University of BRISTOL

# Machine Learning

## Topic Models

Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 24, 2017

http://www.carlhenrik.com

# Introduction

# Process $\rightarrow$ Distribution $\rightarrow$ value

- Each evaluation of a process is a distribution

$$\mathcal{N}(\mu, \Sigma) \sim \mathcal{GP}(m(x), k(\mathbf{x}, \mathbf{x}))$$

- Each evaluation of a distribution is a value

$$y \sim \mathcal{N}(y|0, \Sigma)$$

- Parametric models are definedy by distributions and non-parametric by processes

# Bernoulli Distribution

- Distribution over binary random variable $x \in \{0, 1\}$

$$p(x = 1|\mu) = \mu$$

# Bernoulli Distribution

- Distribution over binary random variable $x \in \{0, 1\}$

$$p(x = 1|\mu) = \mu$$

- Due to binary outcome

$$p(x = 0|\mu) = 1 - \mu$$

# Bernoulli Distribution

- Distribution over binary random variable $x \in \{0, 1\}$

$$p(x = 1|\mu) = \mu$$

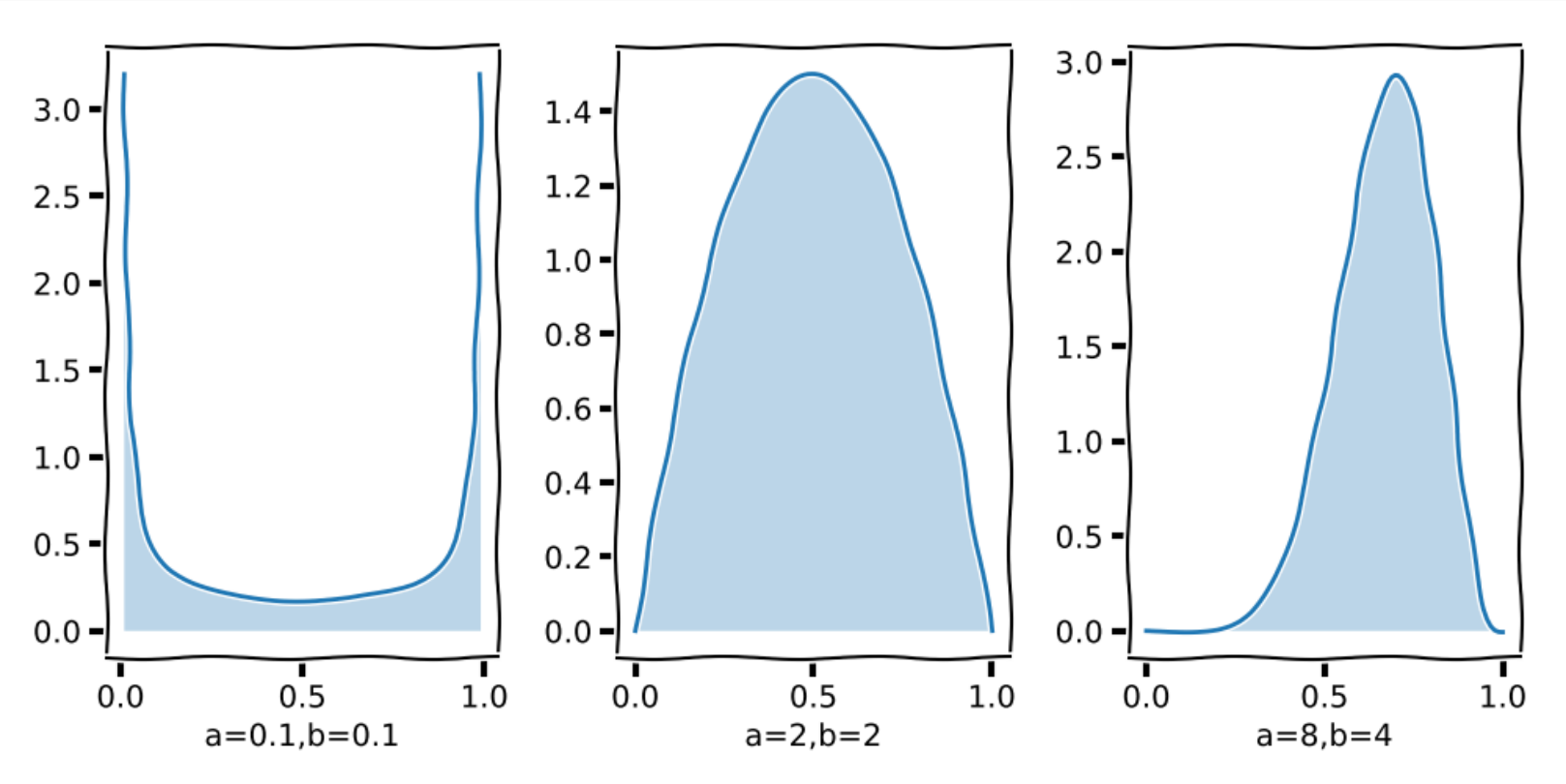
- Due to binary outcome

$$p(x = 0|\mu) = 1 - \mu$$

- Distribution

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}$$

# Beta Distribution




$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

# Multinomial

- If we have a variable that can take $K$ different states

$$\mathbf{x} = [0, 0, 1, 0, 0, 0]^{\mathrm{T}}$$

# Multinomial

- If we have a variable that can take $K$ different states

$$\mathbf{x} = [0, 0, 1, 0, 0, 0]^{\mathrm{T}}$$

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]^{\mathrm{T}}, \sum_k \mu_k = 1$$

# Multinomial

- If we have a variable that can take $K$ different states

$$\mathbf{x} = [0, 0, 1, 0, 0, 0]^{\mathrm{T}}$$

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$\boldsymbol{\mu} = [\mu_1, \ldots, \mu_k]^{\mathrm{T}}, \sum_k \mu_k = 1$$

- Likelihood

$$p(\mathbf{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}}$$

# Dirichlet

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

# Dirichlet

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- Conjugate prior

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

# Dirichlet

- Multinomial

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

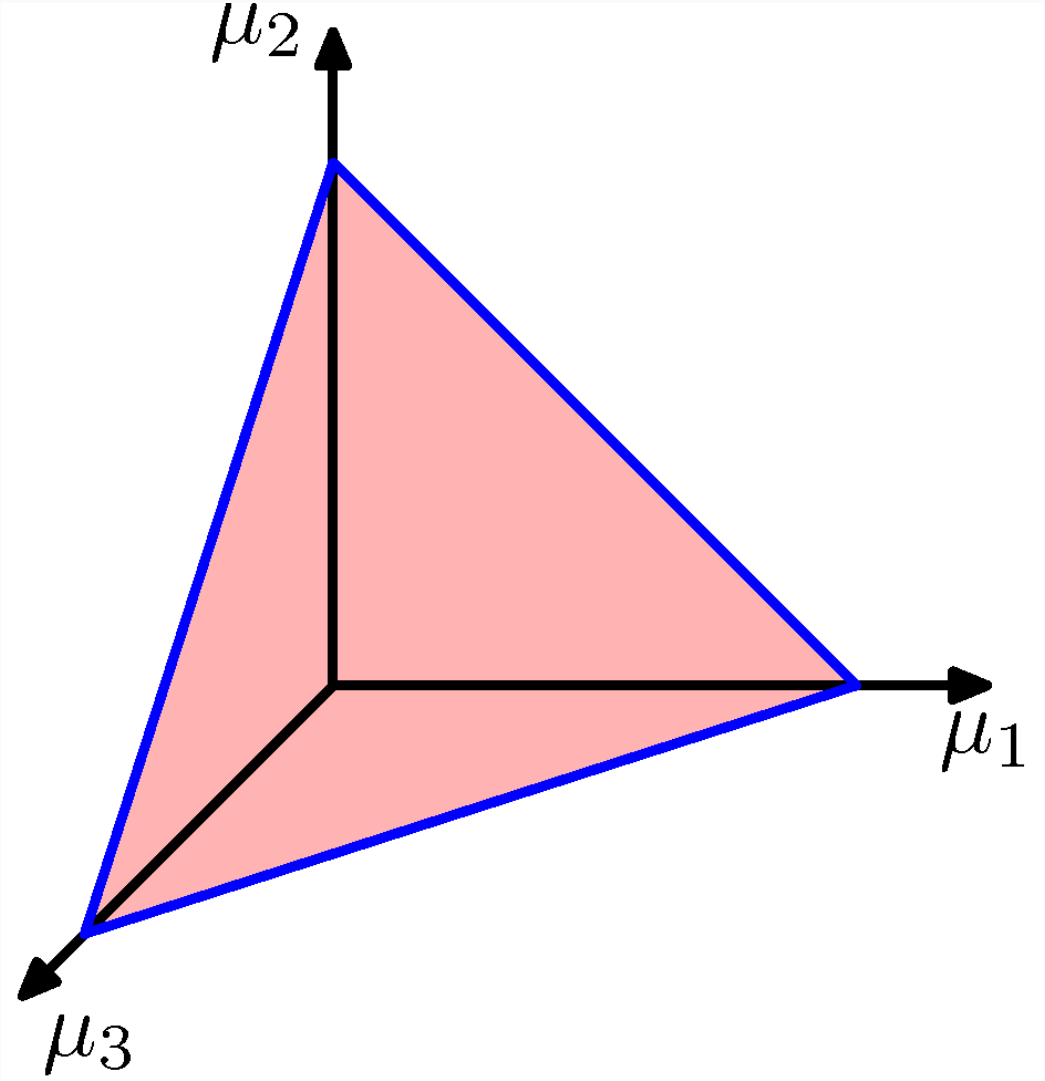
- Conjugate prior

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

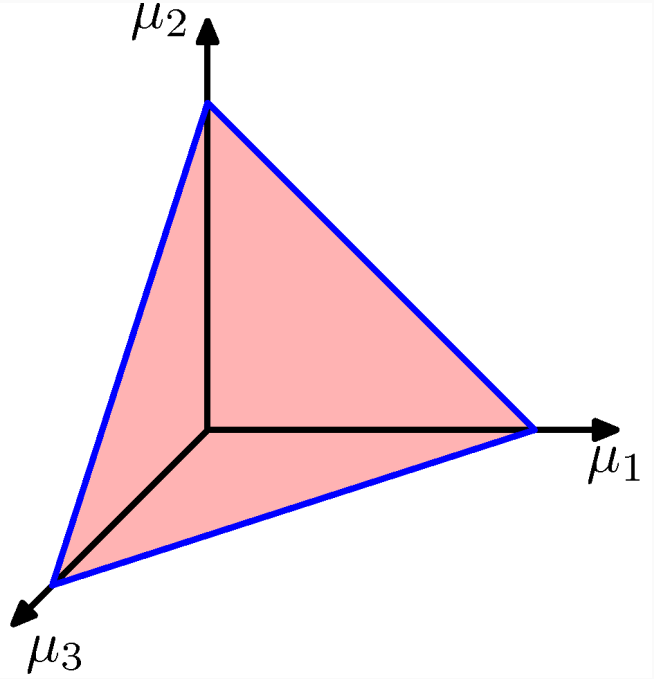
- Dirichlet Distribution

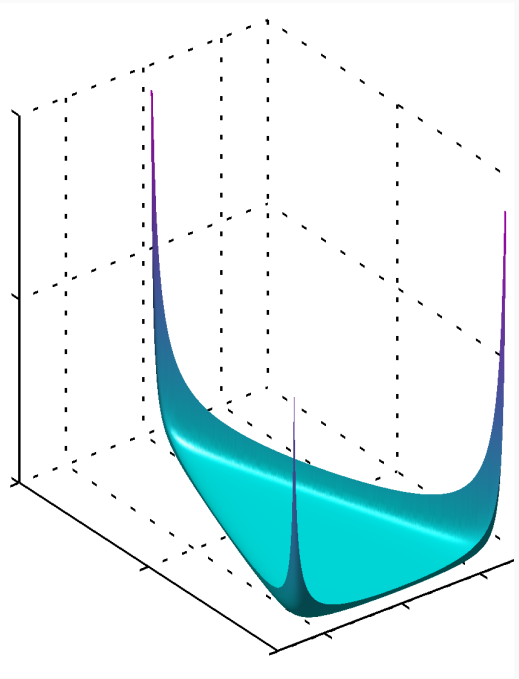
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdot \ldots \cdot \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

# Dirichlet Prior



# Dirichlet Distribution

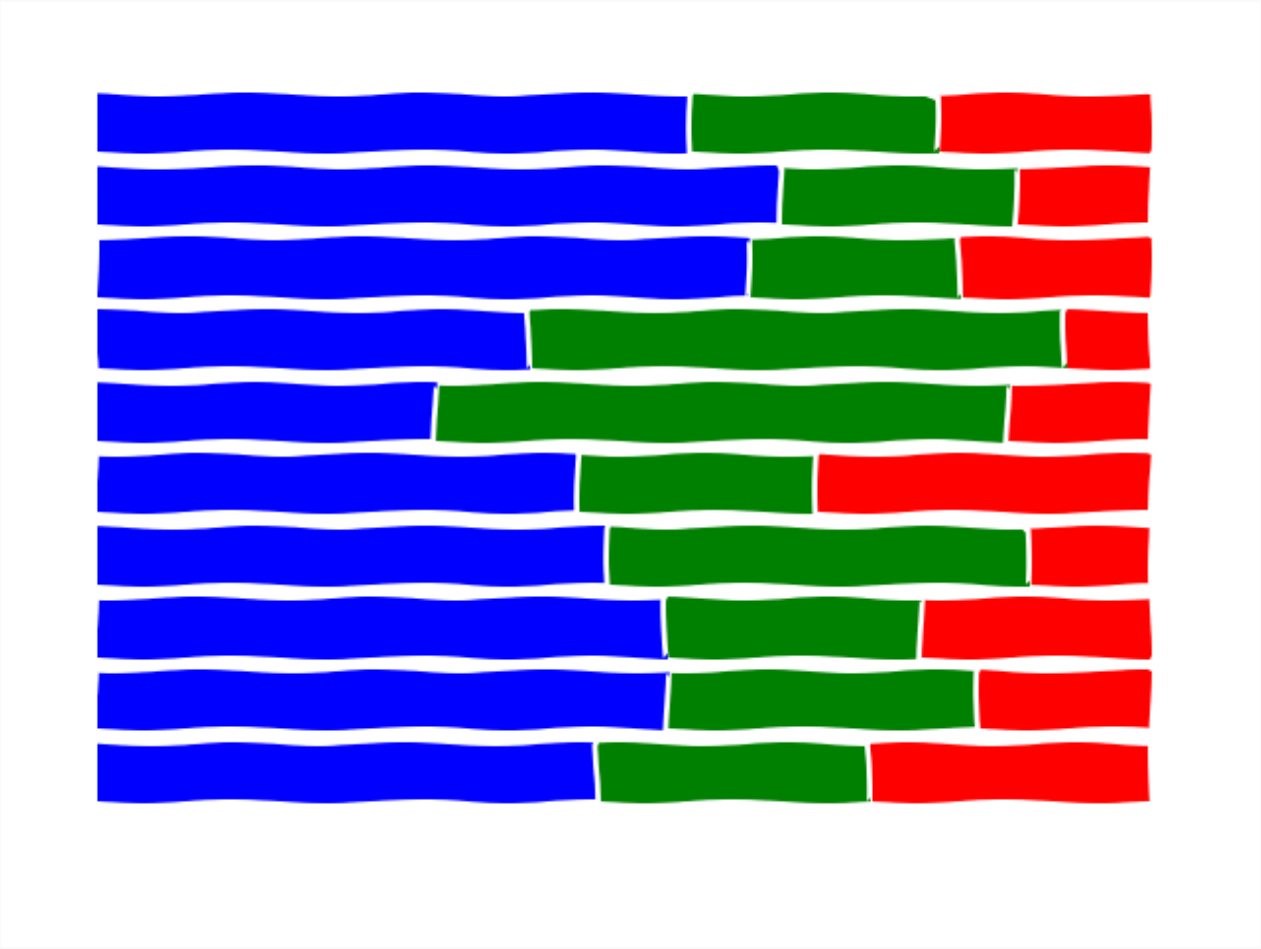


- Conjugate prior to multinomial

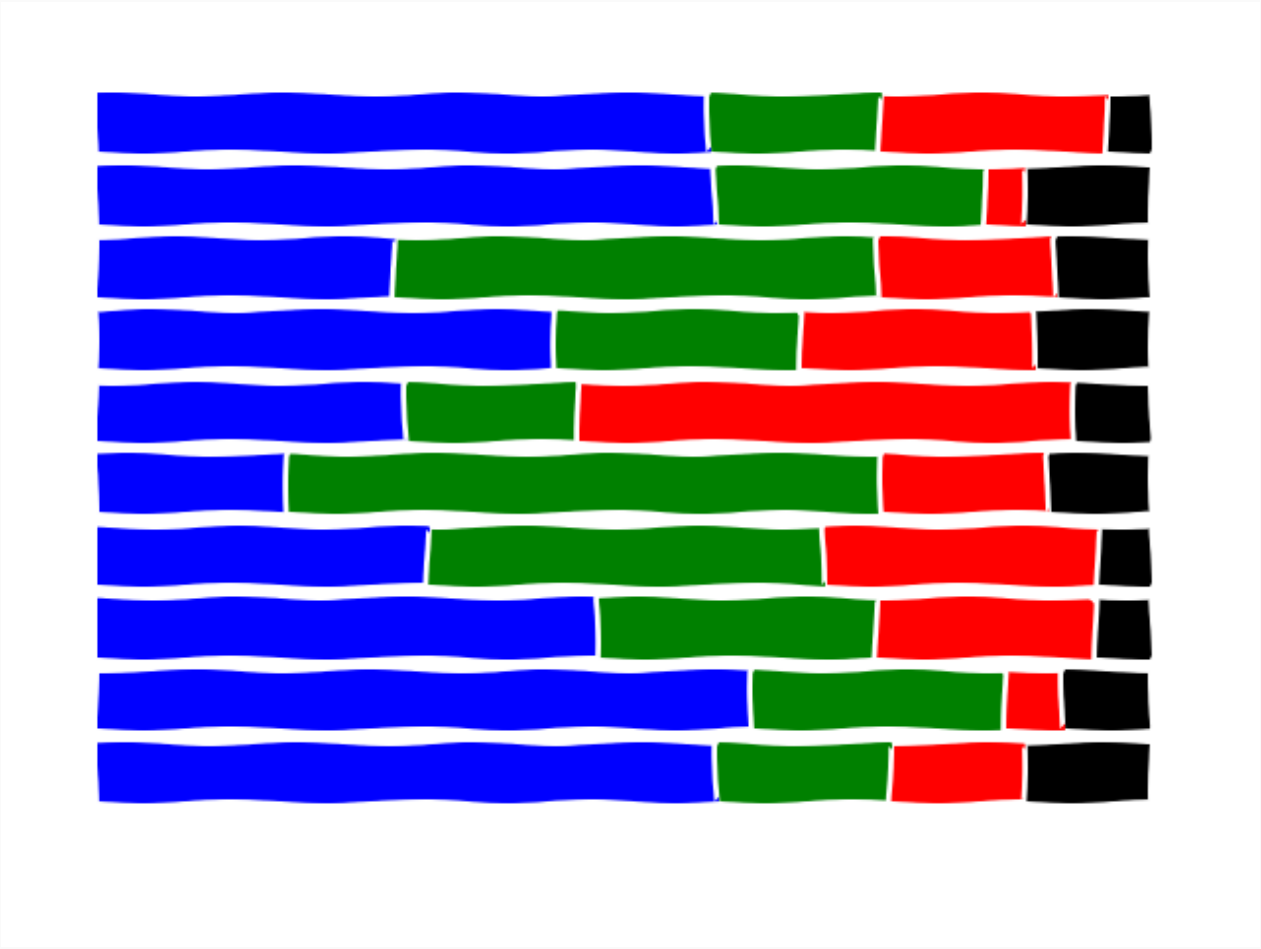
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdot\ldots\cdot\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$$

# Dirichlet Distribution



$\mathrm{Dir}(10, 5, 3)$

# Dirichlet Distribution

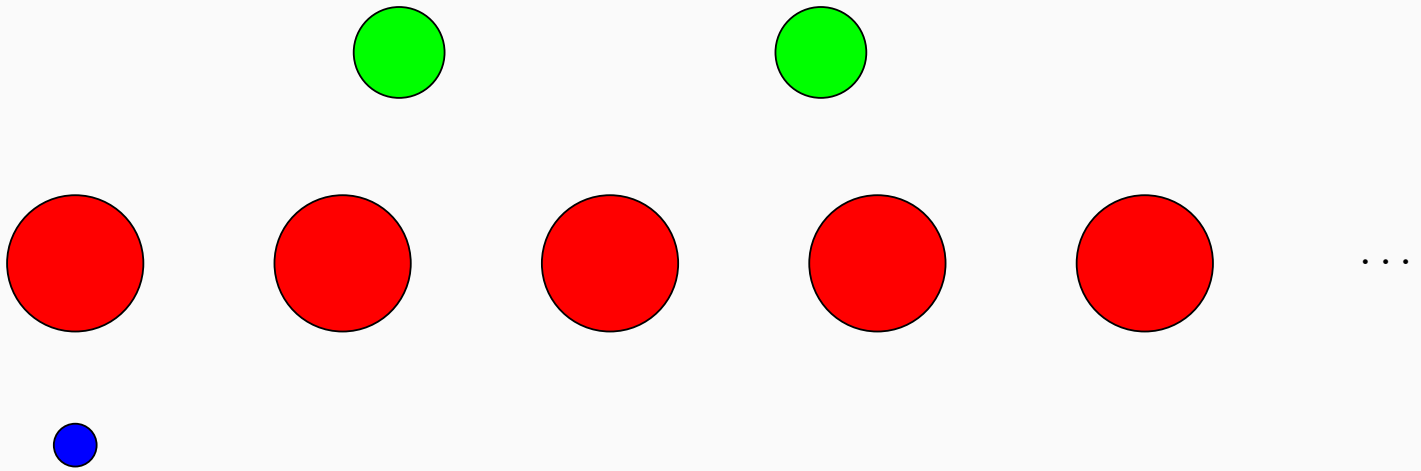


$\mathrm{Dir}(7, 5, 3, 2)$

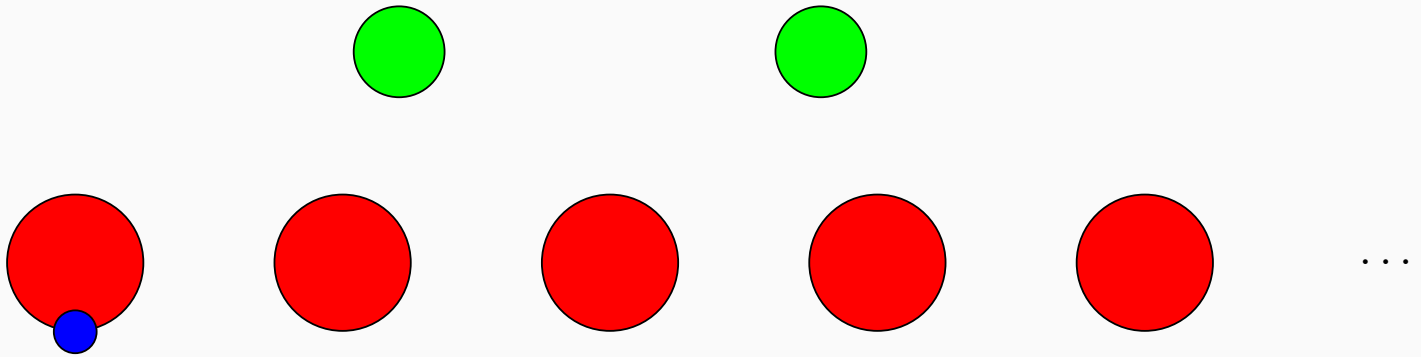
# Dirichlet Process

- Is the infinite dimensional generalisation of a Dirichlet distribution
  - just as Gaussian process is of Gaussian distribution
- Generates a partitioning of (possibly) infinite number of elements
- Not as intuitive to write down
- Best written constructively

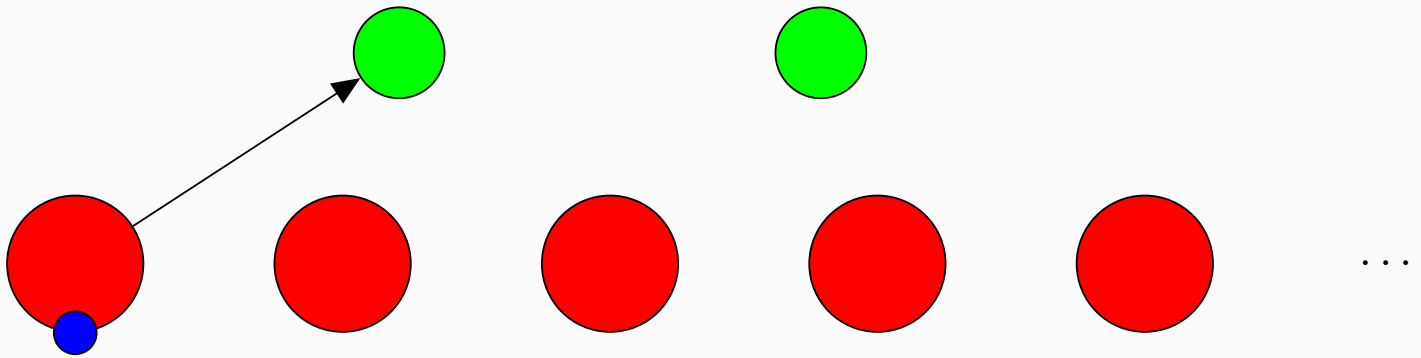
# Chinese Resturant Process



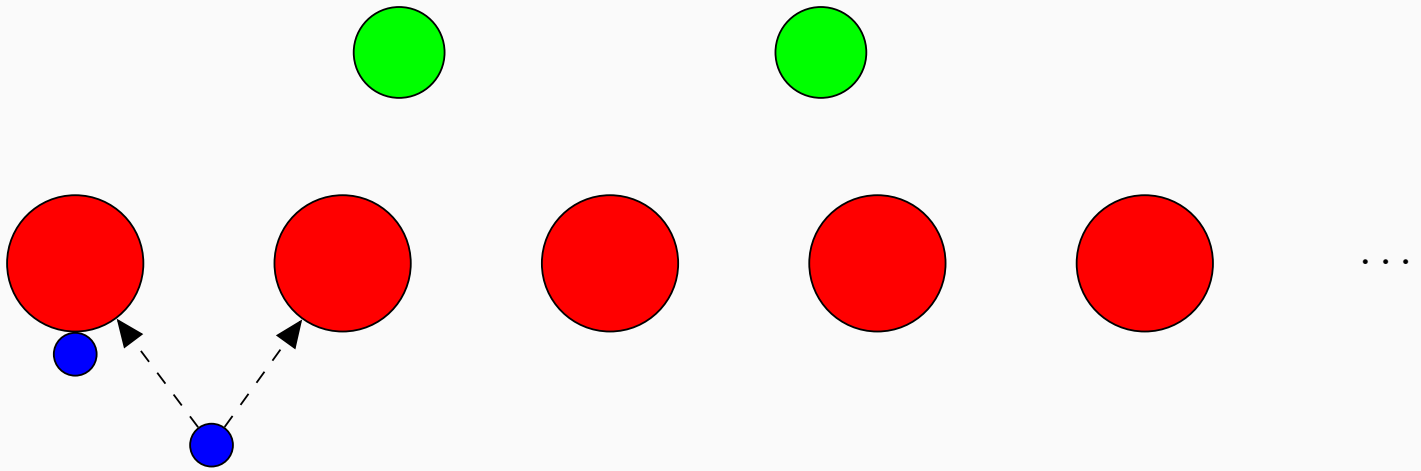
# Chinese Resturant Process



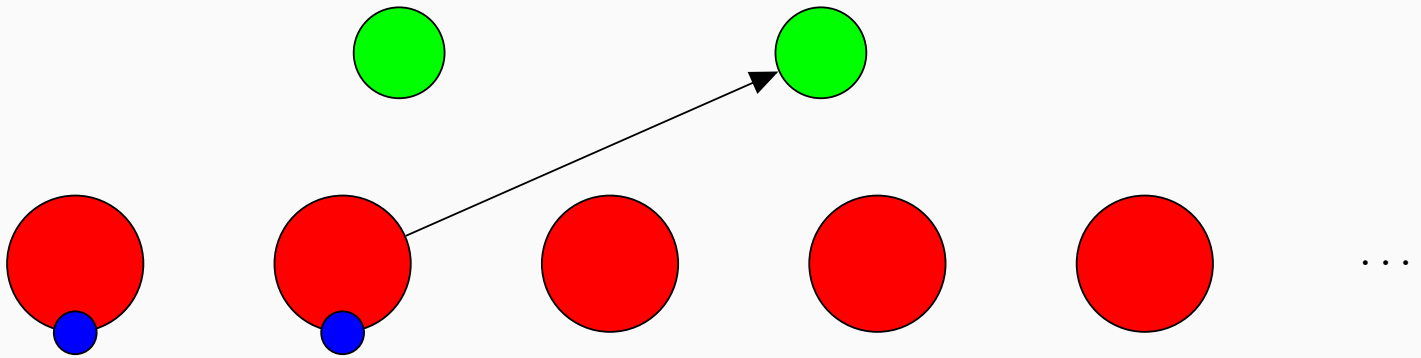
# Chinese Resturant Process



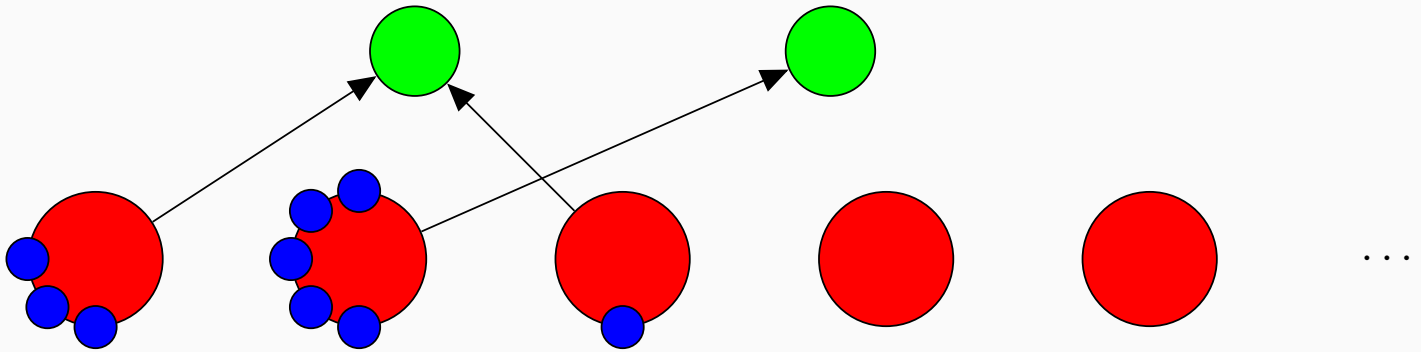
# Chinese Resturant Process



# Chinese Resturant Process

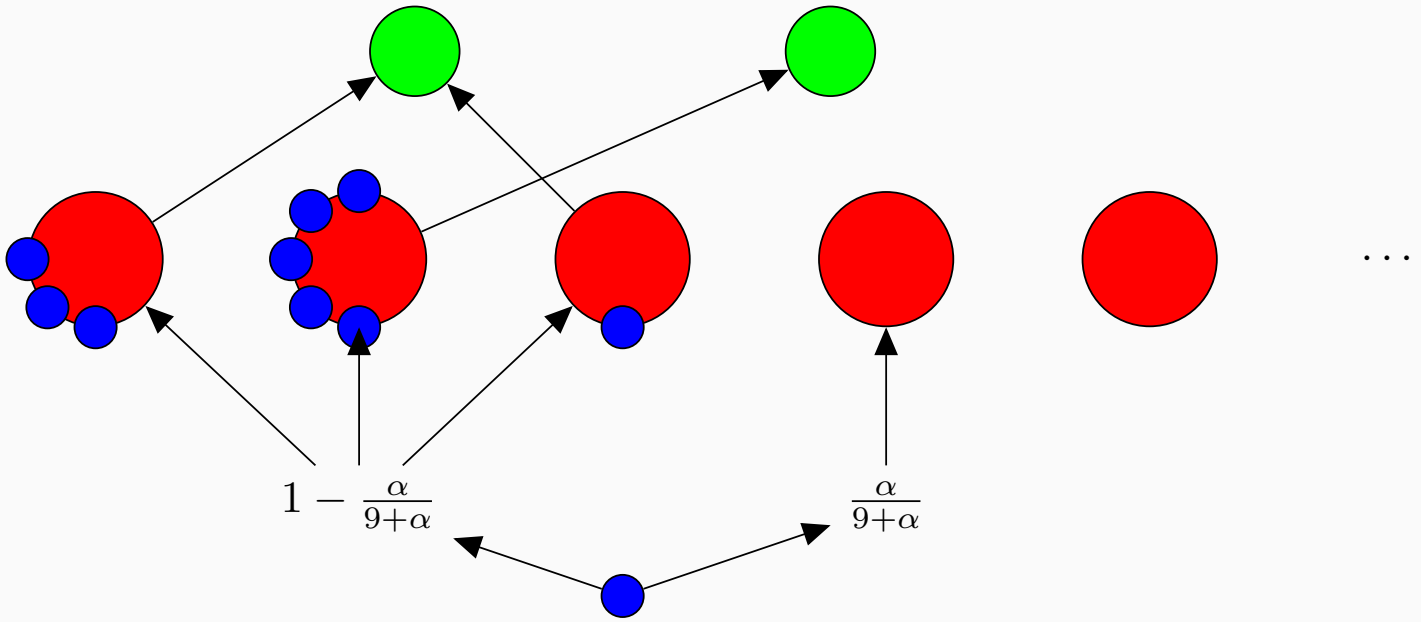


# Chinese Resturant Process

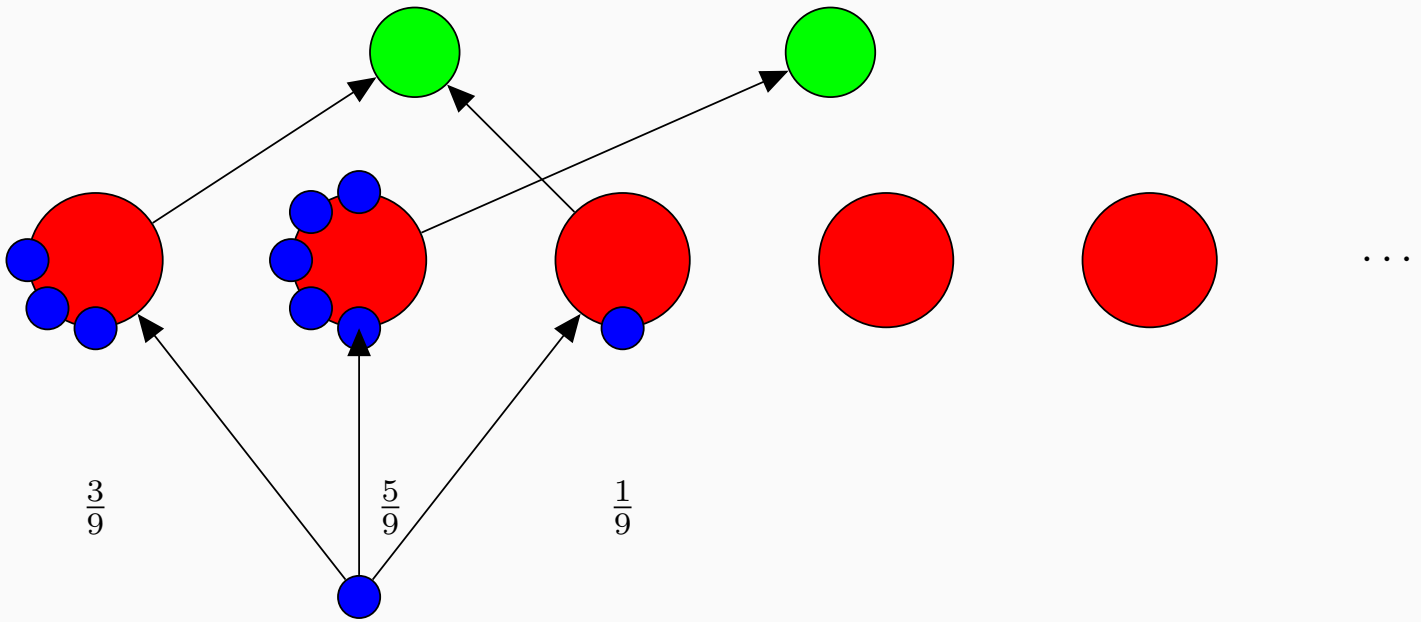


- Go to new table $\frac{\alpha}{N-1+\alpha}$
- If not choose table as $\frac{n_i}{N}$ where $n_i$ number of diners at table *in*

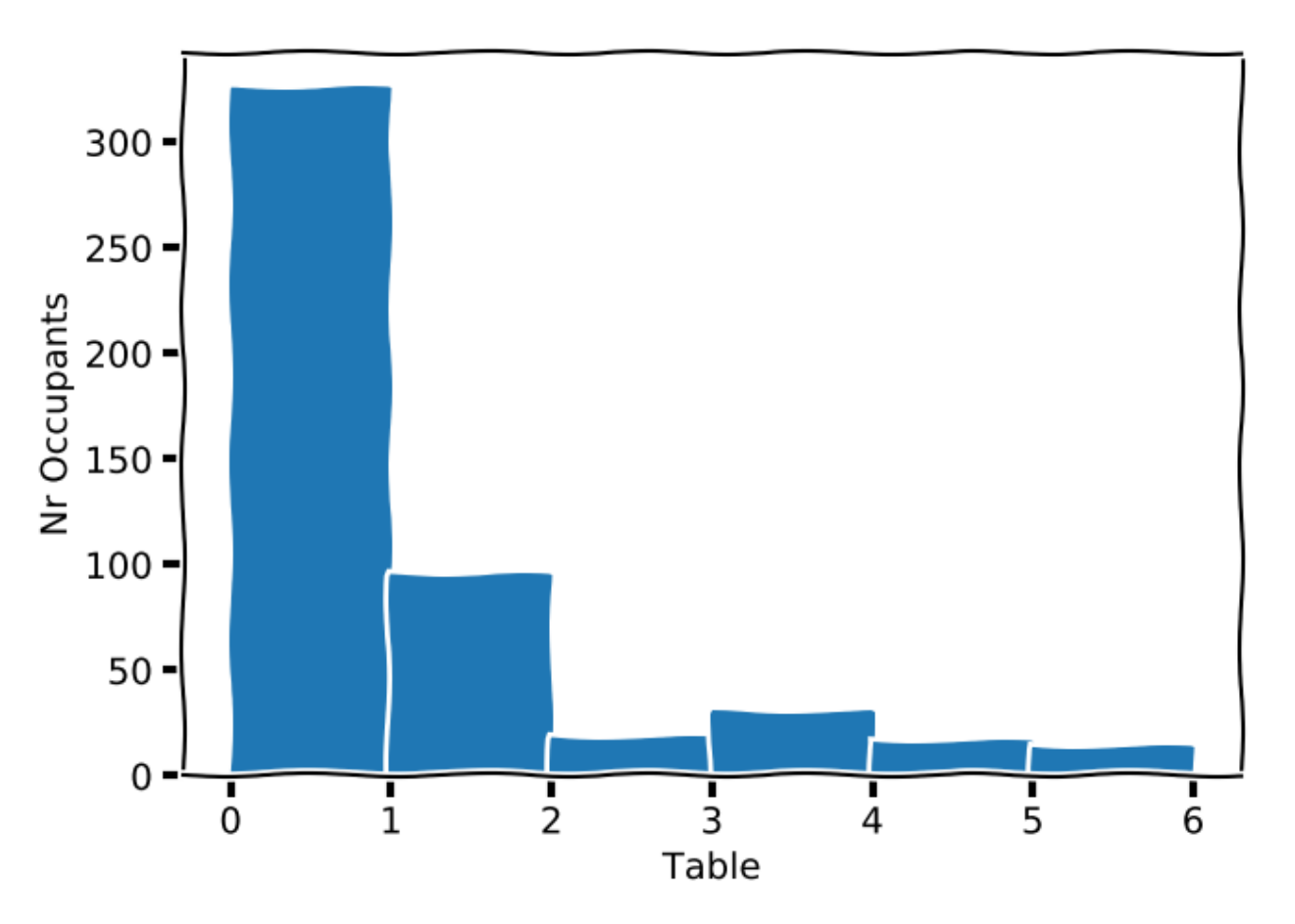
# Chinese Resturant Process



# Chinese Resturant Process

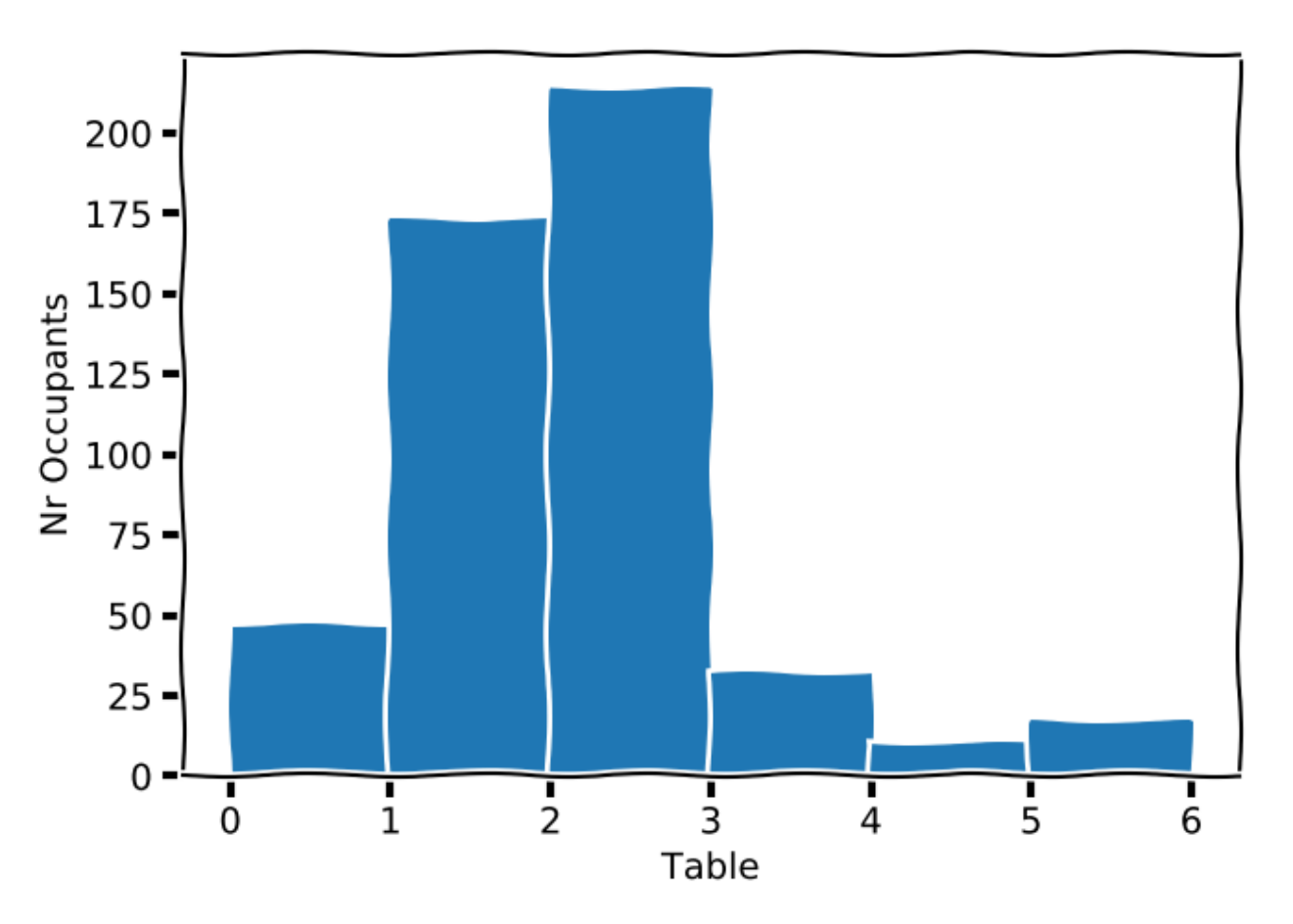


# Chinese Resturant Process



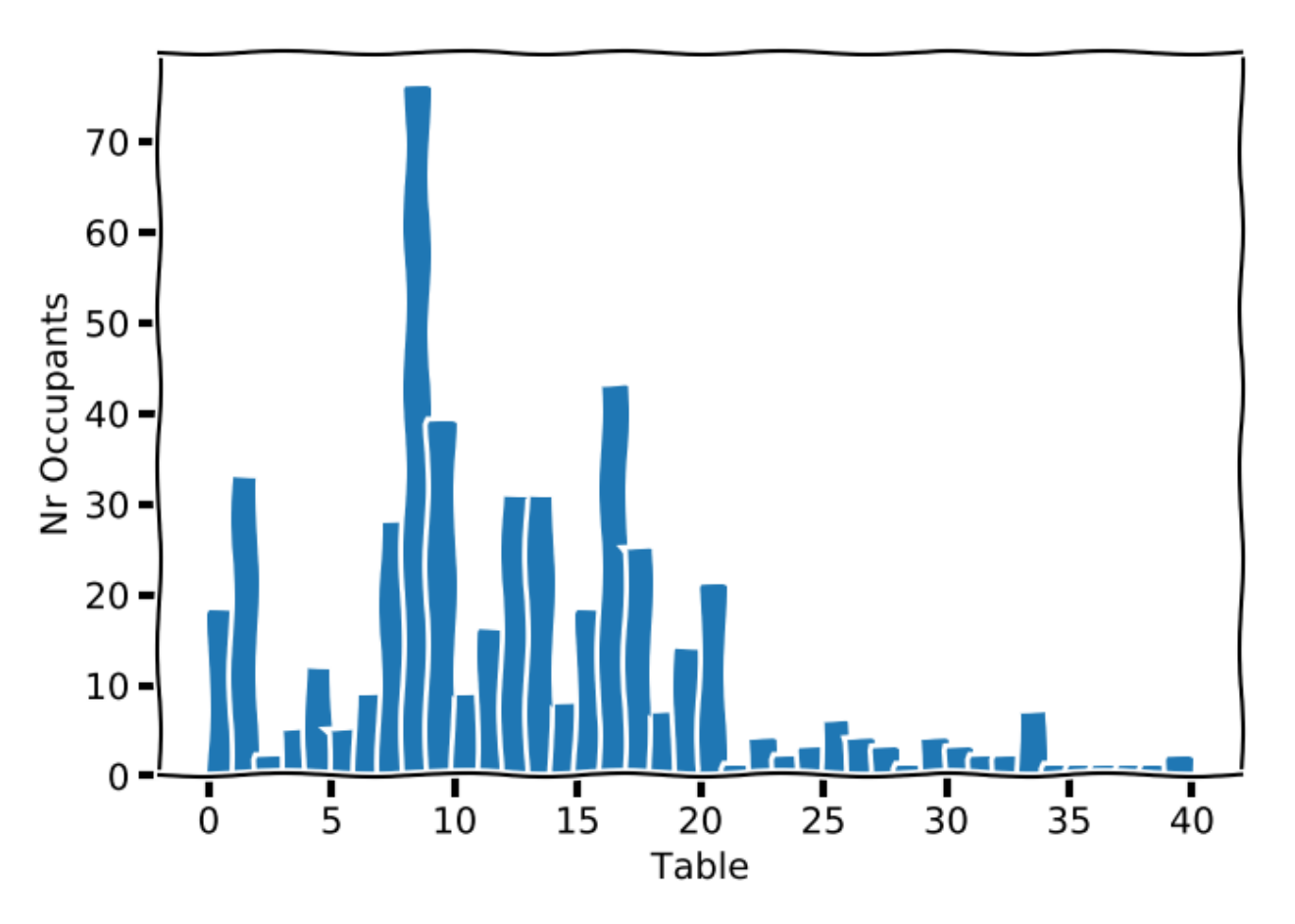

$N = 500 \quad \alpha = 1.0$

# Chinese Resturant Process



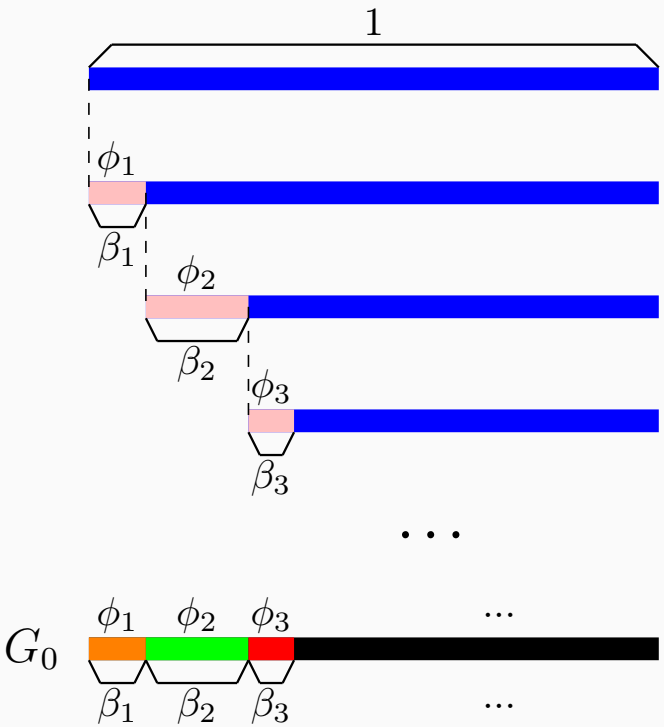

$N = 500 \quad \alpha = 2.0$

# Chinese Resturant Process




$N = 500 \quad \alpha = 10.0$

# Stick-Breaking




$$G \sim \mathrm{DP}(\mathcal{H}, \alpha)$$

$$\hat{\beta}_k \sim \mathrm{Beta}(1, \alpha)$$

$$\beta_k = \hat{\beta}_k \prod_{l=1}^{k-1} (1 - \hat{\beta}_l)$$

$$\Phi \sim \mathcal{H}$$

$$G = \sum_{k=1}^{\infty} \beta_k \Phi_k$$

# Bayesnet

```
\usetikzlibrary{bayesnet}
\begin{tikzpicture}
  \node[obs] (x) {$\mathbf{x}_i$};
  \node[latent, above=of x] (z) {$\mathbf{z}_i$};
  \node[latent,right=of x] (w) {$\mathbf{W}$};
  \node[left=of x] (mu) {$\boldsymbol{\mu}$};
  \node[above left=of x](sigma) {$\sigma^2$};
  \edge {z,sigma,mu,w} {x};
  \plate {} {(x)(z)} {$N$};
\end{tikzpicture}
```

# Infinite Mixture Model

1. Pick first data-point

# Infinite Mixture Model

1. Pick first data-point
2. Pick the first mixture

# Infinite Mixture Model

1. Pick first data-point
2. Pick the first mixture
3. Pick parameters associated with this mixture $\mu_k, \sigma_k$

# Infinite Mixture Model

1. Pick first data-point
2. Pick the first mixture
3. Pick parameters associated with this mixture $\mu_k, \sigma_k$
4. Next data-point

# Infinite Mixture Model

1. Pick first data-point
2. Pick the first mixture
3. Pick parameters associated with this mixture $\mu_k, \sigma_k$
4. Next data-point
5. Associate with a new mixture or create new

# Infinite Mixture Model

1. Pick first data-point
2. Pick the first mixture
3. Pick parameters associated with this mixture $\mu_k, \sigma_k$
4. Next data-point
5. Associate with a new mixture or create new
6. If new, pick new parameters

# Infinite Mixture Model

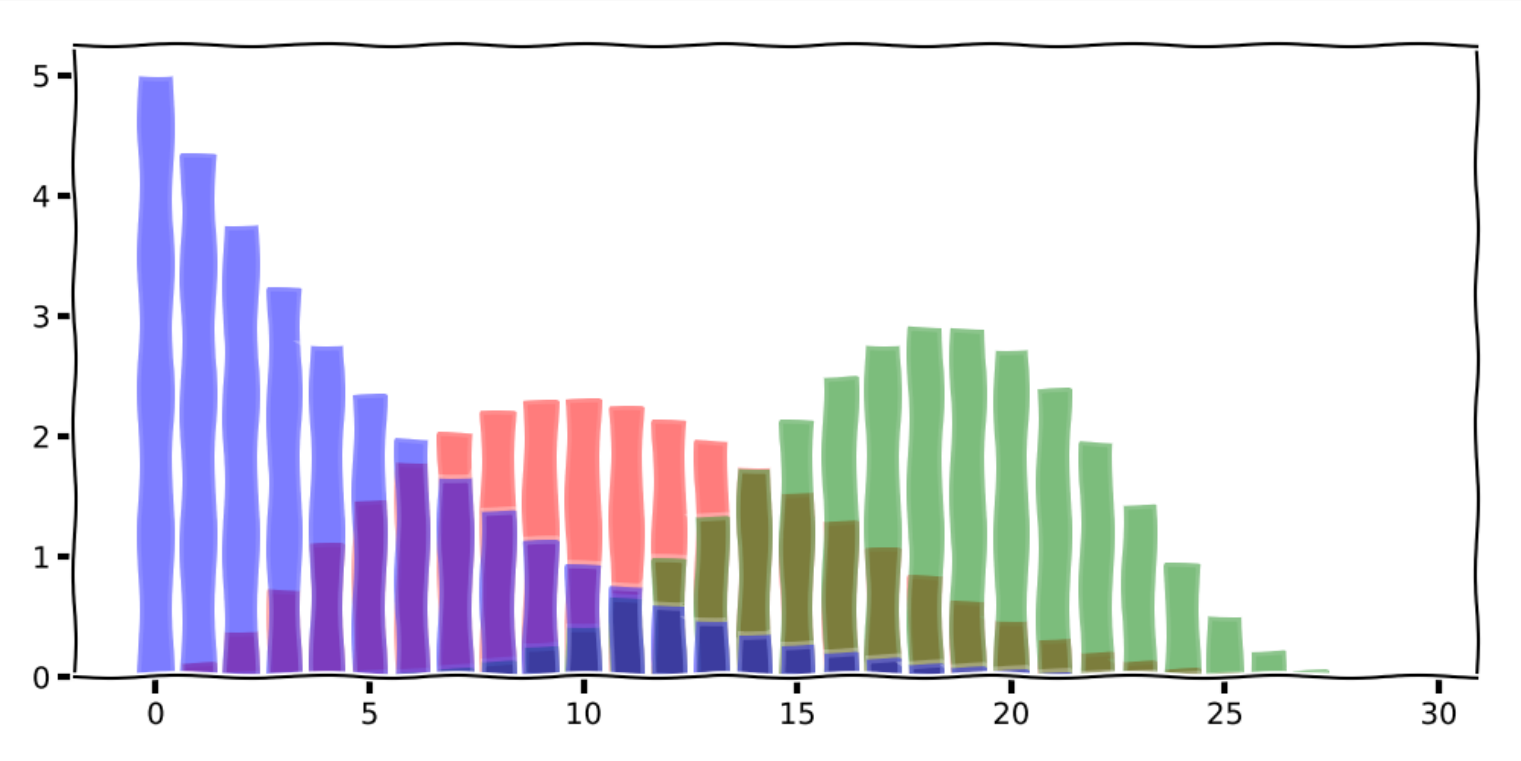
1. Pick first data-point
2. Pick the first mixture
3. Pick parameters associated with this mixture $\mu_k, \sigma_k$
4. Next data-point
5. Associate with a new mixture or create new
6. If new, pick new parameters
7. Repeat

# Latent Dirichlet Allocation

# Text Data

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

# Bag-of-Words



# Topic Model

**Assumptions**

- Documents contains a blend of topics
- Each topic has a characteristic distribution of words
- For simplicity lets assume a finite (known) number of words

# Generative Model

1. Choose a random distribution over topics
2. Randomly choose a topic from the topic distribution
3. Randomly choose a word from the word distribution of this topic

# Topic Distribution

| Topic | $p$(Topic) |
|---|---|
| Computer Science | 0.3 |
| Pugs | 0.5 |
| Bristol City | 0.1 |
| C64 | 0.1 |

$$p(\text{Topic}) = p(\beta)$$

# Word Topic Distribution

| Topic | 8-bit | Tammy | AI | Hugs |
|---|---|---|---|---|
| Computer Science | 0.6 | 0.09 | 0.3 | 0.01 |
| Pugs | 0.1 | 0.1 | 0.3 | 0.5 |
| Bristol City | 0.1 | 0.8 | 0.05 | 0.05 |
| C64 | 0.8 | 0.01 | 0.15 | 0.04 |

$$p(\text{Word}|\text{Topic}) = p(w_{d,n}|\beta_k)$$

# Topic Model

$w_{d,n}$ $n$:th word in $d$:th document

# Topic Model

$w_{d,n}$ $n$:th word in $d$:th document

$\beta_k$ topic-word distribution for $k$:th topic

# Topic Model

$w_{d,n}$ $n$:th word in $d$:th document

$\beta_k$ topic-word distribution for $k$:th topic

$\theta_d$ topic distribution for $d$:th document

# Topic Model

$w_{d,n}$ $n$:th word in $d$:th document

$\beta_k$ topic-word distribution for $k$:th topic

$\theta_d$ topic distribution for $d$:th document

$z_{d,n}$ topic assignment for $n$:th word in $d$:th document

# Topic Model

$w_{d,n}$ $n$:th word in $d$:th document

$\beta_k$ topic-word distribution for $k$:th topic

$\theta_d$ topic distribution for $d$:th document

$z_{d,n}$ topic assignment for $n$:th word in $d$:th document

*Now lets write down the joint distribution of these according to our generative model*

# Topic Model

- To generate the $n$:th word in the $d$:th document I need to know the topic that has been assigned to this word and the topic-word distribution

$$p(w_{d,n}|\beta, z_{d,n})$$

# Topic Model

- To generate the $n$:th word in the $d$:th document I need to know the topic that has been assigned to this word and the topic-word distribution

$$p(w_{d,n}|\beta, z_{d,n})$$

- To generate a topic assignment $z_{d,n}$ I need to know the topic distribution for document $d$

$$p(z_{d,n}|\theta_d)$$

# Topic Model

- To generate the $n$:th word in the $d$:th document I need to know the topic that has been assigned to this word and the topic-word distribution

$$p(w_{d,n}|\beta, z_{d,n})$$

- To generate a topic assignment $z_{d,n}$ I need to know the topic distribution for document $d$

$$p(z_{d,n}|\theta_d)$$

- If we know the topic assignment and the topic-word distribution we can assume the words to be independent

$$p(w_d|\beta, z_d) = \prod_{n=1}^{N} p(w_{d,n}|\beta, z_{d,n})$$

# Topic Model

- We can assume that the topic-document distribution is not dependent between documents

$$p(\theta) = \prod_{d=1}^{D} p(\theta_d)$$

# Topic Model

- We can assume that the topic-document distribution is not dependent between documents

$$p(\theta) = \prod_{d=1}^{D} p(\theta_d)$$

- We can assume that topic word distribution is independent

$$p(\beta) = \prod_{k=1}^{K} p(\beta_k)$$

# Topic Model

$$p(w, z, \theta, \beta) = \underbrace{\prod_{k=1}^{K} p(\beta_k) \underbrace{\prod_{d=1}^{D} p(\theta_d) \underbrace{\prod_{n=1}^{N} p(w_{d,n}|\beta, z_{d,n}) p(z_{d,n}|\theta_d)}_{\text{word}}}_{\text{document}}}_{\text{corpus}}$$

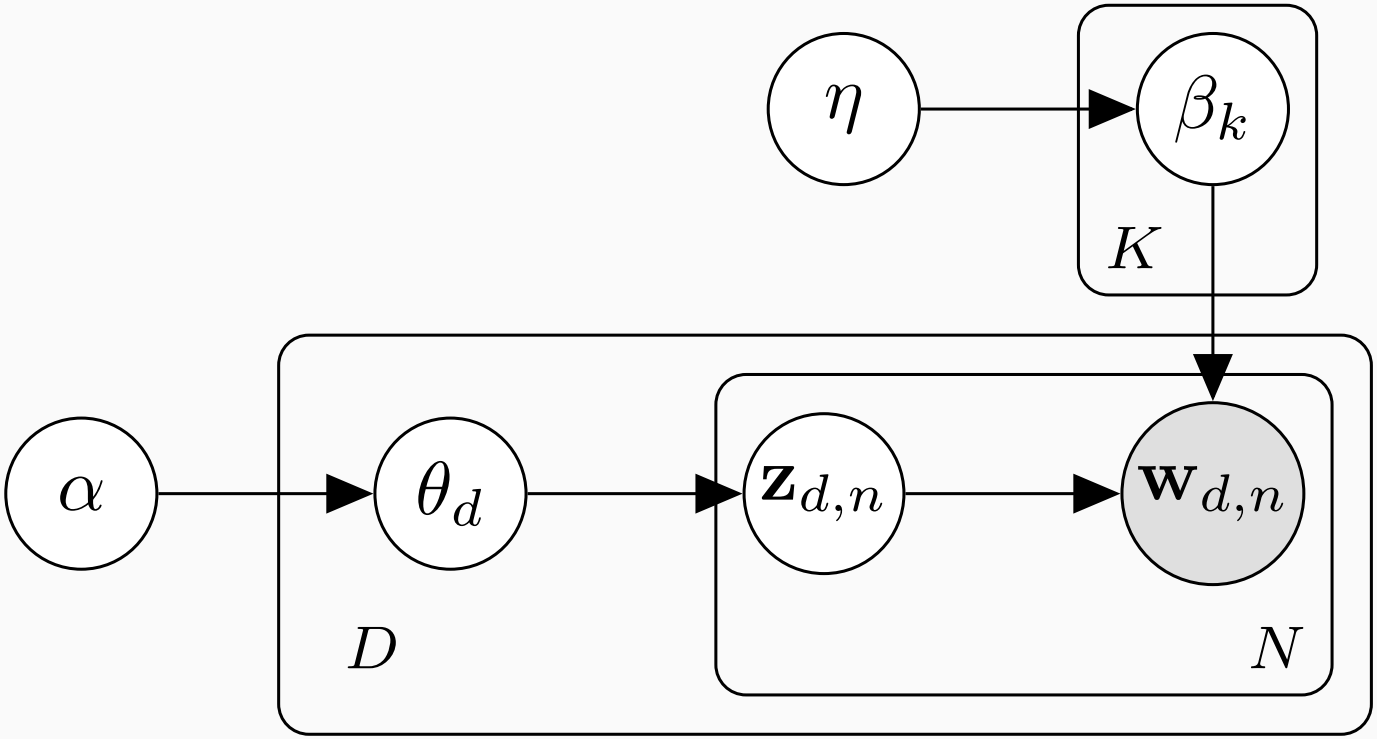
- This is the joint distribution of a specific topic model
- The assumptions we have made are called a Latent Dirichlet Allocation [1]
- This specifies a specific probability distribution by our assumptions

# Topic Model: Learning

$$p(z, \theta, \beta | w) = \frac{p(w, z, \theta, \beta)}{p(w)}$$

- We only observe the words
- Want to infer all the distributions
- Need assumptions

# Graphical Model

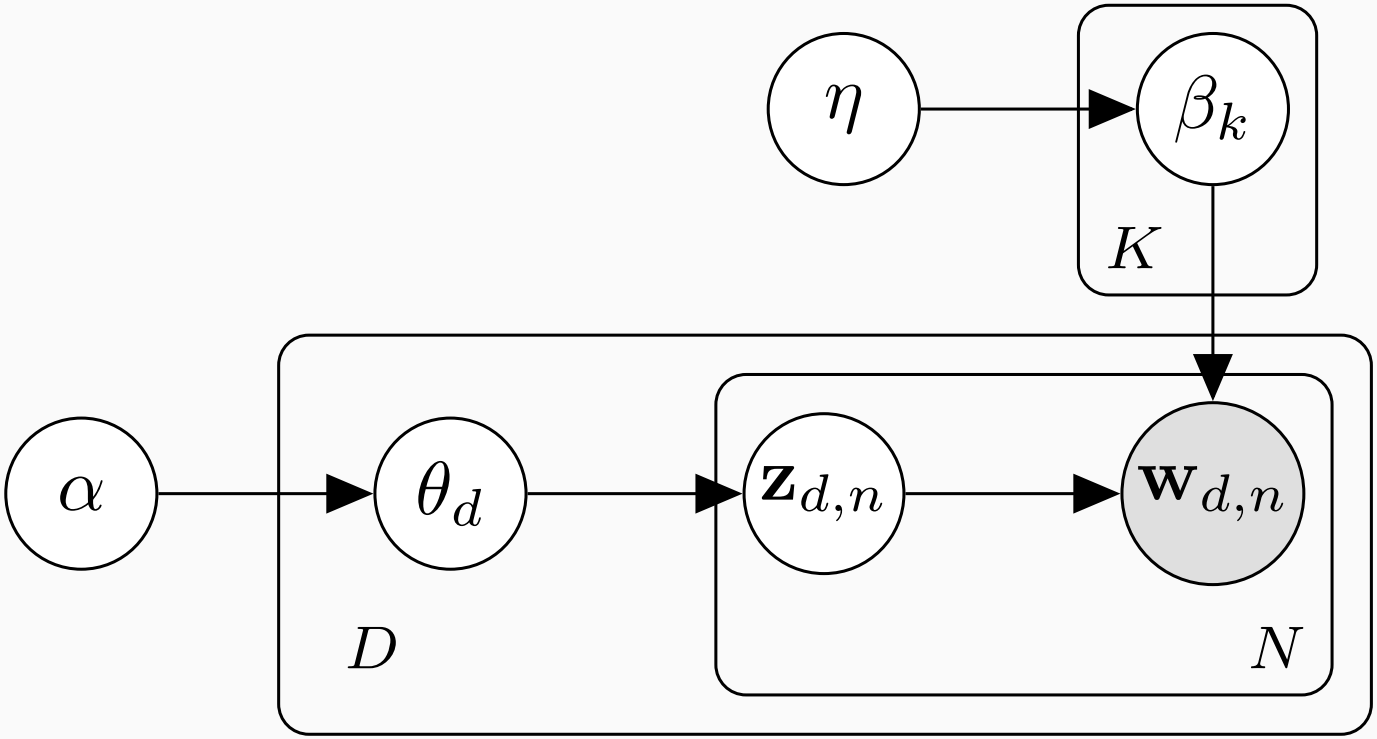


# Topic Model: Evidence

$$
\begin{aligned}
p(w) &= \int p(w, z, \theta, \beta) \mathrm{d}\beta \mathrm{d}\theta dz \\
&= \int \prod_{k=1}^{K} p(\beta_k) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(w_{d,n}|\beta, z_{d,n}) p(z_{d,n}|\theta_d) \mathrm{d}\beta \mathrm{d}\theta dz
\end{aligned}
$$

- Distribution over corpus for *any topic model* weighted by our assumptions

# Graphical Model



# Topic Model: Distributions

- We still have to choose the specific form of the distribution
- Topic-word: $\beta_k \sim \text{Dir}(\eta)$
- Topic-proportions: $\theta_d \sim \text{Dir}(\alpha)$
- Topic assignment: $z_{d,n} \sim \text{Mult}(\theta_d)$
- Word assginment: $w_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$

# Topic Model: Learning

$$p(z, \theta, \beta | w) = \frac{p(w, z, \theta, \beta)}{p(w)}$$

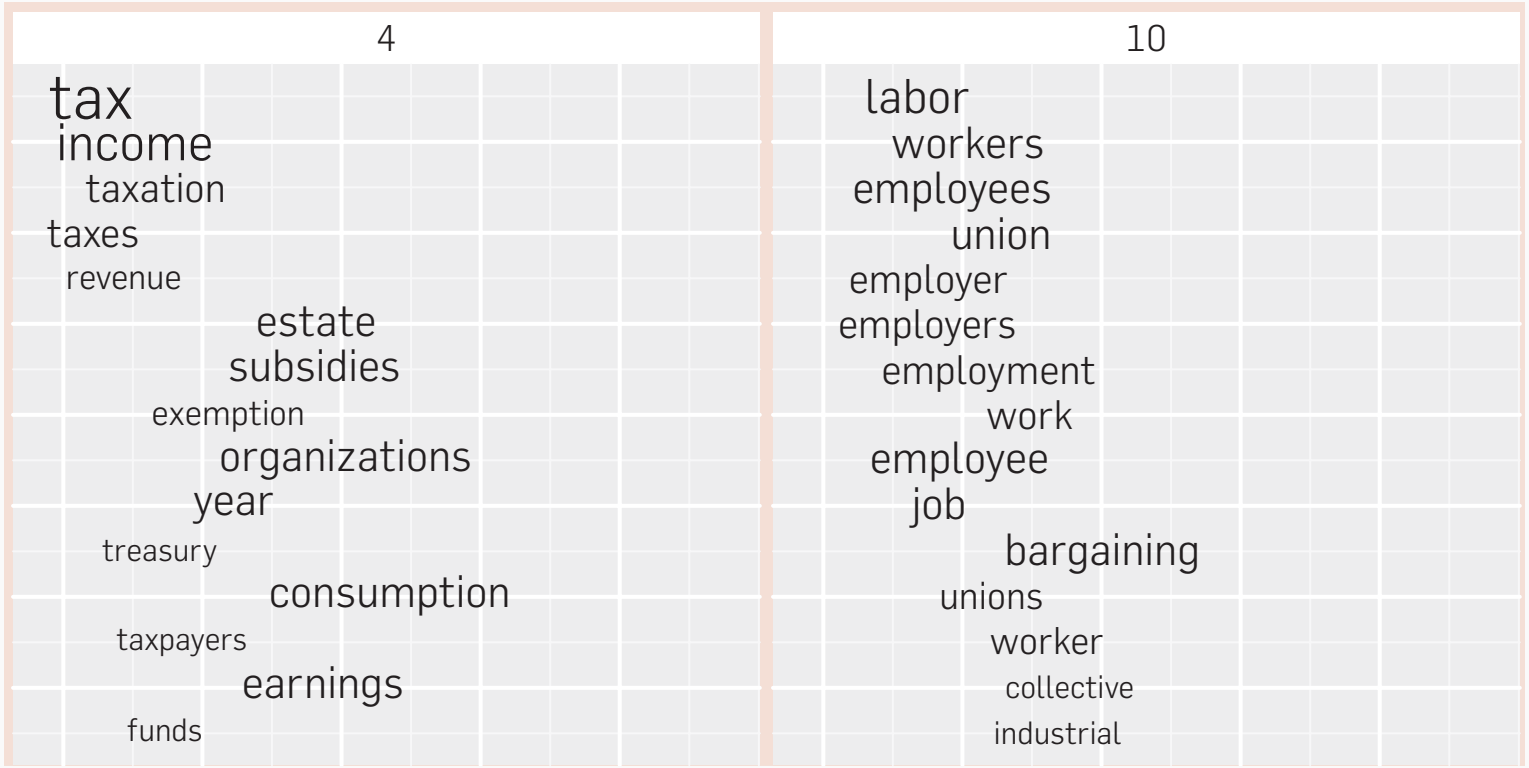
- Completely intractable
- Efficient approximations (Part III)
- Possible to learn from huge amounts of data

# Experiments

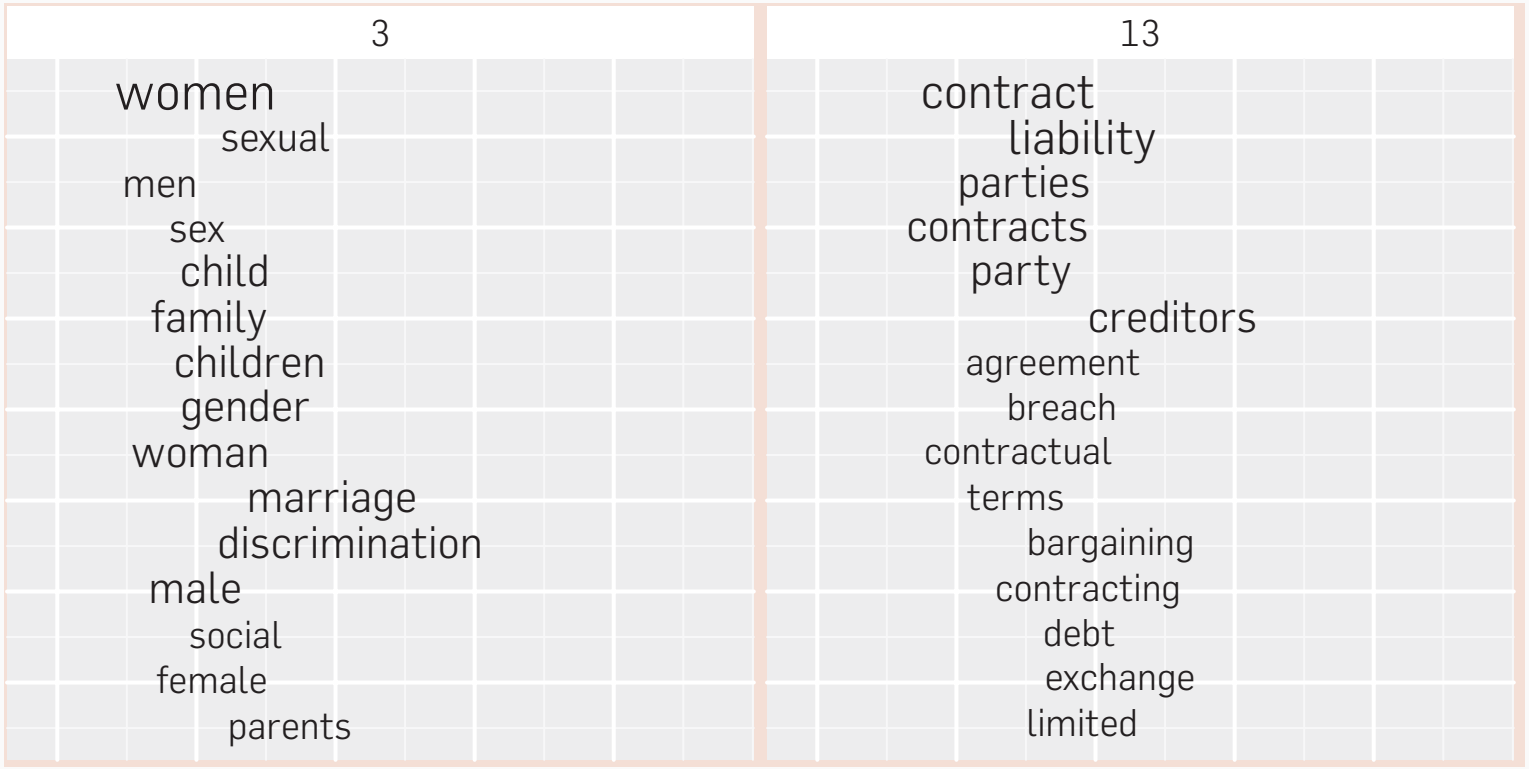
# Science [2]

| **“Genetics”** | **“Evolution”** | **“Disease”** | **“Computers”** |
|---|---|---|---|
| human | evolution | disease | computer |
| genome | evolutionary | host | models |
| dna | species | bacteria | information |
| genetic | organisms | diseases | data |
| genes | life | resistance | computers |
| sequence | origin | bacterial | system |
| gene | biology | new | network |
| molecular | groups | strains | systems |
| sequencing | phylogenetic | control | model |
| map | living | infectious | parallel |
| information | diversity | malaria | methods |
| genetics | group | parasite | networks |
| mapping | new | parasites | software |
| project | two | united | new |
| sequences | common | tuberculosis | simulations |

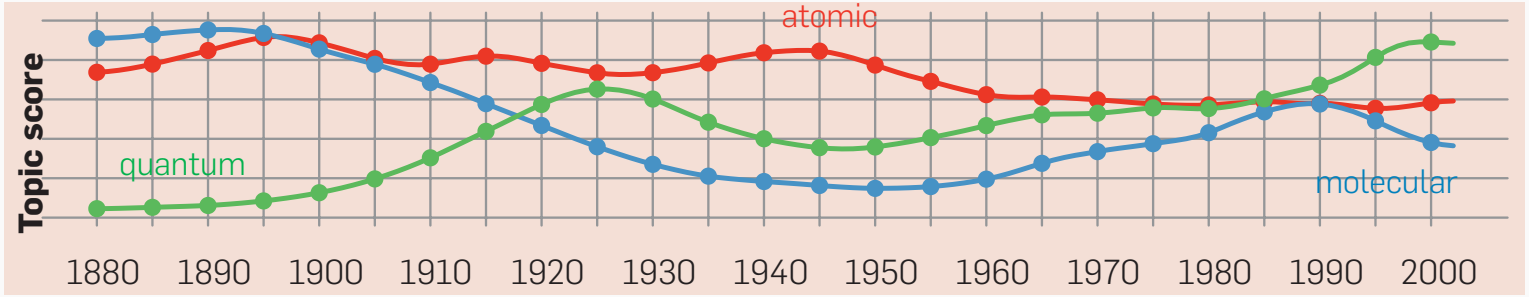
# Yale Law Journal [2]

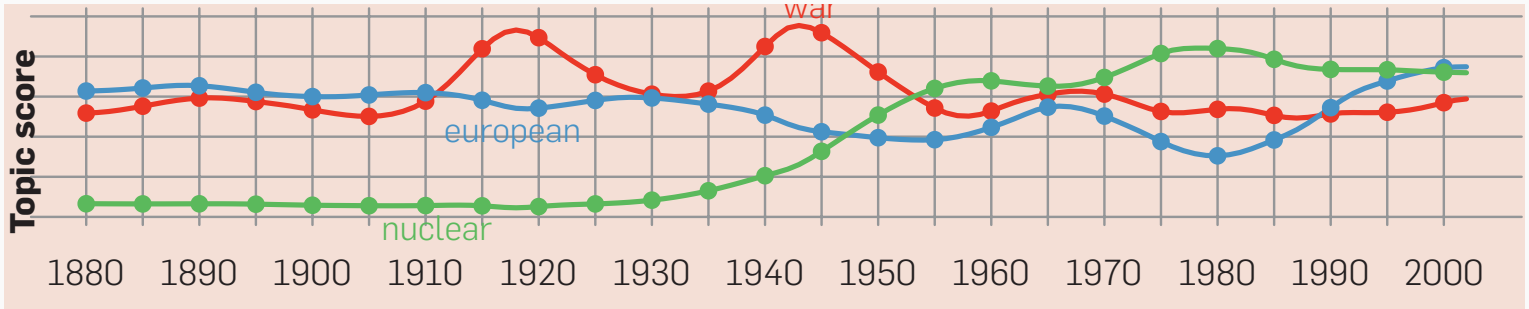


# Yale Law Journal [2]



# Science [2]

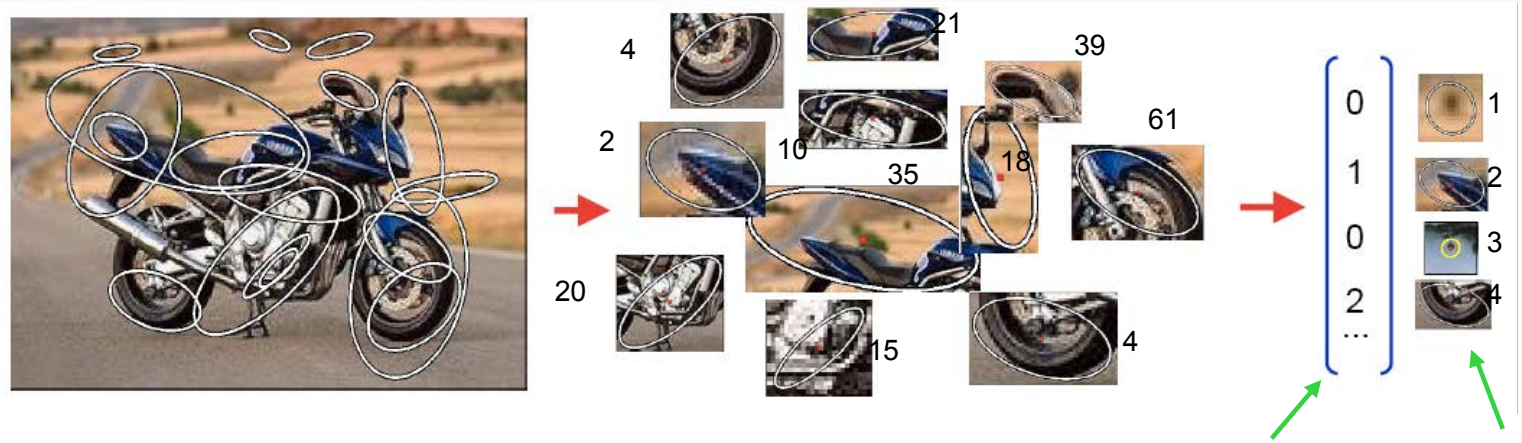

Topic score
atomic
quantum
molecular
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000


Topic score
war
european
nuclear
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000

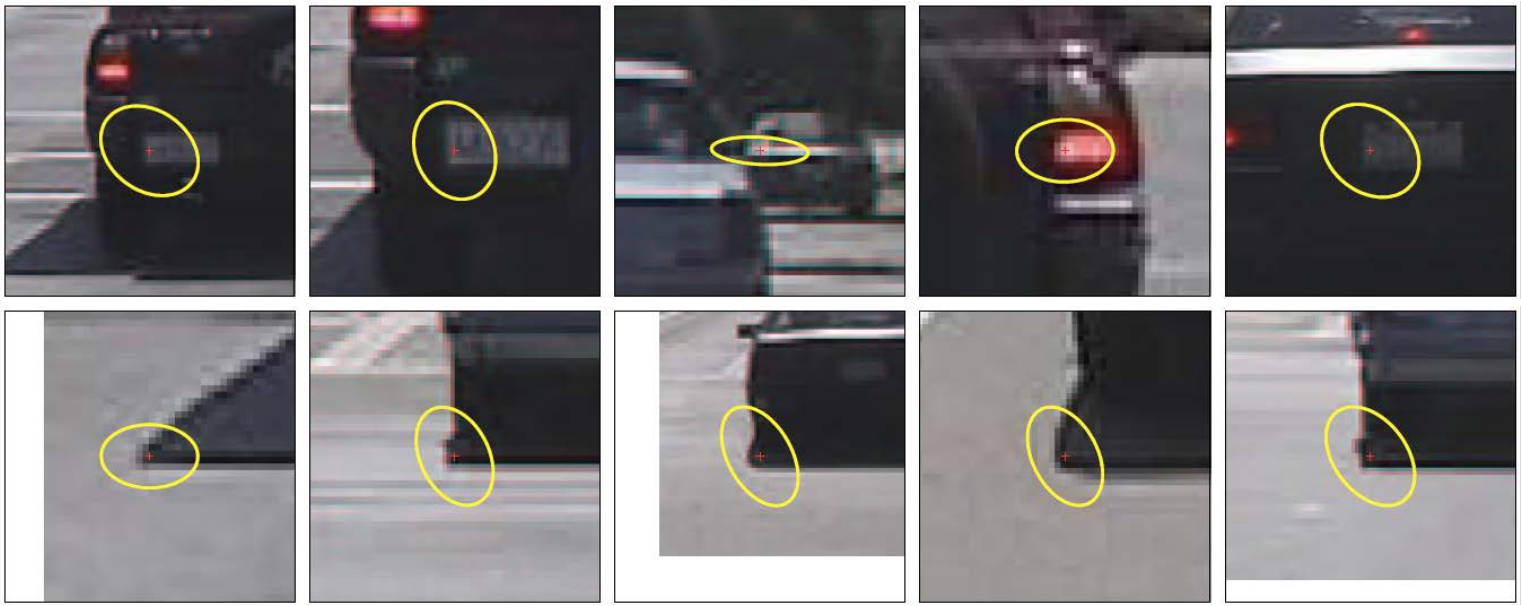
# Beyond words

# Computer Vision

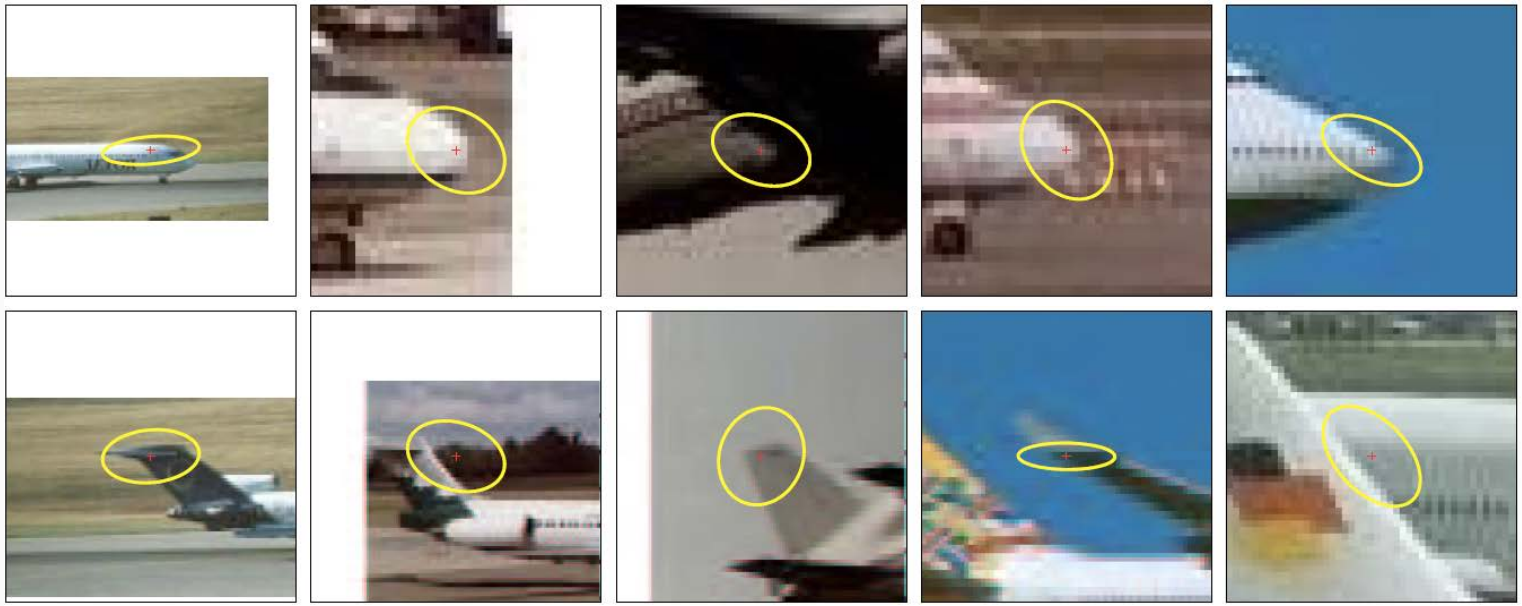
# Computer Vision



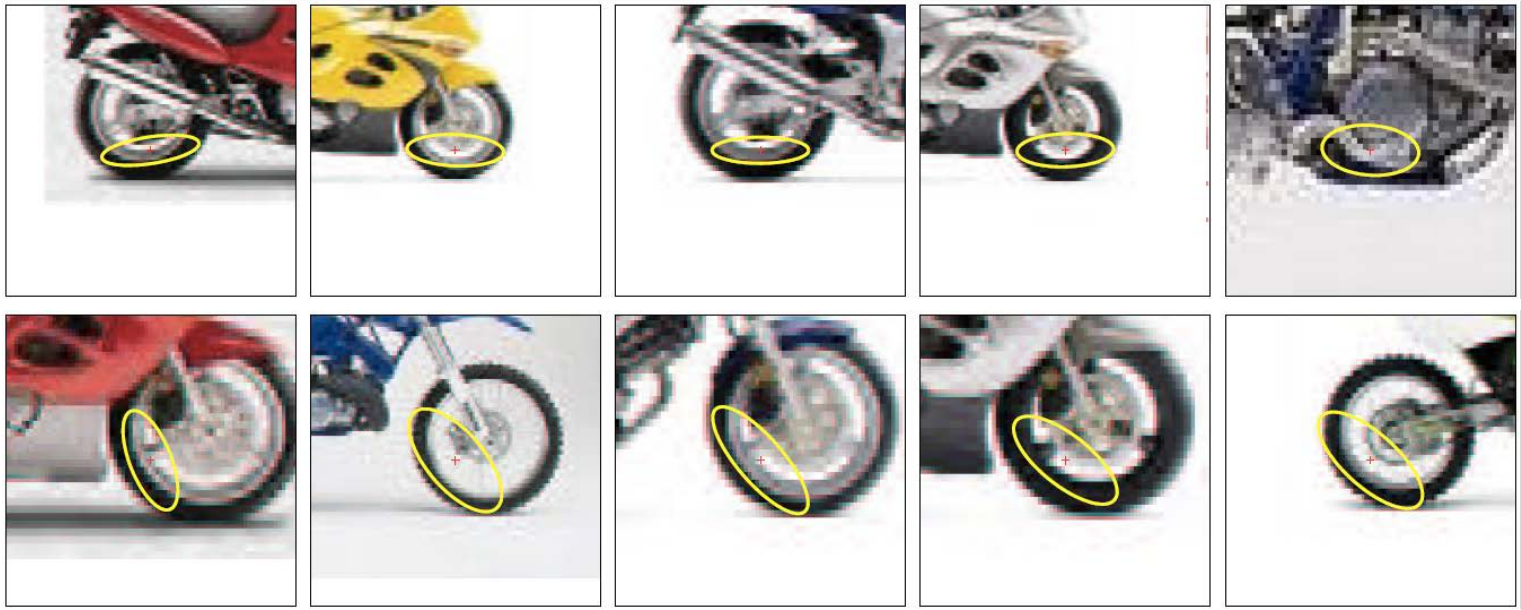
# Computer Vision



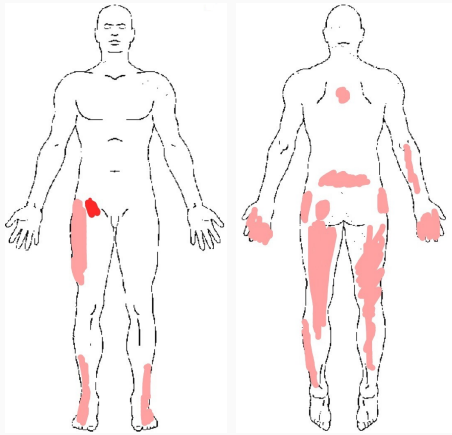
# Computer Vision



# Computer Vision



# Health [3]

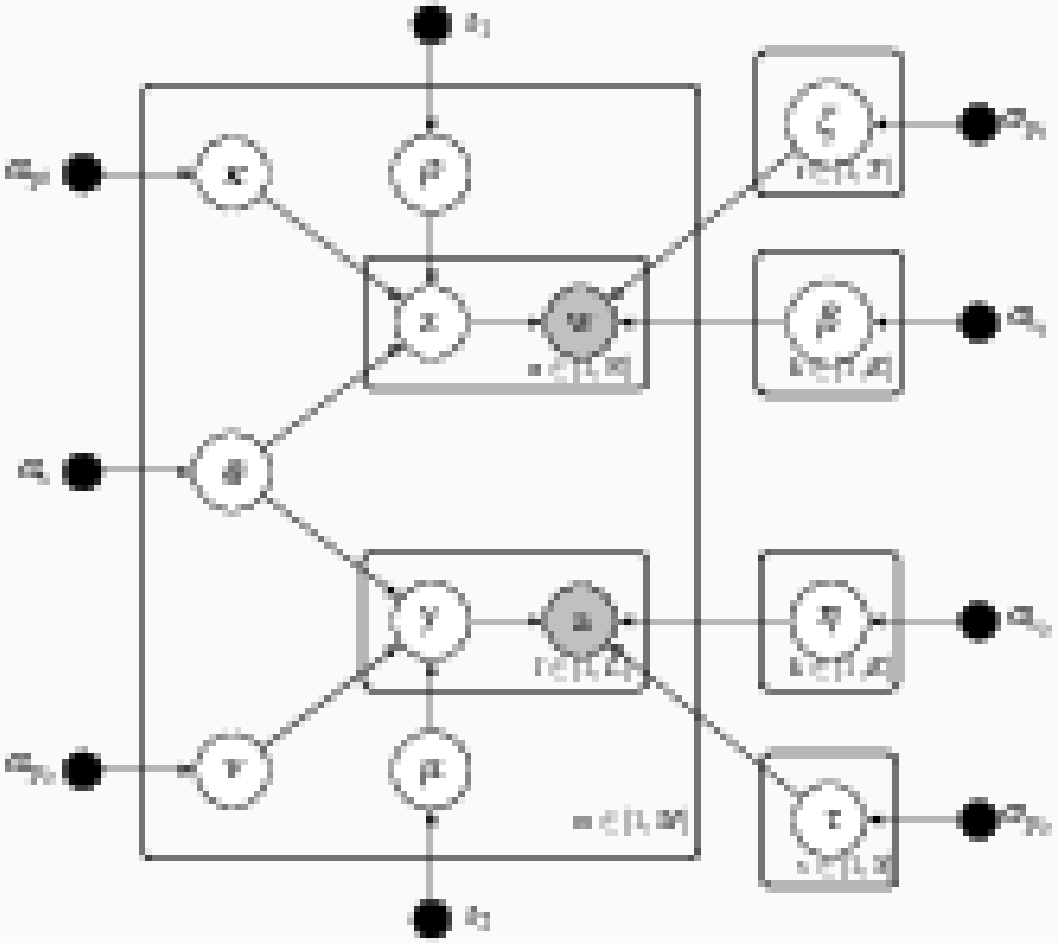


**Symptom diagnoses:** Interscapular discomfort; R arm discomfort; B hands discomfort; Lumbago; B crest of the ilium discomfort; L side thigh discomfort; B back thigh discomfort; B calf discomfort; B achilles tendinitis; B shin discomfort; R inguinal discomfort;

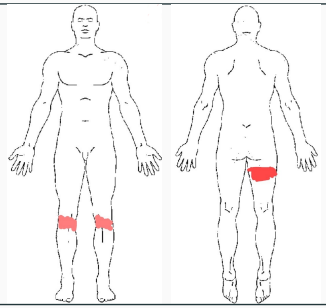
**Pattern diagnoses** B L5 Radiculopathy; B S1 Radiculopathy; B C7 Radiculopathy;

**Pathophysiological diagnoses** DLI L4-L5; DLI S1-S2; DLI C6-C7

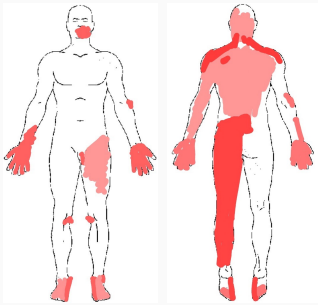
# Health [3]



# Health [3]

| | |
|---|---|
|  | **6 Prd:** R back thigh dcf; L PFS (Patellofemoral pain syndrome); R PFS;<br><br>L L5 Rdc; R L5 Rdc; DLI L4-L5;<br><br>**6 GT:** R back thigh dcf; L PFS; R PFS;<br><br>L L5 Rdc; R L5 Rdc; DLI L4-L5; |

# Health [3]



**50 Prd:** Headache; L neck dcf; R neck dcf; Neck dcf; L upper trapezius dcf; R upper trapezius dcf; L shoulder dcf; L hand dcf; R hand dcf; Interscapular dcf; Lumbago; Lateral abdominal dcf; L groin dcf; L side thigh dcf; R side thigh dcf; L calf dcf; L back thigh dcf; L crest of the ilium dcf; R crest of the ilium dcf; R foot arch dcf; L toe joint dcf; R toe joint dcf; L medial elbow dcf; L ankle dcf; R ankle dcf; L foot arch dcf; L PFS; L dorsal knee dcf; R medial knee dcf;

L C2 Rdc; R C2 Rdc; L C3 Rdc; R C3 Rdc; L C4 Rdc; R C4 Rdc; R C6 Rdc; L C6 Rdc; L C7 Rdc; R C7 Rdc; L L5 Rdc; R L5 Rdc; L S1 Rdc; R S1 Rdc; DLI C2-C3; DLI C3-C4; DLI C5-C6; DLI C6-C7; DLI L4-L5; DLI L5-S1; OB;

**50 GT:** L back headache; R back headache; Neck dcf; L jaw dcf; L upper trapezius dcf; R upper trapezius dcf; L arm dcf; R arm dcf; L lateral elbow dcf; R lateral elbow dcf; L hand joint dcf; R hand joint dcf; L hand dcf; R hand dcf; L thumb dcf; R thumb dcf; L finger dcf; R finger dcf; Lumbago; L groin dcf; L back thigh dcf; L calf dcf; L medial knee dcf; L ankle dcf; R ankle dcf; R medial knee dcf; R big toe dcf; L big toe dcf;

L C2 Rdc; R C2 Rdc; L C3 Rdc; R C3 Rdc; L C4 Rdc; R C4 Rdc; L C5 Rdc; R C5 Rdc; L C6 Rdc; R C6 Rdc; L C7 Rdc; R C7 Rdc; L L4 Rdc; L L5 Rdc; R L5 Rdc; L S1 Rdc; R S1 Rdc; Craniocervical joint injury; DLI C4-C5; DLI L3-L4; DLI L4-L5; DLI L5-S1;

# Summary

# Summary

- How to build up a more complicated model from scratch
- Thinking about conditional distributions as generative models
- Topic models

# Next time

- We are done with Part II
- This was the big part of the unit
- How to build models
- Monday: Summary on Part II
- Final part: how to perform inference

eof

# References

David M Blei, Andrew Y Ng, and Michael I Jordan.
**Latent dirichlet allocation.**
3:993–1022, March 2003.

David M. Blei.
**Probabilistic topic models.**
*Commun. ACM*, 55(4):77–84, April 2012.

Cheng Zhang, Hedvig Kjellström, Carl Henrik Ek, and Bo C. Bertilson.
**Diagnostic prediction using discomfort drawings with IBTM.**
In *Proceedings of the 1st Machine Learning in Health Care, MLHC 2016, Los Angeles, CA, USA, August 19-20, 2016*, pages 226–238, 2016.