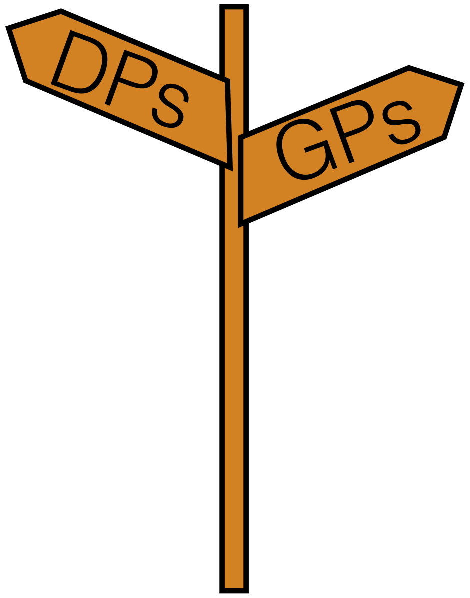
Nonparametric Bayesian Methods: Models, Algorithms, and Applications (Day 5)

Tamara Broderick
ITT Career Development Assistant Professor
Electrical Engineering & Computer Science
MIT

Roadmap

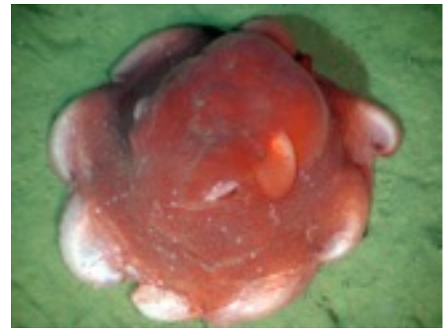


- Bayes Foundations
- Unsupervised Learning
  - Example problem: clustering
  - Example BNP model: Dirichlet process (DP)
  - Chinese restaurant process
- Supervised Learning
  - Example problem: regression
  - Example BNP model: Gaussian process (GP)
- Venture further into the wild world of Nonparametric Bayes

- Big questions
  - Why BNP?
  - What does an infinite/growing number of parameters really mean (in BNP)?
  - Why is BNP challenging but practical?

Applications

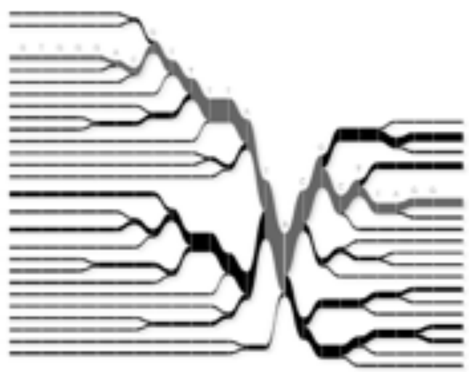
[wikipedia.org]



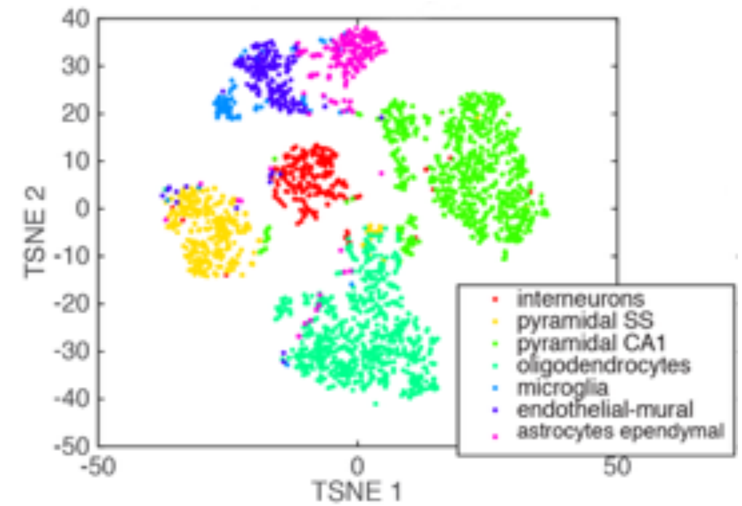
[Ed Bowlby, NOAA]

[Fox et al 2014]

[Lloyd et al 2012; Miller et al 2010]



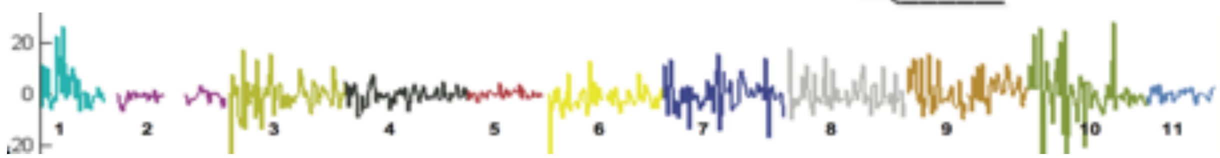
[Ewens 1972; Hartl, Clark 2003]



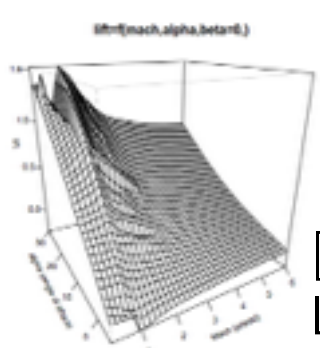
[Prabhakaran, Azizi, Carr, Pe'er 2016]

[Sudderth, Jordan 2009]

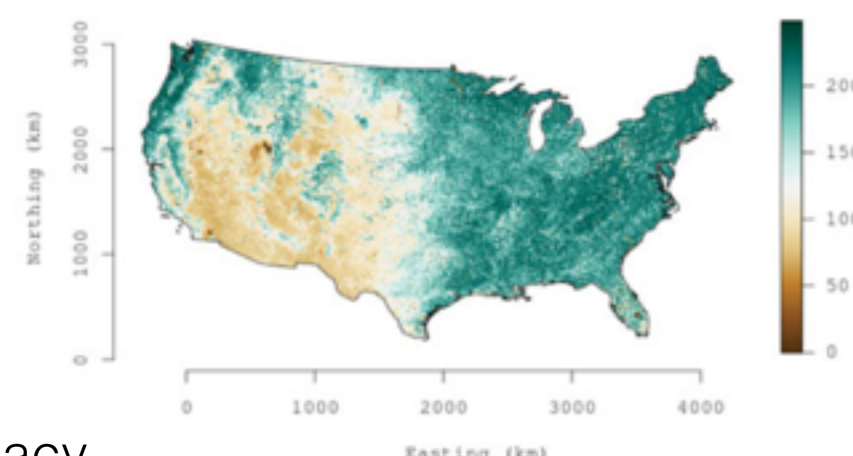
[Saria et al 2010]



[US CDC PHIL; Futoma, Hariharan, Heller 2017]

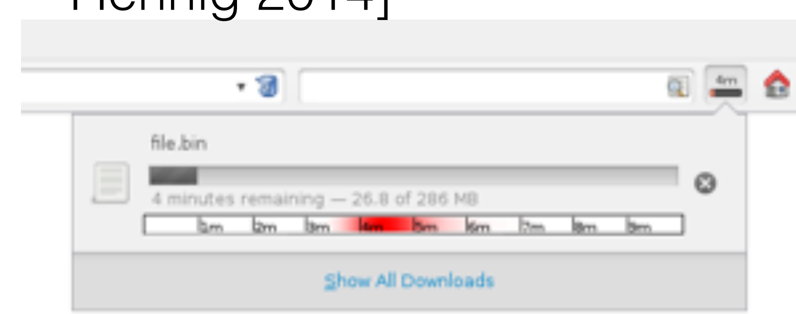


[Gramacy, Lee 2009]



[Datta, Banerjee, Finley, Gelfand 2016]

[Kiefel, Schuler, Hennig 2014]



[Deisenroth, Fox, Rasmussen 2015]

[Chati, Balakrishnan 2017]

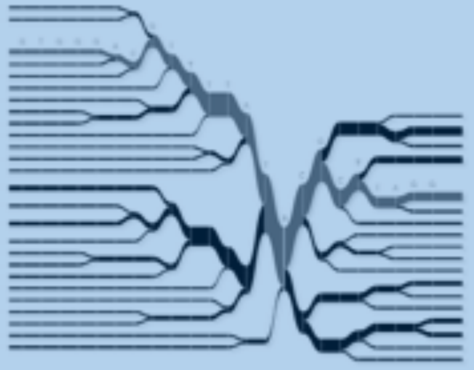
Applications

[wikipedia.org]

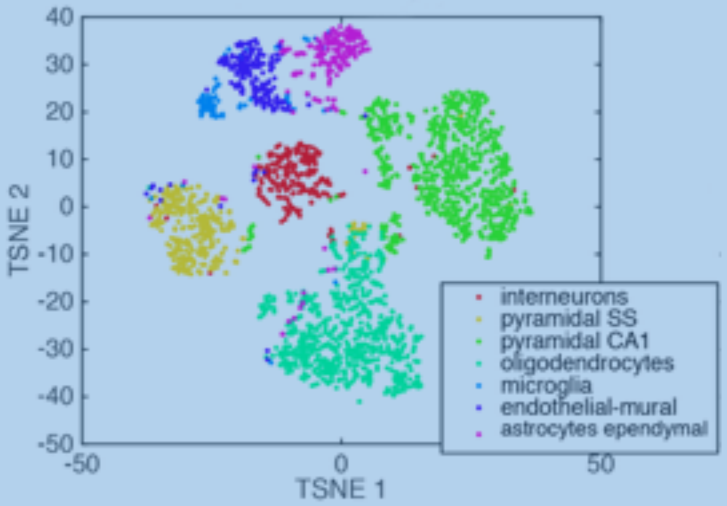
[Ed Bowlby, NOAA]

[Fox et al 2014]

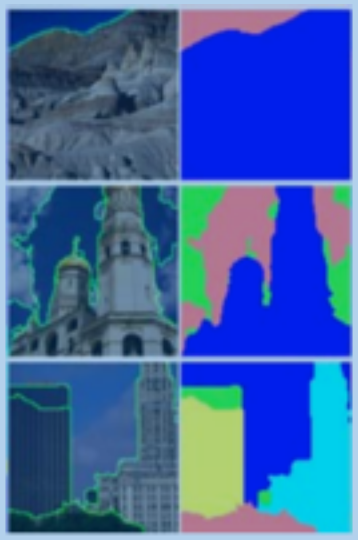
[Lloyd et al 2012; Miller et al 2010]



[Ewens 1972; Hartl, Clark 2003]

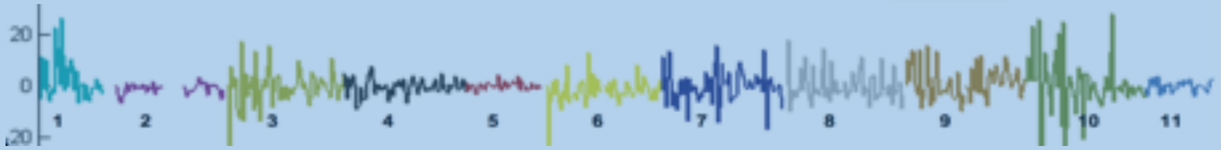


[Prabhakaran, Azizi, Carr, Pe'er 2016]

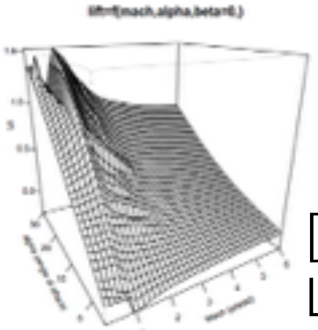


[Sudderth, Jordan 2009]

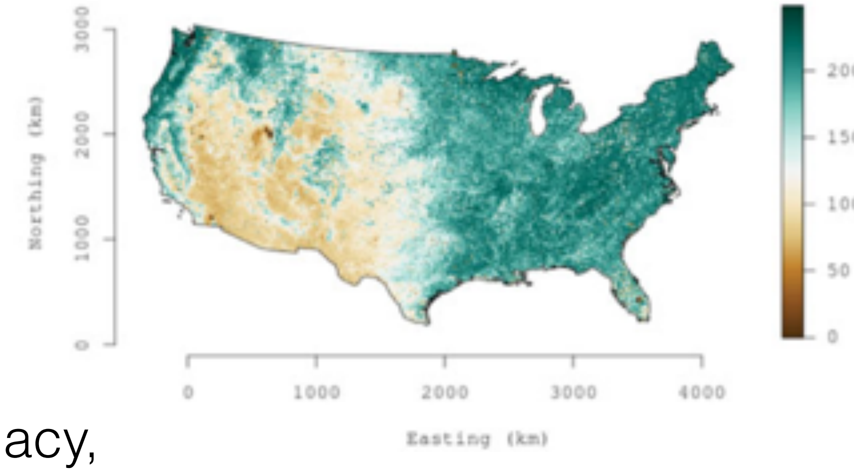
[Saria et al 2010]



[US CDC PHIL; Futoma, Hariharan, Heller 2017]

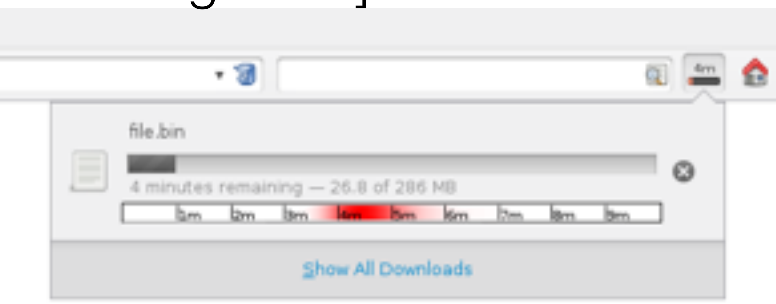


[Gramacy, Lee 2009]



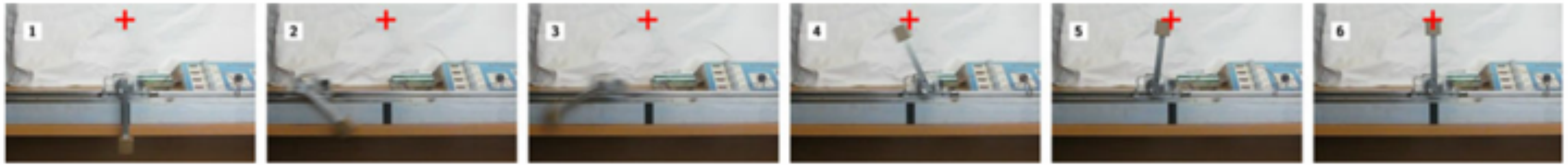
[Datta, Banerjee, Finley, Gelfand 2016]

[Kiefel, Schuler, Hennig 2014]



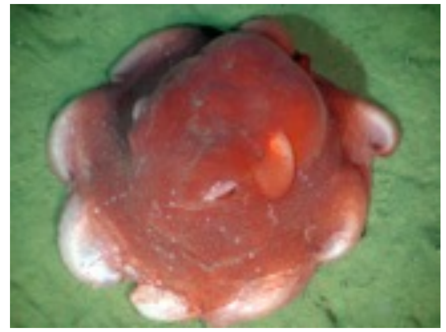
[Chati, Balakrishnan 2017]

[Deisenroth, Fox, Rasmussen 2015]



Applications

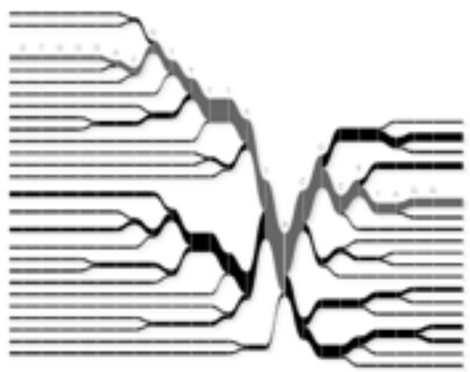
[wikipedia.org]



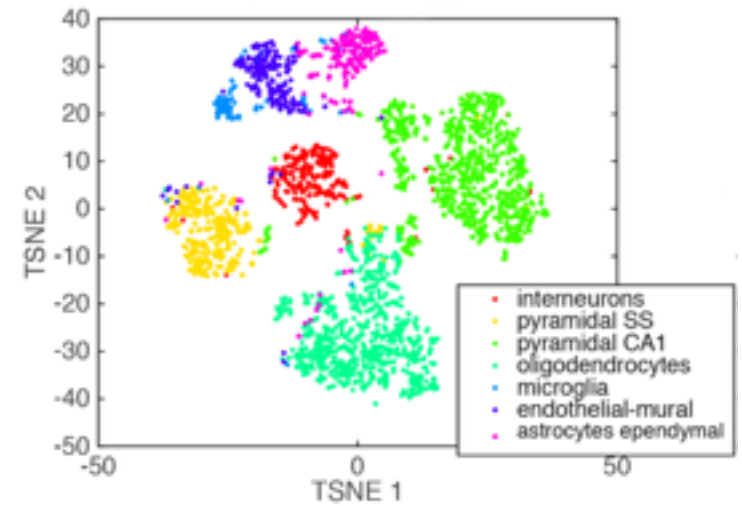
[Ed Bowlby, NOAA]

[Fox et al 2014]

[Lloyd et al 2012; Miller et al 2010]



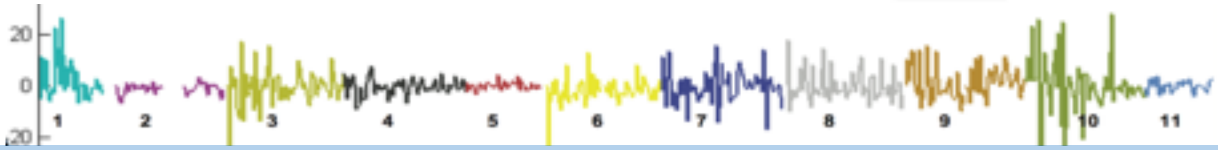
[Ewens 1972; Hartl, Clark 2003]



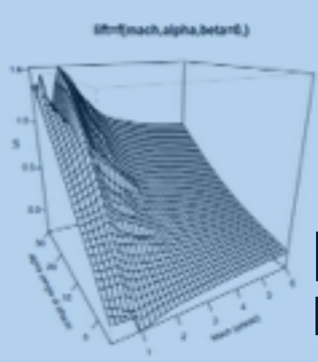
[Prabhakaran, Azizi, Carr, Pe'er 2016]

[Sudderth, Jordan 2009]

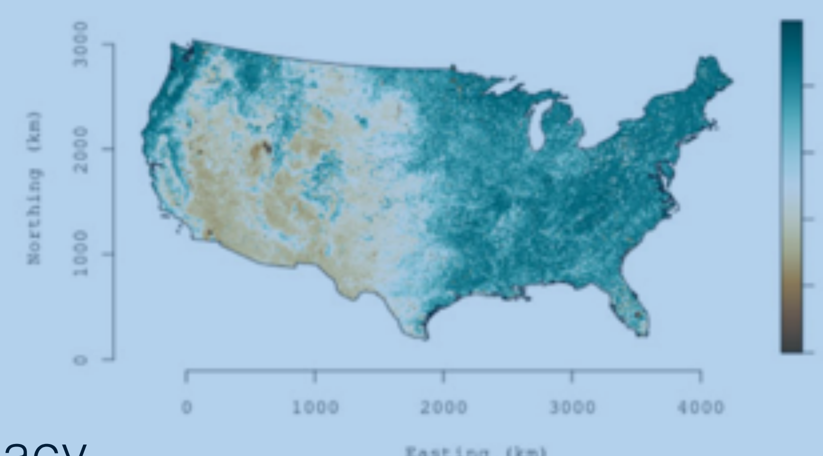
[Saria et al 2010]



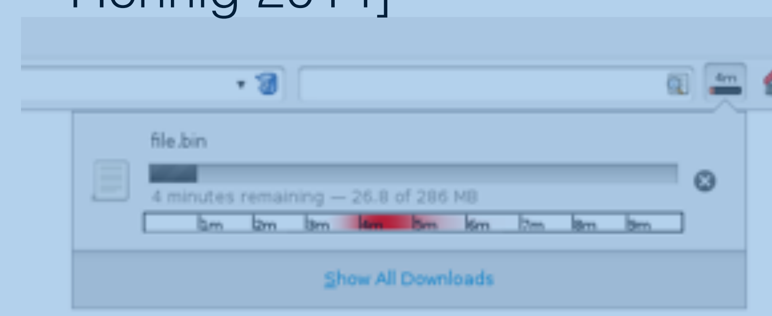
[US CDC PHIL; Futoma, Hariharan, Heller 2017]



[Gramacy, Lee 2009]

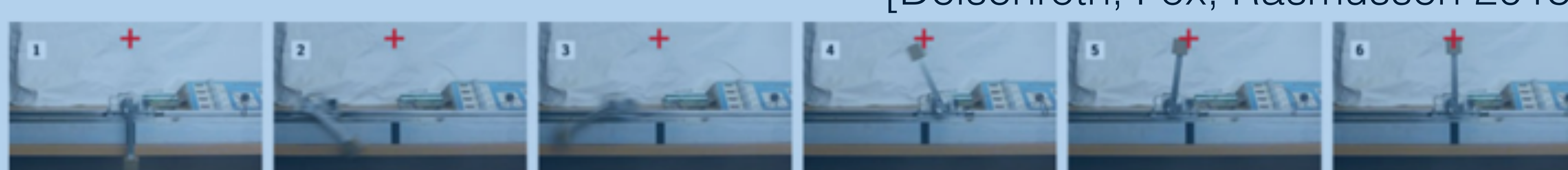


[Datta, Banerjee, Finley, Gelfand 2016]



[Kiefel, Schuler, Hennig 2014]

[Chati, Balakrishnan 2017]



[Deisenroth, Fox, Rasmussen 2015]

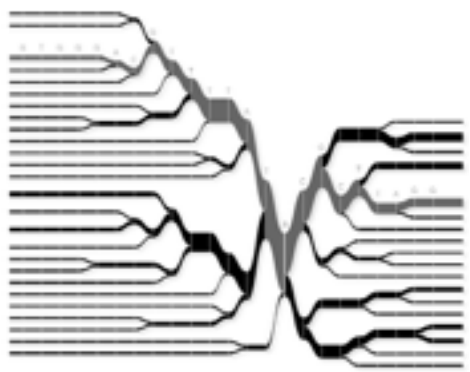
Applications

[wikipedia.org]

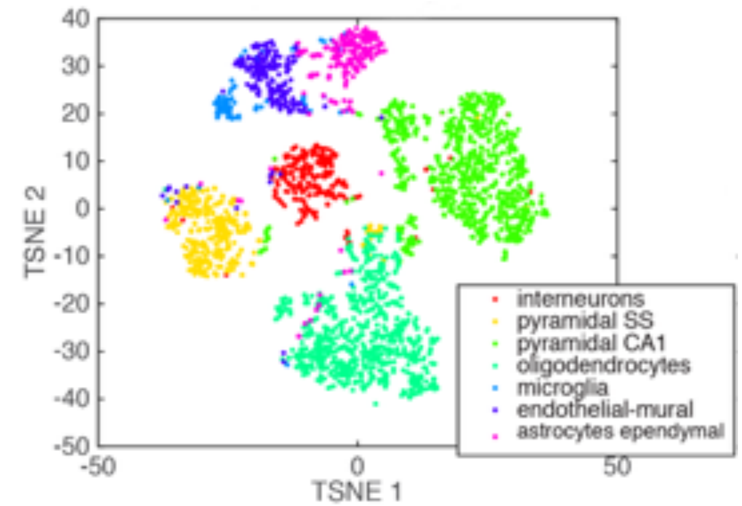
[Ed Bowlby, NOAA]

[Fox et al 2014]

[Lloyd et al 2012; Miller et al 2010]



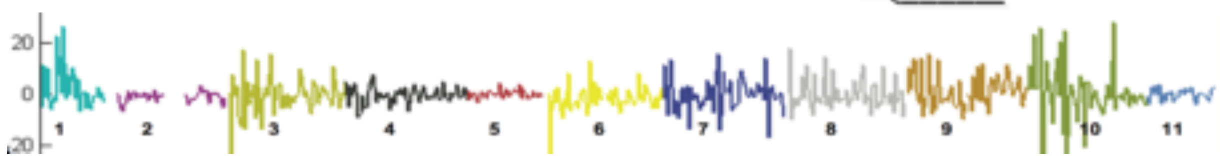
[Ewens 1972; Hartl, Clark 2003]



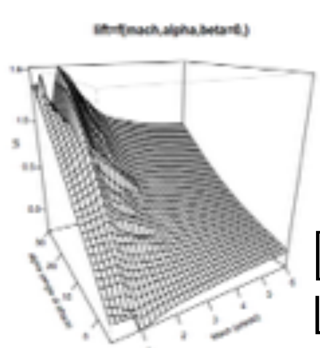
[Prabhakaran, Azizi, Carr, Pe'er 2016]

[Sudderth, Jordan 2009]

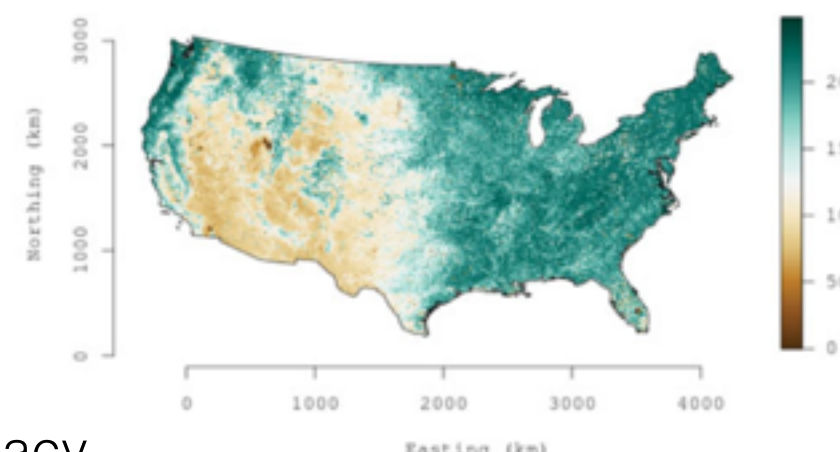
[Saria et al 2010]



[US CDC PHIL; Futoma, Hariharan, Heller 2017]

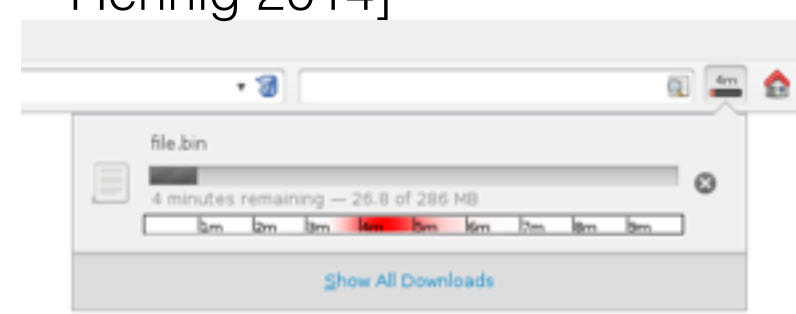


[Gramacy, Lee 2009]



[Datta, Banerjee, Finley, Gelfand 2016]

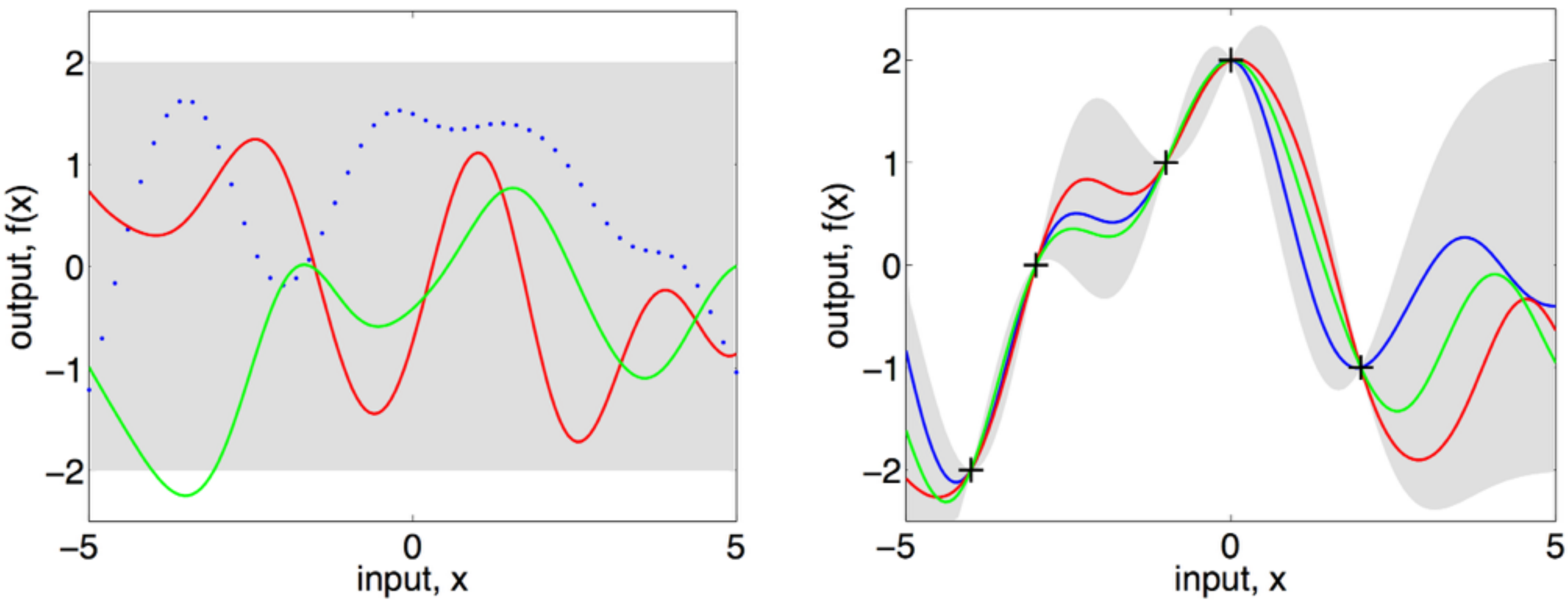
[Kiefel, Schuler, Hennig 2014]



[Chati, Balakrishnan 2017]

[Deisenroth, Fox, Rasmussen 2015]

Regression



(a), prior (b), posterior

C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X. © 2006 Massachusetts Institute of Technology. www.GaussianProcess.org/gpml

[Tiia Monto, https://commons.wikimedia.org/wiki/File:Trail_and_mountain.jpg]

More Markov Chain Monte Carlo

More Markov Chain Monte Carlo

- Slice sampling

More Markov Chain Monte Carlo

- Slice sampling

[Kalli, Griffin, Walker 2011; Broderick, Mackey, Paisley, Jordan 2015]

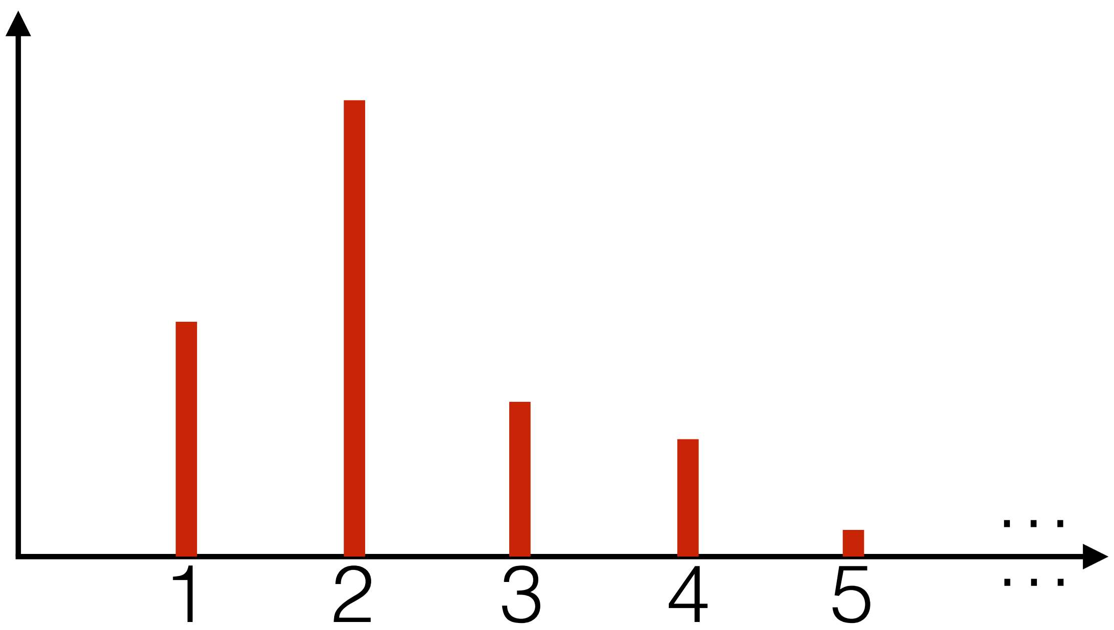
More Markov Chain Monte Carlo

- Slice sampling
  - auxiliary variable ➔ finite conditionals

[Kalli, Griffin, Walker 2011; Broderick, Mackey, Paisley, Jordan 2015]

More Markov Chain Monte Carlo

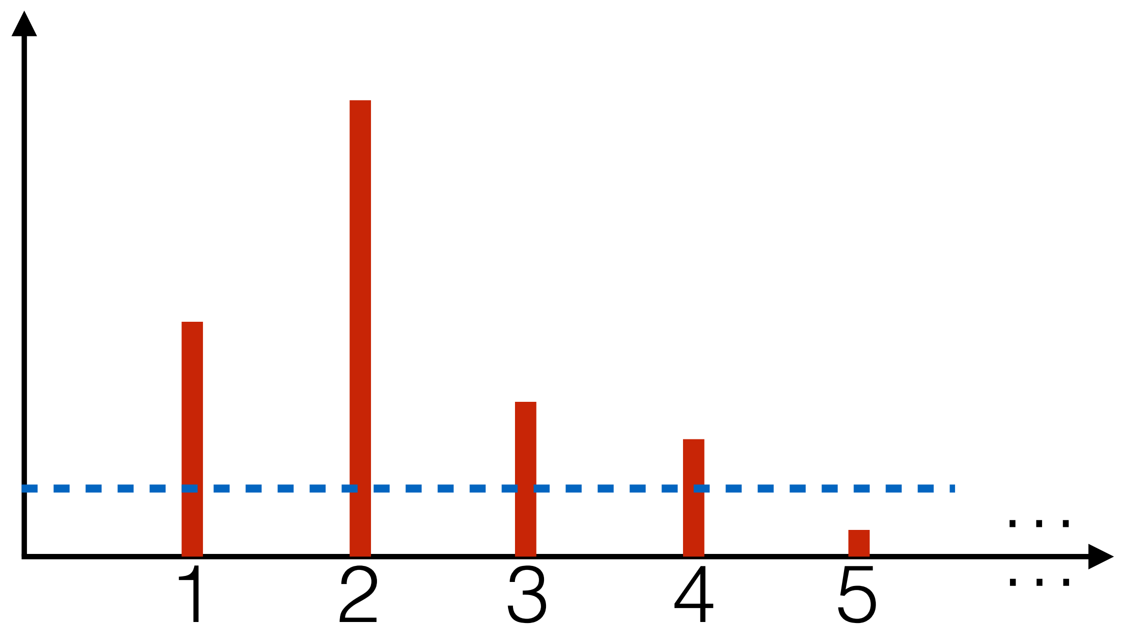
- Slice sampling
  - auxiliary variable ➔ finite conditionals



[Kalli, Griffin, Walker 2011; Broderick, Mackey, Paisley, Jordan 2015]

More Markov Chain Monte Carlo

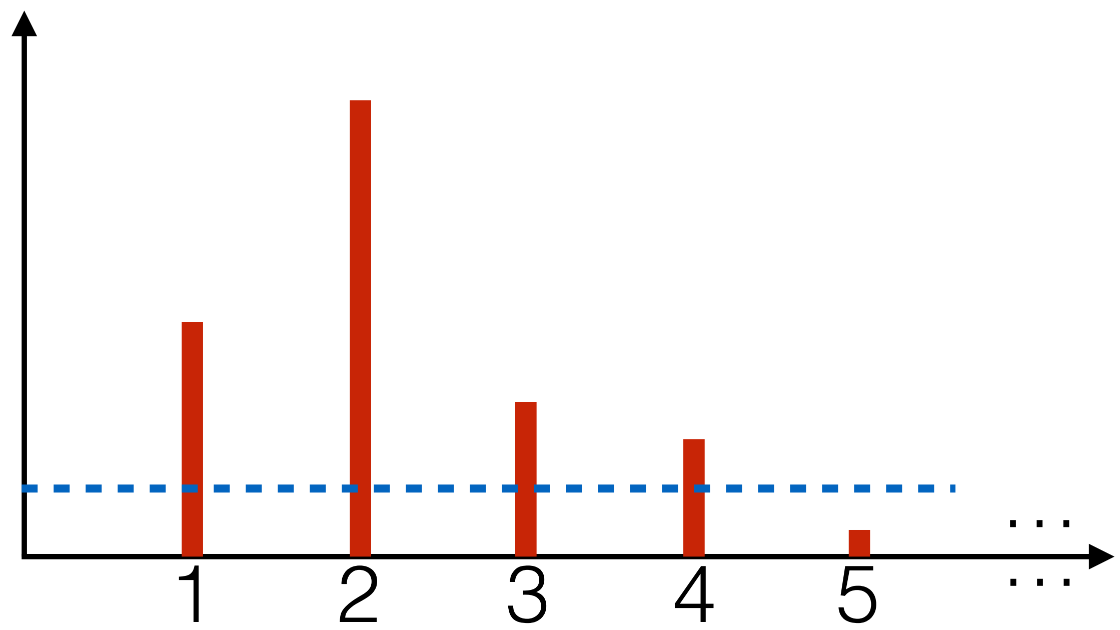
- Slice sampling
  - auxiliary variable ➔ finite conditionals



[Kalli, Griffin, Walker 2011; Broderick, Mackey, Paisley, Jordan 2015]

More Markov Chain Monte Carlo

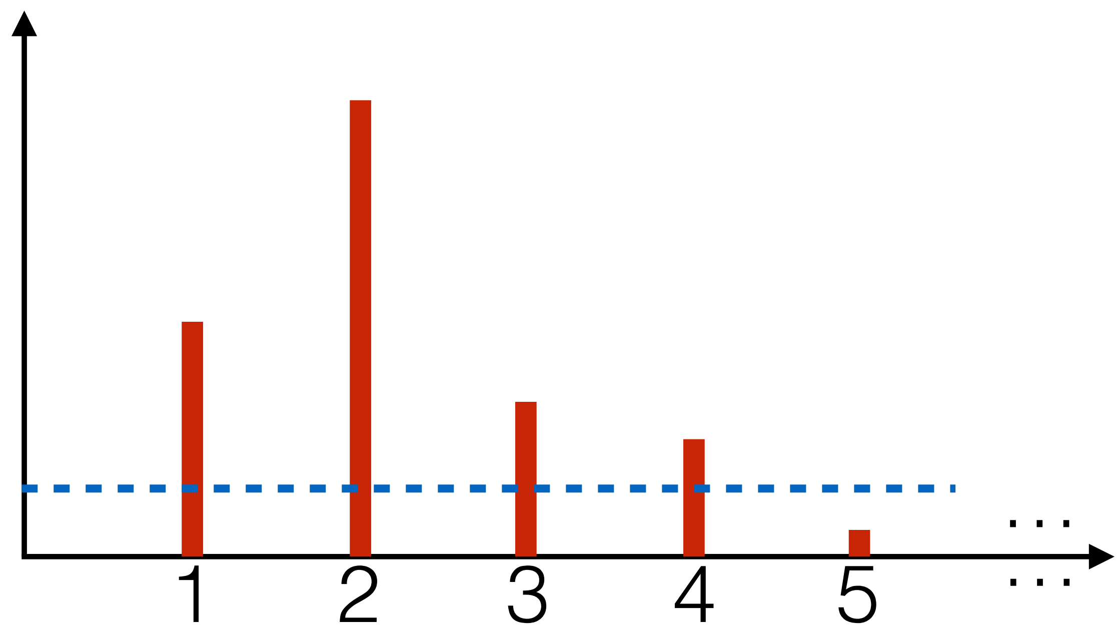
- Slice sampling
  - auxiliary variable ➔ finite conditionals



- Approximate with truncated distribution

More Markov Chain Monte Carlo

- Slice sampling
  - auxiliary variable ➔ finite conditionals

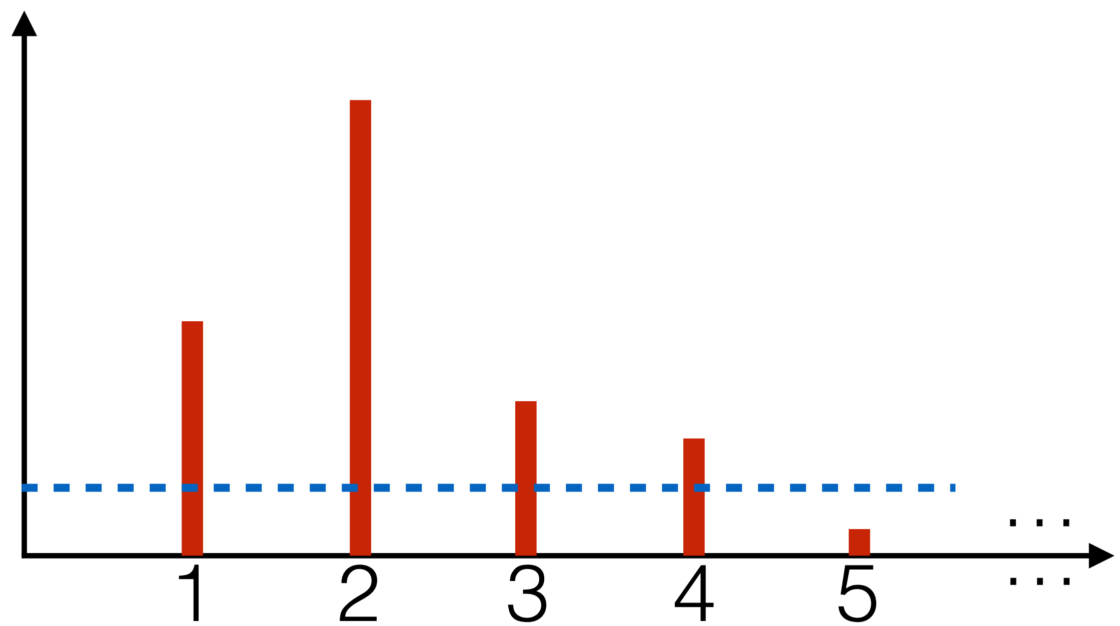


- Approximate with truncated distribution

[Ishwaran, James 2001; Campbell*, Huggins*, Broderick 2016]

More Markov Chain Monte Carlo

- Slice sampling
    - auxiliary variable ➔ finite conditionals

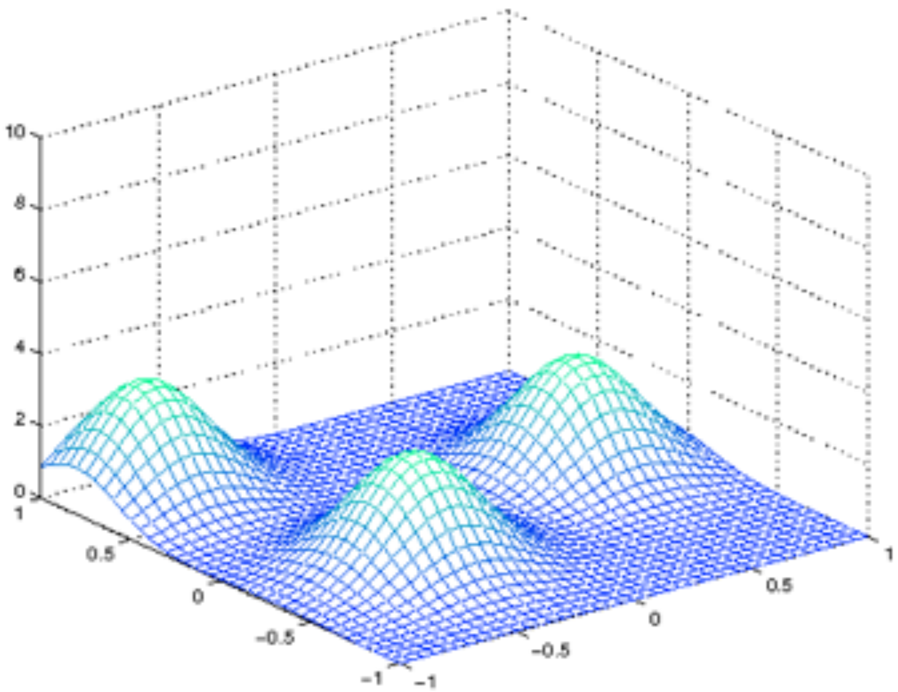


- Approximate with truncated distribution
    - E.g., Hamiltonian Monte Carlo

[Ishwaran, James 2001; Campbell*, Huggins*, Broderick 2016]

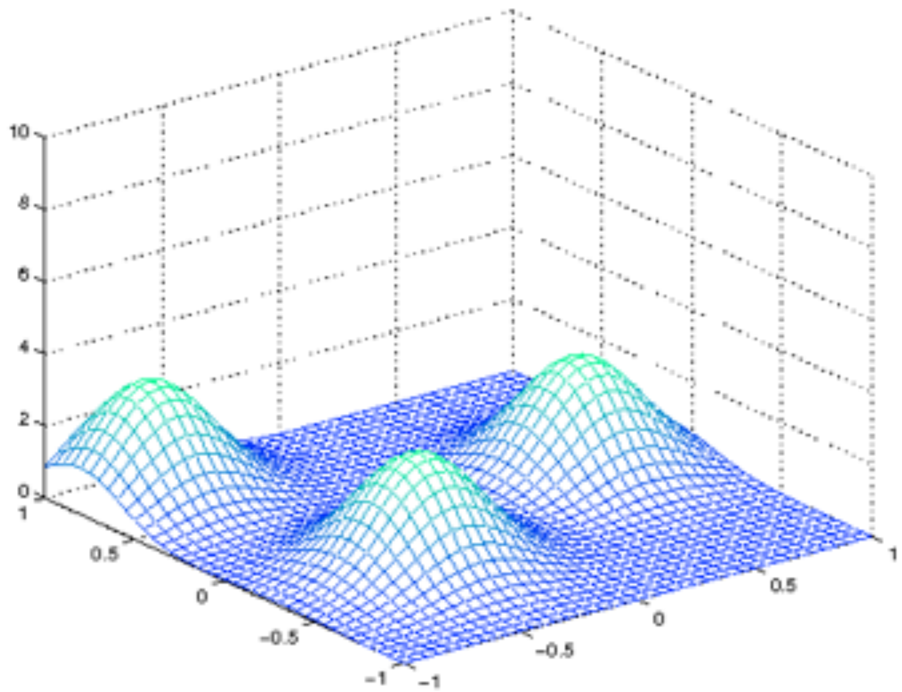
Variational Bayes

Variational Bayes



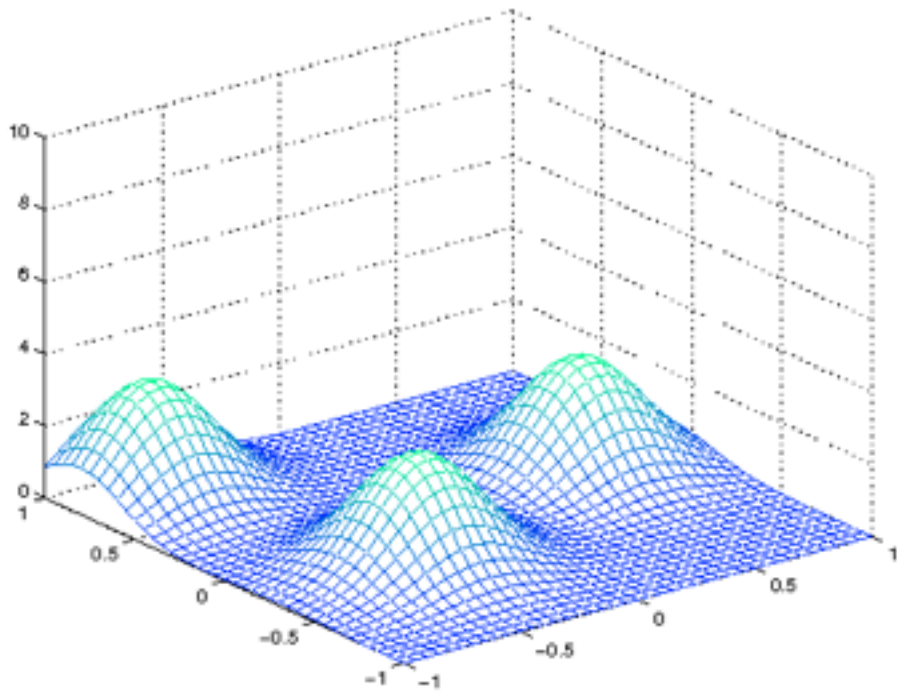
- Variational Bayes (VB)

Variational Bayes

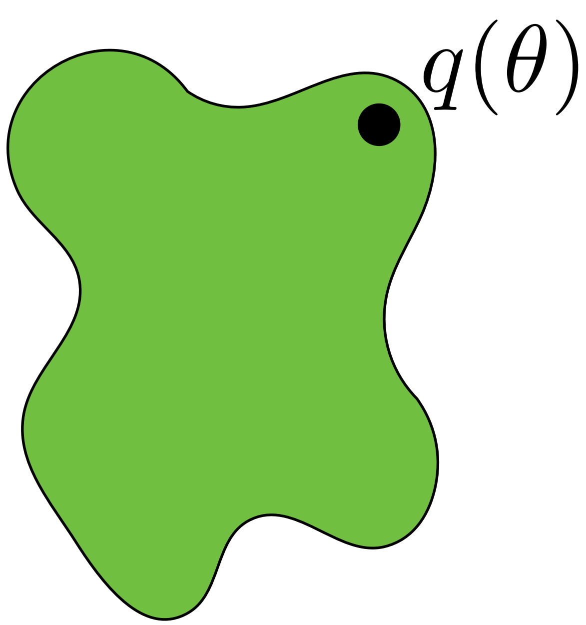


- Variational Bayes (VB)
    - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

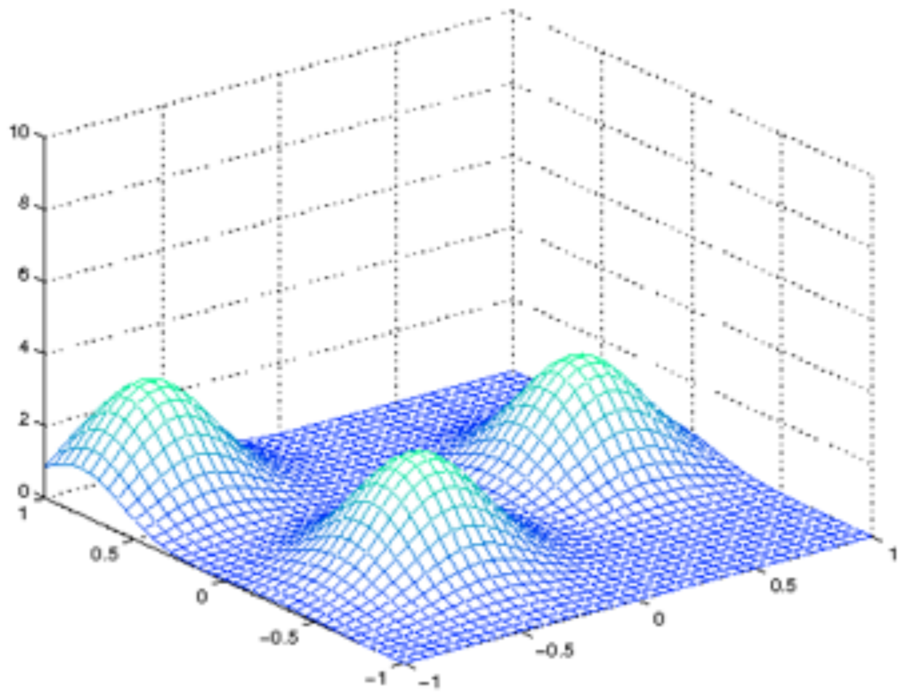
Variational Bayes

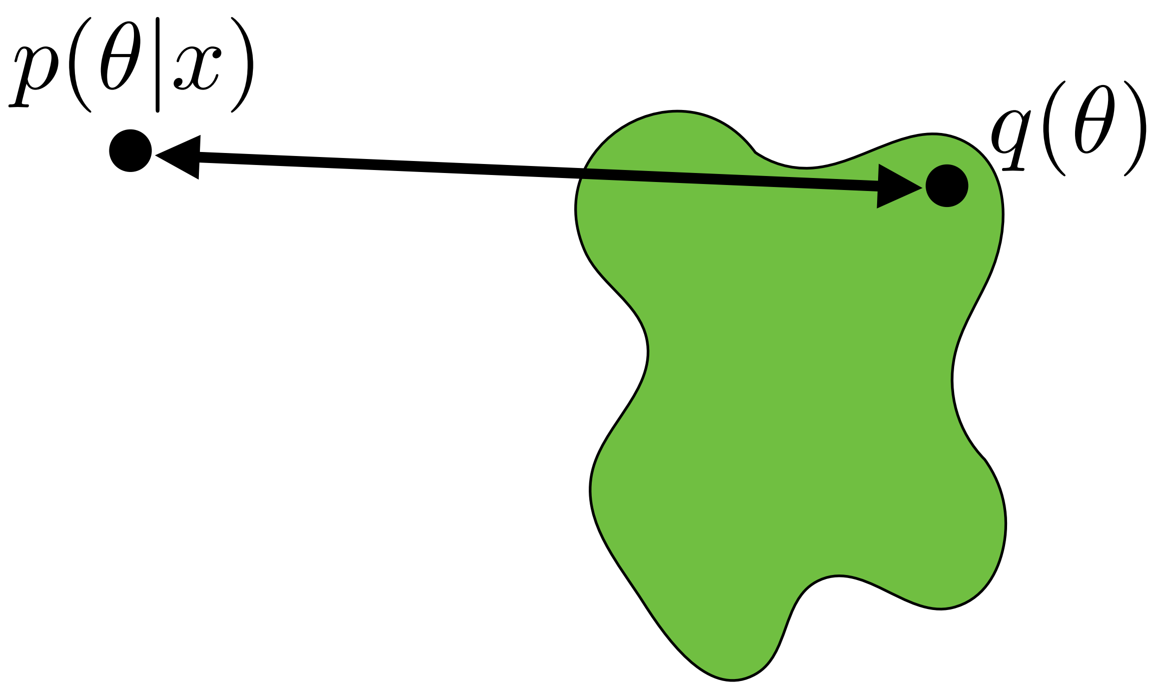


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$



Variational Bayes

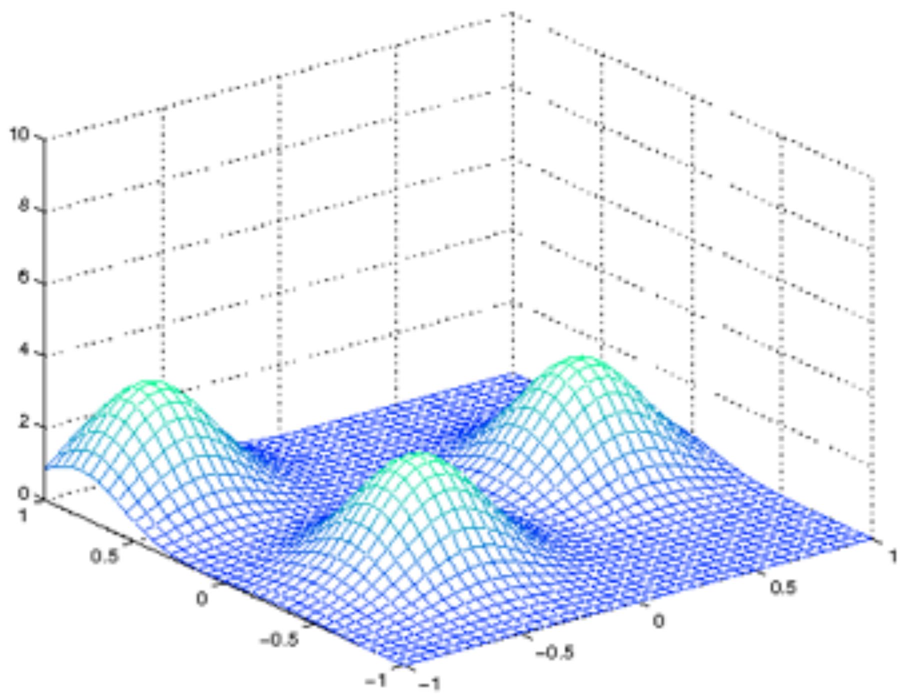


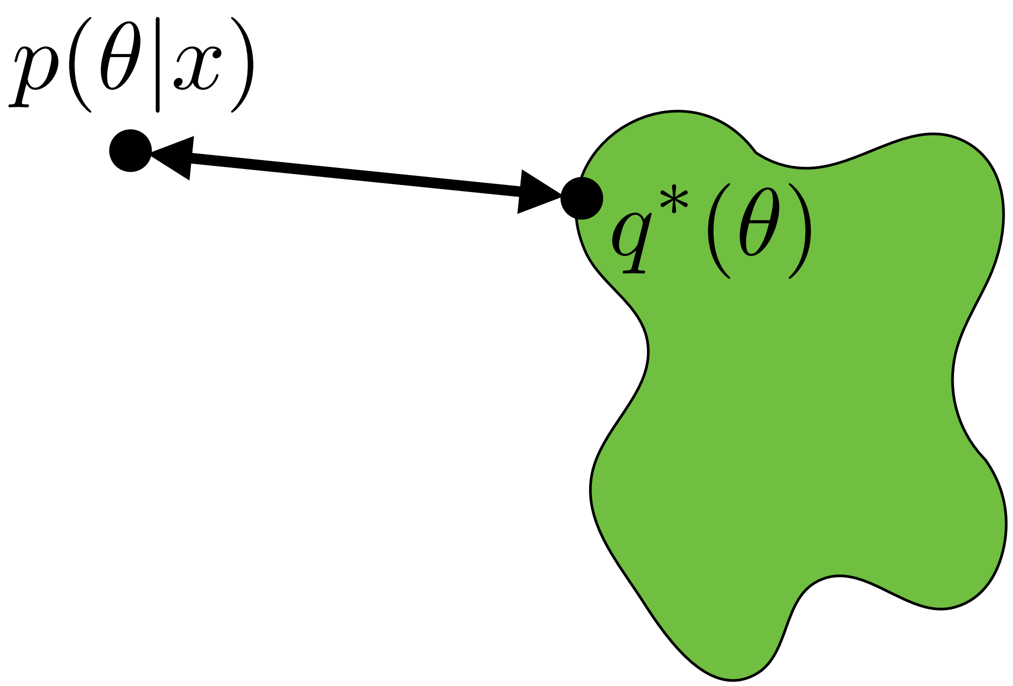


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

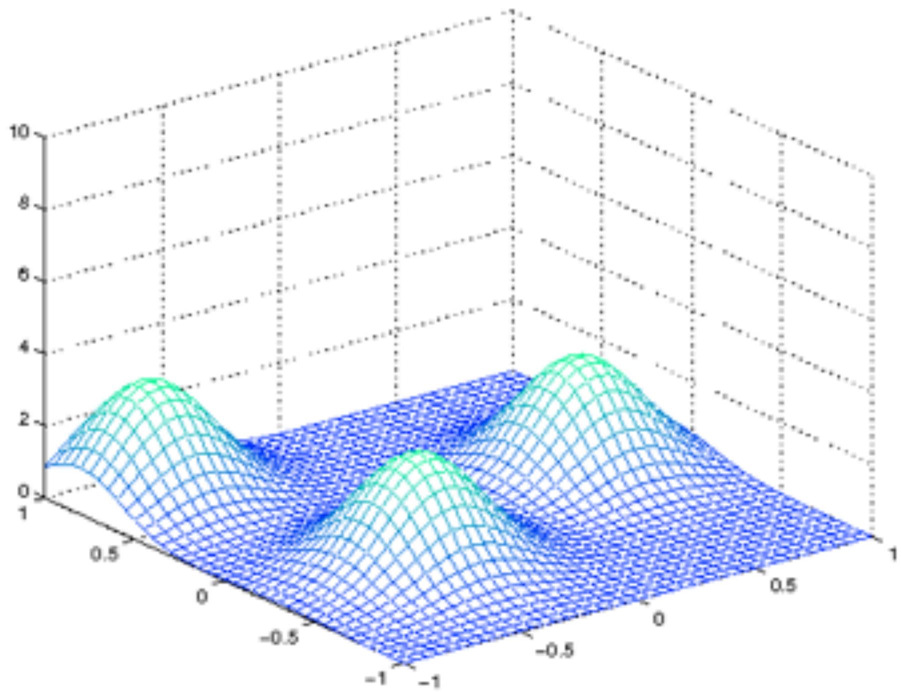
Variational Bayes

- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$

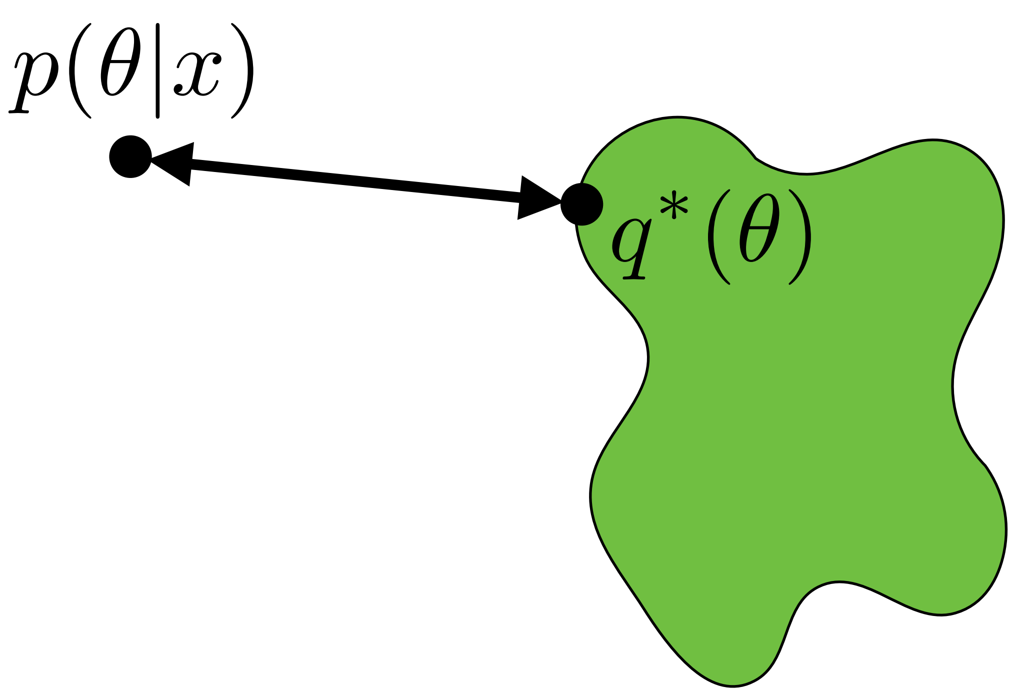




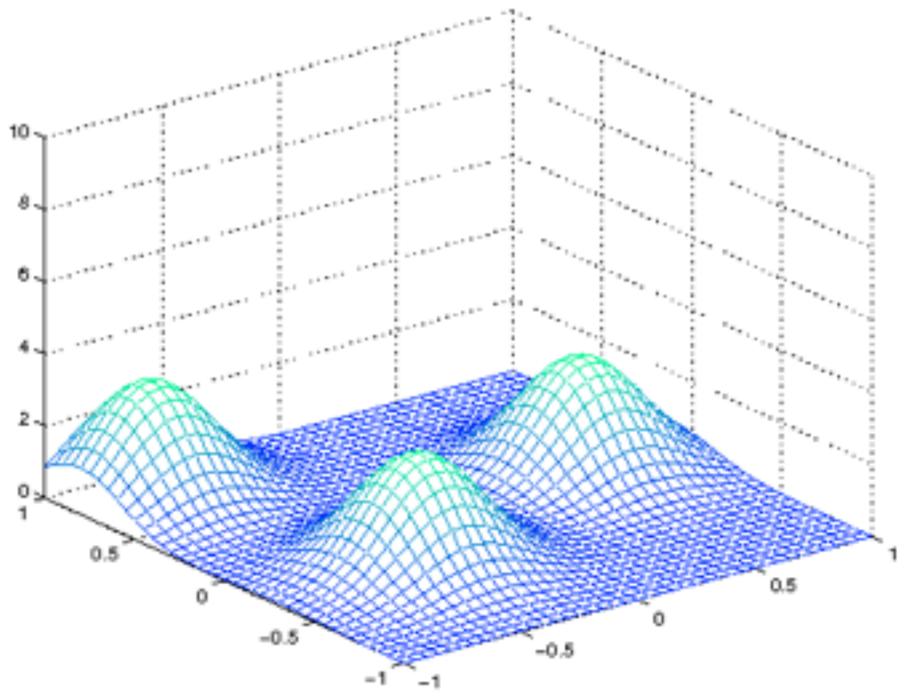
Variational Bayes

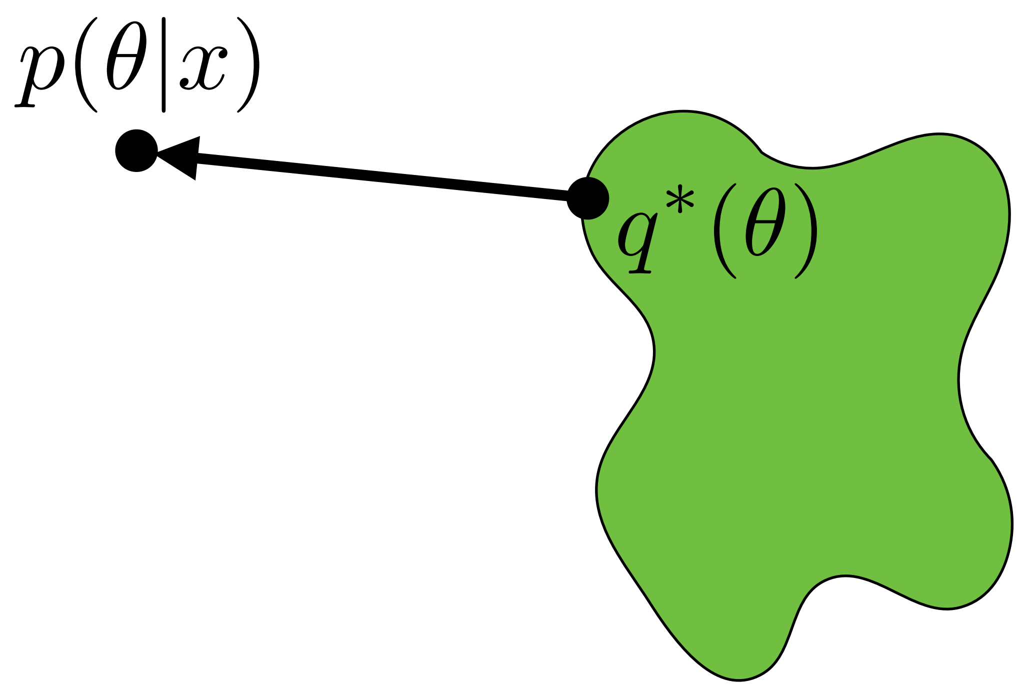


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - "Close": Minimize Kullback-Liebler (KL) divergence:
    $$KL(q\|p(\cdot|x))$$



Variational Bayes

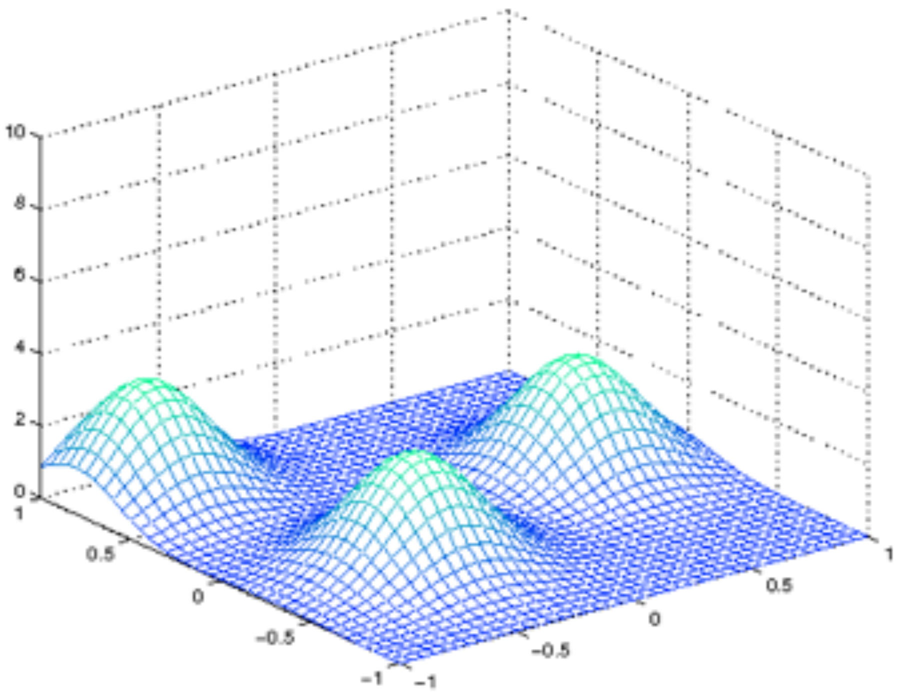


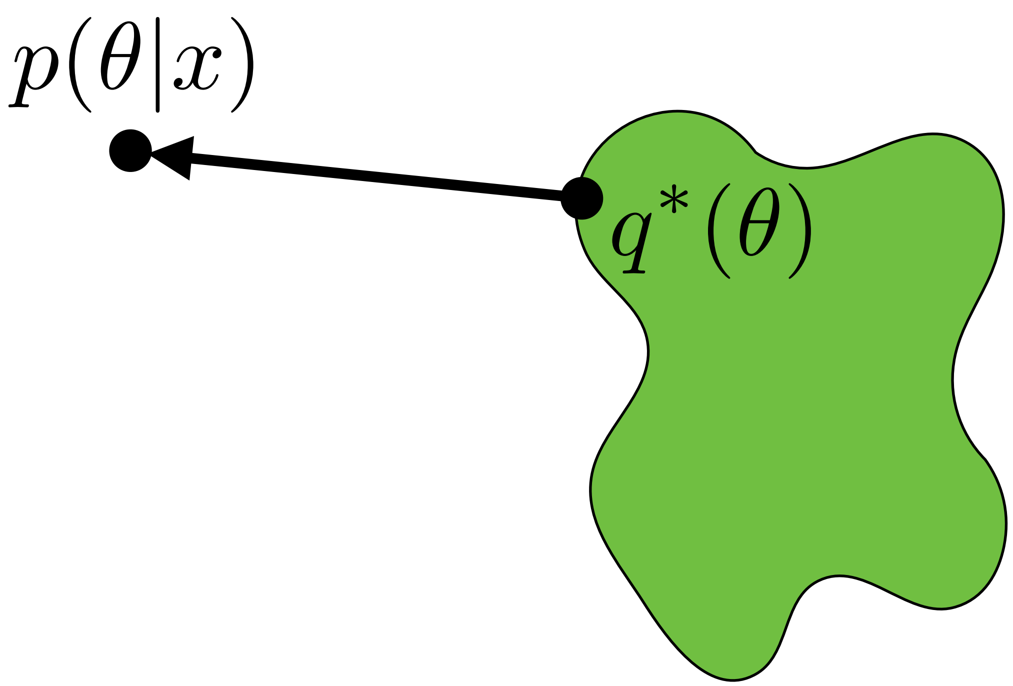


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - "Close": Minimize Kullback-Liebler (KL) divergence:

$$KL(q\|p(\cdot|x))$$

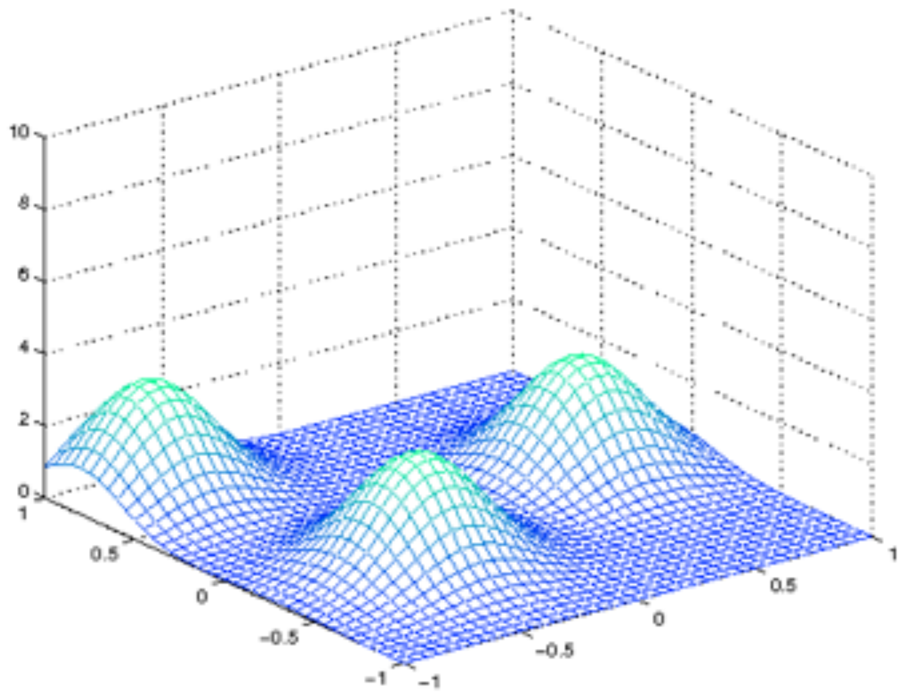
Variational Bayes

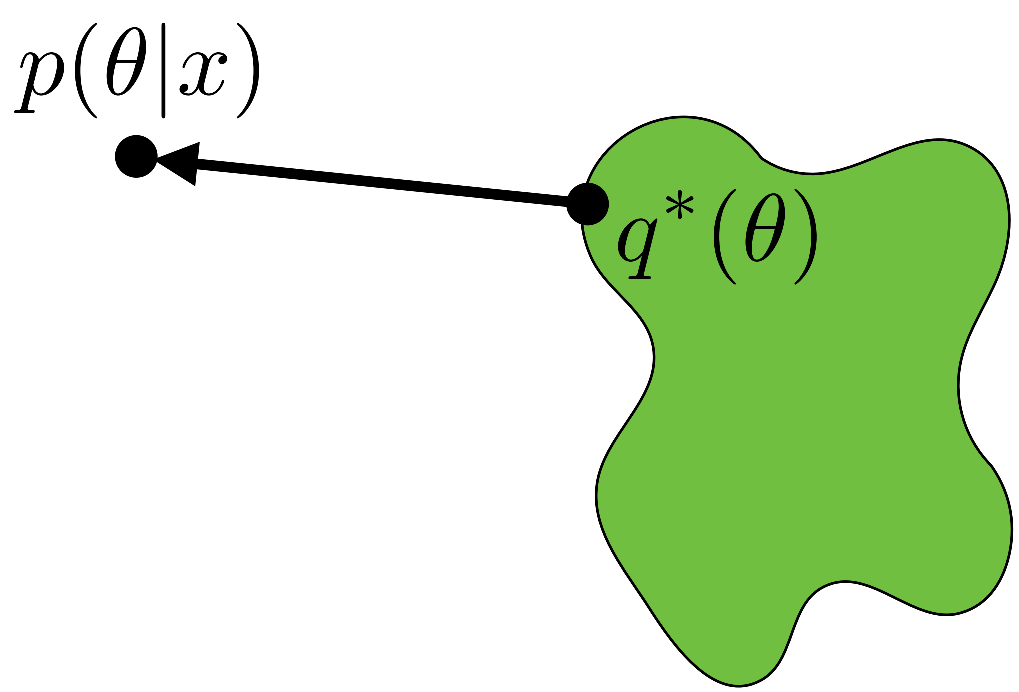




- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - "Close": Minimize Kullback-Liebler (KL) divergence:
    $$KL(q\|p(\cdot|x))$$
  - "Nice": factorizes, exponential family, truncation

Variational Bayes

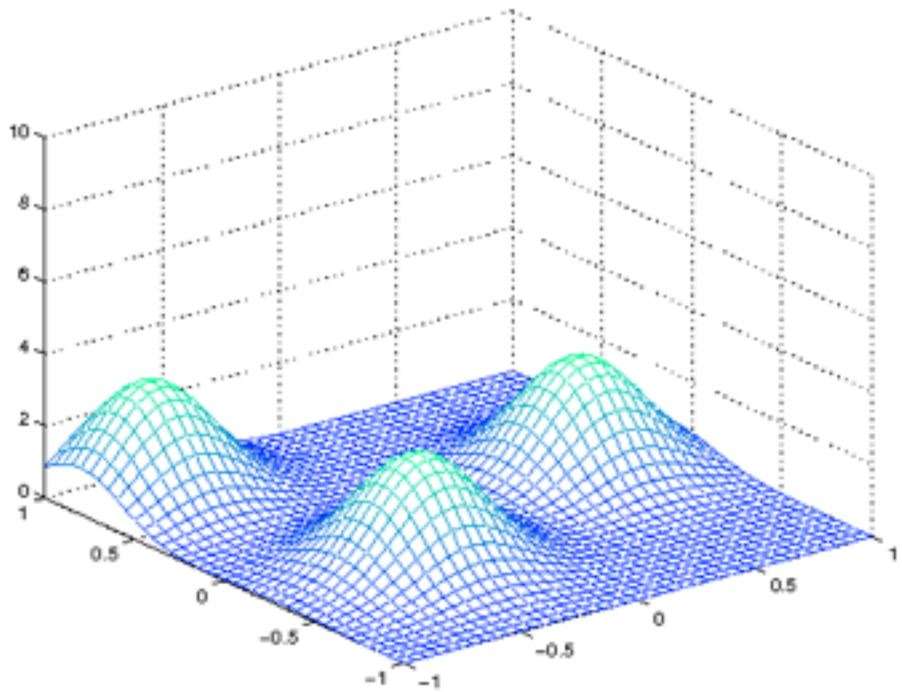


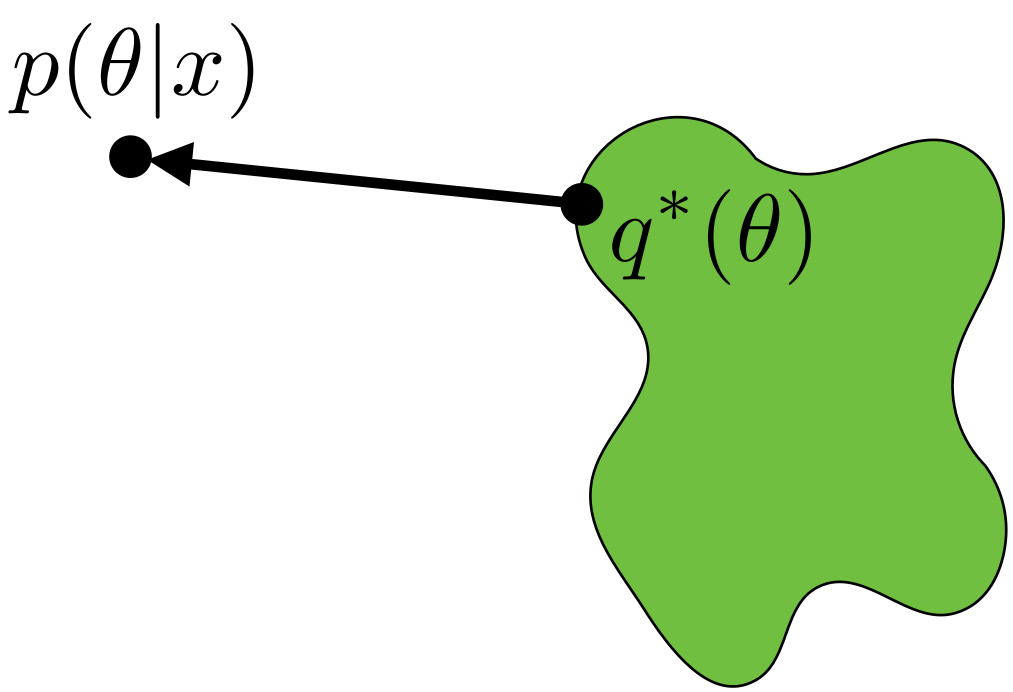


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - "Close": Minimize Kullback-Liebler (KL) divergence:
  $$KL(q\|p(\cdot|x))$$
  - "Nice": factorizes, exponential family, truncation

- VB practical success

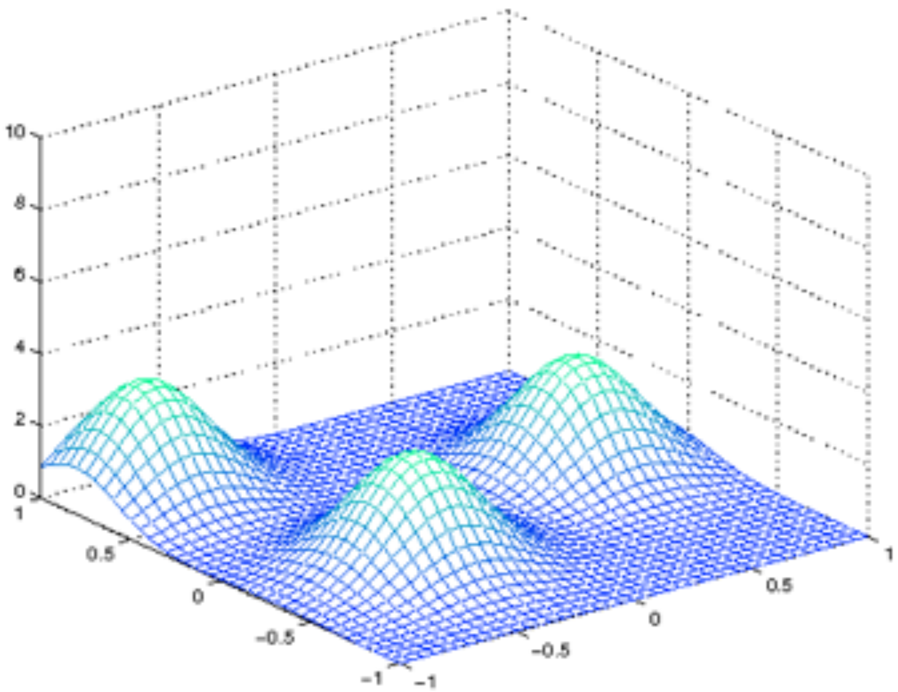
Variational Bayes

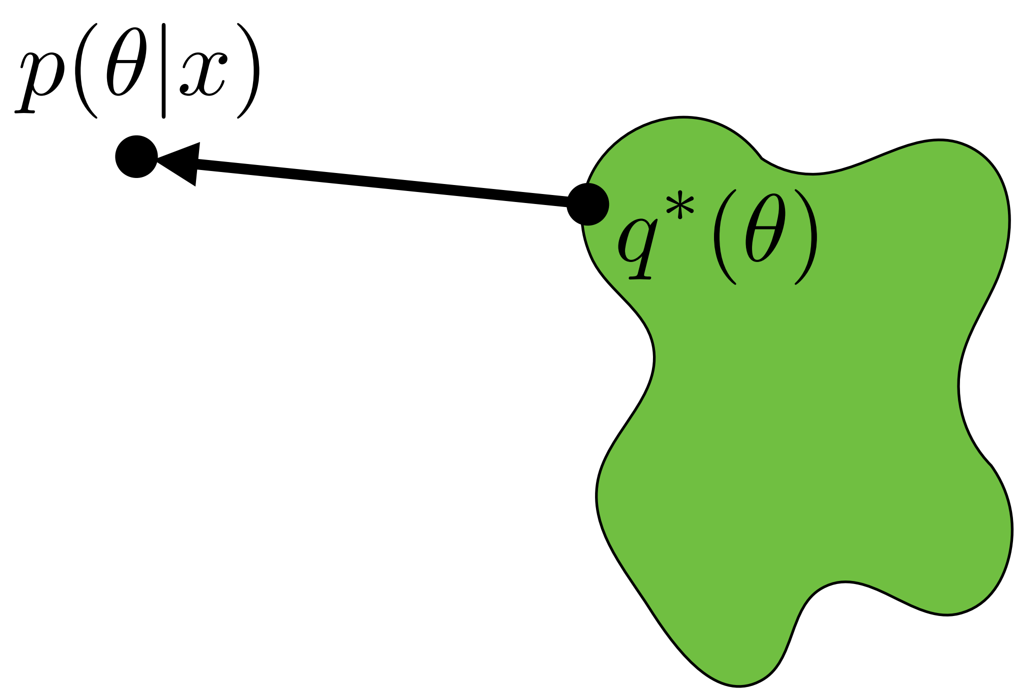




- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - "Close": Minimize Kullback-Liebler (KL) divergence:
    $$KL(q\|p(\cdot|x))$$
  - "Nice": factorizes, exponential family, truncation
- VB practical success
  - point estimates and prediction

Variational Bayes





- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - "Close": Minimize Kullback-Liebler (KL) divergence:

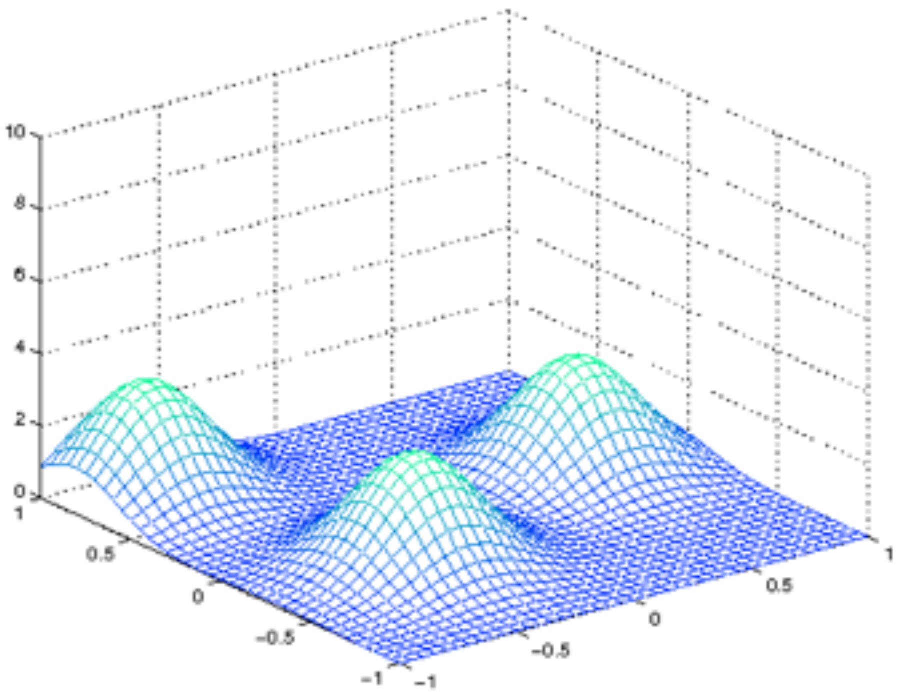
$$KL(q\|p(\cdot|x))$$

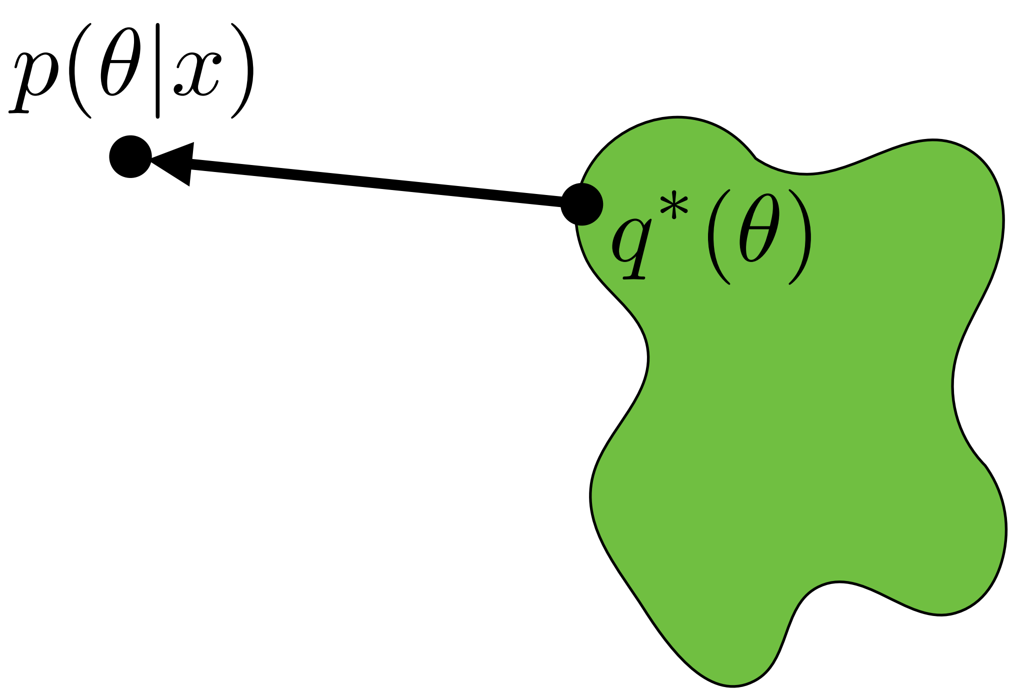
  - "Nice": factorizes, exponential family, truncation

- VB practical success
  - point estimates and prediction
  - fast, streaming, distributed

[Broderick, Boyd, Wibisono, Wilson, Jordan 2013]

Variational Bayes



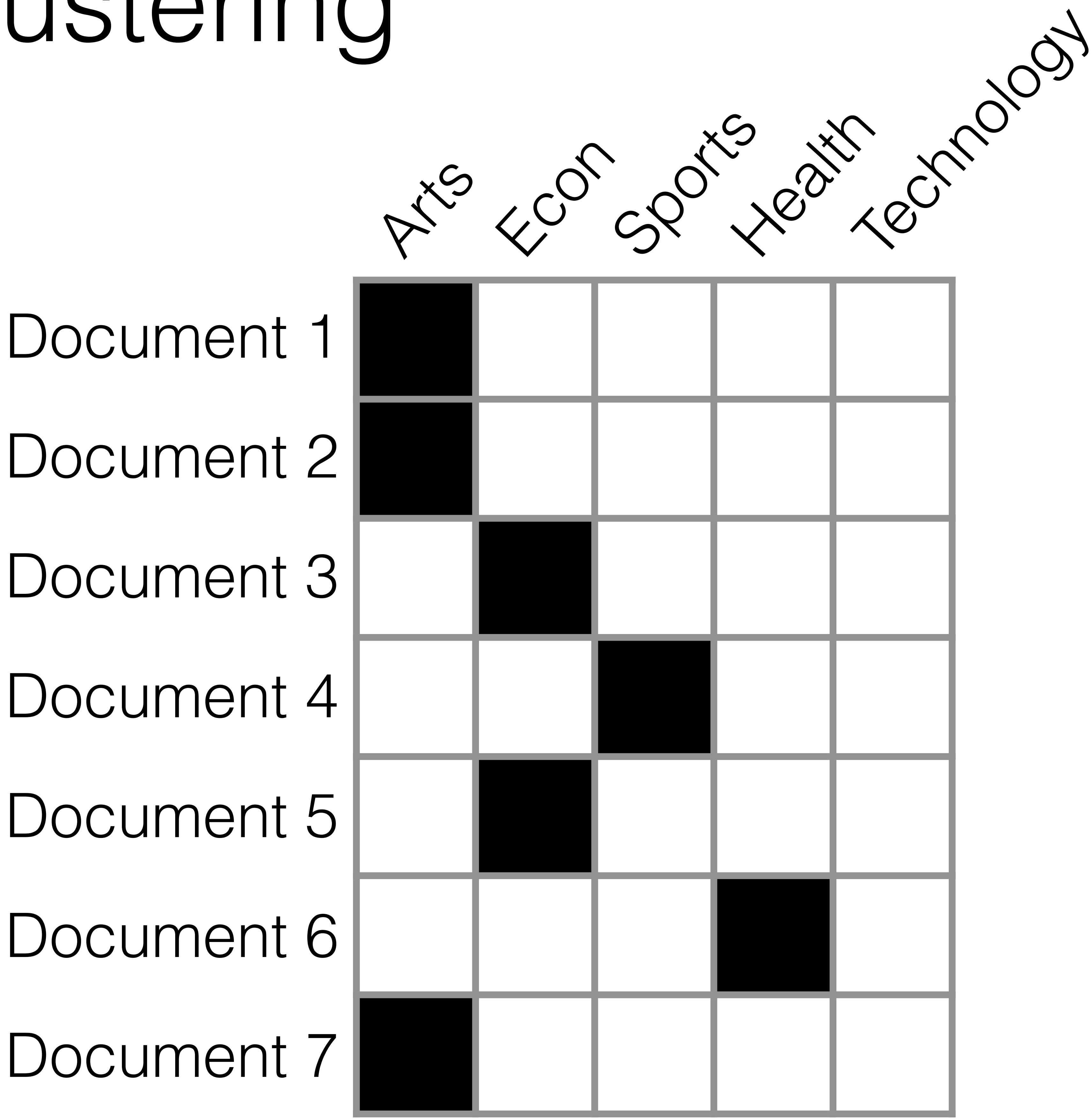


- Variational Bayes (VB)
  - Approximation $q^*(\theta)$ for posterior $p(\theta|x)$
  - "Close": Minimize Kullback-Liebler (KL) divergence:
    $$KL(q\|p(\cdot|x))$$
  - "Nice": factorizes, exponential family, truncation

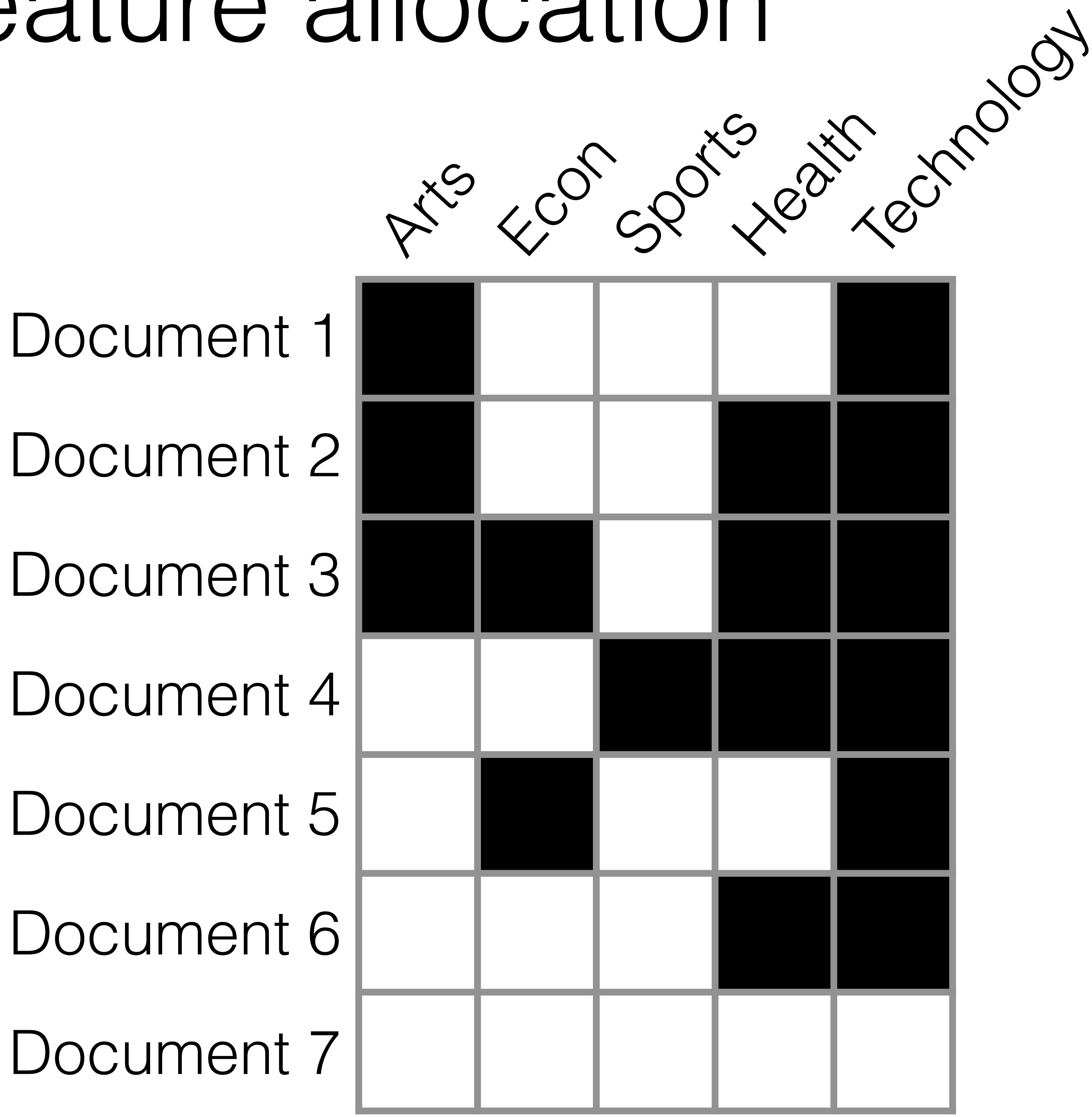
- VB practical success
  - point estimates and prediction
  - fast, streaming, distributed
  - can underestimate uncertainties

[Broderick, Boyd, Wibisono, Wilson, Jordan 2013; Giordano, Broderick, Jordan 2015; Huggins, Campbell, Broderick 2016]

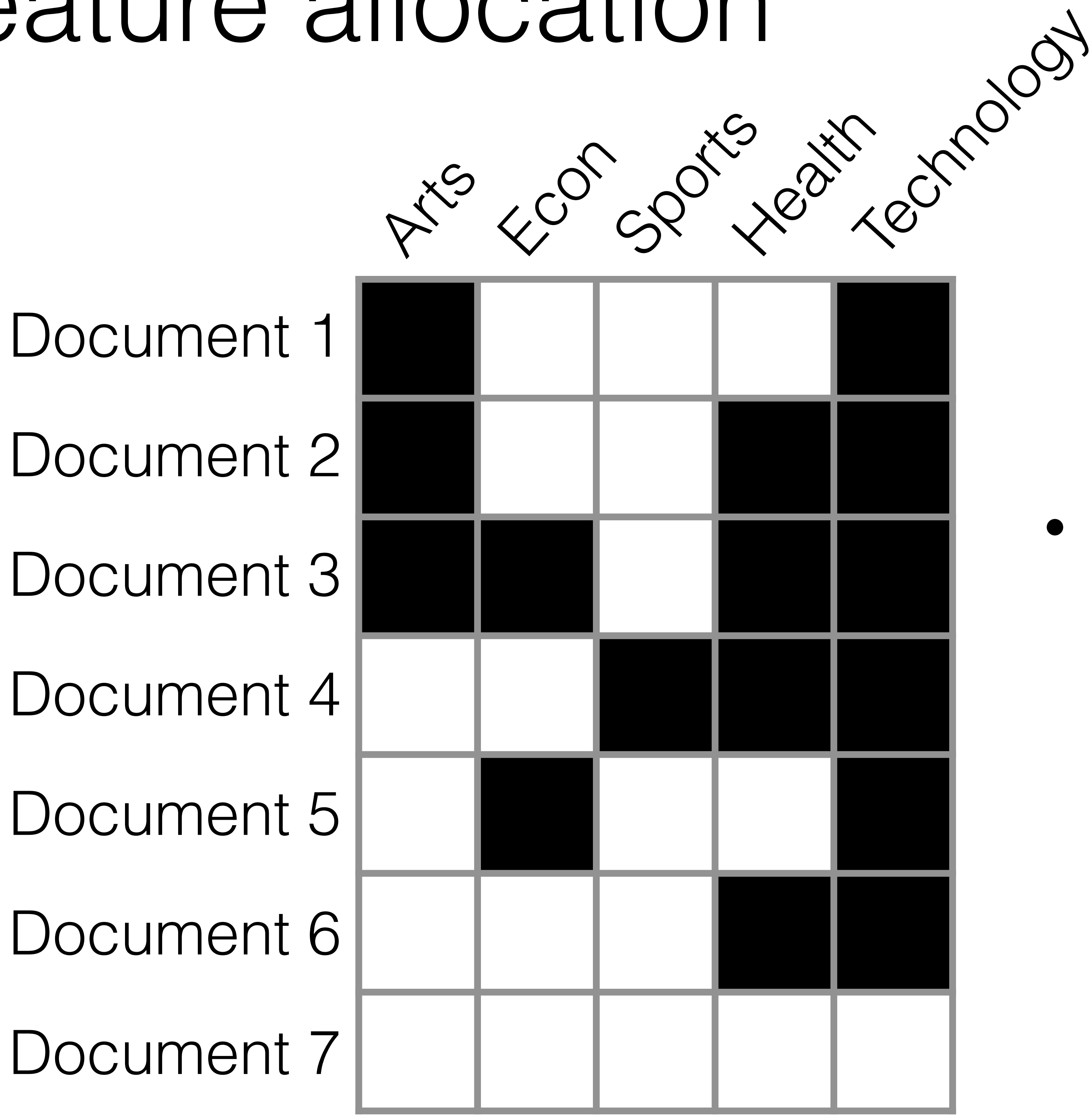
Clustering


Arts
Econ
Sports
Health
Technology
Document 1
Document 2
Document 3
Document 4
Document 5
Document 6
Document 7

Feature allocation

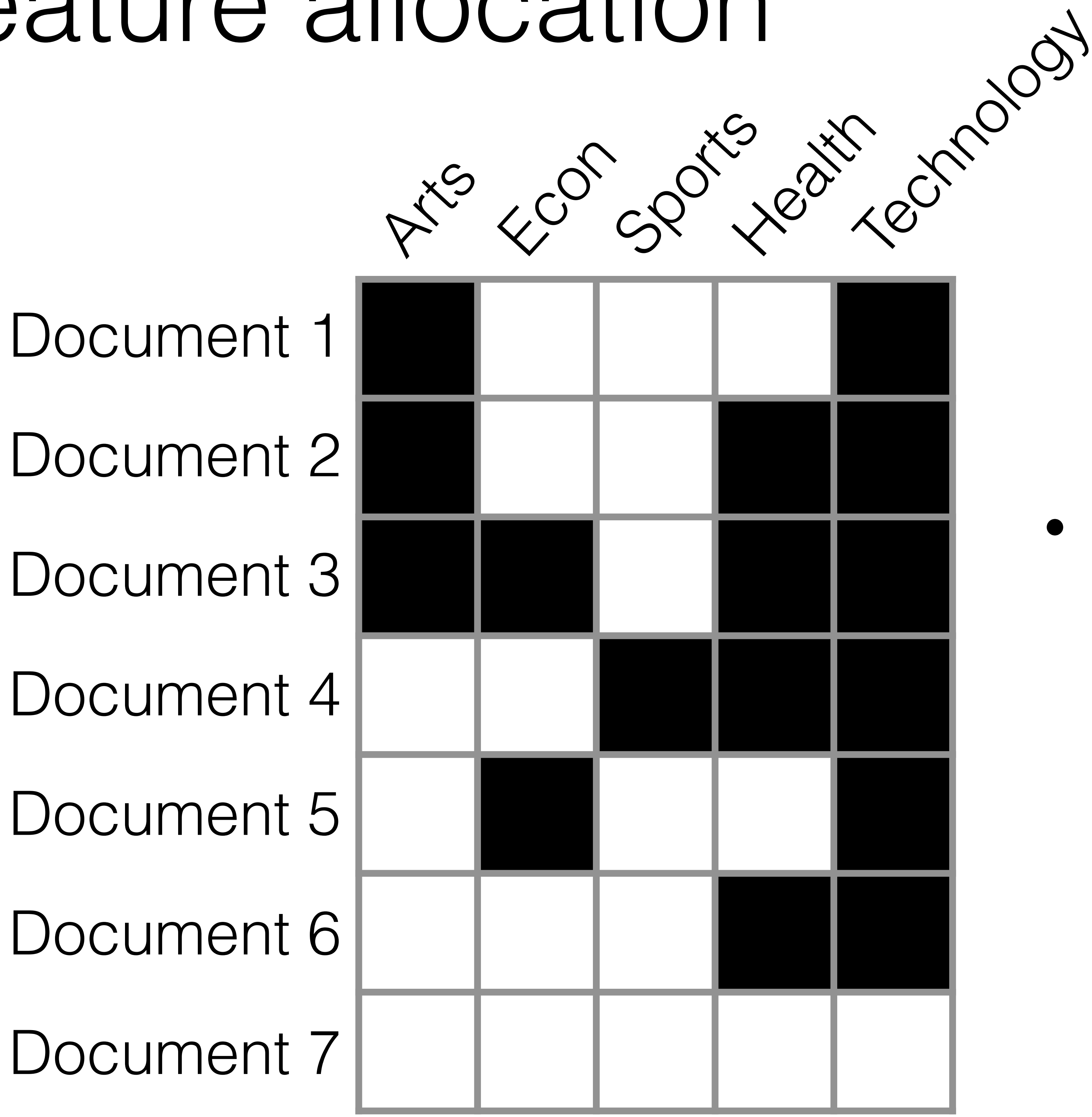


Feature allocation



- Indian buffet process

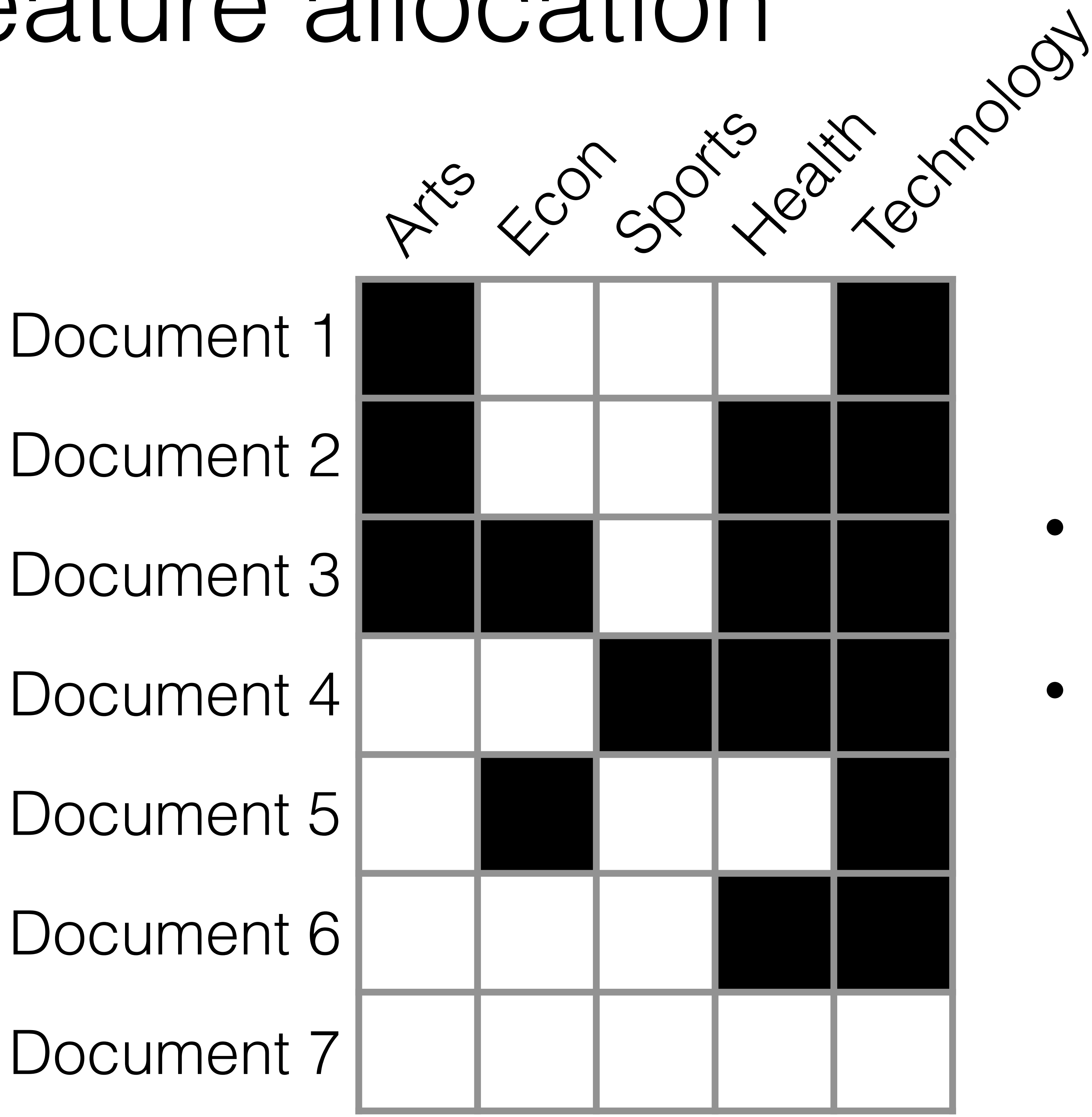
Feature allocation



- Indian buffet process

[Griffiths, Ghahramani 2005

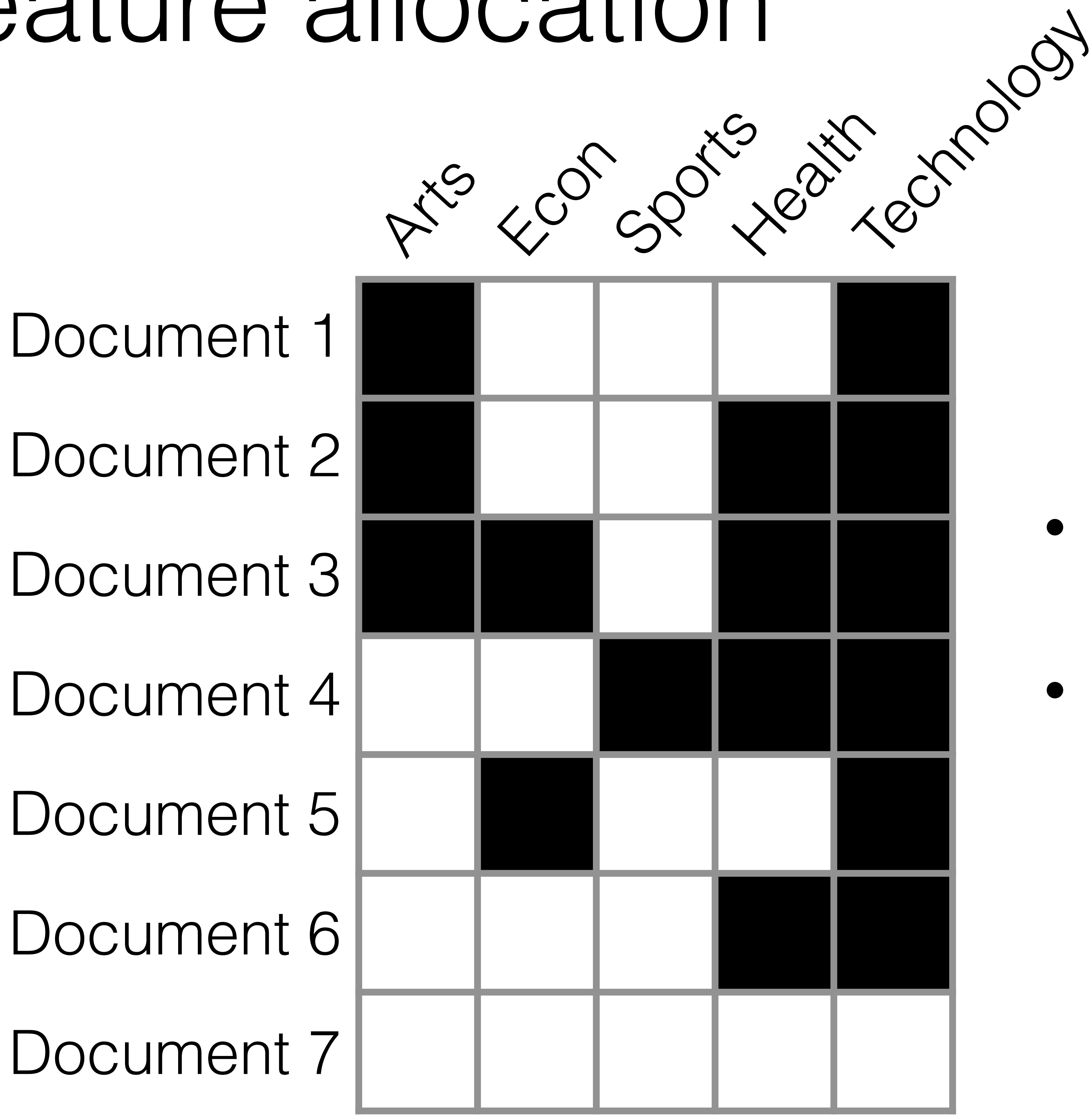
Feature allocation



- Indian buffet process
- Beta process

[Griffiths, Ghahramani 2005

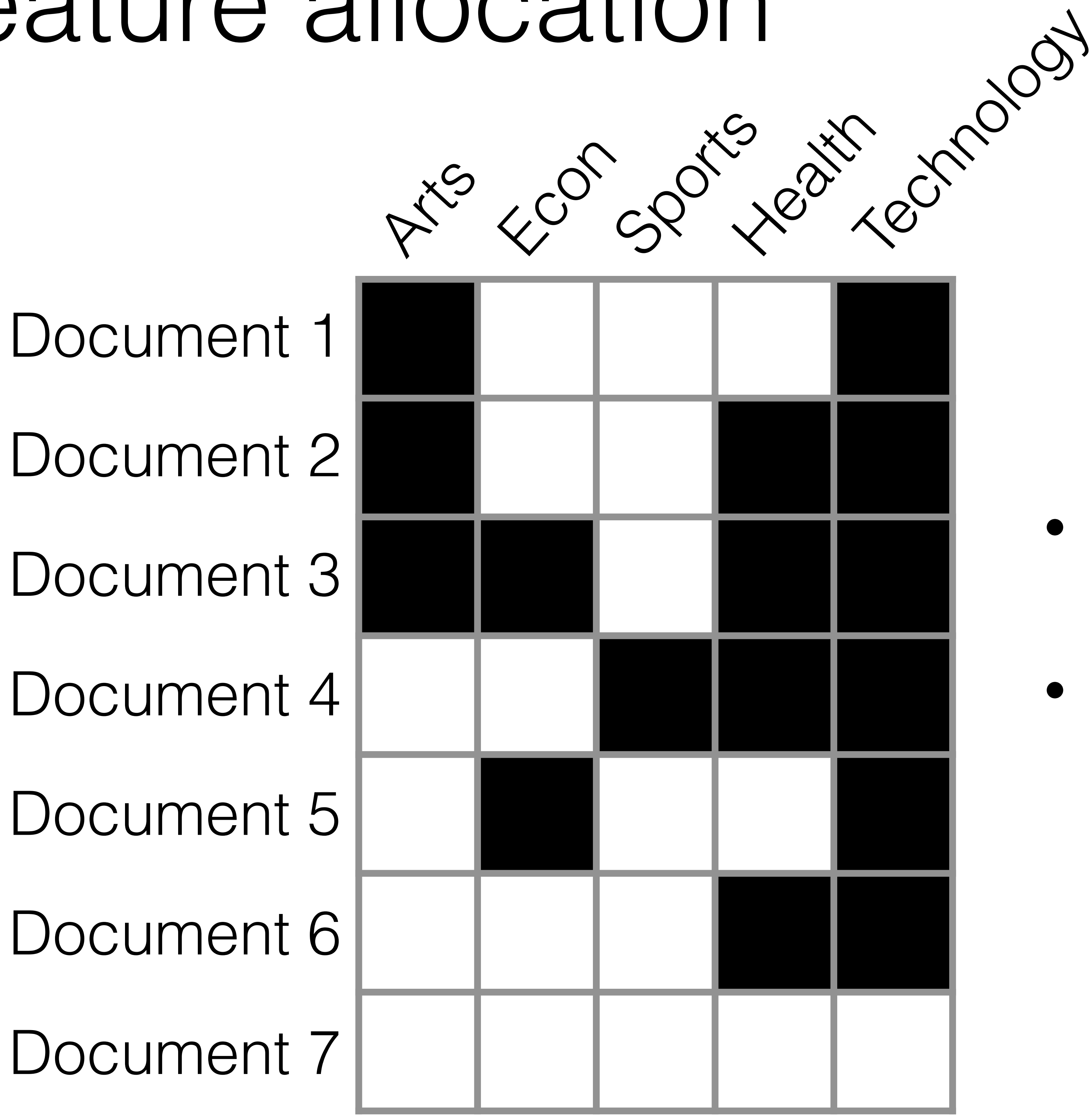
Feature allocation



- Indian buffet process
- Beta process

[Griffiths, Ghahramani 2005, Hjort 1990

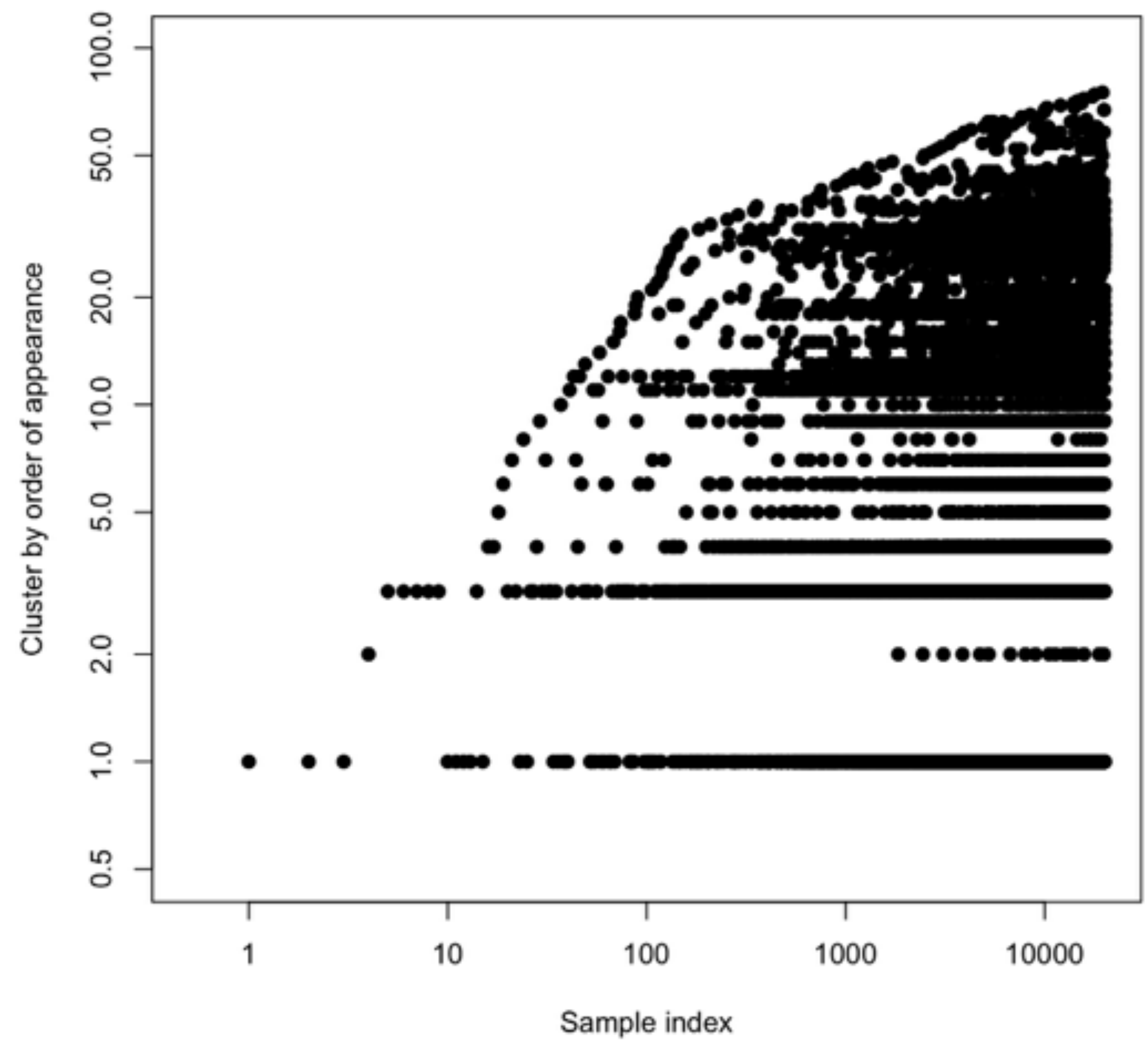
Feature allocation



- Indian buffet process
- Beta process

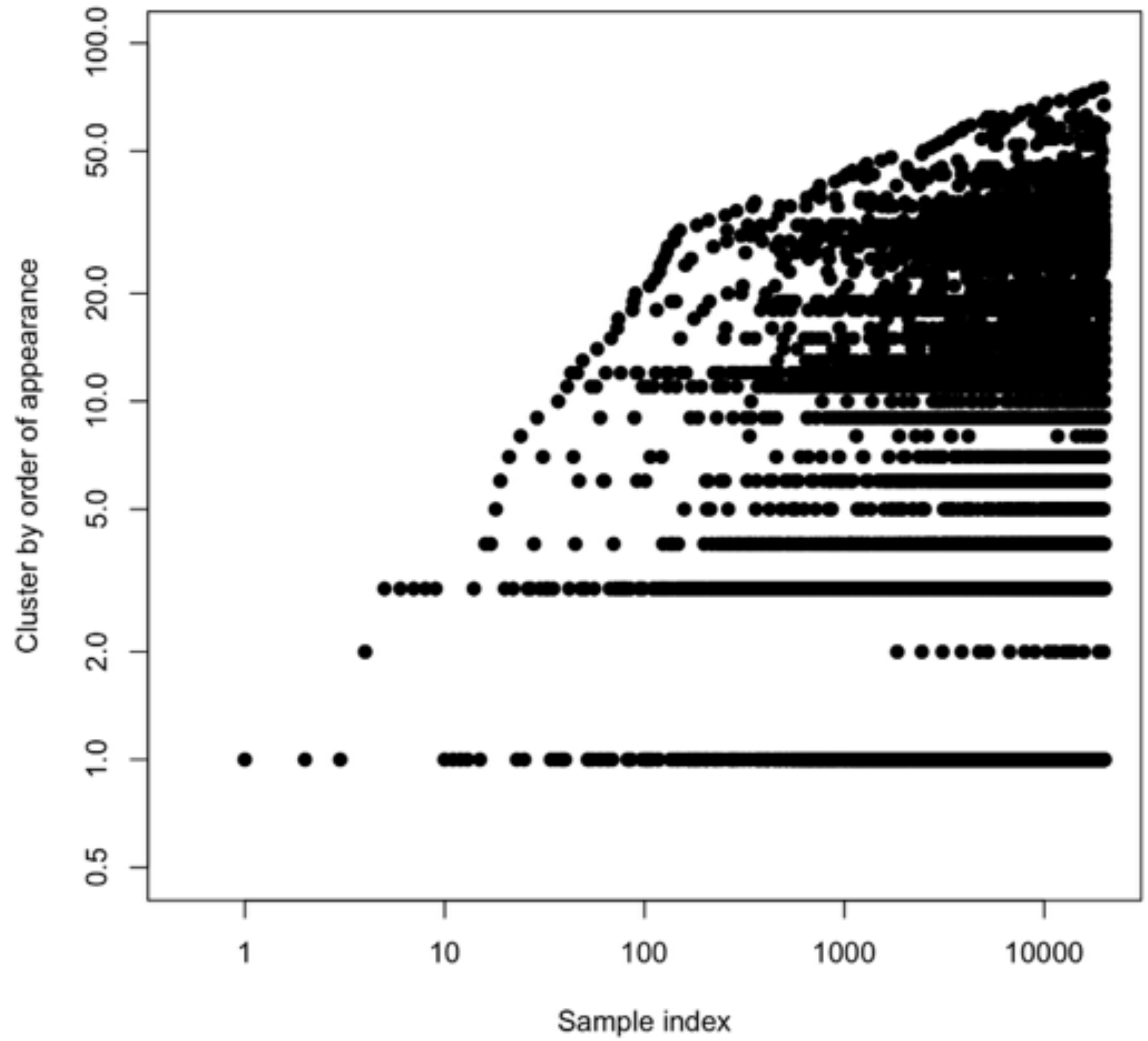
[Griffiths, Ghahramani 2005, Hjort 1990, Kim 1999, Thibaux, Jordan 2007, Broderick, Jordan, Pitman 2013]

Power laws


Cluster by order of appearance
0.5
1.0
2.0
5.0
10.0
20.0
50.0
100.0
1
10
100
1000
10000
Sample index

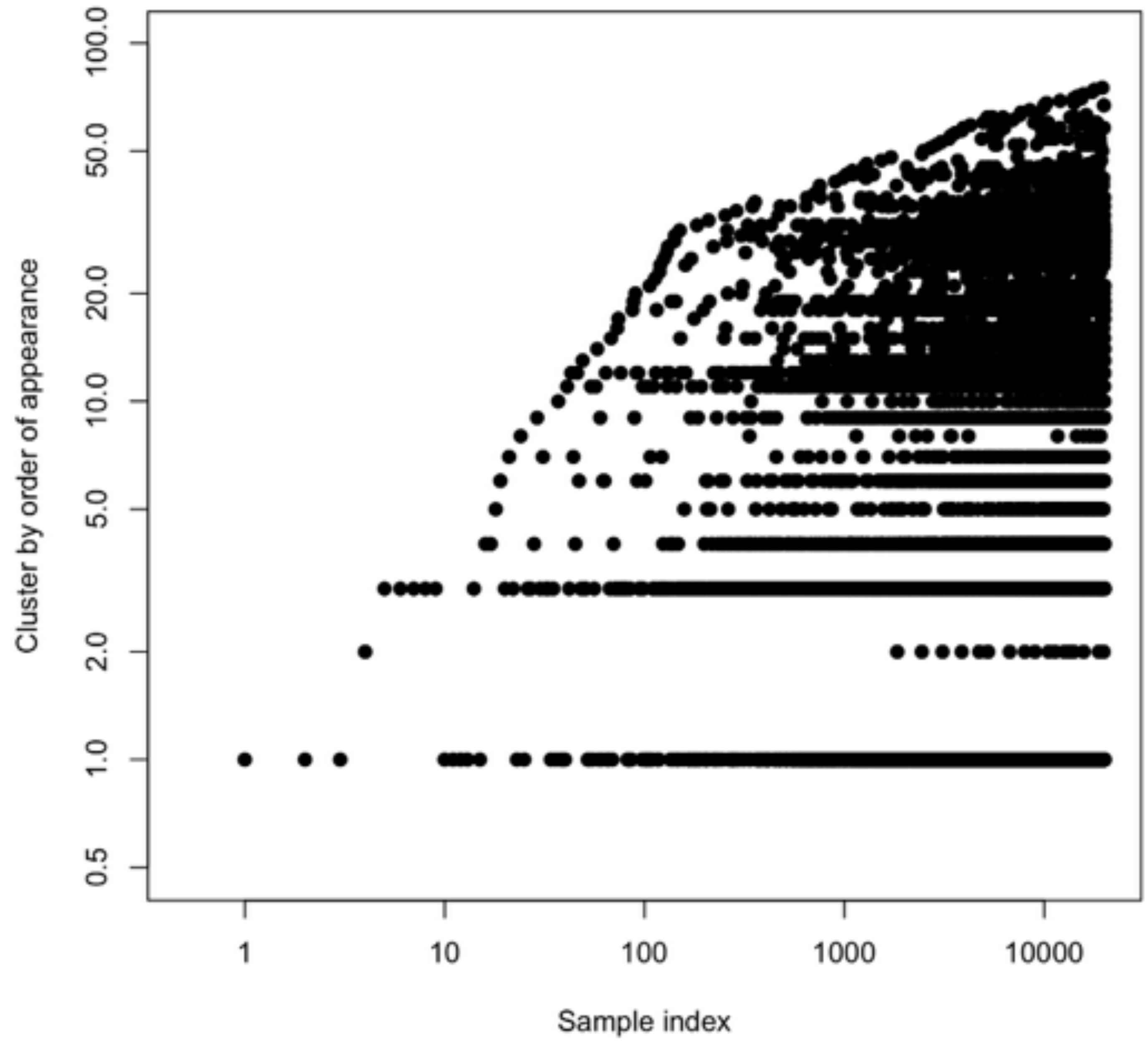
Power laws

- K_N := # clusters occupied by N data points



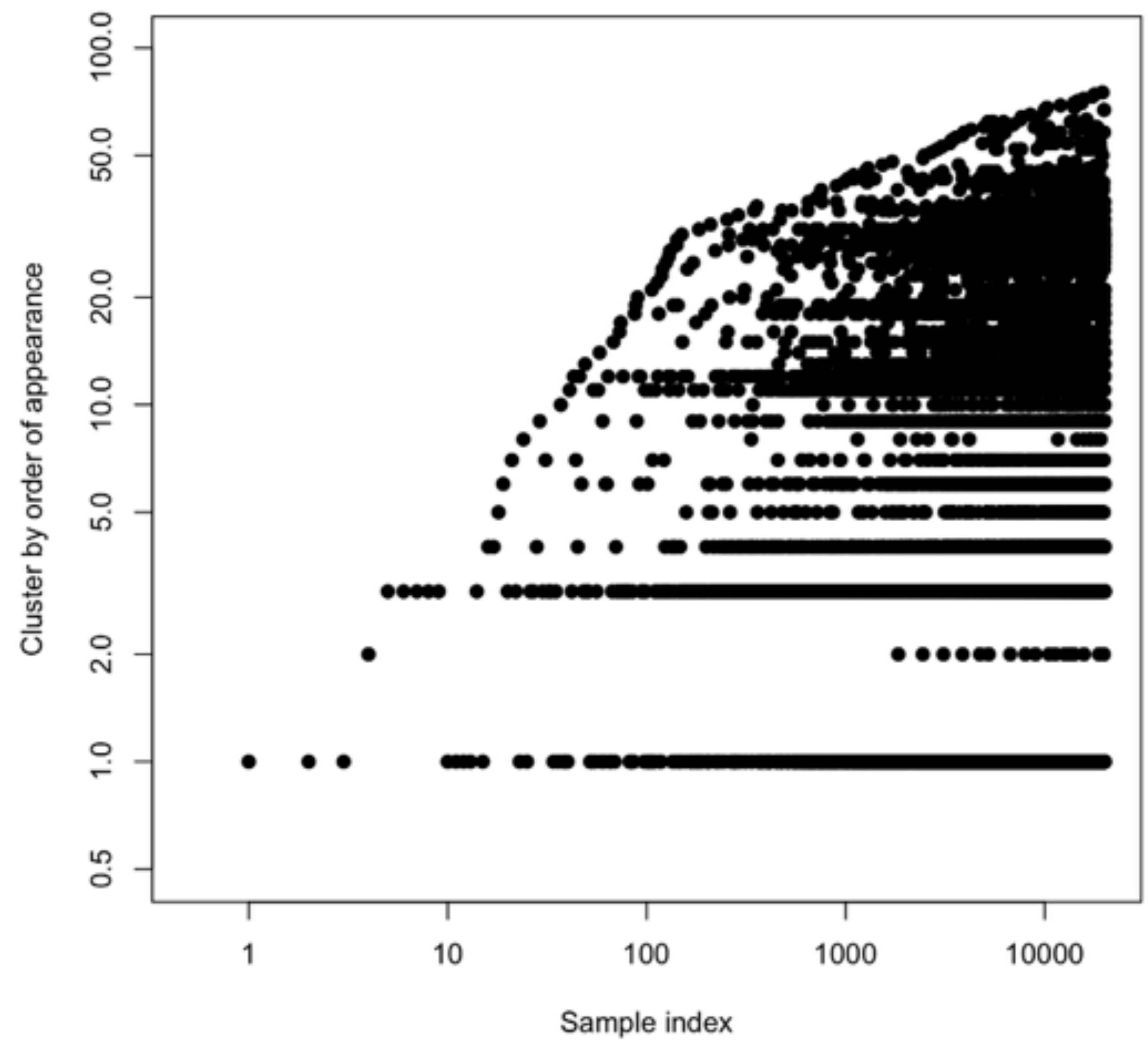
Power laws

- K_N := # clusters occupied by N data points
- CRP: $K_N \sim \alpha \log N \text{ w.p. } 1$



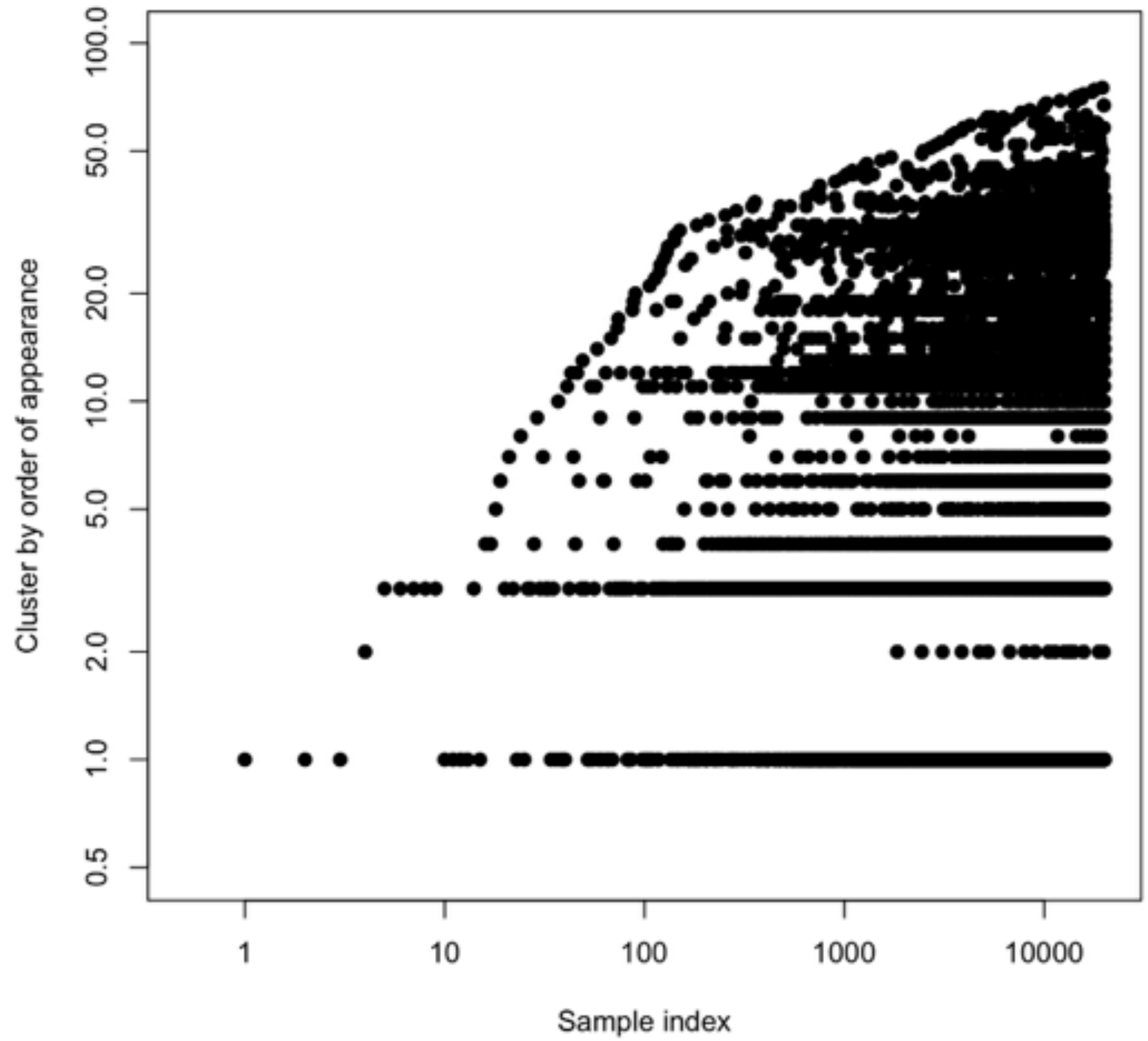
Power laws

- K_N := # clusters occupied by N data points
- CRP: $K_N \sim \alpha \log N \text{ w.p. } 1$
    - vs. Heaps' law, Herdan's law, etc



Power laws

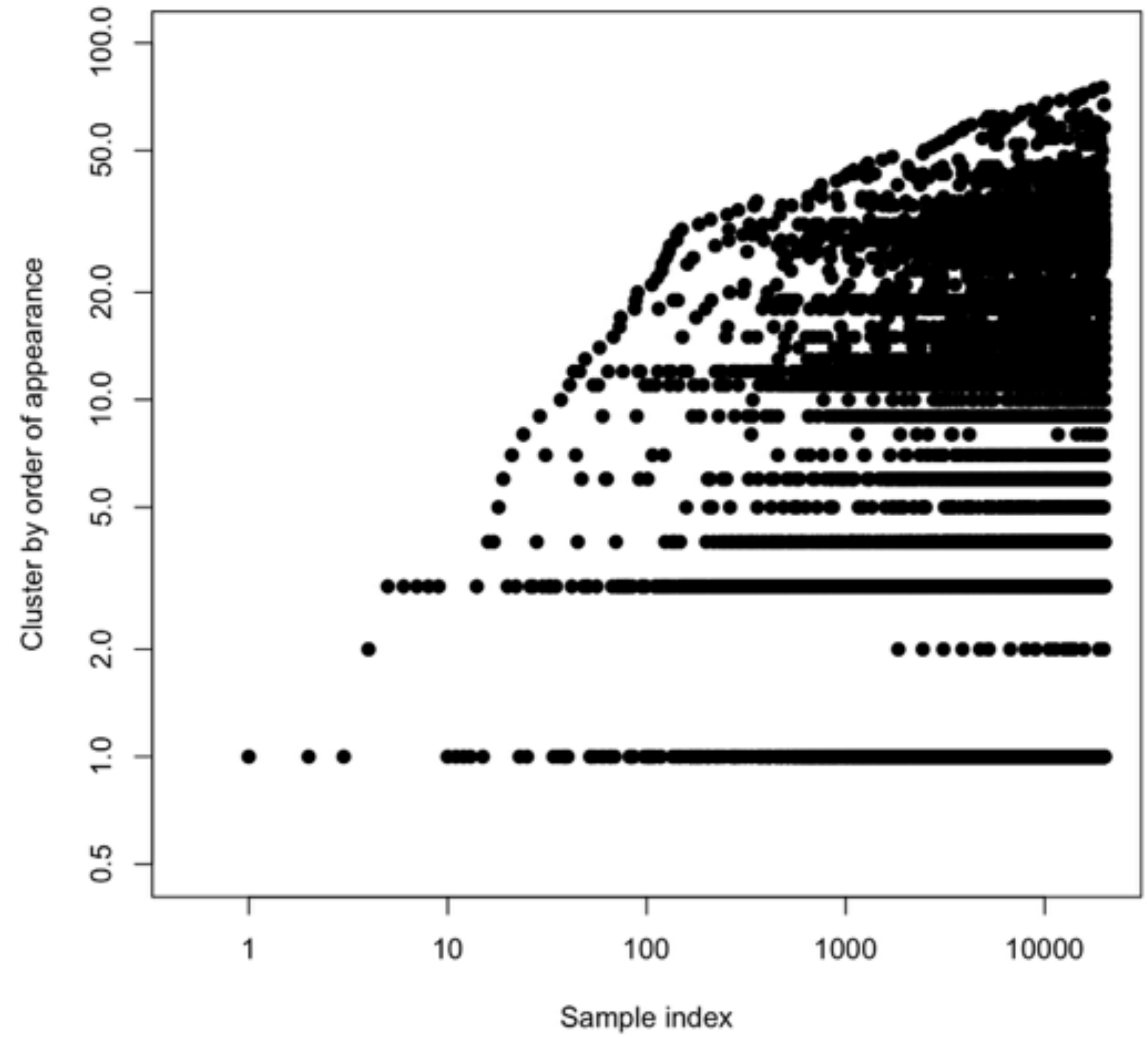
- K_N := # clusters occupied by N data points
- CRP: $K_N \sim \alpha \log N \text{ w.p. } 1$
  - vs. Heaps' law, Herdan's law, etc



[Gnedin, et al 2007

Power laws

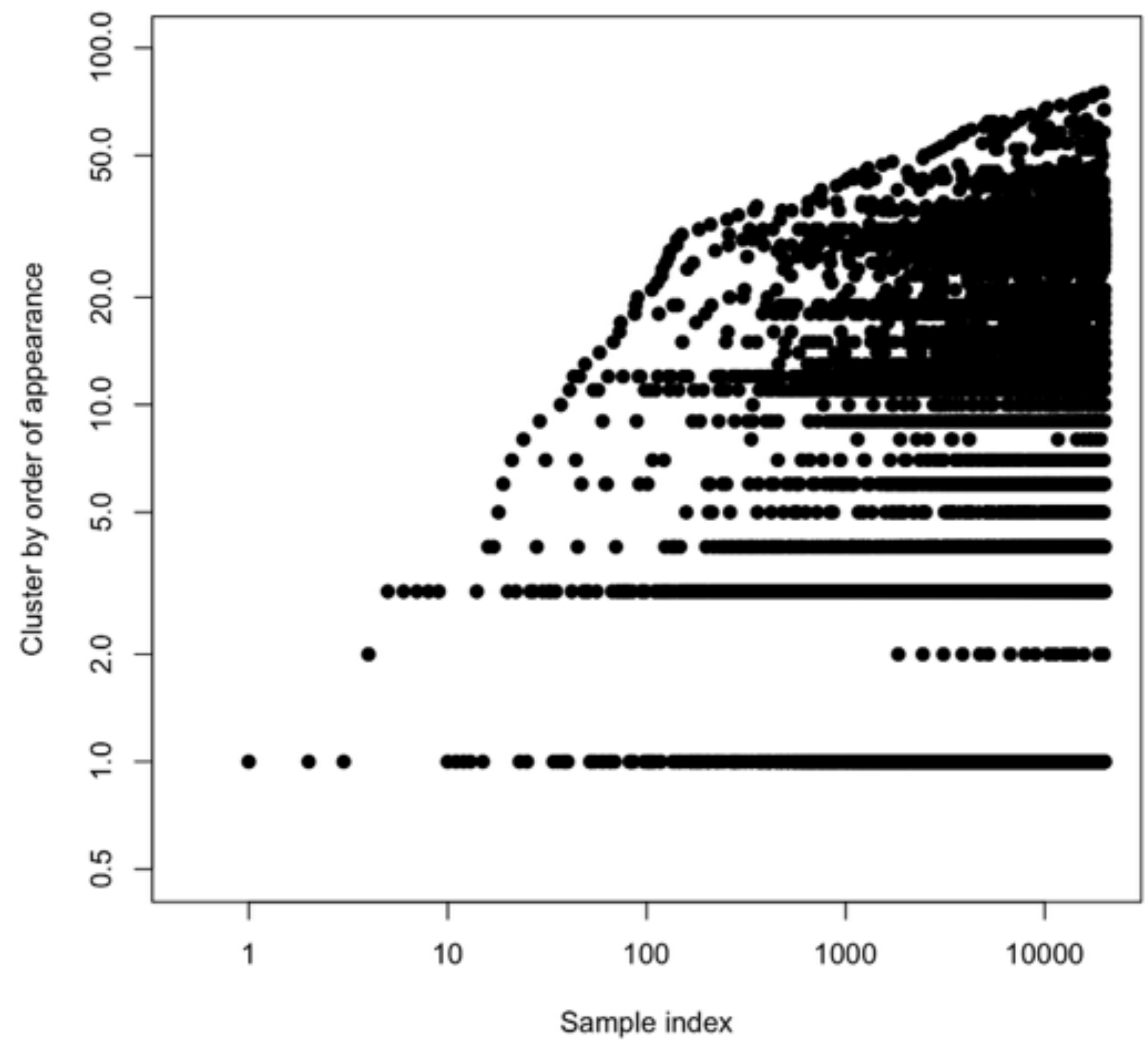
- K_N := # clusters occupied by N data points
- CRP: $K_N \sim \alpha \log N \text{ w.p. } 1$
    - vs. Heaps' law, Herdan's law, etc
- Pitman-Yor process:



[Gnedin, et al 2007

Power laws

- K_N := # clusters occupied by N data points
- CRP: $K_N \sim \alpha \log N \text{ w.p. } 1$
    - vs. Heaps' law, Herdan's law, etc
- Pitman-Yor process:

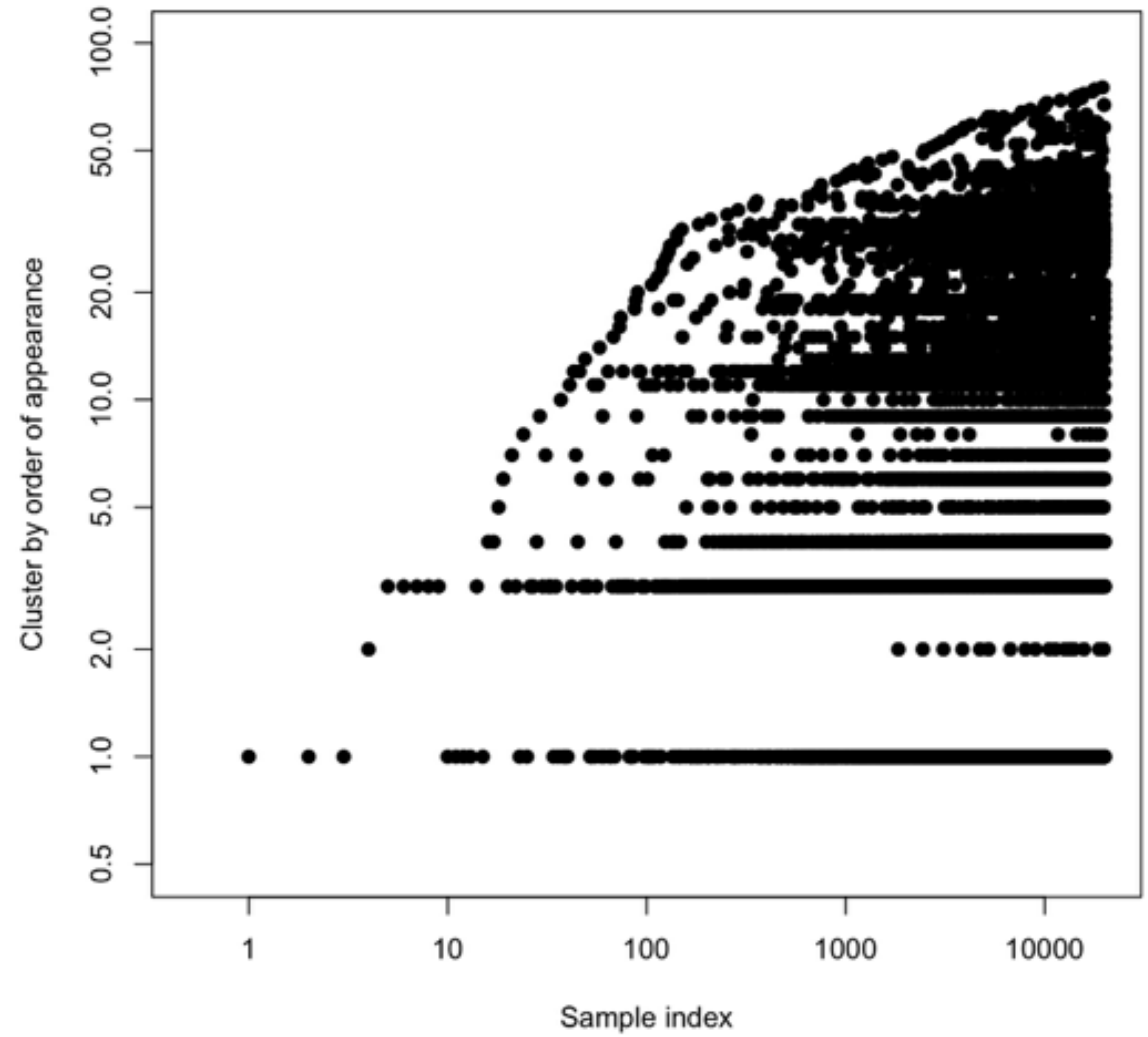


[Gnedin, et al 2007, Pitman, Yor 1997

Power laws

- K_N := # clusters occupied by N data points
- CRP: $K_N \sim \alpha \log N \text{ w.p. } 1$
    - vs. Heaps' law, Herdan's law, etc
- Pitman-Yor process:

$$K_N \sim S_\alpha N^\sigma \text{ w.p. } 1$$

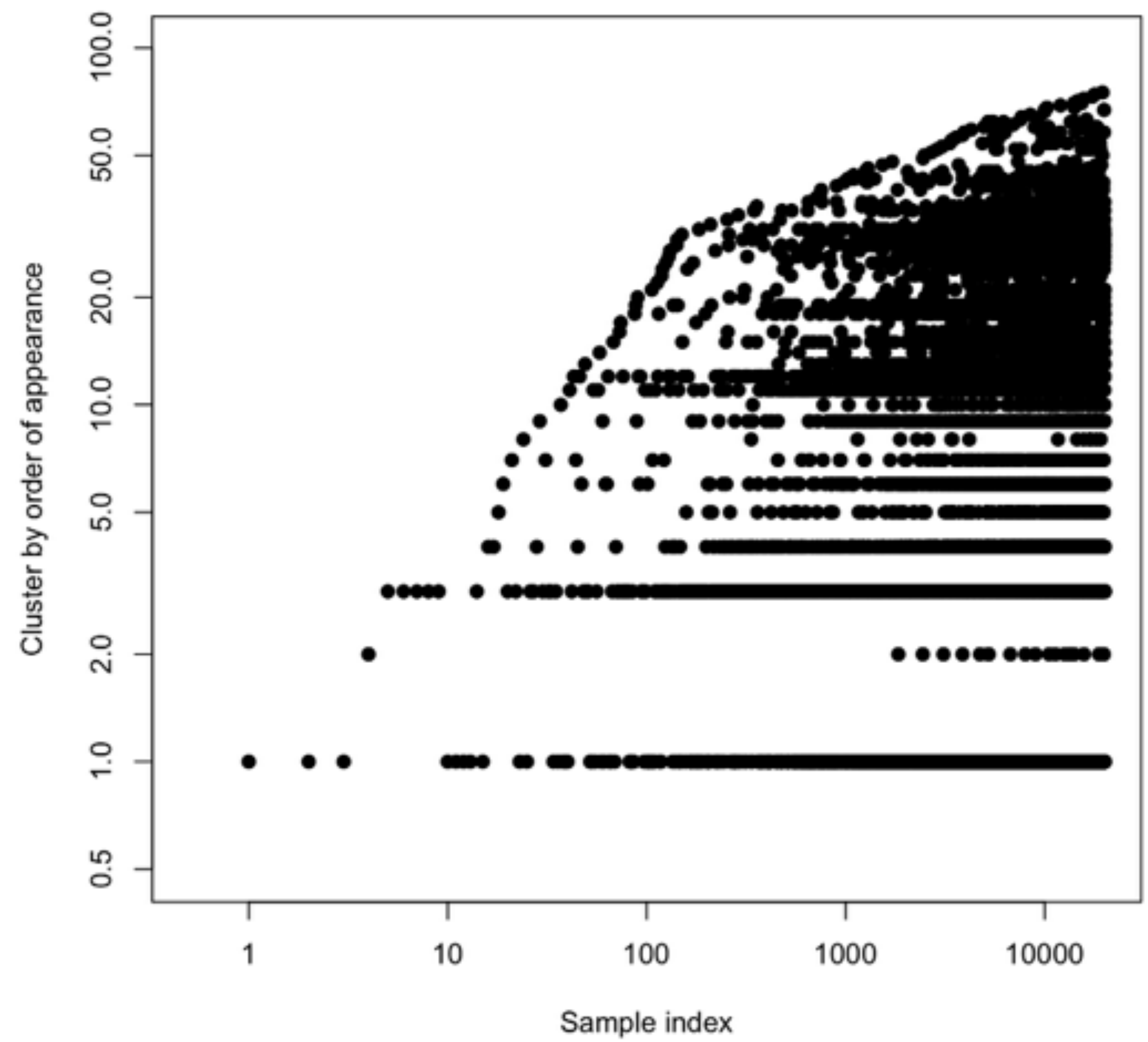


[Gnedin, et al 2007, Pitman, Yor 1997

Power laws

- K_N := # clusters occupied by N data points
- CRP: $K_N \sim \alpha \log N \text{ w.p. } 1$
  - vs. Heaps' law, Herdan's law, etc
- Pitman-Yor process:

$$K_N \sim S_\alpha N^\sigma \text{ w.p. } 1$$



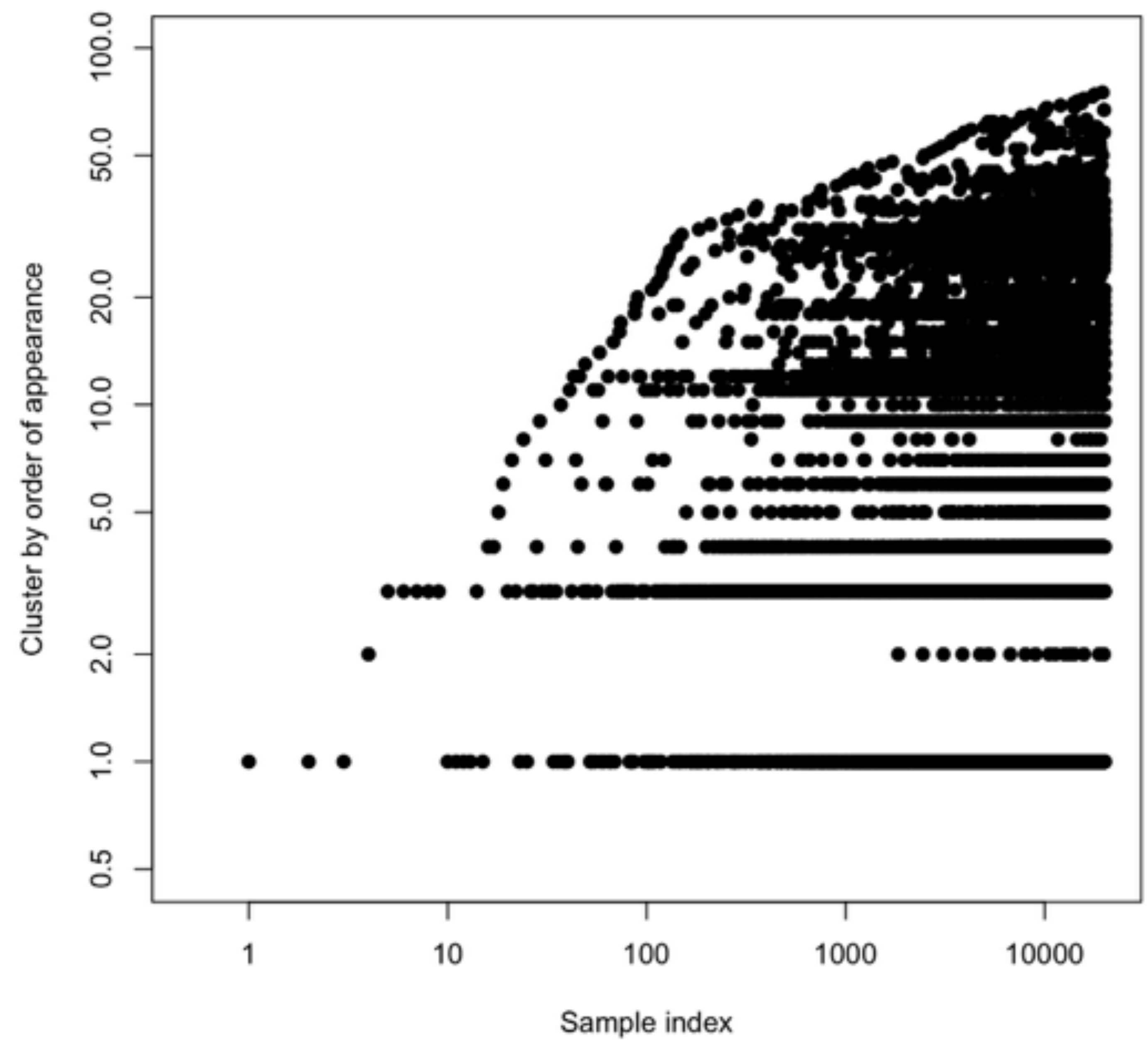
[Gnedin, et al 2007, Pitman, Yor 1997, Goldwater et al 2005, Teh 2006

Power laws

- K_N := # clusters occupied by N data points
- CRP: $K_N \sim \alpha \log N \text{ w.p. } 1$
  - vs. Heaps' law, Herdan's law, etc
- Pitman-Yor process:

  $$K_N \sim S_\alpha N^\sigma \text{ w.p. } 1$$

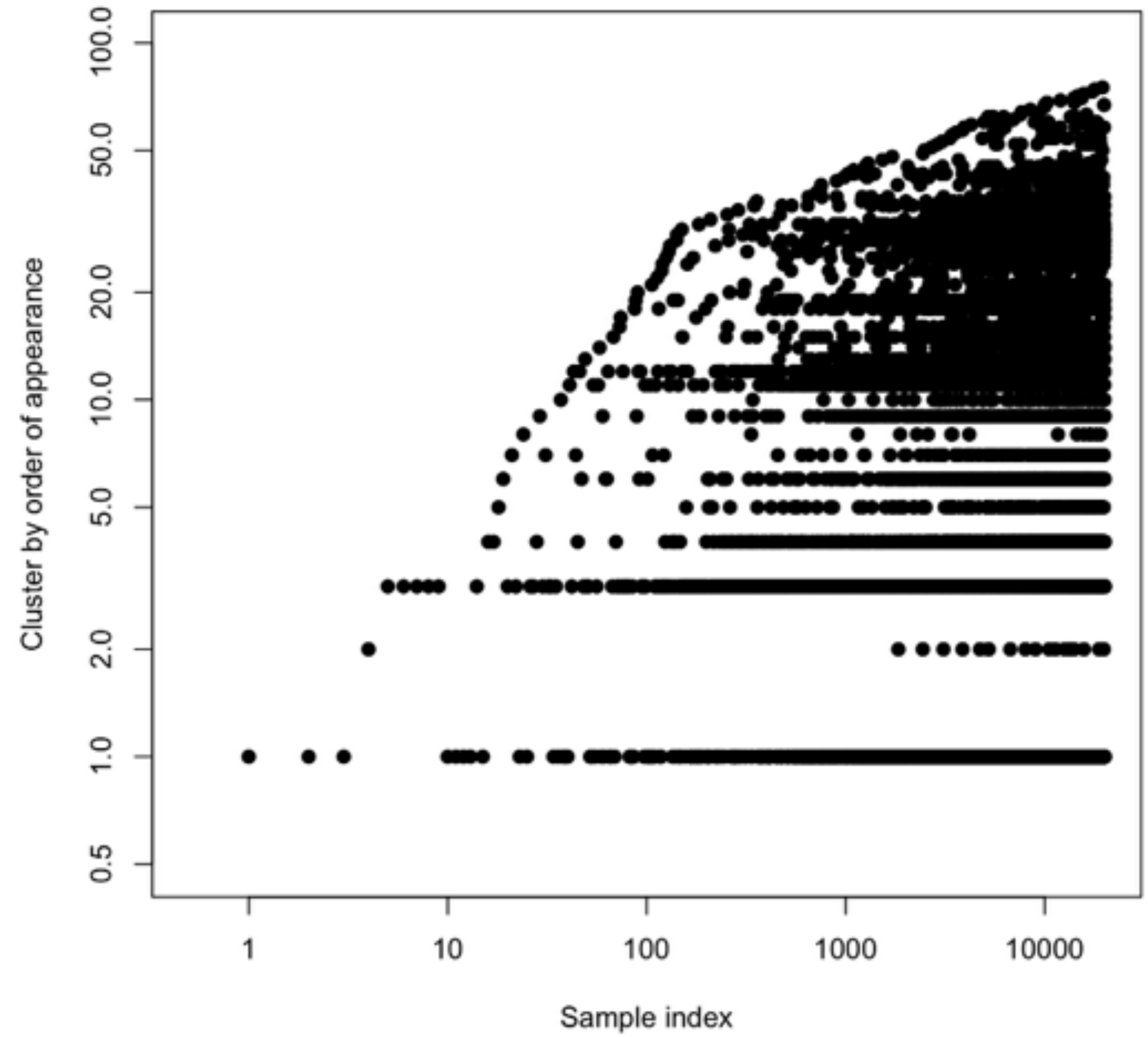
  - related to Zipf's law (ranked frequencies)



[Gnedin, et al 2007, Pitman, Yor 1997, Goldwater et al 2005, Teh 2006

Power laws

- K_N := # clusters occupied by N data points
- CRP: $K_N \sim \alpha \log N \text{ w.p. } 1$
  - vs. Heaps' law, Herdan's law, etc
- Pitman-Yor process:
  $$K_N \sim S_\alpha N^\sigma \text{ w.p. } 1$$
  - related to Zipf's law (ranked frequencies)
- Not just clusters



[Gnedin, et al 2007, Pitman, Yor 1997, Goldwater et al 2005, Teh 2006, Broderick et al 2012]

Hierarchies

Hierarchies

- Hierarchical Dirichlet process

Hierarchies

- Hierarchical Dirichlet process

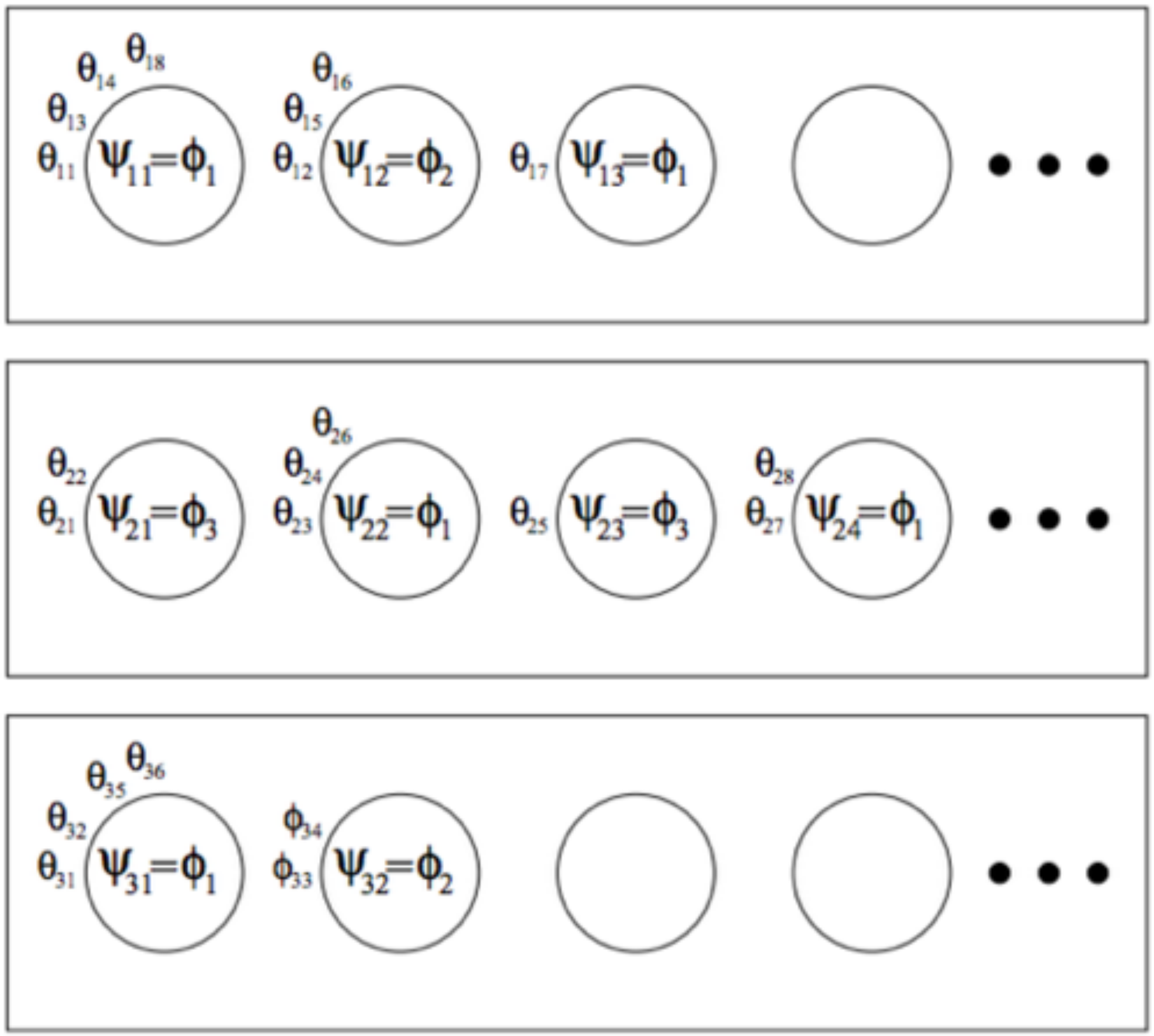
[Teh et al 2006, Rodríguez et al 2008

Hierarchies

- Hierarchical Dirichlet process
- Chinese restaurant franchise

[Teh et al 2006, Rodríguez et al 2008

Hierarchies

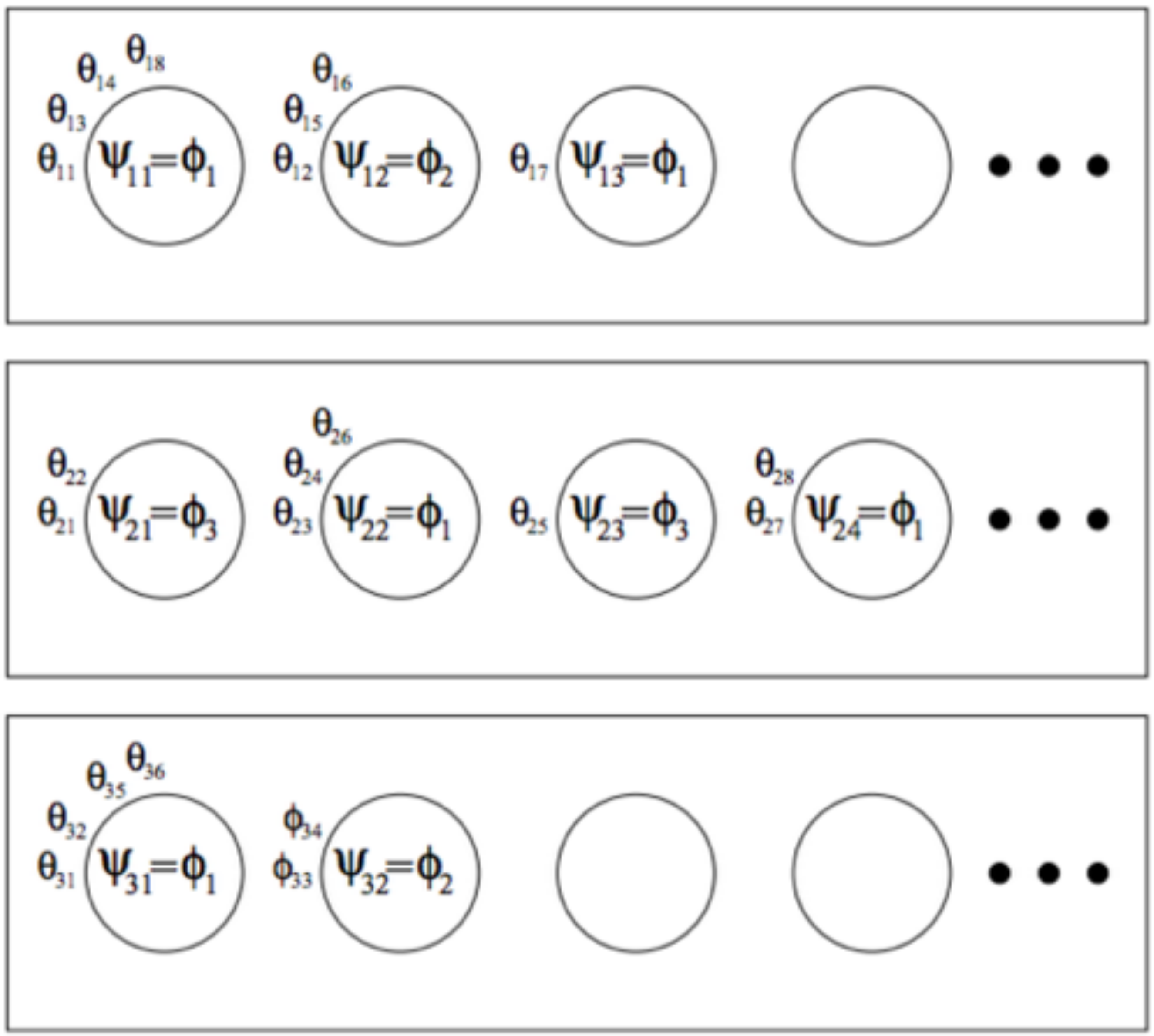


[Teh et al 2006]

- Hierarchical Dirichlet process
- Chinese restaurant franchise

[Teh et al 2006, Rodríguez et al 2008

Hierarchies

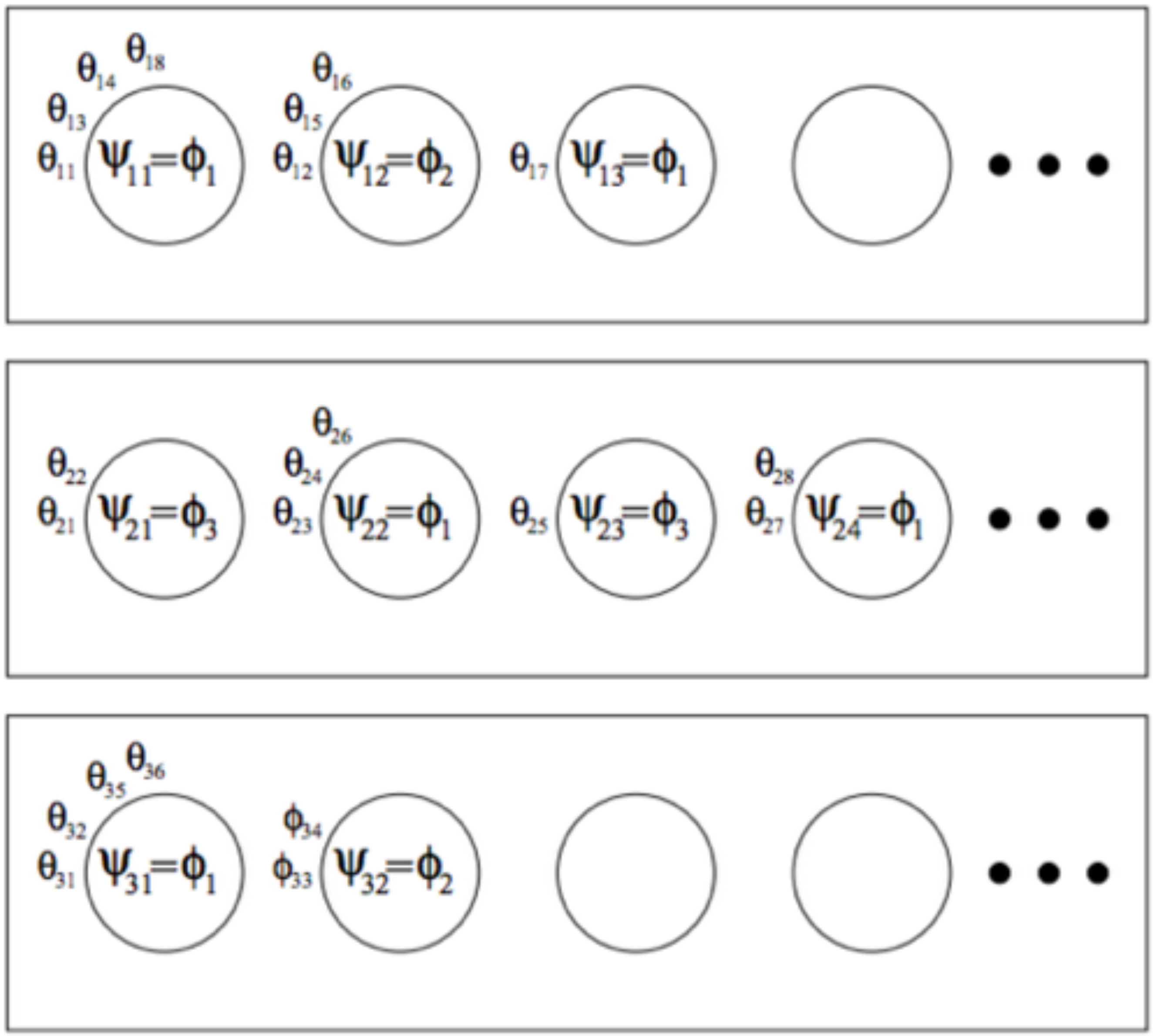


[Teh et al 2006]

- Hierarchical Dirichlet process
- Chinese restaurant franchise
- Hierarchical beta process

[Teh et al 2006, Rodríguez et al 2008

Hierarchies

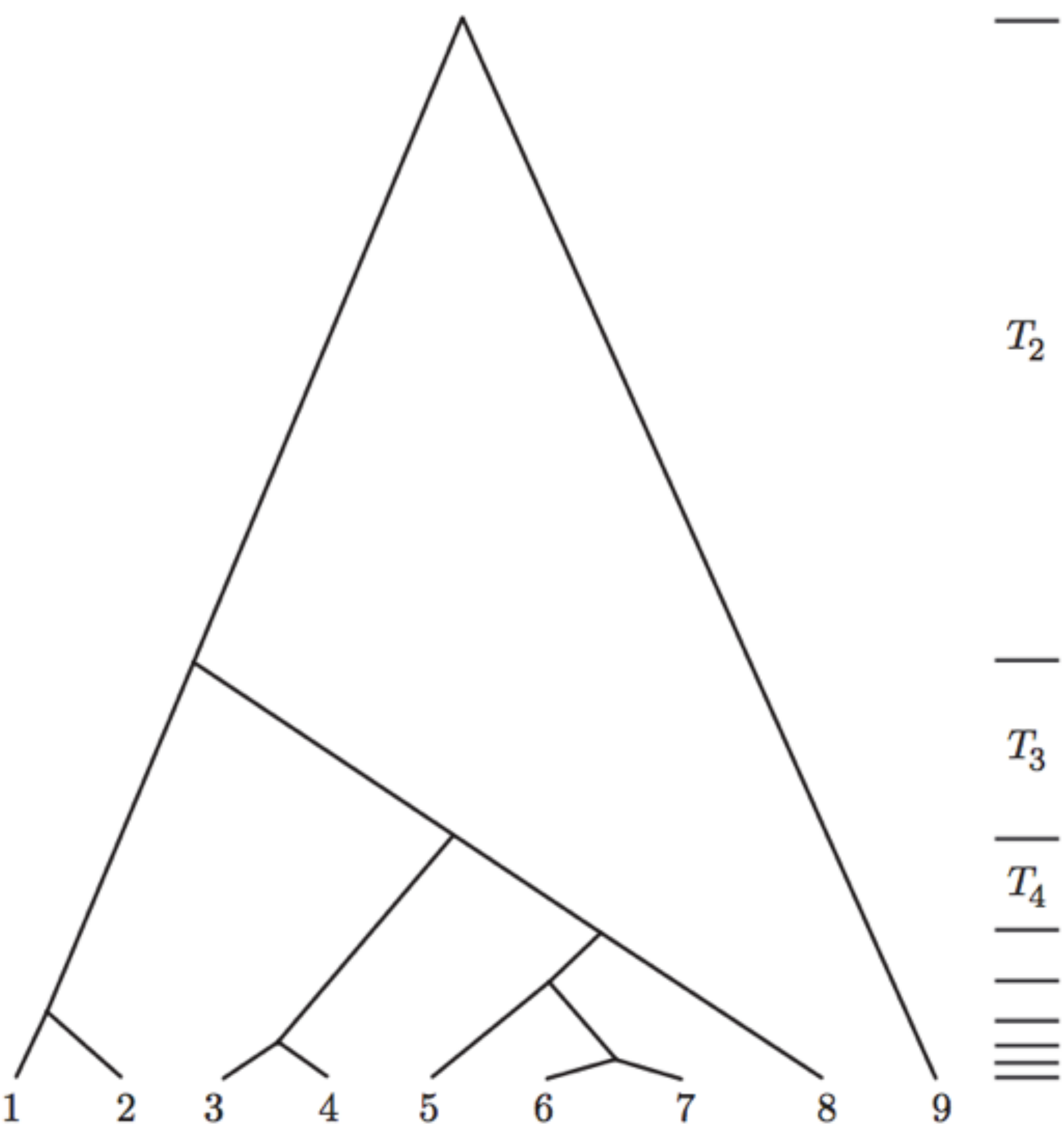


[Teh et al 2006]

- Hierarchical Dirichlet process
- Chinese restaurant franchise
- Hierarchical beta process

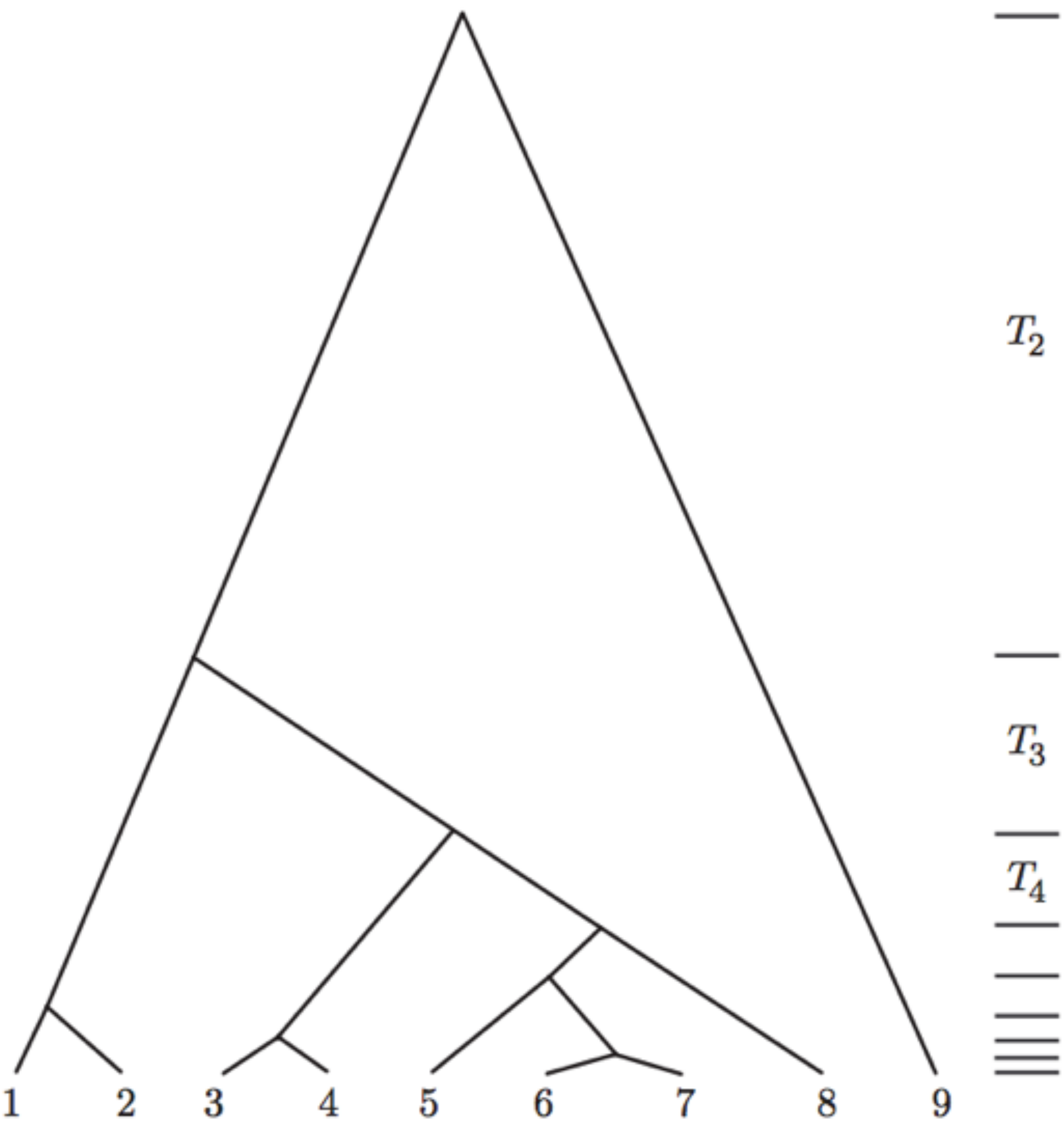
[Teh et al 2006, Rodríguez et al 2008, Thibaux, Jordan 2007]

Genealogy, trees, beyond trees



[Wakeley 2008]

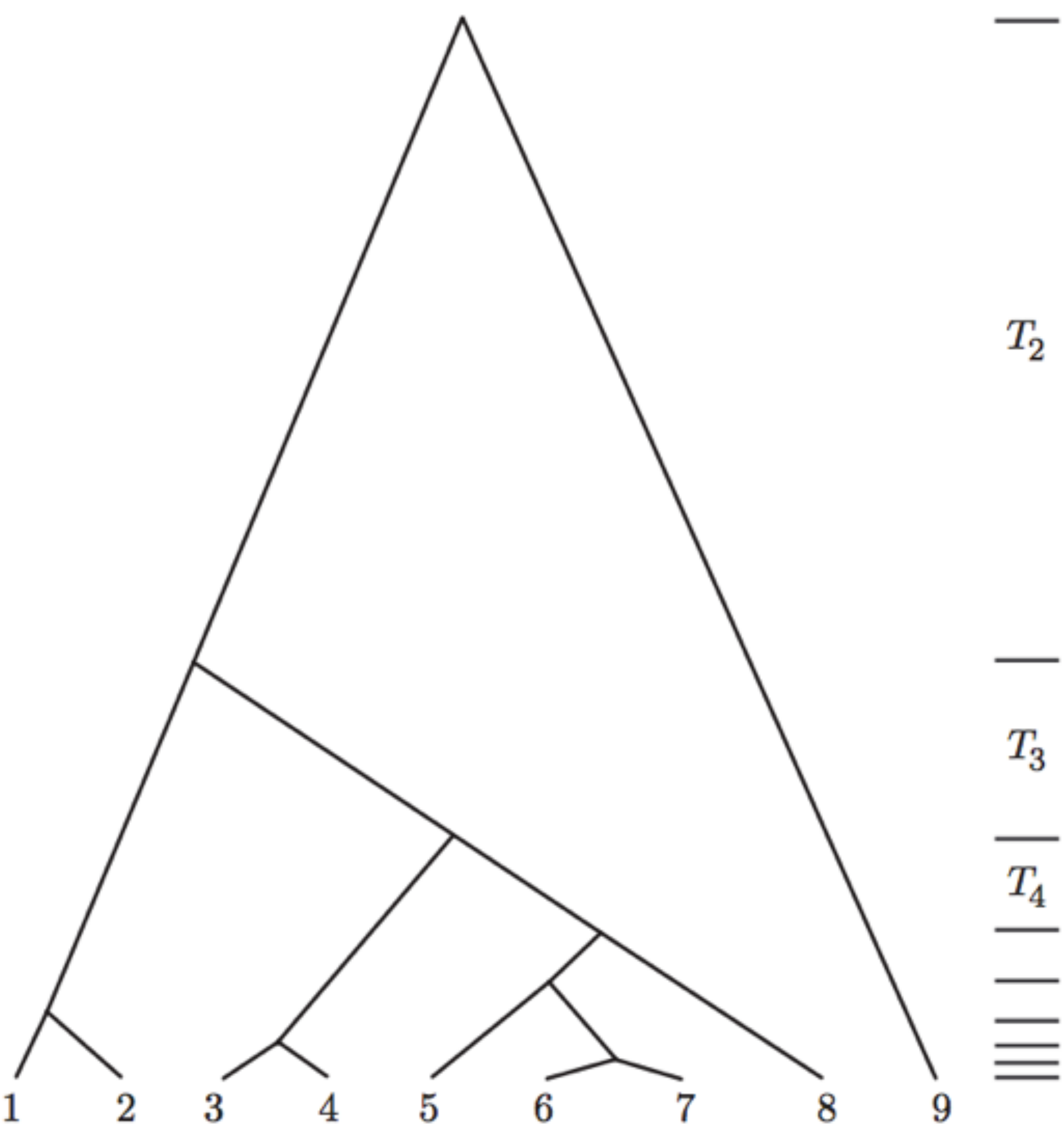
Genealogy, trees, beyond trees



[Wakeley 2008]

- Kingman coalescent

Genealogy, trees, beyond trees

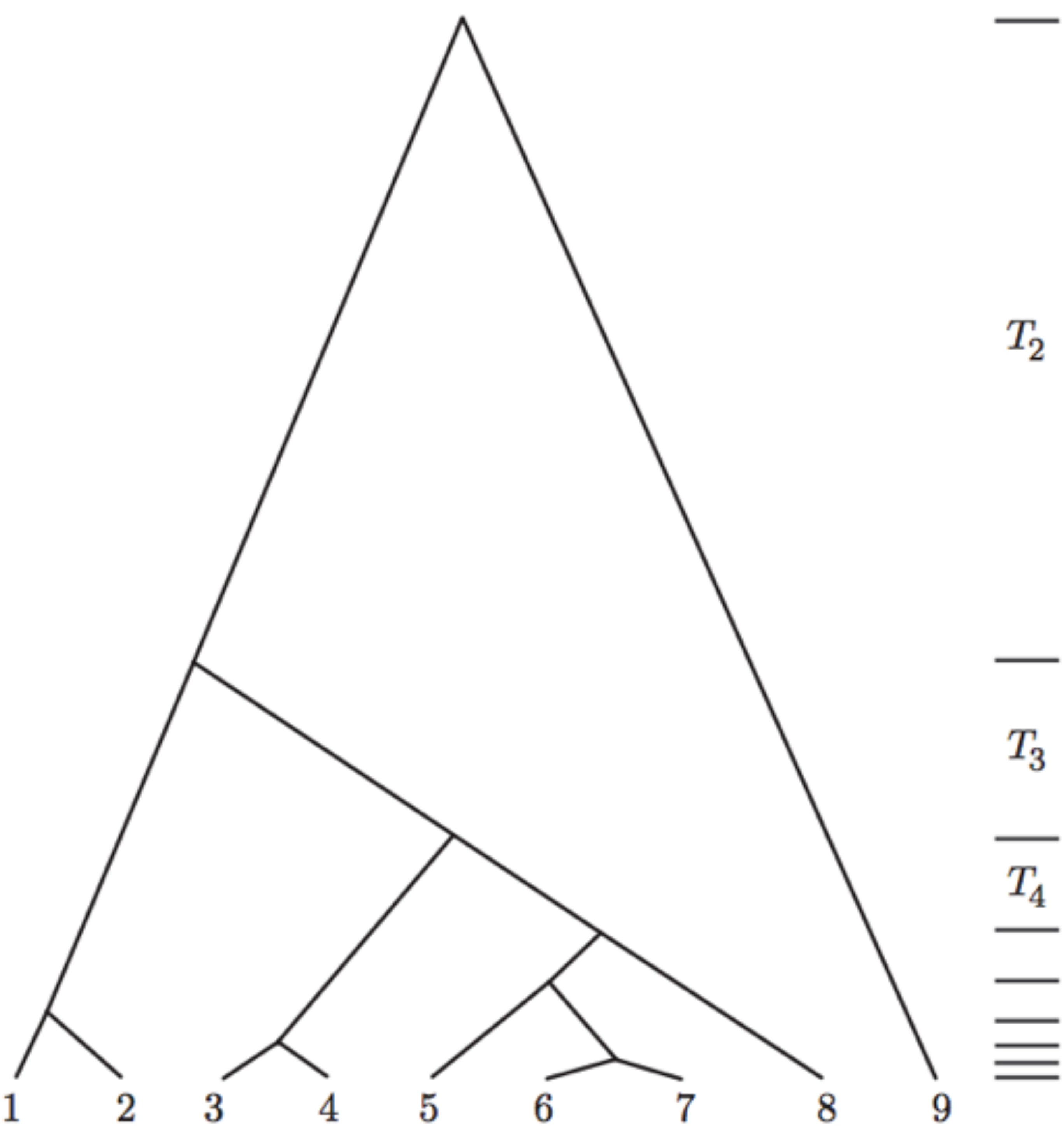


[Wakeley 2008]

- Kingman coalescent

[Kingman 1982

Genealogy, trees, beyond trees

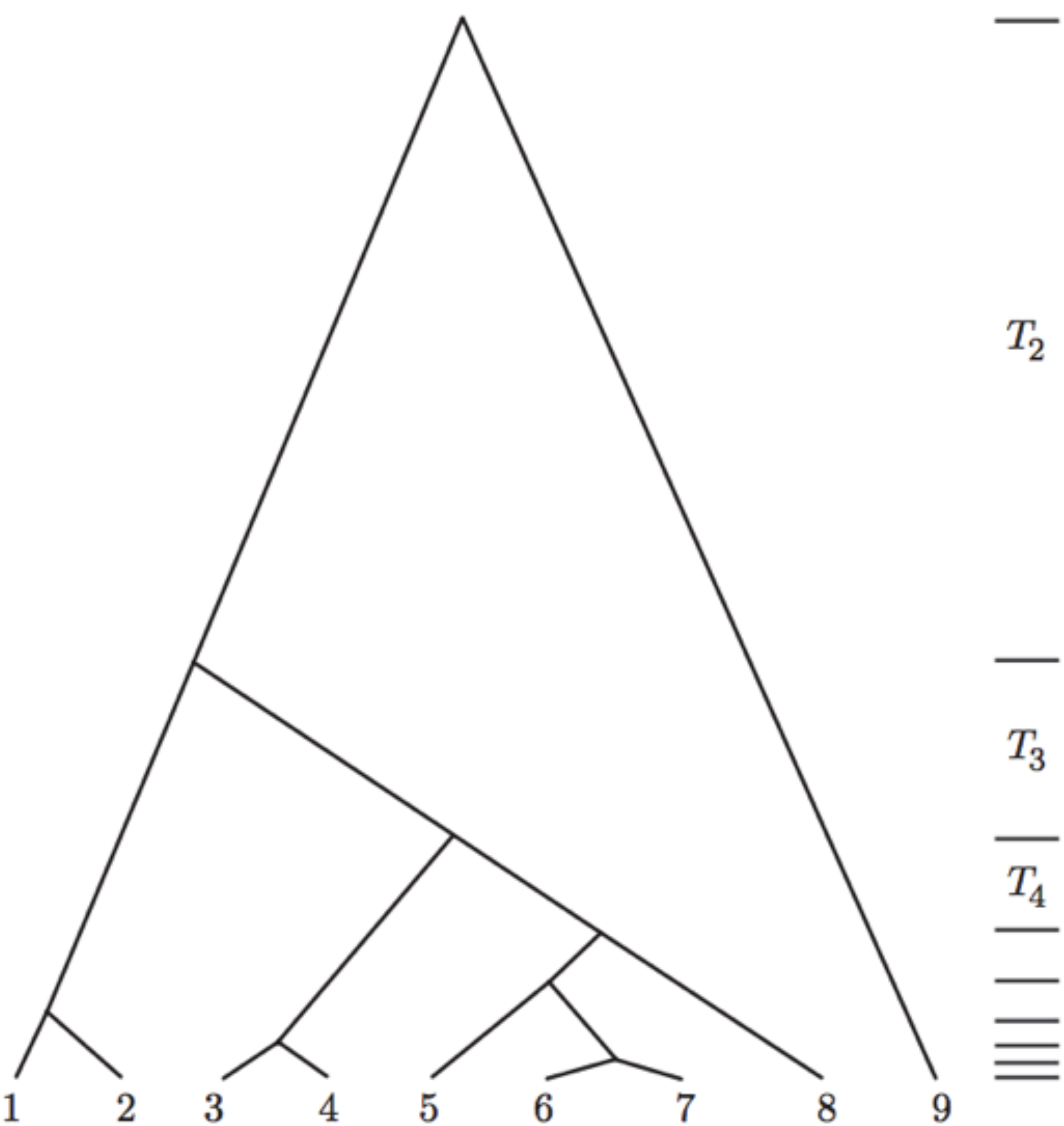


[Wakeley 2008]

- Kingman coalescent
- Fragmentation
- Coagulation

[Kingman 1982

Genealogy, trees, beyond trees

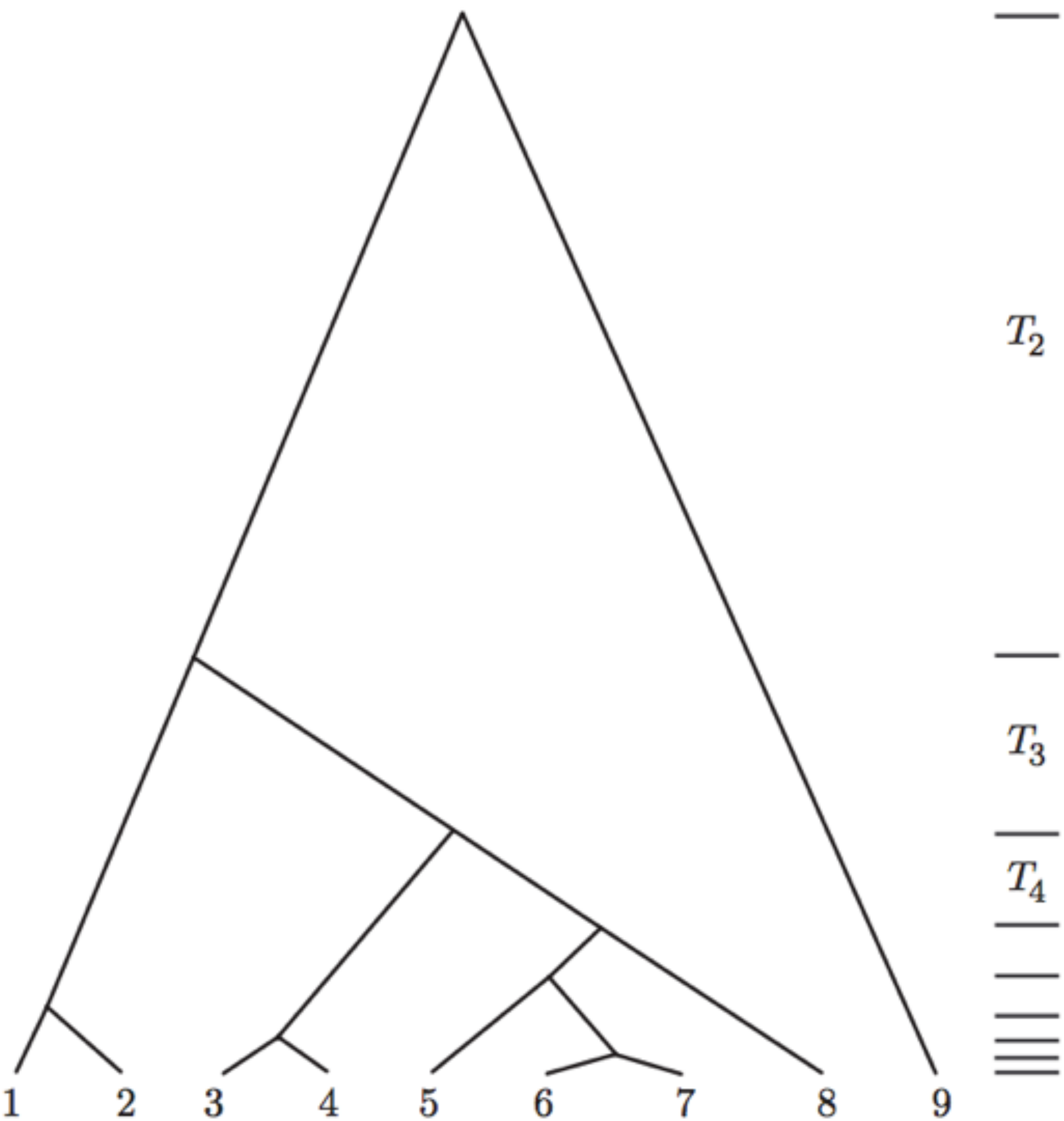


[Wakeley 2008]

- Kingman coalescent
- Fragmentation
- Coagulation

[Kingman 1982, Bertoin 2006, Teh et al 2011

Genealogy, trees, beyond trees

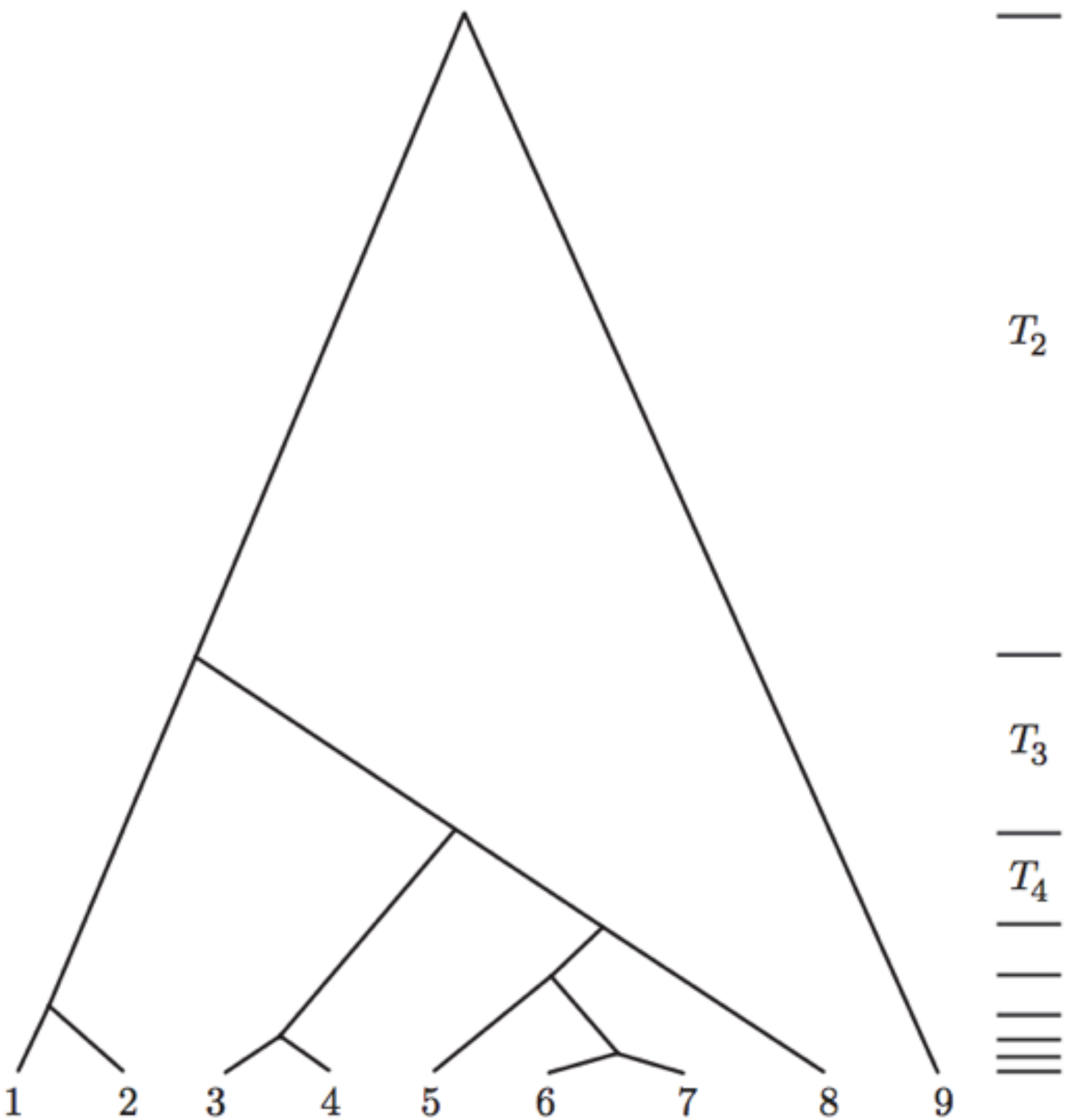


[Wakeley 2008]

- Kingman coalescent
- Fragmentation
- Coagulation
- Dirichlet diffusion tree

[Kingman 1982, Bertoin 2006, Teh et al 2011

Genealogy, trees, beyond trees



[Wakeley 2008]

- Kingman coalescent
- Fragmentation
- Coagulation
- Dirichlet diffusion tree

[Kingman 1982, Bertoin 2006, Teh et al 2011, Neal 2003]

Conjugacy & Poisson point processes

Orbanz 2009, Orbanz 2010

Conjugacy & Poisson point processes

- Beta process, Bernoulli process (Indian buffet)

Orbanz 2009, Orbanz 2010

Conjugacy & Poisson point processes

- Beta process, Bernoulli process (Indian buffet)
- Gamma process, Poisson likelihood process (DP, CRP)

Orbanz 2009, Orbanz 2010

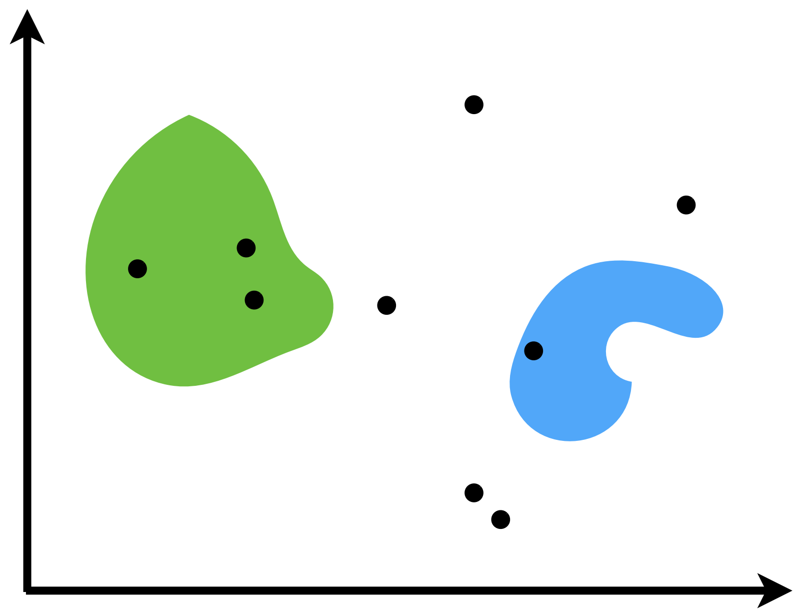
Conjugacy & Poisson point processes

- Beta process, Bernoulli process (Indian buffet)
- Gamma process, Poisson likelihood process (DP, CRP)
- Beta process, negative binomial process

Orbanz 2009, Orbanz 2010

Conjugacy & Poisson point processes

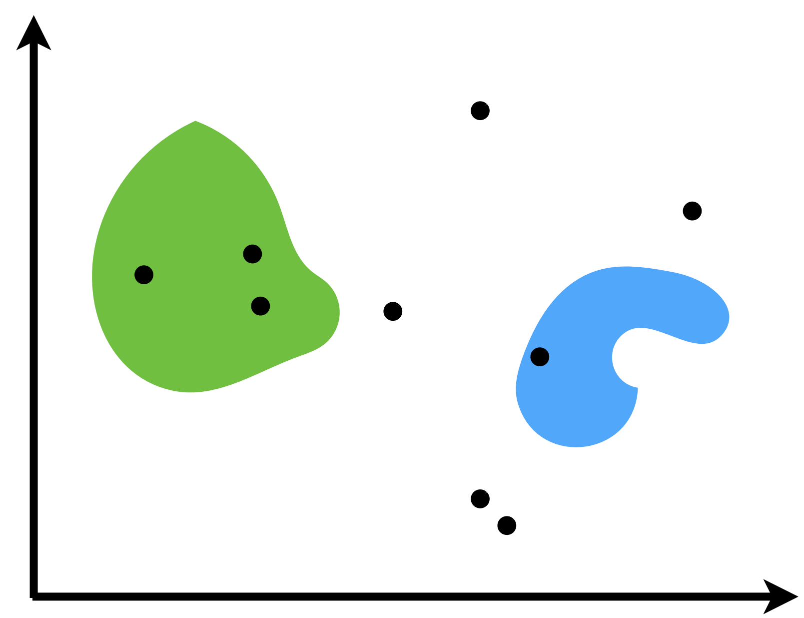
- Beta process, Bernoulli process (Indian buffet)
- Gamma process, Poisson likelihood process (DP, CRP)
- Beta process, negative binomial process



Orbanz 2009, Orbanz 2010

Conjugacy & Poisson point processes

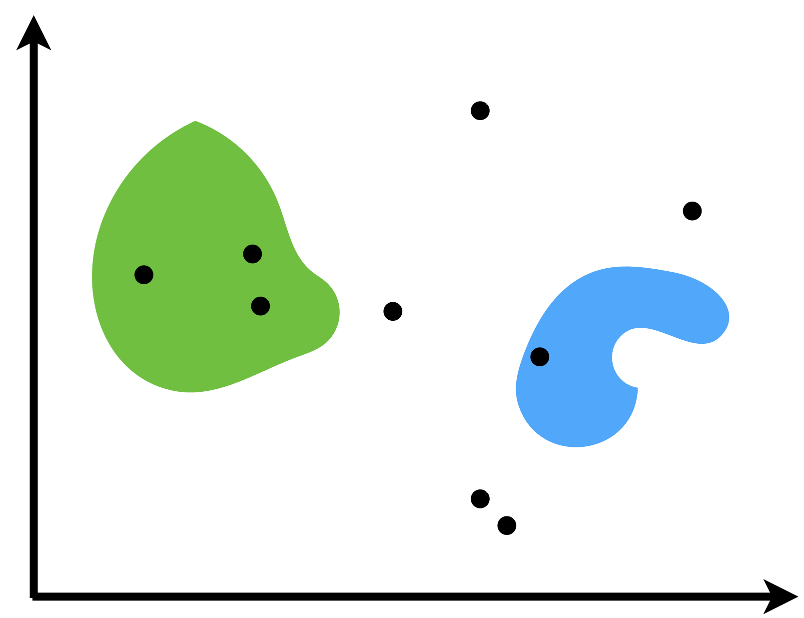
- Beta process, Bernoulli process (Indian buffet)
- Gamma process, Poisson likelihood process (DP, CRP)
- Beta process, negative binomial process



[Kingman 1992, Orbanz 2009, Orbanz 2010

Conjugacy & Poisson point processes

- Beta process, Bernoulli process (Indian buffet)
- Gamma process, Poisson likelihood process (DP, CRP)
- Beta process, negative binomial process

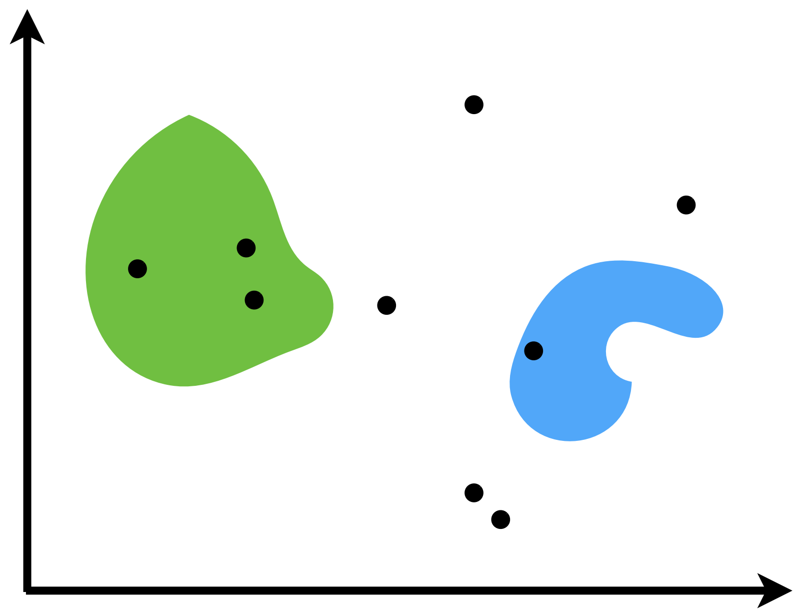


- Posteriors, conjugacy, and exponential families for completely random measures

[Kingman 1992, Orbanz 2009, Orbanz 2010

Conjugacy & Poisson point processes

- Beta process, Bernoulli process (Indian buffet)
- Gamma process, Poisson likelihood process (DP, CRP)
- Beta process, negative binomial process



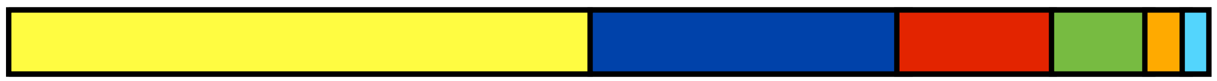
- Posteriors, conjugacy, and exponential families for completely random measures

[Kingman 1992, Orbanz 2009, Orbanz 2010, Broderick et al 2014, James 2014]

De Finetti mixing measures

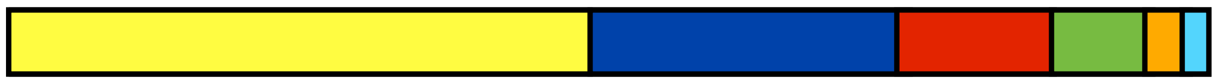
De Finetti mixing measures

- Clustering: Kingman paintbox



De Finetti mixing measures

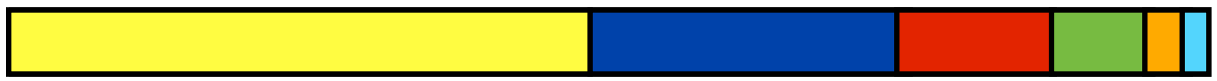
- Clustering: Kingman paintbox



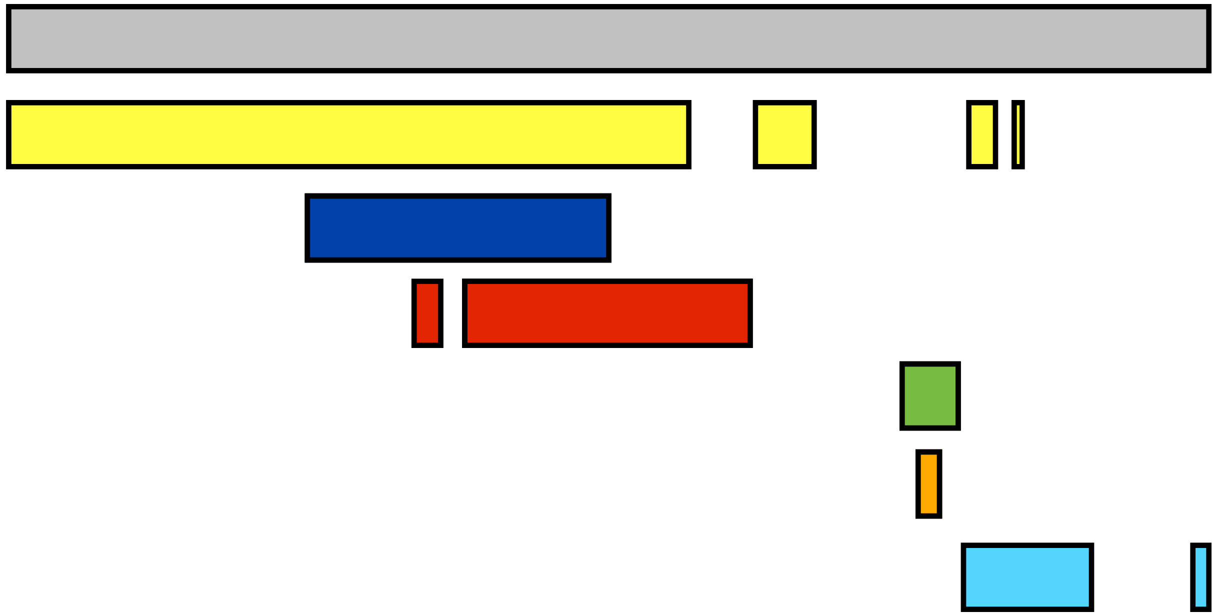
[Kingman 1978]

De Finetti mixing measures

- Clustering: Kingman paintbox



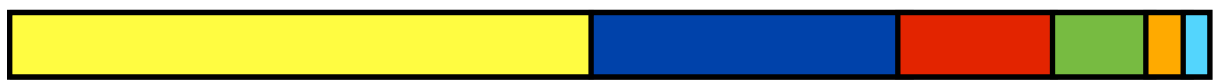
- Feature allocation: Feature paintbox



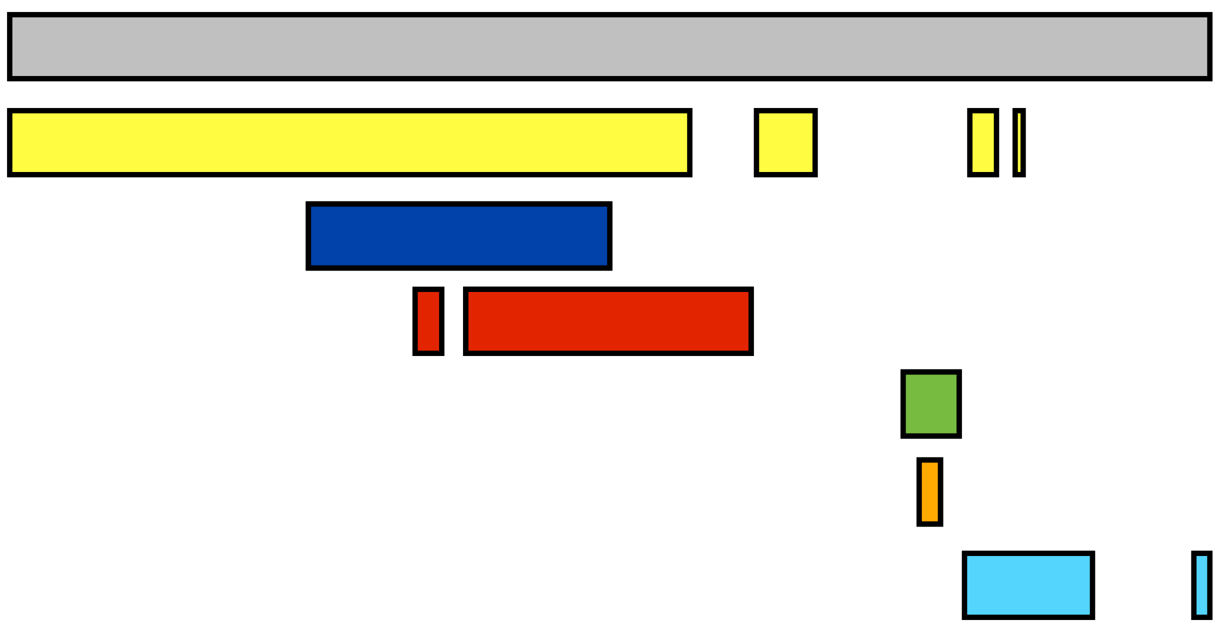
[Kingman 1978]

De Finetti mixing measures

- Clustering: Kingman paintbox

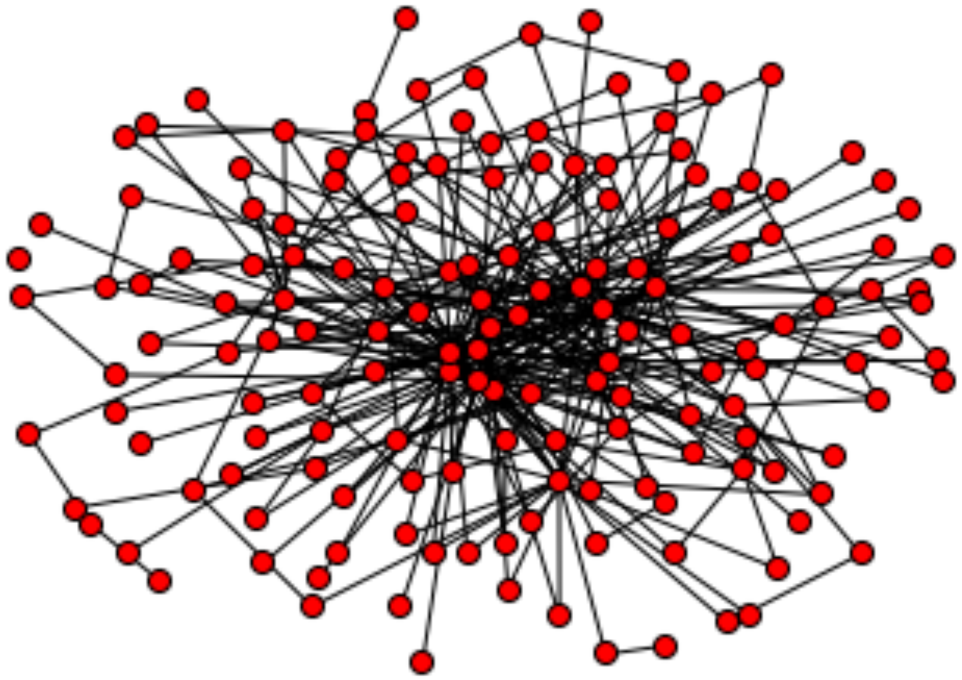


- Feature allocation: Feature paintbox



[Kingman 1978, Broderick, Pitman, Jordan 2013]

Probabilistic models for graphs

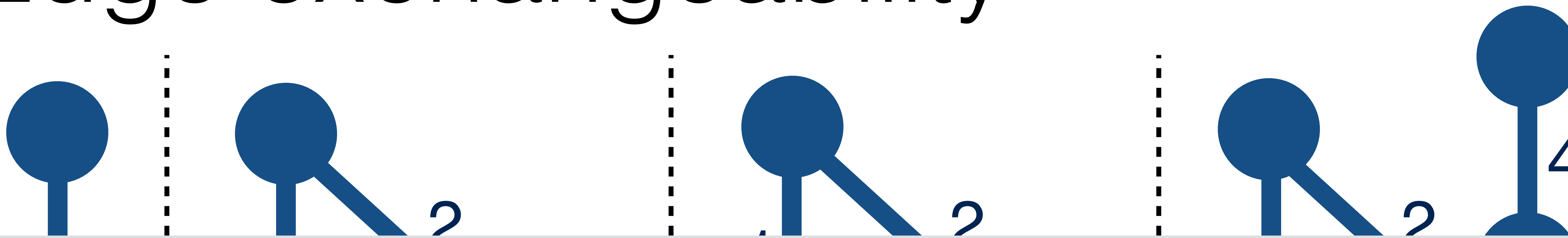
$p($  $)$

E.g. online social networks, biological networks, communication networks, transportation networks

- Rich relationships, coherent uncertainties, prior info
- Stochastic block model, mixed membership stochastic block model, infinite relational model, and many more
- Assume: Adding more data doesn't change distribution of earlier data (*projectivity*)
- **Problem**: model misspecification, dense graphs

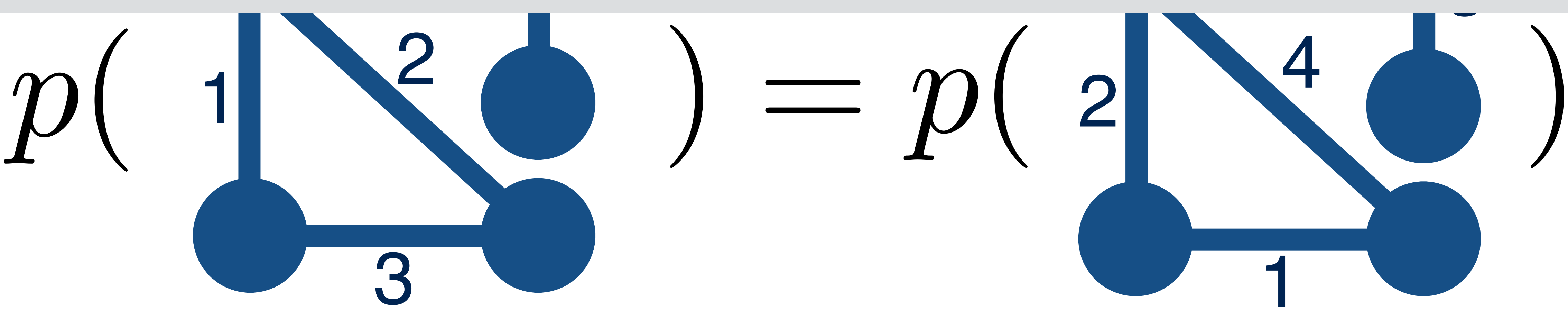
[Holland et al 1983; Kemp et al 2006; Xu et al 2007; Airoldi et al 2008; Lloyd et al 2012]

Edge exchangeability



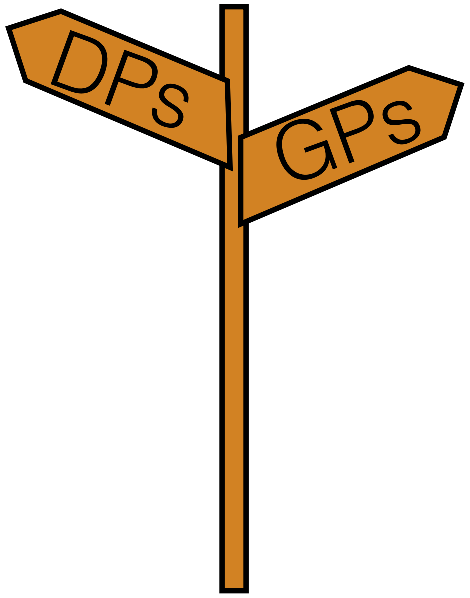
Thm. A wide range of edge-exchangeable graph sequences are sparse

Thm. A paintbox-style characterization for edge-exchangeable graph sequences



[Broderick, Cai 2015; Crane, Dempsey 2015; Crane, Dempsey 2016; Cai, Campbell, Broderick 2016]

Roadmap

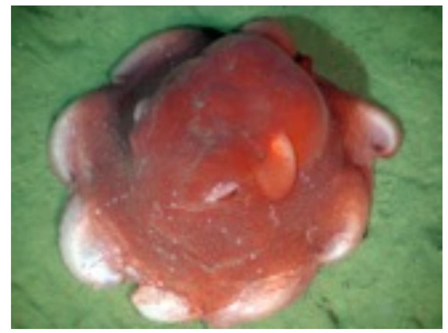


- Bayes Foundations
- Unsupervised Learning
  - Example problem: clustering
  - Example BNP model: Dirichlet process (DP)
  - Chinese restaurant process
- Supervised Learning
  - Example problem: regression
  - Example BNP model: Gaussian process (GP)
- Venture further into the wild world of Nonparametric Bayes

- Big questions
  - Why BNP?
  - What does an infinite/growing number of parameters really mean (in BNP)?
  - Why is BNP challenging but practical?

Applications

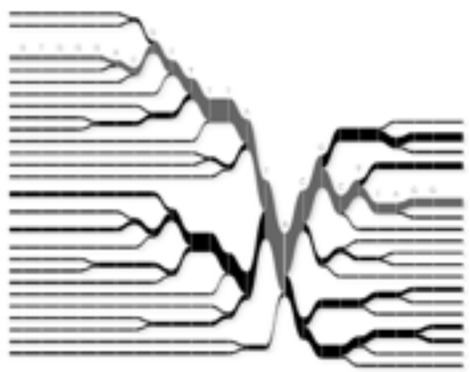
[wikipedia.org]



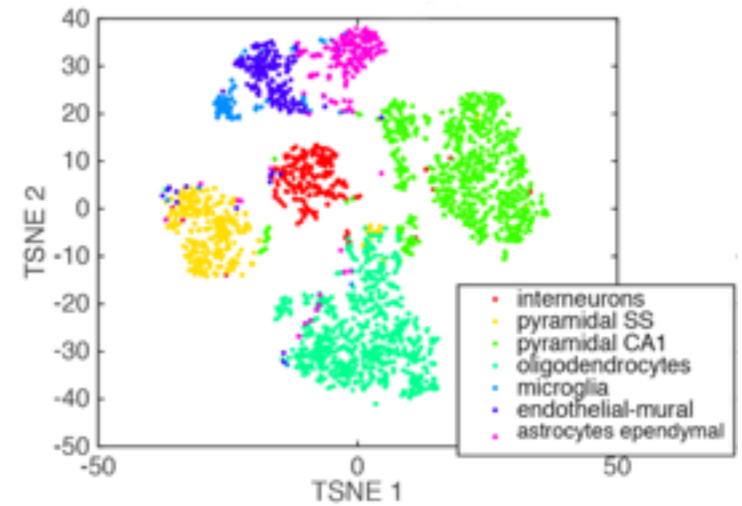
[Ed Bowlby, NOAA]

[Fox et al 2014]

[Lloyd et al 2012; Miller et al 2010]



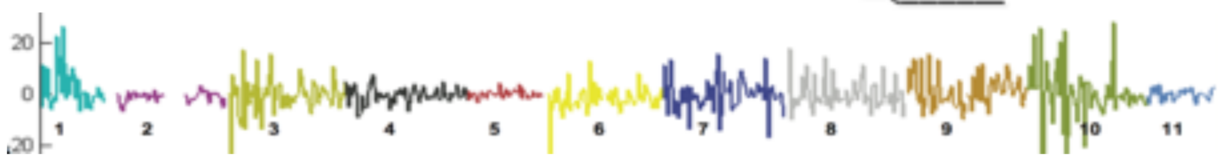
[Ewens 1972; Hartl, Clark 2003]



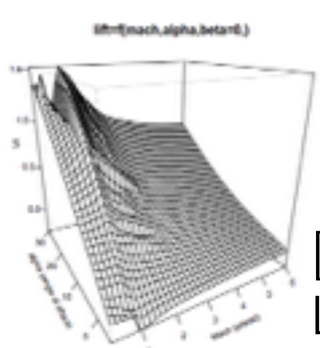
[Prabhakaran, Azizi, Carr, Pe'er 2016]

[Sudderth, Jordan 2009]

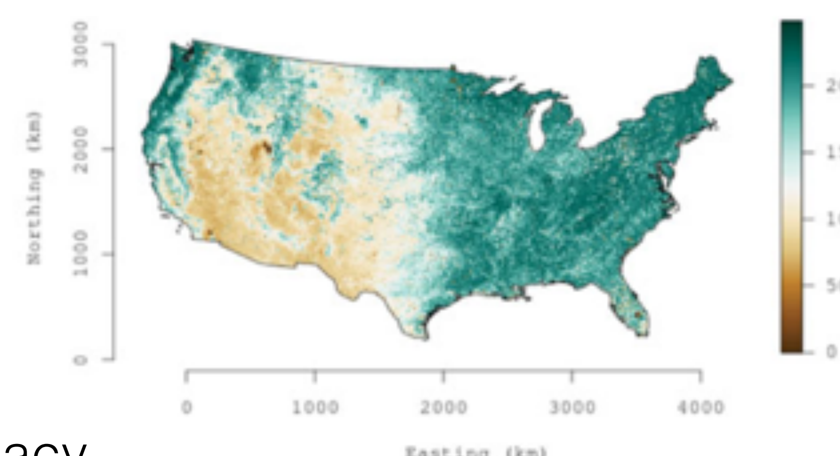
[Saria et al 2010]



[US CDC PHIL; Futoma, Hariharan, Heller 2017]

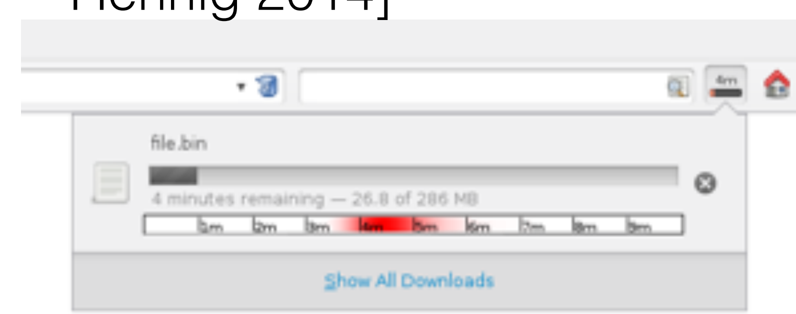


[Gramacy, Lee 2009]



[Datta, Banerjee, Finley, Gelfand 2016]

[Kiefel, Schuler, Hennig 2014]



[Chati, Balakrishnan 2017]

[Deisenroth, Fox, Rasmussen 2015]