RWTH AACHEN UNIVERSITY

Advanced Machine Learning
Lecture 14

Hierarchical Dirichlet Processes II

12.12.2012

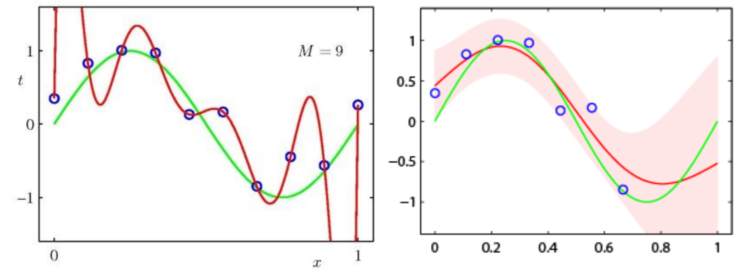
Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de/

leibe@vision.rwth-aachen.de

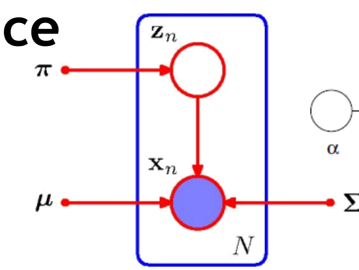
This Lecture: *Advanced Machine Learning*

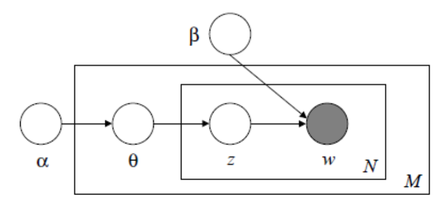
- Regression Approaches
  - Linear Regression
  - Regularization (Ridge, Lasso)
  - Kernels (Kernel Ridge Regression)
  - Gaussian Processes

$$f : \mathcal{X} \rightarrow \mathbb{R}$$



- Bayesian Estimation & Bayesian Non-Parametrics
  - Prob. Distributions, Approx. Inference
  - Mixture Models & EM
  - Dirichlet Processes
  - Latent Factor Models
  - Beta Processes





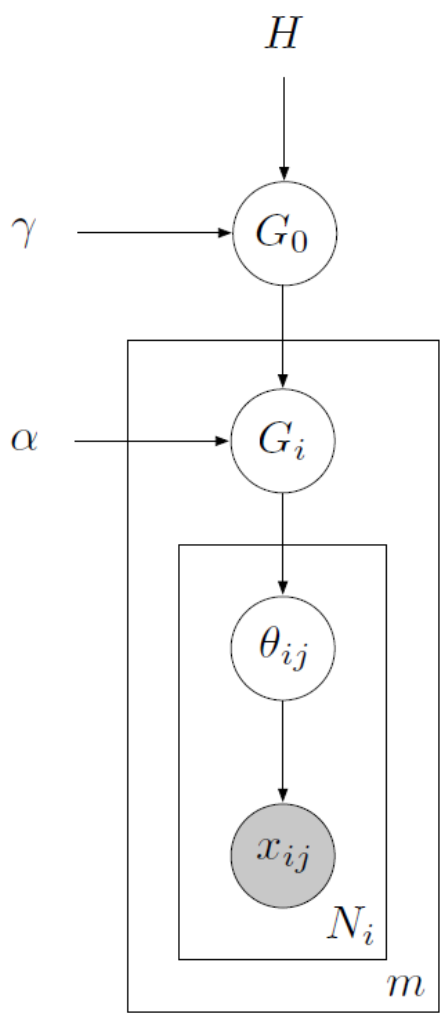
- SVMs and Structured Output Learning
  - SV Regression, SVDD
  - Large-margin Learning

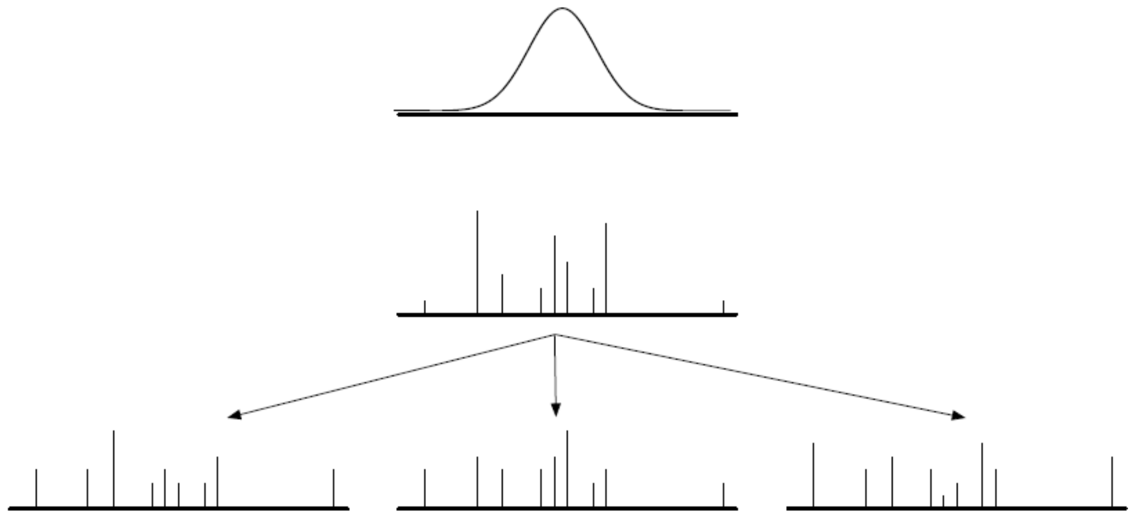
$$f : \mathcal{X} \rightarrow \mathcal{Y}$$

Topics of This Lecture

- Hierarchical Dirichlet Processes
  - Recap
  - Chinese Restaurant Franchise
  - Gibbs sampling for HDPs
  - CRF Sampler
- Applications
  - Example: Document topic modeling
  - Latent Dirichlet Allocation (LDA)

Recap: Hierarchical Dirichlet Processes

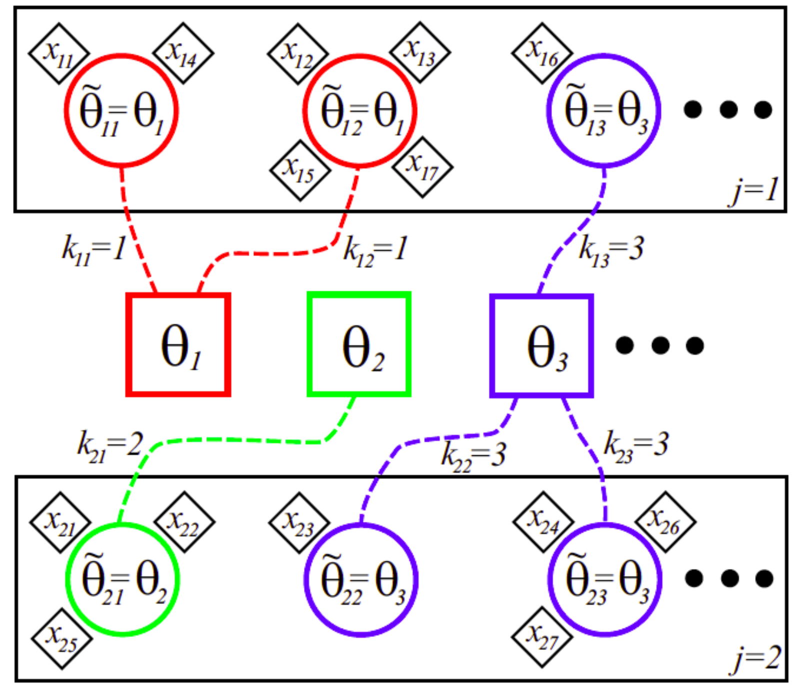




$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma, H) \\ G_i &\sim \mathrm{DP}(\alpha, G_0) \\ \theta_{ij} | G_i &\sim G_i \\ x_{ij} | \theta_{ij} &\sim p(x_{ij} | \theta_{ij}) \end{aligned}$$

Recap: Chinese Restaurant Franchise (CRF)

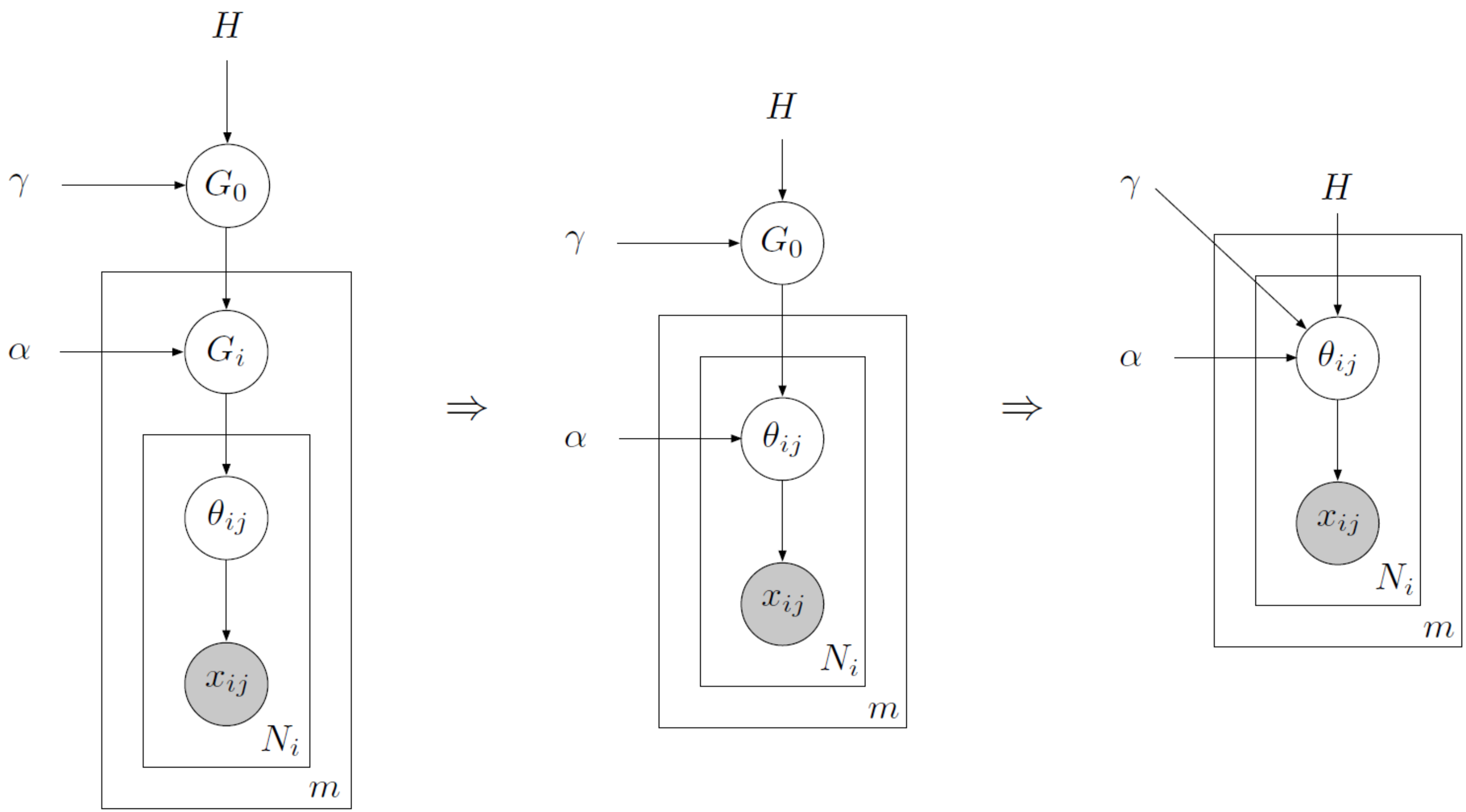
- **Chain of Chinese restaurants**
  - Each restaurant has an unbounded number of tables.
  - There is a global menu with an unbounded number of dishes.
  - The first customer at a table selects the dish for that table from the global menu.



- **Reinforcement effects**
  - Customers prefer to sit at tables with many other customers, and prefer dishes that are chosen by many other customers.
  - Dishes are chosen with probability proportional to the number of tables (franchise-wide) that have previously served that dish.

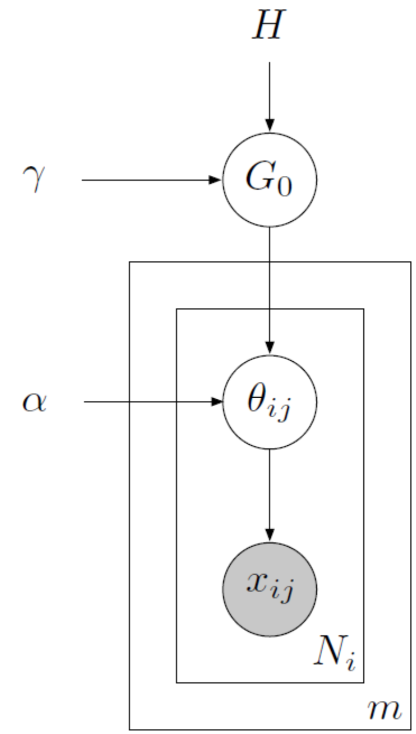
Chinese Restaurant Franchise (CRF)

- **Examine marginal properties of HDP**
  - First integrate out G_i, then G_0.



Chinese Restaurant Franchise (CRF)

- **Step 1: Integrate out G_i:**
  - **Variable definitions**
    - θ_{ij} : RV for customer i in restaurant j.
    - θ_{jt}^* : RV for table t in restaurant j.
    - θ_k^{**} : RV for dish k.
    - m_{jk}: number of tables in rest. j serving dish k.
    - n_{jtk}: number of customers in rest. j sitting at table t and being served dish k.
    - We denote marginal counts by dots, e.g. $m_{j\cdot} = \sum_{k=1}^{K} m_{jk}$



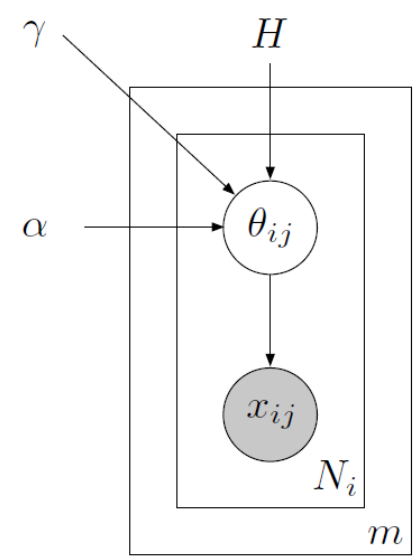
  - **Integration yields a set of conditional distributions described by a Polya urn scheme**

$$\theta_{ij} | \theta_{1j}, \ldots, \theta_{i-1,j}, \alpha, G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{\alpha + n_{j\cdot\cdot}} \delta_{\theta^*_{jt}} + \frac{\alpha}{\alpha + n_{j\cdot\cdot}} G_0$$

Image source: Kurt Miller

Chinese Restaurant Franchise (CRF)

- **Step 2: Integrate out G_0:**
  - **Variable definitions**
    - θ_{ij} : RV for customer i in restaurant j.
    - θ_{jt}^{*} : RV for table t in restaurant j.
    - θ_{k}^{**} : RV for dish k.
    - m_{jk}: number of tables in rest. j serving dish k.
    - n_{jtk}: number of customers in rest. j sitting at table t and being served dish k.
    - We denote marginal counts by dots, e.g. $m_{j\cdot} = \sum_{k=1}^{K} m_{jk}$



  - **Again, we get a Polya urn scheme**

$$\theta_{jt}^{*} | \theta_{11}^{*}, \ldots, \theta_{1,m_{1,\cdot}}^{*}, \ldots, \theta_{j,t-1}^{*}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{\gamma + m_{\cdot\cdot}} \delta_{\theta_k^{**}} + \frac{\gamma}{\gamma + m_{\cdot\cdot}} H$$

Image source: Kurt Miller

Inference for HDP: CRF Sampler

- Using the CRF representation of the HDP
  - Customer i in restaurant j is associated with i.i.d draw from G_i and sits at table t_{ij}.
  - Table t in restaurant j is associated with i.i.d draw from G_0 and serves dish k_{jt}.
  - Dish k is associated with i.i.d draw from H.

- Gibbs sampling approach
  - Iteratively sample the table and dish assignment variables, conditioned on the state of all other variables.
  - The parameters θ_{ij} are integrated out analytically (assuming conjugacy).
  - To resample, make use of exchangeability.

  $\Rightarrow$ Imagine each customer i being the last to enter restaurant j.

Inference for HDP: CRF Sampler

- **Procedure**
  1. **Resample t_{ij} according to the following distribution**

$$\begin{cases} t_{ij} = t & \text{with prob.} \quad \propto \frac{n_{jt\cdot}^{\neg ij}}{n_{j\cdot\cdot}^{\neg ij}+\alpha} f_{k_{jt}}(\{x_{ij}\}) \\ t_{ij} = t^{\text{new}}, k_{jt^{\text{new}}} = k & \text{with prob.} \quad \propto \frac{\alpha}{n_{j\cdot\cdot}^{\neg ij}+\alpha} \frac{m_{\cdot k}^{\neg ij}}{m_{\cdot\cdot}^{\neg ij}+\gamma} f_k(\{x_{ij}\}) \\ t_{ij} = t^{\text{new}}, k_{jt^{\text{new}}} = k^{\text{new}} & \text{with prob.} \quad \propto \frac{\alpha}{n_{j\cdot\cdot}^{\neg ij}+\alpha} \frac{\gamma}{m_{\cdot\cdot}^{\neg ij}+\gamma} f_{k^{\text{new}}}(\{x_{ij}\}) \end{cases}$$

where $\neg ij$ denotes counts in which customer i in restaurant j is removed from the CRF. (If this empties a table, we also remove the table from the CRF, along with the dish on it.)

  - **The terms $f_k(\{x_{ij}\})$ are defined as follows**

$$f_k(\{x_{ij}\}_{ij\in D}) = \frac{\int h(\theta) \prod_{i'j' \in D_k \cup D} p(x_{i'j'}|\theta)\mathrm{d}\theta}{\int h(\theta) \prod_{i'j' \in D_k \setminus D} p(x_{i'j'}|\theta)\mathrm{d}\theta}$$

where D_K denotes the set of indices associated with dish k.

Inference for HDP: CRF Sampler

- **Procedure (cont'd)**
  2. **Resample k_{jt} (Gibbs update for the dish)**

$$k_{jt} = \begin{cases} k & \text{with prob.} \quad \propto \frac{m_{\cdot k}^{\neg jt}}{m_{\cdot\cdot}^{\neg jt} + \gamma} f_k(\{x_{ij} : t_{ij} = t\}) \\ k^{\text{new}} & \text{with prob.} \quad \propto \frac{\gamma}{m_{\cdot\cdot}^{\neg jt} + \gamma} f_{k^{\text{new}}}(\{x_{ij} : t_{ij} = t\}) \end{cases}$$

- **Remarks**
  - Computational cost of Gibbs updates is dominated by computation of the marginal conditional probabilities $f_k(\cdot)$.
  - Still, the number of possible events that can occur at one Gibbs step is one plus the total number of tables and dishes in all restaurants that are ancestors of j.
  - This number can get quite large in deep or wide hierarchies...

Topics of This Lecture

- Hierarchical Dirichlet Processes
  - Recap
  - Chinese Restaurant Franchise
  - Gibbs sampling for HDPs
  - CRF Sampler
- Applications
  - Example: Document topic modeling
  - Latent Dirichlet Allocation (LDA)

Applications

- **Example: Document topic modelling**
  - Topic: probability distribution over a set of words
  - Model each document as a probability distribution over topics.

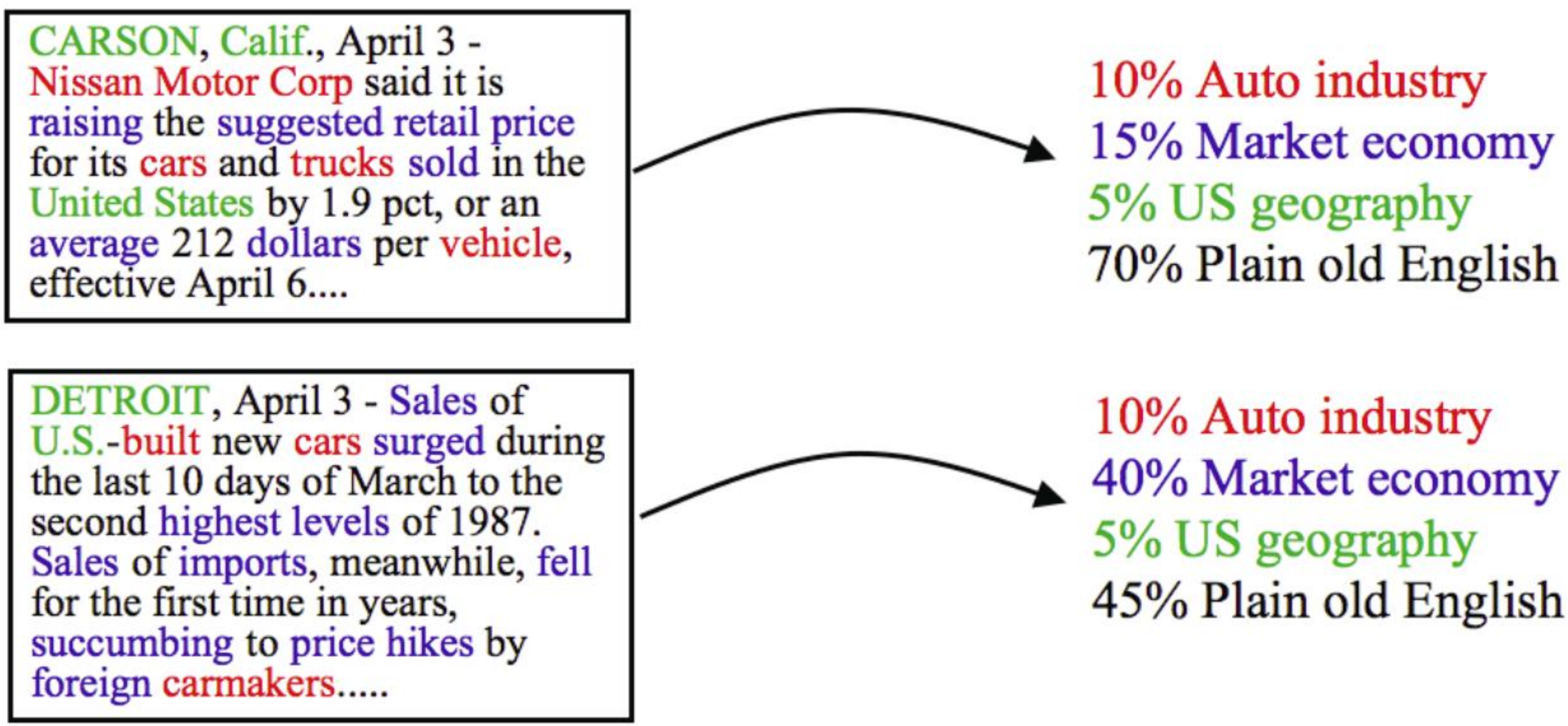
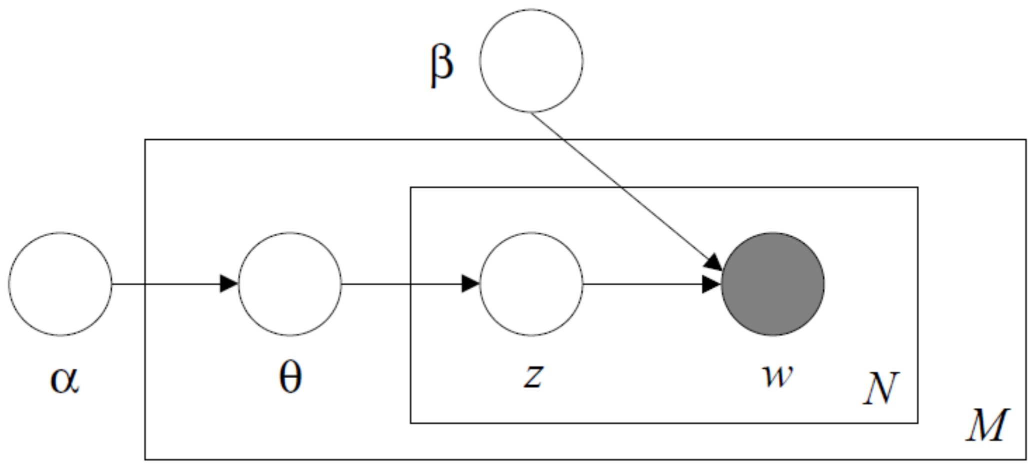


Image source: Yee Whye Teh

Applications

- **Latent Dirichlet Allocation** [Blei *et al.*, 2003]
  - Popular topic modelling approach with fixed number of topics k

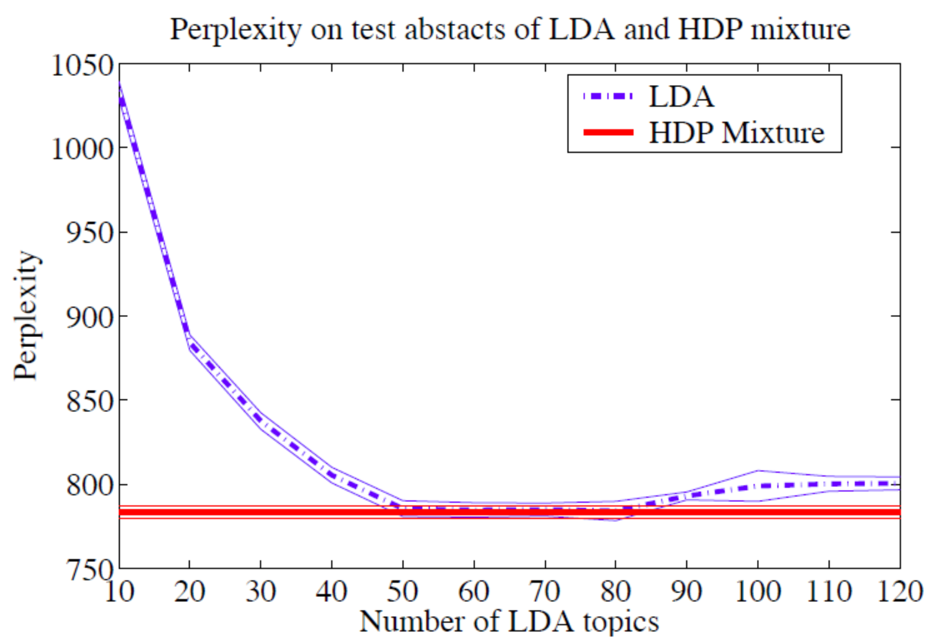


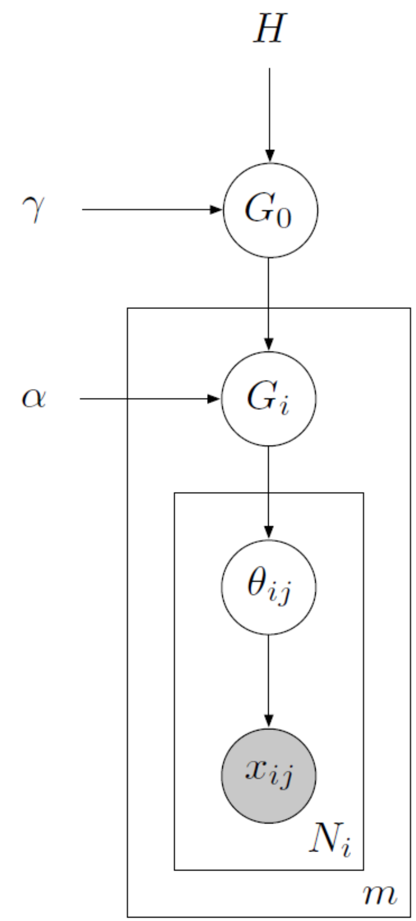
  - Random variables
    - A word is represented as a multinomial random variable w
    - A topic is represented as a multinomial random variable z
    - A document is represented as a Dirichlet random variable θ

Image source: Mike Jordan

Applications

- **HDPs can be used to define a BNP version of LDA**
  - **Number of topics is open-ended**
  - **Multiple infinite mixture models, linked via shared topic distribution.**





$\Rightarrow$ **HDP-LDA avoids the need for model selection.**

Image source: Mike Jordan

Applications

- **Model the evolution of topics over time**

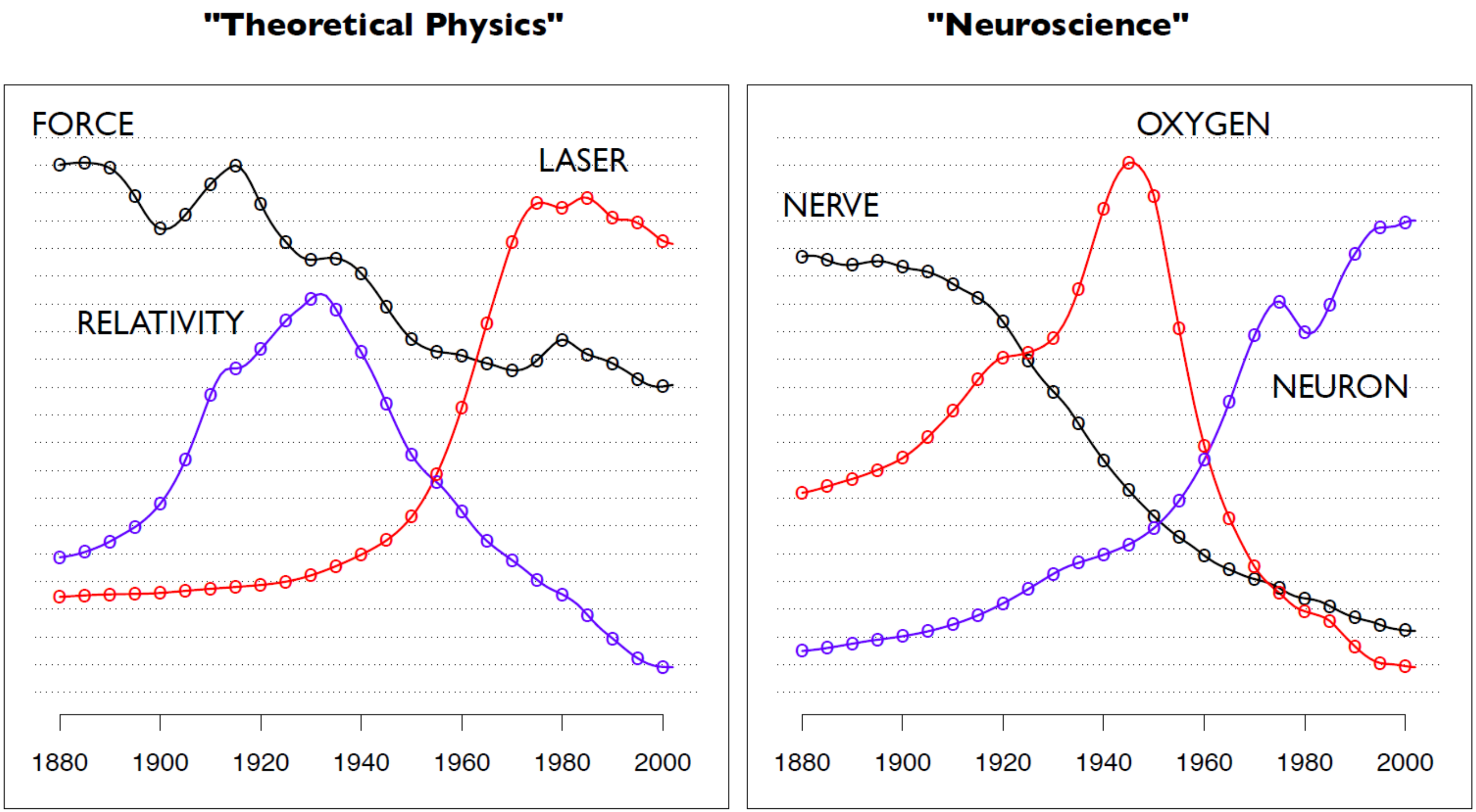


Image source: David Blei

Applications

- **Model connection between topics**

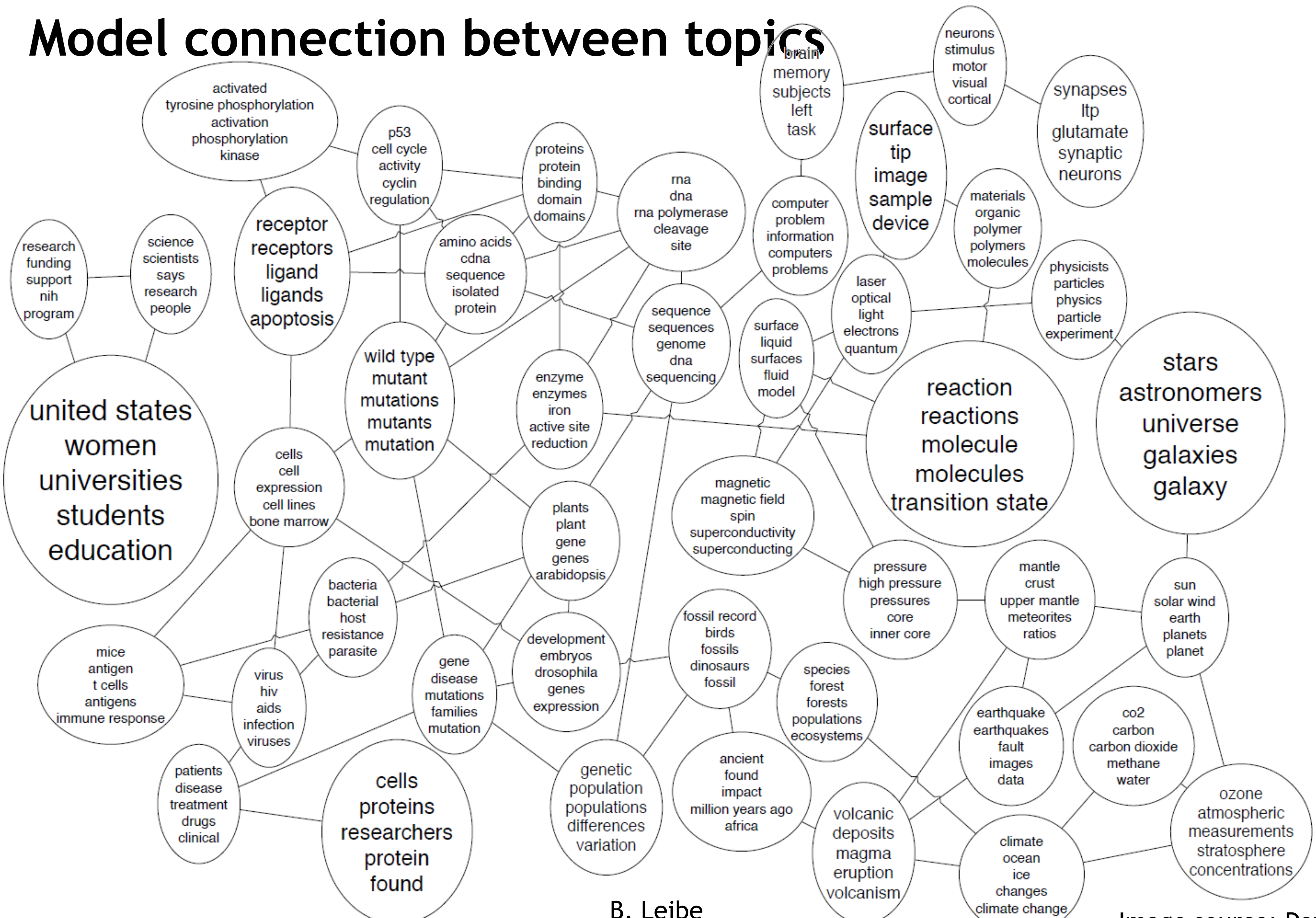


Image source: David Blei

Applications

- **Image auto-annotation**

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE

SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

Image source: David Blei

Applications

- There are many other generalizations I didn’t talk about
  - Dependent DPs
  - Nested DPs
  - Pitman-Yor Processes (2-parameter extension of DPs)
  - Infinite HMMs
  - ...

- And some that I will talk about in Lecture 16...
  - Infinite Latent Factor Models
  - Beta Processes
  - Indian Buffet Process
  - Hierarchical Beta Process

References and Further Reading

- Unfortunately, there are currently no good introductory textbooks on Dirichlet Processes. We will therefore post a number of tutorial papers on their different aspects.
    - One of the best available general introductions
        - E.B. Sudderth, "Graphical Models for Visual Object Recognition and Tracking", PhD thesis, Chapter 2, Section 2.5, 2006.
    - A tutorial on Hierarchical DPs
        - Y.W. Teh, M.I. Jordan, Hierarchical Bayesian Nonparametric Models with Applications. Bayesian Nonparametrics, Cambridge Univ. Press, 2010.
    - Good overview of MCMC methods for DPMMs
        - R. Neal, Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics, Vol. 9(2), p. 249-265, 2000.