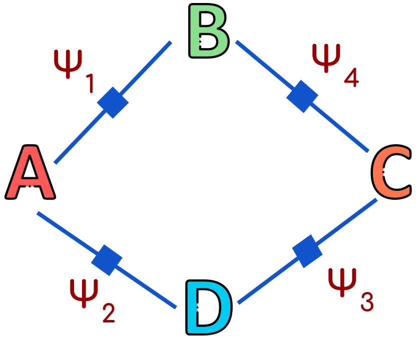
$\propto$ Ψ(Smokes{A , C } and Does not Smoke {B , D })

= exp(Ψ_1(A, B)) x exp(Ψ_2(A , D)) exp(Ψ_3(B , C)) exp(Ψ_4(C , D))

= exp(0) x exp(0.5) x exp(0.5) x exp(0) = exp(1)

ψ_1(A , B) =

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits



P(Smokes{ A , B } and Does not Smoke { C , D })

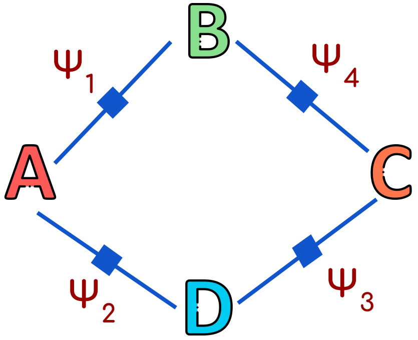
IS ~7.4 times more likely (exp(2) = exp(3) / exp(1)) than

P(Smokes{ A, C } and Does not Smoke { B , D})

Factor Graphs and Compatibility

$\Psi_{1,\,friendship}$(A , B) =

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits

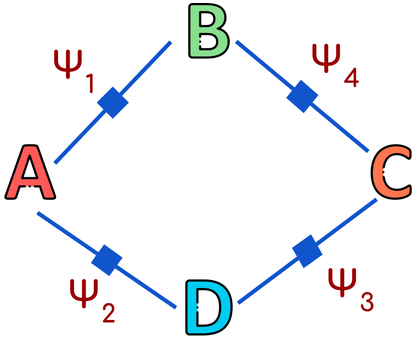


$\Psi_{1,\ \text{friendship}}$(A ,B) =

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits

$\Psi_{1,\ \text{couple}}$ (A ,B) =

- 5 if A and B are couples and both smoke
- 0.5 if A and B are not couples
- 0 if A and B are couples and don't have same smoking habits

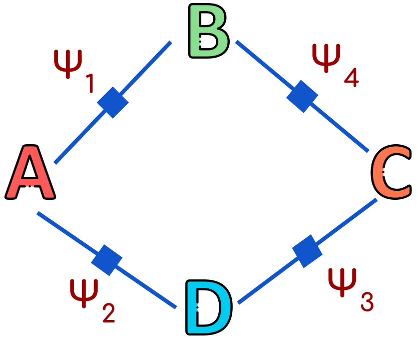


$\Psi_{1,\ \text{friendship}}(A, B) =$

- 1 if A and B are friends and both smoke or both don't smoke
- 0.5 if A and B are are not friends
- 0 if A and B are friends and don't have same smoking habits

$\Psi_{1,\ \text{couple}}(A, B) =$

- 5 if A and B are couples and both smoke
- 0.5 if A and B are not couples
- 0 if A and B are couples and don't have same smoking habits



$\Psi_1(A, B) = \Psi_{1,\ \text{friendship}}(A, B) + \Psi_{1,\ \text{couple}}(A, B)$

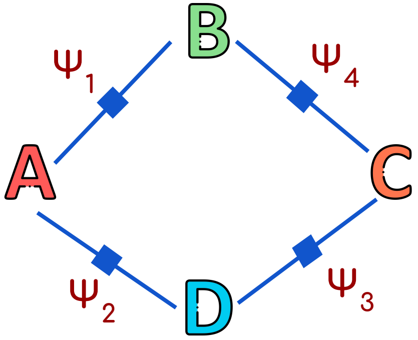
Factor Graphs and Compatibility

$\Psi_{1,\ \mathrm{friendship}}$(A , B) =

- 1 if A and B are friends and **both smoke or both don't smoke**
- 0.5 if A and B are **are not friends**
- 0 if A and B are friends and **don't have same smoking habits**

$\Psi_{1,\ \mathrm{couple}}$ (A , B) =

- 5 if A and B are couples and both smoke
- 0.5 if A and B are not couples
- 0 if A and B are couples and **don't have same smoking habits**



Ψ_1(A , B) = $\Psi_{1,\ \mathrm{friendship}}$(A , B) + $\Psi_{1,\ \mathrm{couple}}$(A , B)

Learning by maximizing the likelihood

Suppose I show you a biased coin:  with parameter θ where P(Head) = θ

I flip the coin 100 times in front of you, finally we get 30 heads and 70 tails.

Learning by maximizing the likelihood

Suppose I show you a biased coin: 

with parameter θ where P(Head) = θ

I flip the coin 100 times in front of you, finally we get 30 heads and 70 tails.

Intuitively, what is an estimate for θ?

Suppose I show you a biased coin:  with parameter θ where P(Head) = θ

I flip the coin 100 times in front of you, finally we get 30 heads and 70 tails.

Intuitively, what is an estimate for θ?

=> θ = 30 / 100 = 0.3

Learning by maximizing the likelihood

Suppose I show you a biased coin:  with parameter θ where P(Head) = θ

I flip the coin 100 times in front of you, finally we get 30 heads and 70 tails.

Intuitively, what is an estimate for θ?

=> θ = 30 / 100 = 0.3

P(data; θ) = P(30 heads and 70 tails; θ) = P(30 heads; θ) x P(70 tails; θ)

Learning by maximizing the likelihood

Suppose I show you a biased coin: 

with parameter θ where P(Head) = θ

I flip the coin 100 times in front of you, finally we get 30 heads and 70 tails.

Intuitively, what is an estimate for θ?

=> $\theta = 30 / 100 = 0.3$

$P(\text{data}; \theta) = P(30 \text{ heads and } 70 \text{ tails}; \theta) = P(30 \text{ heads}; \theta) \times P(70 \text{ tails}; \theta)$

$= \theta^{30}(1 - \theta)^{70}$

Learning by maximizing the likelihood

Suppose I show you a biased coin:  with parameter θ where P(Head) = θ

I flip the coin 100 times in front of you, finally we get 30 heads and 70 tails.

Intuitively, what is an estimate for θ?

=> $\theta = 30 / 100 = 0.3$

$P(\text{data}; \theta) = P(30 \text{ heads and } 70 \text{ tails}; \theta) = P(30 \text{ heads}; \theta) \times P(70 \text{ tails}; \theta)$

$= \theta^{30}(1 - \theta)^{70}$

Maximizing $\theta^{30}(1 - \theta)^{70}$ is equivalent to maximizing $\log(\theta^{30}(1 - \theta)^{70}) =$

Learning by maximizing the likelihood

Suppose I show you a biased coin: 

with parameter θ where P(Head) = θ

I flip the coin 100 times in front of you, finally we get 30 heads and 70 tails.

Intuitively, what is an estimate for θ?

=> $\theta = 30 / 100 = 0.3$

$P(\text{data}; \theta) = P(30 \text{ heads and } 70 \text{ tails}; \theta) = P(30 \text{ heads}; \theta) \times P(70 \text{ tails}; \theta)$

$= \theta^{30}(1 - \theta)^{70}$

Maximizing $\theta^{30}(1 - \theta)^{70}$ is equivalent to maximizing $\log(\theta^{30}(1 - \theta)^{70}) = 30\log(\theta) + 70\log(1-\theta)$

Learning by maximizing the likelihood

Suppose I show you a biased coin:  with parameter θ where P(Head) = θ

I flip the coin 100 times in front of you, finally we get 30 heads and 70 tails.

Intuitively, what is an estimate for θ?

=> $\theta = 30 / 100 = 0.3$

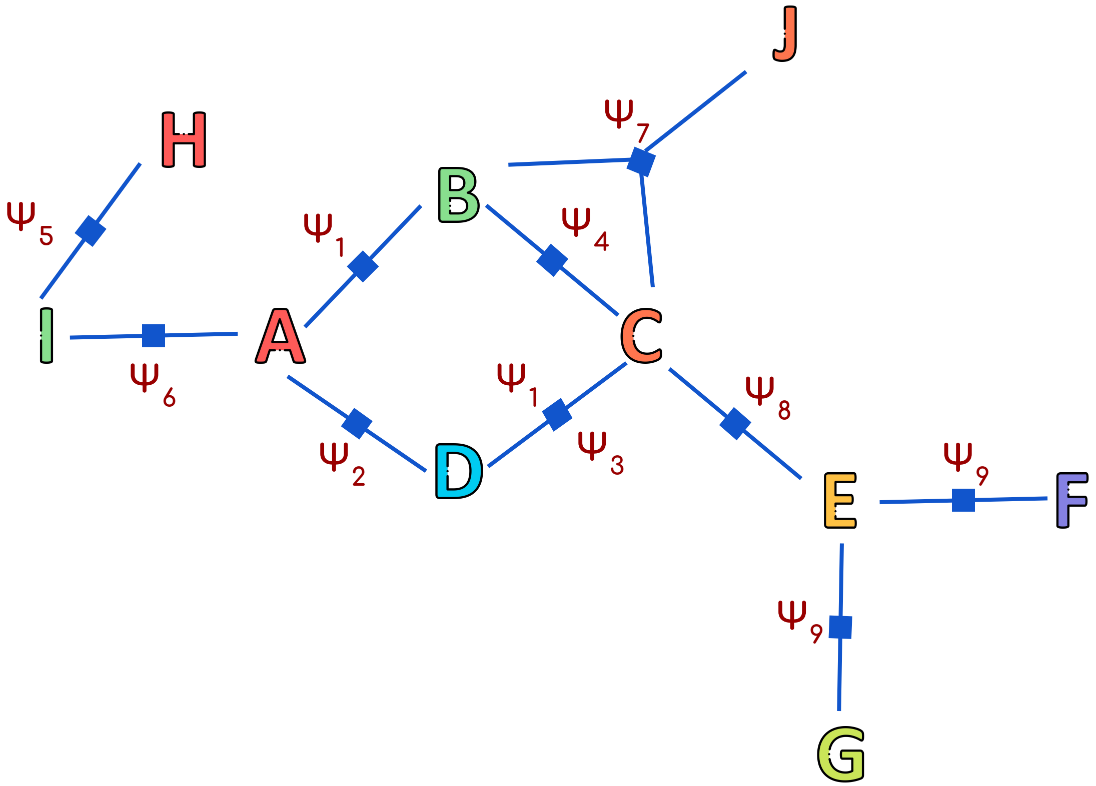
P(data; θ) = P(30 heads and 70 tails; θ) = P(30 heads; θ) x P(70 tails; θ)

$= \theta^{30}(1 - \theta)^{70}$

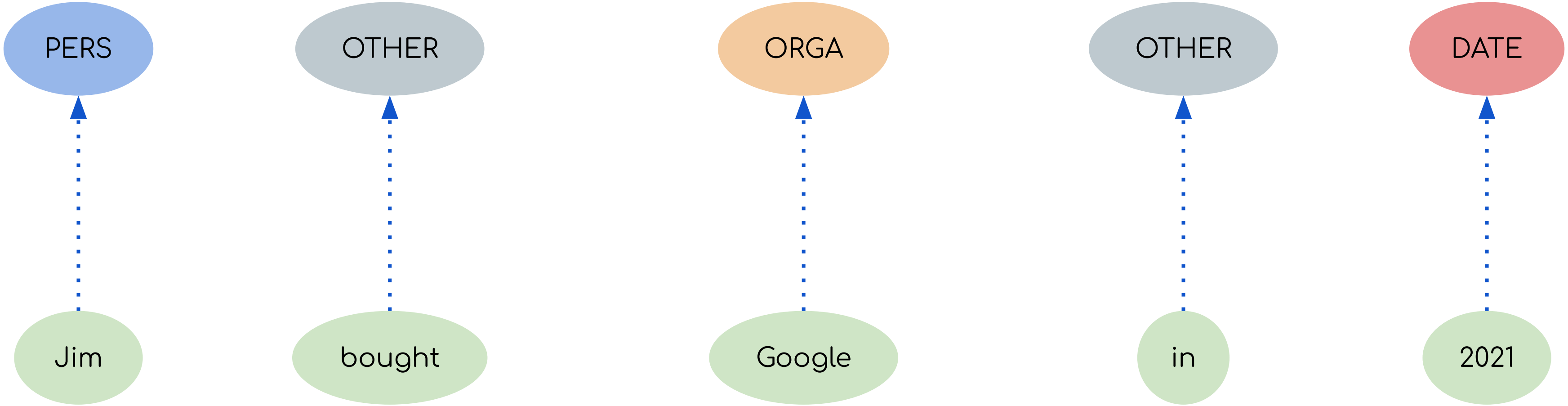
Maximizing $\theta^{30}(1 - \theta)^{70}$ is equivalent to maximizing $\log(\theta^{30}(1 - \theta)^{70}) = 30\log(\theta) + 70\log(1-\theta)$

Derivative => $30/\theta - 70/(1-\theta)$ => it hits zero when $\theta = 30/100$

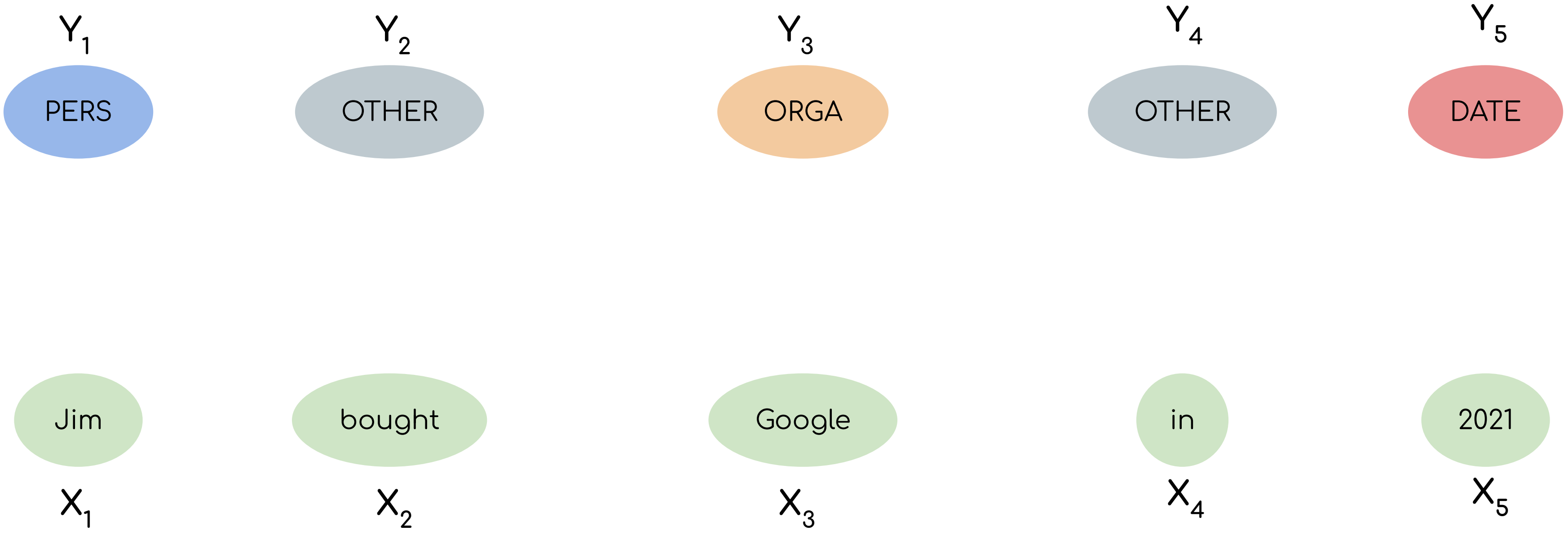
Learning by maximizing the likelihood



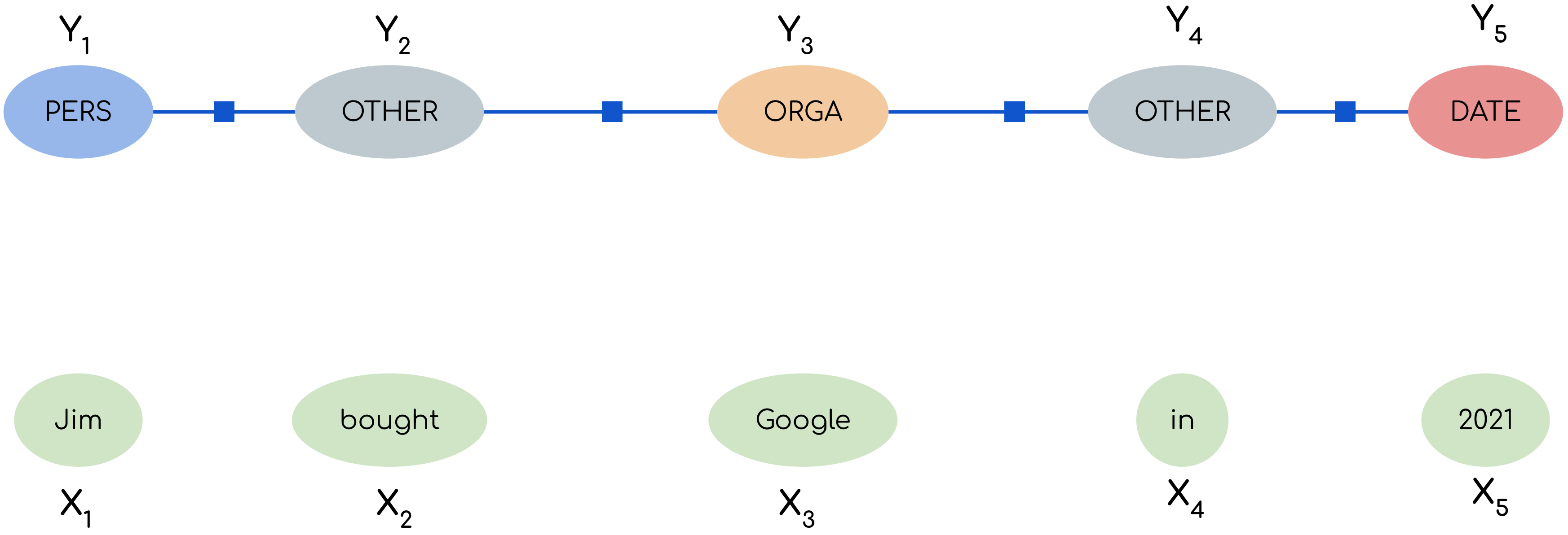
Conditional Random Fields in NER



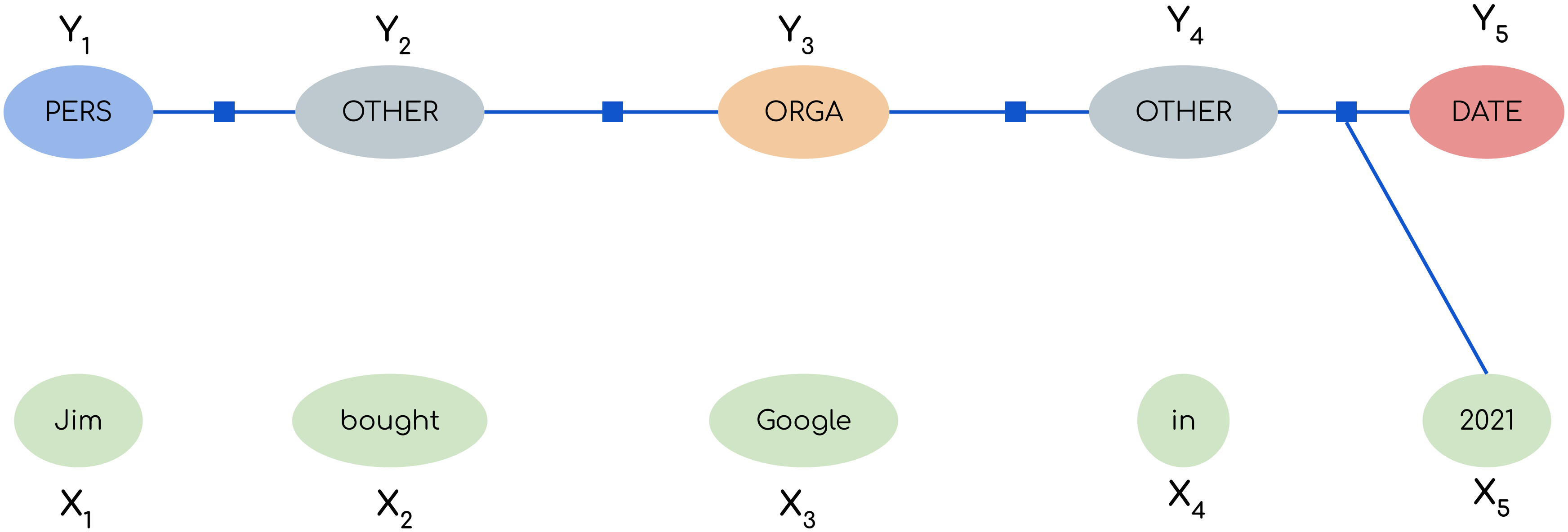
Conditional Random Fields in NER



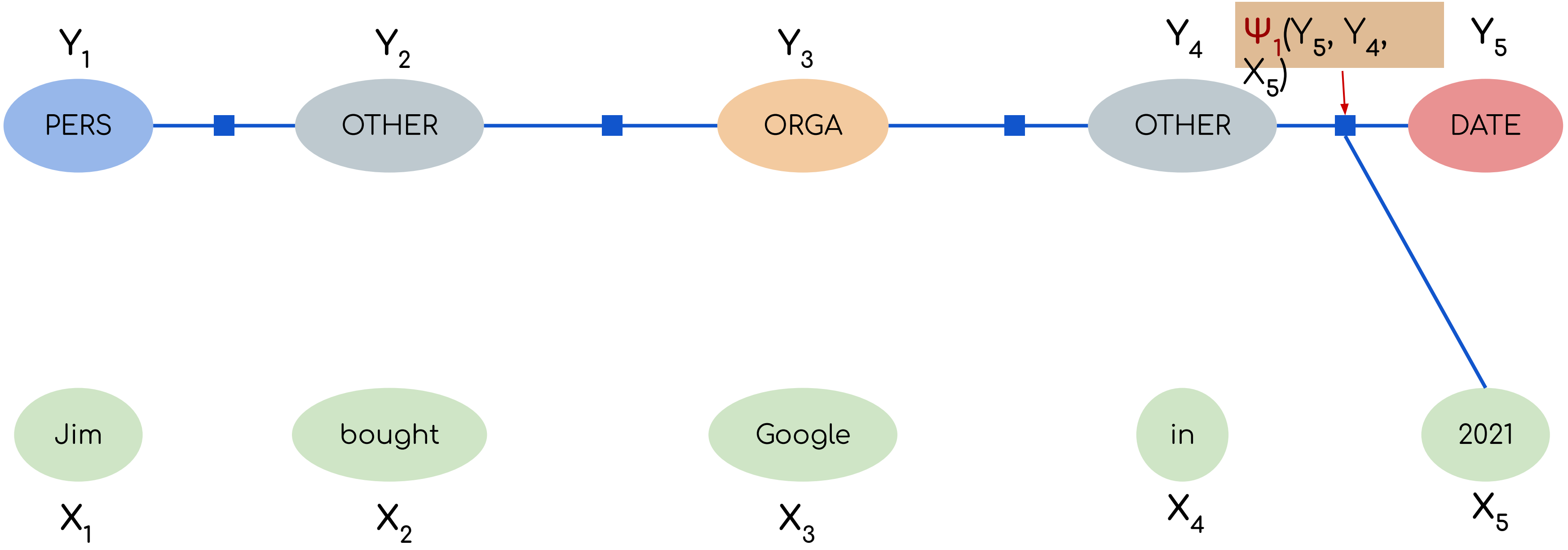
Conditional Random Fields in NER



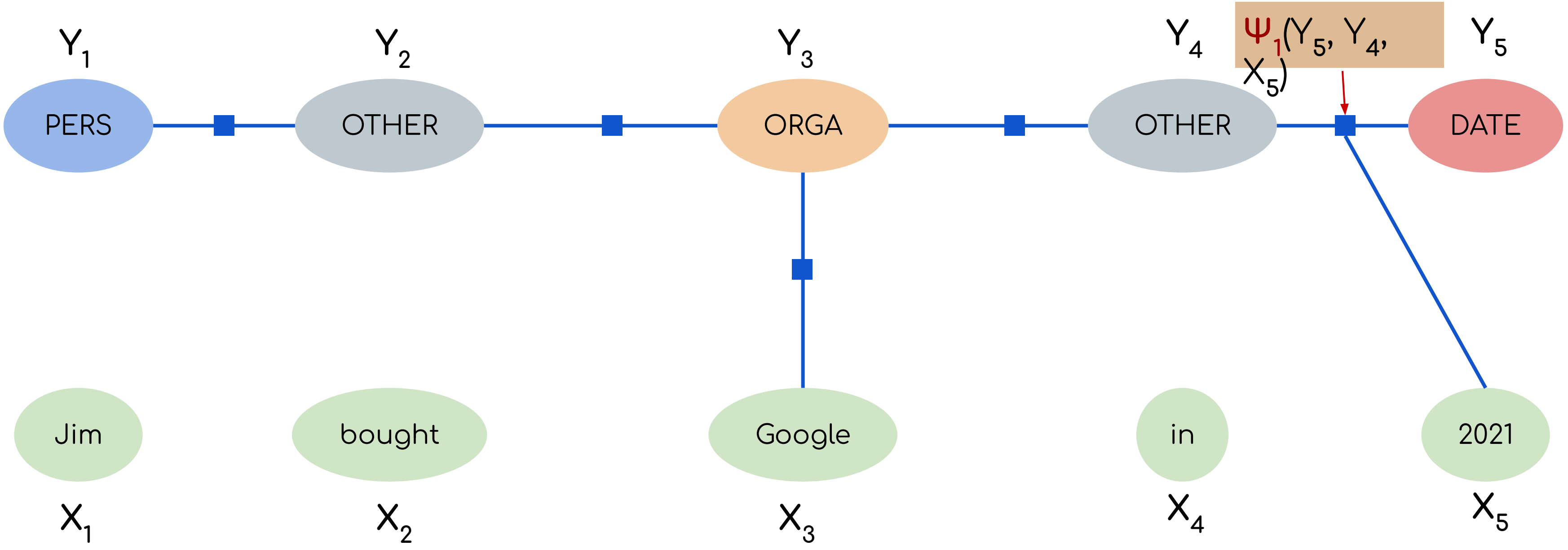
Conditional Random Fields in NER



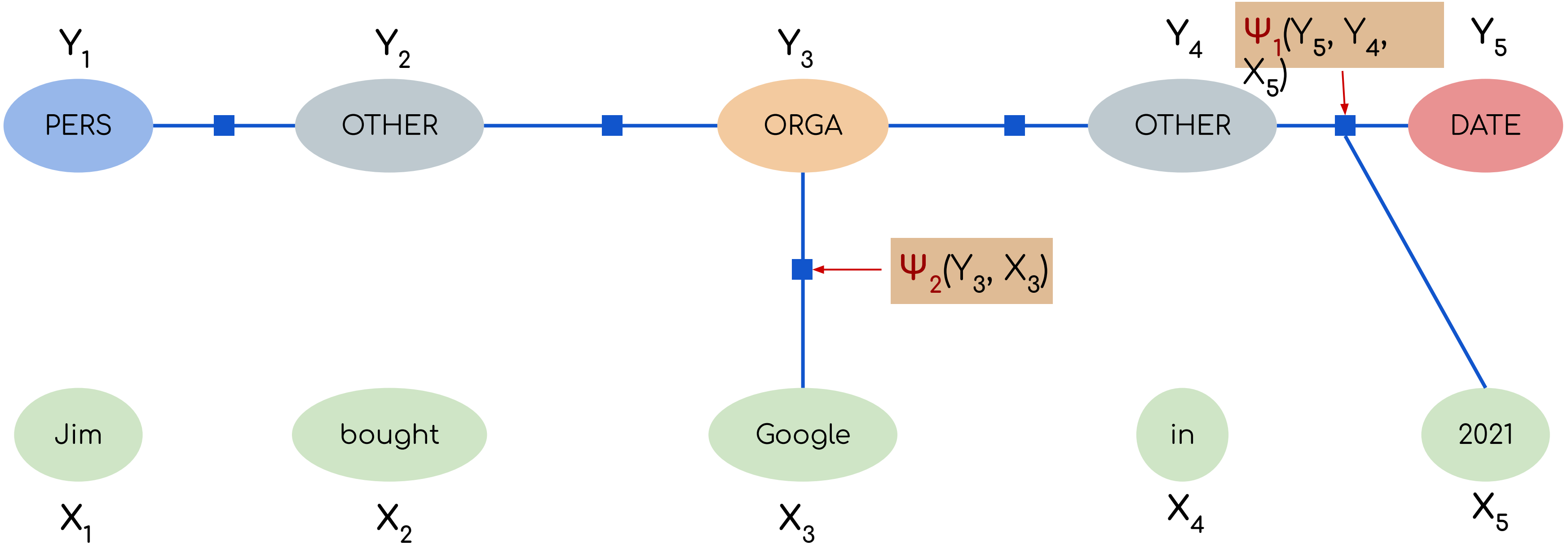
Conditional Random Fields in NER



$\Psi_1(Y_5, Y_4, X_5) = \theta_1$ if Y_5 = 'DATE' and Y_4 = 'OTHER' and X_5 is a number else 0.

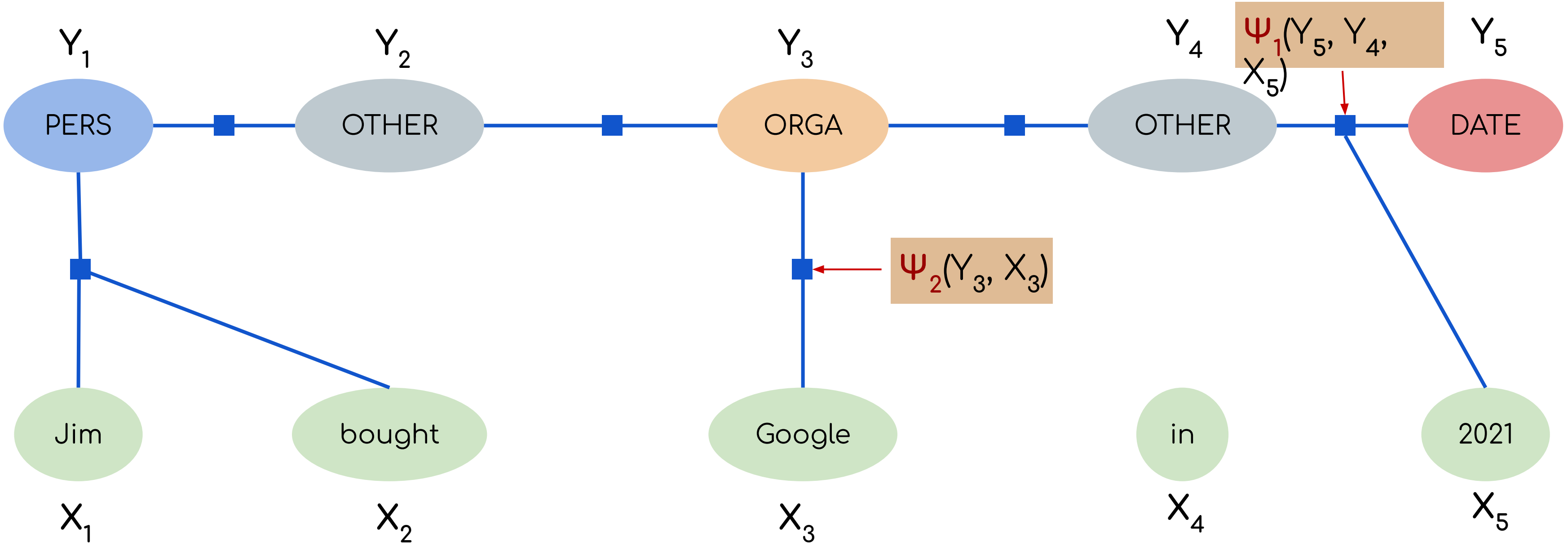

$\Psi_1(Y_5, Y_4, X_5)$
Y_1
PERS
Y_2
OTHER
Y_3
ORGA
Y_4
OTHER
Y_5
DATE
Jim
X_1
bought
X_2
Google
X_3
in
X_4
2021
X_5

Conditional Random Fields in NER

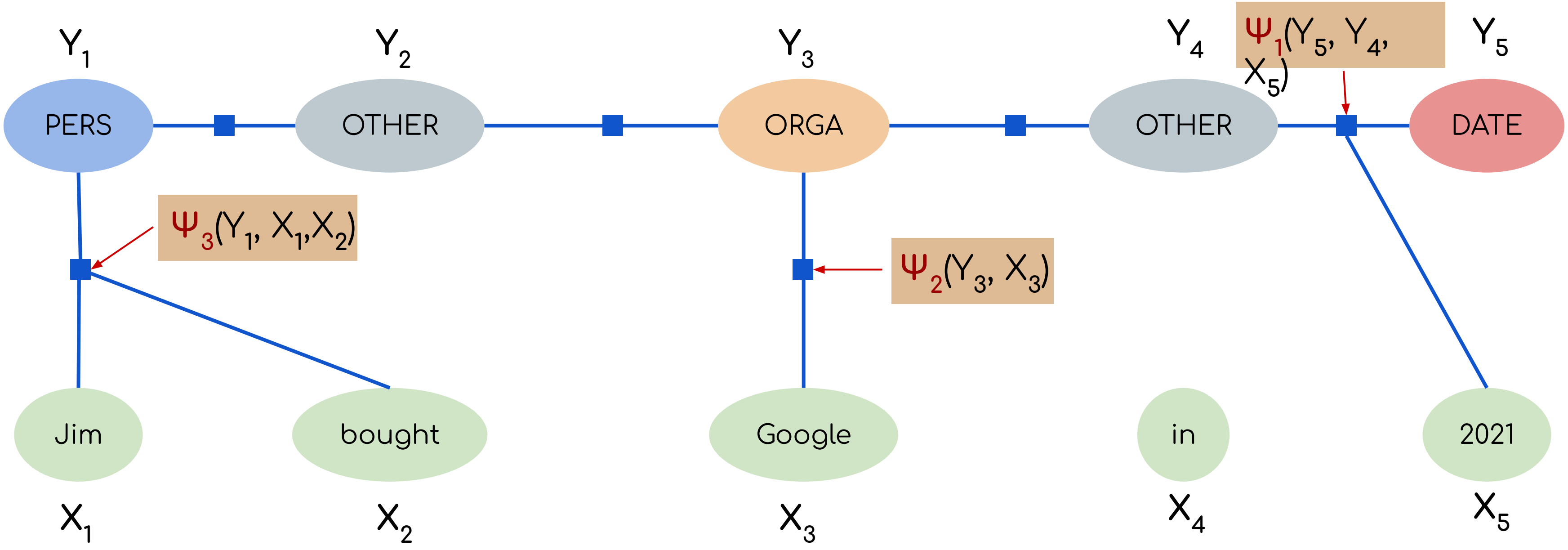


$\Psi_2(Y_3, X_3) = \theta_2$ if Y_3 = 'ORGA' and X_3 starts with a capital letter, else **0**.

Conditional Random Fields in NER

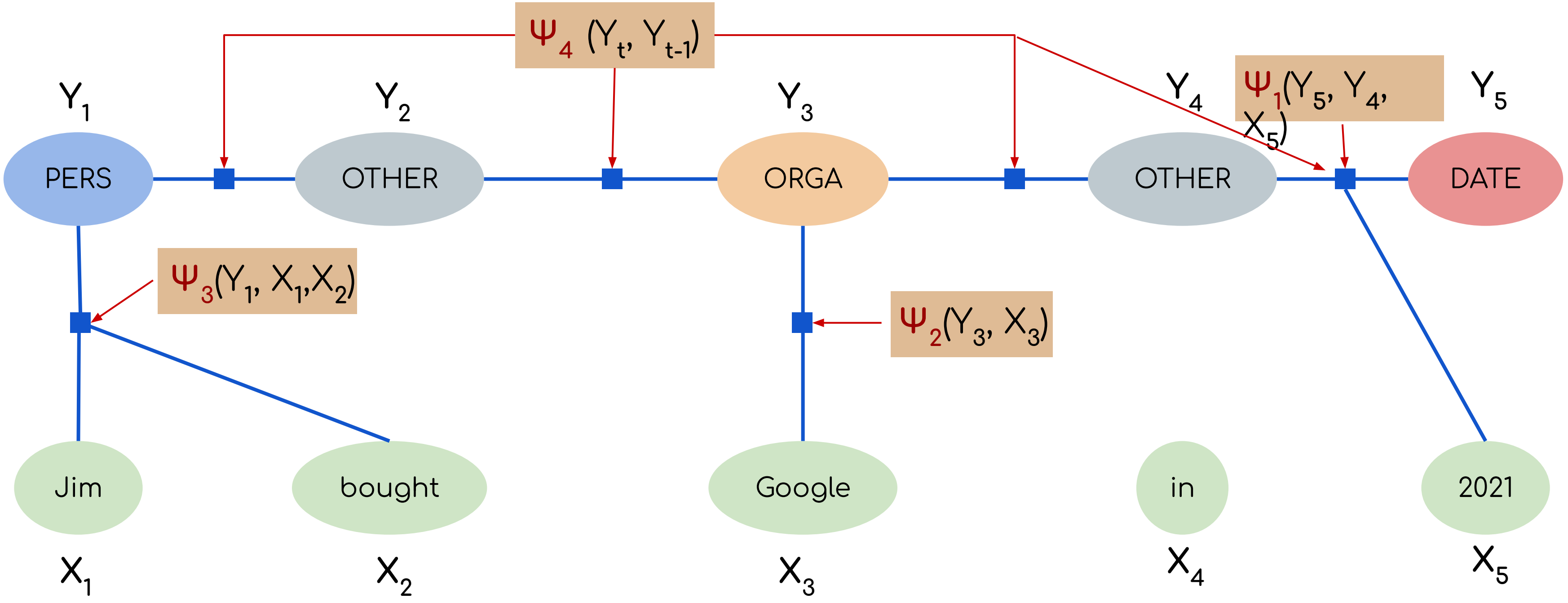


Conditional Random Fields in NER



$\Psi_3(Y_1, X_1, X_2) = \theta_3$ if Y_1 = 'PERS' and X_1 belongs to a dictionary of names and X_2 is a verb

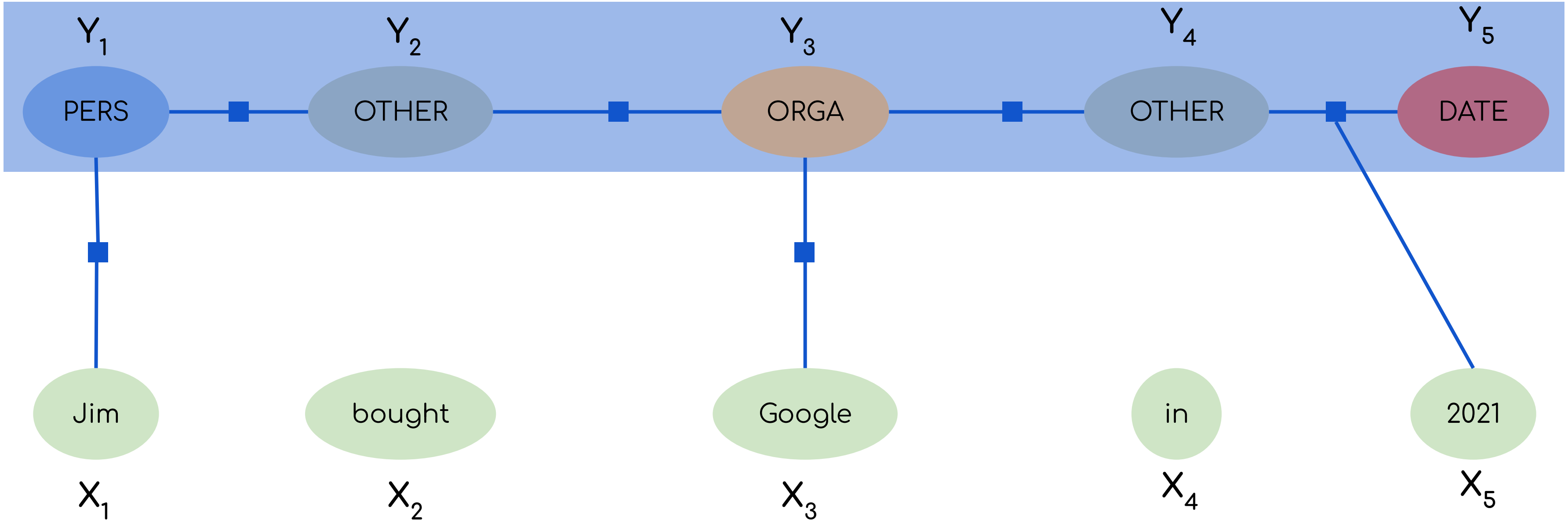
Conditional Random Fields in NER



$\Psi_4(Y_3, Y_2) = \theta_4$ if Y_2 = ‘OTHER’ and Y_3 = ‘ORGA’

Conditional Random Fields in NER

We are interested in modeling P(Y | X) and not P(Y,X) => CONDITIONAL



$\Psi_4(Y_3, Y_2) = \theta_4$ if Y_2 = 'OTHER' and Y_3 = 'ORGA'

Take away points:

- Factor models are a pattern recognition machine learning model for structured prediction

- CRF's model P(Y | X) where X is always observable and does not care about P(X) or P(X | Y) => A discriminative model, can use complex features X and more efficient, BUT always needs complete features X => does not handle incomplete data very well

- We use linear chain CRF's (Every Y_t depends only on Y_{t-1} and Y_{t+1}) for sequential tasks such as NER, Speech to text, …

Conditional Random Fields in NER