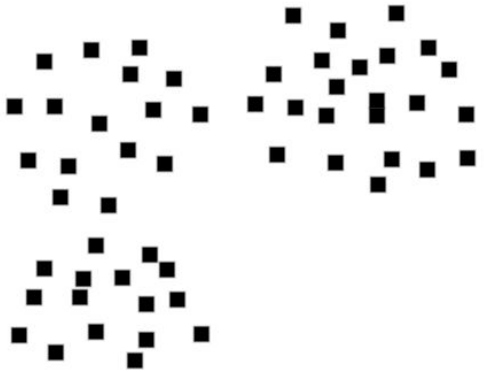
Graphical Models and Causal Inference

Many material from http://ai.stanford.edu/~paskin/gm-short-course/lec2.pdf

Clustering

- Discover underlying structure of data



- Learn the hidden class labels in the data.

Simple clustering

- K-means
  - Find the means, and assign
- GMM
  - Find the means, covariance, then soft assignment
- K-means vs GMM
  - Spherical covariance vs any covariance
  - Hard vs soft assignment

Can we give priors knowledge to clustering?

Latent Dirichlet Allocation (LDA)

- Another method that learns the hidden labels.
- Mostly used for topic modeling and document classification

  - A kind of probabilistic graphical model
    - In this class we will only talk about Bayesian network, Bayes net, directed graphical model, or Belief network

- Let's cover some basics

Directed graphical models

Looks like this

Tells relationship

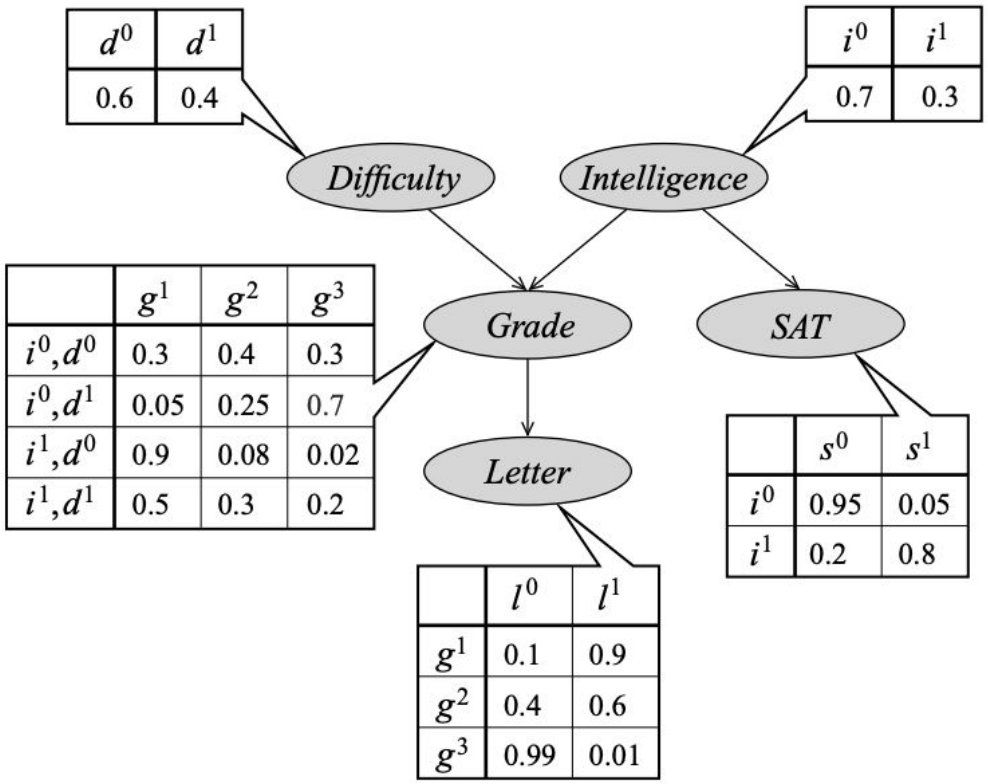


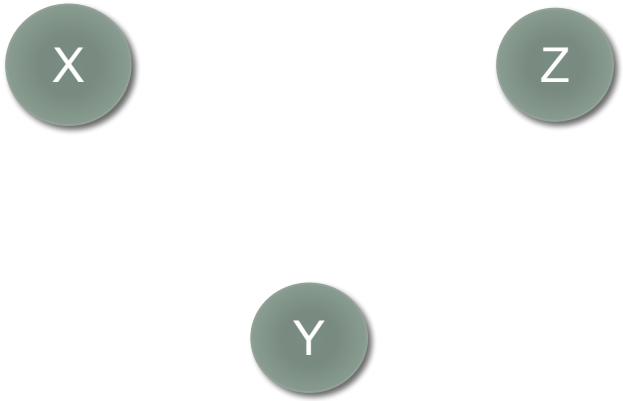
Figure 3.4 Student **Bayesian network** $\mathcal{B}^{student}$ **with CPDs**

Koller et al. Probabilistic graphical models principles and techniques.

Arrows

Arrows implies direct influence. Preferably causal relationship, but not required for Bayesian networks.

We only allow acyclic graph.

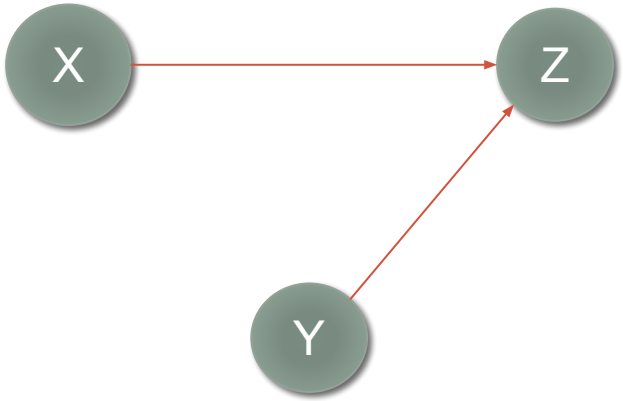


X - Nationality,
Z - food preference
Y- Weight

Arrows

Arrows implies direct influence. Preferably causal relationship, but not required for Bayesian networks.

We only allow acyclic graph.

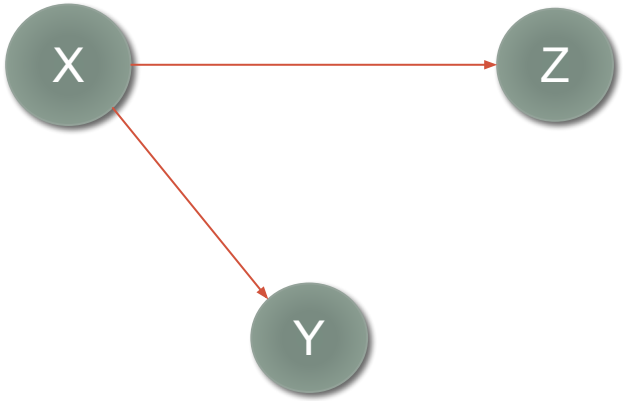


X - Nationality,
Z - food preference
Y- Weight

Arrows

Arrows implies direct influence. Preferably causal relationship, but not required for Bayesian networks.

We only allow acyclic graph.



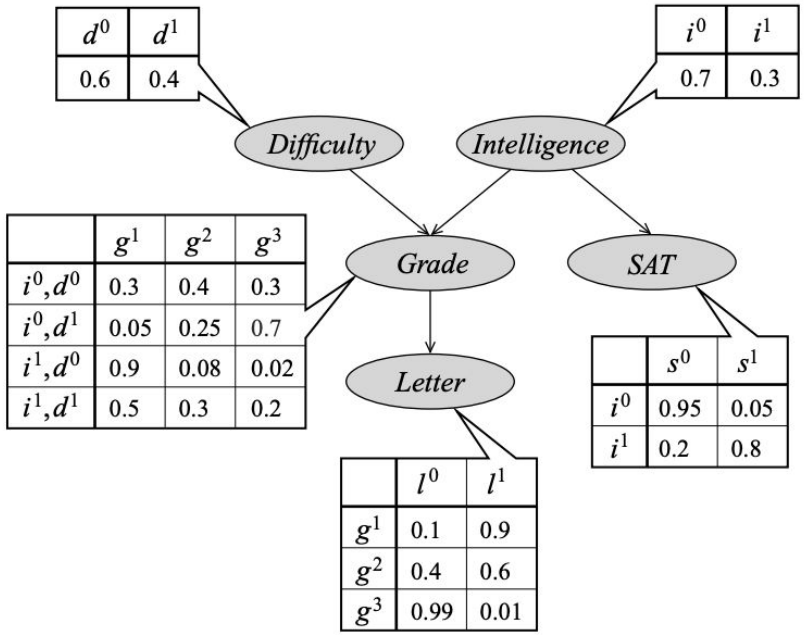
X - Nationality,
Z - food preference
Y- Weight

Arrows and distribution modeling

Each arrow implies a distribution that need to be modelled

This factorizes the distribution

P(D, I, G, S, L) = P(D) P(I) P(G | I, D) P (S | I) P(L| G)



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Figure 3.4 Student **Bayesian network** $\mathcal{B}^{student}$ **with CPDs**

Probabilistic models and independence assumptions

- Implies

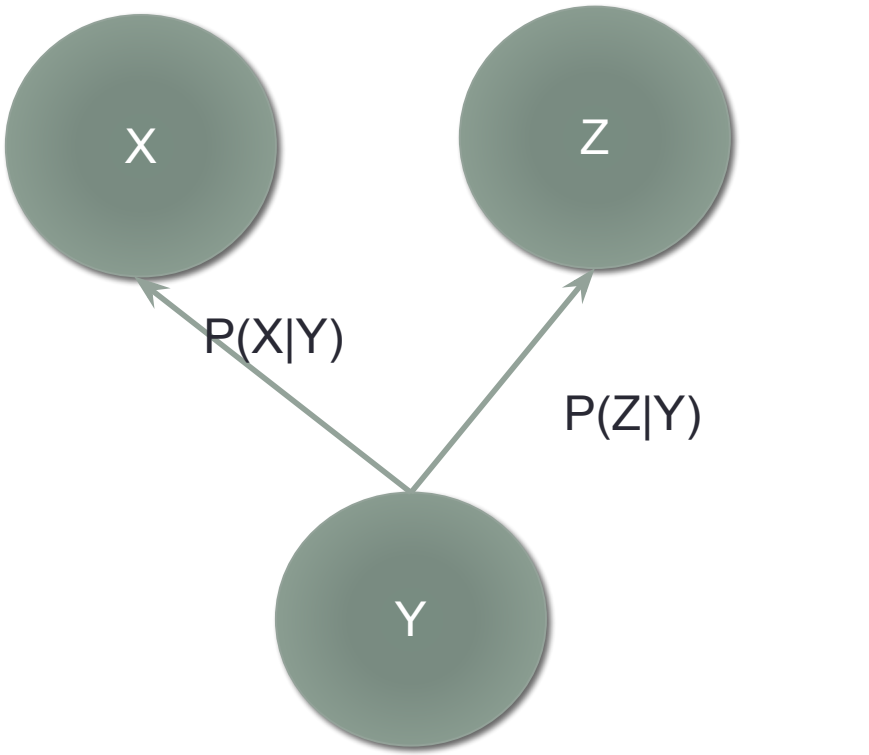
$X \perp\!\!\!\perp Z \mid Y$

How?

$P(X,Y,Z) = P(X,Z|Y)P(Y)$

$= P(X|Y)P(Z|Y)P(Y)$

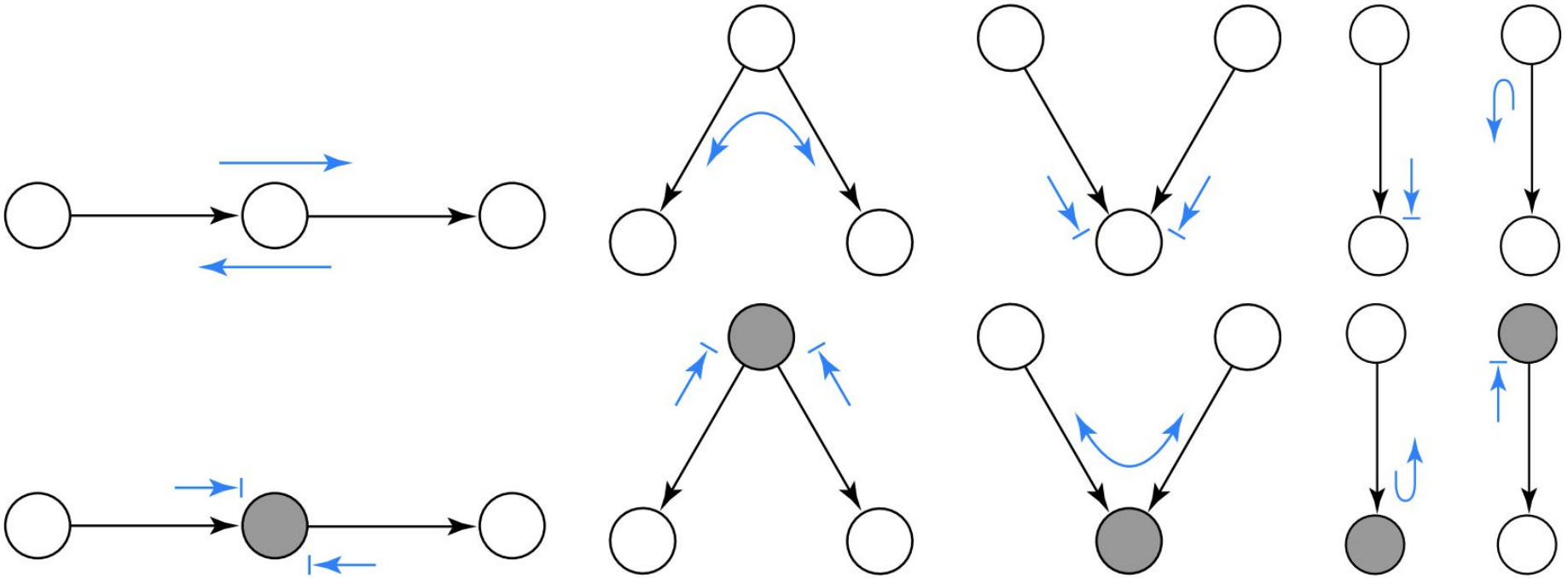
Graphical representation of the relationship



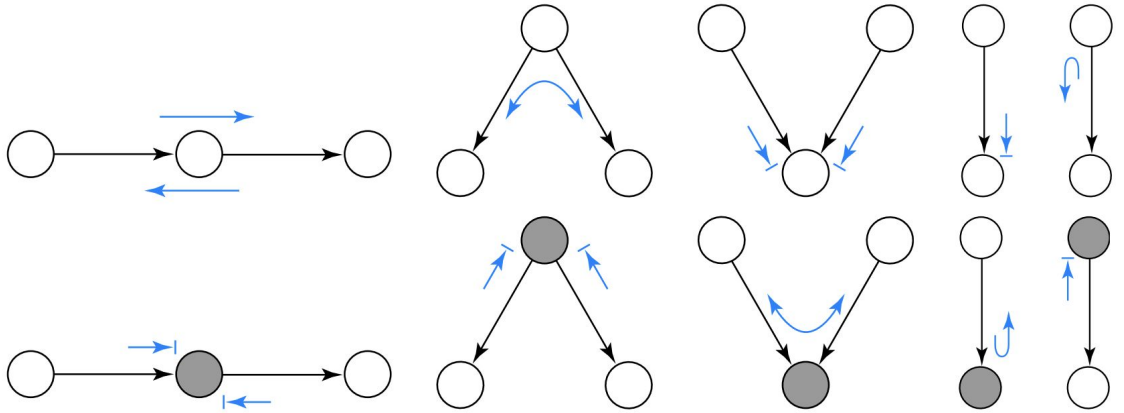
Bayes Ball for independence check

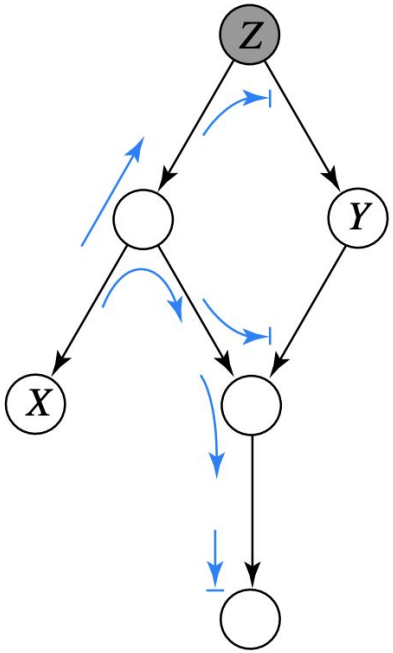
We can “read off” independence and conditional independence assumptions from the graph

X and Y are independent given Z_i if there are no path between them. Z_i are shaded



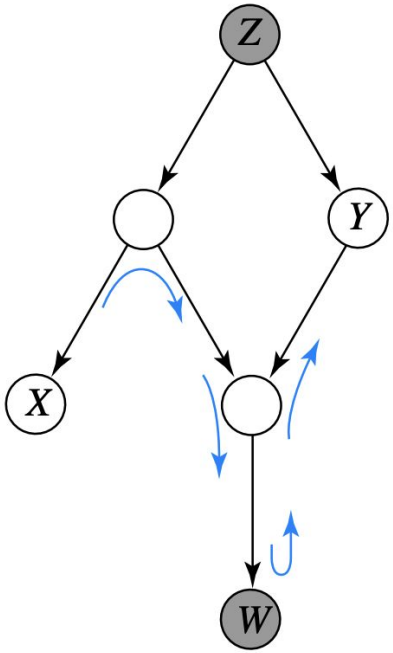
Example





no active paths

$X \perp\!\!\!\perp Y \mid Z$



one active path

$X \not\perp\!\!\!\perp Y \mid \{W, Z\}$

Bayes Ball note

This conditional independence on a directed graph is also called d-separation (direction dependent separation).

Determining the exact graph can be done by exhaustively searching for a graph structure that maximizes the likelihood given the data.

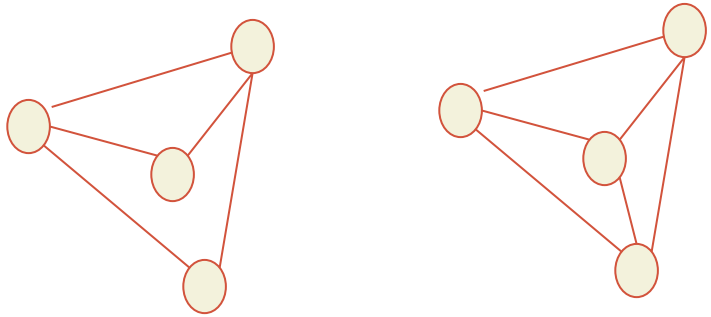
Another method is to use domain knowledge and just draw the graph.

Undirected graphical models

A different view which use undirected graph.

Built from cliques

Cliques - a group of nodes where every pair are adjacent



How many cliques?

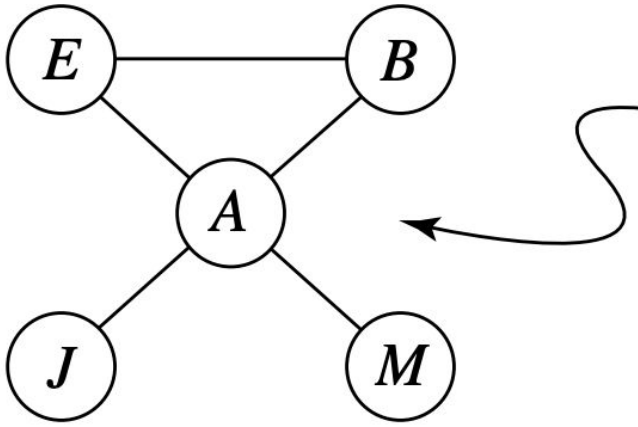
Undirected graphical models

A different view which use undirected graph.

Built from cliques

Cliques - a group of nodes where every pair are adjacent

Each cliques has a potential function of its member which is combined.



Not probability, really just any function

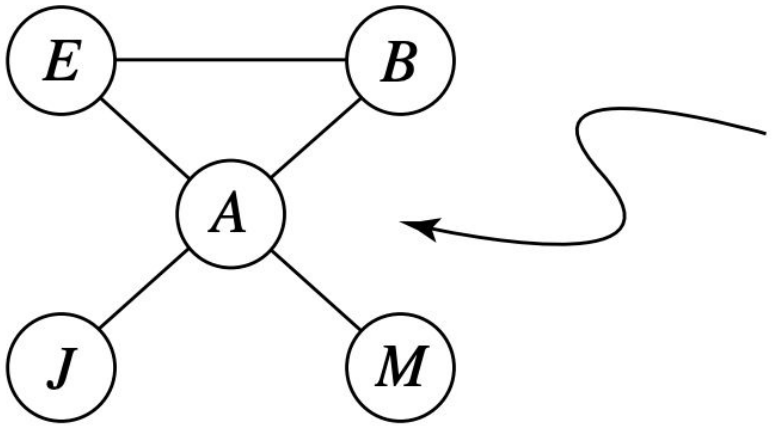
$$p_{EBAJM} = Z^{-1} \times \psi_E \times \psi_B \times \psi_{AEB} \times \psi_{JA} \times \psi_{MA}$$

Scaling factor

Undirected graphical model and independence

Independence in undirected graph is done via graph-separation.

X and Y are independent given Z_i if there is no path between X and Y when we remove all Z_i from the graph



$$p_{EBAJM} = Z^{-1} \times \psi_E \times \psi_B \times \psi_{AEB} \times \psi_{JA} \times \psi_{MA}$$

Comparison between undirected and directed graphical models

Directed - each relationship is a (conditional) probability function

Undirected - just a function, harder to interpret

Directed - can be easily sampled by going from all the edges in order

Undirected - cannot be sampled (discriminative model)

Directed and undirected describes different independencies. Some relationships can only be described in one but not the other.

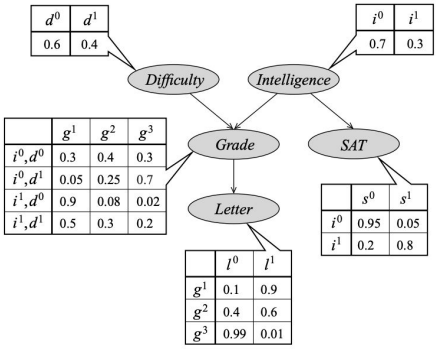


Figure 3.4 Student Bayesian network $\mathcal{B}^{student}$ with CPDs

Observed vs latent

We colorcode the unobserved (latent) variables with white, and observed with black

X - Nationality,
Z - food preference
Y- Weight

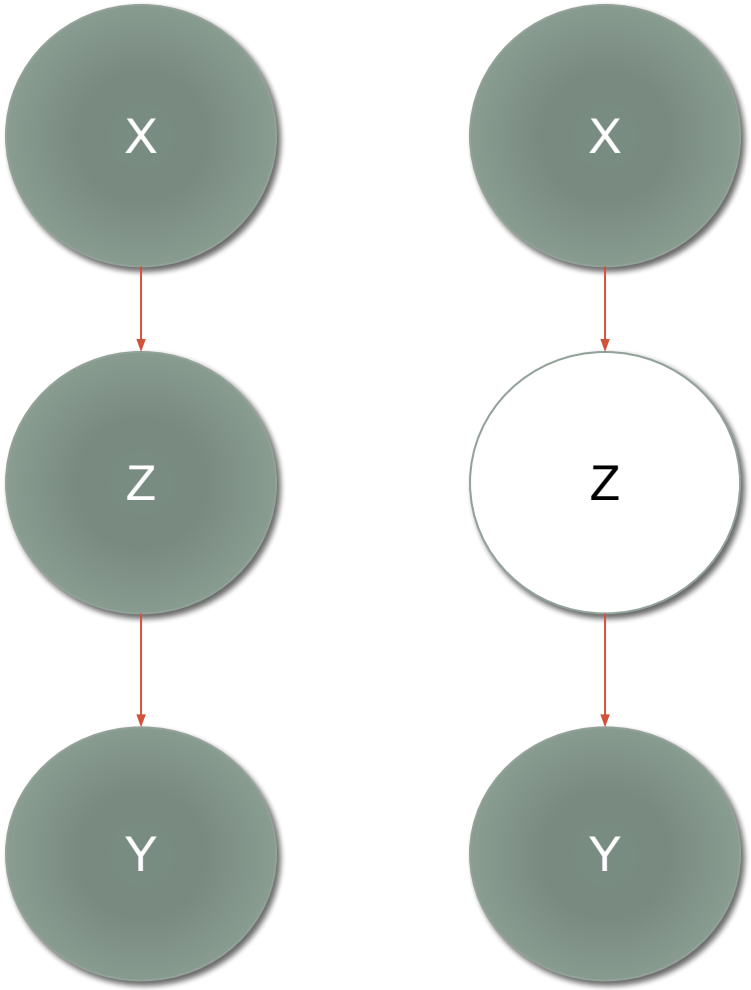


Plate notation

- Sometimes you have too many relationships.

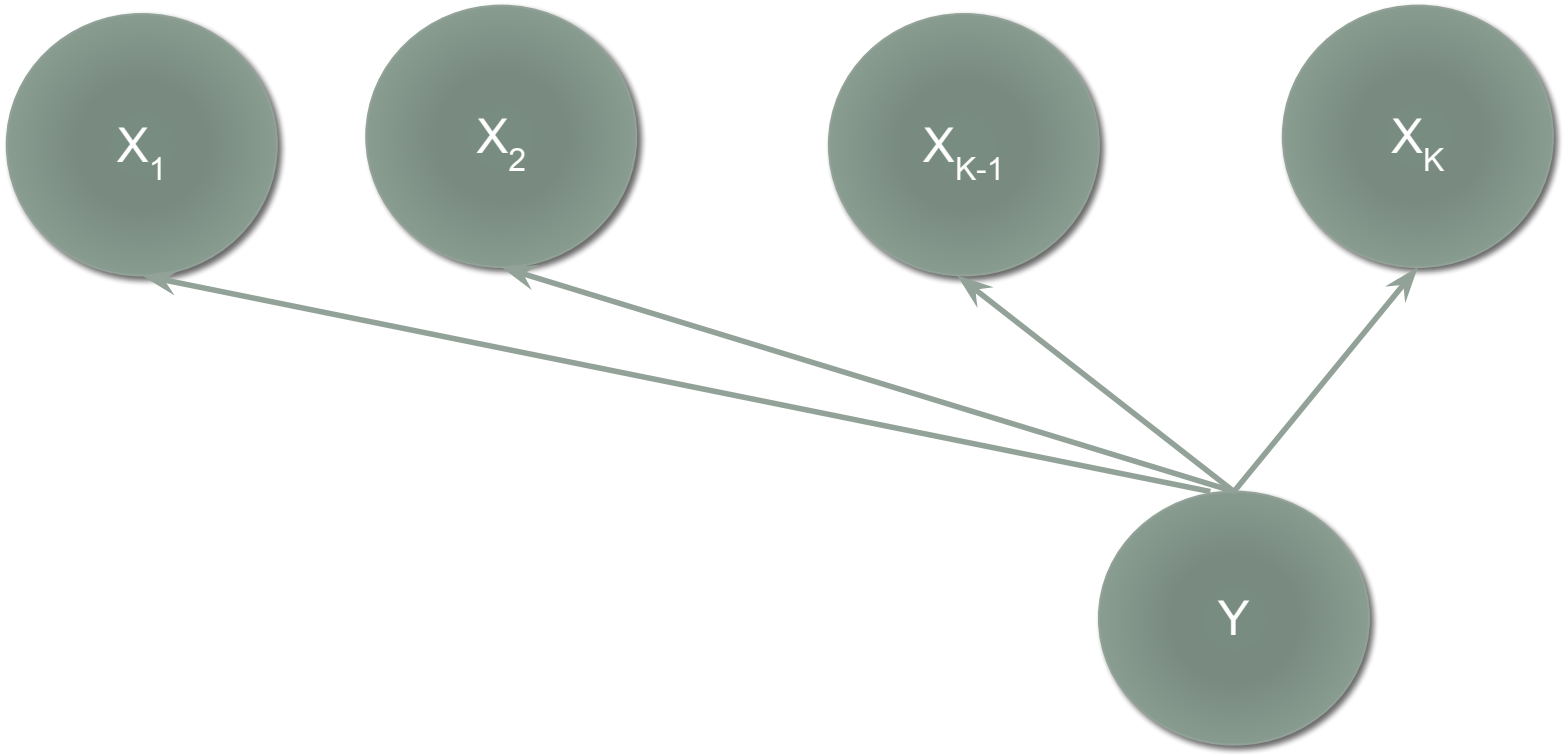
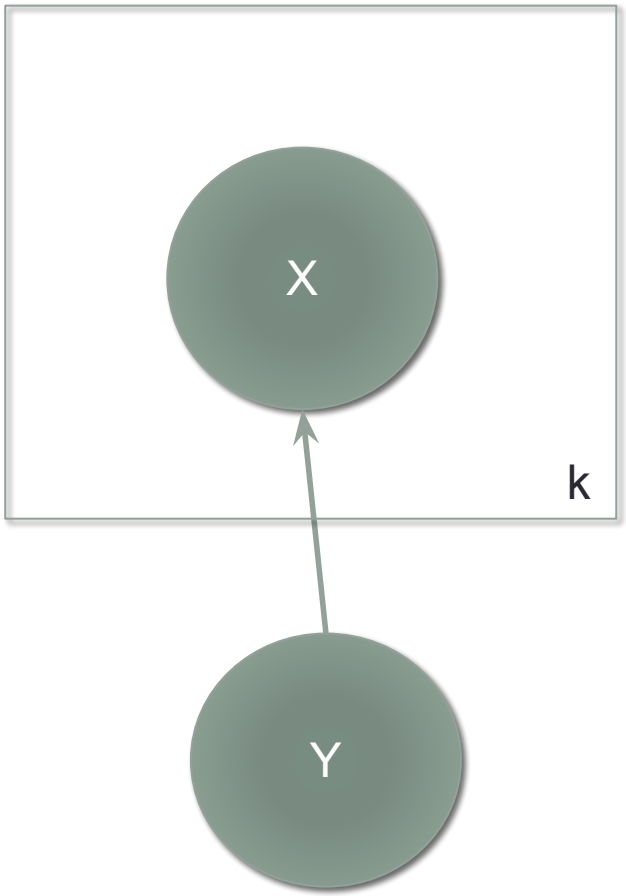


Plate notation

- Summarize by using a square box with number



Topic modeling motivation

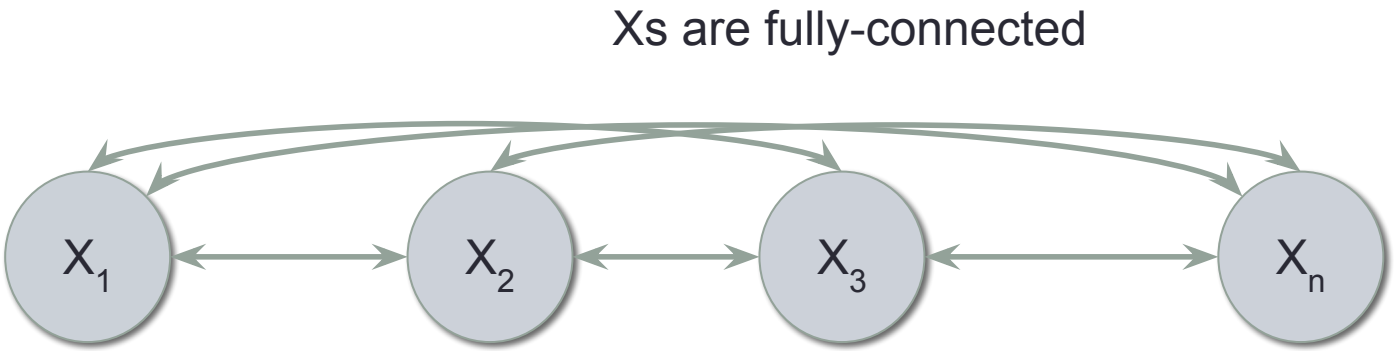
- I want to be able to give probabilities to words/sentences/documents.

- $P(X_1 = \text{Rosses}, X_2 = \text{are}, X_3 = \text{red}, \ldots)$
  - This probability distribution is intractable

- Assume independence
  - $P(X_1)\ P(X_2)\ P(X_3)\ldots$
  - Lost all relational structure. Can we do better?
    - Assume words come from topics. How?

Graphical view of language modeling

- P("Roses are red. Violets are blue.") = ?

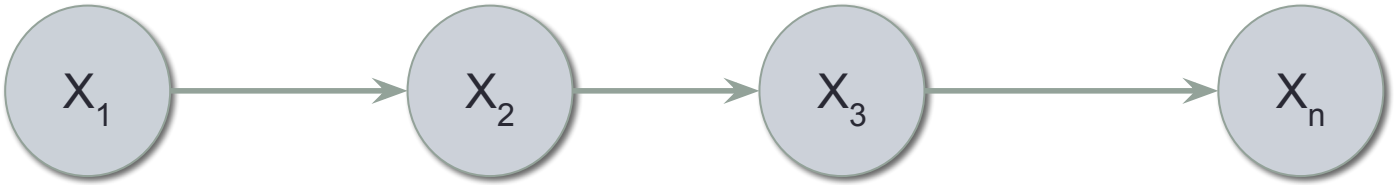


$P(X_1 = \text{Roses}, X_2 = \text{are}, \ldots, X_n = \text{blue})$

Graphical view of language modeling

- P("Roses are red. Violets are blue.") = ?

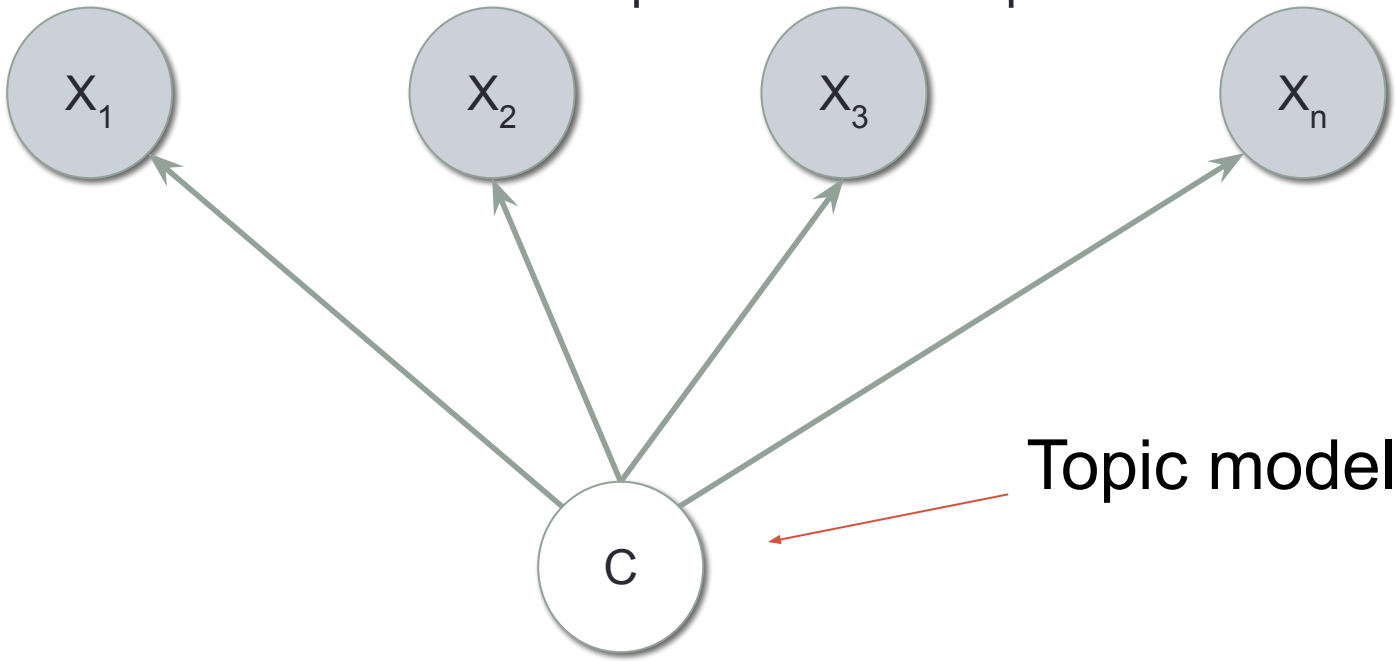
With markov assumption



$P(X_1 = \text{Roses})P(X_2 = \text{are} \mid X_1 = \text{Roses})\ P(X_3 = \text{red} \mid X_2 = \text{are}) \ldots P(X_n = \text{blue} \mid X_{n-1} = \text{are})$

Graphical view of language modeling

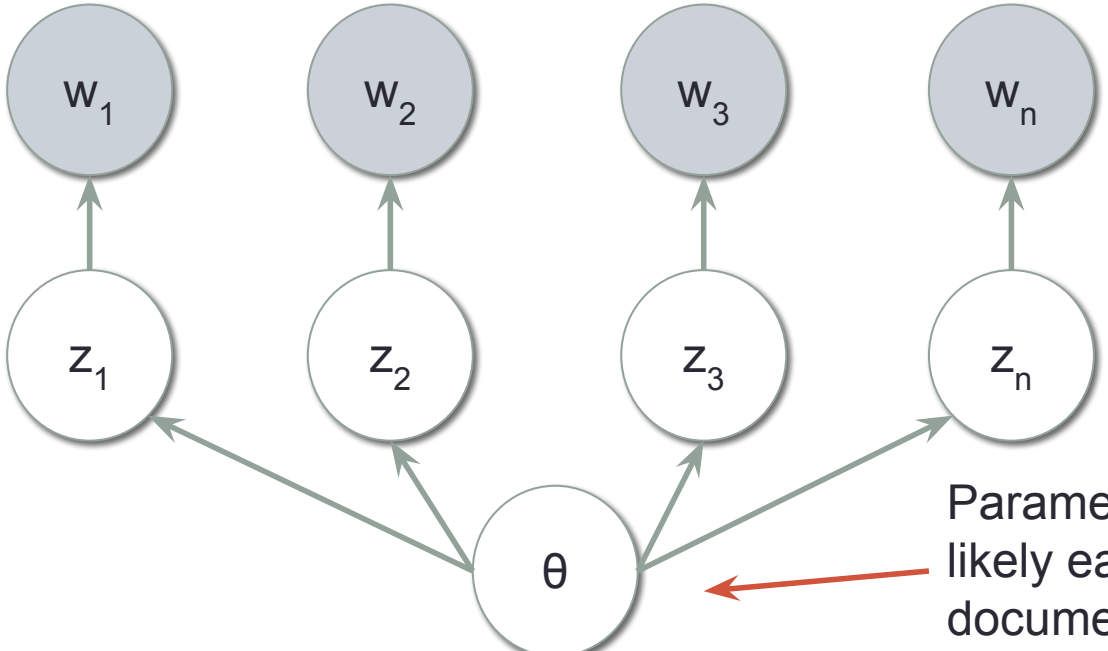
- P("Roses are red. Violets are blue.") = ?
- Let's assume words are generated from a topic
  - Sentence comes from topic C = Shakespeare



$P(X_1 = \text{Roses} \mid c = \text{Shakespeare}) P(X_2 = \text{are} \mid c = \text{Shakespeare}) \ldots$
$P(X_n = \text{blue} \mid c = \text{Shakespeare}) P(c = \text{Shakespeare})$

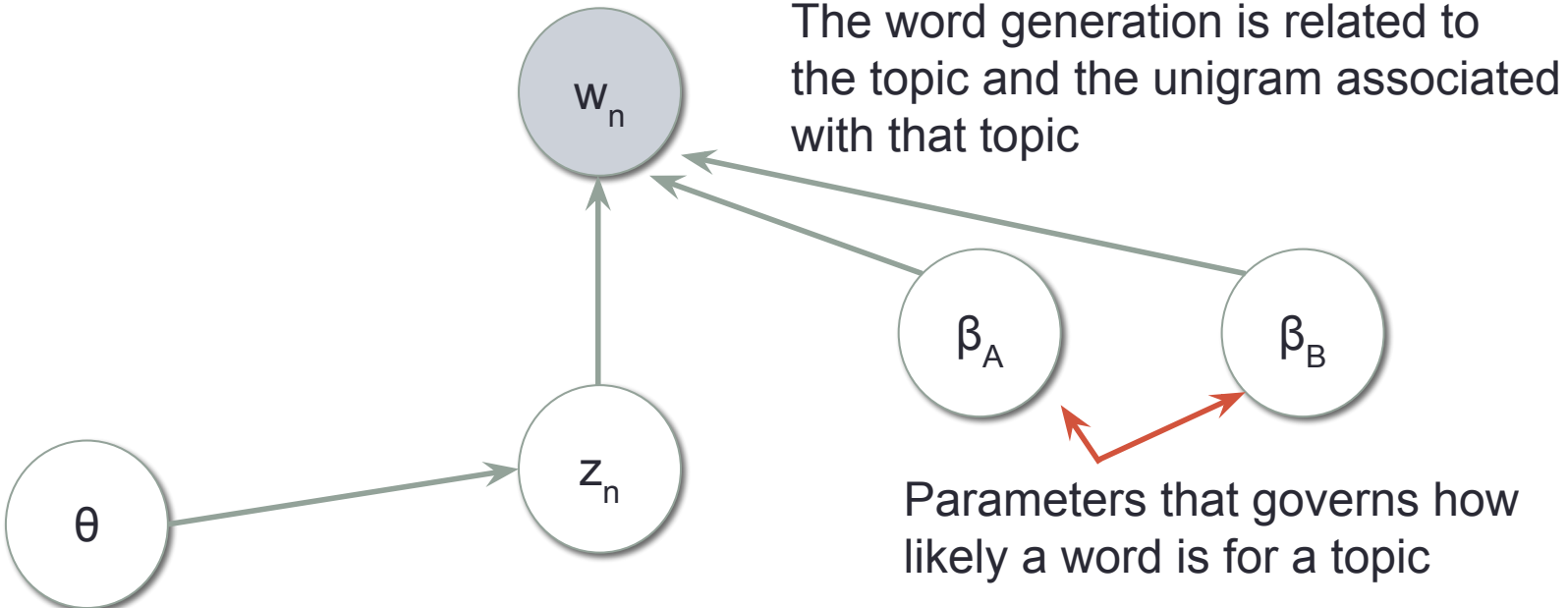
Multiple topics

- Most document have multiple topics. Our model assumes 1 document 1 topic.
  - Let a document be a mixture of topics. Each word has its own topic, z.
    - P(w) = P(z = A) P(w | z = A) + P(z = B) P(w | z = B)
    - P(z = A) + P(z = B) = 1, θ = P(z = A)



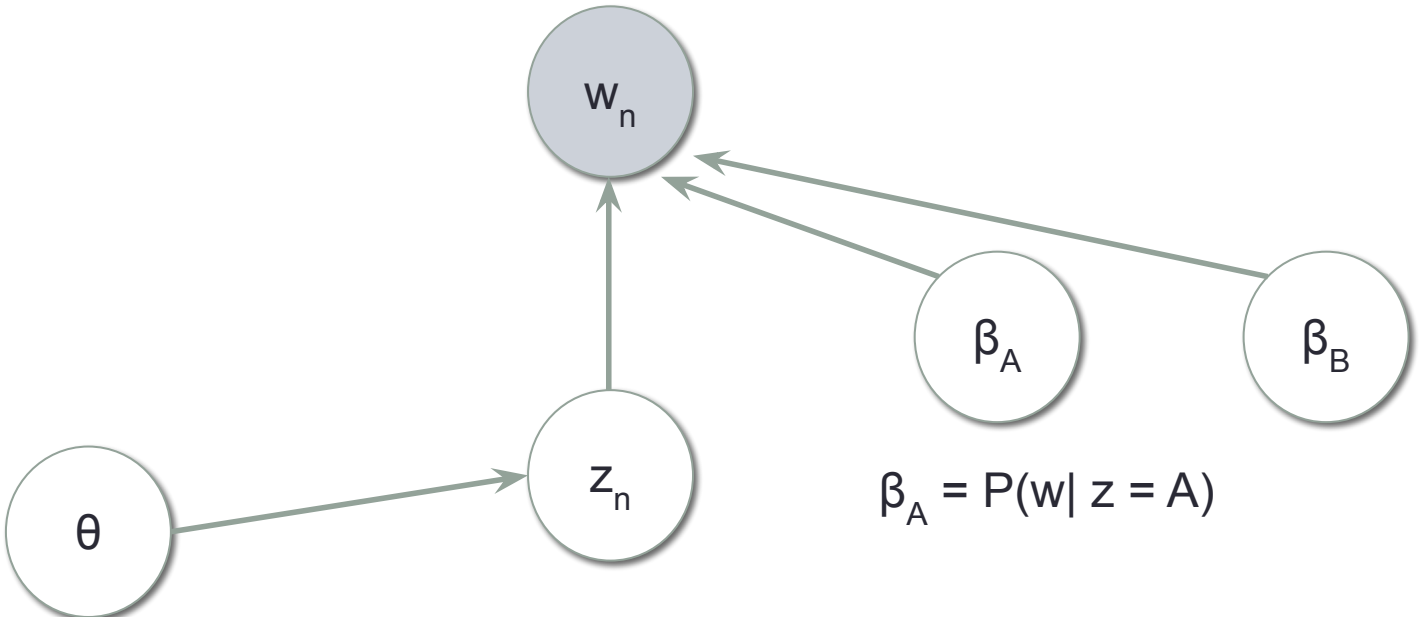
Topic modeling

- Most document have multiple topics. Our model assumes 1 document 1 topic.
  - Let a document be a mixture of topics (language model interpolation). Each word has its own topic, z.
    - P(w) = P(z = A) P(w | z = A) + P(z = B) P(w | z = B)
    - P(z = A) + P(z = B) = 1, $\theta = P(z = A)$ $\beta_A = P(w| z = A)$



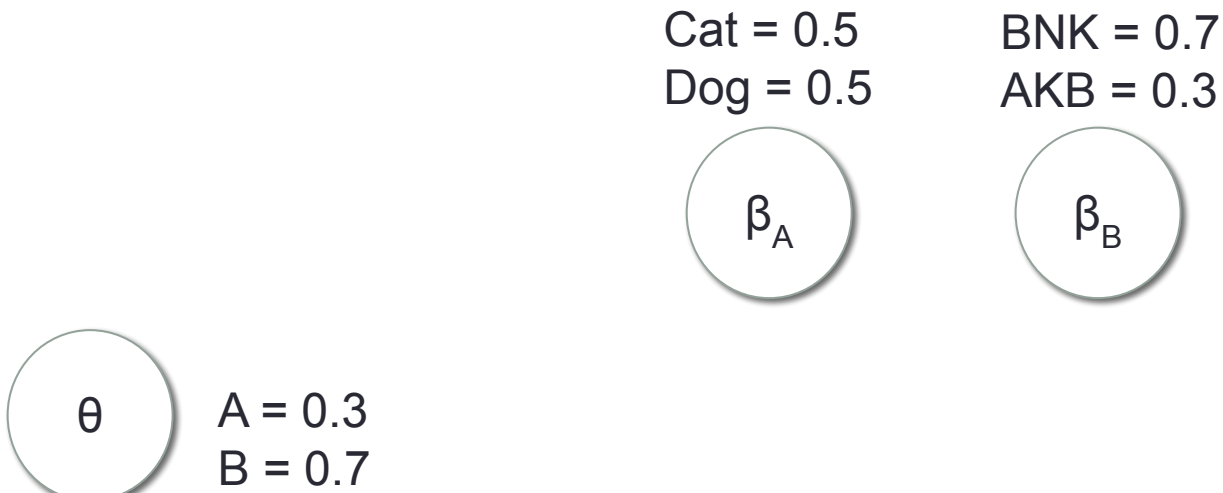
Graphical model and generation

- You can generate a sample from a graphical model by following the arrows



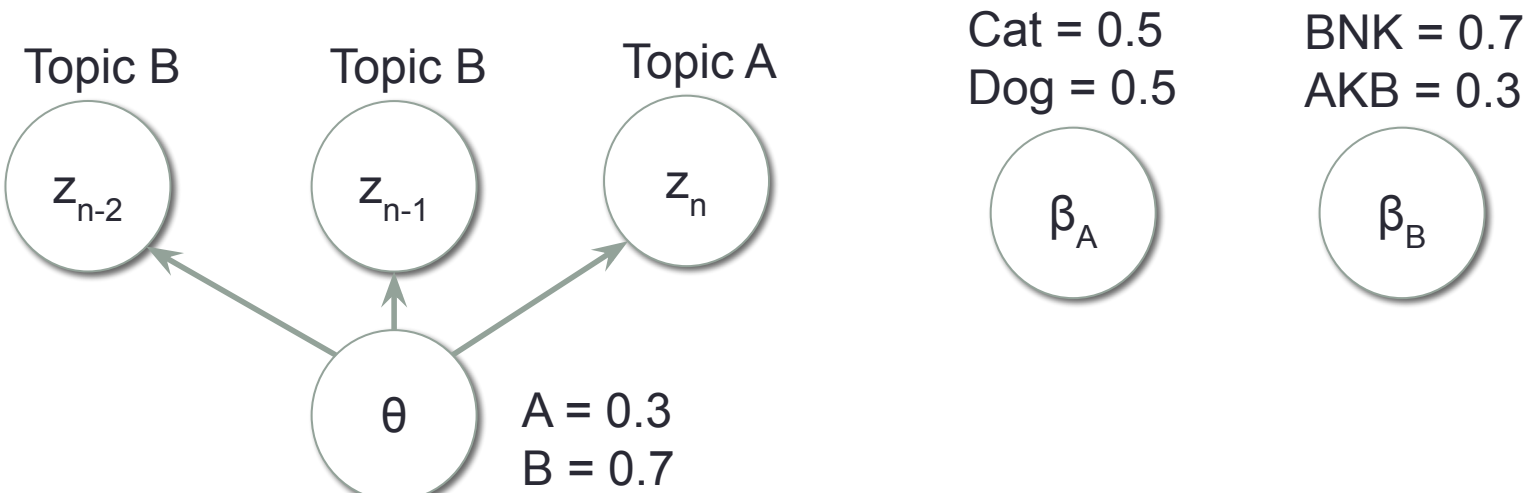
Graphical model and generation

- Given θ, β_A, β_B



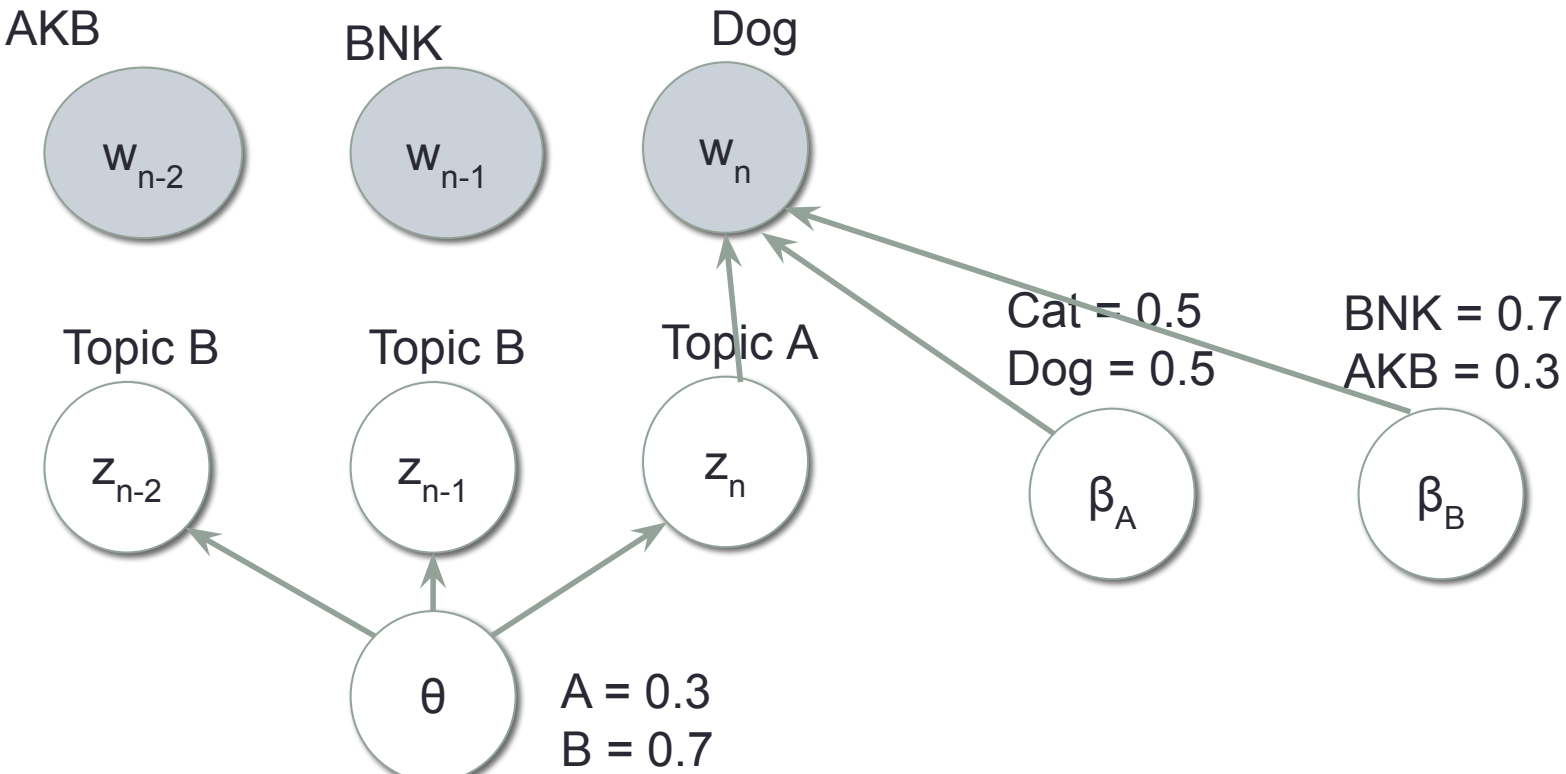
Graphical model and generation

- Given θ, β_A, β_B
- Generate (randomly create) z from θ



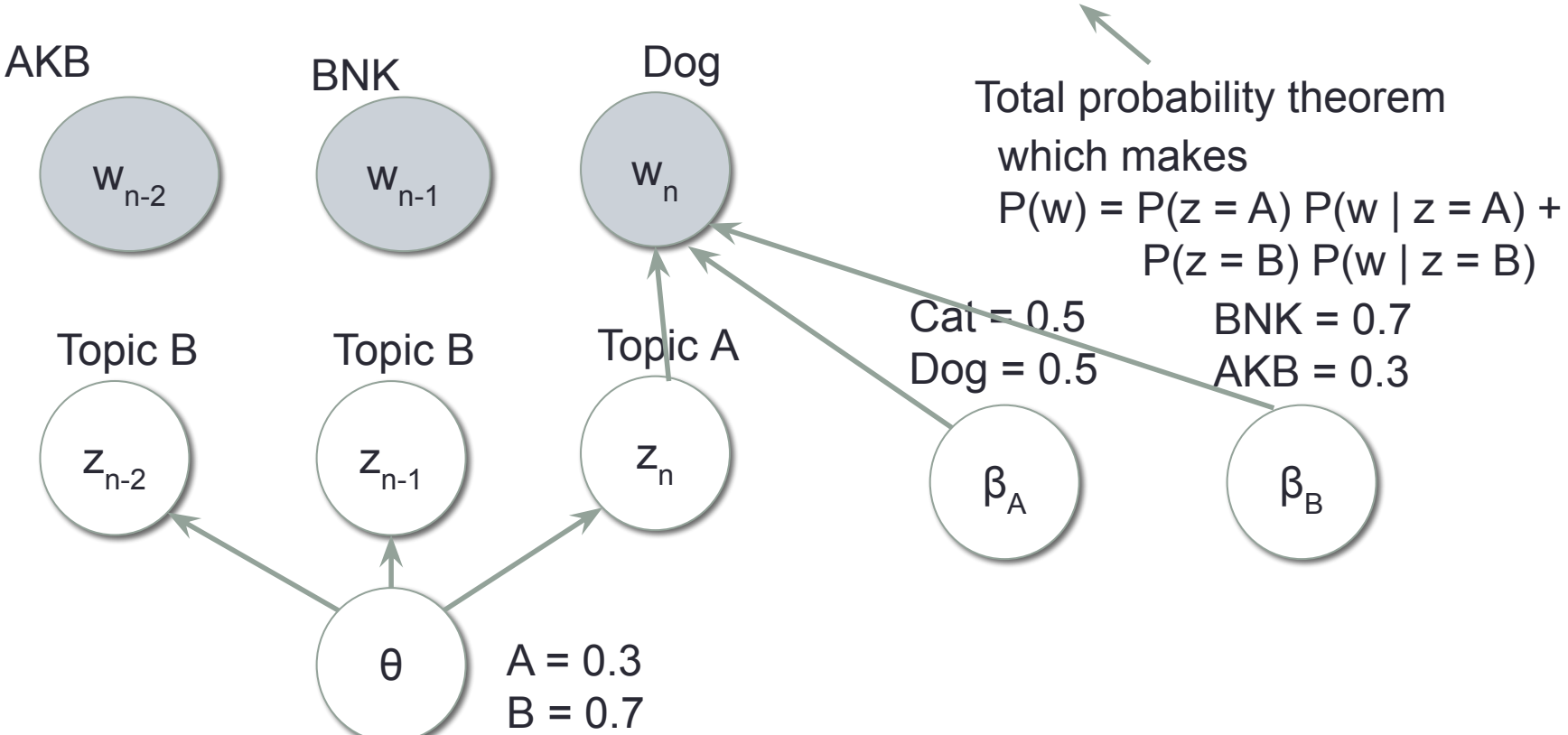
Graphical model and generation

- Given θ, β_A, β_B
- Generate (randomly create) z from θ
- Generate w_n from z_n β_A, β_B

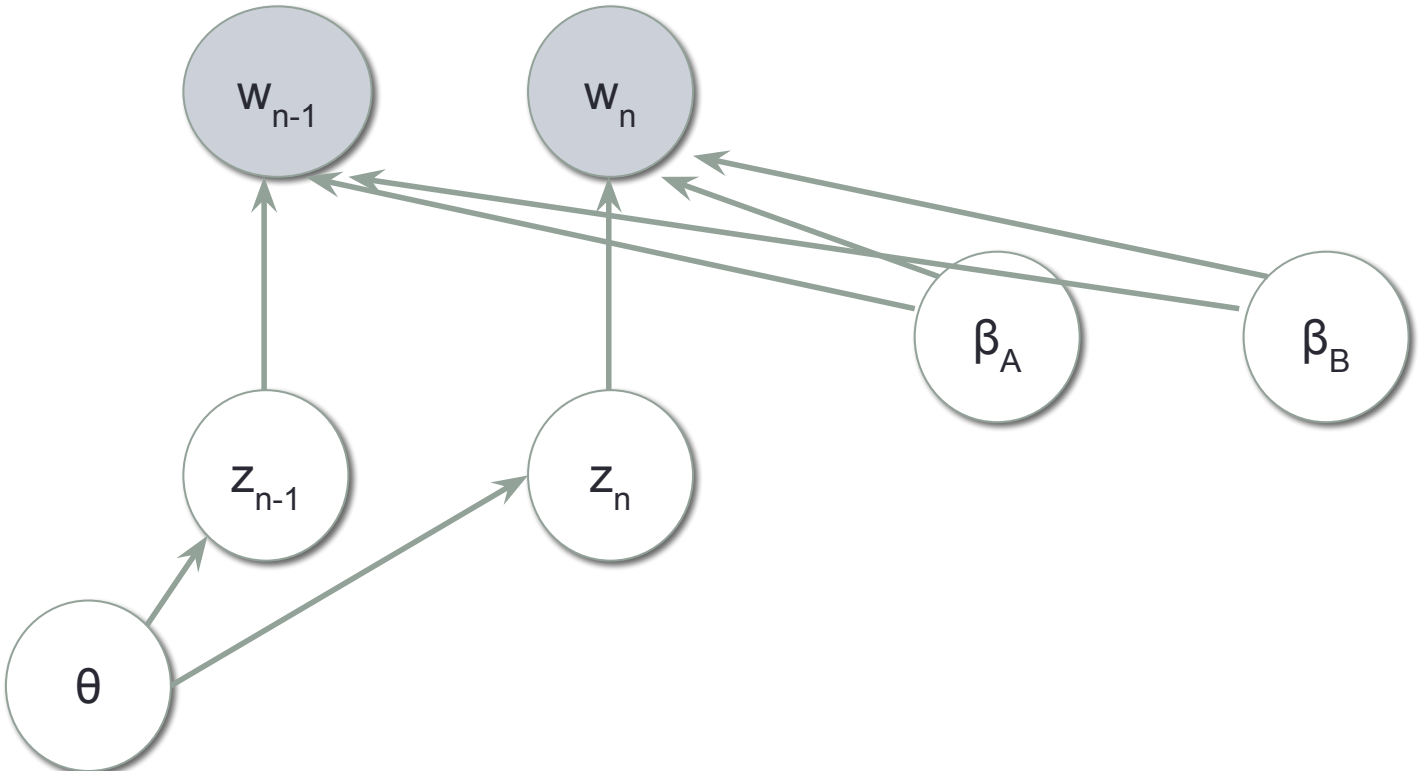


Graphical model and generation

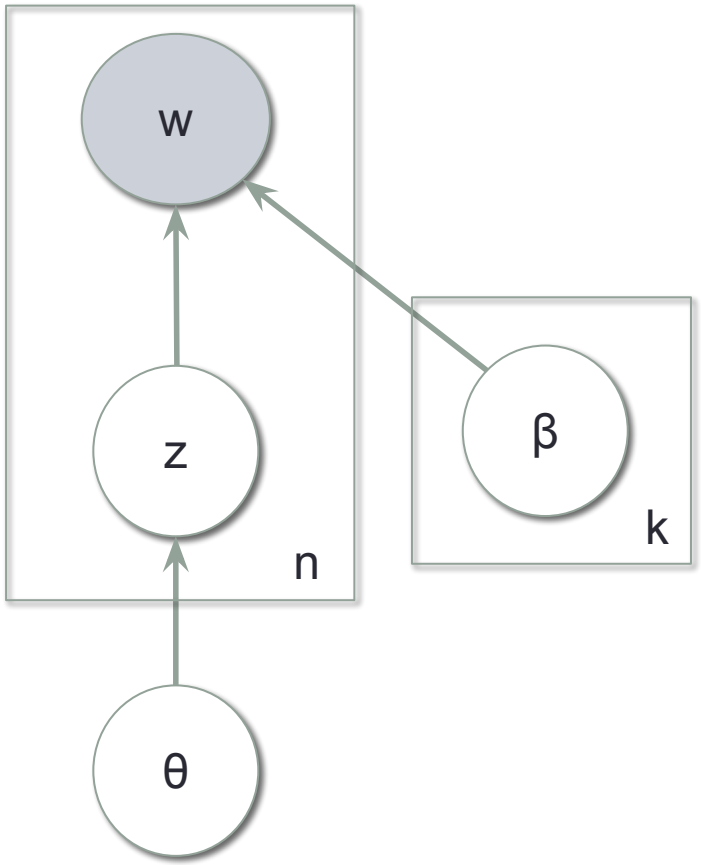
How likely a sentence is likely to be generated follows this generation process
P(AKB,BNK,Dog,B,B,A) = P(B)P(B)P(A)P(AKB|B)P(BNK|B)P(Dog|A)
Note
P(AKB,BNK,Dog) = P(AKB,BNK,Dog,A,A,A) + P(AKB,BNK,Dog,A,A,B)
P(AKB,BNK,Dog,A,B,A) + P(AKB,BNK,Dog,A,B,B) +...



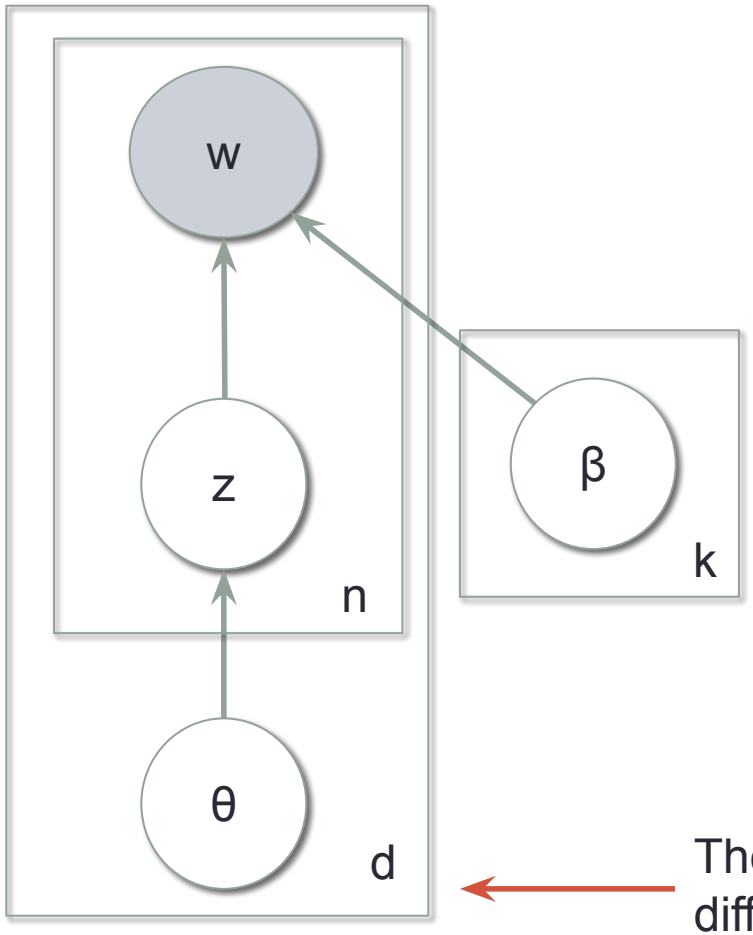
Topic modeling with plate notation



Topic modeling with plate notation

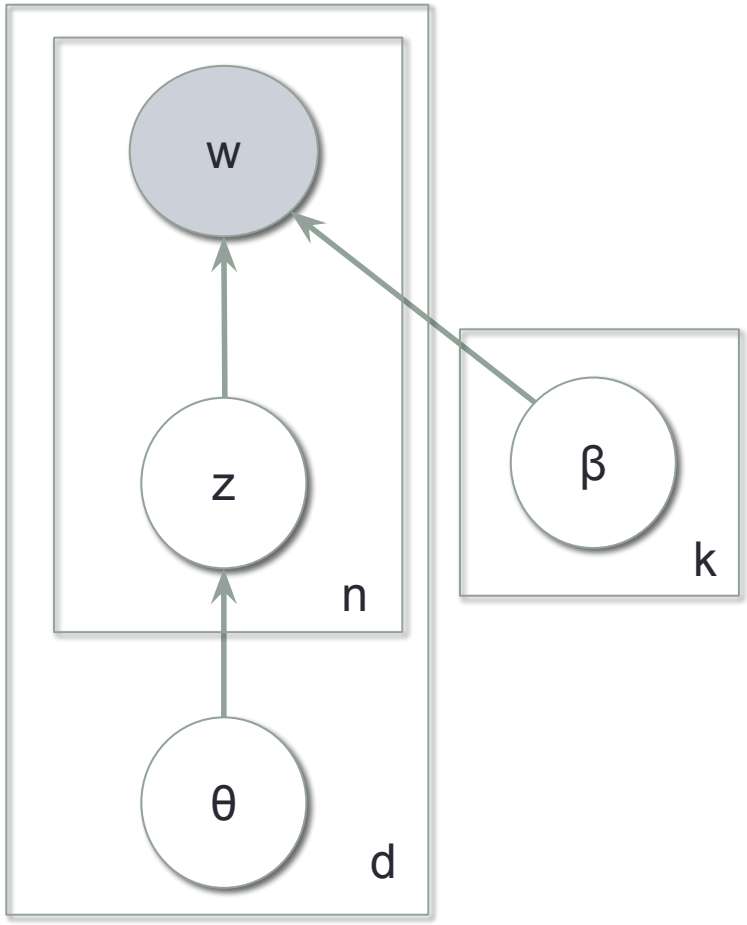


Topic modeling with plate notation



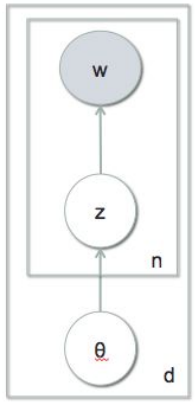
There are d documents each with different topic distributions

Topic modeling with plate notation



Called pLSA
probabilistic Latent Semantic Analysis

Note: if you look at other textbooks, you will see a slightly different picture



Learning topic latent model parameters

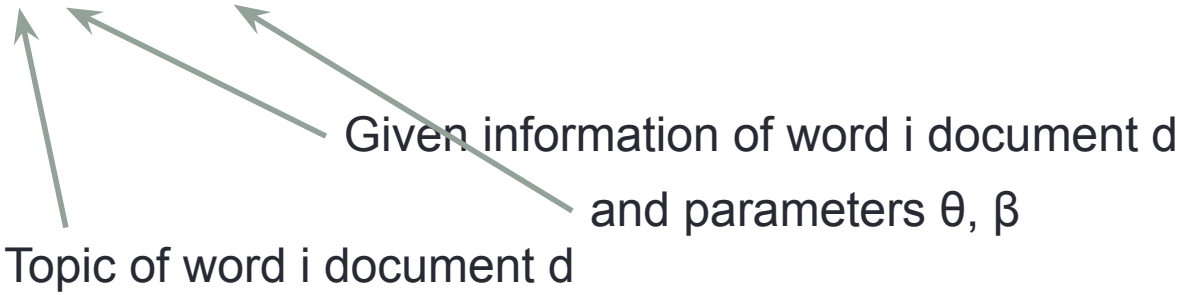
- How to find θ and β?
  - “Cat,Dog,BNK”, “Cat, cat, cat, BNK”, “Dog, dog, BNK, dog, AKB”
- If we know, the latent topic z for each word we can use the counts
  - “Cat_A, Dog_A, BNK_B”
- $P(Cat|A) = \frac{count(Cat\_A)}{count(A)}$
- $P_1(A) = \frac{count(A)}{count(\text{all words in document 1})}$
- But we don’t know the topic z for each word

Expectation maximization (EM)

- A method to iteratively maximize the likelihood of a model on training data
  - Expectation step (E step) – guess latent variables from model parameters (get soft counts)
  - Maximization step (M step) – re-estimate model parameters from latent variables (counts)

E-step

- Find an estimate for the latent variable given parameters θ, β
- $p(z_{di} \mid w_{di}, \theta, \beta)$



E-step

- Find an estimate for the latent variable given parameters θ, β

$$p(z_{di}|w_{di},\theta,\beta) = \frac{p(z_{di},w_{di},\theta,\beta)}{\sum_{z'=1}^{k} p(z'_{di},w_{di},\theta,\beta)}$$

p(w, θ, β)

E-step

- Find an estimate for the latent variable given parameters θ, β

$$p(z_{di}|w_{di},\theta,\beta) = \frac{p(z_{di},w_{di},\theta,\beta)}{\sum_{z'=1}^{k} p(z'_{di},w_{di},\theta,\beta)}$$

$$= \frac{\theta_{z|d}\beta_{w|z}}{\sum_{z'=1}^{k} \theta_{z'|d}\beta_{w|z'}}$$

Index di for z and w dropped for clarity

E-step

- Find an estimate for the latent variable given parameters θ, β

$$p(z_{di}|w_{di},\theta,\beta) = \frac{p(z_{di},w_{di},\theta,\beta)}{\sum_{z'=1}^{k} p(z'_{di},w_{di},\theta,\beta)}$$

$$= \frac{\theta_{z|d}\beta_{w|z}}{\sum_{z'=1}^{k} \theta_{z'|d}\beta_{w|z'}}$$

P(Word is from topic A| word is cat from document 1)
Probability that the word is from each topic. Use as counts

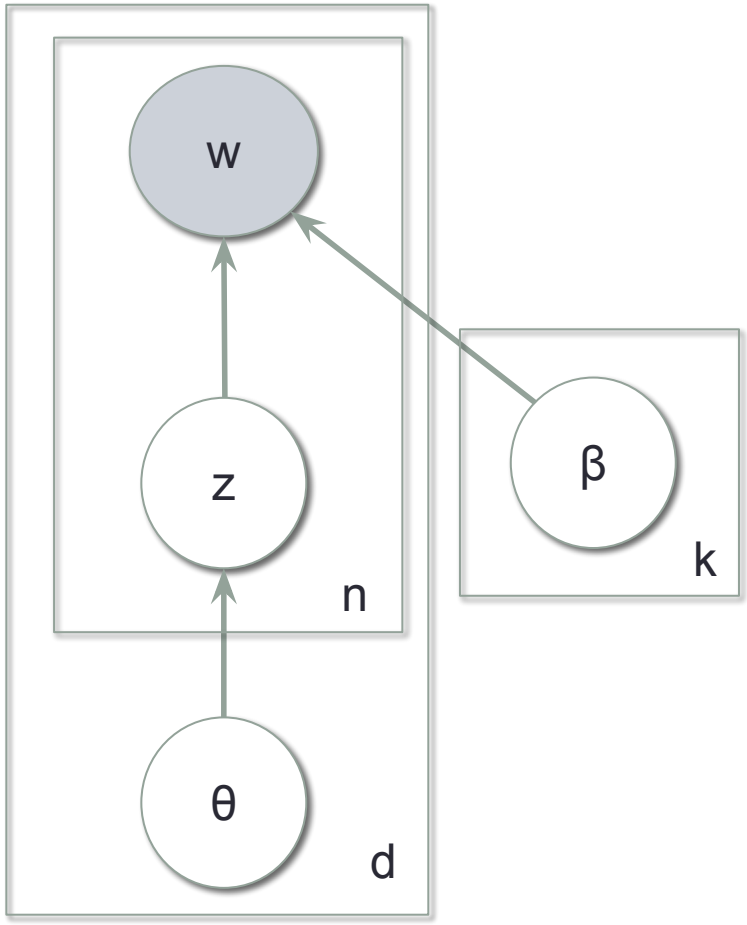
M-Step

- Instead of real counts use $P(z_{di})$ as the topic label
  - $P(Cat|A) = \frac{count(Cat\_A)}{count(A)}$
  - $P_1(A) = \frac{count(A)}{count(all\ words\ in\ document\ 1)}$

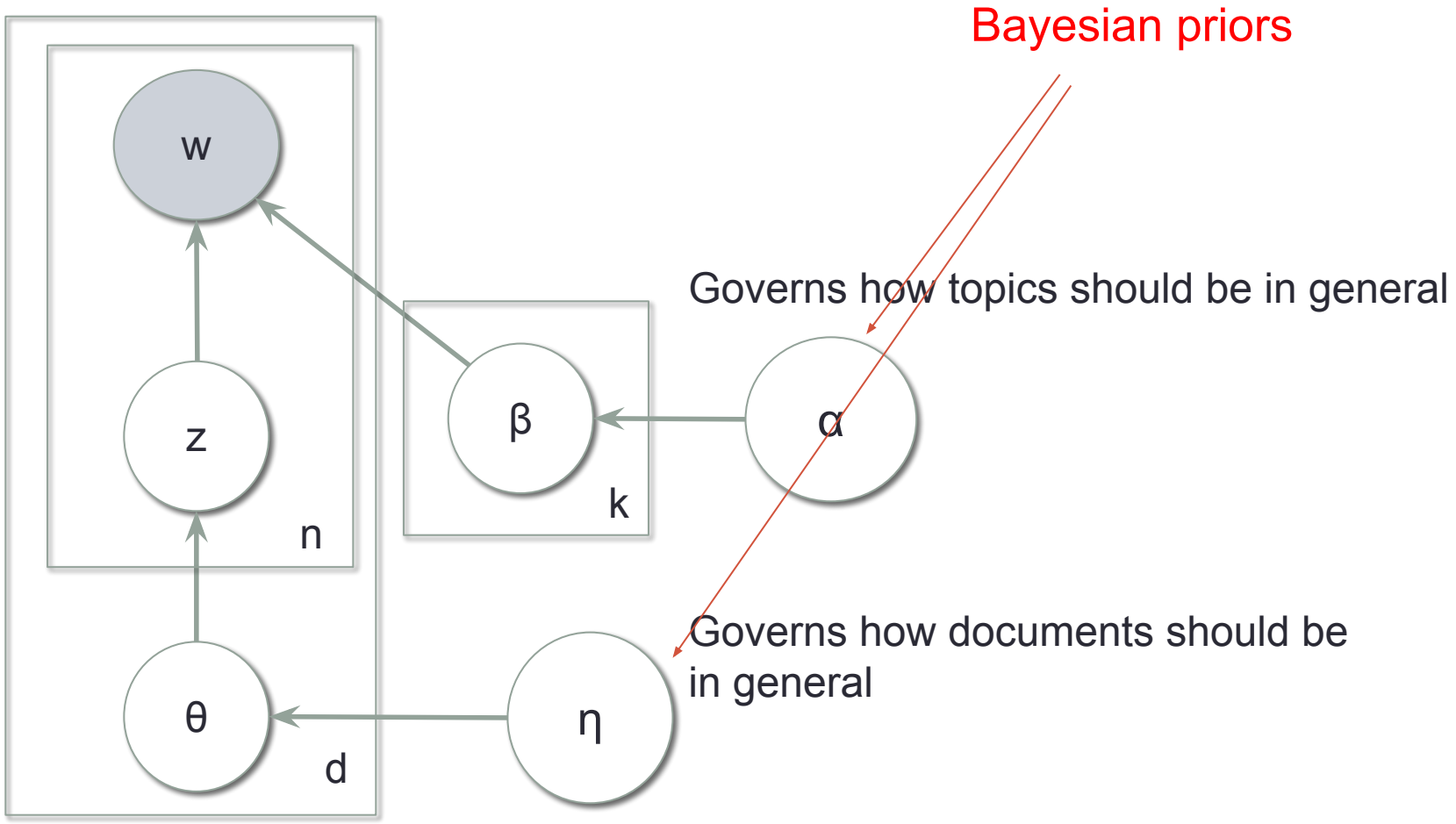
pLSA

- pLSA automatically clusters words into topic unigrams
  - Requires user to specify number of topics
- Automatically learn document representation based on the learned topics
  - DocA = [0.7 0.3] DocB = [0.2 0.8] DocC = [0.5 0.5]
- Overfits easily to data outside of the training set
  - Nothing that ties all document together
  - A document from a document collection should be have topic distributions that are similar
- Solution: LDA (Latent Dirichlet Allocation)

pLSA



LDA


Bayesian priors
w
z
n
β
k
α
Governs how topics should be in general
θ
d
η
Governs how documents should be in general

α and η

- α is a Dirichlet distribution
- Or a distribution of distributions…

- Example: rolling a die
- P(x=1) = 0.3, P(x=2) = 0.1,…
- This is a distribution. (Multinomial distribution)
- But what if I have a bag of dices, each with different distributions.
- α tells me the probability of what kind of die I will get

Dirichlet distribution

- pdf

Parameters giving preference to each side of die/topic

$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

Probability of each side of die, probability of each topic

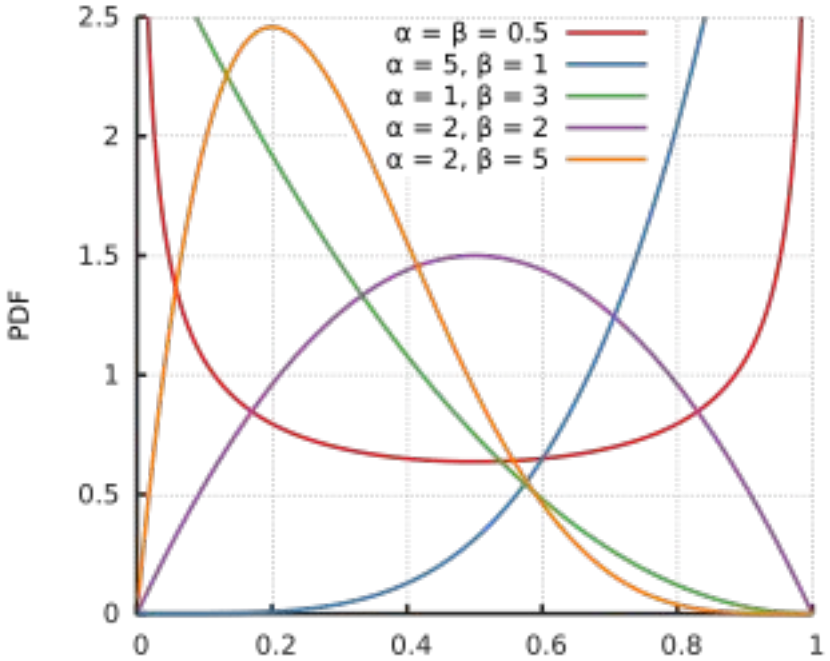
Dirichlet distribution

- pdf

Parameters giving preference to each side of die/topic

$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

Probability of each side of die, probability of each topic

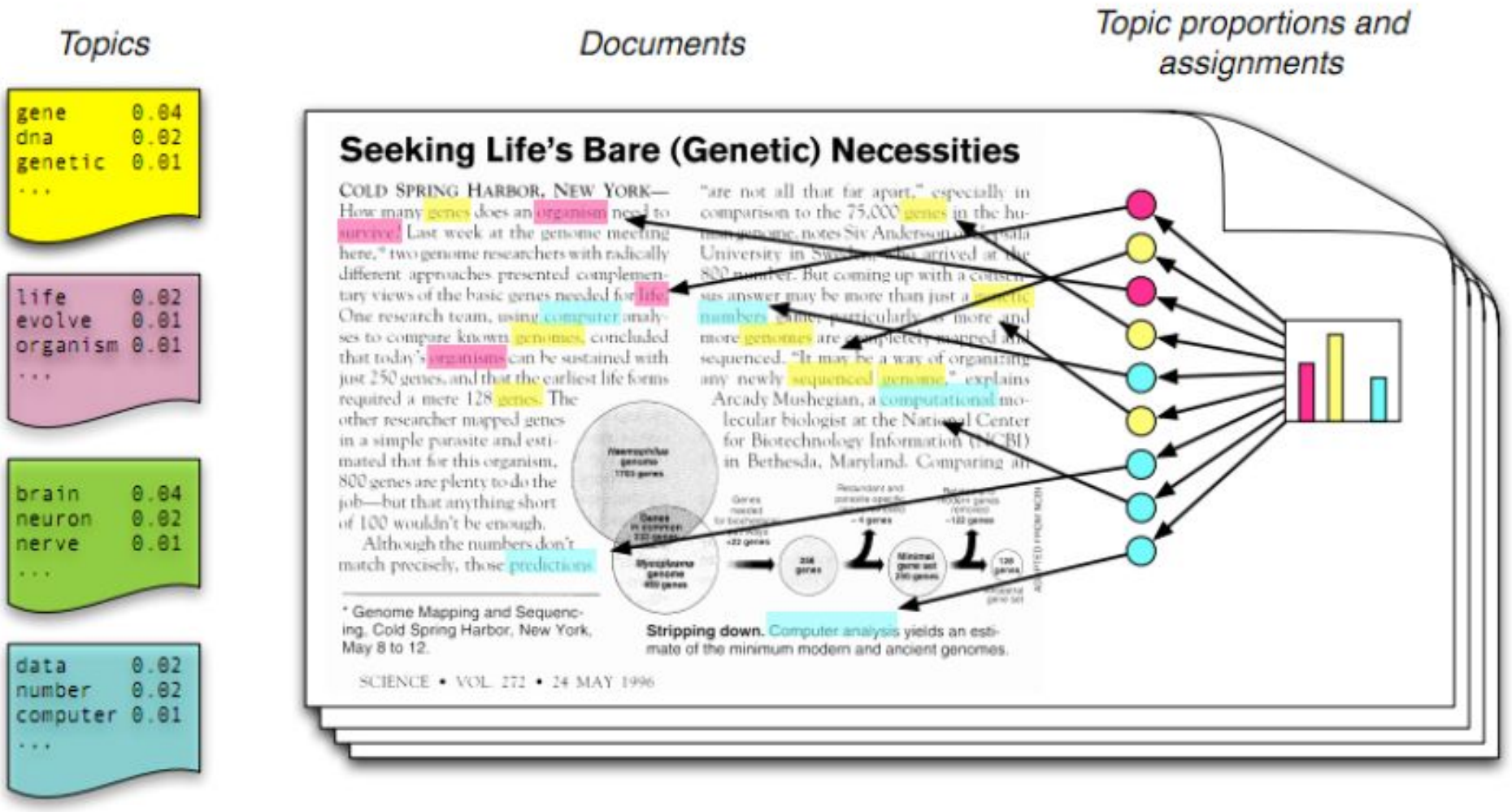


$$p(x; \alpha, \boldsymbol{\beta}) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Beta distribution
A prior for binomial trails

Parameter learning

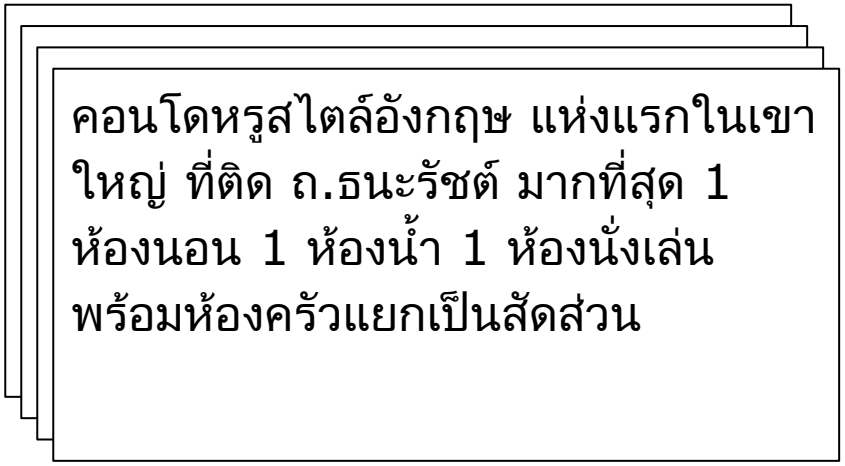
- How do we actually learn these parameters and latent variables?
  - Gibbs sampling
  - Variational methods



Introduction to Probabilistic Topic Models, Blei 2011
http://menome.com/wp/wp-content/uploads/2014/12/Blei2011.pdf

Unsupervised topic modeling for real estate

Can we learn real estate characteristics from unstructured data?

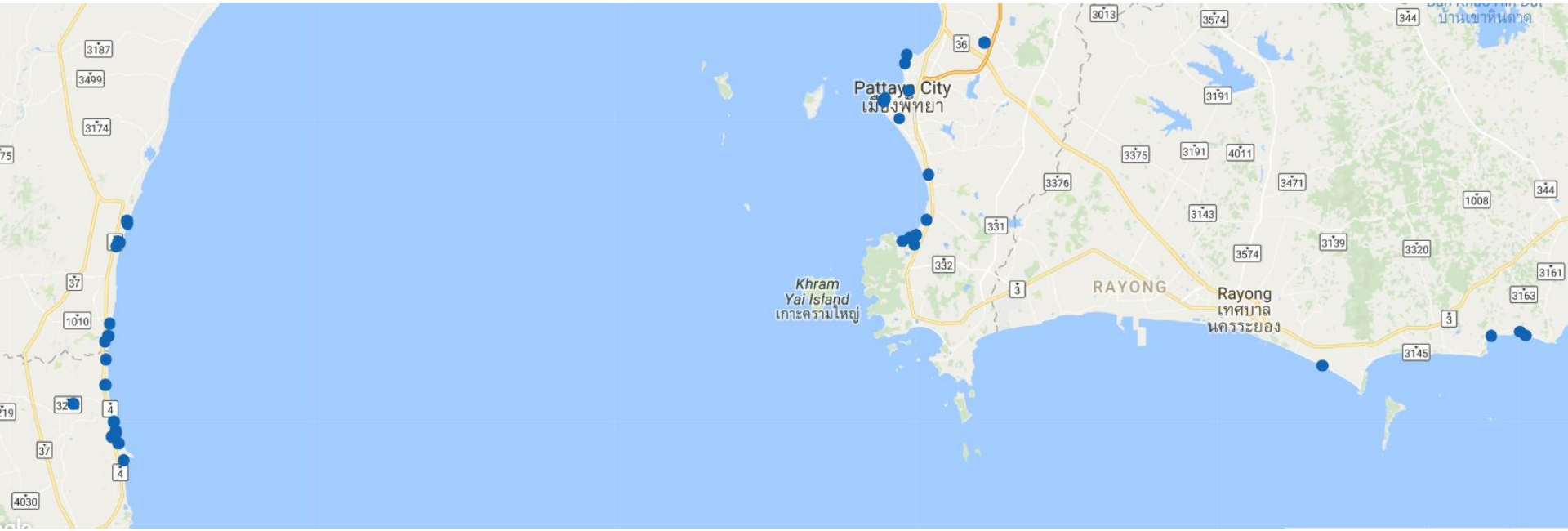


Just give it a bunch of descriptions

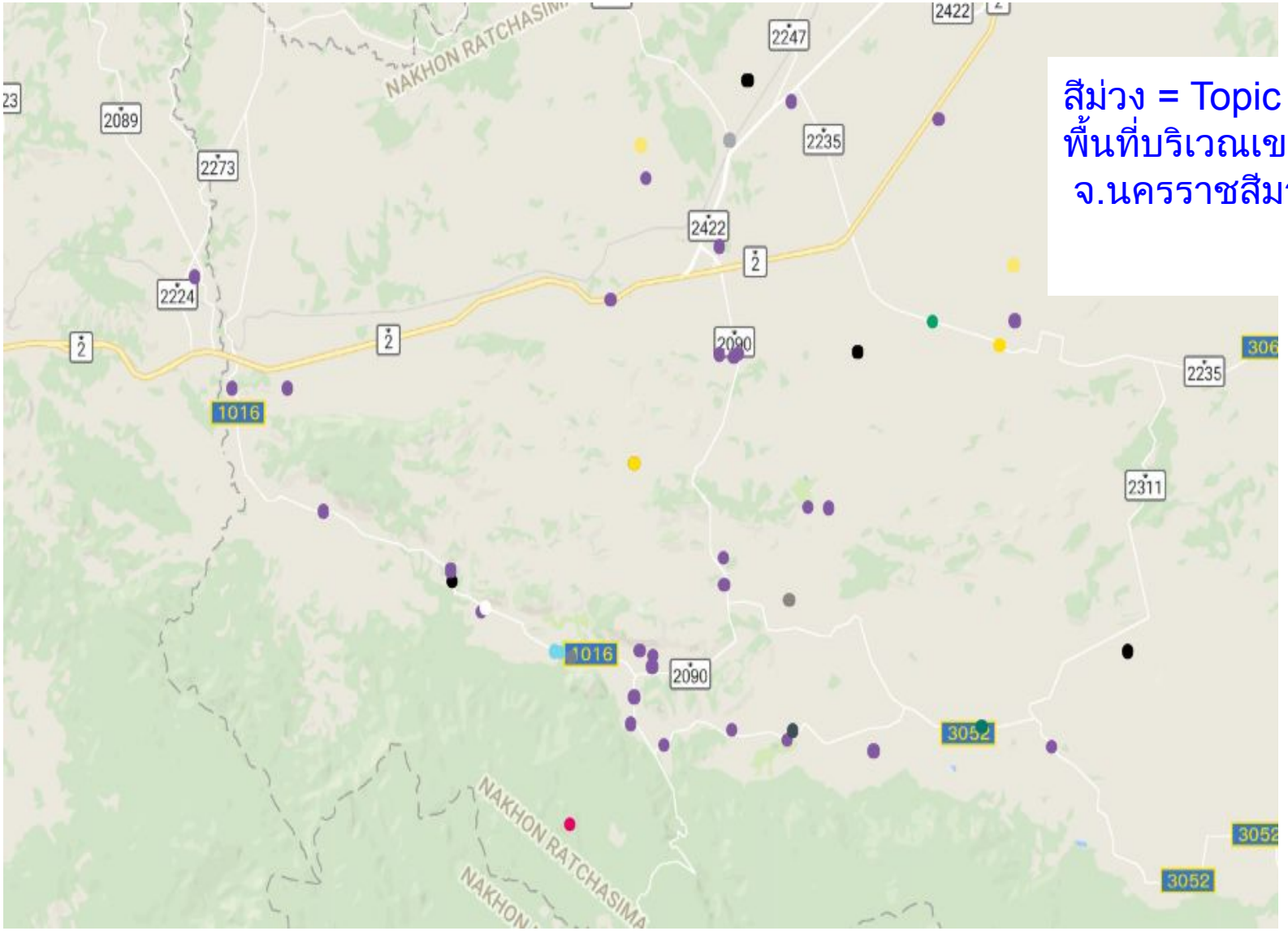
HOME DOT TECH

LDA Examples

Topic 28
0.068*"วิว" + 0.058*"ทะเล" + 0.038*"คอนโด" + 0.029*"หัว" + 0.027*"คอนโดมิเนียม" + 0.025*"มองเห็น" + 0.023*"ทัศนียภาพ" + 0.022*"ชายหาด"



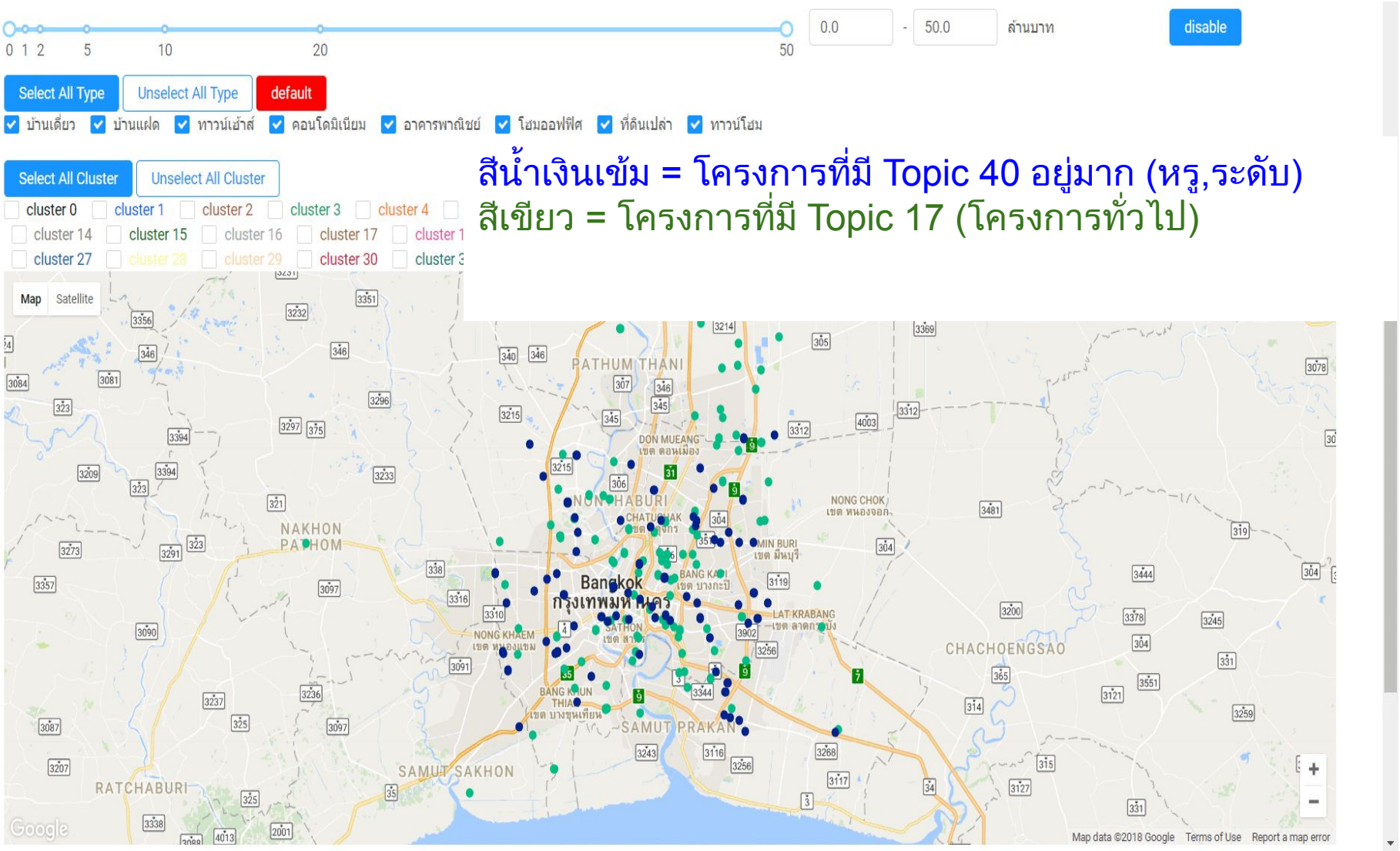
Topic 9
0.071*"ธรรมชาติ" + 0.031*"บรรยากาศ" + 0.028*"ร่มรื่น" + 0.027*"บ้าน" + 0.025*"ท่ามกลาง" + 0.025*"สวน" + 0.025*"สัมผัส" + 0.021*"พื้นที่"



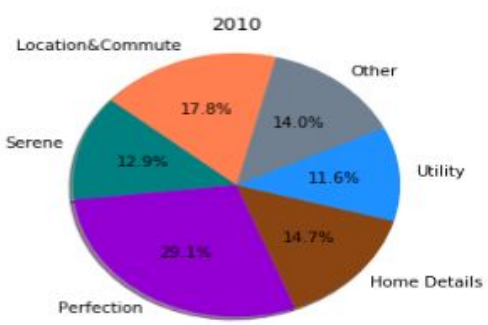
สีม่วง = Topic 9
พื้นที่บริเวณเขาใหญ่
จ.นครราชสีมา

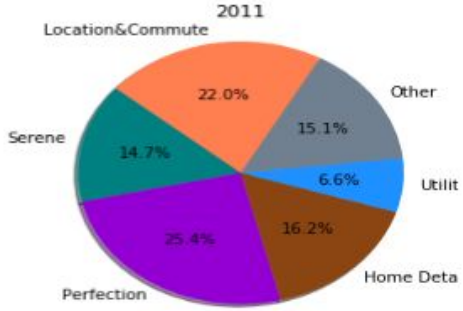
Topic 40
0.115*"ระดับ" + 0.066*"เหนือ" + 0.046*"หรู" + 0.031*"ทำเล" + 0.026*"ชีวิต" + 0.026*"ใช้ชีวิต" + 0.016*"สไตล์" + 0.016*"สะท้อน"

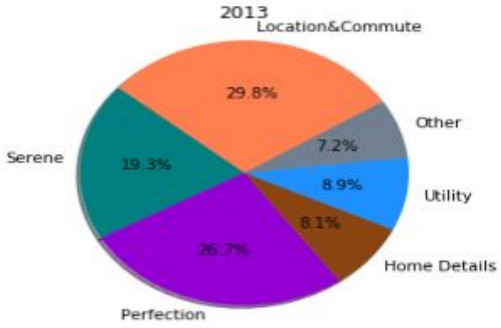
Topic 17
0.077*"พื้นที่" + 0.060*"ออกแบบ" + 0.045*"โล่ง" + 0.039*"โปร่ง" + 0.038*"ใช้สอย" + 0.020*"ประโยชน์" + 0.018*"ห้อง" + 0.017*"อาคาร"

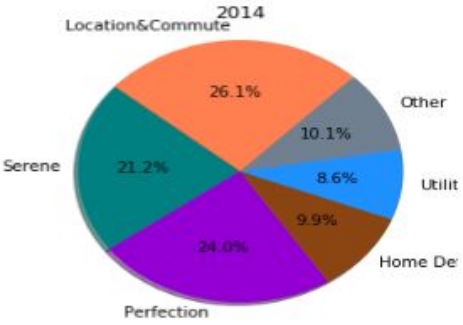


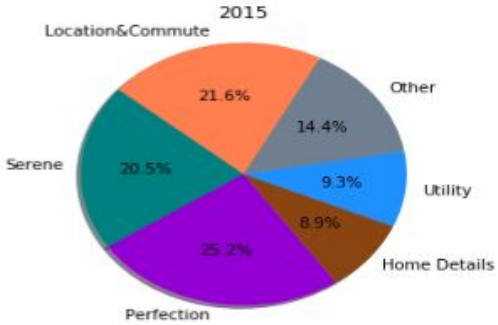
Time and advertisement trends

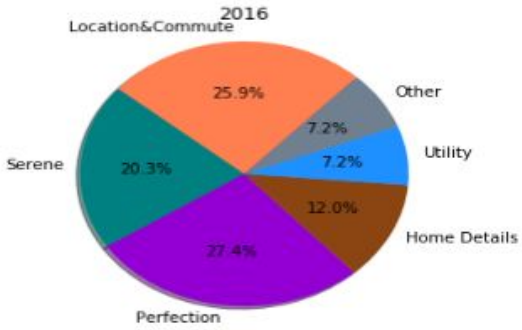


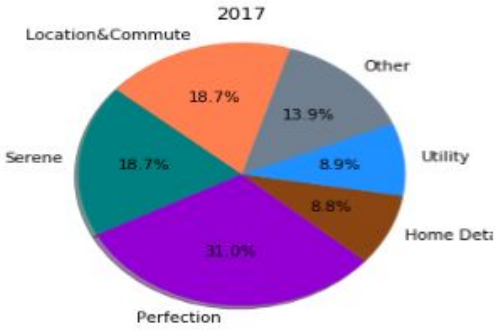


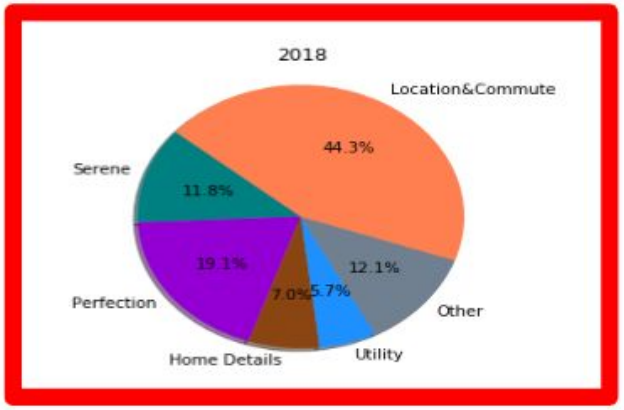






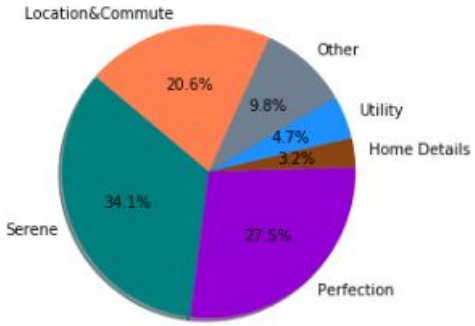


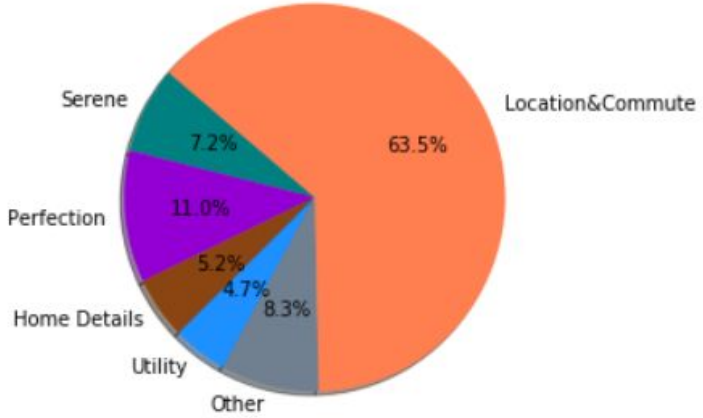


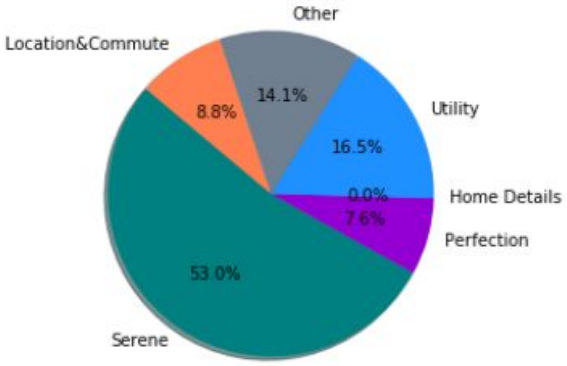


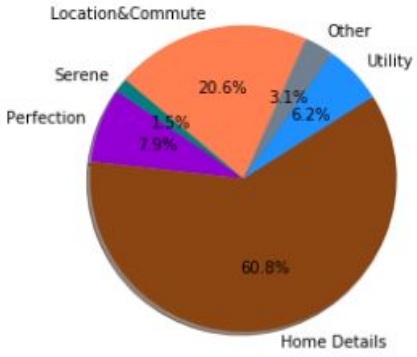
Advertisement niche of each developer

Each real estate developer has its own style



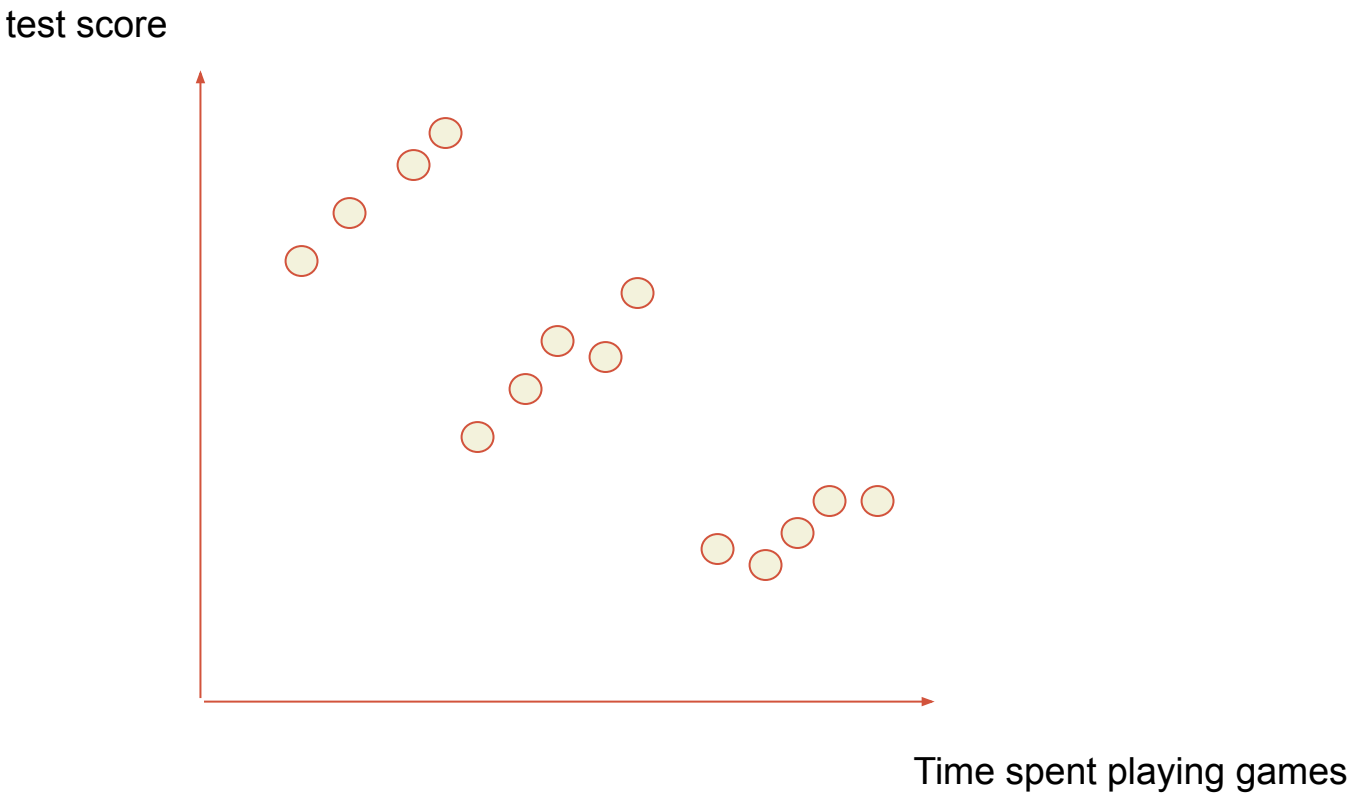






Causal inference motivation

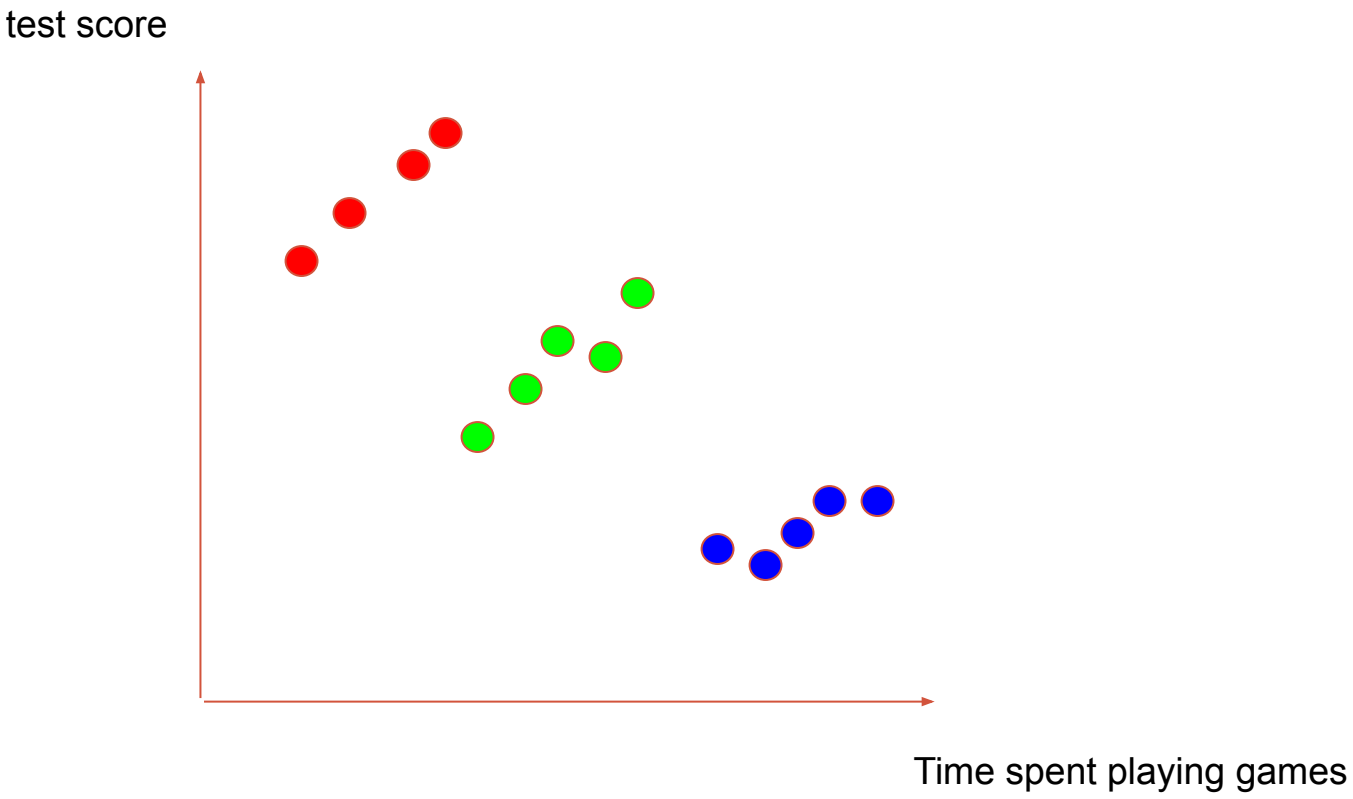
Motivation: correlation is not causation



A model learned on correlation (our current probabilistic framework) might give misleading results

Causal inference motivation

Motivation: correlation is not causation



A model learned on correlation (our current probabilistic framework) might give misleading results

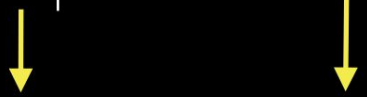
Simpson's paradox

	surv. (Y)	not-surv. (¬Y)		Recovery Rate
drug (X)	20	20	40	50%
no-drug (¬X)	16	24	40	40%
	36	44		

Many figures in this section is from Elias Bareinboim slides at MLRS

Simpson's paradox

	surv. (Y)	not-surv. (¬Y)		Recovery Rate
drug (X)	20	20	40	50%
no-drug (¬X)	16	24	40	40%
	36	44		



male (¬F)	surv. (Y)	not-surv.		Recovery Rate
drug (X)	18	12	30	60%
no-drug (¬X)	7	3	10	70%
	25	15	40	

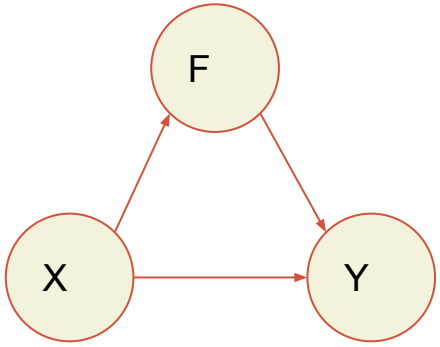
female (F)	surv. (Y)	not-surv. (¬Y)		Recovery Rate
drug (X)	2	8	10	20%
no-drug (¬X)	9	21	30	30%
	11	29	40	

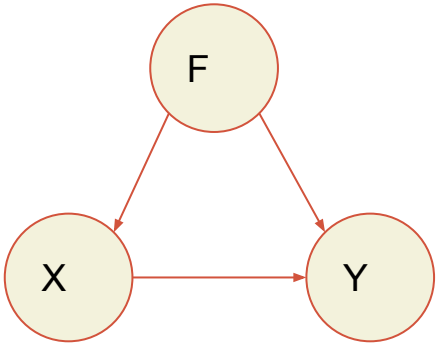
Is this a problem with the graphical model?

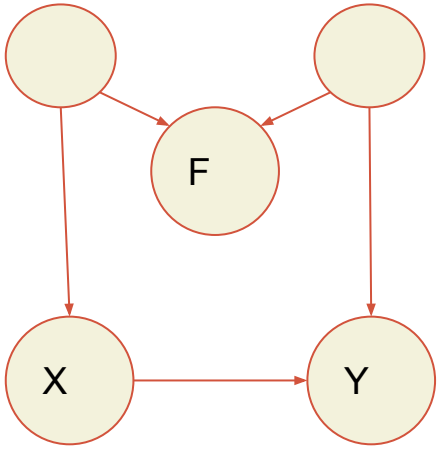
	surv. (Y)	not-surv. (¬Y)		Recovery Rate
drug (X)	20	20	40	50%
no-drug (¬X)	16	24	40	40%
	36	44		

male (¬F)	surv. (Y)	not-surv.		Recovery Rate
drug (X)	18	12	30	60%
no-drug (¬X)	7	3	10	70%
	25	15	40	

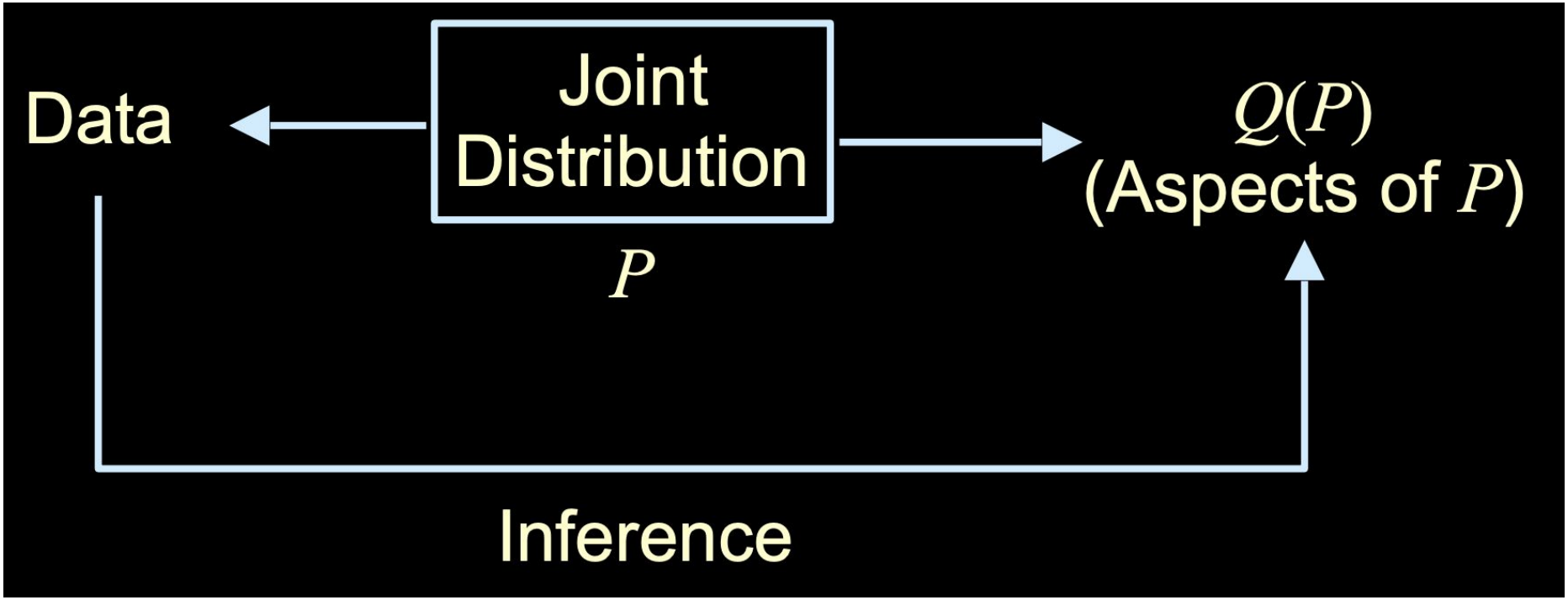
female (F)	surv. (Y)	not-surv. (¬Y)		Recovery Rate
drug (X)	2	8	10	20%
no-drug (¬X)	9	21	30	30%
	11	29	40	





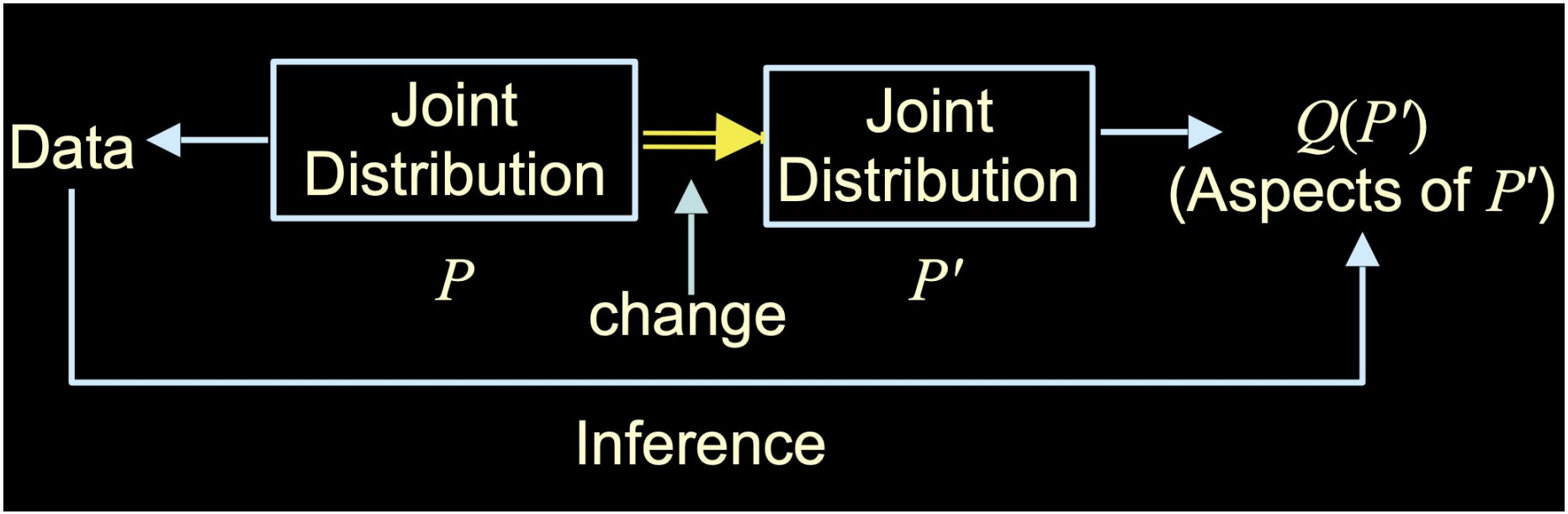


Traditional ML



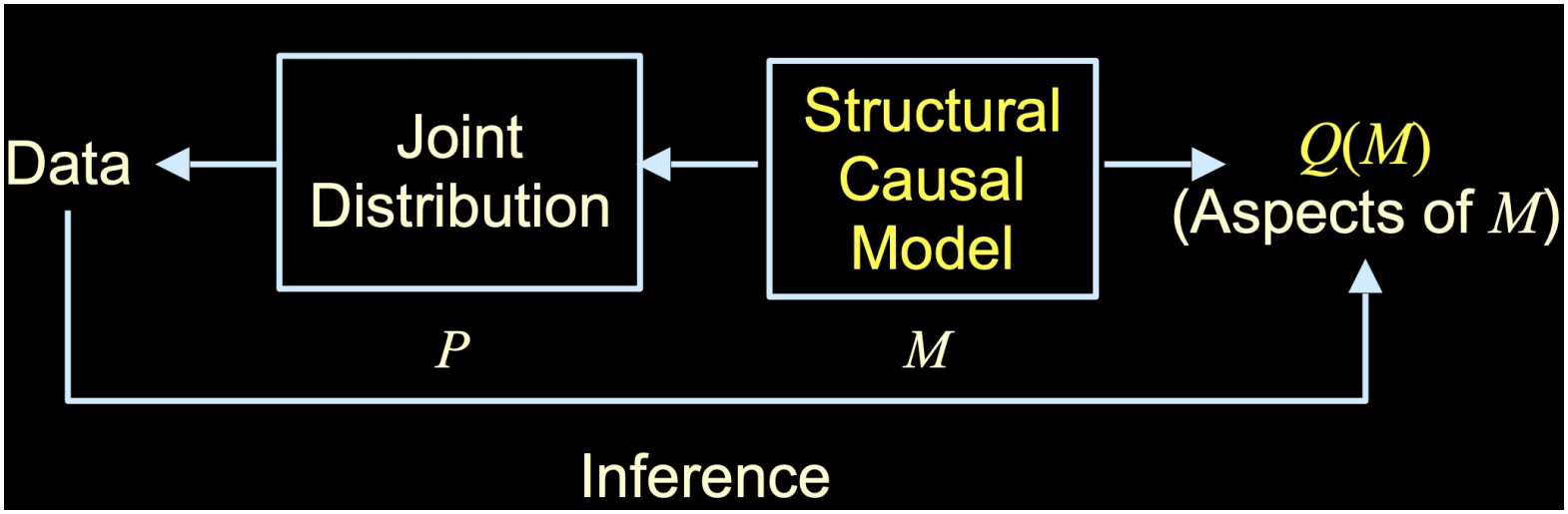
Traditional ML learns P from data

Causal analysis



What happens if we do P'?
Problem: we have a shift in data

Causal modeling paradigm



M - model of nature which captures all the cause and effect

P - observed data

Causal inference is about learning M

Level of causal hierarchy

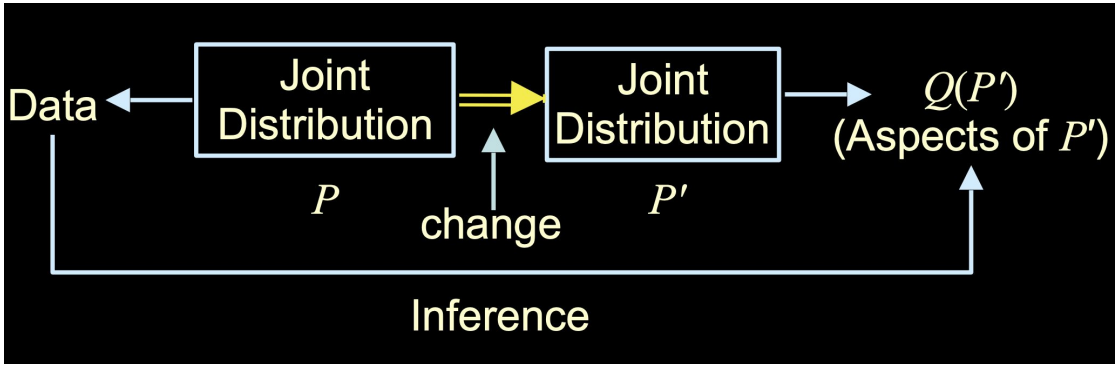
	Level (Symbol)	Typical Activity	Typical Question	Examples
1 	Associational $P(y \mid x)$	Seeing ML - (Un)Supervised DNN, Bayes Net Regression	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?
2 	Interventional $P(y \mid do(x), c)$	Doing ML - Reinforcement Causal Bayes Net	What if? What if I do X?	What if I take aspirin, will my headache be cured?
3 	Counterfactual $P(y_x \mid x', y')$	Imagining, Retrospection	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?

Level of causal hierarchy

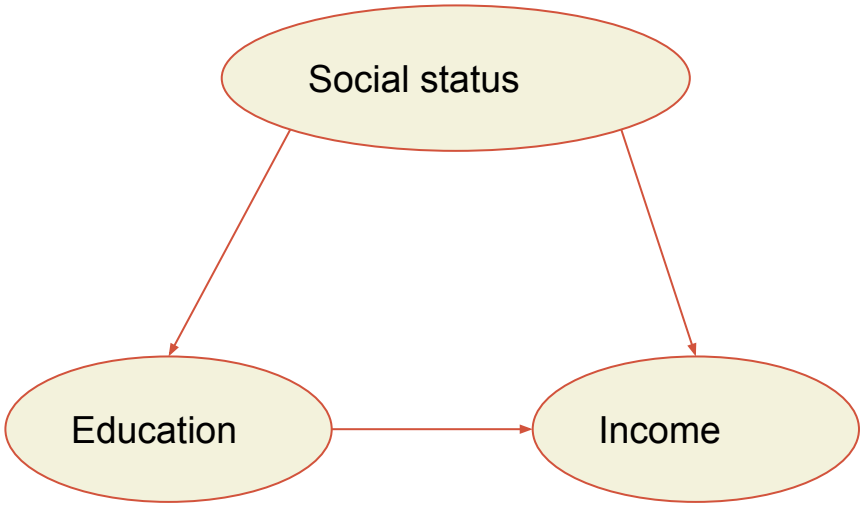
	Level (Symbol)	Typical Activity	Typical Question	Examples
1	Associational $P(y \mid x)$	Seeing ML - (Un)Supervised DNN, Bayes Net Regression	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?
2	Interventional $P(y \mid do(x), c)$	Doing ML - Reinforcement Causal Bayes Net	What if? What if I do X?	What if I take aspirin, will my headache be cured?
3	Counterfactual $P(y_x \mid x', y')$	Imagining, Retrospection	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?

The challenge of interventional inference

We never really observe the effect of an action



regular graphical model thinking



What happens if we make education compulsory?

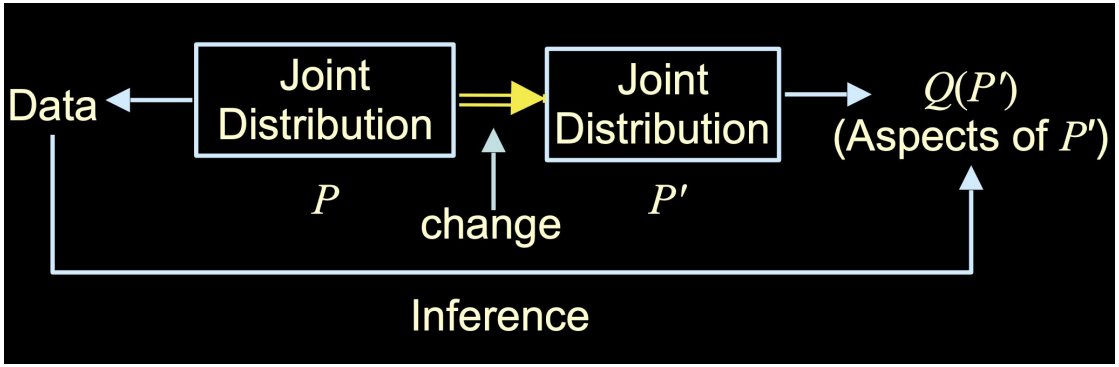
- Pin education = university

Education is tied with social status. If we change one, we effect the other.

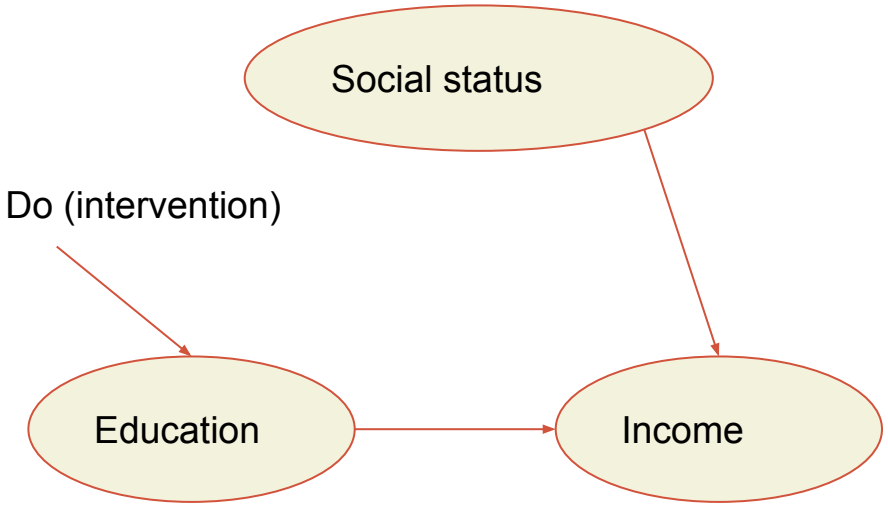
Arrow is relationship not cause effect

The challenge of interventional inference

We never really observe the effect of an action



Causal inference thinking



Ideally we want to disconnect social status and education to perform the analysis

…

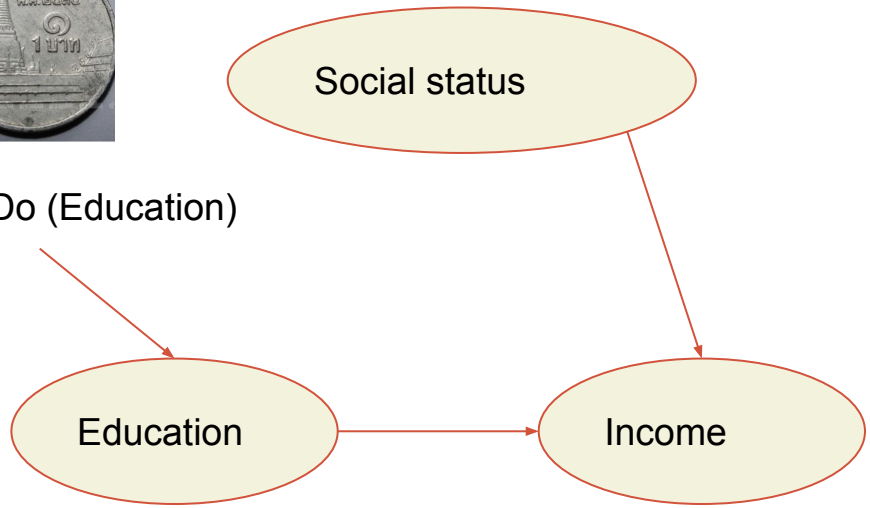
How?

Randomized Control Trial (RCT)

Disconnect by only interacting directly with education.

For a random person, toss a coin, give education to him if the coin is head…

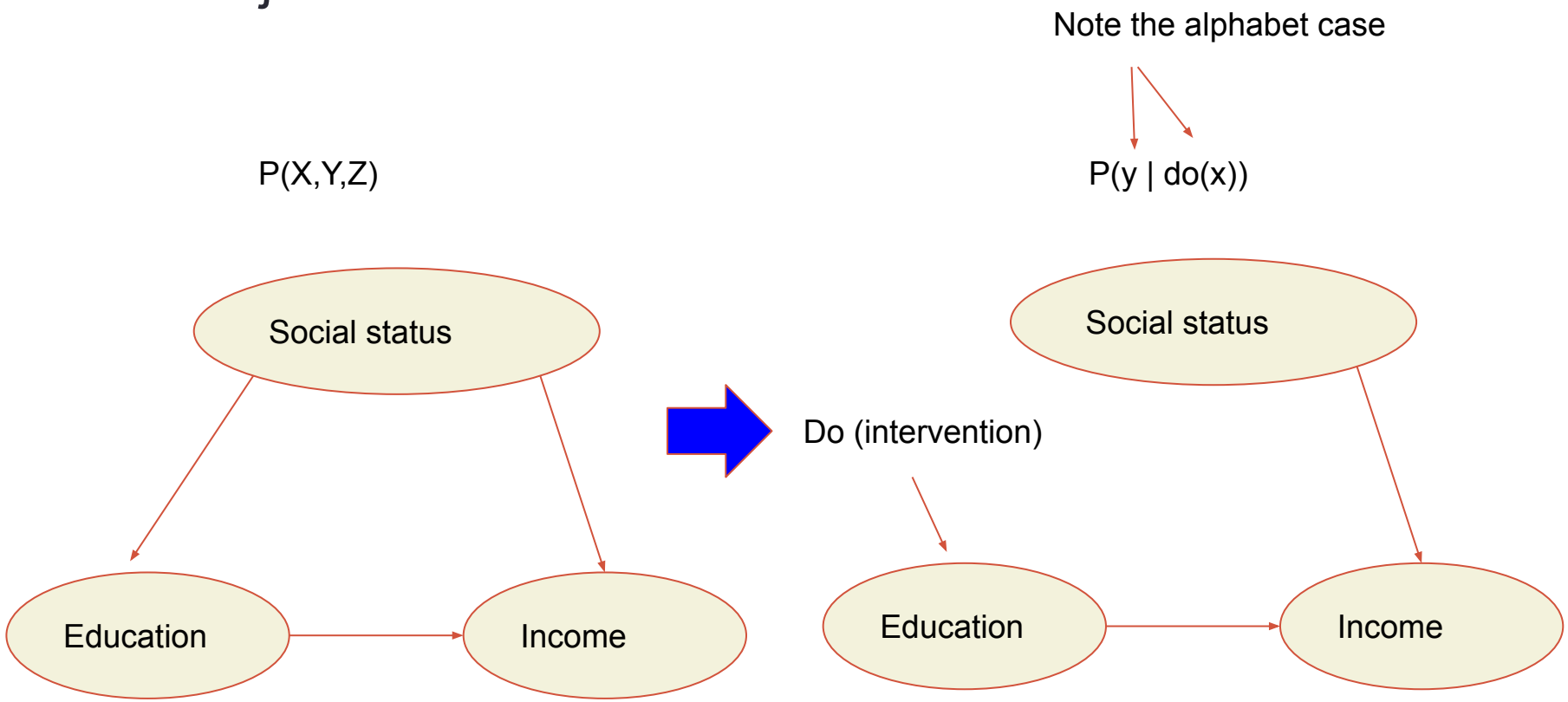
This is the standard for many social studies today.



Problems with RCT

Well...we need to perform it

What if we just have normal data?



Converting to P(y|do(x))

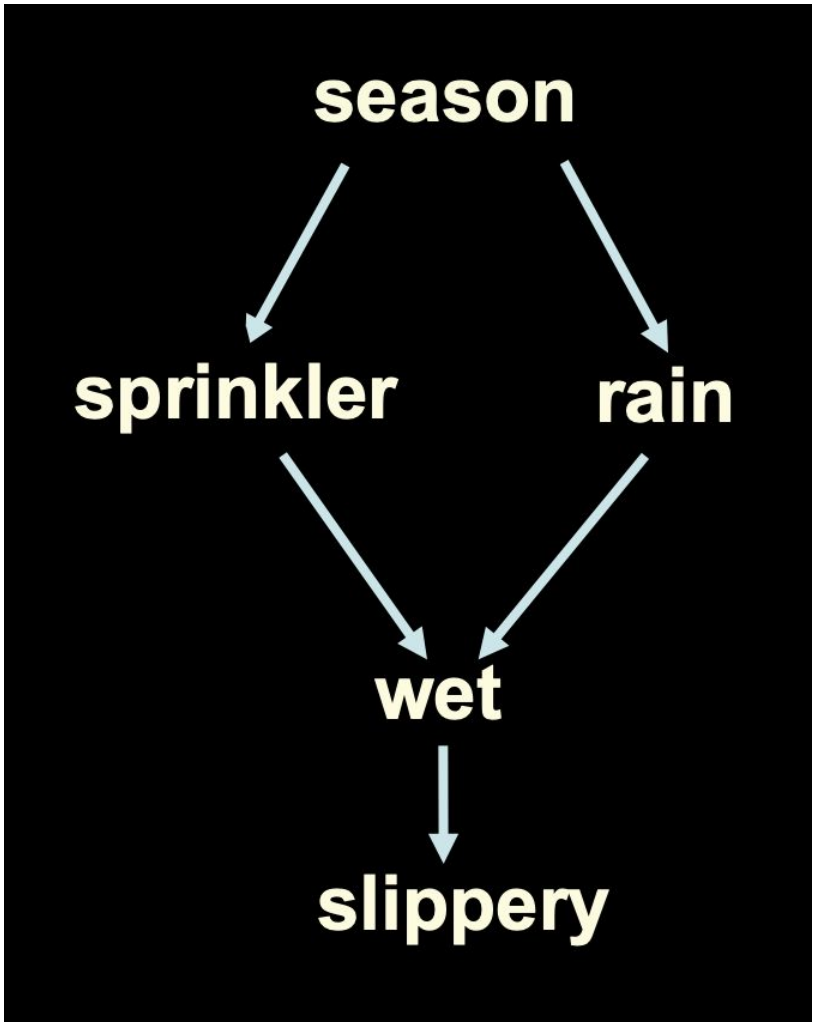
P(y|do(x)) can be found by deleting all factors with X in the original graphical model and replacing conditionals with x

Called the Adjustment formula

Example

P(W| Sp) = ?

P(W = yes| do(Sprinkler = yes) = ?



More

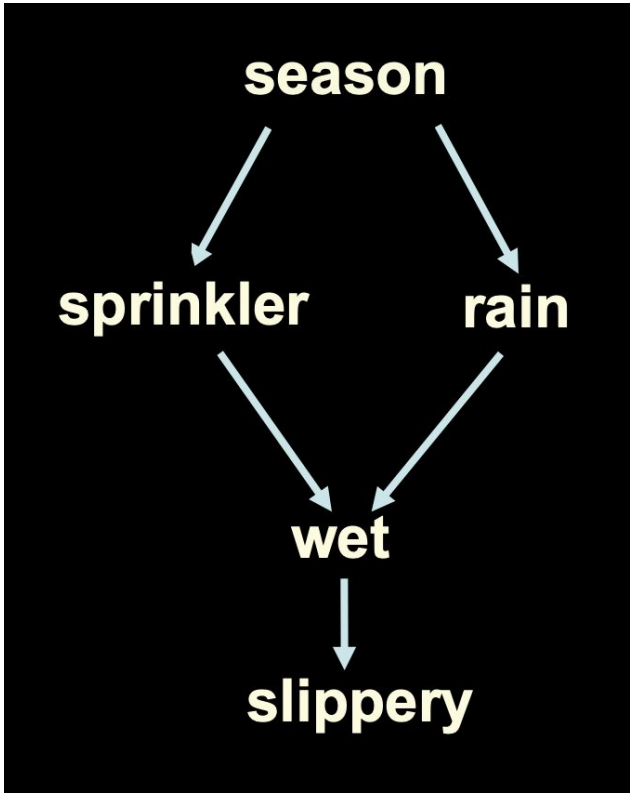
This only works if the parents of X and Y are observed (Season is observed).

Other theorems to deal with this

- backdoor criterion
- frontdoor criterion

Further reading

http://bayes.cs.ucla.edu/home.htm



Solving Simpson's paradox with do(x) notation

Before we found

$P(Y|X) < P(Y|\sim X)$

But

$P(Y|X,F) > P(Y|\sim X,F)$ and

$P(Y|X,\sim F) > P(Y|\sim X,\sim F)$

Let's try

$P(Y|do(X))$ instead

Proof

$$P(Y|do(X)) = P(Y|do(X),F)P(F|do(X)) + P(Y|do(X),\sim F)P(\sim F|do(X))$$
$$= P(Y|do(X),F)P(F) + P(Y|do(X),\sim F)P(\sim F)$$

Similarly

$$P(Y|do(\sim X)) = P(Y|do(\sim X),F)P(F) + P(Y|do(\sim X),\sim F)P(\sim F)$$

If

$$P(Y|do(X),F) > P(Y|do(\sim X),F) \text{ and}$$

$$P(Y|do(X),\sim F) > P(Y|do(\sim X),\sim F)$$

Then,

$$P(Y|do(X)) > P(Y|do(\sim X))$$

Proof

$$P(Y|do(X)) = P(Y|do(X),F)P(F|do(X)) + P(Y|do(X),\sim F)P(\sim F|do(X))$$
$$= P(Y|do(X),F)P(F) + P(Y|do(X),\sim F)P(\sim F)$$

Similarly

$$P(Y|do(\sim X)) = P(Y|do(\sim X),F)P(F) + P(Y|do(\sim X),\sim F)P(\sim F)$$

If

$$P(Y|do(X),F) > P(Y|do(\sim X),F) \text{ and}$$

$$P(Y|do(X),\sim F) > P(Y|do(\sim X),\sim F)$$

Then,

$$P(Y|do(X)) > P(Y|do(\sim X))$$

Summary

Graphical models

- DAG vs Undirected graph
- LDA

Causal inference