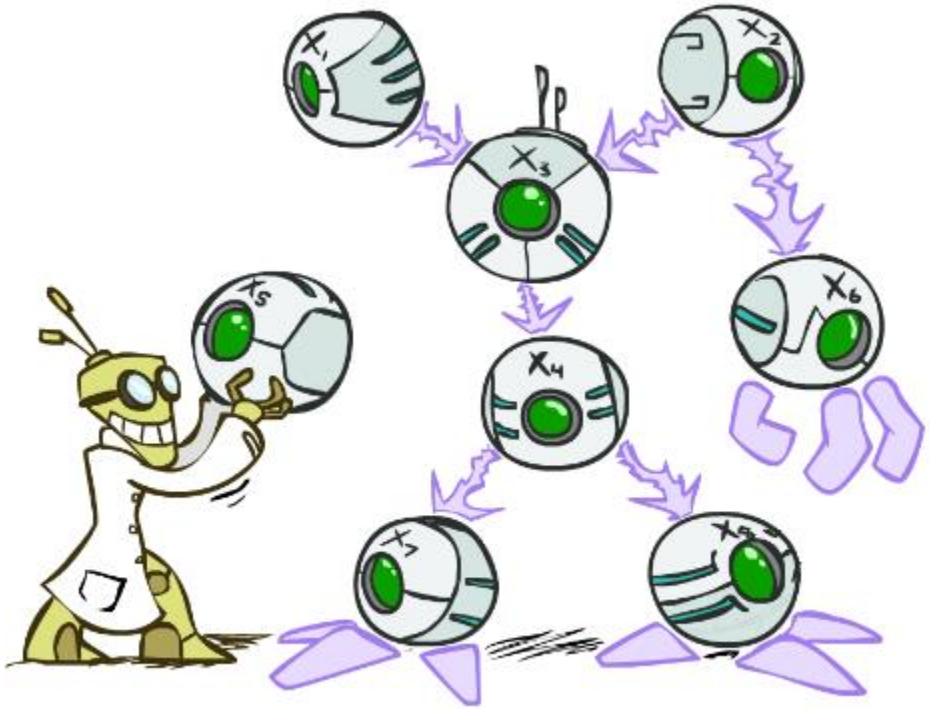
Bayesian Networks



AIMA Chapter 14.1, 14.2, PRML Chapter 8

[Adapted from slides by Dan Klein and Pieter Abbeel at UC Berkeley (ai.berkeley.edu)
and from slides by Markus Svensén & Christopher M Bishop]

Example Application: Topic Modeling

[Containing some materials from slides by David Blei at Princeton and by Hongning Wang at UVa]

Introduction

- A large body of text available online
  - It is difficult to find and discover what we need.
- Topic models
  - Approaches to discovering the main themes of a large unstructured collection of documents
  - Can be used to automatically organize, understand, search, and summarize large electronic archives
  - Latent Dirichlet Allocation (LDA) is the most popular

Plate Notation

- Representation of repeated subgraphs in a Bayesian network

Plate Notation

- Representation of repeated subgraphs in a Bayesian network

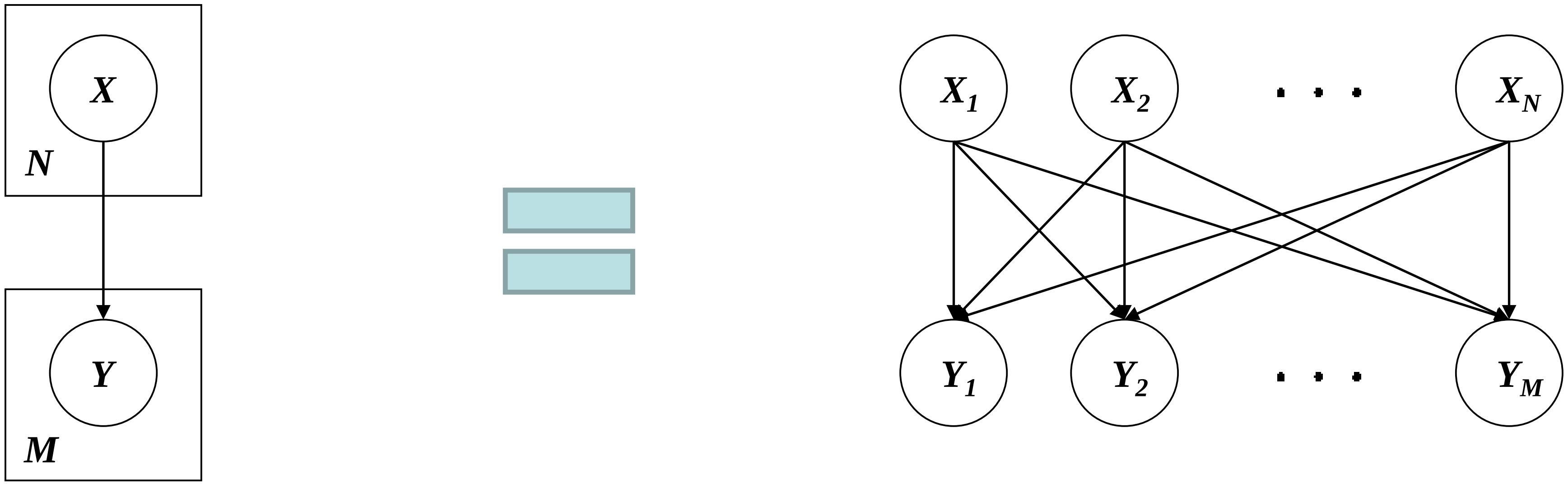
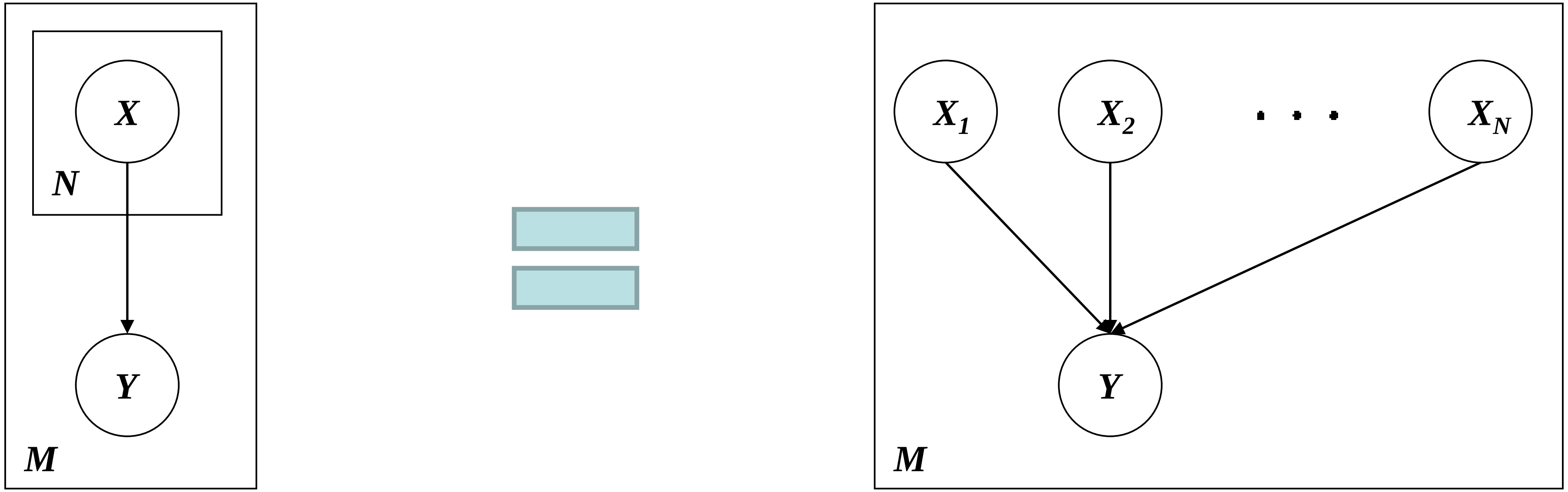
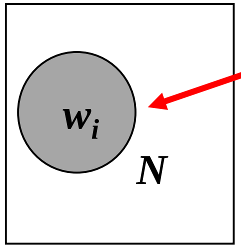


Plate Notation

- Representation of repeated subgraphs in a Bayesian network



How to generate a document



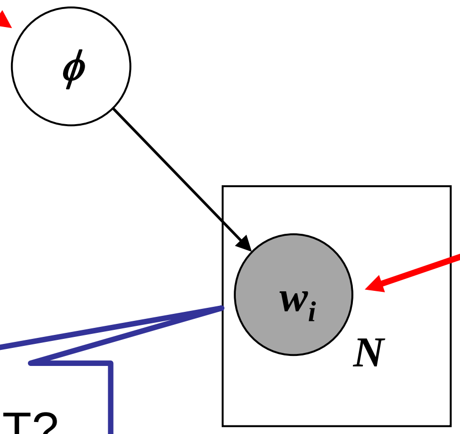
word ***e.g.*, "how", "to", "get", …**

How to generate a document

distribution over words

e.g., [0.1, 0.3, 0.2, …]

how	0.1
to	0.3
get	0.2
…	…



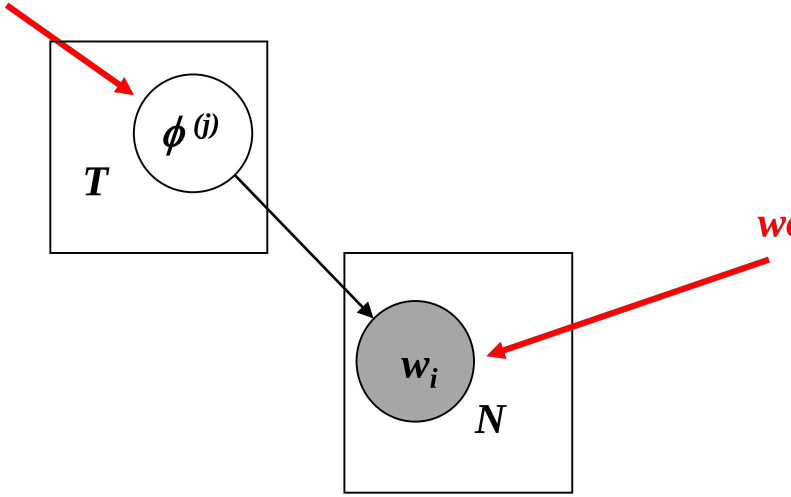
word

What is the CPT?

How to generate a document

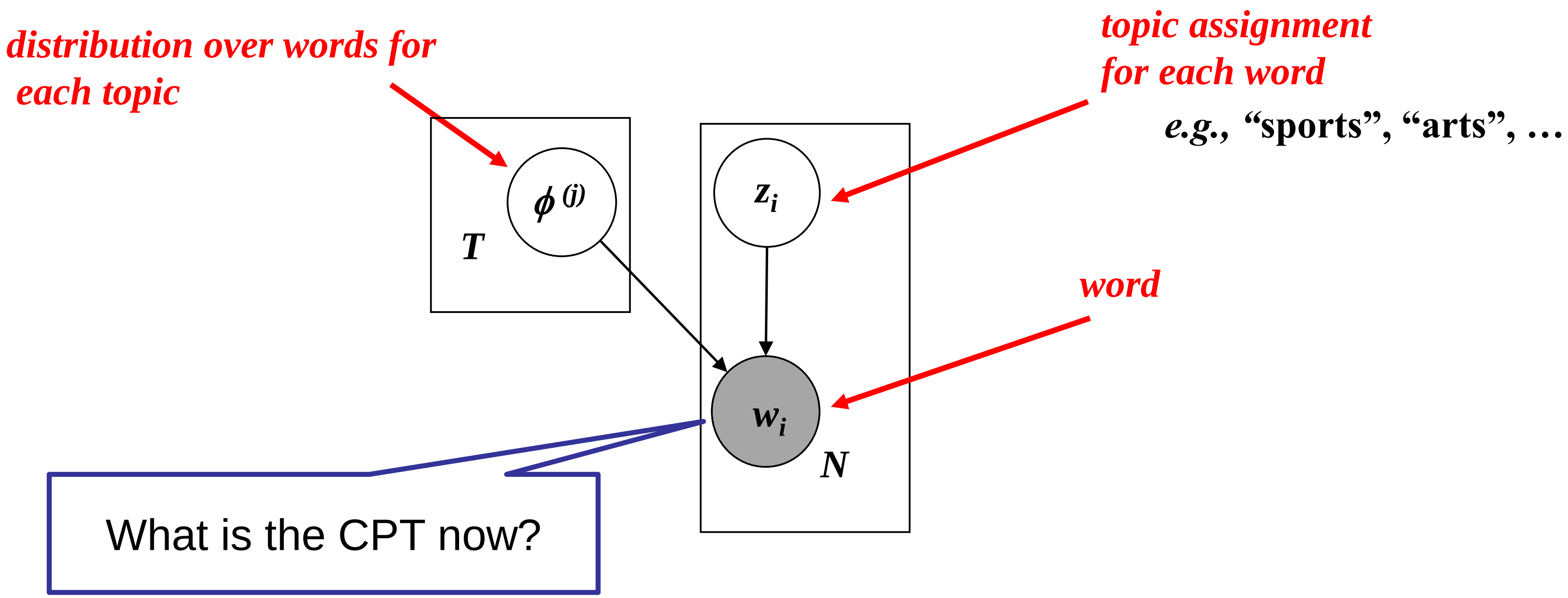
distribution over words for each topic

e.g., sports: [0.1, 0.3, 0.2, …]
arts: [0.2, 0.3, 0.1, …]
science: [0.3, 0.4, 0.1, …]
……

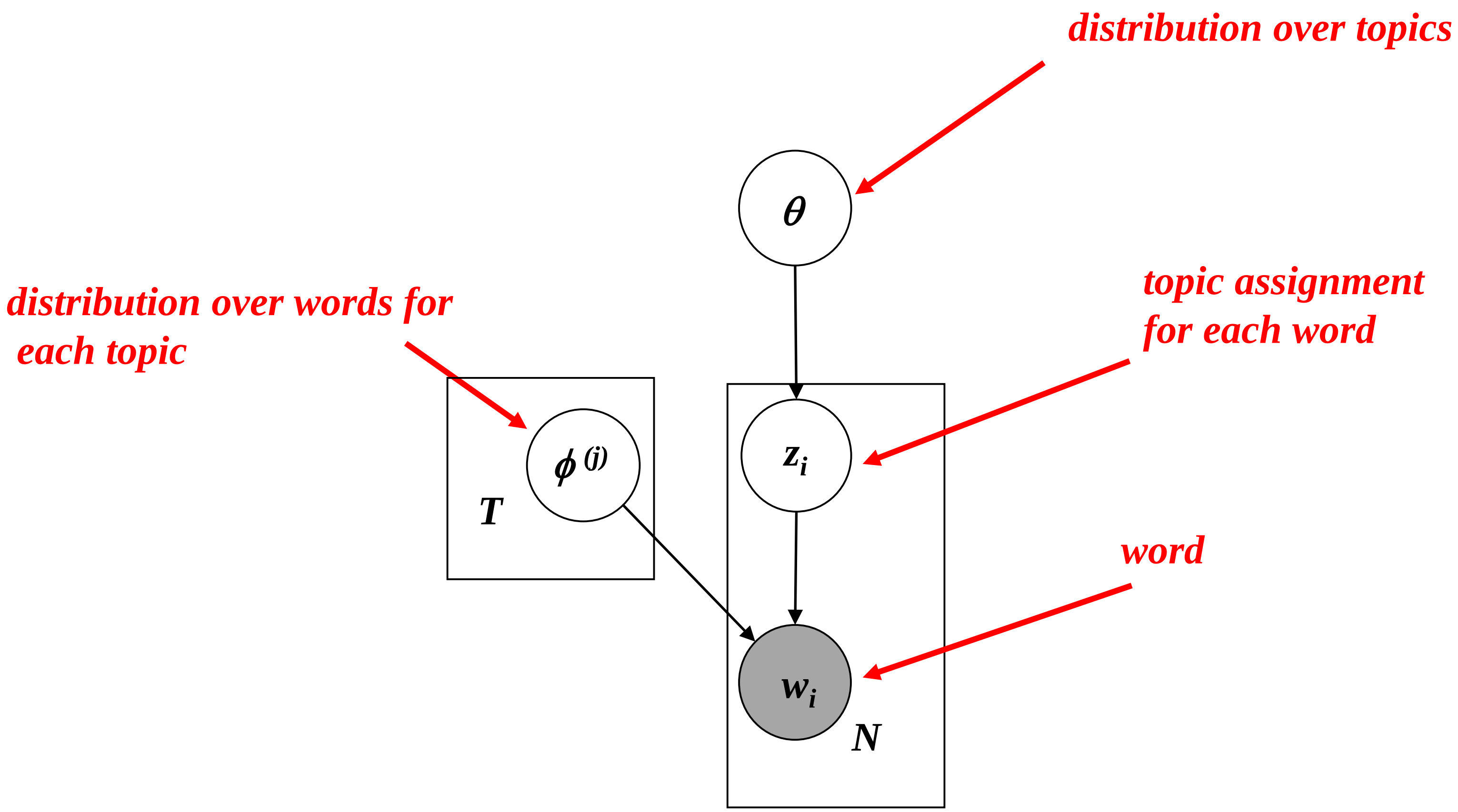


word

How to generate a document



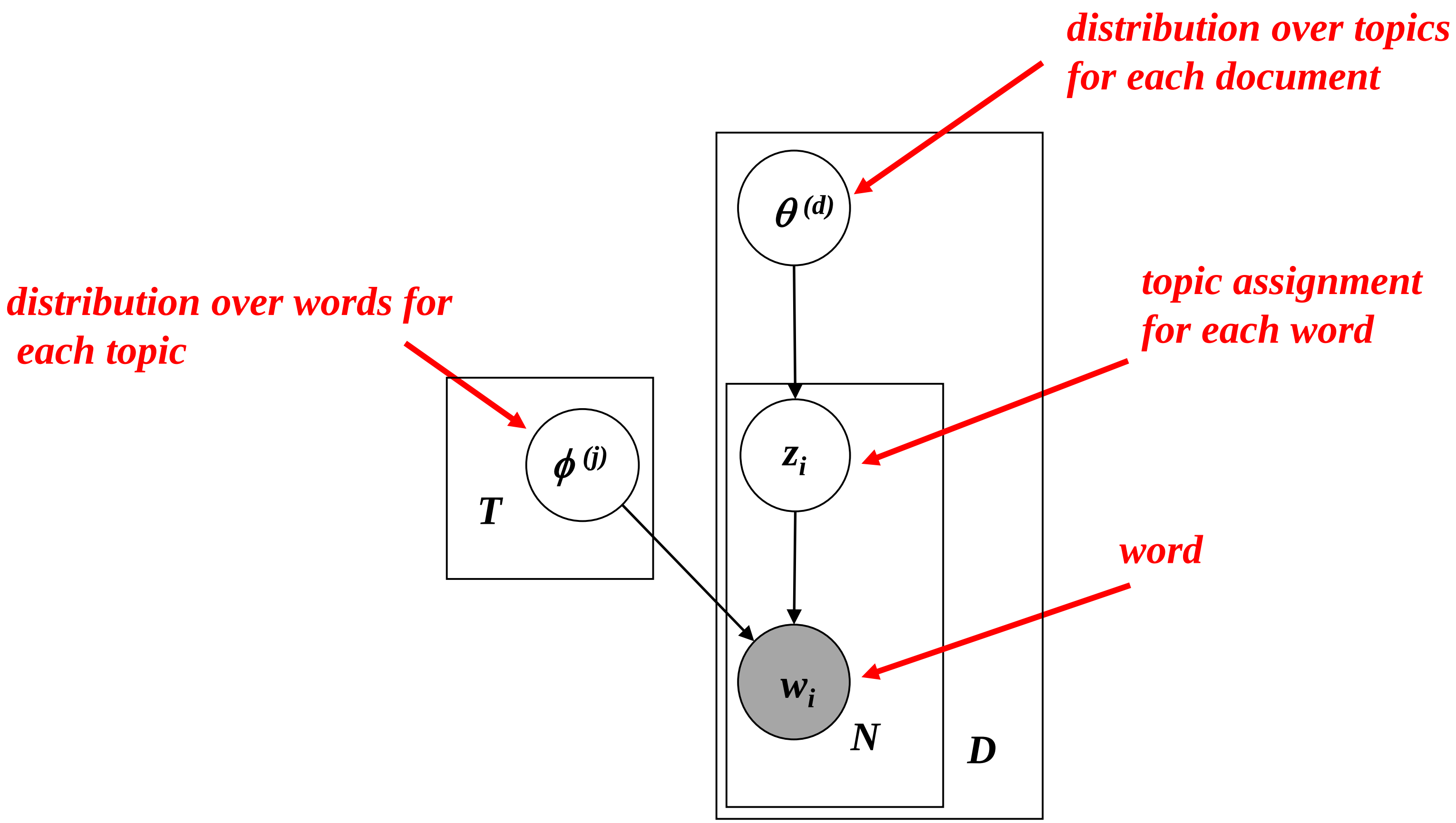
How to generate a document



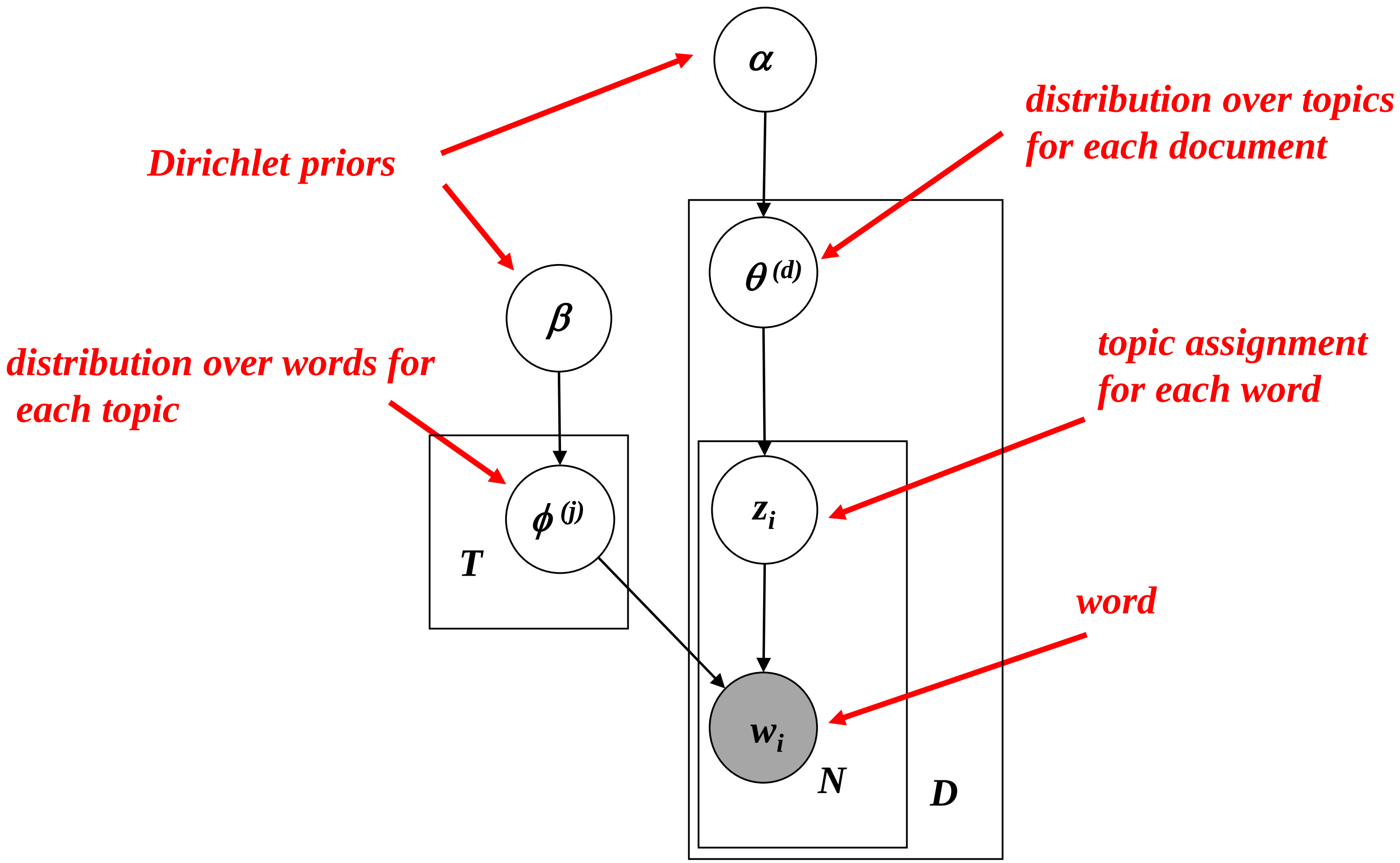
***e.g.*, [0.2, 0.3, 0.2, …]**

sports	**0.1**
arts	**0.3**
science	**0.2**
…	**…**

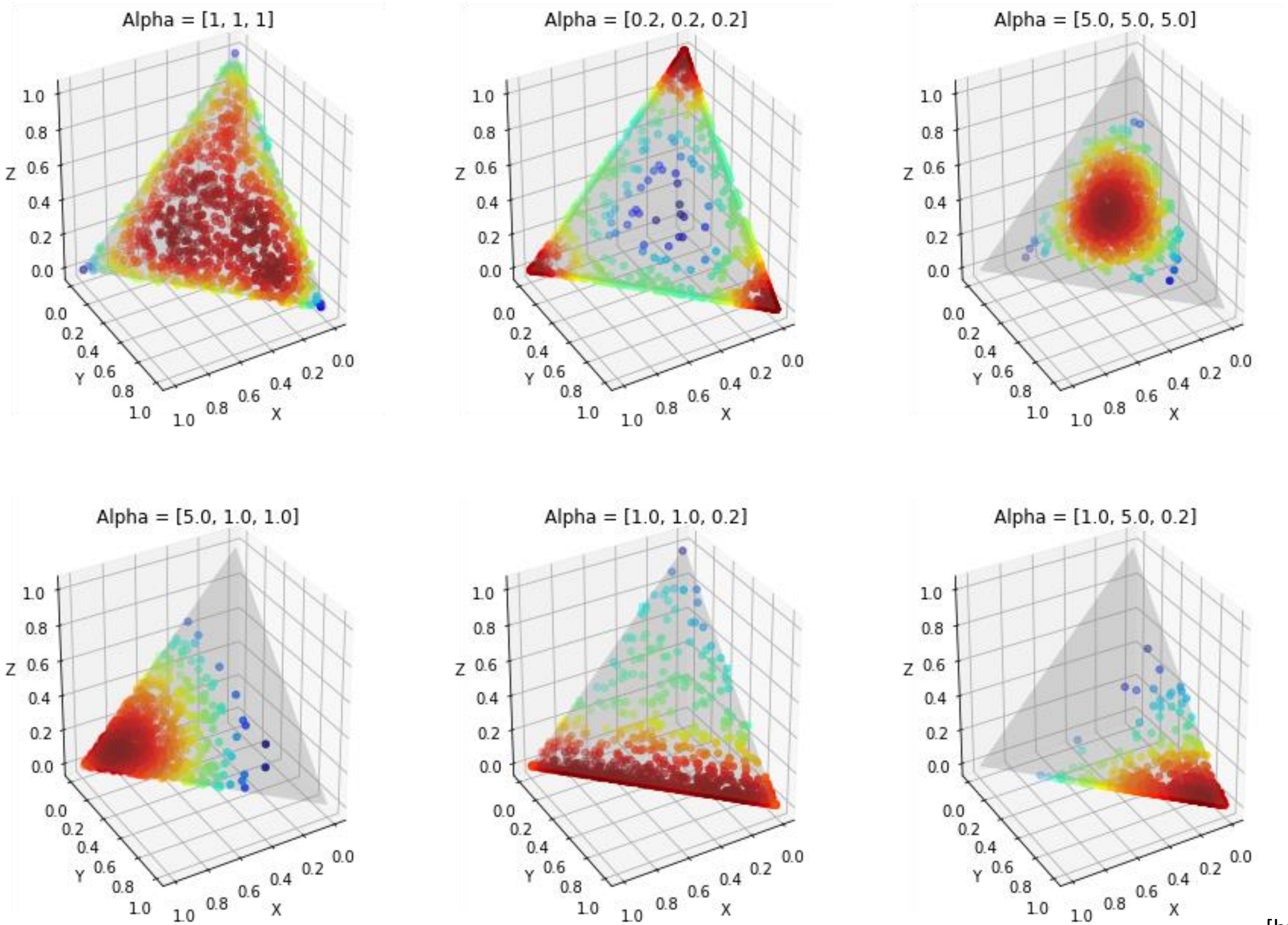
How to generate documents



How to generate documents

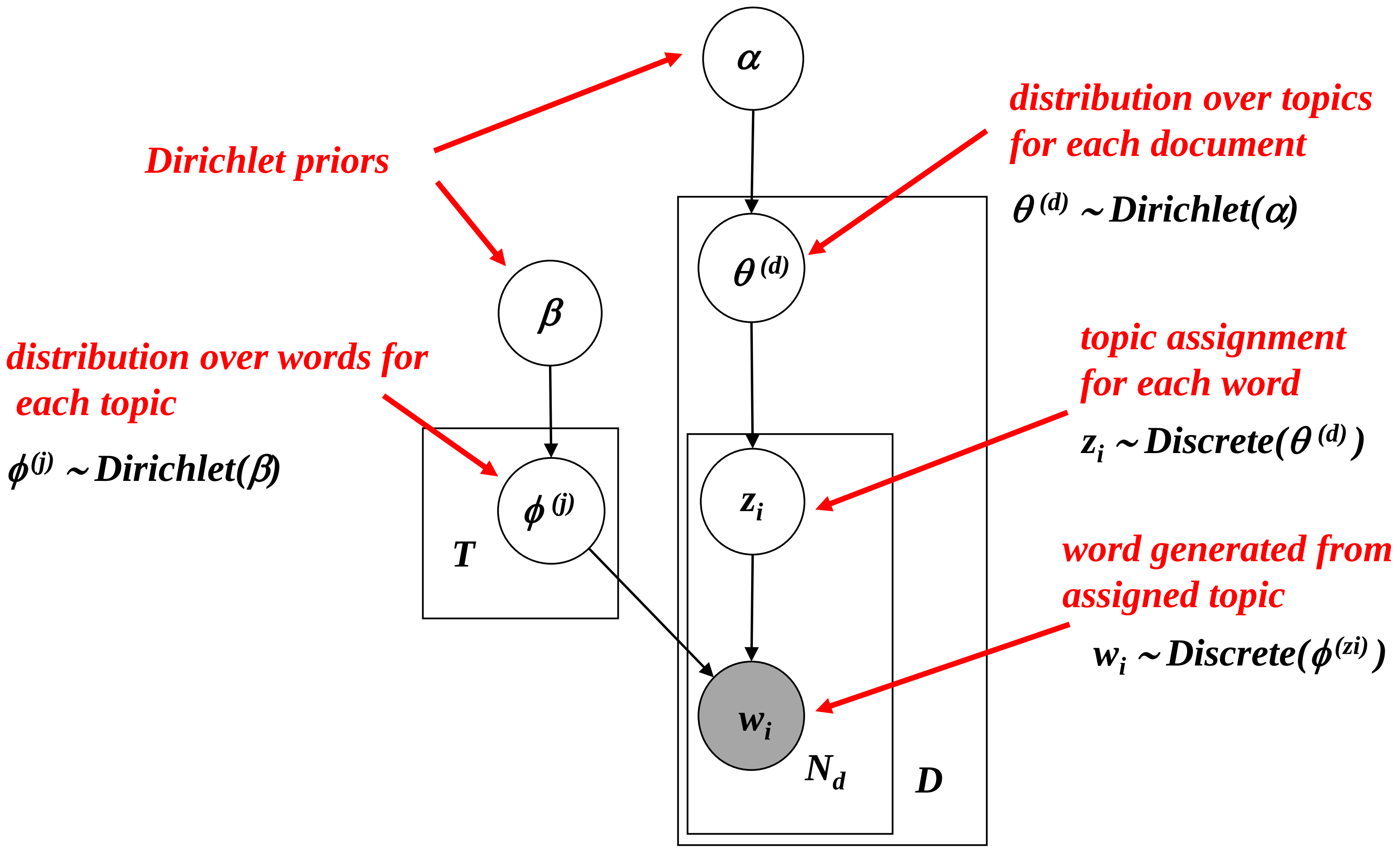


Dirichlet Distribution

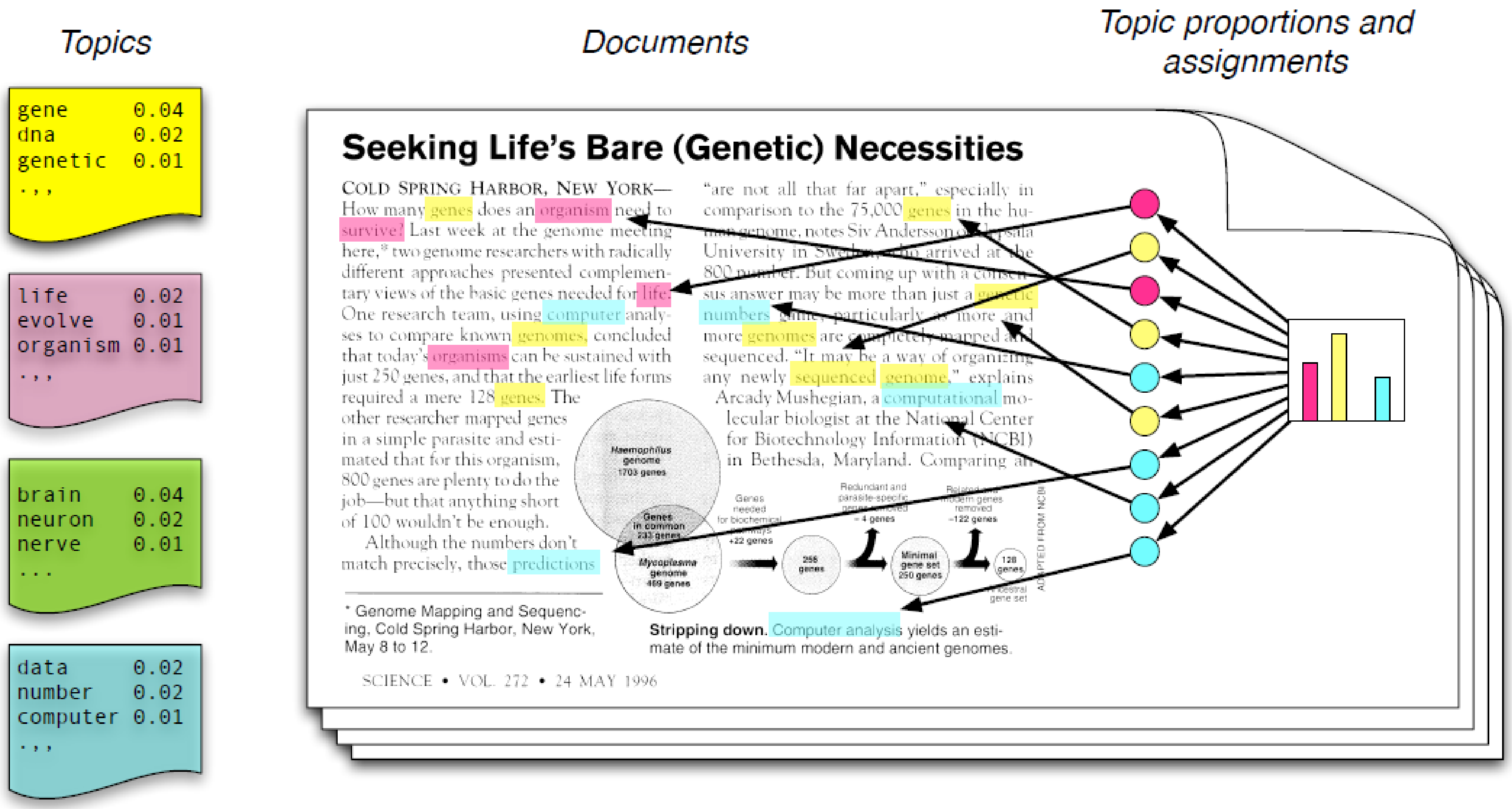


[by Thushan Ganegedara at USydney]

Latent Dirichlet Allocation (LDA)

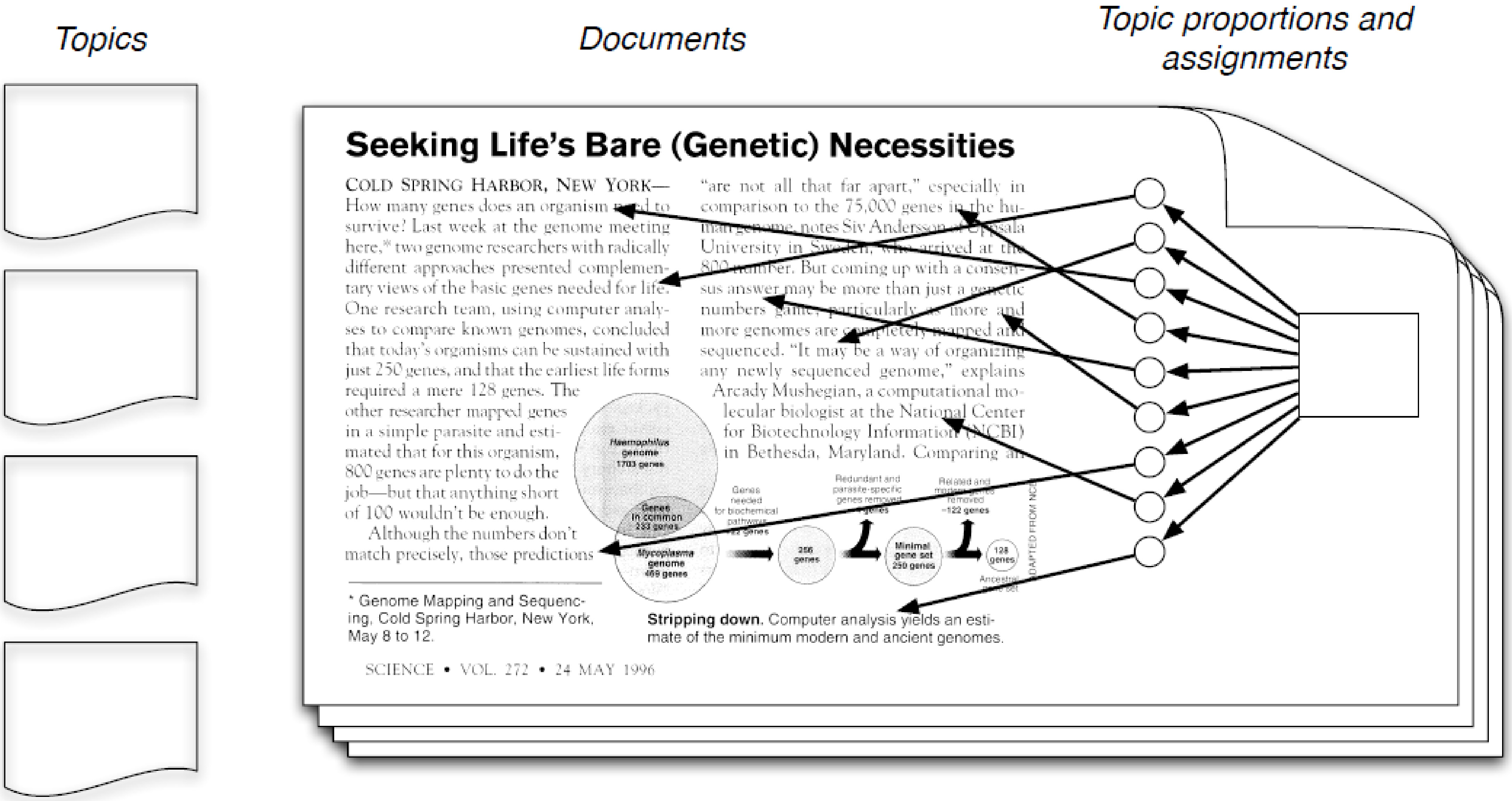


Illustration



- Each **topic** is a distribution of words; each **document** is a mixture of corpus-wide topics; and each **word** is drawn from one of those topics.

Illustration

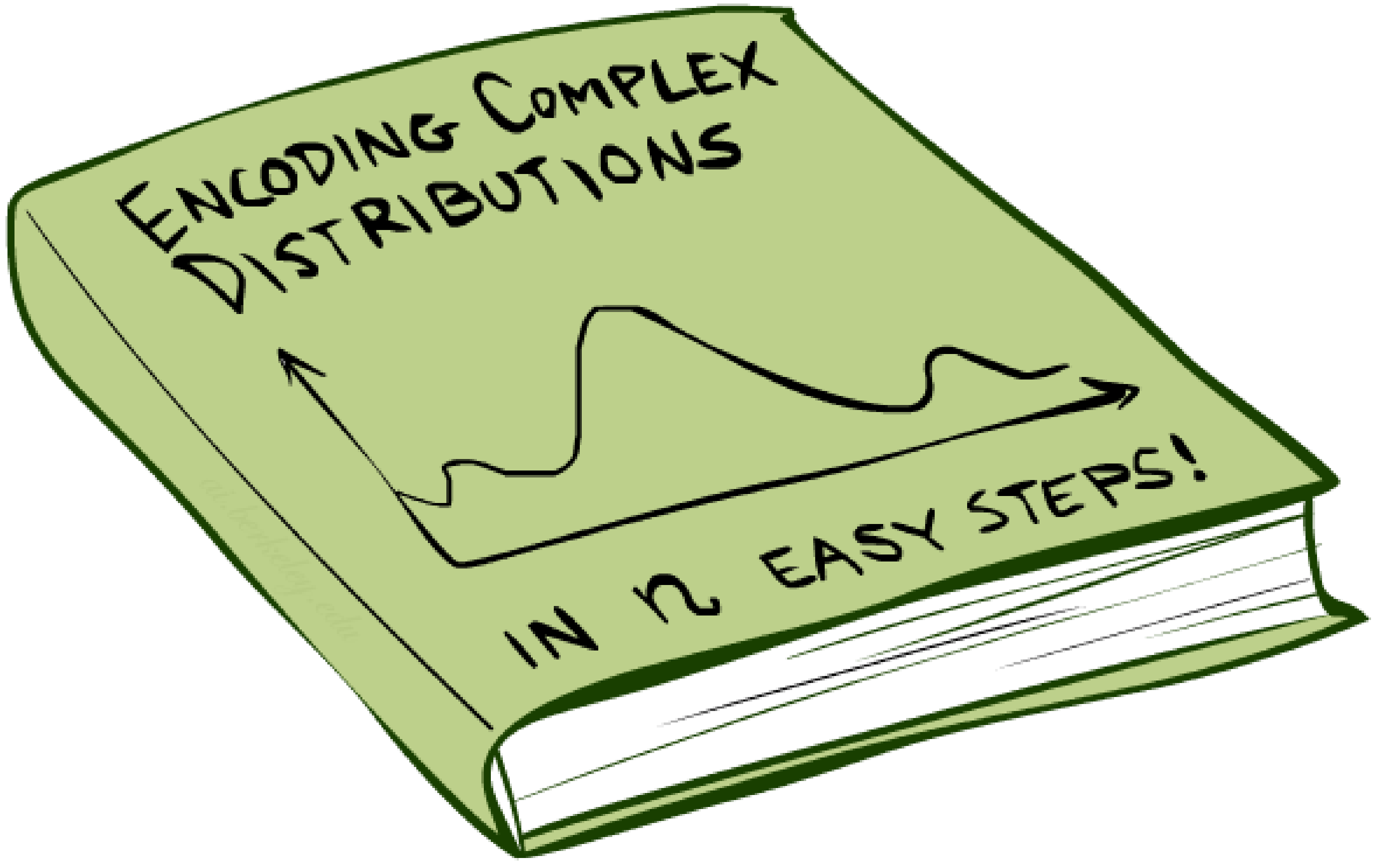


- In reality, we only observe documents. The other structures are hidden variables that must be inferred. (We will discuss inference later.)

Topics inferred by LDA

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Markov Networks



Markov Networks

- A Bayesian network encodes a joint distribution with a directed acyclic graph
  - A CPT captures uncertainty between a node and its parents

- A Markov network (or Markov random field) encodes a joint distribution with an undirected graph
  - A potential function captures uncertainty between a clique of nodes

Markov Networks

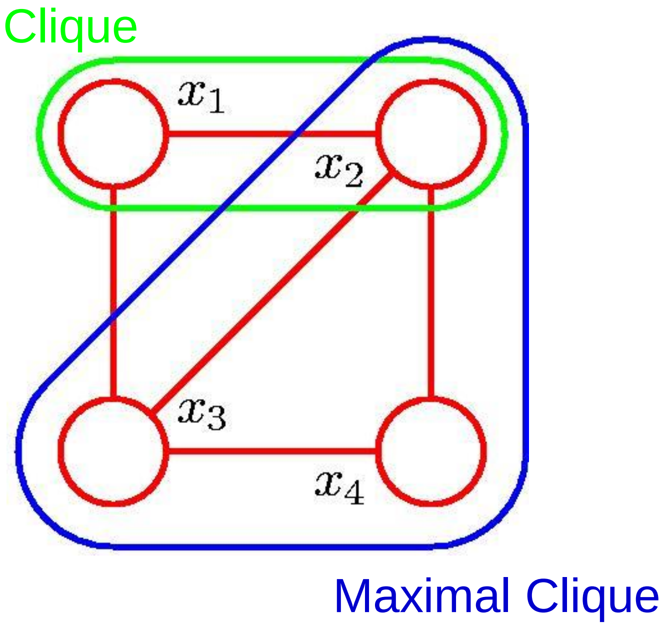
- Markov network = undirected graph + potential functions
  - For each clique (or max clique), a potential function is defined
    - A potential function is not locally normalized, i.e., it doesn't encode probabilities
  - A joint probability is proportional to the product of potentials

$$p(\mathbf{x}) = \frac{1}{Z}\prod_{C}\psi_C(\mathbf{x}_C)$$

where $\psi_C(\mathbf{x}_C)$ is the potential over clique C and

$$Z = \sum_{\mathbf{x}}\prod_{C}\psi_C(\mathbf{x}_C)$$

is the normalization coefficient (aka. partition function).



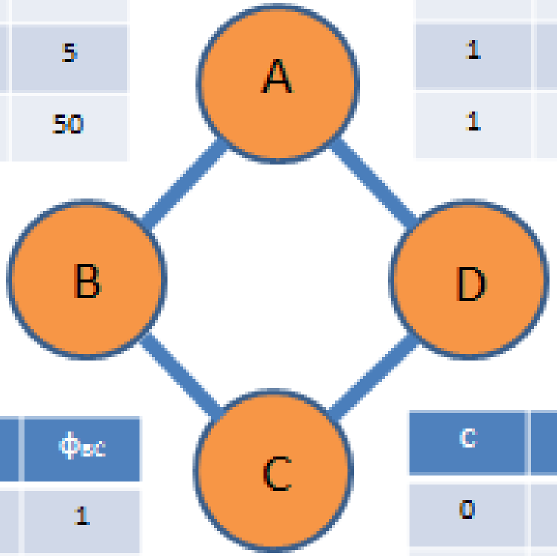
Markov Networks

A	B	C	D	$\phi_{AB}\phi_{BC}\phi_{CD}\phi_{AD}$
0	0	0	0	250
0	0	0	1	37500
0	0	1	0	50000
0	0	1	1	625000
0	1	0	0	1125
0	1	0	1	168750
0	1	1	0	50000
0	1	1	1	625000
1	0	0	0	250
1	0	0	1	375
1	0	1	0	50000
1	0	1	1	6250
1	1	0	0	112500
1	1	0	1	168750
1	1	1	0	5000000
1	1	1	1	625000

$Z = 7520750$

A	B	ϕ_{AB}
0	0	50
0	1	5
1	0	5
1	1	50

A	D	ϕ_{AD}
0	0	5
0	1	50
1	0	50
1	1	5

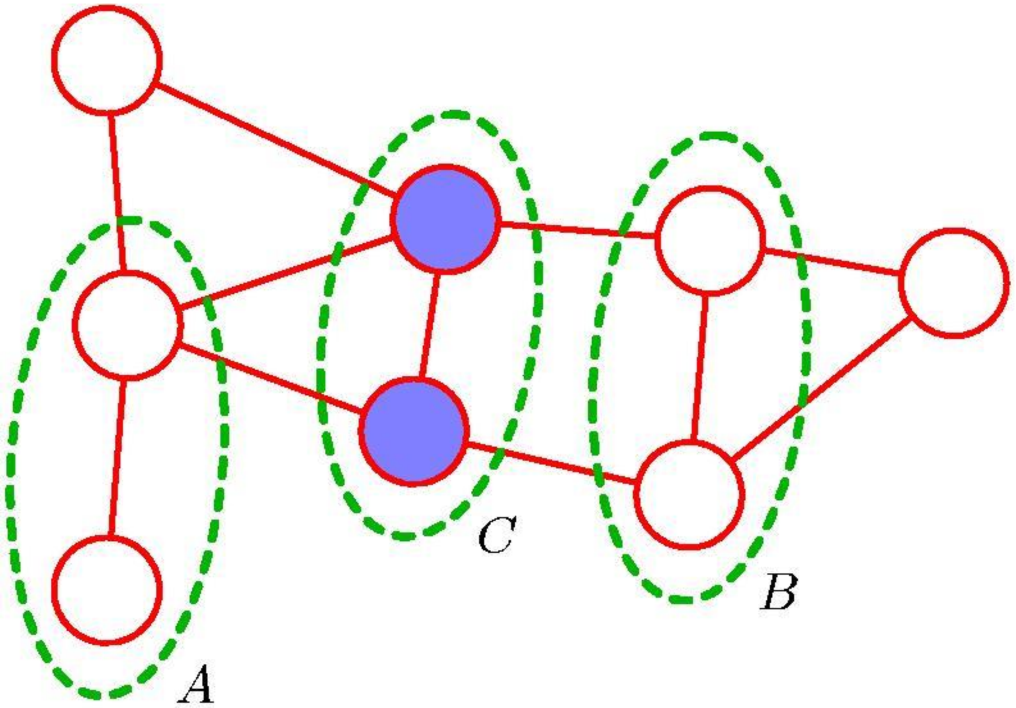


B	C	ϕ_{BC}
0	0	1
0	1	5
1	0	45
1	1	50

C	D	ϕ_{CD}
0	0	1
0	1	15
1	0	40
1	1	50

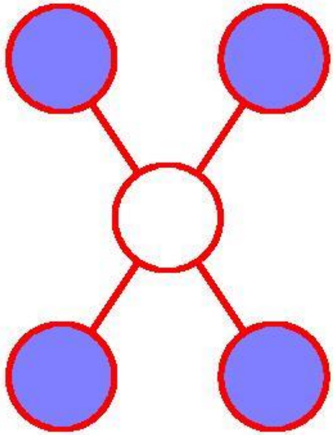
Markov Networks

- Conditional independence and Markov blanket in MN



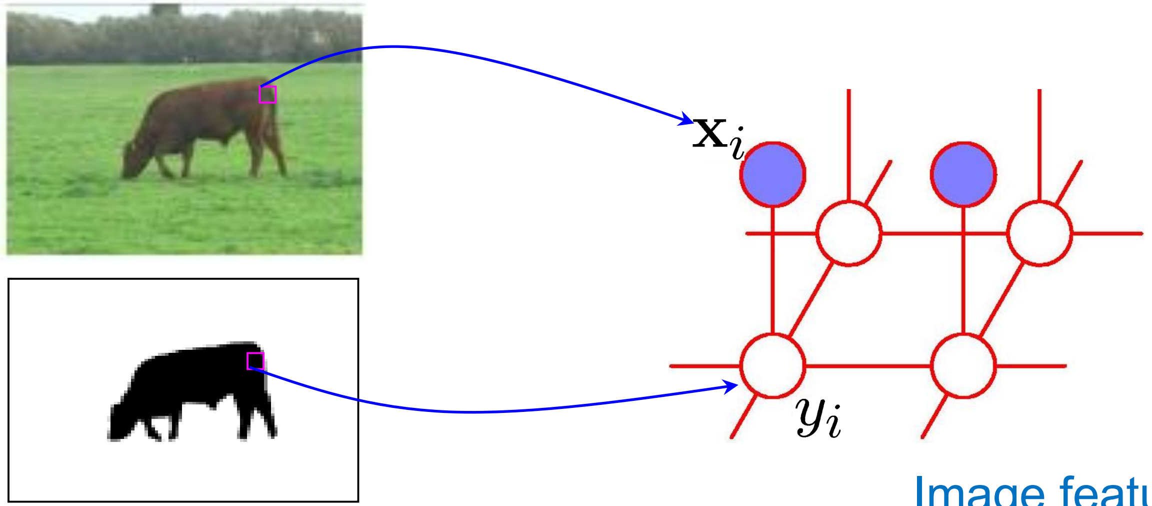
$$A \perp\!\!\!\perp B \mid C$$

Markov Blanket



An example – foreground object

- Binary segmentation



$\mathbf{x}_i$ Image inputs

$y_i \in \{0, 1\}$

Image features (color, texture,…)

Indicator function

$$\hat{y}_i = \delta[\mathbf{w}^T \phi(\mathbf{x}_i) > 0]$$

Weight parameter

- A "local" solution

Toy example:

$$\phi(\mathbf{x}_i) = [green(\mathbf{x}_i), brown(\mathbf{x}_i)]^T$$
$$\mathbf{w} = [-1, 1]^T$$

An example cont'd

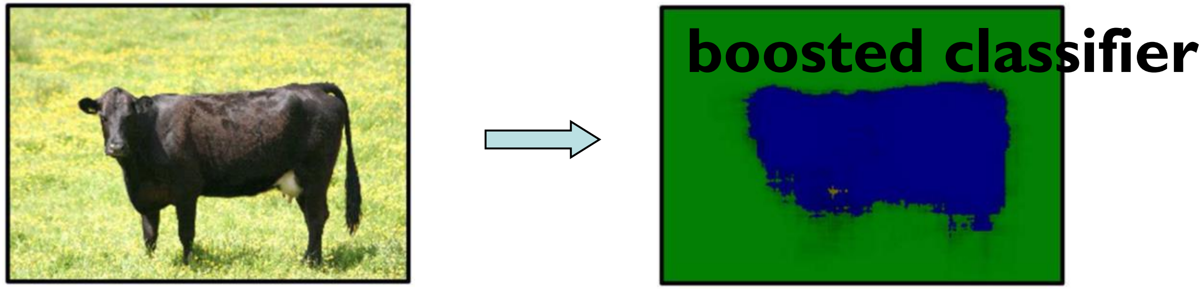
- A score maximization view

$$\hat{y}_i = \delta[\mathbf{w}^T\phi(\mathbf{x}_i) > 0] \iff \hat{y}_i = \arg\max_{y_i} \; y_i\mathbf{w}^T\phi(\mathbf{x}_i)$$

$$= \arg\max_{y_i} \; \underbrace{\mathbf{w}^T\tilde{\phi}(\mathbf{x}_i, y_i)}_{\text{Score function}}$$

  - Predicted label has a higher score.

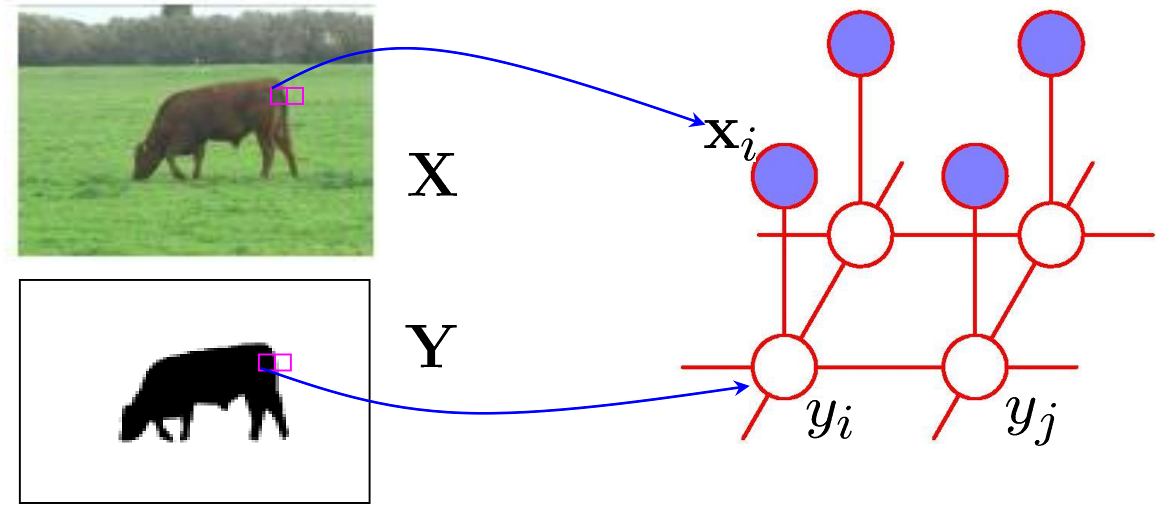
- Problem?



(Shotton et al, ECCV 2006)

An example cont’d

- Incorporating spatial context
  - Labels are generally spatially smooth



$\mathbf{x}_i$ Image inputs

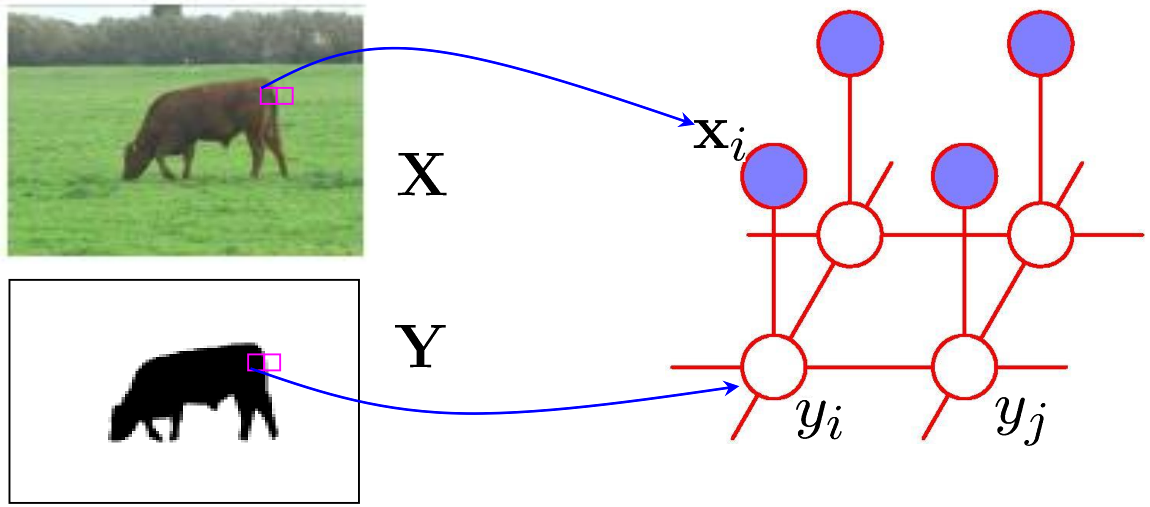
$y_i \in \{0, 1\}$

$$F(\mathbf{X}, \mathbf{Y}; \mathbf{W}) = \sum_i \mathbf{w}^T \phi(y_i, \mathbf{x}_i)$$

“Local” image cues

An example cont'd

- A simple smooth model



$\mathbf{x}_i$ Image inputs

$y_i \in \{0, 1\}$

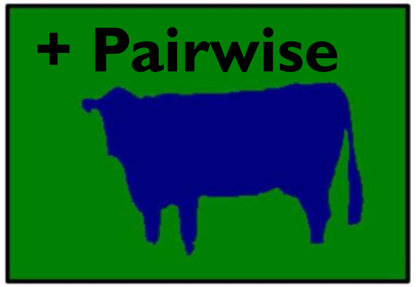
$$\psi(y_i, y_j, |\mathbf{x}_i - \mathbf{x}_j|) = \delta[y_i = y_j] e^{\{-|\mathbf{x}_i - \mathbf{x}_j|\}}$$

- Same labeling for neighboring pixels unless an intensity gradient exists

An example cont'd

- Inferring the scene properties (i.e., foreground mask) globally

$$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}} F(\mathbf{X}, \mathbf{Y}; \mathbf{W})$$
$$= \arg\max_{\mathbf{Y}} \sum_i \mathbf{w}^T \phi(y_i, \mathbf{x}_i) + \alpha \sum_{i,j} \psi(y_i, y_j, |\mathbf{x}_i - \mathbf{x}_j|)$$



Structured prediction framework

- Input $\mathbf{X} = \{x_i\}_{i=1}^n, \quad x_i \in \mathbf{R}^d$
- Output $\mathbf{Y} = \{y_i\}_{i=1}^n, \quad y_i \in \mathbf{L}, \mathbf{L} = \{1, \cdots, K\}$
- Structured prediction model

$$\hat{\mathbf{Y}} = \operatorname{arg\,max}_{\mathbf{Y}} F(\mathbf{X}, \mathbf{Y}, \mathbf{W})$$

$$F(\mathbf{X}, \mathbf{Y}, \mathbf{W}) = \sum_i \phi(x_i, y_i; \mathbf{w}_u)$$ Unary potential

$$+ \sum_{i,j} \psi(\mathbf{x}_{ij}, y_i, y_j; \mathbf{w}_p)$$ Pairwise potential

$$+ \sum_c \psi_c(\mathbf{x}_c, \mathbf{y}_c; \mathbf{w}_c)$$ Higher-order potential

- Examples: surface contour, object class, depth, pose, …

Structured prediction framework

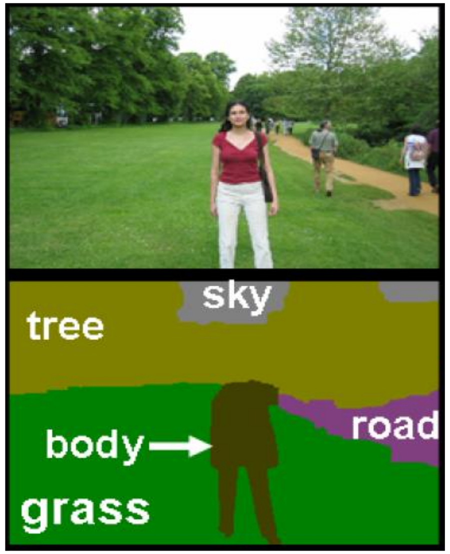
- A probabilistic view – (conditional) random field
  - Conditional probability

$$P(\mathbf{Y}|\mathbf{X};\mathbf{W}) = \frac{1}{Z_{\mathbf{X},\mathbf{W}}} \exp\{F(\mathbf{X},\mathbf{Y},\mathbf{W})\}$$

  - Label prediction: MAP estimation
- Main question in scene modeling
  - What are the potential functions?
  - Hand-crafted features, deep neural networks, …

Multiclass scene labeling

- TextonBoost CRF (Shotton et al., ECCV 2006)
  - Simultaneous recognition and segmentation
  - Explain every pixel (dense features)



Model output

Model overview – TextonBoost CRF

- What are useful cues for object classification?
  - Appearance (color, texture,…)
  - Shape
  - Object location
  - Spatial context
- Incorporating those factors into a score function:

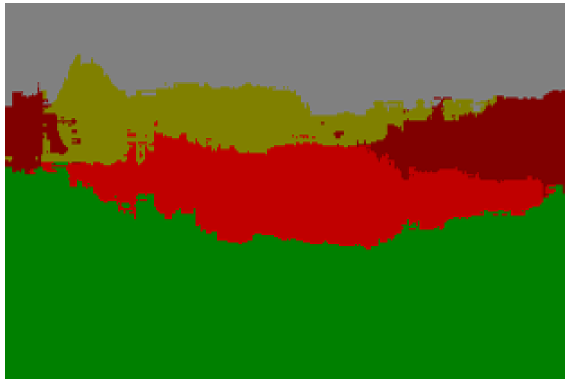
F = shape-texture term (A) + color term (B)
+ location term (C) + spatial context term (D)

A. Shape-texture potential

$$F(\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}) = \underbrace{\sum_i \psi_i(y_i, \mathbf{x}; \boldsymbol{\theta}_\psi)}_{\text{shape-texture potentials}}$$

jointly across all pixels

- Shape-texture potentials
  - broad intra-class appearance distribution
  - log boosted classifier
  - parameters $\boldsymbol{\theta}_\psi$ learned offline



shape-texture potentials

B. Color potential

$$F(\mathbf{X},\mathbf{Y},\boldsymbol{\theta}) \;=\; \sum_i \psi_i(y_i,\mathbf{x};\boldsymbol{\theta}_\psi) + \overbrace{\pi(y_i,\mathbf{x}_i;\boldsymbol{\theta}_\pi)}^{\text{colour potentials}}$$

- Colour potentials
  - compact appearance distribution
  - Gaussian mixture model

intra-class
appearance variations

C. Location potential

$$F(\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}) \;=\; \sum_i \psi_i(y_i, \mathbf{x}; \boldsymbol{\theta}_\psi) + \pi(y_i, \mathbf{x}_i; \boldsymbol{\theta}_\pi) + \overbrace{\lambda(y_i, i; \boldsymbol{\theta}_\lambda)}^{\text{location potentials}}$$

- Capture prior on absolute image location

tree sky road

D. Spatial context

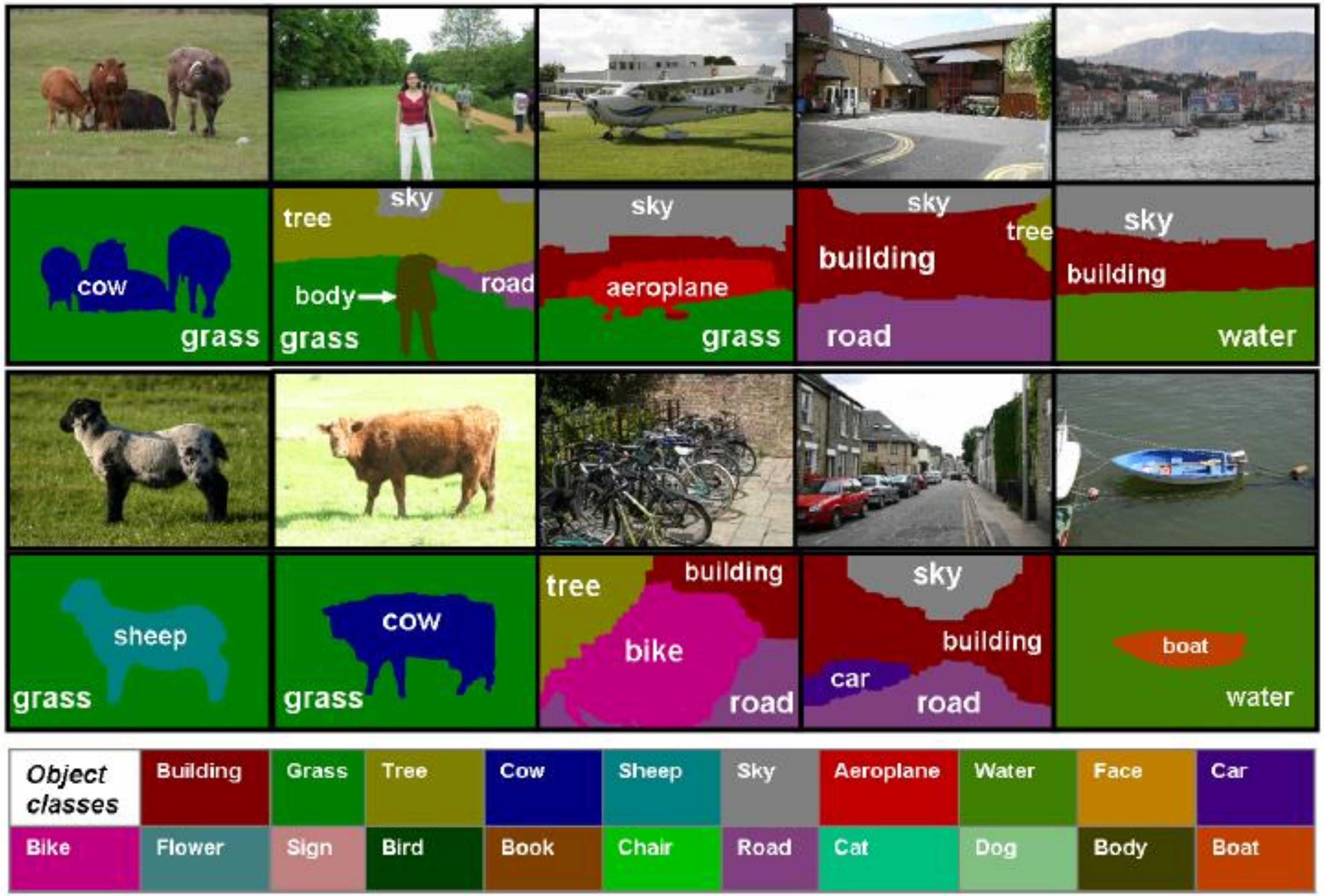
$$F(\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}) = \sum_i \psi_i(y_i, \mathbf{x}; \boldsymbol{\theta}_\psi) + \pi(y_i, \mathbf{x}_i; \boldsymbol{\theta}_\pi) + \lambda(y_i, i; \boldsymbol{\theta}_\lambda) + \sum_{(i,j)\in\mathcal{E}} \underbrace{\phi(y_i, y_j, \mathbf{g}_{ij}(\mathbf{x}); \boldsymbol{\theta}_\phi)}_{\text{edge potentials}}$$

sum over neighbouring pixels

- Potts model
  - encourages neighbouring pixels to have same label
- Contrast sensitivity
  - encourages segmentation to follow image edges

image edge map

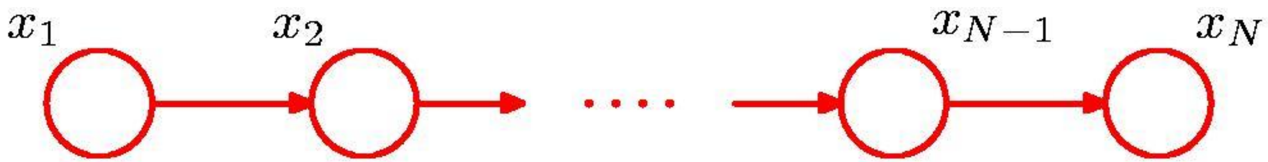
Good Results



Graphical Models

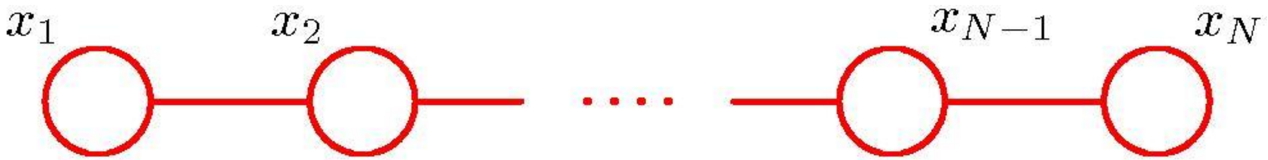
- A graphical model is a probabilistic model for which a graph expresses conditional dependence between random variables
  - Bayesian networks: directed acyclic graph
  - Markov networks: undirected graph
  - Factor graphs, conditional random fields, etc.

Converting Directed to Undirected Graphs (1)



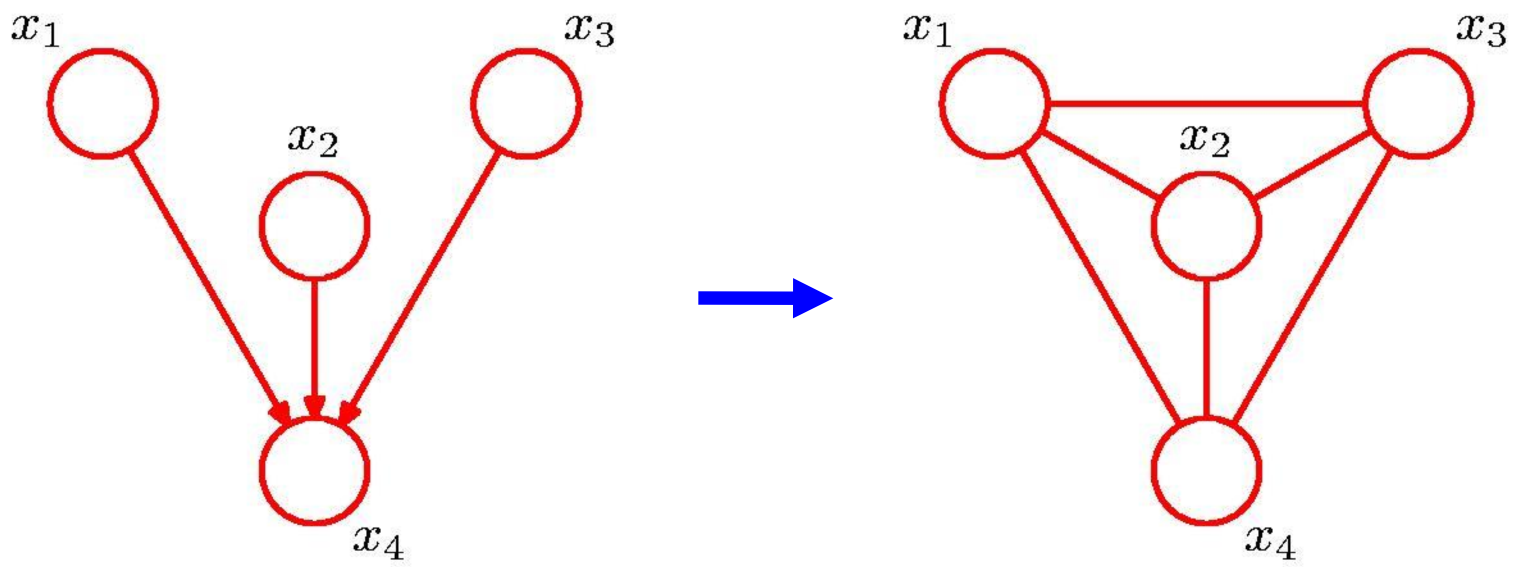
$$p(\mathbf{x}) = \underbrace{p(x_1)p(x_2|x_1)}\, p(x_3|x_2) \cdots p(x_N|x_{N-1})$$

$$p(\mathbf{x}) = \frac{1}{Z}\, \psi_{1,2}(x_1, x_2)\, \psi_{2,3}(x_2, x_3) \cdots \psi_{N-1,N}(x_{N-1}, x_N)$$



Converting Directed to Undirected Graphs (2)

- Additional links (moralization)

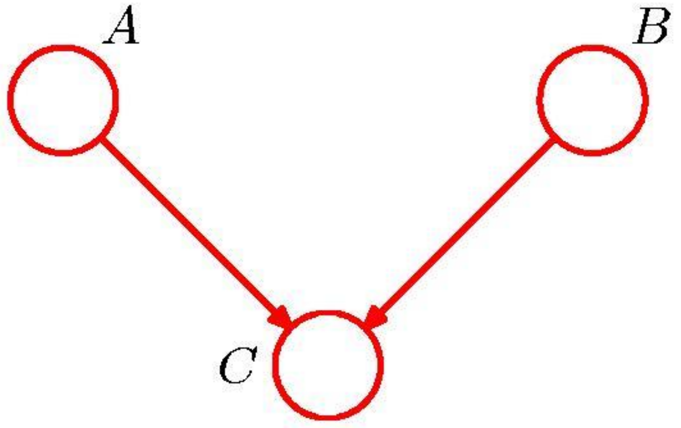


$$
\begin{aligned}
p(\mathbf{x}) &= p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3) \\
&= \frac{1}{Z}\psi(x_1,x_2,x_3,x_4)
\end{aligned}
$$

Bayesian Network → Markov Network

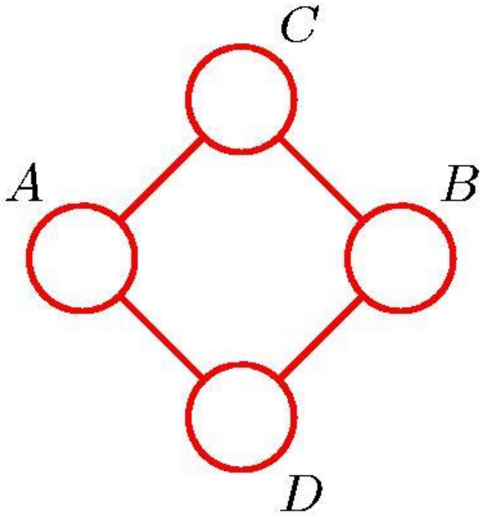
- Steps
  1. Moralization
  2. Construct potential functions from CPTs
- The BN and MN encode the same distribution
- Do they encode the same set of conditional independence?

Encoding Conditional Independence



$$A \perp\!\!\!\perp B \mid \emptyset$$

$$A \not\!\perp\!\!\!\perp B \mid C$$

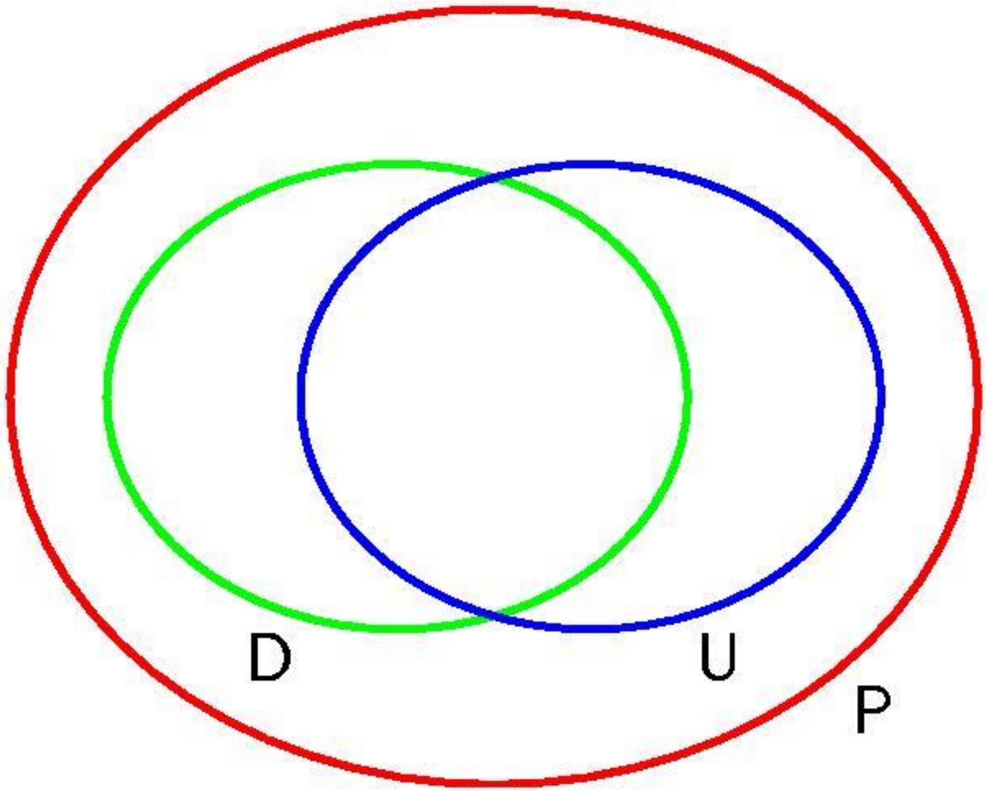


$$A \not\!\perp\!\!\!\perp B \mid \emptyset$$

$$A \perp\!\!\!\perp B \mid C \cup D$$

$$C \perp\!\!\!\perp D \mid A \cup B$$

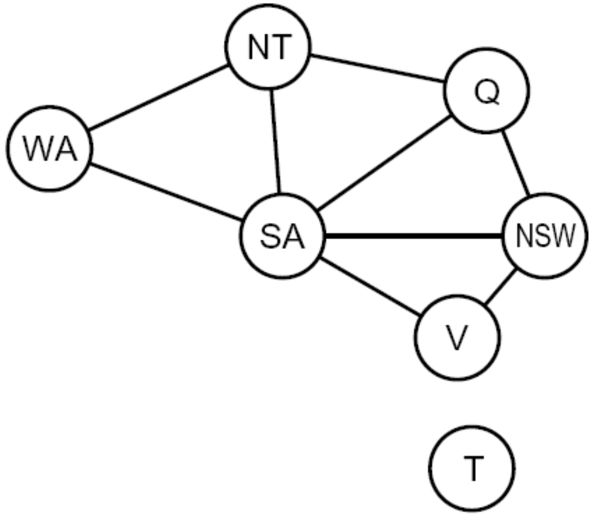
Encoding Conditional Independence

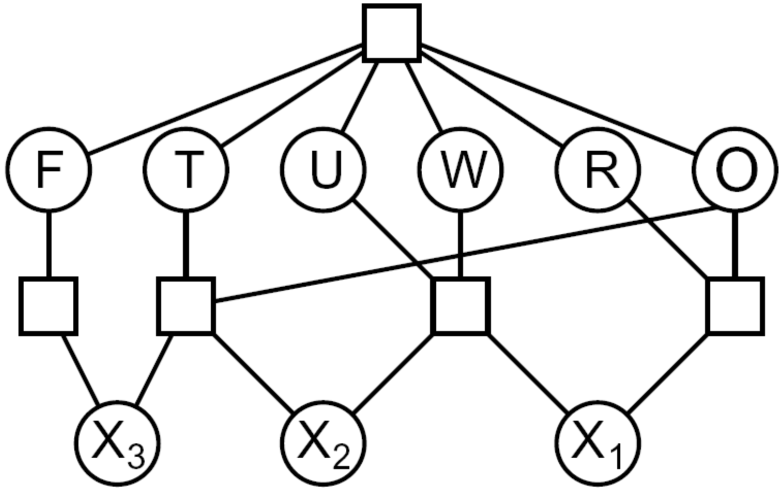


The set of distributions whose conditional independence can be exactly (i.e., no more, no less) represented by a **directed**/**undirected** graph

Markov networks vs. Constraint graphs

- Constraint graphs can be seen as Markov networks with 0/1 potentials





BN/MN vs. Logic

- Which logic is BN/MN more similar to: PL? FOL?
  - Boolean nodes represent propositions
  - No explicit representation of objects, relations, quantifiers
- BN/MN can be seen as a probabilistic extension of PL
- PL can be seen as BN/MN with deterministic CPTs/potentials