10-418/10-618 Machine Learning for Structured Data

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Topic Modeling
+
Convolutional Neural Networks

Matt Gormley
Lecture 15
Oct. 26, 2022

Reminders

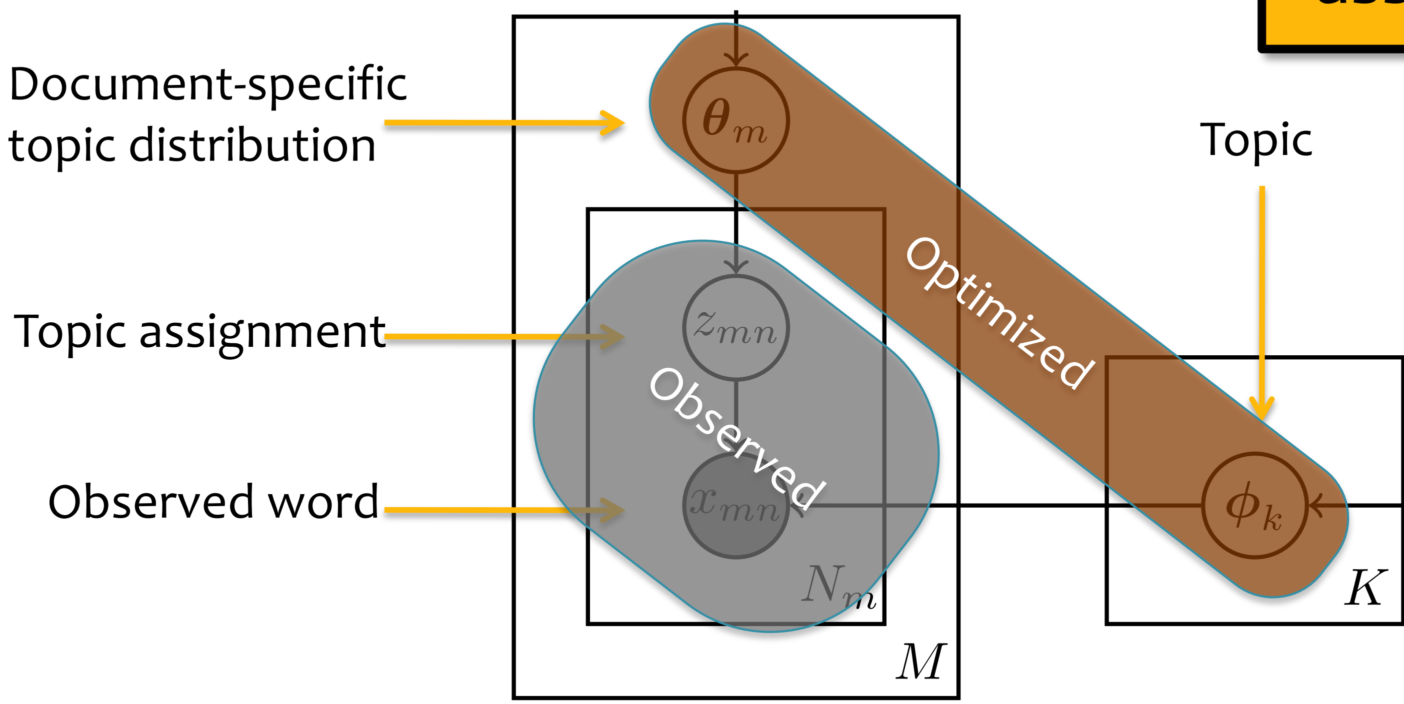
- **Homework 4: MCMC**
  - **Out: Mon, Oct 24**
  - **Due: Fri, Nov 3 at 11:59pm**

BAYESIAN INFERENCE FOR PARAMETER ESTIMATION

LDA Inference

- Fully Observed MLE

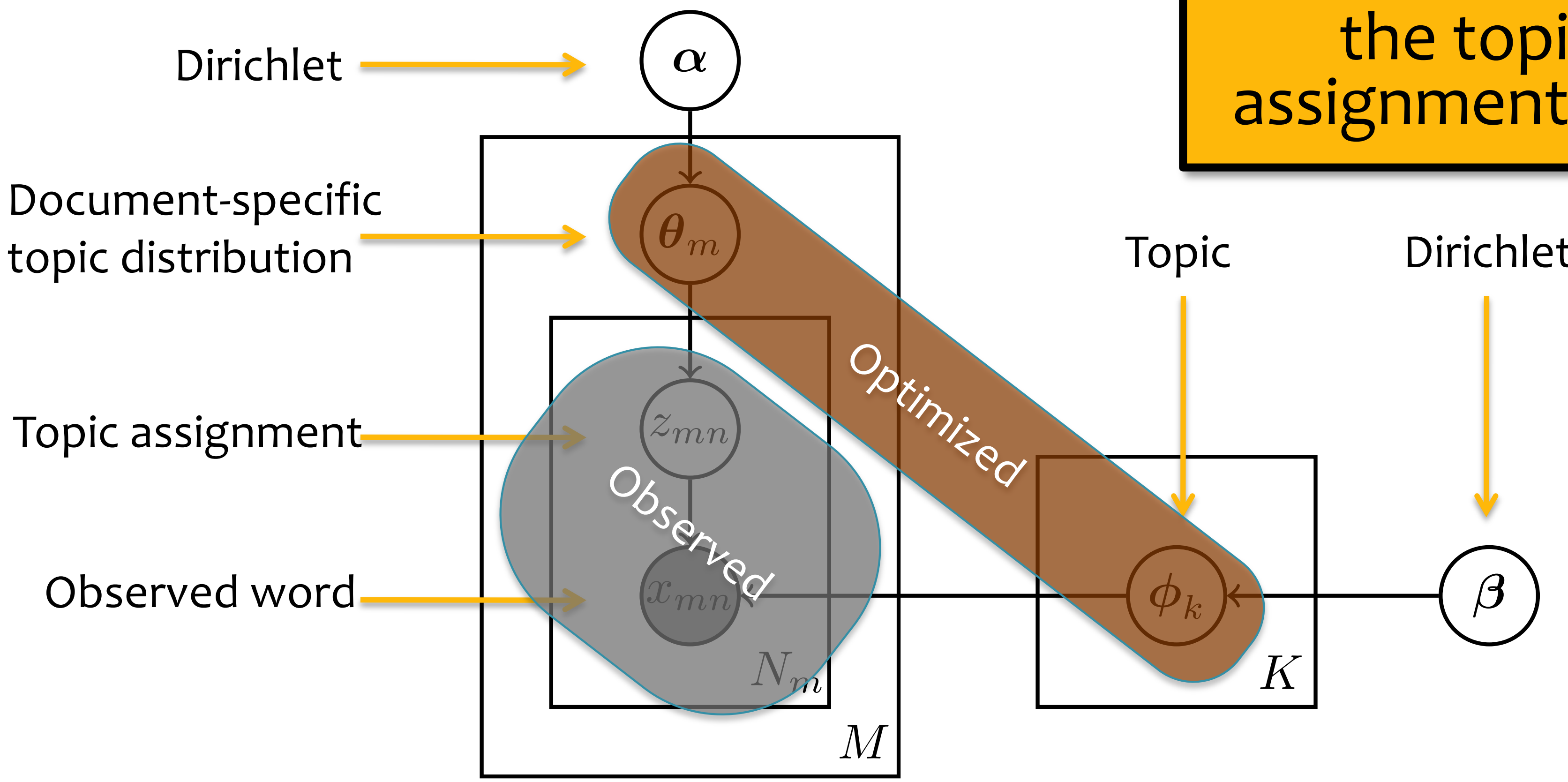
Learning like this would be easy, but in practice we do not observe the topic assignments z_{mn}



LDA Inference

- Full Observed MAP Estimation

Learning like this would be easy, but in practice we do not observe the topic assignments z_{mn}



Unsupervised Learning

Three learning paradigms:

1. Maximum likelihood estimation (MLE)

$$\arg\max_{\theta} p(X|\theta)$$

2. Maximum a posteriori (MAP) estimation

$$\arg\max_{\theta} p(\theta|X) \propto p(X|\theta)p(\theta)$$

3. Bayesian approach

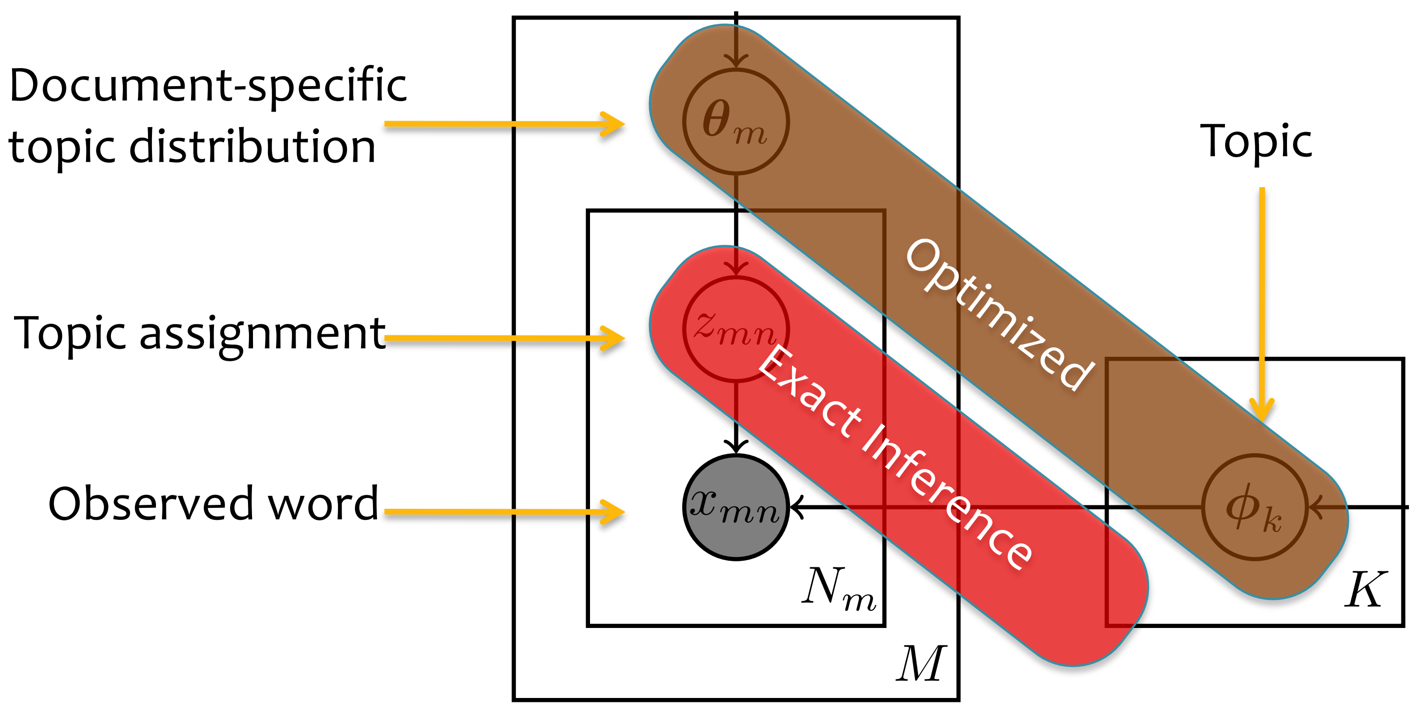
Estimate the posterior:

$$p(\theta|X) = \;\ldots$$

LDA Inference

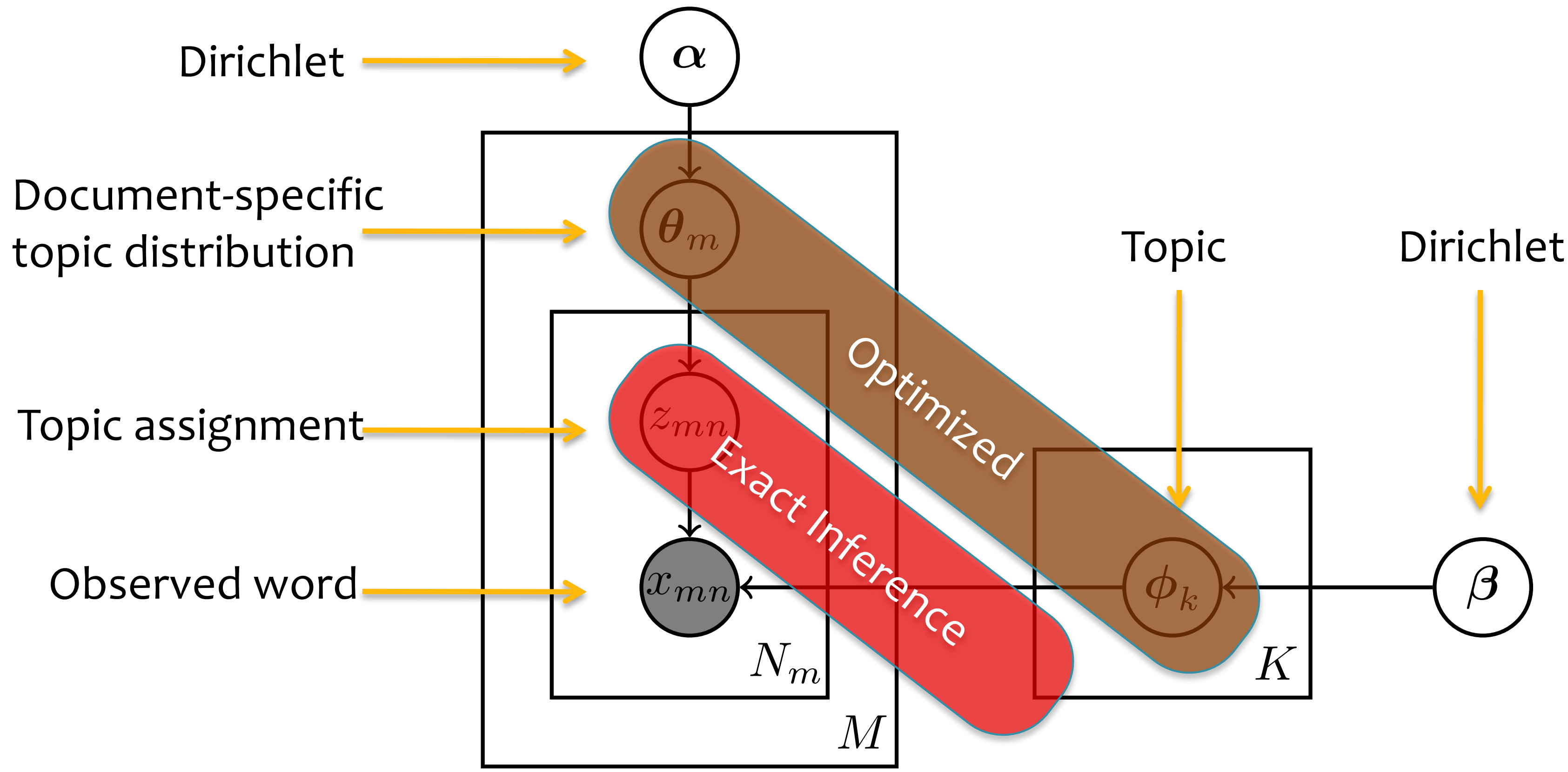
- Standard EM (MLE)

$P_{\theta,\phi}(x) = \sum_z P_{\theta,\phi}(x,z)$



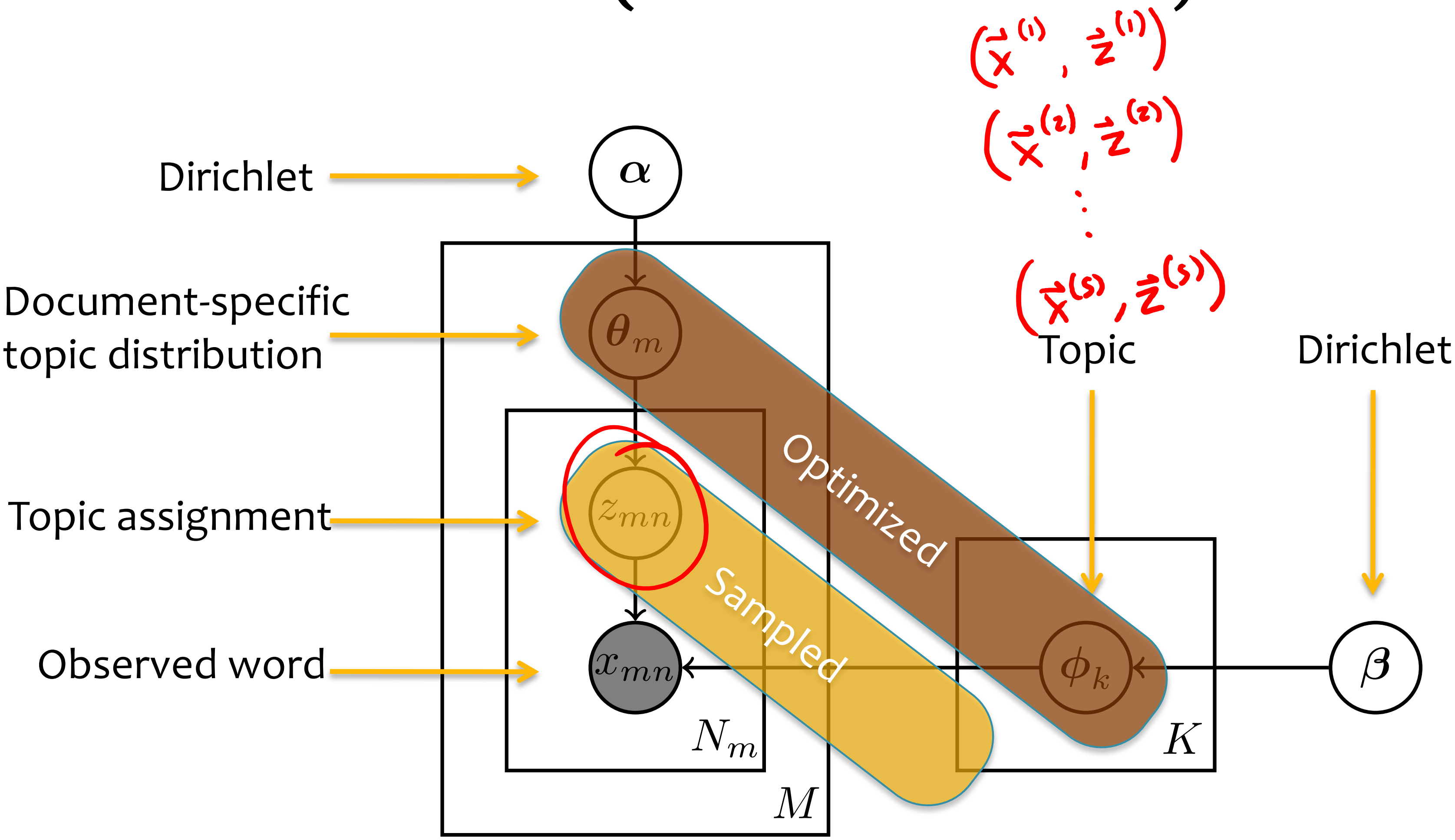
LDA Inference

- Standard EM (MAP Estimation)



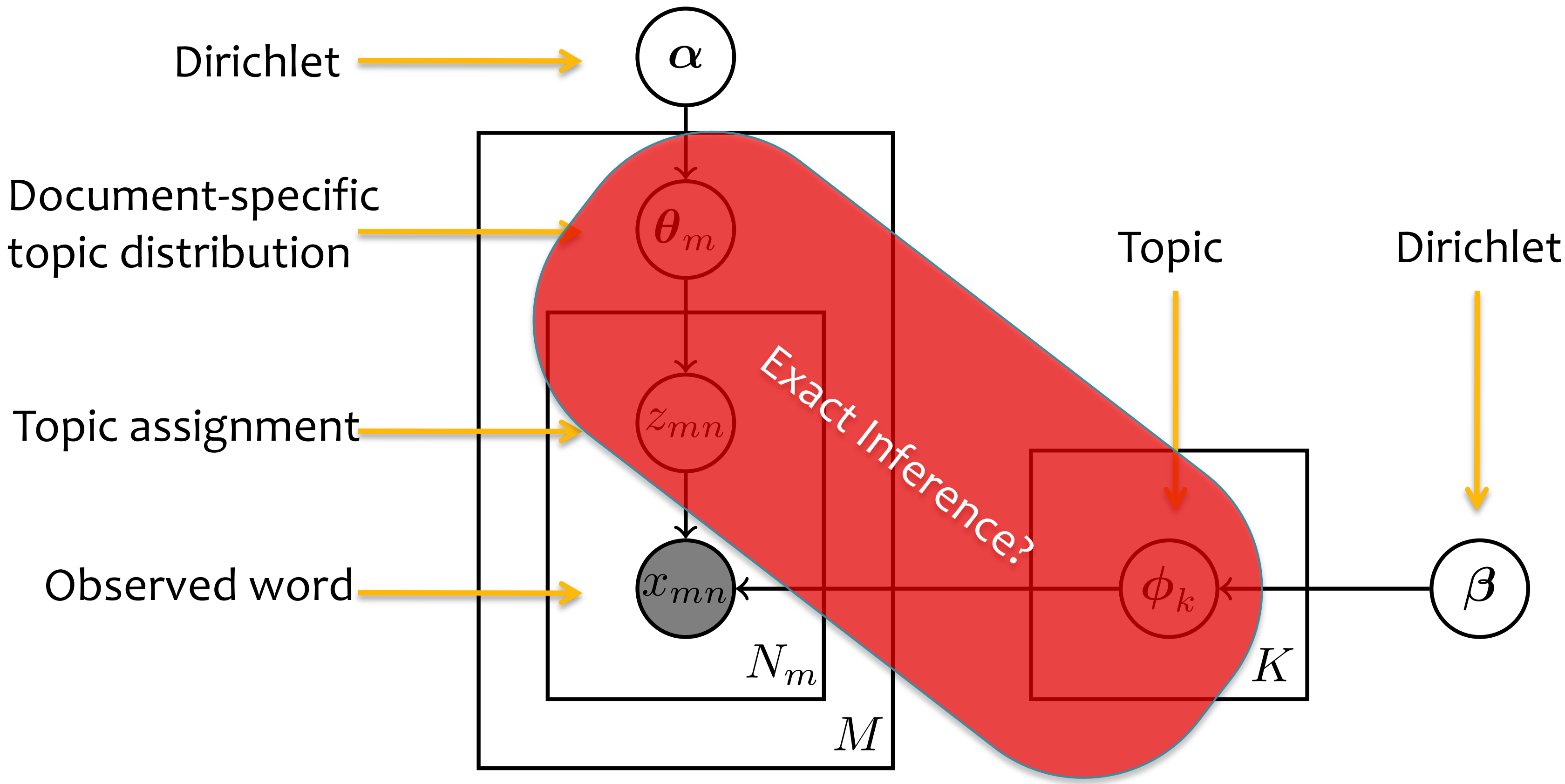
LDA Inference

- Monte Carlo EM (MAP Estimation)



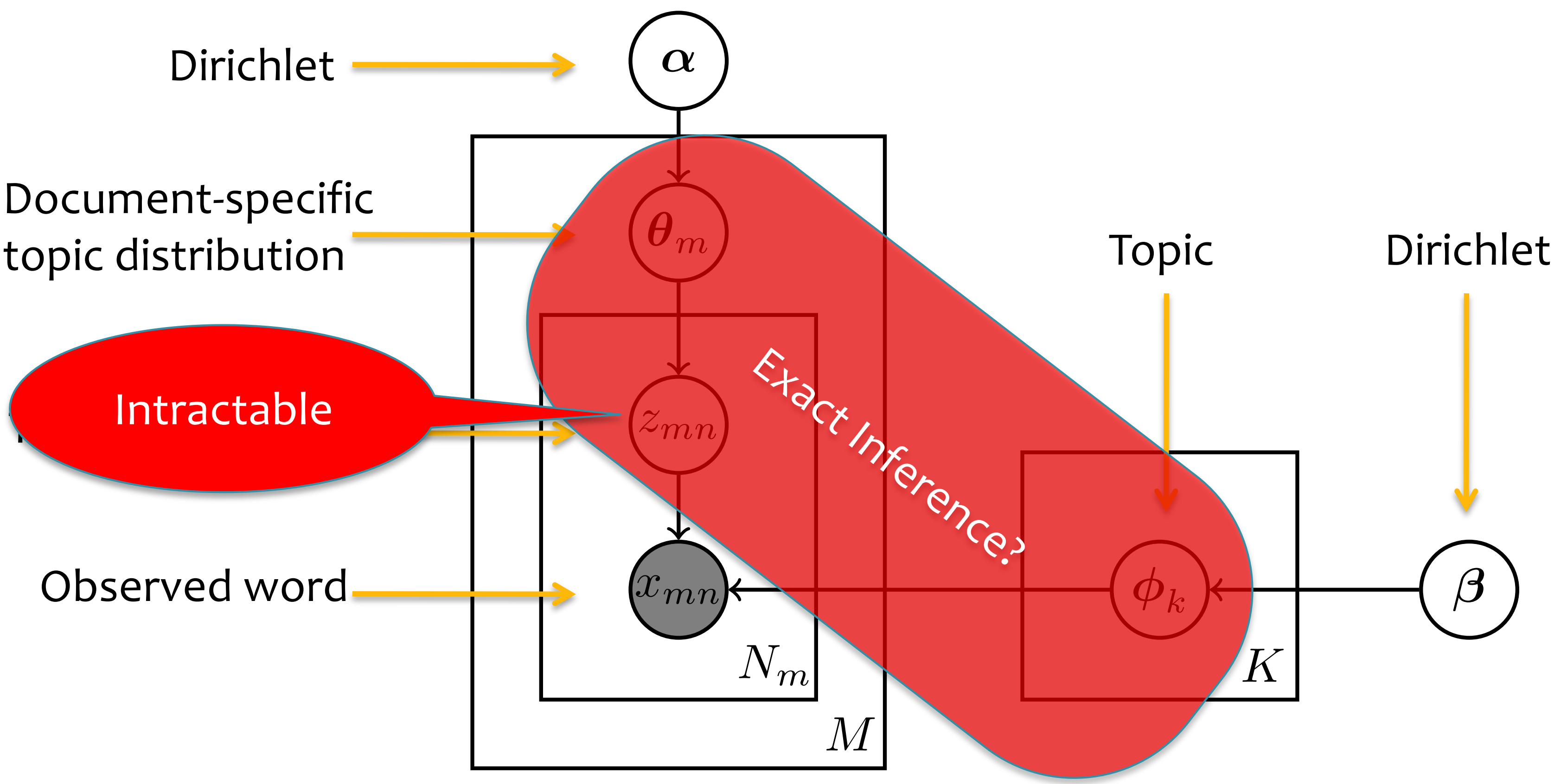
LDA Inference

- Bayesian Approach



LDA Inference

- Bayesian Approach

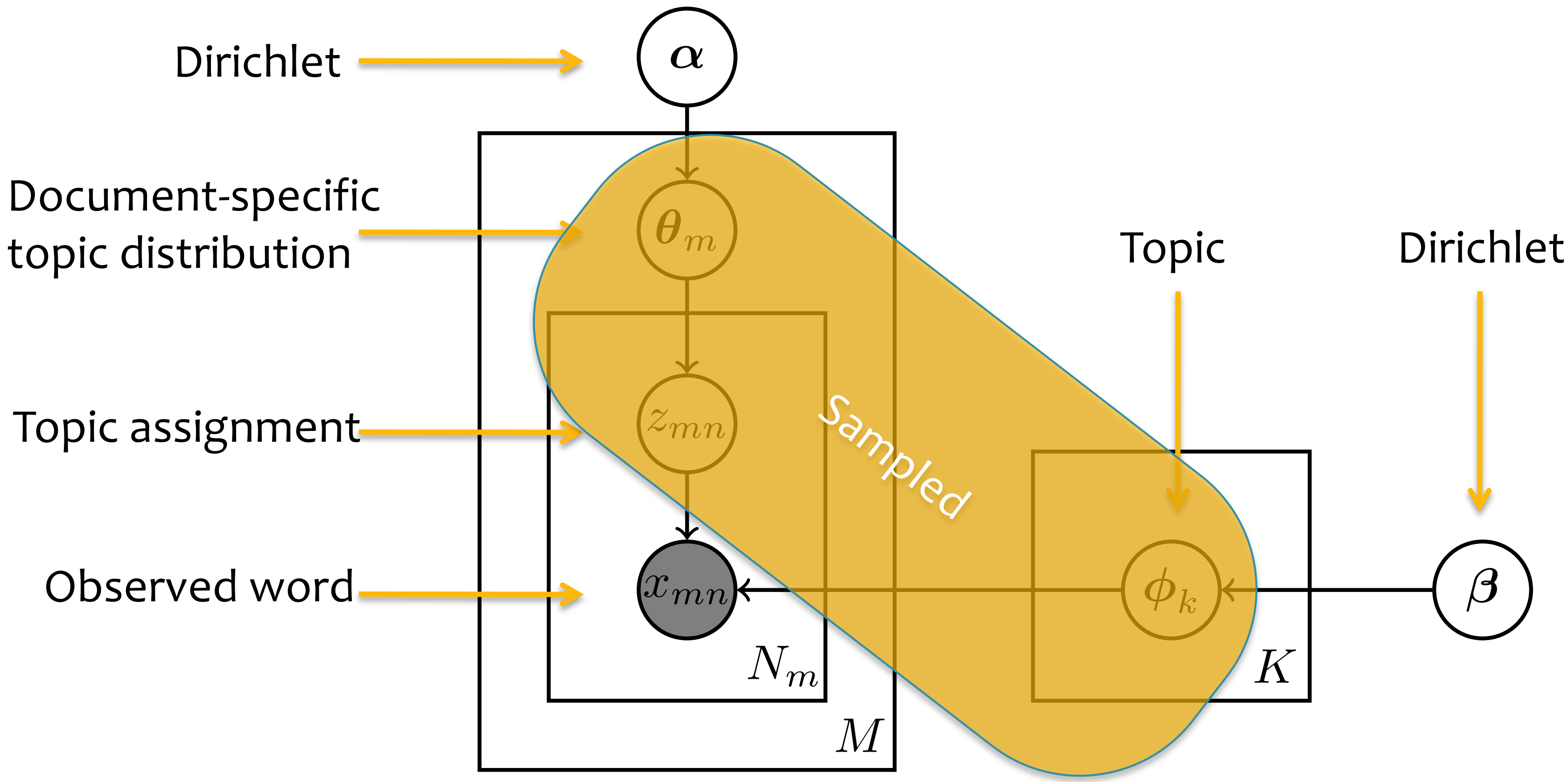


Exact Inference in LDA

- Exactly computing the posterior is intractable in LDA
  - Junction tree algorithm: exact inference in general graphical models
    1. "moralization" converts directed to undirected
    2. "triangulation" breaks 4-cycles by adding edges
    3. Cliques arranged into a junction tree
  - Time complexity is exponential in size of cliques
  - LDA cliques will be large (at least O(# topics)), so complexity is $O(2^{\#\text{ topics}})$
- Exact MAP inference in LDA is NP-hard for a large number of topics (Sontag & Roy, 2011)

LDA Inference

- Explicit Gibbs Sampler



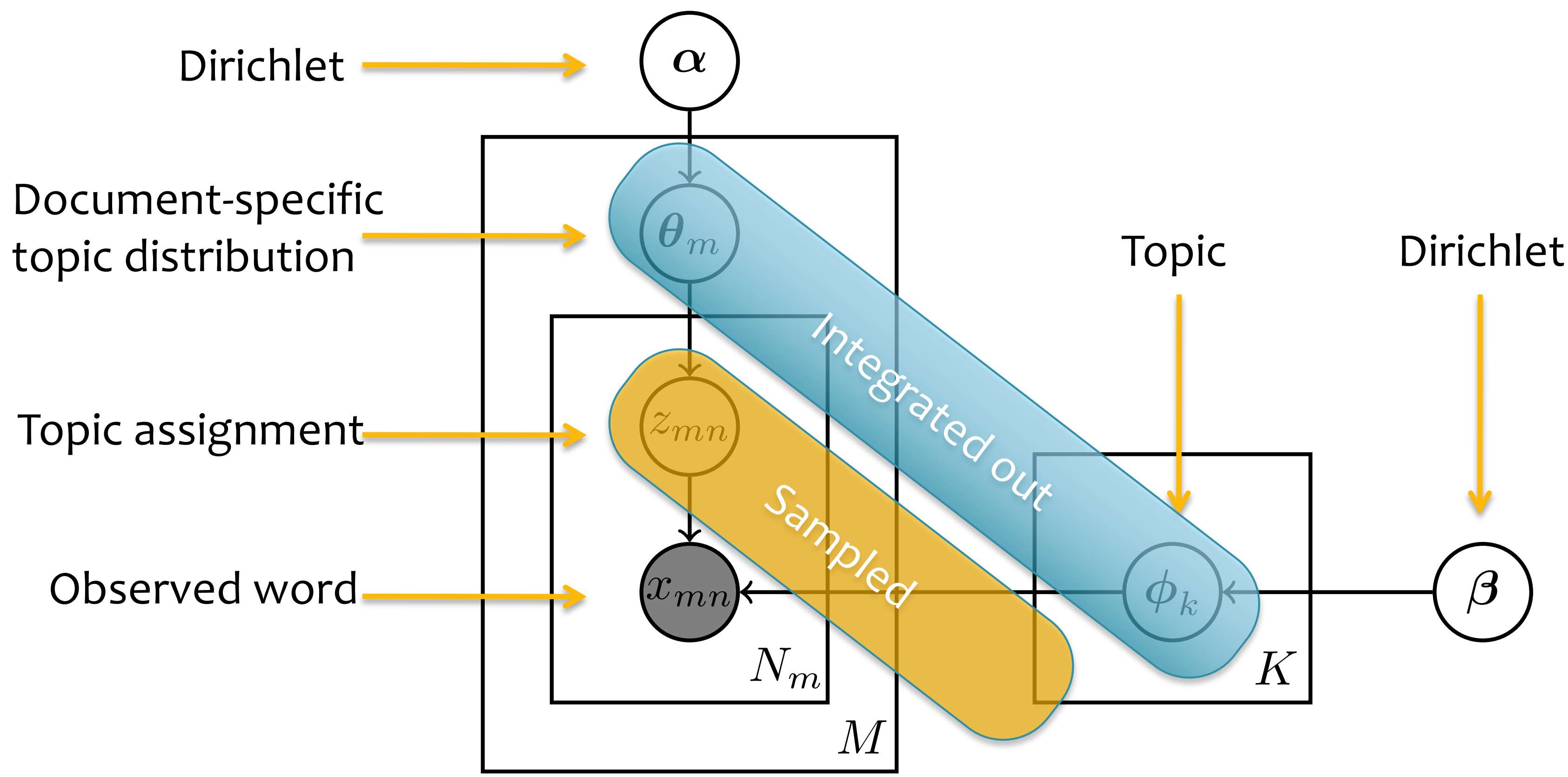
LDA Inference

Whiteboard:

– Explicit Gibbs Sampler for LDA

LDA Inference

- Collapsed Gibbs Sampler



LDA Inference

Whiteboard:

– Collapsed Gibbs Sampler for LDA

COLLAPSED GIBBS SAMPLER FOR LDA

Collapsed Gibbs Sampler for LDA

Goal:

- Draw samples from the posterior $p(Z|X, \alpha, \beta)$
- Integrate out topics ϕ and document-specific distribution over topics θ

Algorithm:

- While not done…
  - For each document, m:
    - For each word, n:
      - » Resample a single topic assignment using the full conditionals for z_{mn}

Collapsed Gibbs Sampler for LDA

- What can we do with samples of z_{mn}?
  - Mean of z_{mn}
  - Mode of z_{mn}
  - Estimate posterior over z_{mn}
  - Estimate of topics ϕ and document-specific distribution over topics θ

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t},$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k}.$$

Collapsed Gibbs Sampler for LDA

$Z_{mn} \longrightarrow Z_i$ where $i = (m,n)$

- Full conditionals

all other topic assignments — all words — hyperparameters

$$p(z_i = k | Z^{-i}, X, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \overbrace{\frac{n_{kt}^{-i} + \beta_t}{\sum_{v=1}^{T} n_{kv}^{-i} + \beta_v}}^{\text{how popular is word } t \text{ for topic } k} \cdot \overbrace{\frac{n_{mk}^{-i} + \alpha_k}{\sum_{j=1}^{K} n_{mj}^{-i} + \alpha_j}}^{\text{how popular is topic } k \text{ in doc } m}$$

where t, m are given by i

$N_m - 1$

n_{kt} = # times topic k appears with type t

n_{mk} = # times topic k appears in document m

Collapsed Gibbs Sampler for LDA

Whiteboard:

- Efficient computation of count variables

Collapsed Gibbs Sampler for LDA

- Sketch of the derivation of the full conditionals

$$
\begin{aligned}
p(z_i = k | Z^{-i}, X, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{p(X, Z | \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(X, Z^{-i} | \boldsymbol{\alpha}, \boldsymbol{\beta})} \\
&\propto p(X, Z | \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= p(X | Z, \boldsymbol{\beta}) p(Z | \boldsymbol{\alpha}) \\
&= \int_\Phi p(X | Z, \Phi) p(\Phi | \boldsymbol{\beta}) \, d\Phi \int_\Theta p(Z | \Theta) p(\Theta | \boldsymbol{\alpha}) \, d\Theta \\
&= \left(\prod_{k=1}^{K} \frac{B(\vec{n}_k + \boldsymbol{\beta})}{B(\boldsymbol{\beta})} \right) \left(\prod_{m=1}^{M} \frac{B(\vec{n}_m + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \right) \\
&= \frac{n_{kt}^{-i} + \beta_t}{\sum_{v=1}^{T} n_{kv}^{-i} + \beta_v} \cdot \frac{n_{mk}^{-i} + \alpha_k}{\sum_{j=1}^{K} n_{mj}^{-i} + \alpha_j}
\end{aligned}
$$

where t, m are given by i

Dirichlet-Multinomial Model

- The Dirichlet is conjugate to the Multinomial

$\boldsymbol{\phi} \sim \text{Dir}(\boldsymbol{\beta})$ [*draw distribution over words*]
For each word $n \in \{1, \ldots, N\}$
  $x_n \sim \text{Mult}(1, \boldsymbol{\phi})$ [*draw word*]

- The posterior of ϕ is $p(\phi|X) = \frac{p(X|\phi)p(\phi)}{P(X)}$
- Define the count vector $\boldsymbol{n}$ such that n_t denotes the number of times word t appeared
- Then the posterior is also a Dirichlet distribution: $p(\phi|X) \sim \text{Dir}(\boldsymbol{\beta} + \boldsymbol{n})$

Dirichlet-Multinomial Model

- Why conjugacy is so useful

$$p(X|\boldsymbol{\alpha}) = \int_\phi p(X|\vec{\phi})p(\vec{\phi}|\boldsymbol{\alpha})\, d\phi$$

Mult Dir

$$= \int_\phi \left(\prod_{v=1}^V \phi_v^{n_v}\right)\left(\frac{1}{B(\boldsymbol{\alpha})}\prod_{v=1}^V \phi_v^{\alpha_v-1}\right) d\phi$$

$$= \frac{1}{B(\boldsymbol{\alpha})}\int_\phi \prod_{v=1}^V \phi_v^{n_v+\alpha_v-1}\, d\phi$$

mult by 1.0

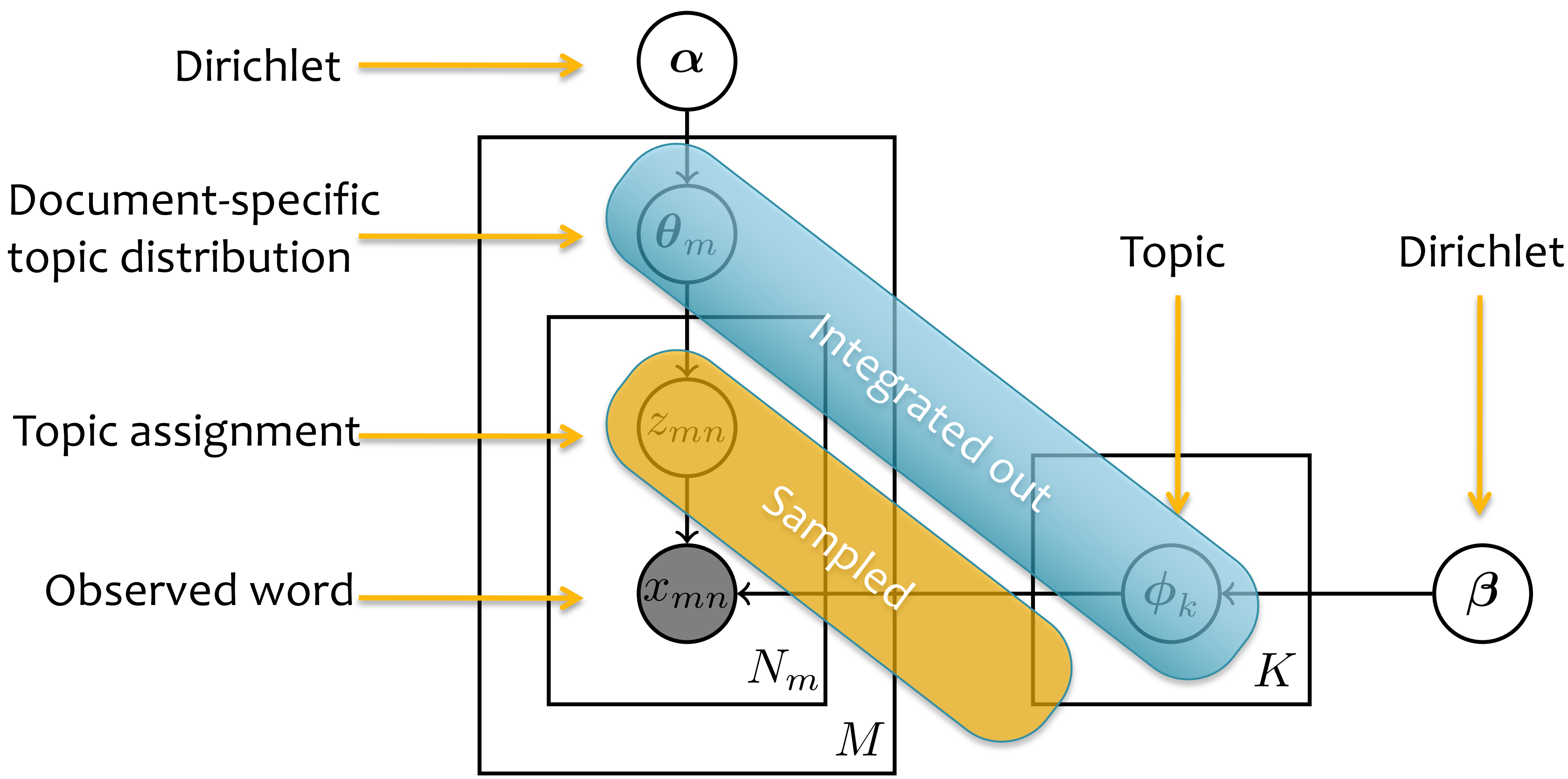
$$= \frac{1}{B(\boldsymbol{\alpha})}\int_\phi \boxed{\frac{B(\vec{n}+\boldsymbol{\alpha})}{B(\vec{n}+\boldsymbol{\alpha})}}\prod_{v=1}^V \phi_v^{n_v+\alpha_v-1}\, d\phi$$

$$= \frac{B(\vec{n}+\boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}\int_\phi \underbrace{\frac{1}{B(\vec{n}+\boldsymbol{\alpha})}\prod_{v=1}^V \phi_v^{n_v+\alpha_v-1}}_{Dir(\vec{n}+\boldsymbol{\alpha})}\, d\phi$$

$$= \frac{B(\vec{n}+\boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}$$

LDA Inference

- Collapsed Gibbs Sampler



Collapsed Gibbs Sampler for LDA

Algorithm

```
// initialisation
```

zero all count variables, $n_m^{(k)}, n_m, n_k^{(t)}, n_k$

for all documents $m \in [1, M]$ **do**

  for all words $n \in [1, N_m]$ in document m **do**

    sample topic index $z_{m,n}=k \sim \text{Mult}(1/K)$ ← randomly initialize variable assignments

    increment document–topic count: $n_m^{(k)}$ += 1

    increment document–topic sum: n_m += 1

    increment topic–term count: $n_k^{(t)}$ += 1

    increment topic–term sum: n_k += 1

Collapsed Gibbs Sampler for LDA

Algorithm

```
// Gibbs sampling over burn-in period and sampling period
```
while not finished **do**

  for all documents $m \in [1, M]$ **do**

    for all words $n \in [1, N_m]$ in document m **do**

      `// for the current assignment of` k `to a term` t `for word` $w_{m,n}$`:`

      decrement counts and sums: $n_m^{(k)} \mathrel{-}= 1; n_m \mathrel{-}= 1; n_k^{(t)} \mathrel{-}= 1; n_k \mathrel{-}= 1$

      `// multinomial sampling acc. to Eq. 78 (decrements from previous step):`

      sample topic index $\tilde{k} \sim p(z_i|\vec{z}_{\neg i}, \vec{w})$

      `// for the new assignment of` $z_{m,n}$ `to the term` t `for word` $w_{m,n}$`:`

      increment counts and sums: $n_m^{(\tilde{k})} \mathrel{+}= 1; n_m \mathrel{+}= 1; n_{\tilde{k}}^{(t)} \mathrel{+}= 1; n_{\tilde{k}} \mathrel{+}= 1$

Collapsed Gibbs Sampler for LDA

Whiteboard:

- Q: How to recover parameter estimates from the collapsed Gibbs sampler?
- Dirichlet distribution over parameters
- Expected values of the parameters

Why does Gibbs sampling work?

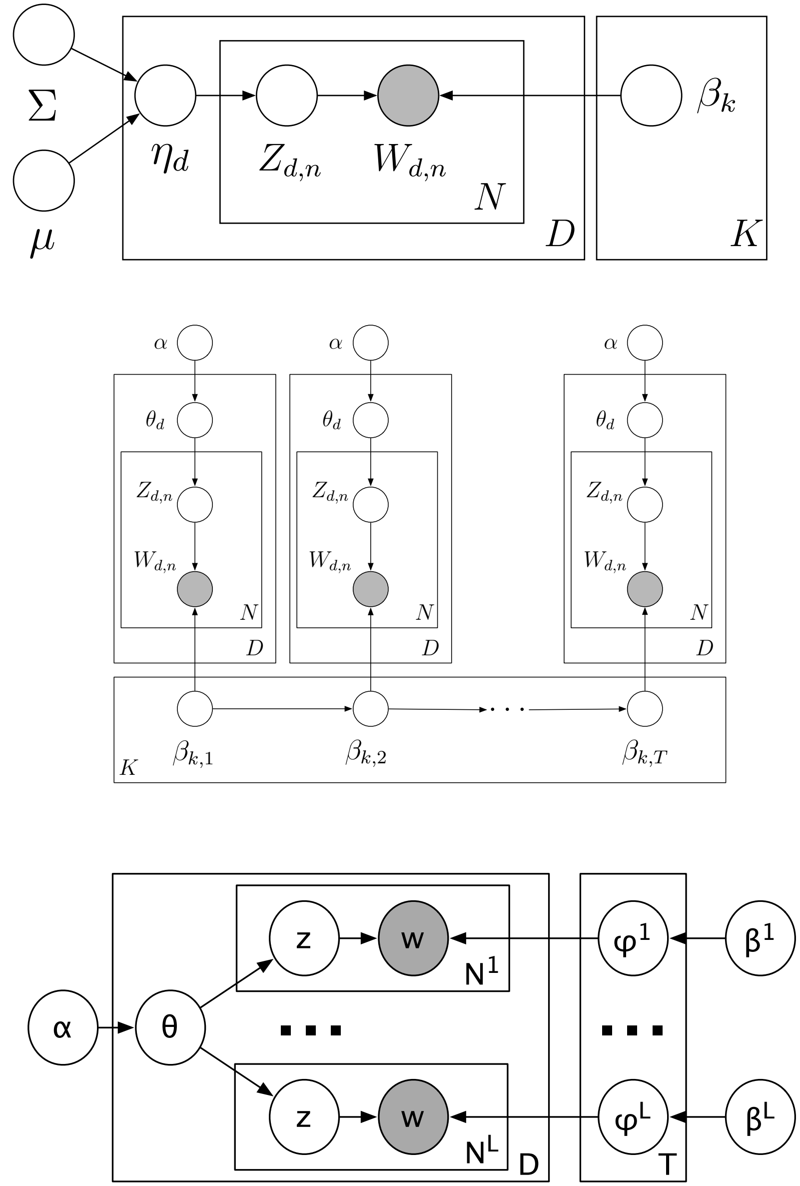
- Metropolis-Hastings
  - Markov chains
  - Stationary distribution
  - MH Algorithm
    - Constructs a Markov chain whose stationary distribution is the desired distribution
  - Proof that samples will be from desired distribution:
    - Sufficient conditions for constructing a markov chain with desired stationary distribution:
      - ergodicity
      - detailed balance (stronger, than what we need, but easier for the proof)
- Gibbs Sampling is a special case of Metropolis-Hastings
  - a special proposal distribution, which ensures the hastings ratio is always 1.0

EXTENSIONS OF LDA

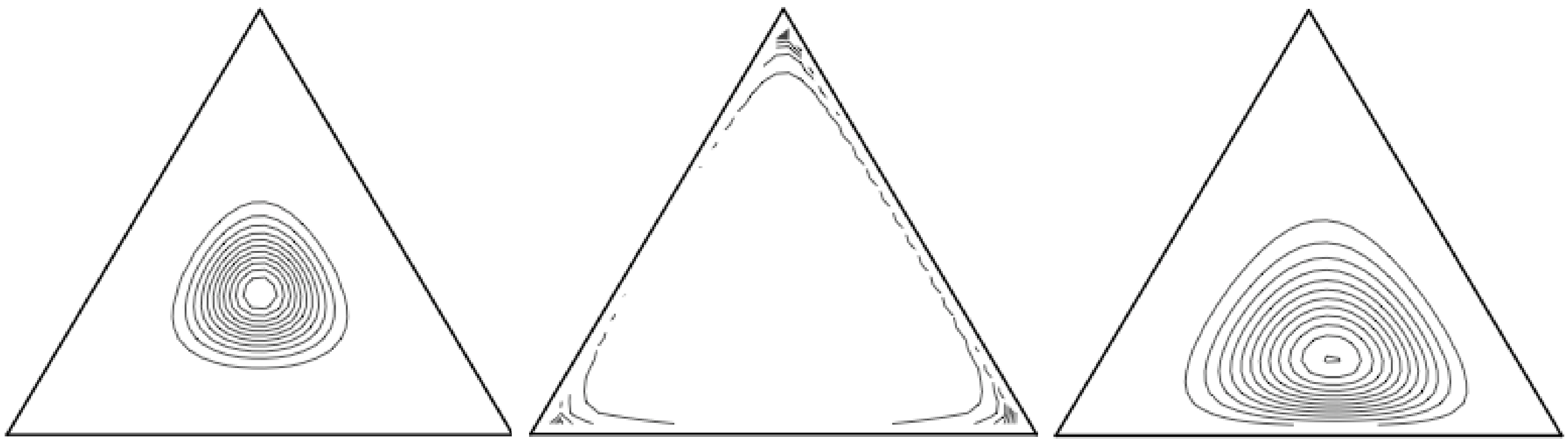
Extensions to the LDA Model

- Correlated topic models
  - Logistic normal prior over topic assignments
- Dynamic topic models
  - Learns topic changes over time
- Polylingual topic models
  - Learns topics aligned across multiple languages

…



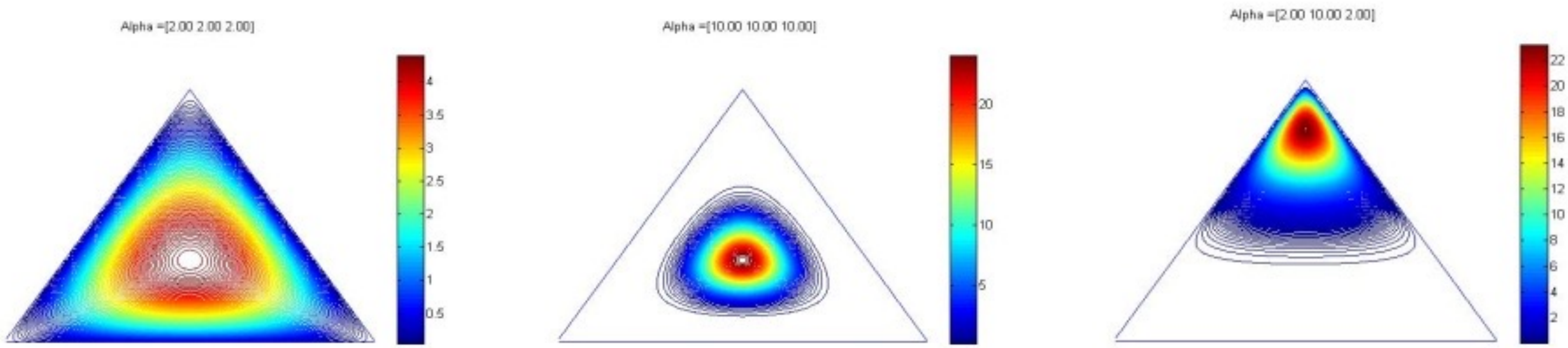
Correlated Topic Models



- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

Slide from David Blei, MLSS 2012

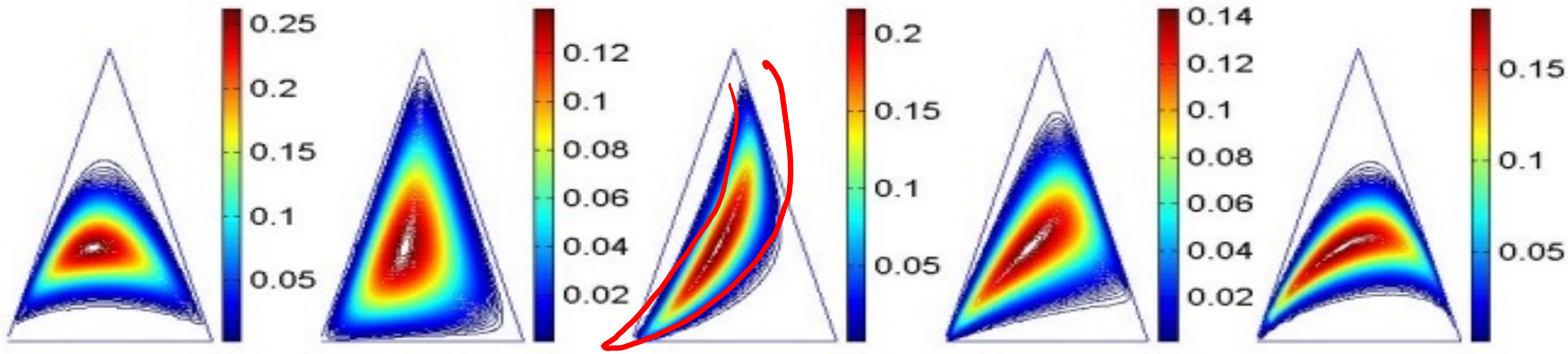
Correlated Topic Models



- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

(Blei & Lafferty, 2004)

Correlated Topic Models



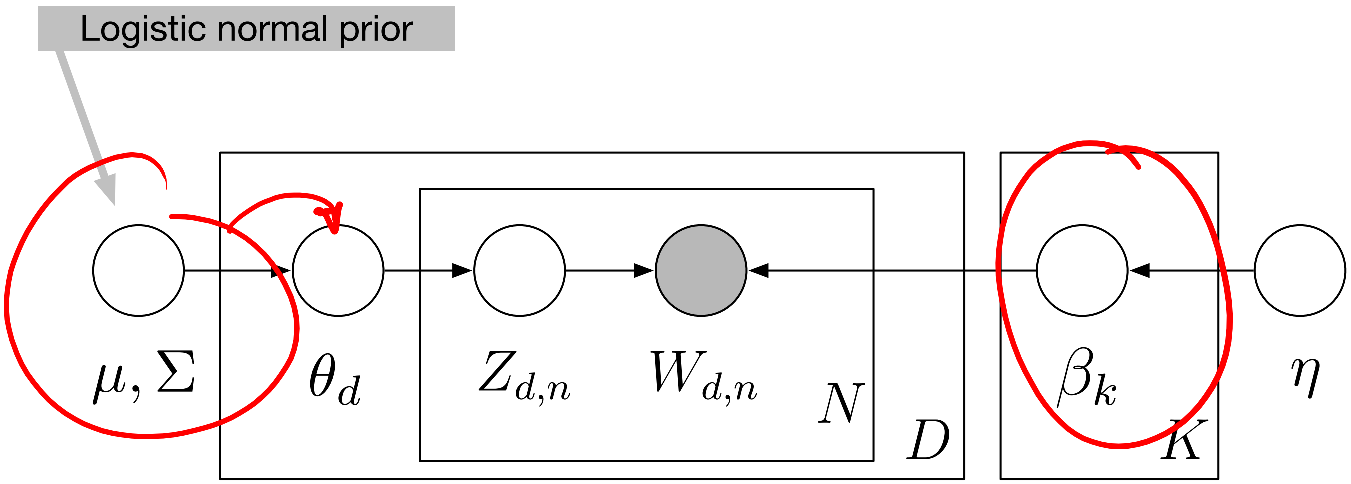
- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

$$\begin{aligned} X &\sim \mathcal{N}_K(\mu, \Sigma) \\ \theta_i &\propto \exp\{x_i\}. \end{aligned}$$

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models

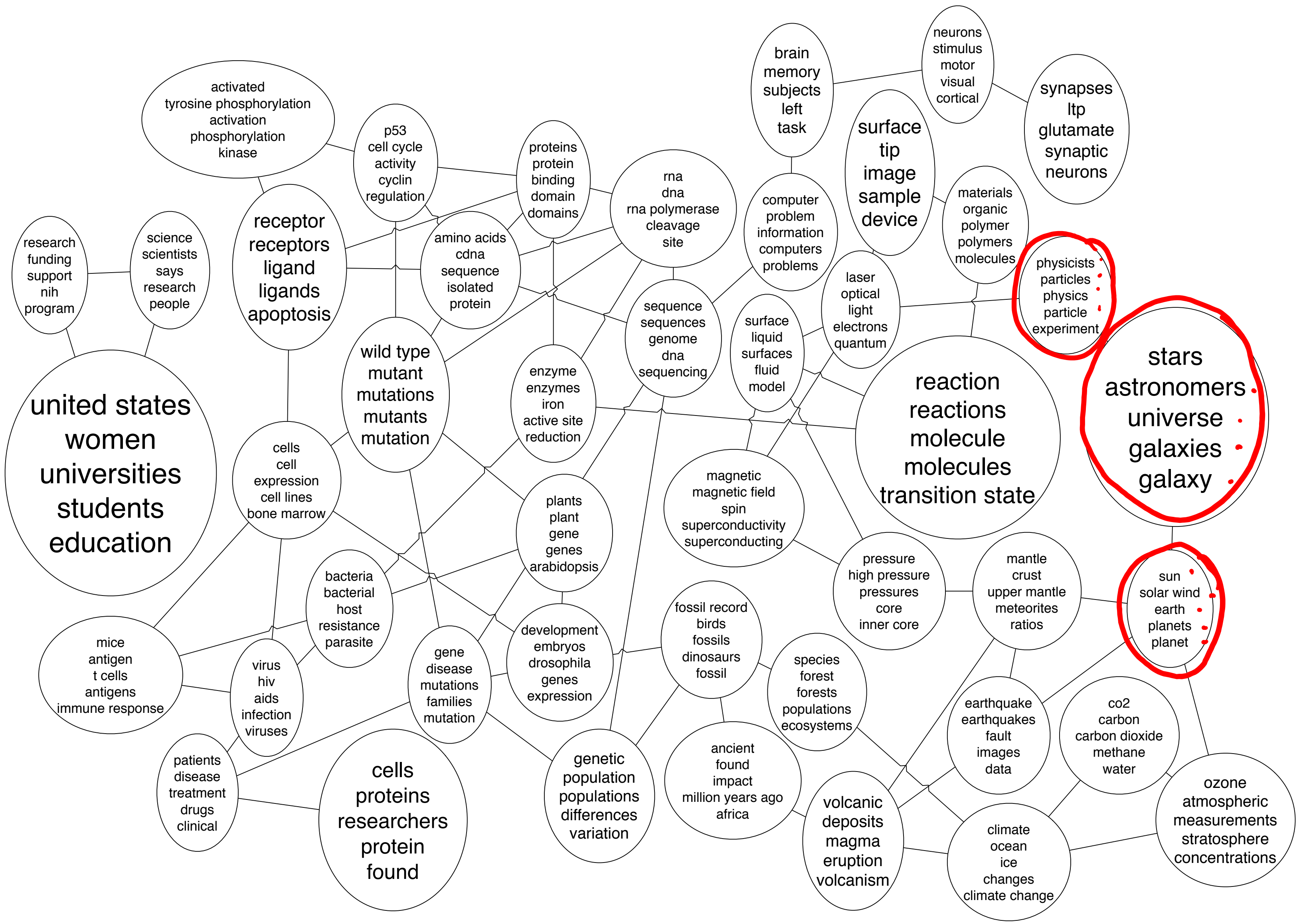


- Draw topic proportions from a logistic normal
- This allows topic occurrences to exhibit correlation.
- Provides a "map" of topics and how they are related
- Provides a better fit to text data, but computation is more complex

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models



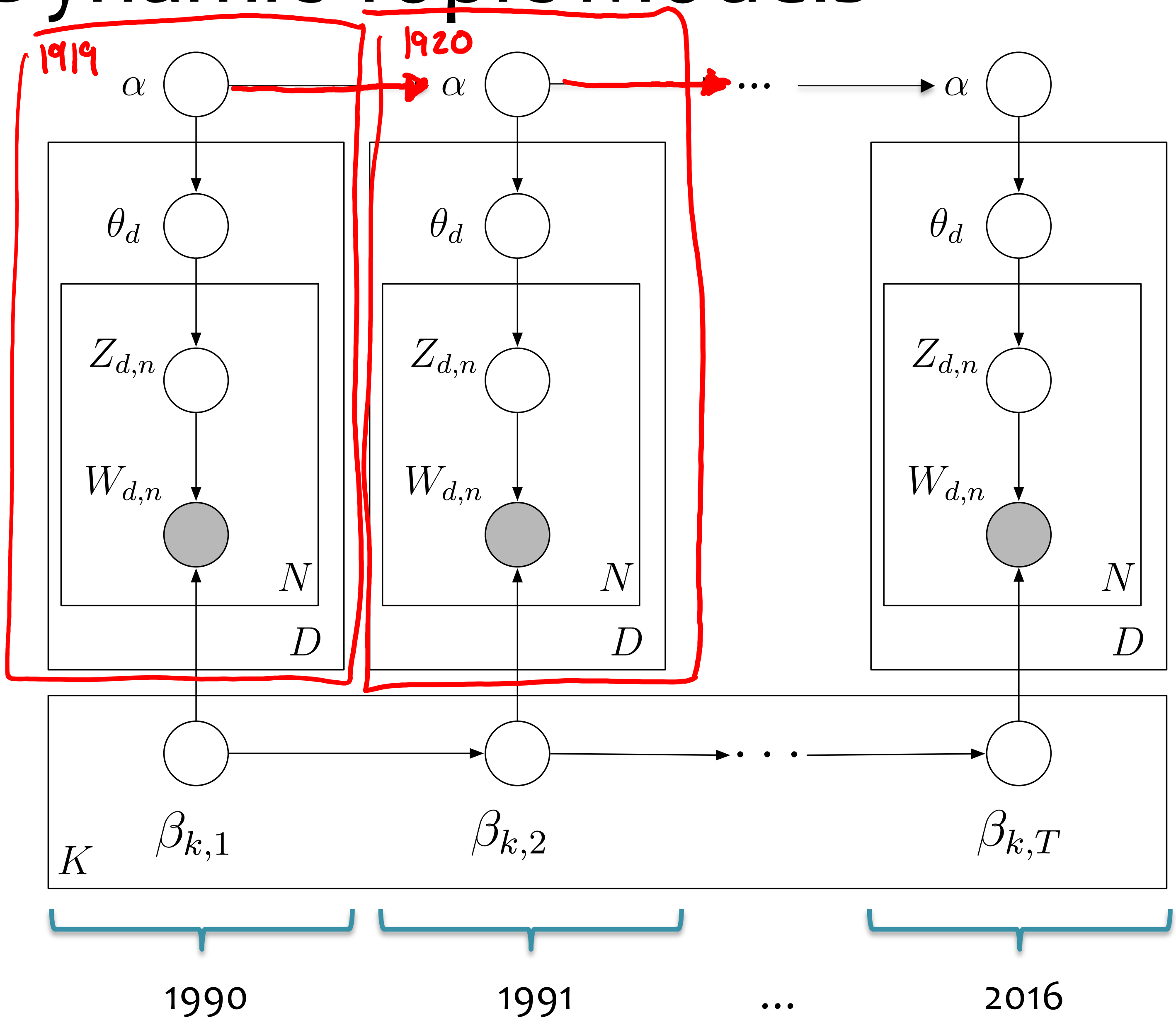
Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2006)

Dynamic Topic Models

High-level idea:

- Divide the documents up by year
- Start with a separate topic model for each year
- Then add a dependence of each year on the previous one



(Blei & Lafferty, 2006)

Dynamic Topic Models

1789

Inaugural addresses

2009

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

- LDA assumes that the order of documents does not matter.
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- Further, we may want to track how language changes over time.
- Dynamic topic models let the topics *drift* in a sequence.

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2006)

Dynamic Topic Models

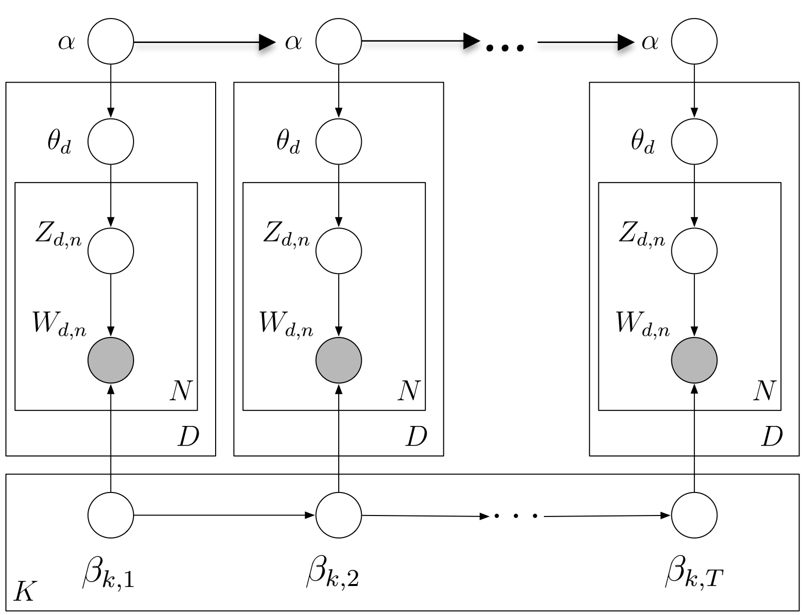
Generative Story

1. Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
   (b) For each word:
      i. Draw $Z \sim Mult(\pi(\eta))$.
      ii. Draw $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$.

Logistic-normal priors

The pi function maps from the natural parameters to the mean parameters:

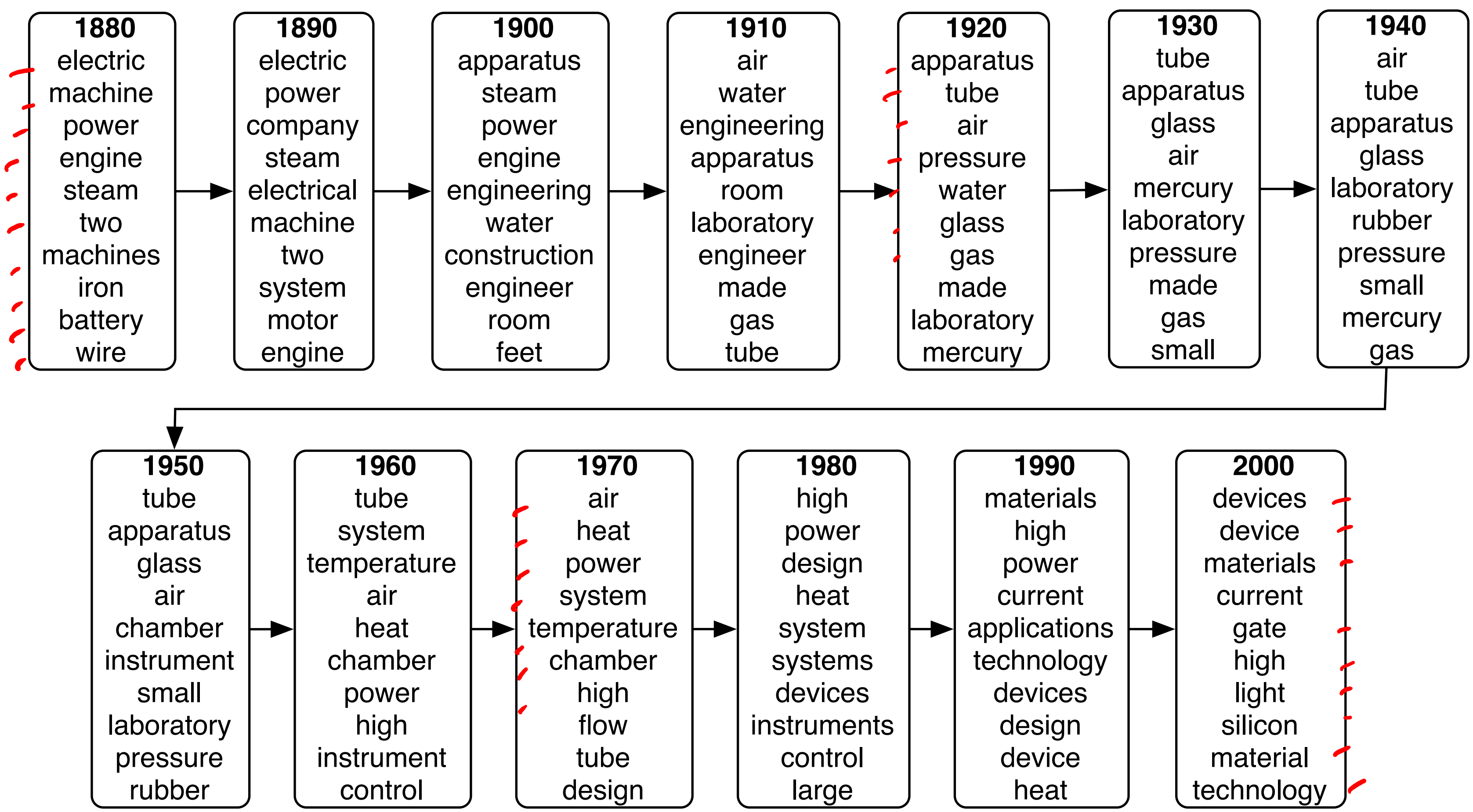
$$\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})}.$$



(Blei & Lafferty, 2006)

Dynamic Topic Models

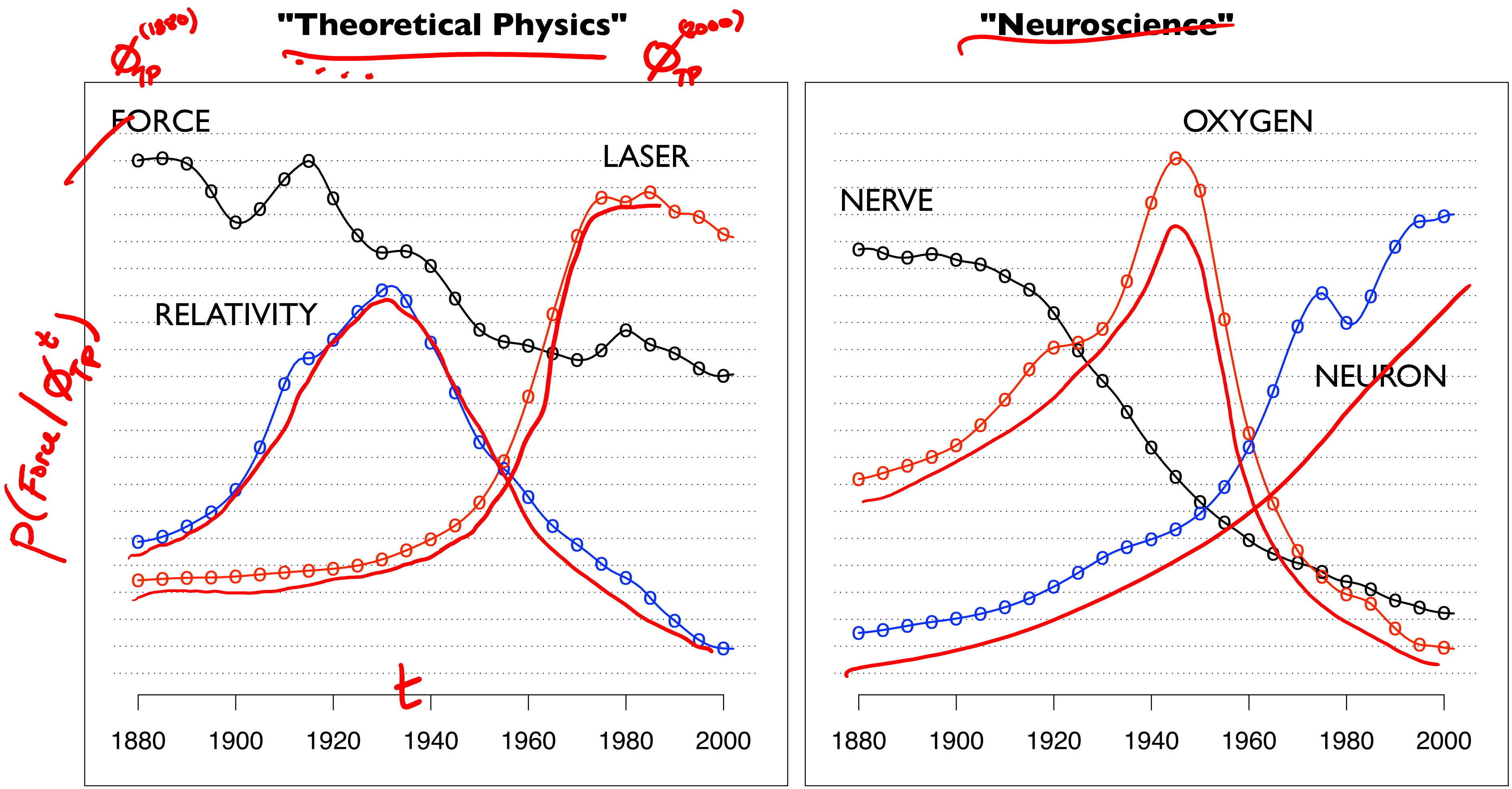
Top ten most likely words in a "drifting" topic shown at 10-year increments



(Blei & Lafferty, 2006)

Dynamic Topic Models

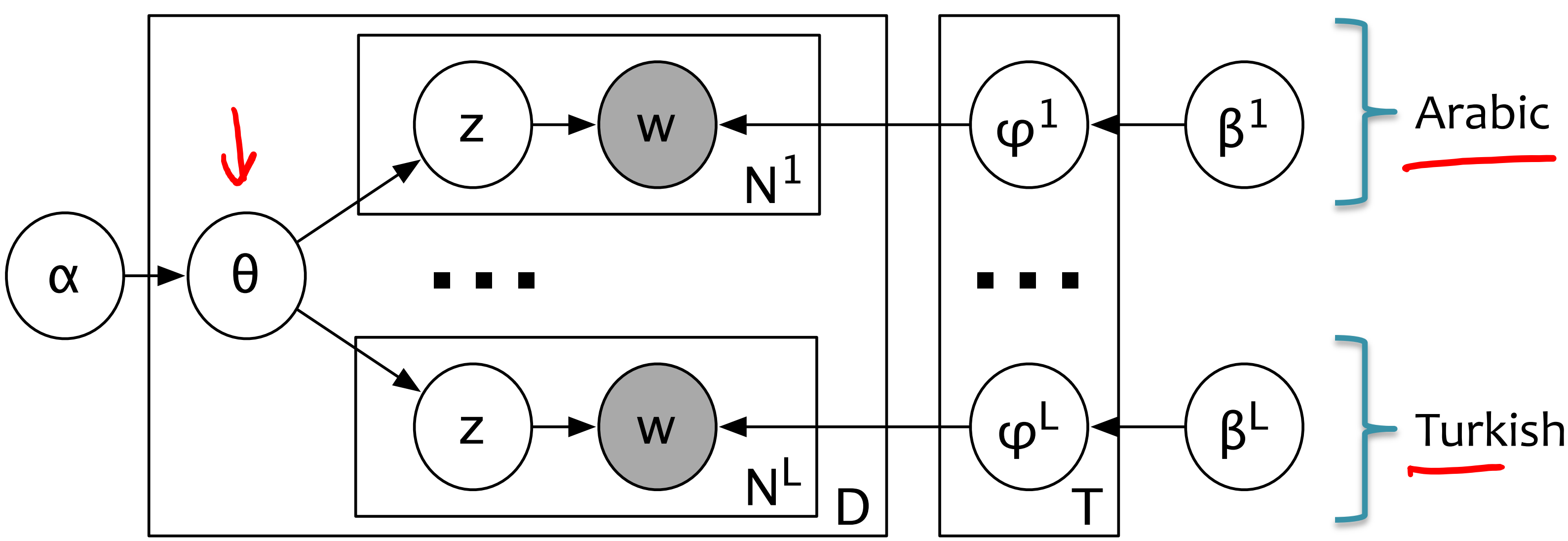
Posterior estimate of **word frequency as a function of year** for three words each in two separate topics:



(Mimno et al., 2009)

Polylingual Topic Models

- **Data Setting:** Comparable versions of each document exist in multiple languages (e.g. the Wikipedia article for "Barak Obama" in twelve languages)
- **Model:** Very similar to LDA, except that the topic assignments, z, and words, w, are sampled separately for each language.



(Mimno et al., 2009)

Polylingual Topic Models

Topic 1 (twelve languages)

CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضایی ماموریت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

(Mimno et al., 2009)

Polylingual Topic Models

Topic 2 (twelve languages)

CY	sbaen madrid el la josé sbaeneg
DE	de spanischer spanischen spanien madrid la
EL	ισπανίας ισπανία de ισπανός ντε μαδρίτη
EN	**de spanish spain la madrid y**
FA	ترین de اسپانیا اسپانیایی کوبا مادرید
FI	espanja de espanjan madrid la real
FR	espagnol espagne madrid espagnole juan y
HE	ספרד ספרדית דה מדריד הספרדית קובה
IT	de spagna spagnolo spagnola madrid el
PL	de hiszpański hiszpanii la juan y
RU	де мадрид испании испания испанский de
TR	ispanya ispanyol madrid la küba real

(Mimno et al., 2009)

Polylingual Topic Models

Topic 3 (twelve languages)

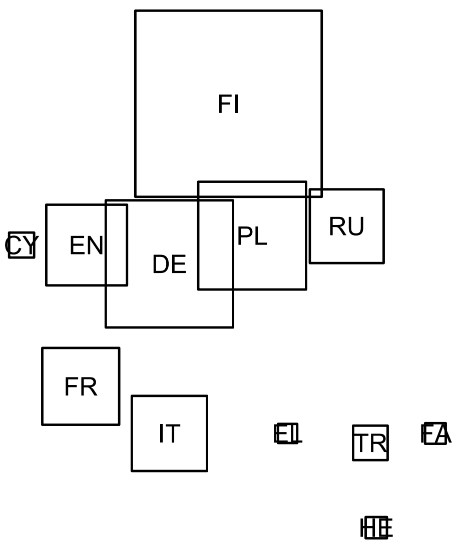
CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses
HE	משורר ספרות שירה סופר שירים המשורר
IT	poeta letteratura poesia opere versi poema
PL	poeta literatury poezji pisarz in jego
RU	поэт его писатель литературы поэзии драматург
TR	şair edebiyat şiir yazar edebiyatı adlı

(Mimno et al., 2009)

Polylingual Topic Models

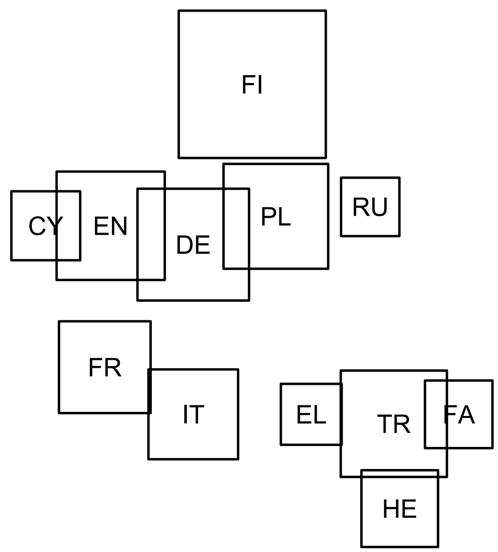
Size of each square represents proportion of tokens assigned to the specified topic.

world
ski
km
won



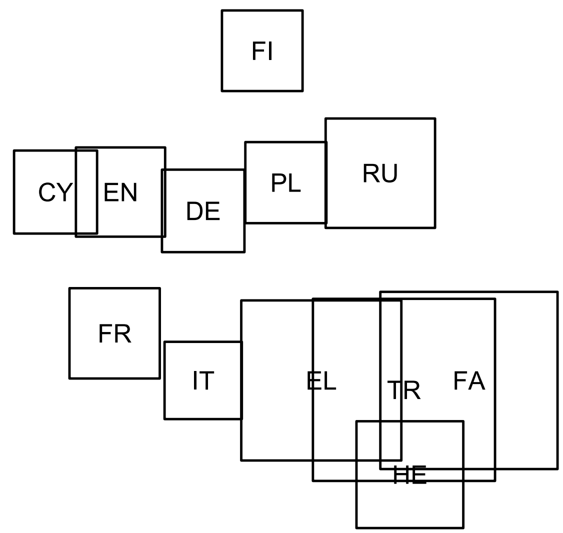
Analysis: mostly Nordic countries

actor
role
television
actress



Analysis: uniform across countries

ottoman
empire
khan
byzantine



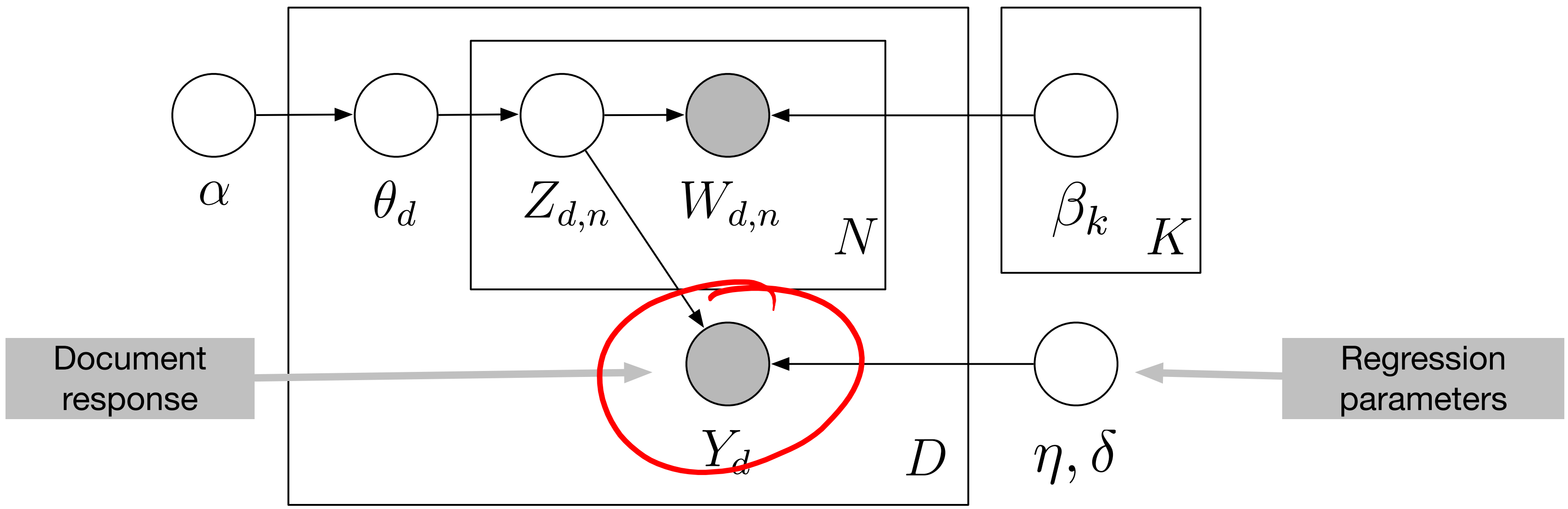
Analysis: mostly countries near Istanbul

Supervised LDA

- LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?
- Many data are paired with **response variables**.
  - User reviews paired with a number of stars
  - Web pages paired with a number of “likes”
  - Documents paired with links to other documents
  - Images paired with a category
- **Supervised LDA** are topic models of documents and responses. They are fit to find topics predictive of the response.

Slide from David Blei, MLSS 2012

Supervised LDA



1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \mathrm{N}\left(\eta^\top \bar{z}, \sigma^2\right)$, where

$$\bar{z} = (1/N) \sum_{n=1}^{N} z_n.$$

Slide from David Blei, MLSS 2012

Summary: Topic Modeling

- **The Task of Topic Modeling**
  - Topic modeling enables the **analysis of large** (possibly unannotated) **corpora**
  - Applicable to more than just bags of words
  - Extrinsic evaluations are often appropriate for these unsupervised methods
- **Constructing Models**
  - LDA is comprised of **simple building blocks** (Dirichlet, Multinomial)
  - LDA itself can act as a building block **for other models**
- **Approximate Inference**
  - Many different approaches to inference (and learning) can be applied to the same model

What if we don't know the number of topics, K, ahead of time?

Solution: Bayesian Nonparametrics

- New modeling constructs:
  - Chinese Restaurant Process (Dirichlet Process)
  - Indian Buffet Process
- e.g. an **infinite number of topics** in a finite amount of space

Summary: Approximate Inference

- Markov Chain Monte Carlo (MCMC)
  - Metropolis-Hastings, Gibbs sampling, Hamiltonion MCMC, slice sampling, etc.
- Variational inference
  - Minimizes KL(q||p) where q is a simpler graphical model than the original p
- Loopy Belief Propagation
  - Belief propagation applied to general (loopy) graphs
- Expectation propagation
  - Approximates belief states with moments of simpler distributions
- Spectral methods
  - Uses tensor decompositions (e.g. SVD)

CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Network (CNN)

- Typical layers include:
  - Convolutional layer
  - Max-pooling layer
  - Fully-connected (Linear) layer
  - ReLU layer (or some other nonlinear activation function)
  - Softmax
- These can be arranged into arbitrarily deep topologies

Architecture #1: LeNet-5

PROC. OF THE IEEE, NOVEMBER 1998 7

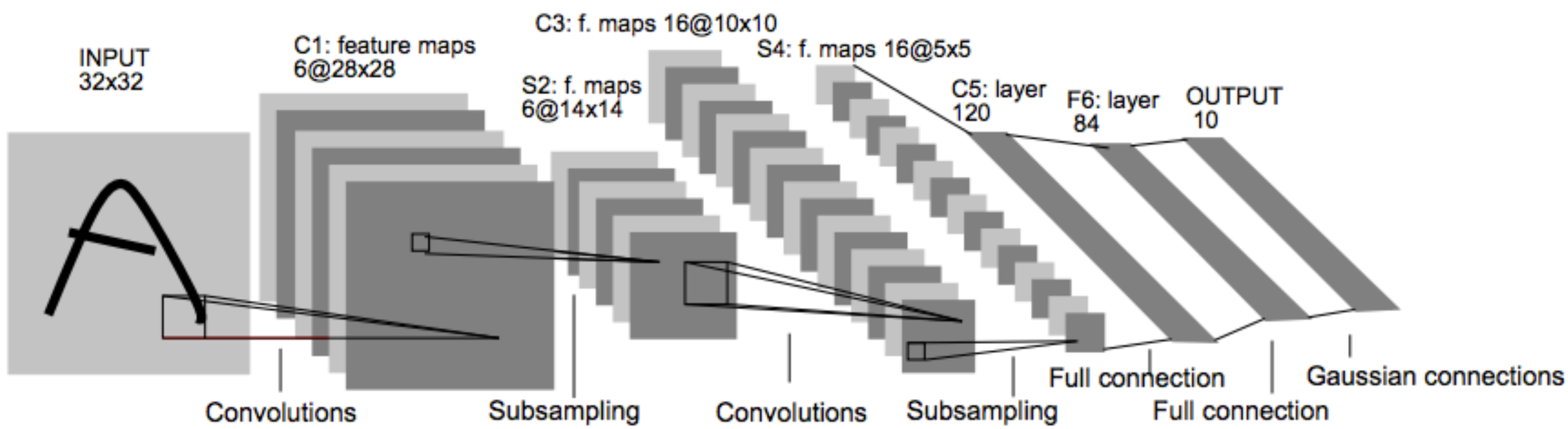


Fig. 2. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Convolutional Neural Network (CNN)

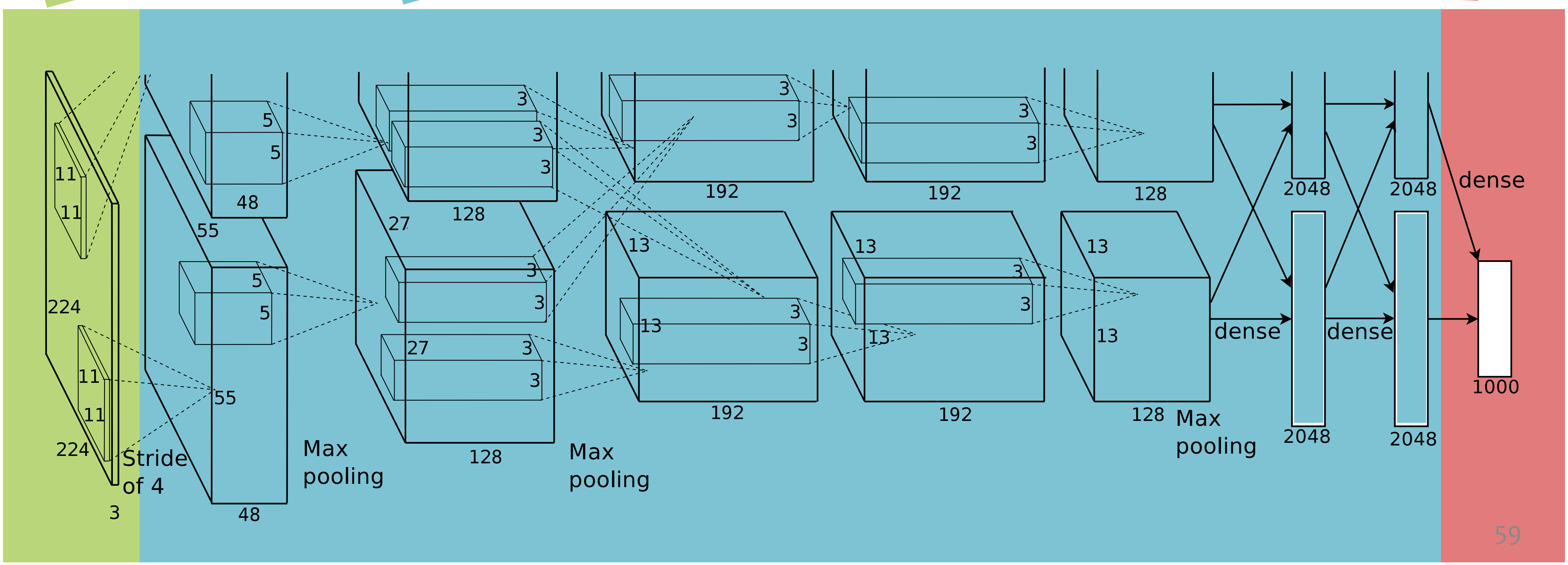
Architecture #2: AlexNet

CNN for Image Classification
(Krizhevsky, Sutskever & Hinton, 2012)
15.3% error on ImageNet LSVRC-2012 contest

Input image (pixels)

- Five convolutional layers (w/max-pooling)
- Three fully connected layers

1000-way softmax



Convolutional Neural Network (CNN)

Typical Architectures

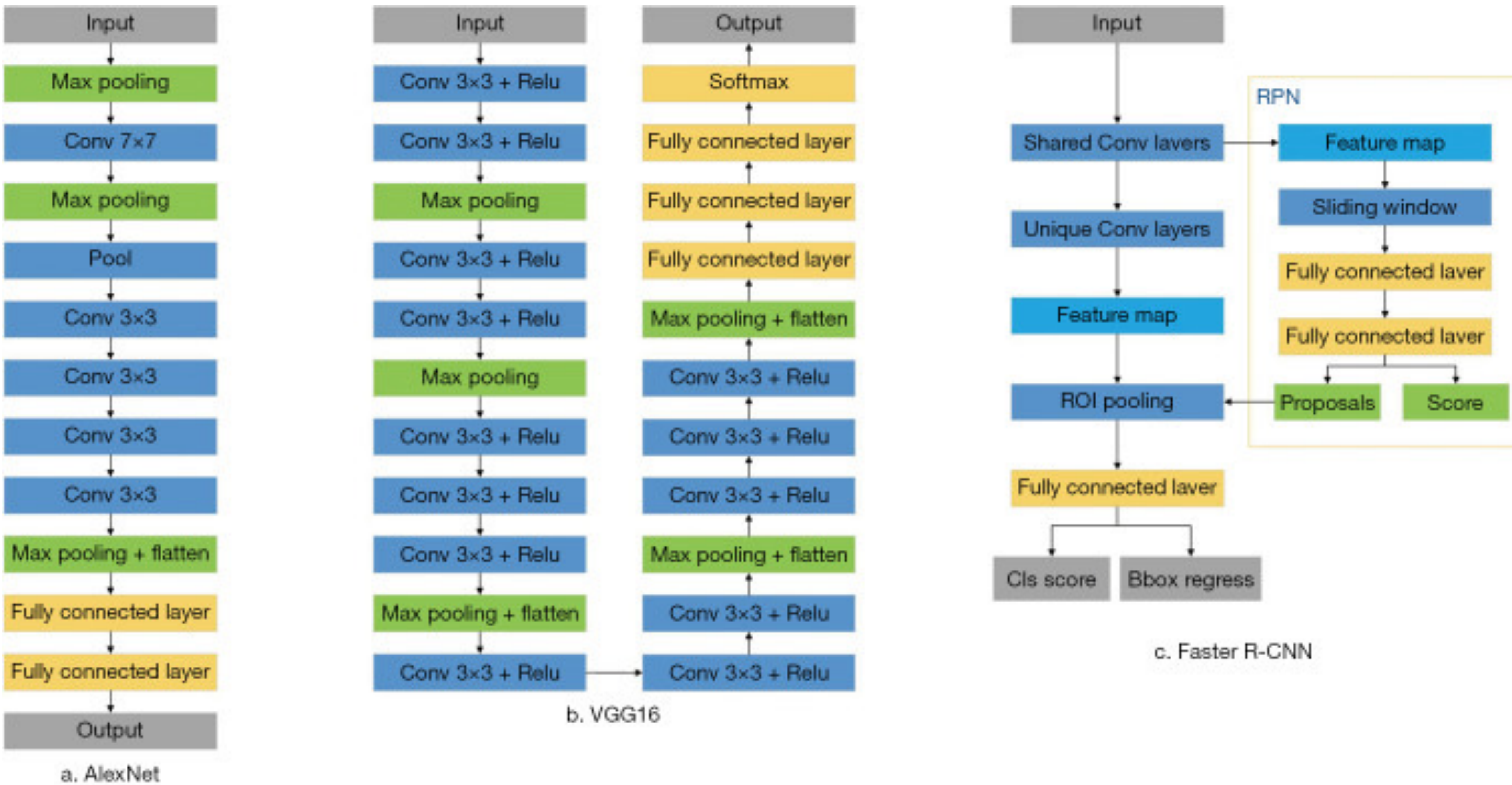


Figure from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7327346/

Convolutional Neural Network (CNN)

Typical Architectures

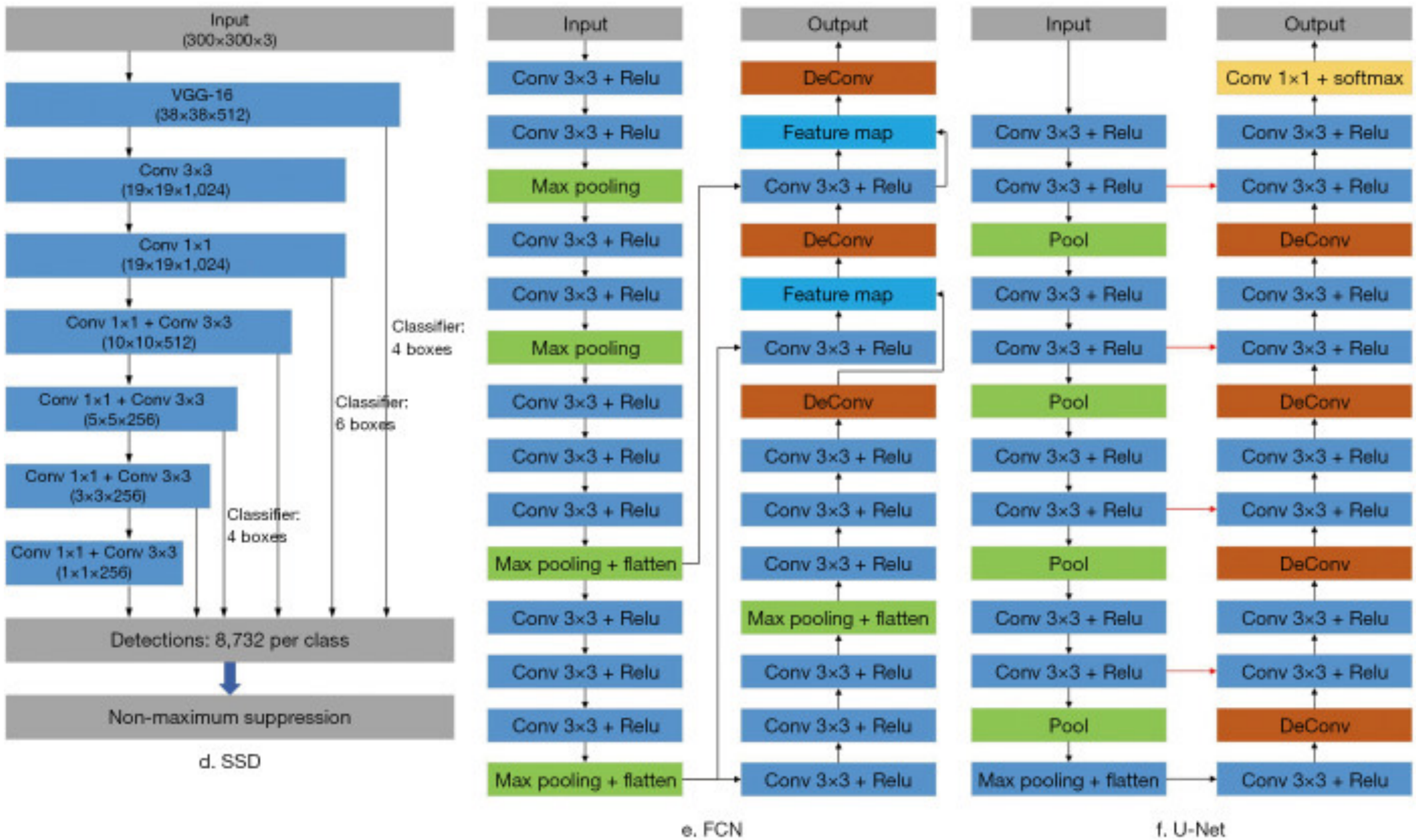


Figure from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7327346/

Slides in this section are from Kaiming He, “Deep Residual Learning”, ICCV 2015

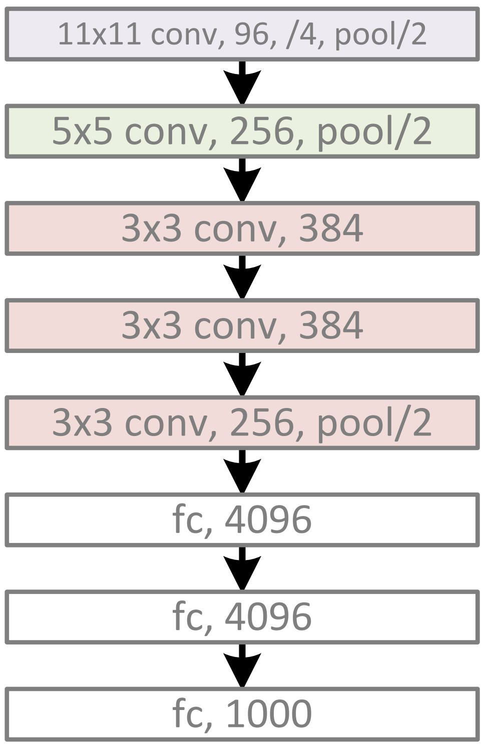
RESNET

ResNet

Microsoft Research

Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)



ICCV15 International Conference on Computer Vision

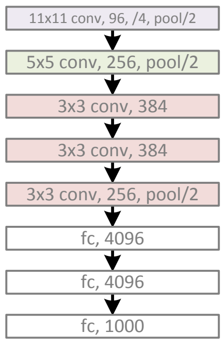
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

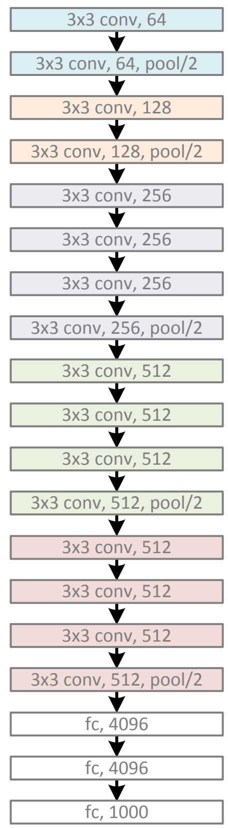
Microsoft Research

Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)



GoogleNet, 22 layers
(ILSVRC 2014)

ICCV15 International Conference on Computer Vision

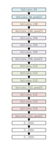
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

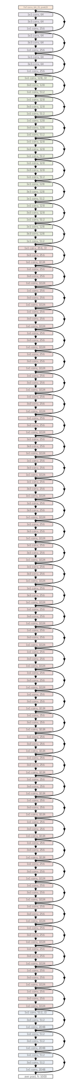
Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)

VGG, 19 layers
(ILSVRC 2014)



ResNet, 152 layers
(ILSVRC 2015)

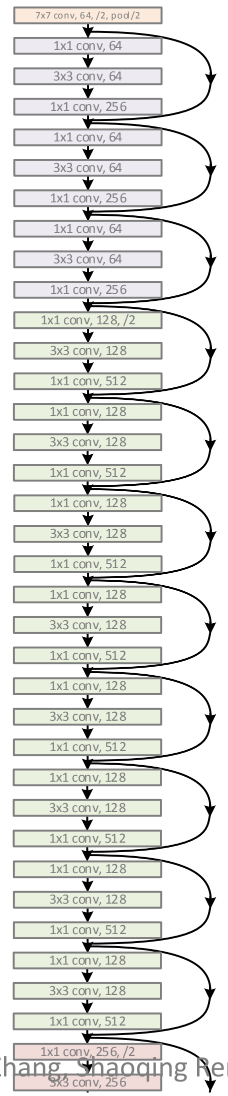


Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

Revolution of Depth

ResNet, 152 layers



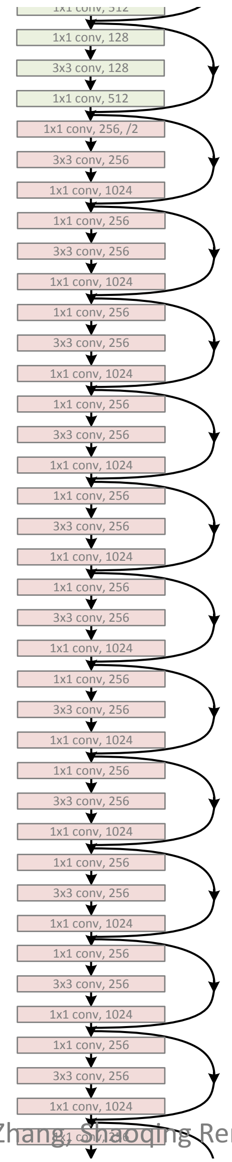
(there was an animation here)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

Revolution of Depth

ResNet, 152 layers



(there was an animation here)

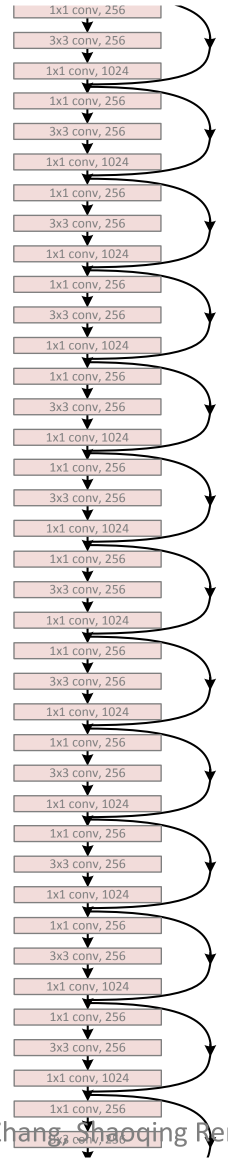
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

Microsoft Research

Revolution of Depth

ResNet, 152 layers



(there was an animation here)

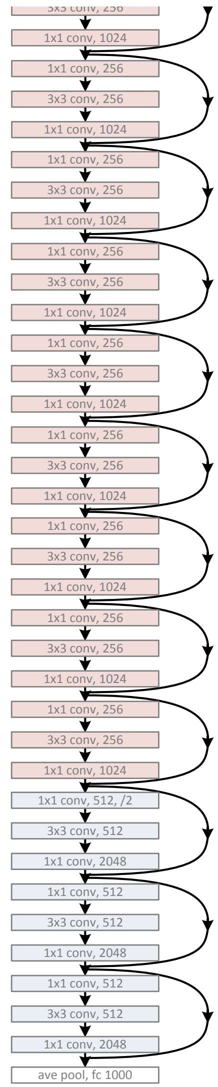
ICCV15 International Conference on Computer Vision

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

Revolution of Depth

ResNet, 152 layers



(there was an animation here)

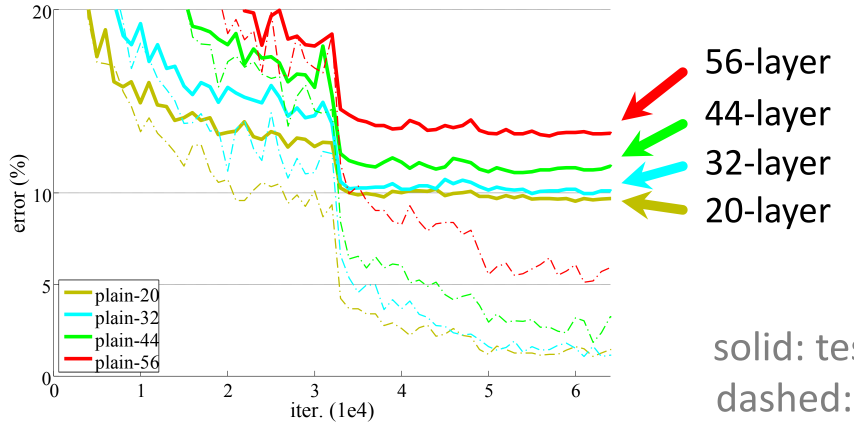
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

Microsoft Research

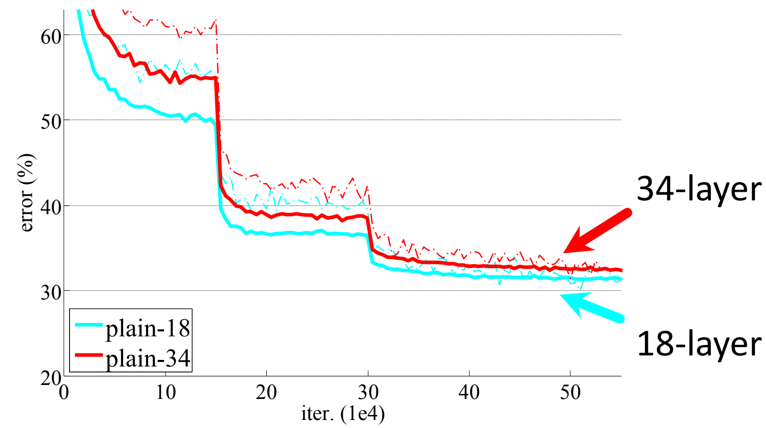
Simply stacking layers?

CIFAR-10



solid: test/val
dashed: train

ImageNet-1000



- “Overly deep” plain nets have **higher training error**
- A general phenomenon, observed in many datasets

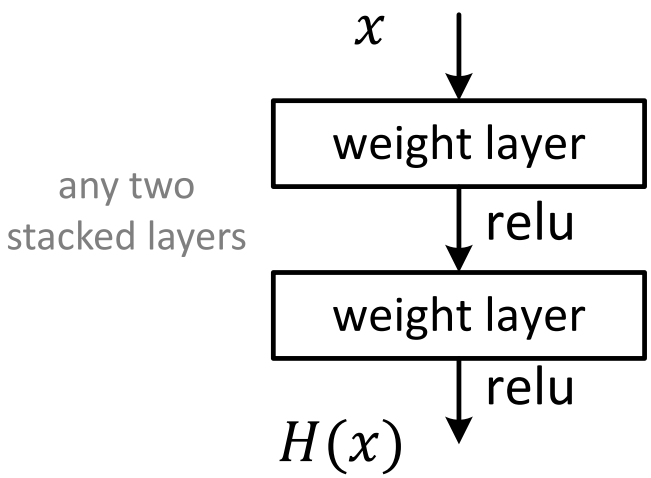
ICCV15 International Conference on Computer Vision

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

Deep Residual Learning

- Plaint net



$H(x)$ is any desired mapping,

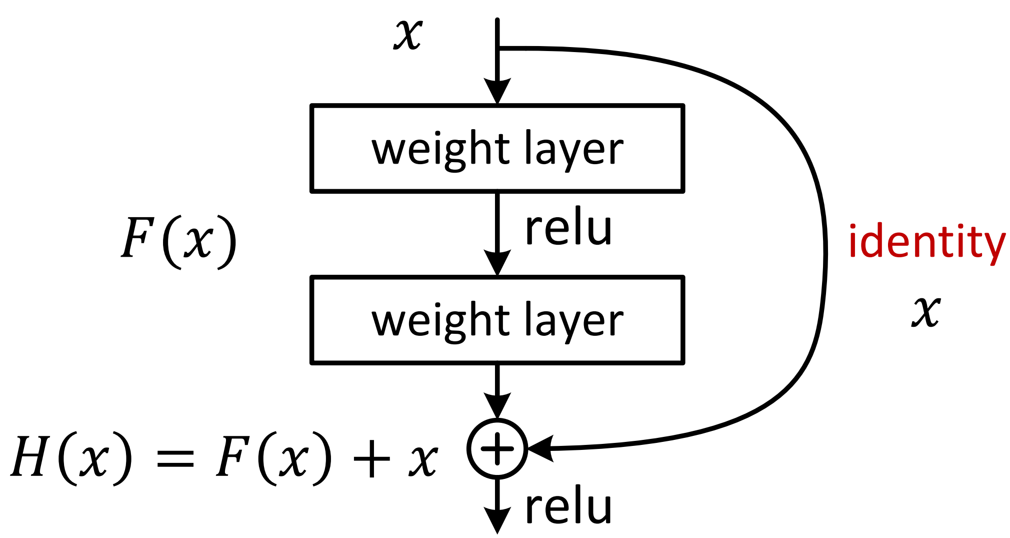
hope the 2 weight layers fit $H(x)$

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

ResNet

Deep Residual Learning

- Residual net



$H(x)$ is any desired mapping,

~~hope the 2 weight layers fit $H(x)$~~

hope the 2 weight layers fit $F(x)$

let $H(x) = F(x) + x$

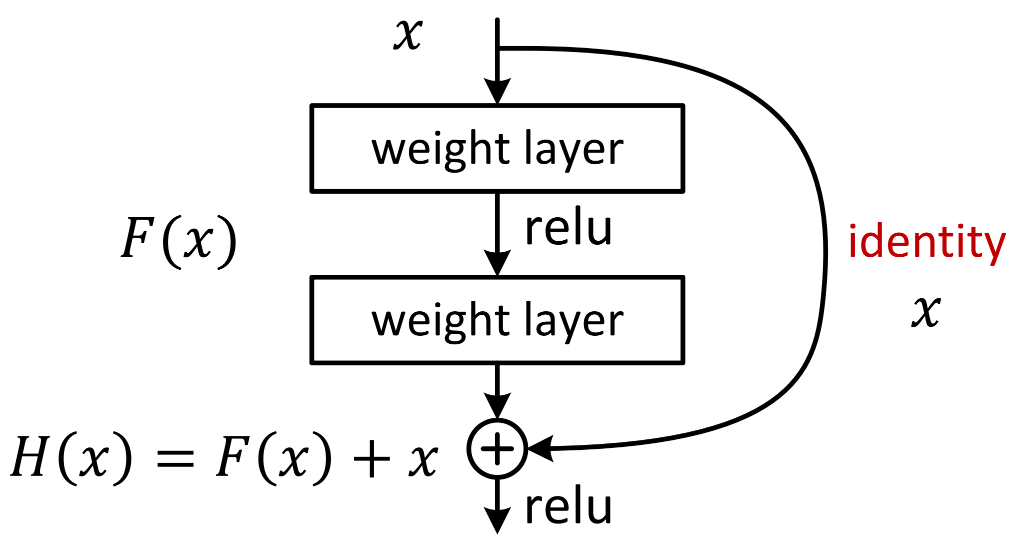
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

Microsoft Research

Deep Residual Learning

- $F(x)$ is a residual mapping w.r.t. identity



- If identity were optimal, easy to set weights as 0
- If optimal mapping is closer to identity, easier to find small fluctuations

ICCV15 International Conference on Computer Vision

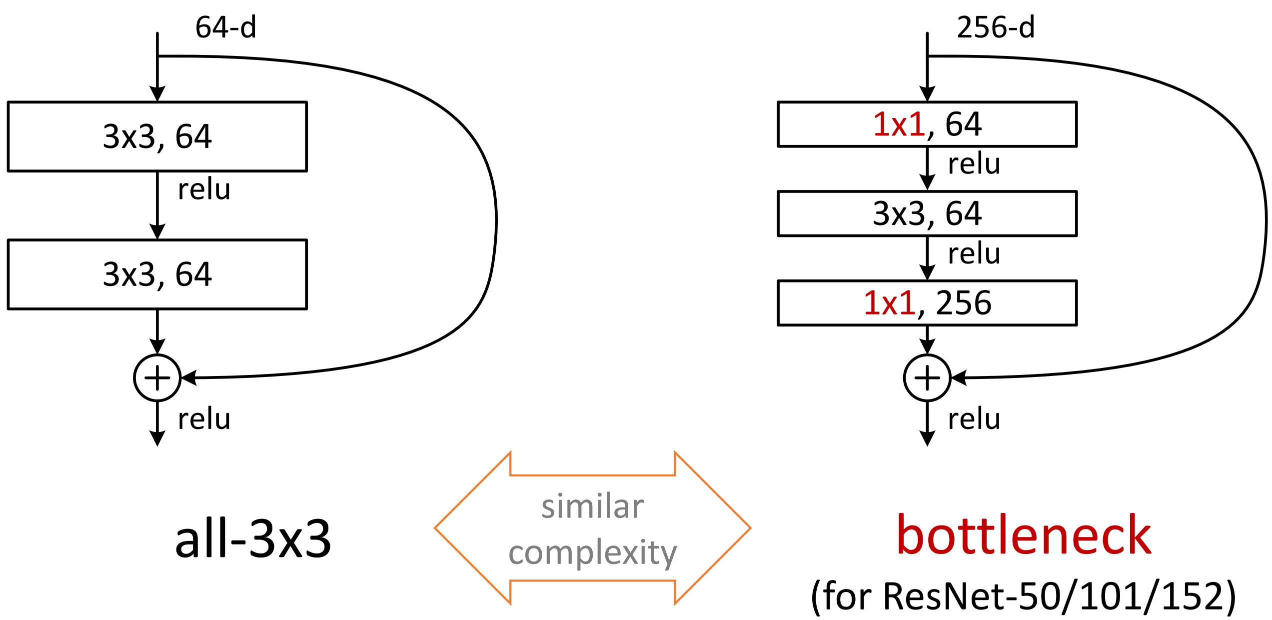
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

ResNet

Microsoft Research

ImageNet experiments

- A practical design of going deeper



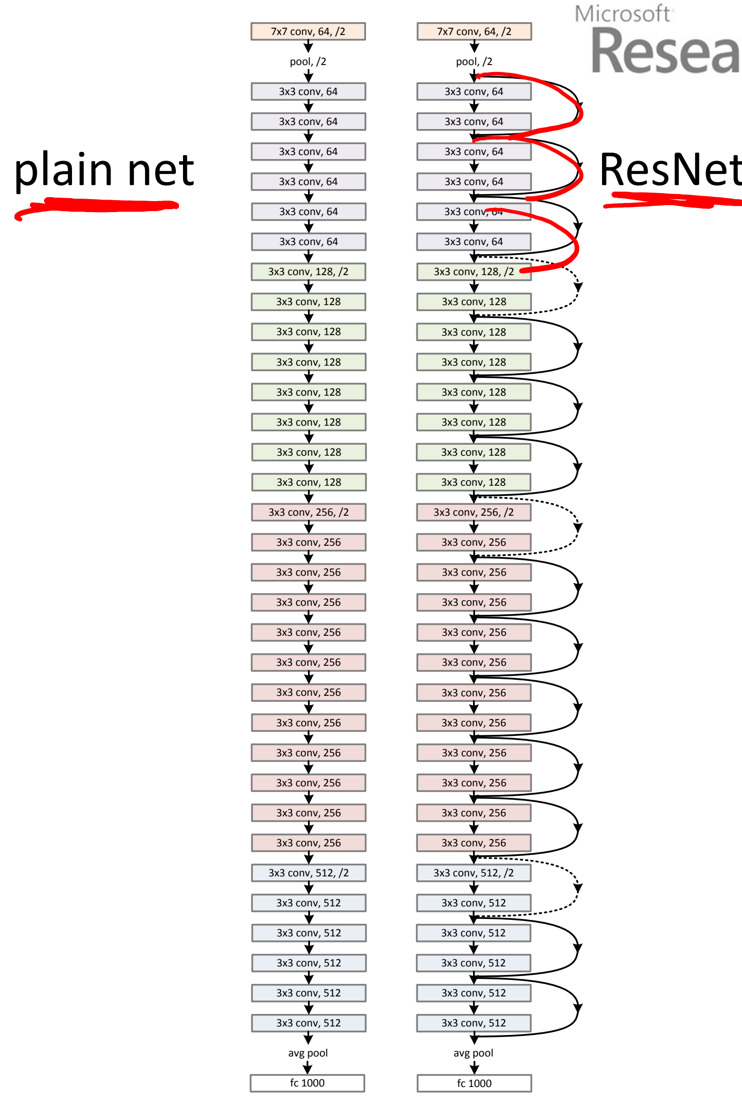
ICCV15 International Conference on Computer Vision

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

Network “Design”

- Keep it simple
- Our basic design (VGG-style)
  - all 3x3 conv (almost)
  - spatial size /2 => # filters x2
  - Simple design; just deep!
- Other remarks:
  - no max pooling (almost)
  - no hidden fc
  - no dropout



Microsoft Research

ICCV15 International Conference on Computer Vision

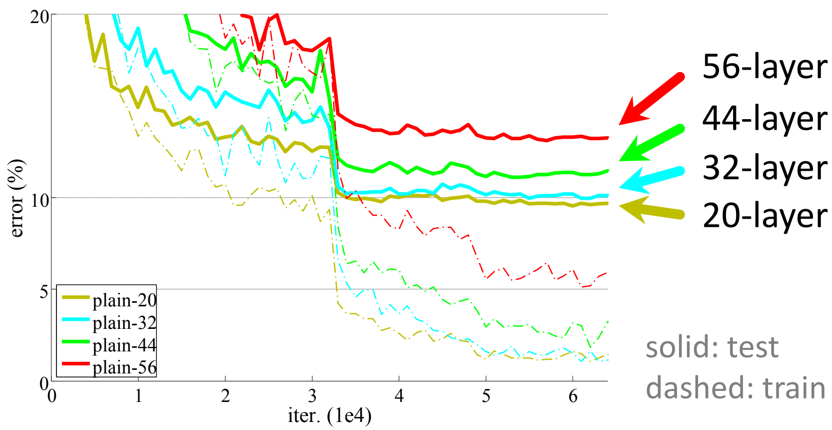
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

ResNet

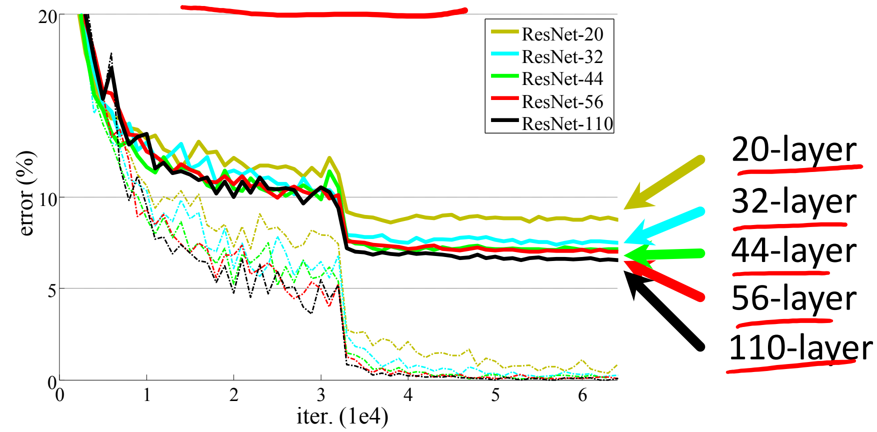
Microsoft Research

CIFAR-10 experiments

CIFAR-10 plain nets



CIFAR-10 ResNets



- Deep ResNets can be trained without difficulties
- Deeper ResNets have **lower training error**, and also lower test error

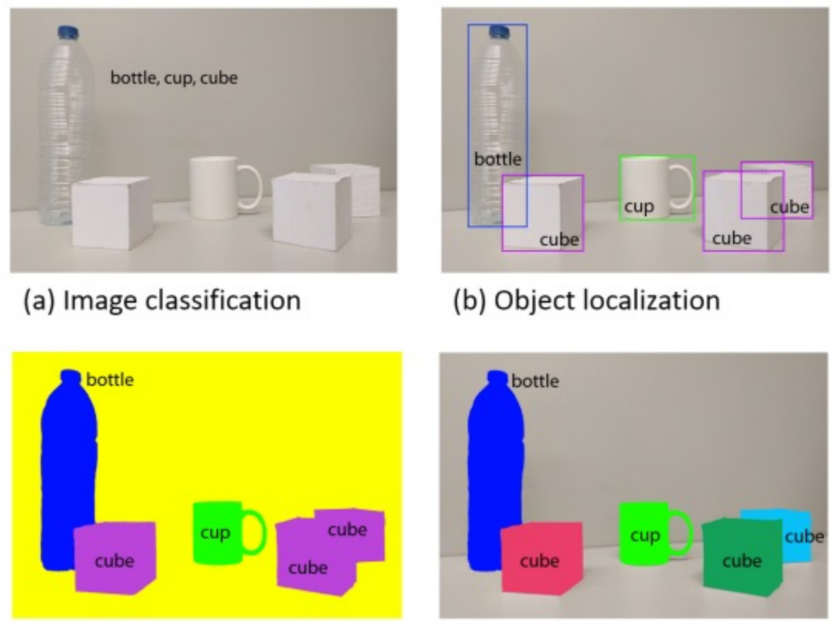
ICCV15 International Conference on Computer Vision

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

COMPUTER VISION

Common Tasks in Computer Vision

1. Image Classification
2. Image Classification + Localization
3. Human Pose Estimation
4. Semantic Segmentation
5. Object Detection
6. Instance Segmentation
7. Image Captioning



(a) Image classification

(b) Object localization

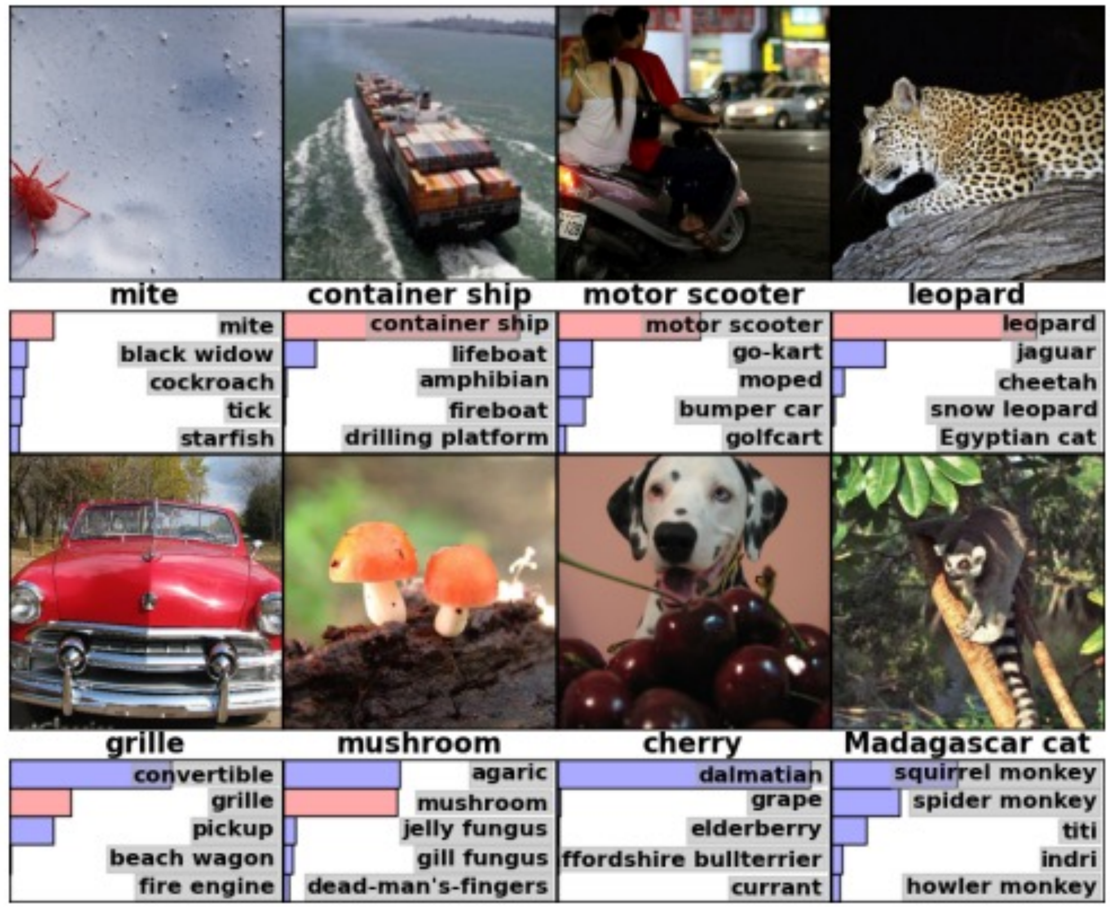
(c) Semantic segmentation

(d) Instance segmentation

Figure from https://arxiv.org/pdf/1704.06857.pdf

Image Classification

- Given an image, predict a single label
- A multi-class classification problem



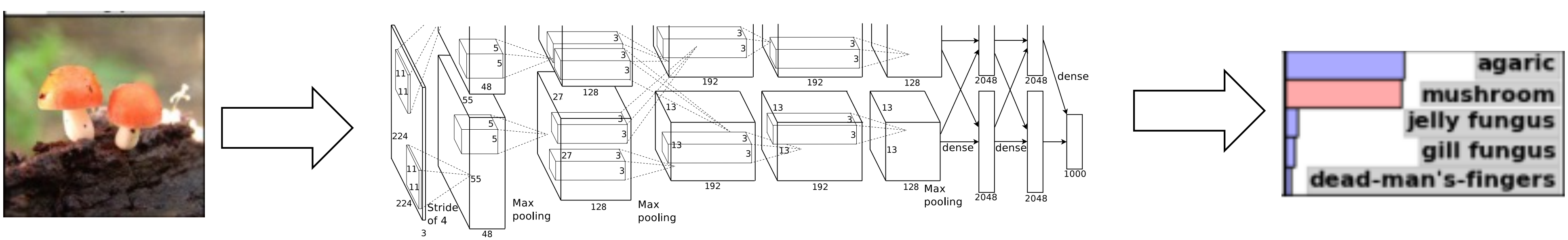


Figure from https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Image Classification + Localization

- Given an image, predict a single label and a bounding box for the object
- Bounding box is represented as (x, y, h, w), position (x,y) and height/width (h,w)

Figure from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5459257

Human Pose Estimation

- Given an image of a human, predict the position of several keypoints (left hand, right hand, left elbow, … , right foot)
- This is a multiple regression problem, where each keypoint has a corresponding position (x_i, y_i)

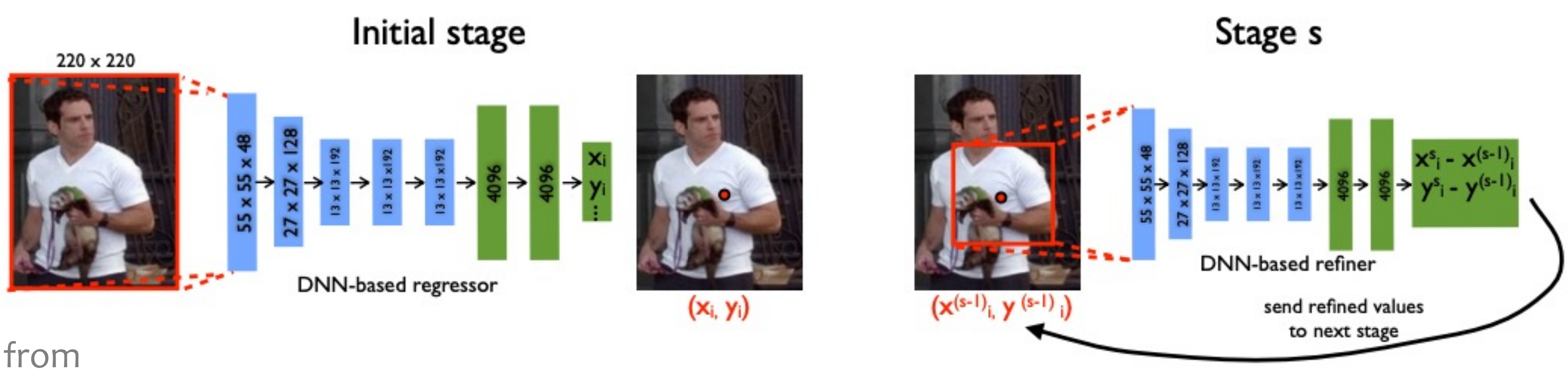
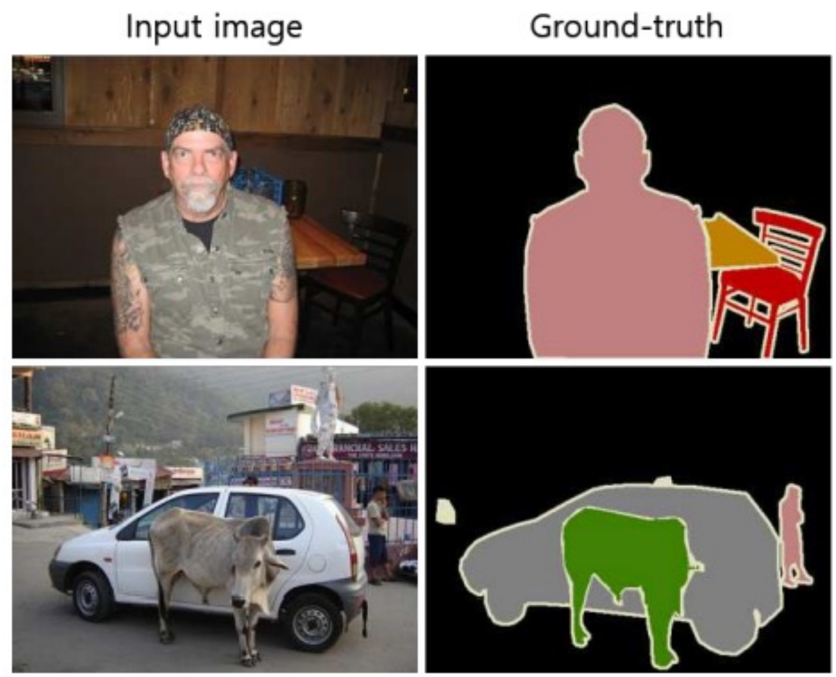


Figure from https://openaccess.thecvf.com/content_cvpr_2014/papers/Toshev_DeepPose_Human_Pose_2014_CVPR_paper.pdf

Semantic Segmentation

- Given an image, predict a label for every pixel in the image
- Not merely a classification problem, because there are strong correlations between pixel-specific labels



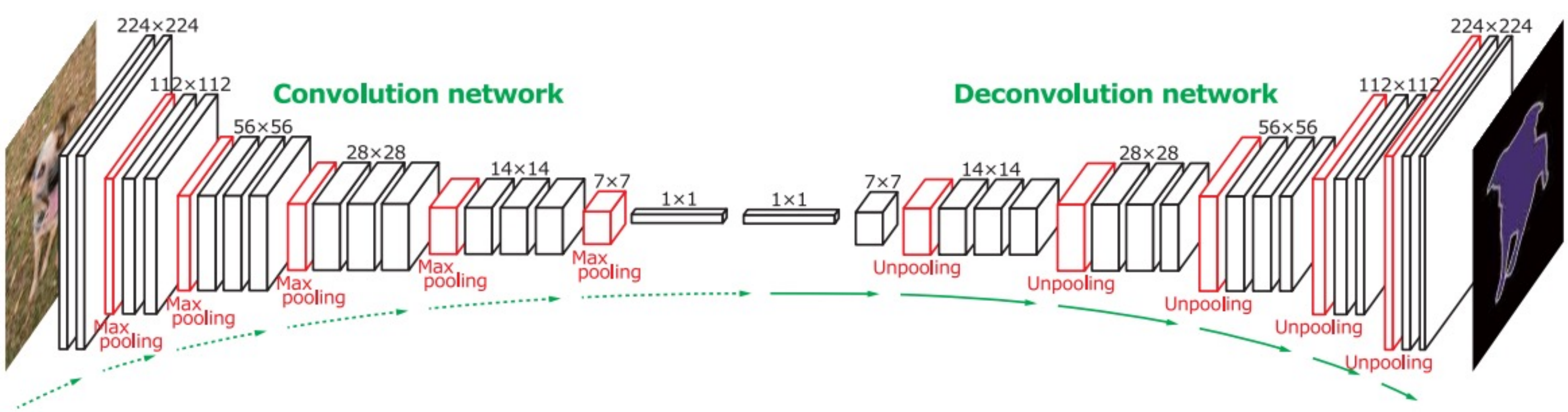


Figure from https://openaccess.thecvf.com/content_iccv_2015/papers/Noh_Learning_Deconvolution_Network_ICCV_2015_paper.pdf

Object Detection

- Given an image, for each object predict a bounding box and a label (x,y,w,h,l)
- Example: R-CNN
  - (x=110, y=13, w=50, h=72, l=person)
  - (x=90, y=55, w=81, h=87, l=horse)
  - (x=421, y=533, w=24, h=30, l=chair)
  - (x=2, y=25, w=51, h=121, l=gate)

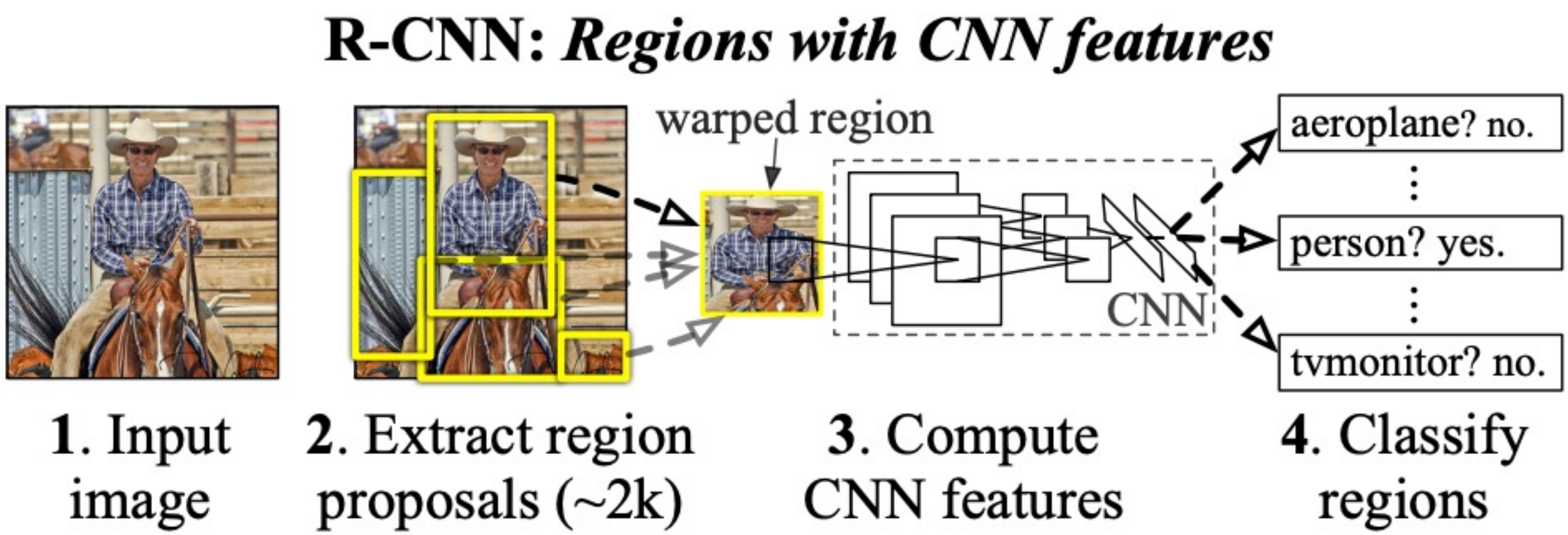


Figure from https://openaccess.thecvf.com/content_cvpr_2014/papers/Girshick_Rich_Feature_Hierarchies_2014_CVPR_paper.pdf

Instance Segmentation

- Predict per-pixel labels as in semantic segmentation, but differentiate between different instances of the same label
- *Example*: if there are two people in the image, one person should be labeled **person-1** and one should be labeled **person-2**

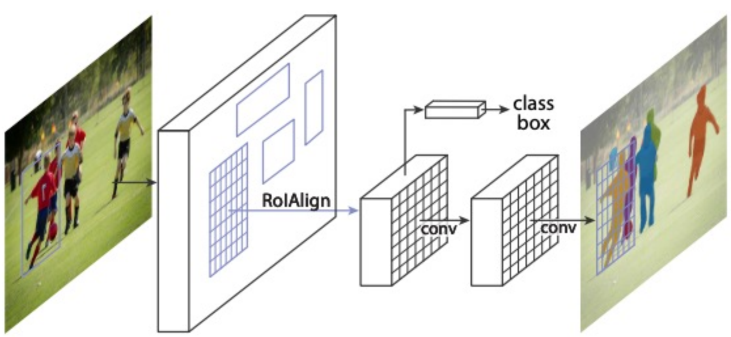
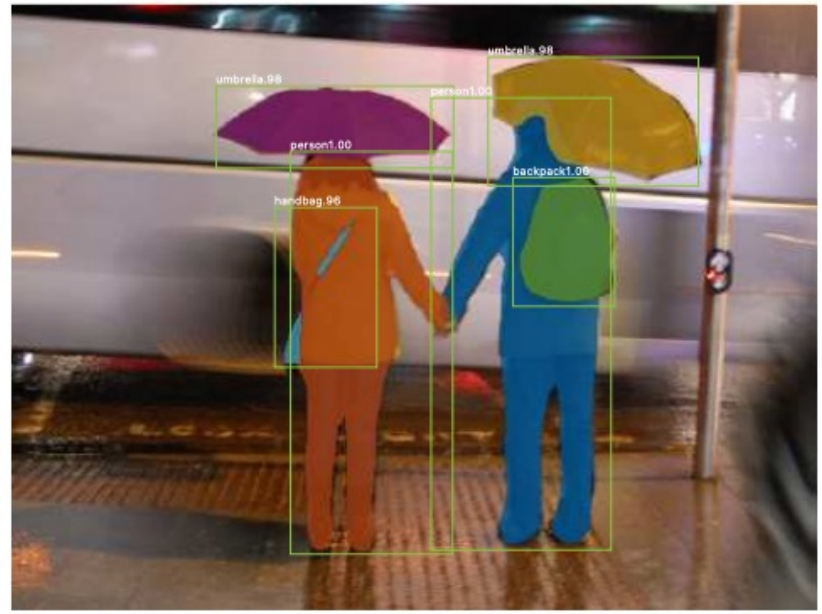


Figure 1. The **Mask R-CNN** framework for instance segmentation.



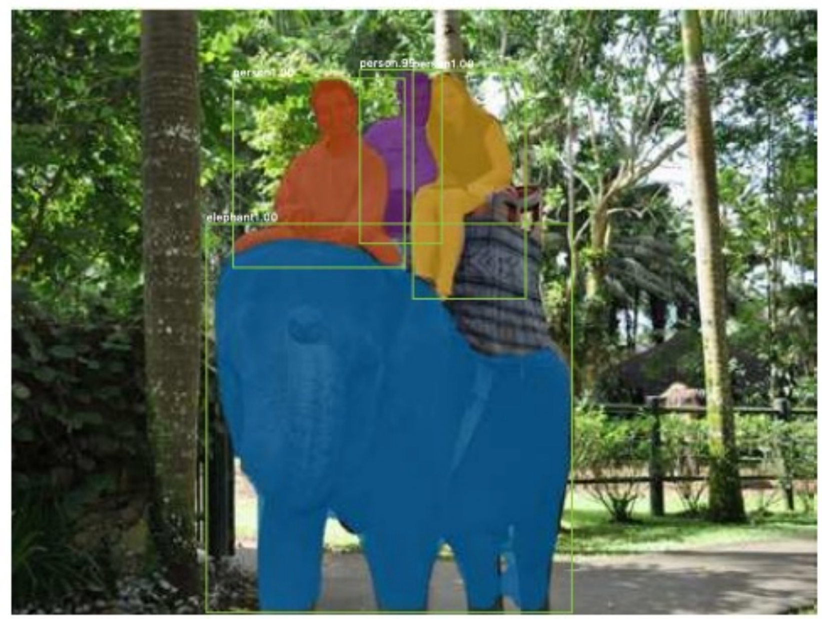
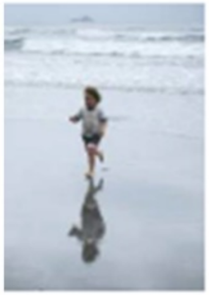


Figure from https://openaccess.thecvf.com/content_ICCV_2017/papers/He_Mask_R-CNN_ICCV_2017_paper.pdf

Image Captioning



Ground Truth Caption: A little boy runs away from the approaching waves of the ocean.

Generated Caption: A young boy is running on the beach.

Ground Truth Caption: A brunette girl wearing sunglasses and a yellow shirt.

Generated Caption: A woman in a black shirt and sunglasses smiles.

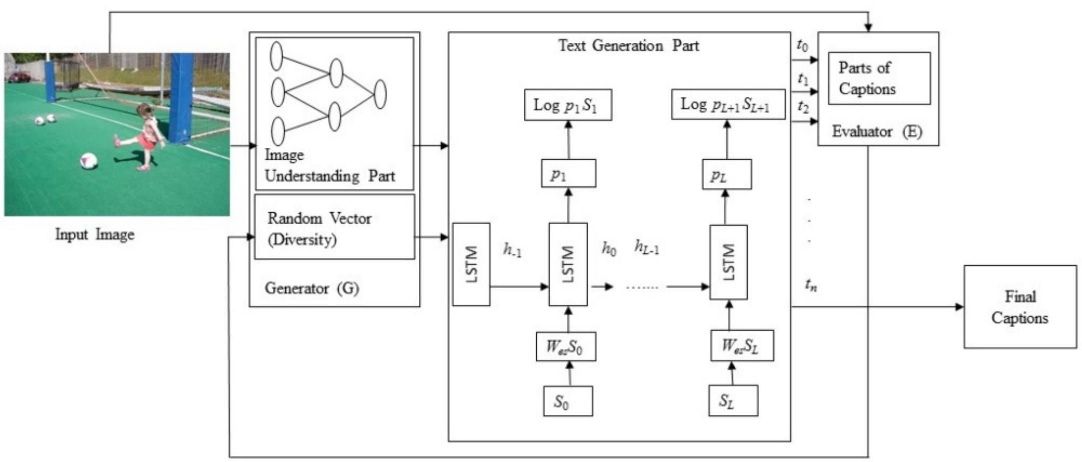


Fig. 3. A block diagram of other deep-learning-based captioning.

- Take an image as input, and generate a sentence describing it as output (i.e. the caption)
- Typical methods include a deep CNN/transformer and a RNN-like language model
- (The task of *Dense Captioning* is to generate one caption per bounding box)

Figure from https://dl.acm.org/doi/pdf/10.1145/3295748

Image Captioning

Table 1. An Overview of the Deep-Learning-Based Approaches for Image Captioning

Reference	Image Encoder	Language Model	Category
Kiros et al. 2014 [69]	AlexNet	LBL	MS, SL, WS, EDA
Kiros et al. 2014 [70]	AlexNet, VGGNet	1. LSTM 2. SC-NLM	MS, SL, WS, EDA
Mao et al. 2014 [95]	AlexNet	RNN	MS, SL, WS
Karpathy et al. 2014 [66]	AlexNet	DTR	MS, SL, WS, EDA
Mao et al. 2015 [94]	AlexNet, VGGNet	RNN	MS, SL, WS
Chen et al. 2015 [23]	VGGNet	RNN	VS, SL, WS, EDA
Fang et al. 2015 [33]	AlexNet, VGGNet	MELM	VS, SL, WS, CA
Jia et al. 2015 [59]	VGGNet	LSTM	VS, SL, WS, EDA
Karpathy et al. 2015 [65]	VGGNet	RNN	MS, SL, WS, EDA
Vinyals et al. 2015 [142]	GoogLeNet	LSTM	VS, SL, WS, EDA
Xu et al. 2015 [152]	AlexNet	LSTM	VS, SL, WS, EDA, AB
Jin et al. 2015 [61]	VGGNet	LSTM	VS, SL, WS, EDA, AB
Wu et al. 2016 [151]	VGGNet	LSTM	VS, SL, WS, EDA, AB
Sugano et at. 2016 [129]	VGGNet	LSTM	VS, SL, WS, EDA, AB
Mathews et al. 2016 [97]	GoogLeNet	LSTM	VS, SL, WS, EDA, SC
Wang et al. 2016 [144]	AlexNet, VGGNet	LSTM	VS, SL, WS, EDA
Johnson et al. 2016 [62]	VGGNet	LSTM	VS, SL, DC, EDA
Mao et al. 2016 [92]	VGGNet	LSTM	VS, SL, WS, EDA
Wang et al. 2016 [146]	VGGNet	LSTM	VS, SL, WS, CA
Tran et al. 2016 [135]	ResNet	MELM	VS, SL, WS, CA
Ma et al. 2016 [90]	AlexNet	LSTM	VS, SL, WS, CA
You et al. 2016 [156]	GoogLeNet	RNN	VS, SL, WS, EDA, SCB
Yang et al. 2016 [153]	VGGNet	LSTM	VS, SL, DC, EDA
Anne et al. 2016 [6]	VGGNet	LSTM	VS, SL, WS, CA, NOB
Yao et al. 2017 [155]	GoogLeNet	LSTM	VS, SL, WS, EDA, SCB
Lu et al. 2017 [88]	ResNet	LSTM	VS, SL, WS, EDA, AB
Chen et al. 2017 [21]	VGGNet, ResNet	LSTM	VS, SL, WS, EDA, AB
Gan et al. 2017 [41]	ResNet	LSTM	VS, SL, WS, CA, SCB
Pedersoli et al. 2017 [112]	VGGNet	RNN	VS, SL, WS, EDA, AB
Ren et al. 2017 [119]	VGGNet	LSTM	VS, ODL, WS, EDA
Park et al. 2017 [111]	ResNet	LSTM	VS, SL, WS, EDA, AB
Wang et al. 2017 [148]	ResNet	LSTM	VS, SL, WS, EDA
Tavakoli et al. 2017 [134]	VGGNet	LSTM	VS, SL, WS, EDA, AB
Liu et al. 2017 [84]	VGGNet	LSTM	VS, SL, WS, EDA, AB
Gan et al. 2017 [39]	ResNet	LSTM	VS, SL, WS, EDA, SC
Dai et al. 2017 [26]	VGGNet	LSTM	VS, ODL, WS, EDA
Shetty et al. 2017 [126]	GoogLeNet	LSTM	VS, ODL, WS, EDA
Liu et al. 2017 [85]	Inception-V3	LSTM	VS, ODL, WS, EDA
Gu et al. 2017 [51]	VGGNet	1. Language CNN 2. LSTM	VS, SL, WS, EDA
Yao et al. 2017 [154]	VGGNet	LSTM	VS, SL, WS, CA, NOB

(Continued)

- Take an image as input, and generate a sentence describing it as output (i.e. the caption)
- Typical methods include a deep CNN/transformer and a RNN-like language model
- (The task of *Dense Captioning* is to generate one caption per bounding box)

Table from https://dl.acm.org/doi/pdf/10.1145/3295748

Medical Image Analysis

Notice that **most** of these tasks are structured prediction problems, not merely classification

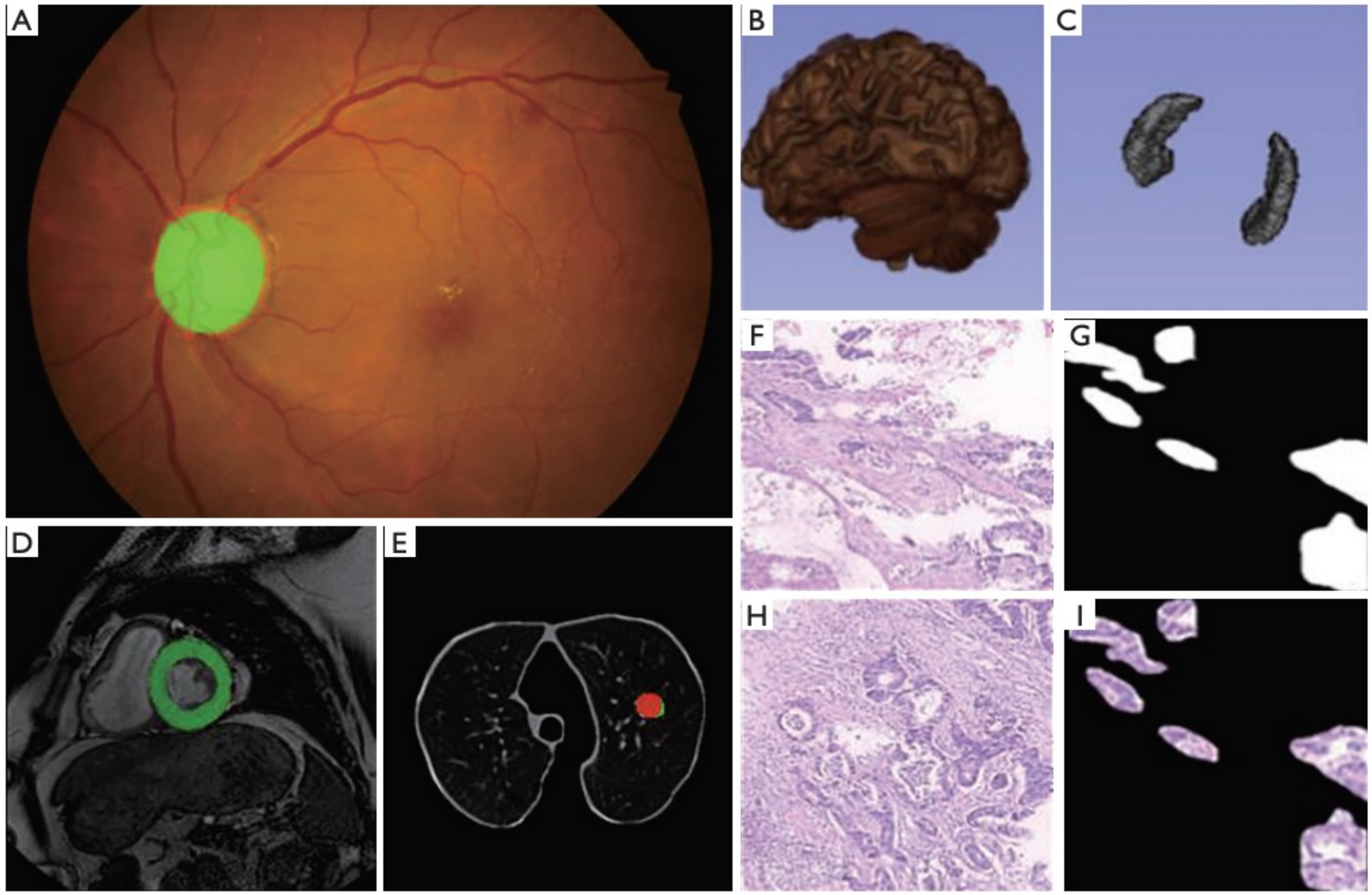


Figure 2 Deep learning application in medical image analysis. (A) Fundus detection; (B,C) hippocampus segmentation; (D) left ventricular segmentation; (E) pulmonary nodule classification; (F,G,H,I) gastric cancer pathology segmentation. The staining method is H&E, and the magnification is ×40.

Figure from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7327346/

SEMANTIC SEGMENTATION

Case Study: Image Segmentation

- Image segmentation (FG/BG) by modeling of interactions btw RVs
  – Images are noisy.
  – Objects occupy continuous regions in an image.

[Nowozin,Lampert 2012]

Input image

Pixel-wise separate optimal labeling

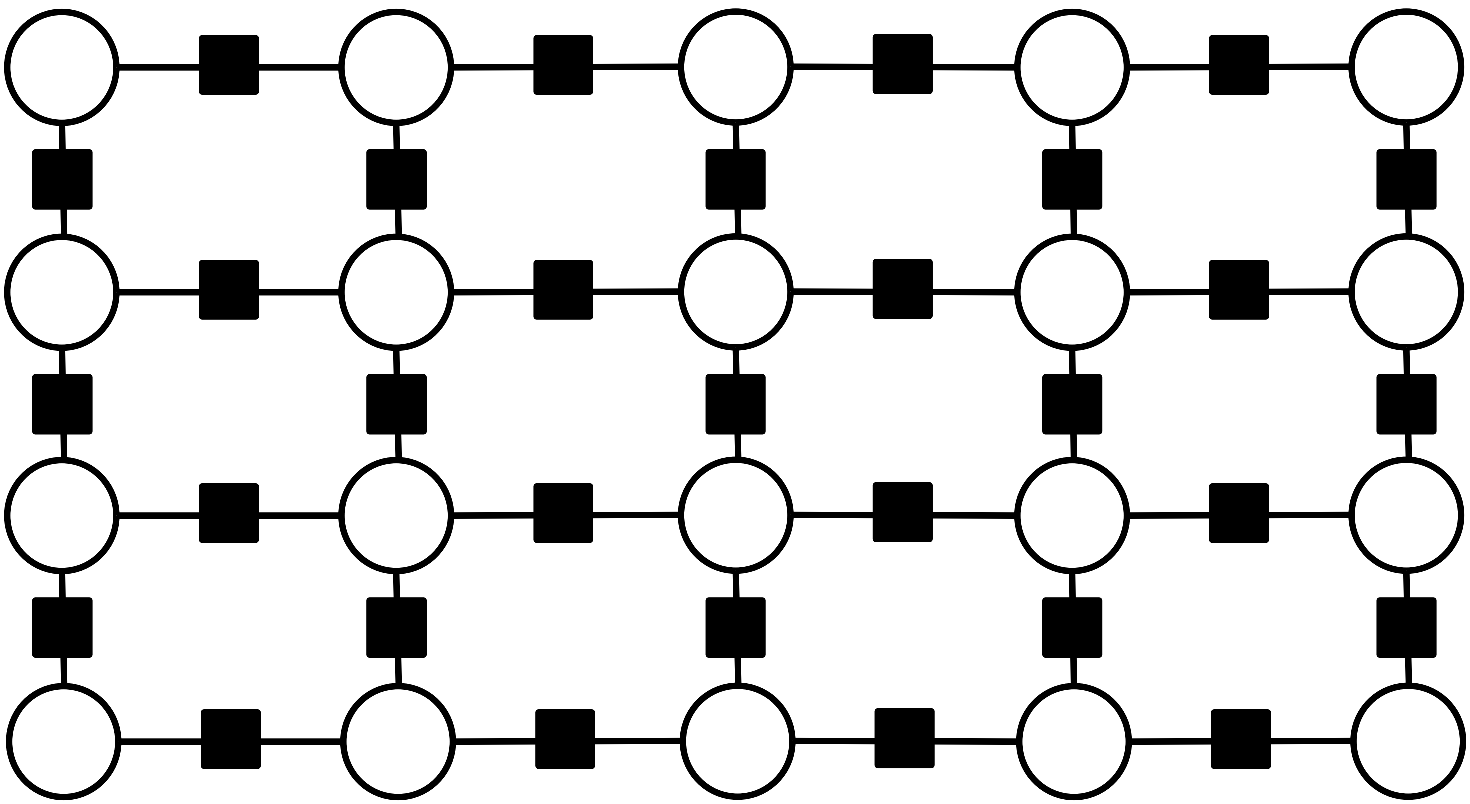
Locally-consistent joint optimal labeling

$$Y^* = \arg\max_{y\in\{0,1\}^n} \left[\overbrace{\sum_{i\in S} V_i(y_i, X)}^{\text{Unary Term}} + \overbrace{\sum_{i\in S}\sum_{j\in N_i} V_{i,j}(y_i, y_j)}^{\text{Pairwise Term}} \right].$$

Y: labels
X: data (features)
S: pixels
N_i: neighbors of pixel i

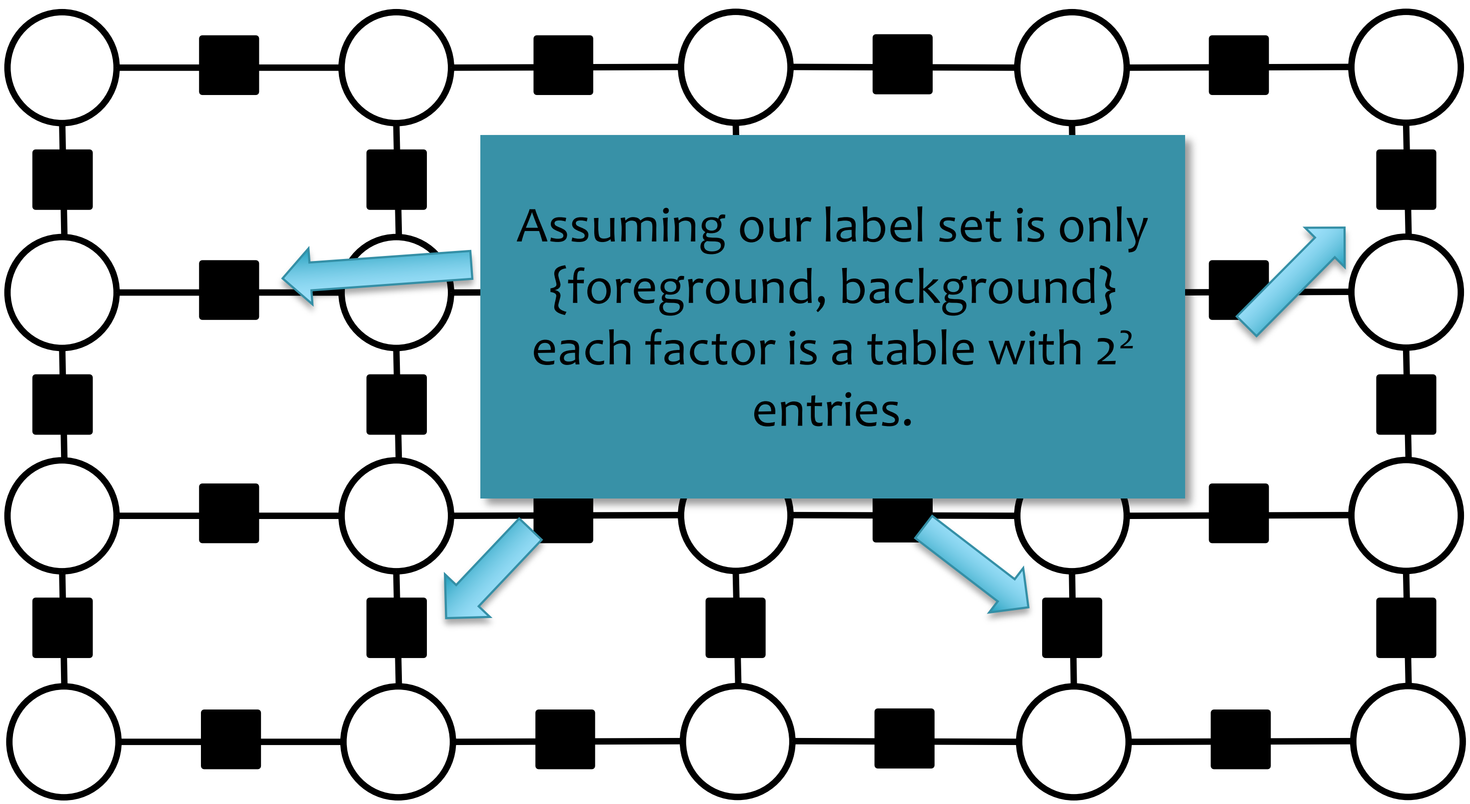
Grid CRF

- Suppose we want to image segmentation using a grid model



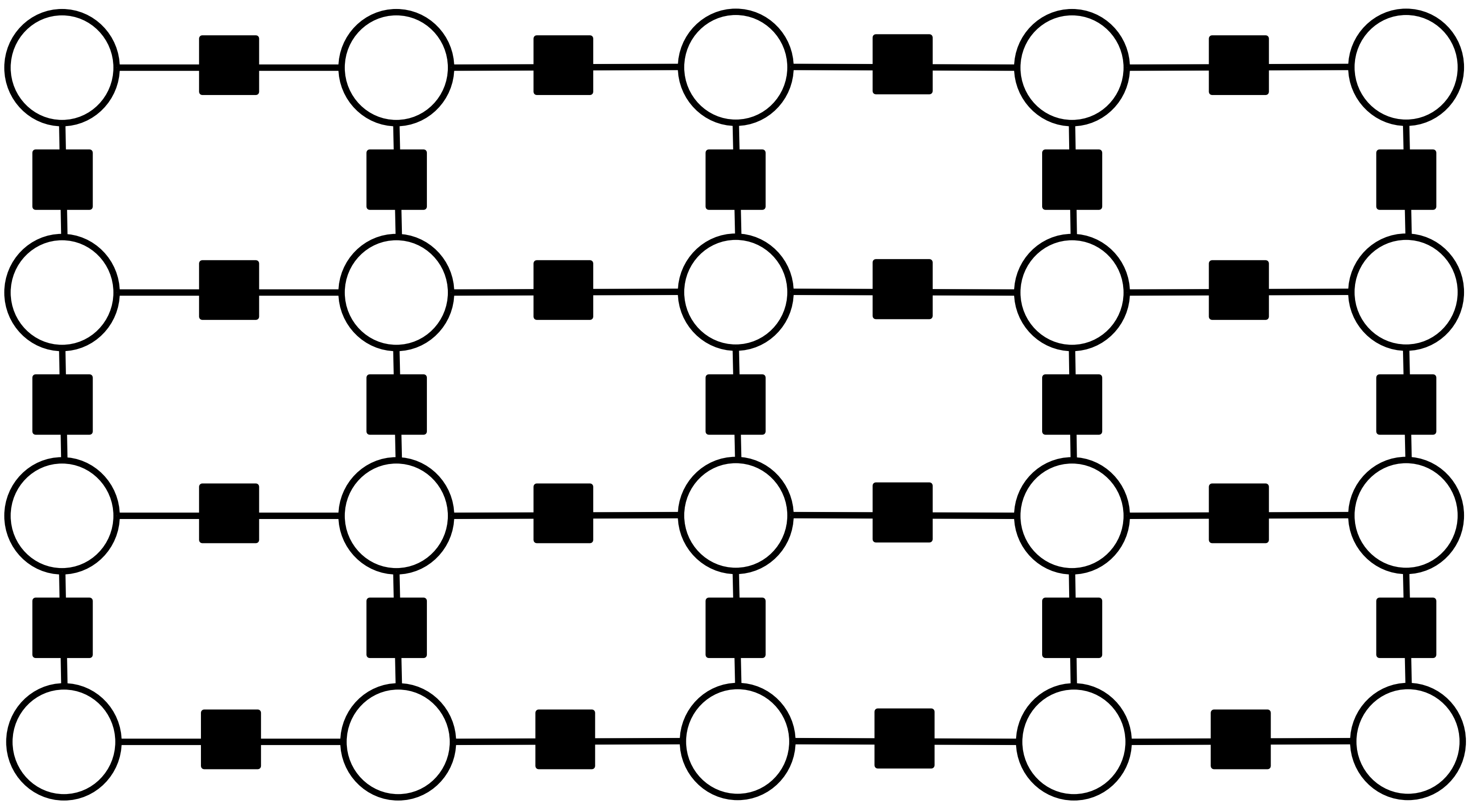
Grid CRF

- Suppose we want to image segmentation using a grid model



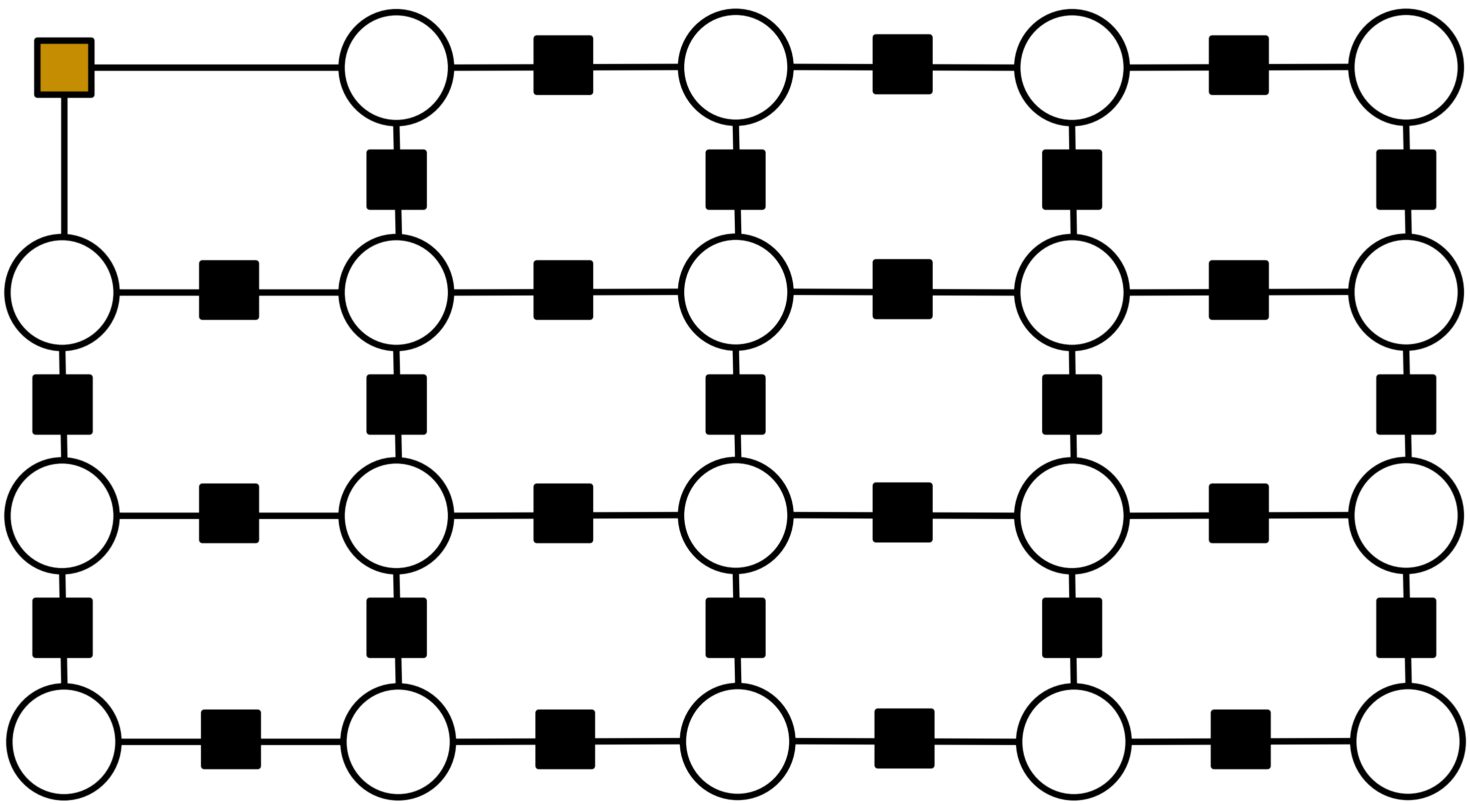
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



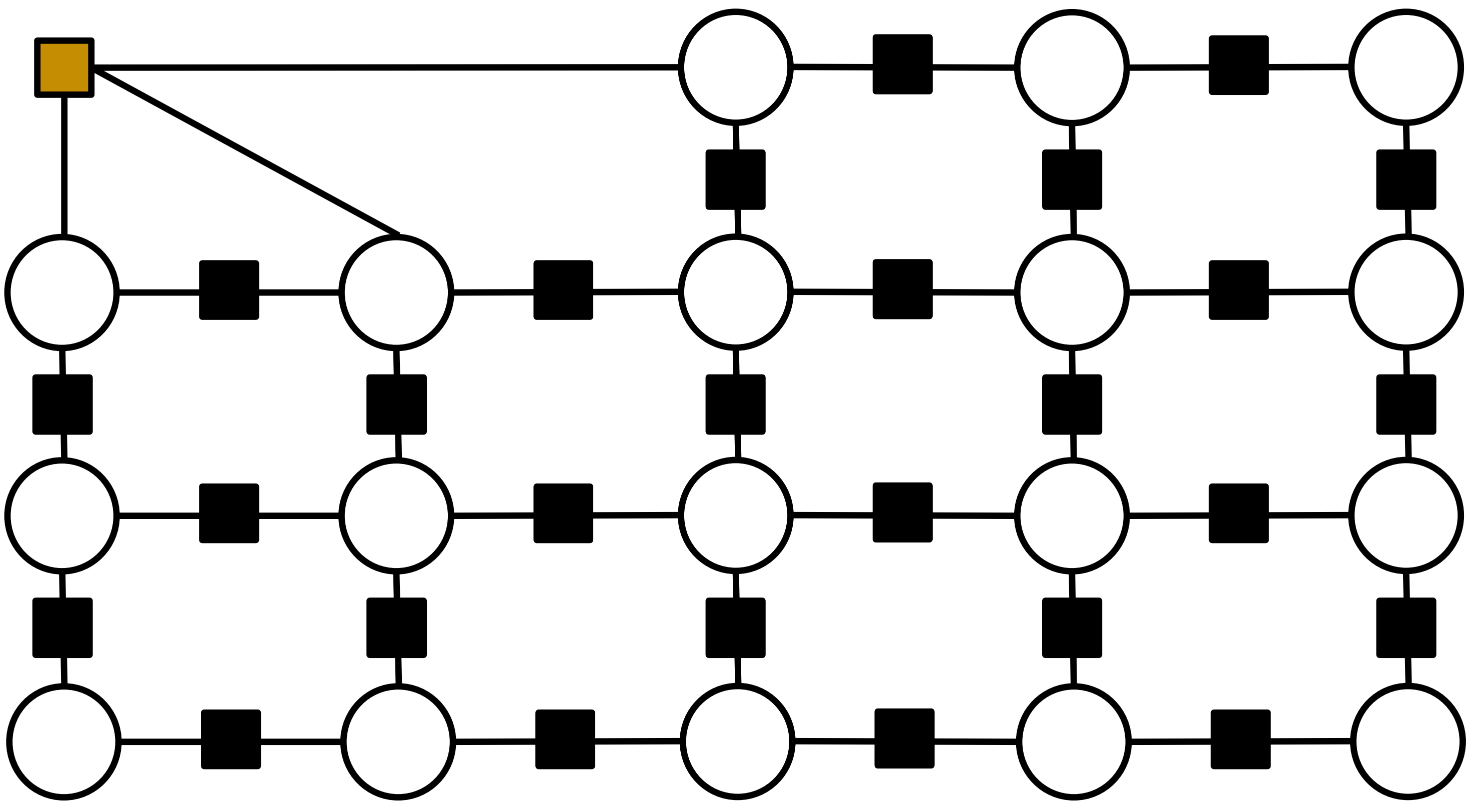
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



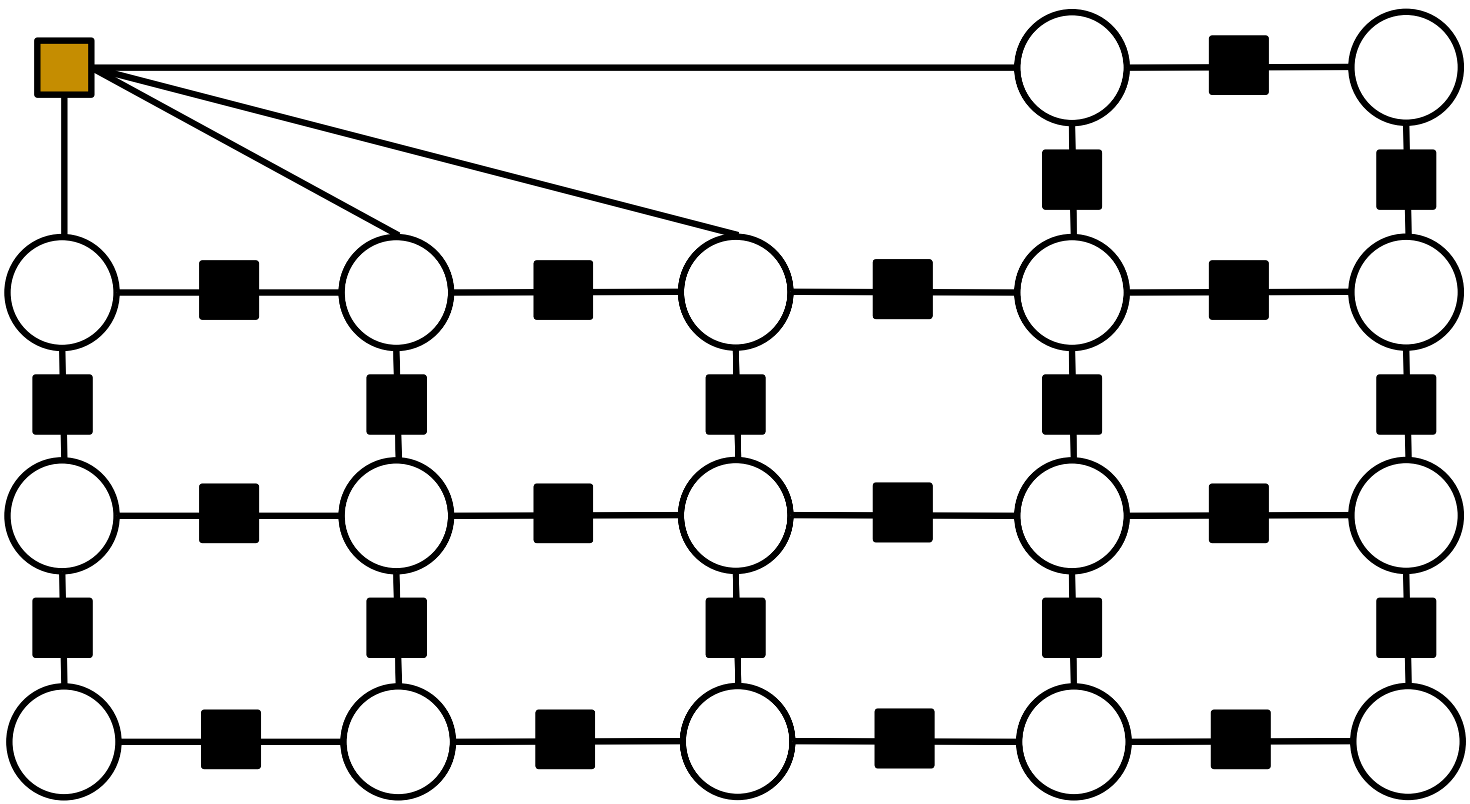
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



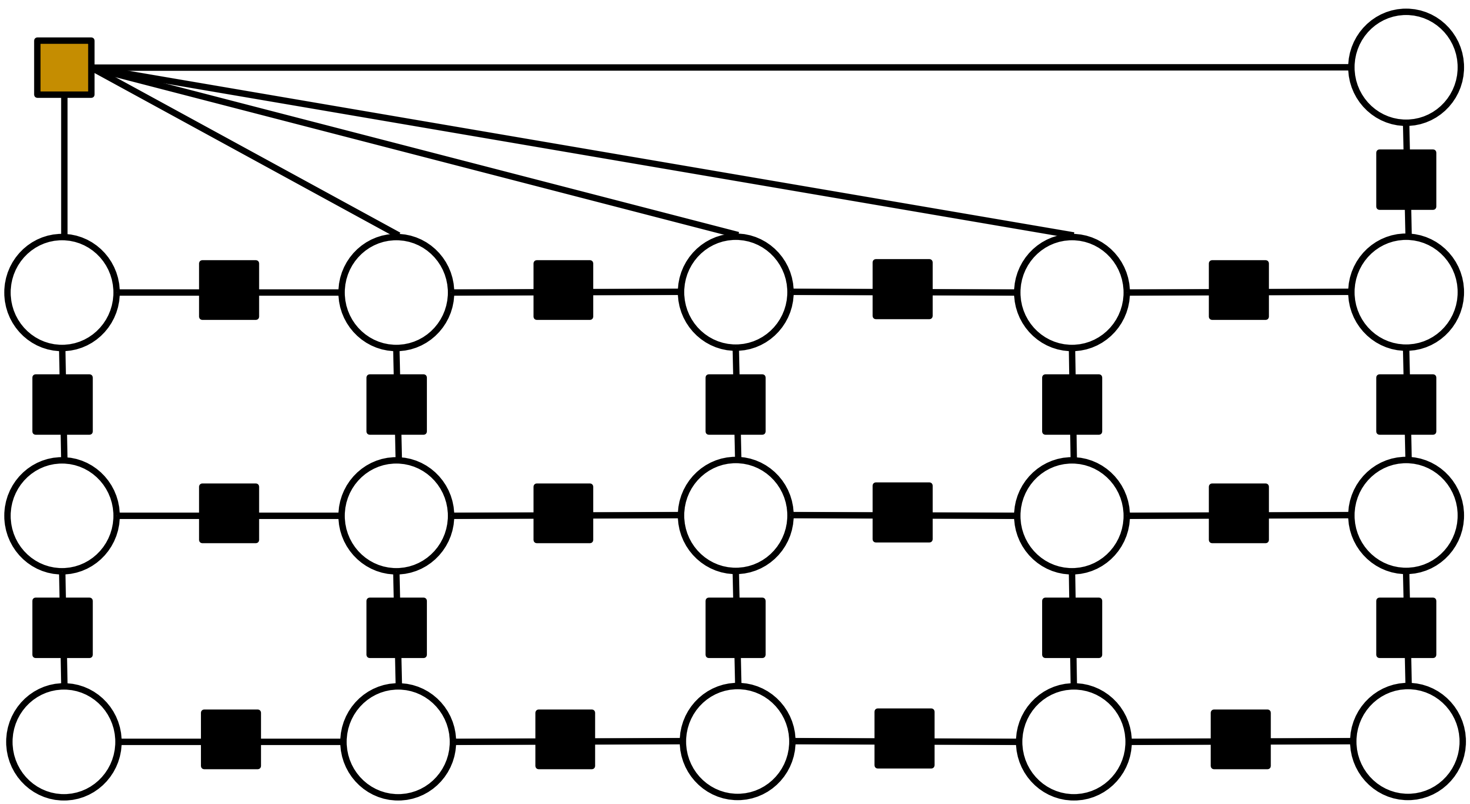
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



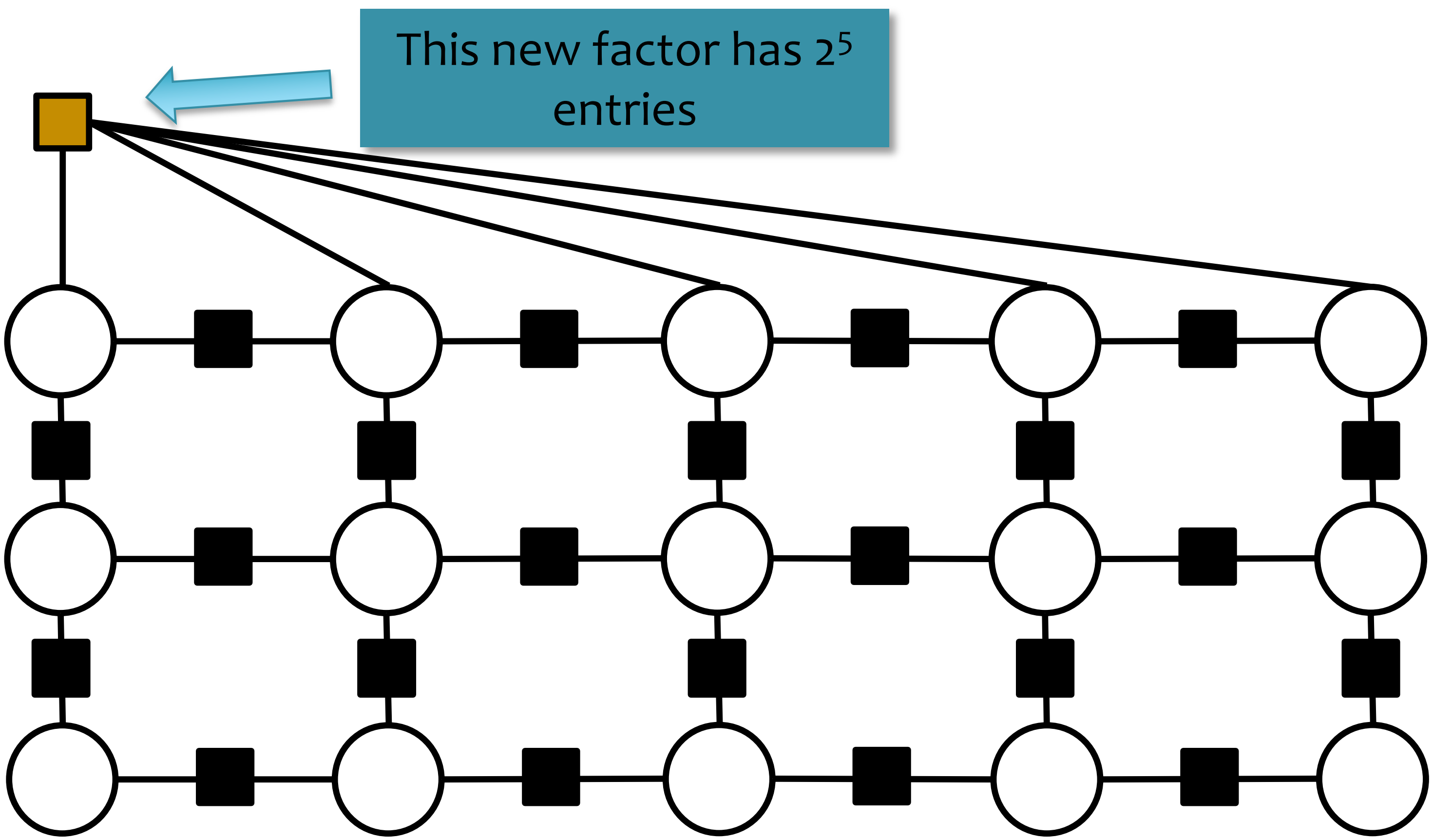
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



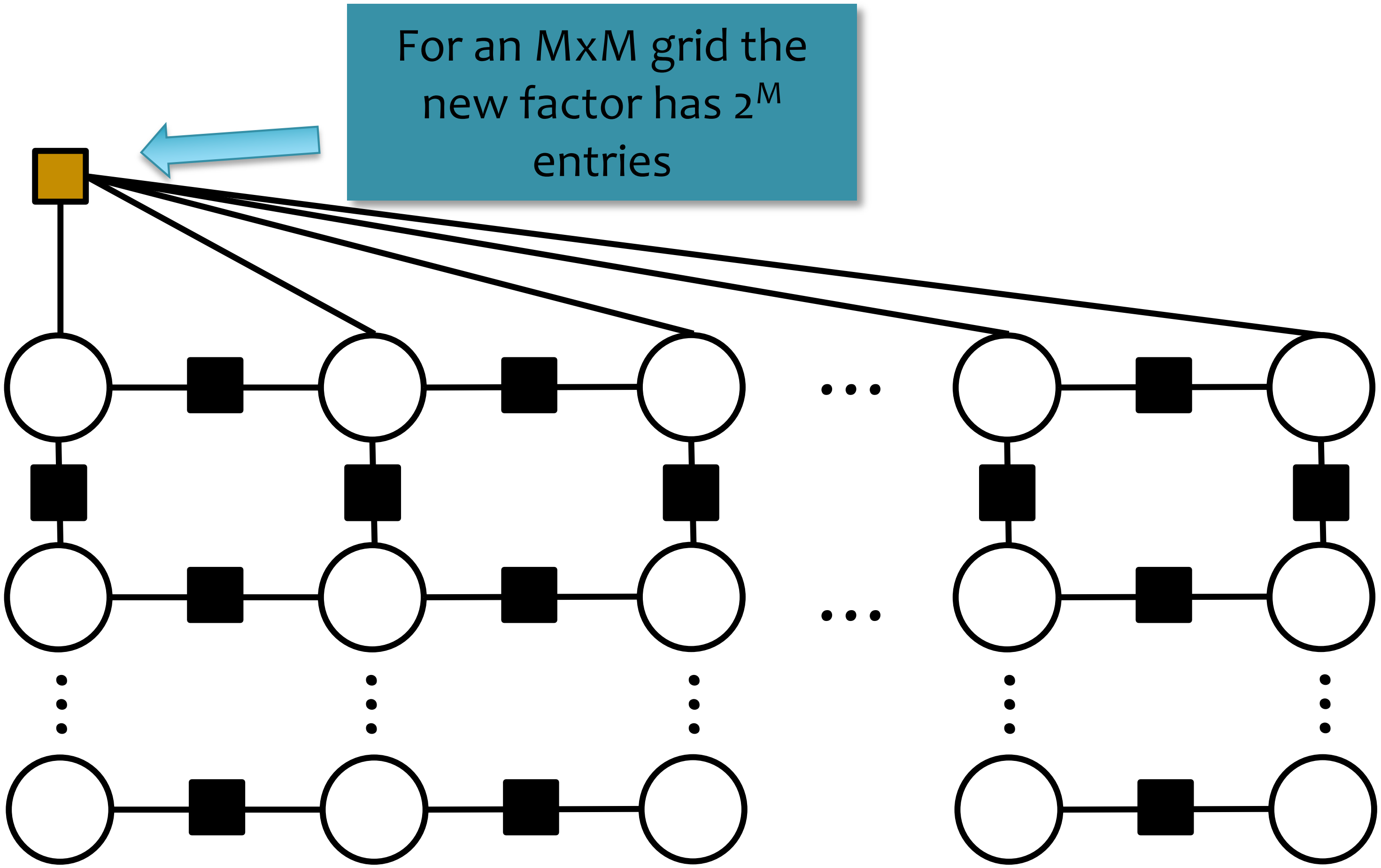
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



Grid CRF

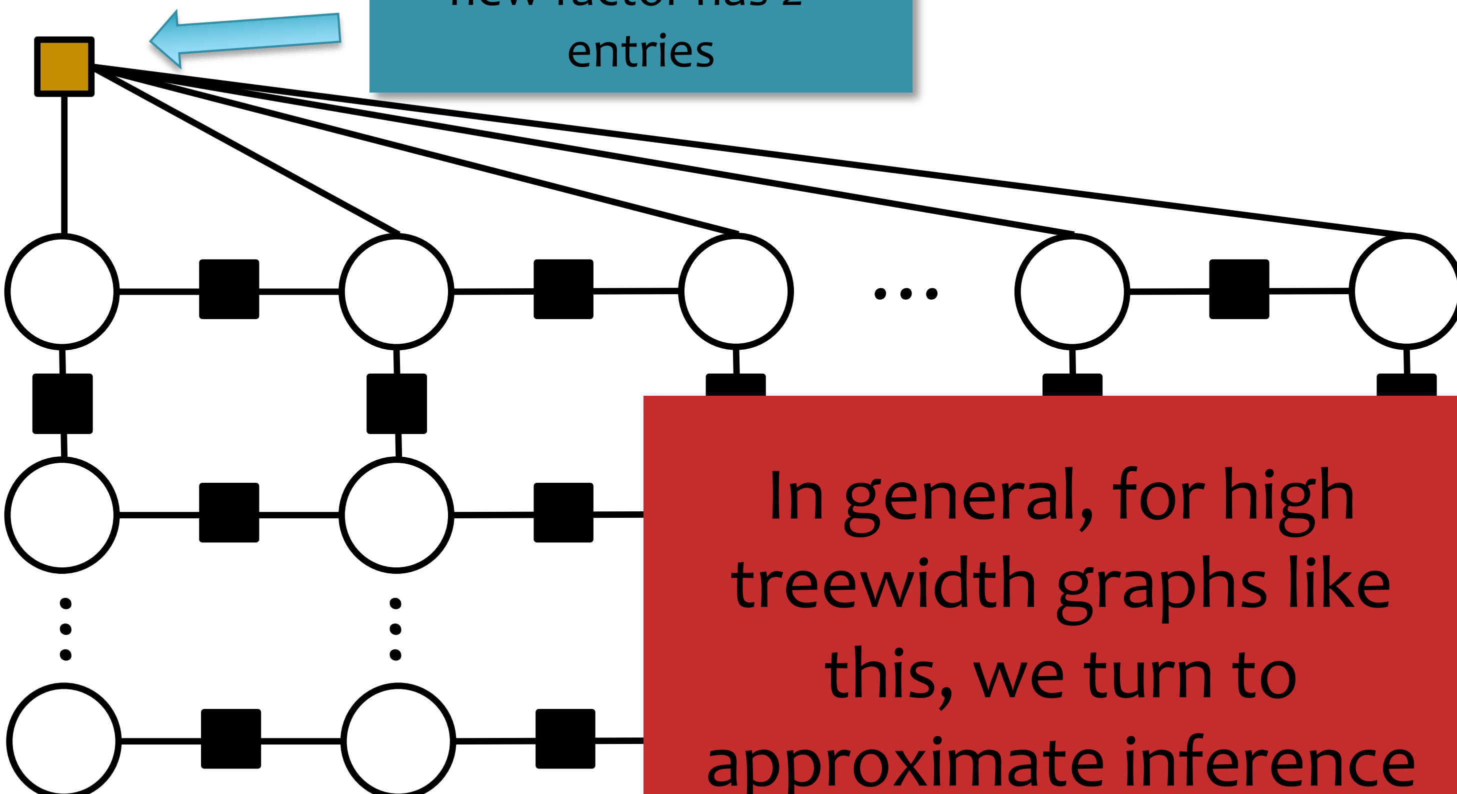
- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?

For an MxM grid the new factor has 2^M entries



In general, for high treewidth graphs like this, we turn to approximate inference (which we'll cover soon!)

Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?
- Can we instead run belief propagation to do exact inference?