YAHOO!®

Graphical Models I

Alexander Smola, Yahoo! Labs

smola@yahoo-inc.com, alex.smola.org, yim:alex.smola, skype:smolix

Statistics 101

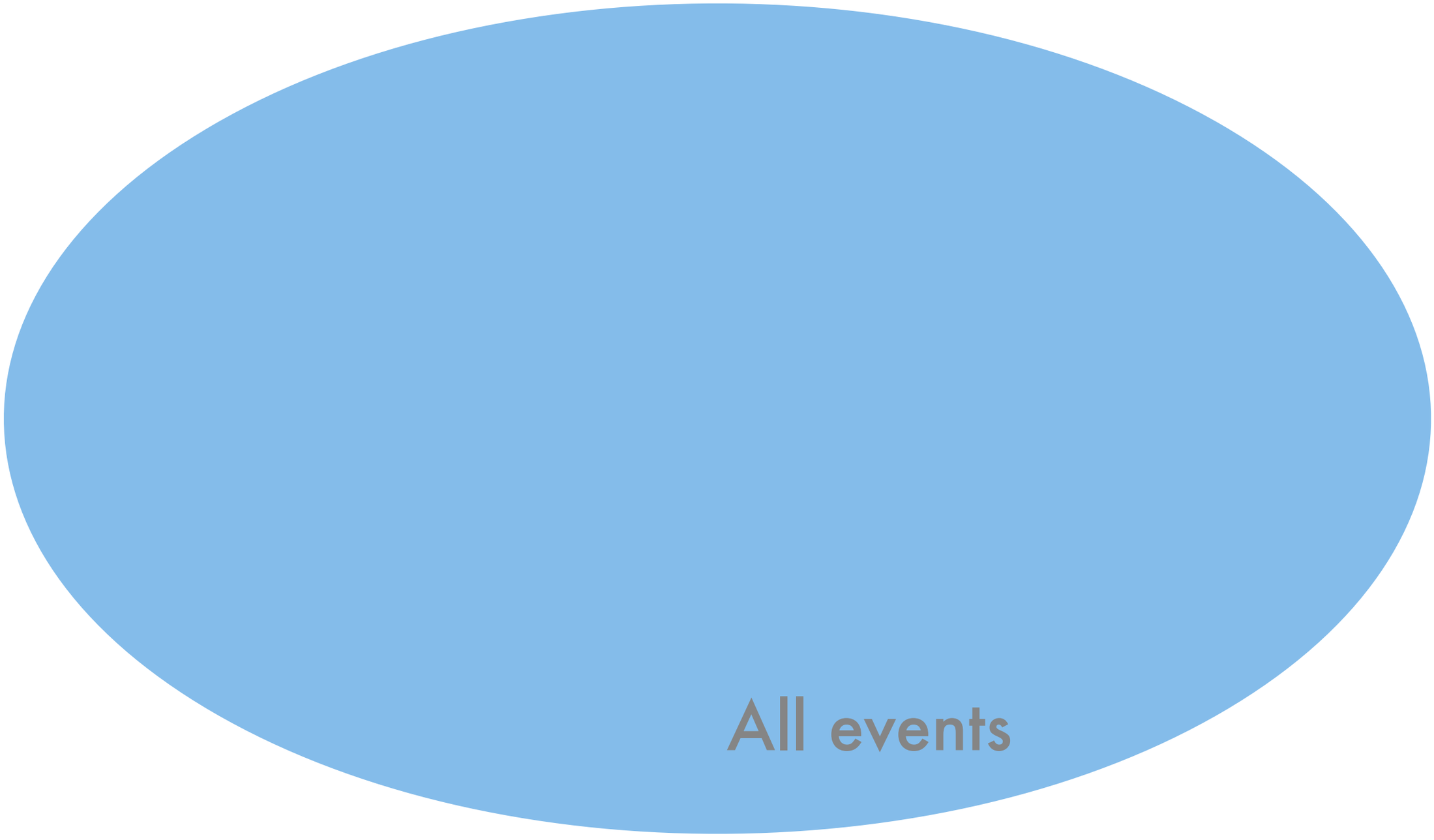
Probability

- Space of events X
  - server working; slow response; server broken
  - income of the user (e.g. $95,000)
  - query text for search (e.g. "statistics tutorial")
- Probability axioms (Kolmogorov)

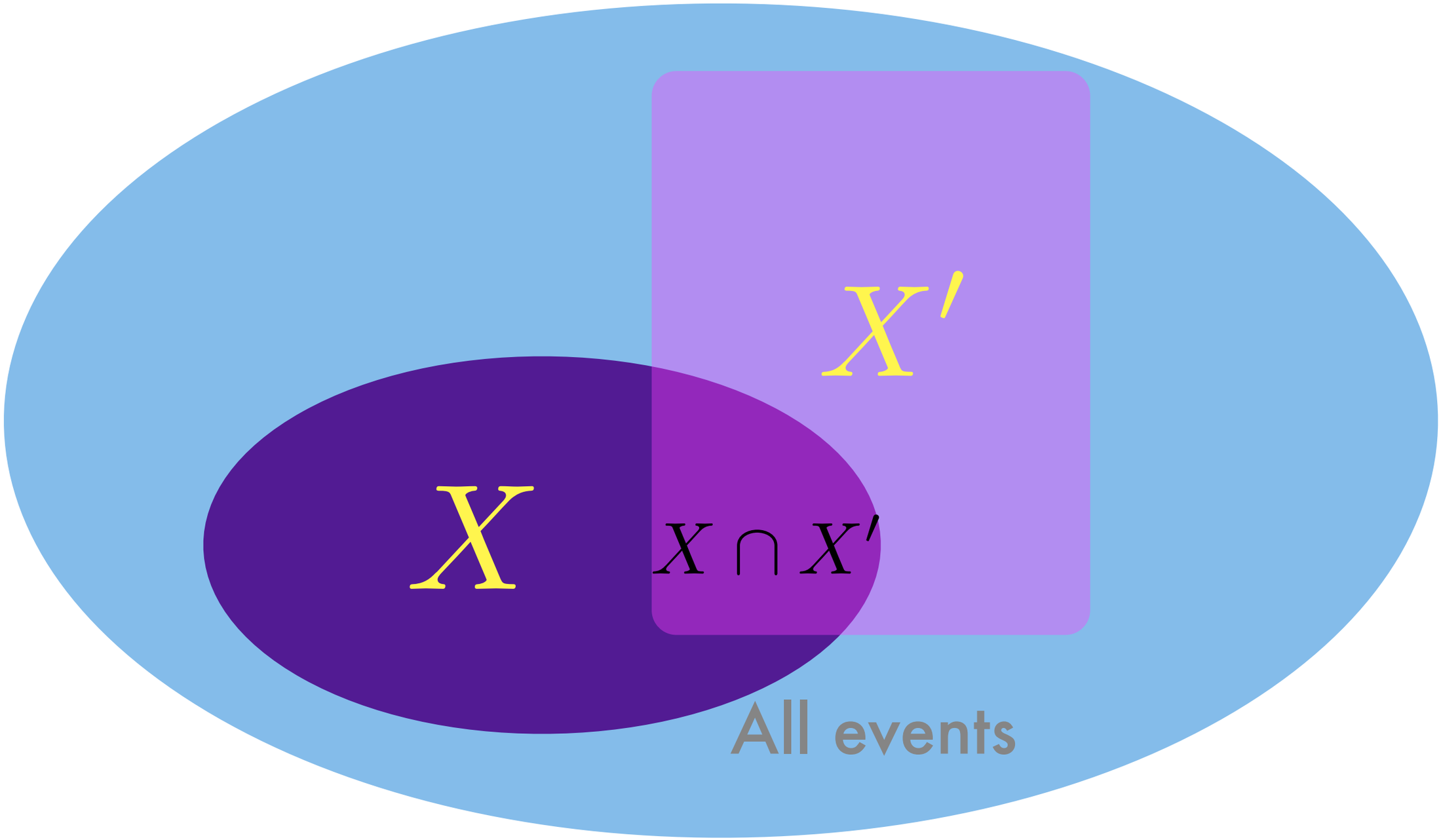
$$\Pr(X) \in [0, 1], \Pr(\mathcal{X}) = 1$$
$$\Pr(\cup_i X_i) = \sum_i \Pr(X_i) \text{ if } X_i \cap X_j = \emptyset$$

- Example queries
  - P(server working) = 0.999
  - P(90,000 < income < 100,000) = 0.1

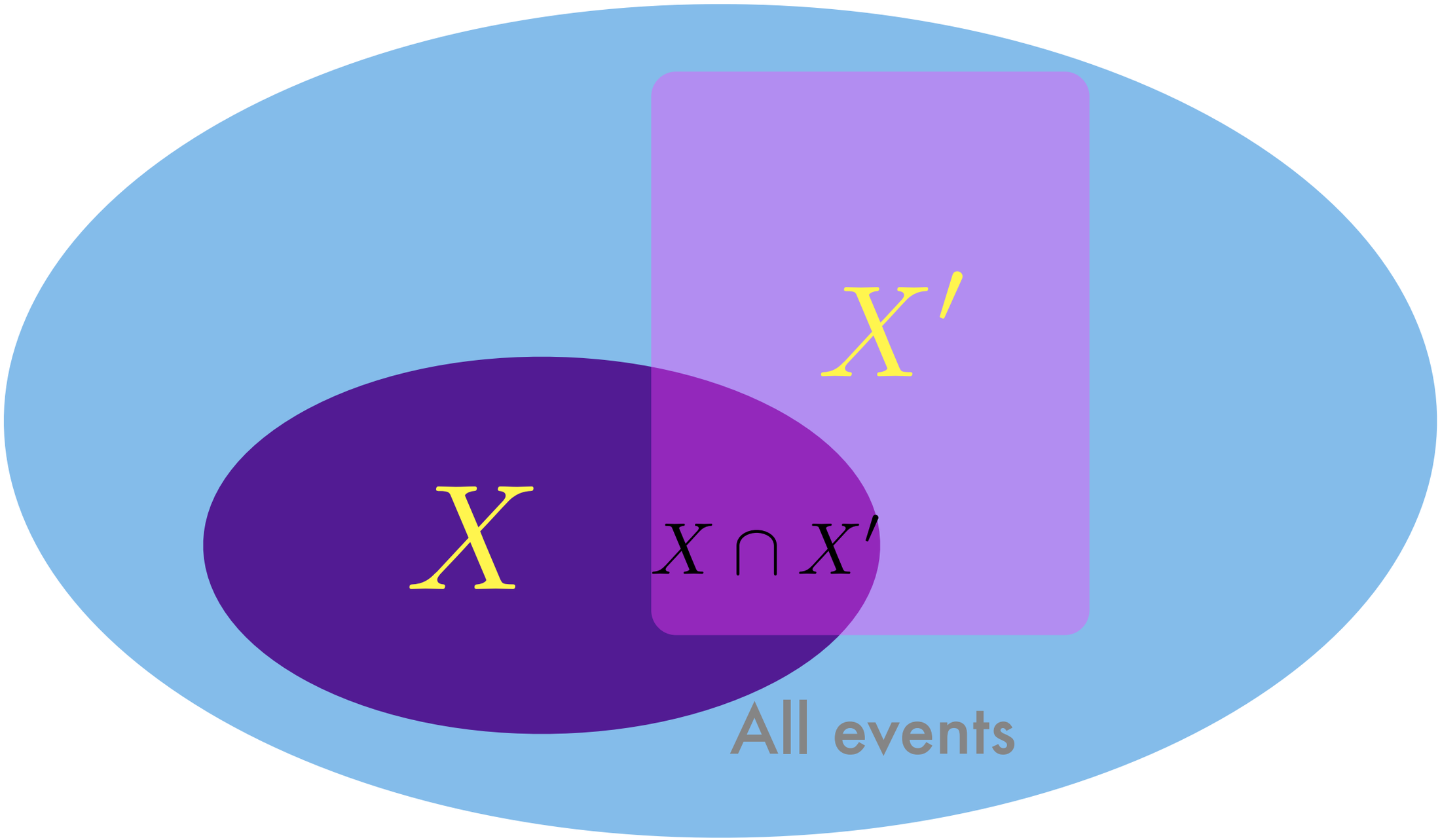
Venn Diagram



Venn Diagram



Venn Diagram



$$\Pr(X \cup X') = \Pr(X) + \Pr(X') - \Pr(X \cap X')$$

(In)dependence

- Independence $\Pr(x, y) = \Pr(x) \cdot \Pr(y)$
  - Login behavior of two users (approximately)
  - Disk crash in different colos (approximately)
- Dependent events $\Pr(x, y) \neq \Pr(x) \cdot \Pr(y)$
  - Emails
  - Queries
  - News stream / Buzz / Twitter feeds
  - IM communication
  - Russian Roulette

(In)dependence

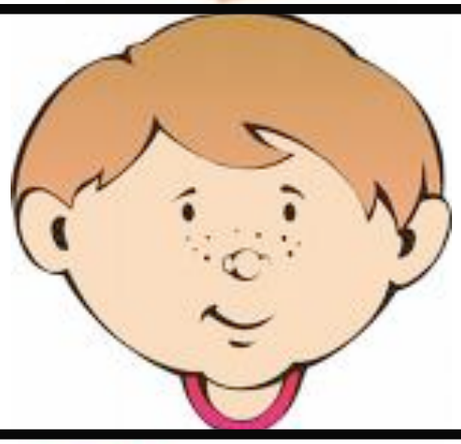
- Independence $\Pr(x, y) = \Pr(x) \cdot \Pr(y)$
  - Login behavior of two users (approximately)
  - Disk crash in different colos (approximately)
- Dependent events
  - Emails
  - Queries
  - News stream / Buzz / T
  - IM communication
  - Russian Roulette

$\Pr(x, y) \neq$ Everywhere $\Pr(x) \cdot \Pr(y)$

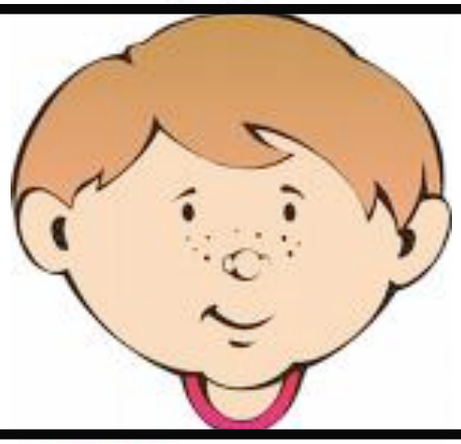
in

Yahoo!

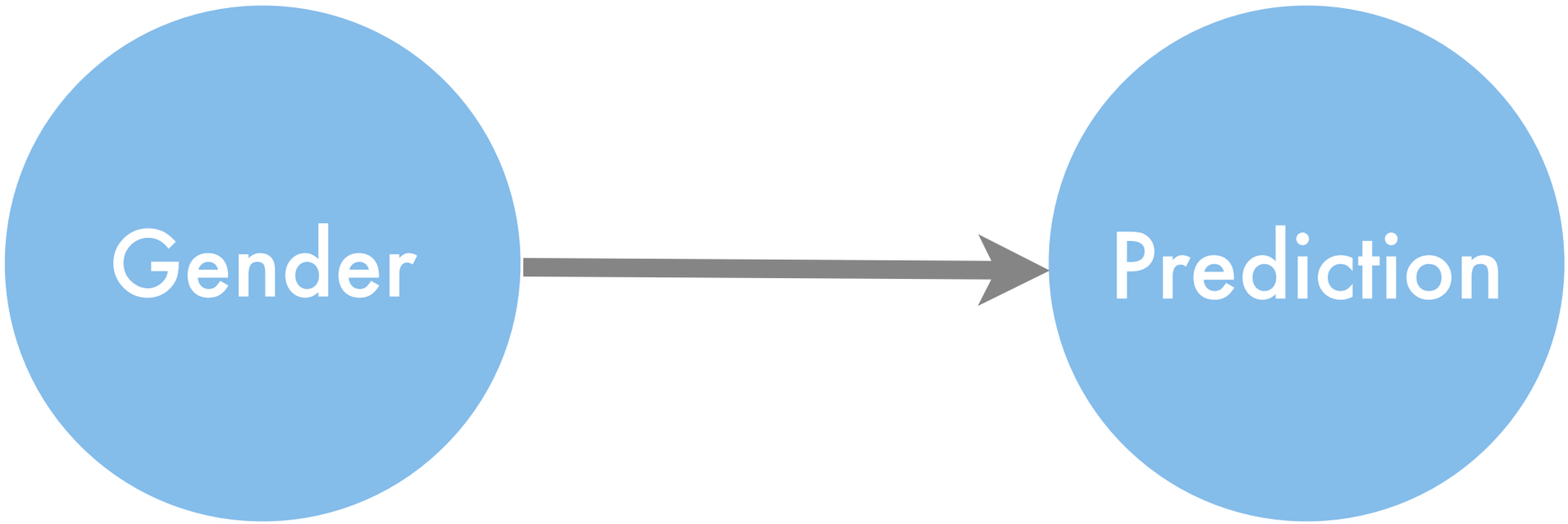
Independence

	0.25	0.25
	0.25	0.25

Dependence

	0.45	0.05
	0.05	0.45

A Graphical Model



$$p(\text{gender}, \text{pred}) = p(\text{gender})p(\text{pred}|\text{gender})$$

Bayes Rule

- Joint Probability

$$\Pr(X, Y) = \Pr(X|Y)\Pr(Y) = \Pr(Y|X)\Pr(X)$$

- Bayes Rule

$$\Pr(X|Y) = \frac{\Pr(Y|X) \cdot \Pr(X)}{\Pr(Y)}$$

  - Hypothesis testing
  - Reverse conditioning

AIDS test (Bayes rule)

- Data
  - Approximately 0.1% are infected
  - Test detects all infections
  - Test reports positive for 1% healthy people
- Probability of having AIDS if test is positive

$$\Pr(a=1|t) = \frac{\Pr(t|a=1)\cdot\Pr(a=1)}{\Pr(t)}$$

$$= \frac{\Pr(t|a=1)\cdot\Pr(a=1)}{\Pr(t|a=1)\cdot\Pr(a=1)+\Pr(t|a=0)\cdot\Pr(a=0)}$$

$$= \frac{1\cdot 0.001}{1\cdot 0.001 + 0.01\cdot 0.999} = 0.091$$

Improving the diagnosis

- Use a follow-up test
  - Test 2 reports positive for 90% infections
  - Test 2 reports positive for 5% healthy people

$$\frac{0.01 \cdot 0.05 \cdot 0.999}{1 \cdot 0.9 \cdot 0.001 + 0.01 \cdot 0.05 \cdot 0.999} = 0.357$$

- Why can't we use Test 1 twice?
  Outcomes are **not** independent but tests 1 and 2 are **conditionally independent**

$$p(t_1, t_2|a) = p(t_1|a) \cdot p(t_2|a)$$

Application: Naive Bayes

Naive Bayes Spam Filter

- Key assumption
  Words occur independently of each other given the label of the document

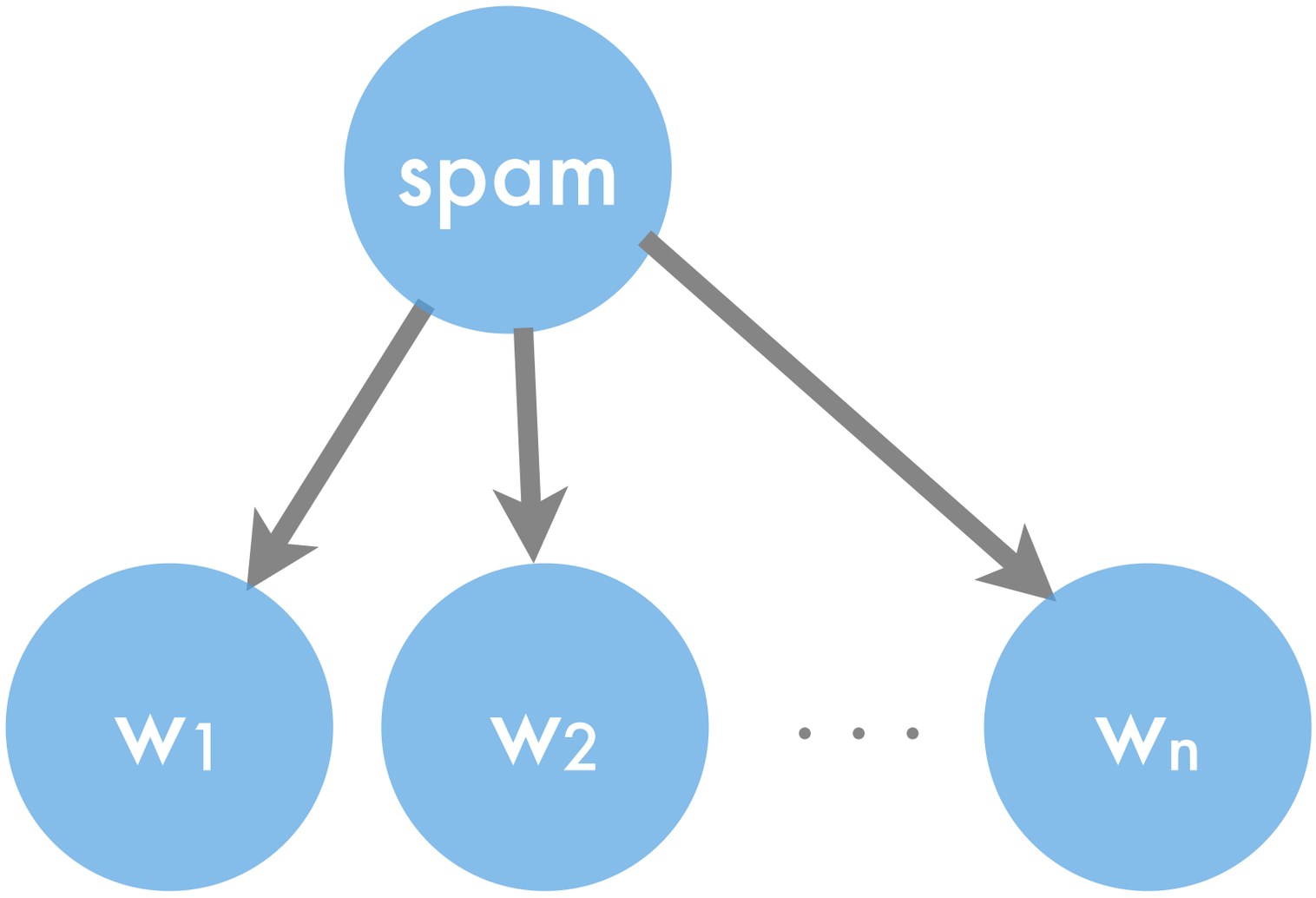
$$p(w_1, \ldots, w_n|\text{spam}) = \prod_{i=1}^{n} p(w_i|\text{spam})$$

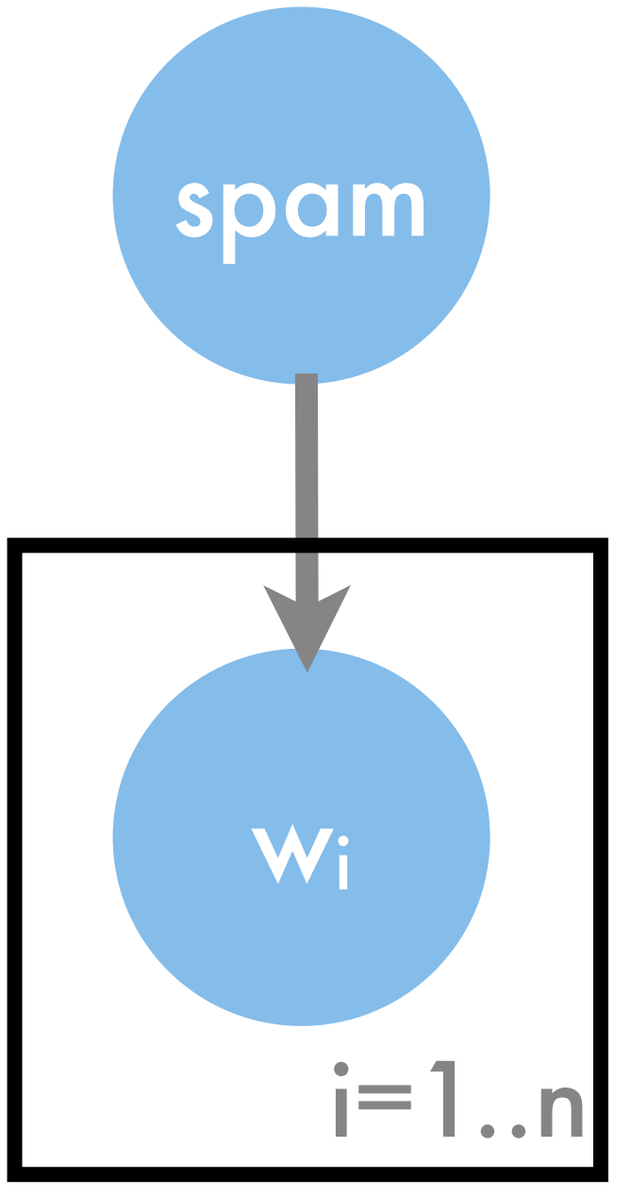
- Spam classification via Bayes Rule

$$p(\text{spam}|w_1, \ldots, w_n) \propto p(\text{spam}) \prod_{i=1}^{n} p(w_i|\text{spam})$$

- Parameter estimation
  Compute spam probability and word distributions for spam and ham

A Graphical Model





$$p(w_1, \ldots, w_n|\text{spam}) = \prod_{i=1}^{n} p(w_i|\text{spam})$$

Estimating Probabilities

Binomial Distribution

- Two outcomes (head, tail); (0,1)
- Data likelihood

$$p(X;\pi) = \pi^{n_1}(1-\pi)^{n_0}$$

- Maximum Likelihood Estimation
- Constrained optimization problem $\pi \in [0,1]$
- Incorporate constraint via $p(x;\theta) = \frac{e^{x\theta}}{1+e^{\theta}}$
- Taking derivatives yields

$$\theta = \log \frac{n_1}{n_0+n_1} \Longleftrightarrow p(x=1) = \frac{n_1}{n_0+n_1}$$

... in detail ...

$$p(X;\theta) = \prod_{i=1}^{n} p(x_i;\theta) = \prod_{i=1}^{n} \frac{e^{\theta x_i}}{1+e^{\theta}}$$

$$\Longrightarrow \log p(X;\theta) = \theta \sum_{i=1}^{n} x_i - n \log\left[1+e^{\theta}\right]$$

$$\Longrightarrow \partial_\theta \log p(X;\theta) = \sum_{i=1}^{n} x_i - n\frac{e^{\theta}}{1+e^{\theta}}$$

$$\Longleftrightarrow \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{e^{\theta}}{1+e^{\theta}} = p(x=1)$$

… in detail …

$$p(X;\theta) = \prod_{i=1}^{n} p(x_i;\theta) = \prod_{i=1}^{n} \frac{e^{\theta x_i}}{1+e^{\theta}}$$

$$\Longrightarrow \log p(X;\theta) = \theta \sum_{i=1}^{n} x_i - n \log\left[1+e^{\theta}\right]$$

$$\Longrightarrow \partial_\theta \log p(X;\theta) = \sum_{i=1}^{n} x_i - n \frac{e^{\theta}}{1+e^{\theta}}$$

$$\Longleftrightarrow \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{e^{\theta}}{1+e^{\theta}} = p(x=1)$$

empirical probability of x=1

Discrete Distribution

- n outcomes (e.g. USA, Canada, India, UK, NZ)
- Data likelihood

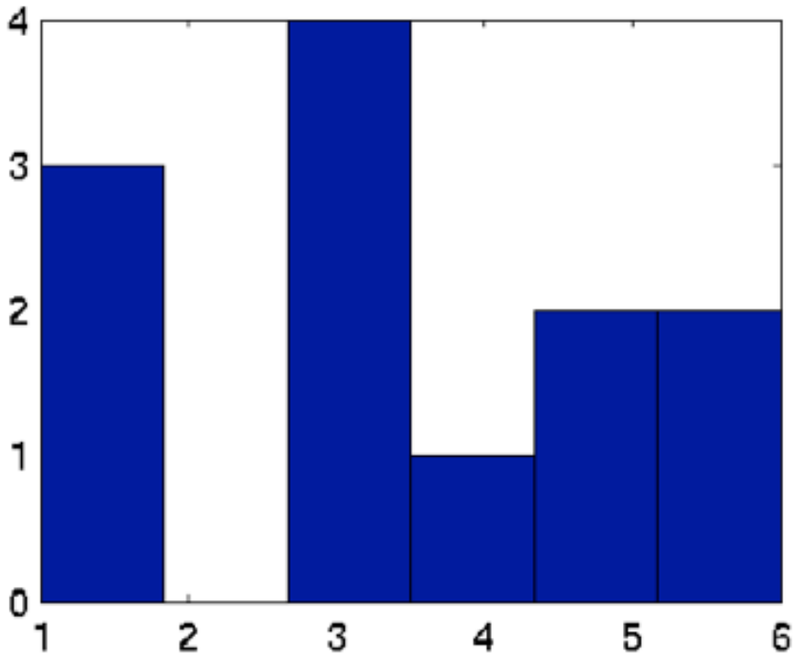
$$p(X;\pi) = \prod_i \pi_i^{n_i}$$

- Maximum Likelihood Estimation
  - Constrained optimization problem ... or ...
  - Incorporate constraint via $p(x;\theta) = \frac{\exp \theta_x}{\sum_{x'} \exp \theta_{x'}}$
  - Taking derivatives yields

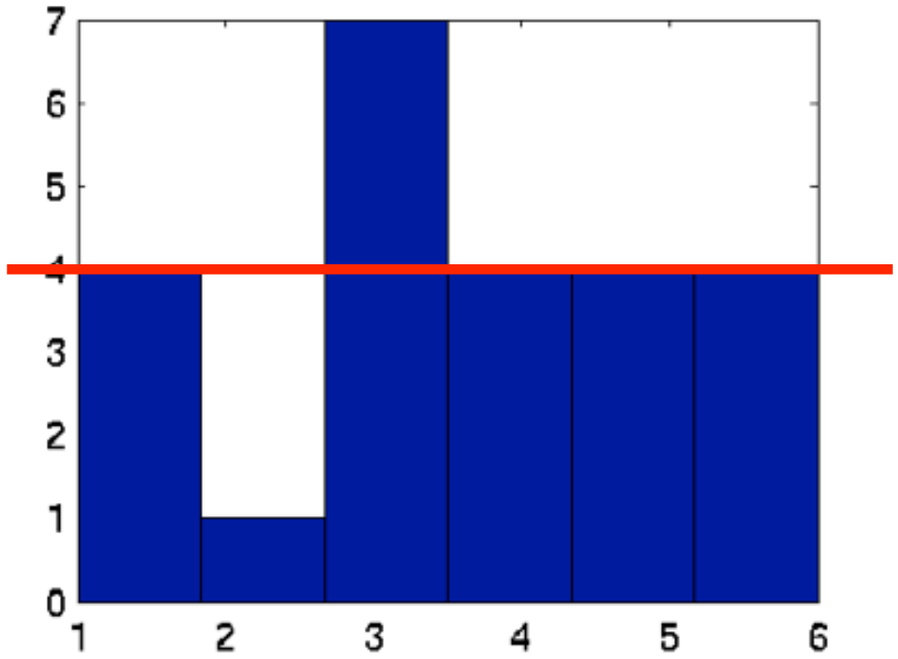
$$\theta_i = \log \frac{n_i}{\sum_j n_j} \iff p(x=i) = \frac{n_i}{\sum_j n_j}$$

Tossing a Dice

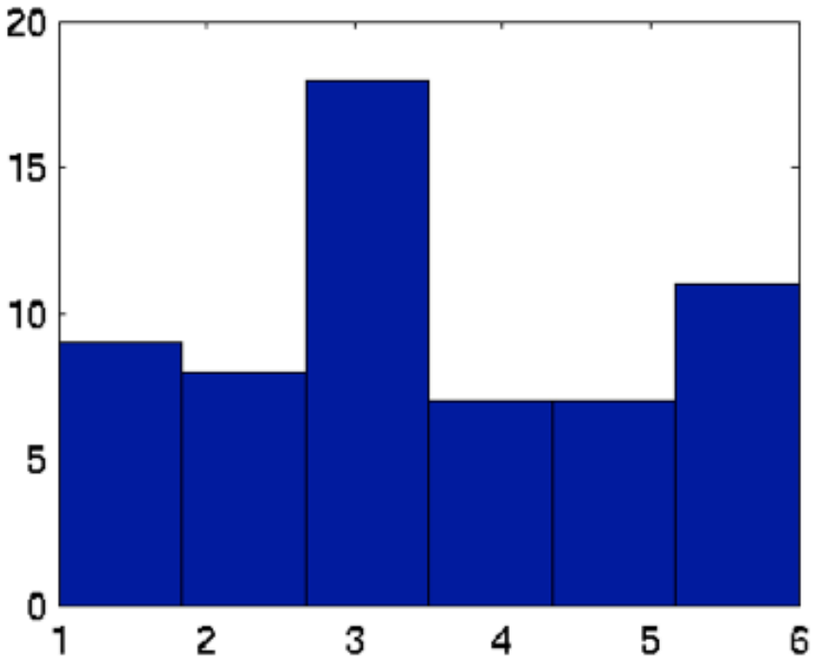
12



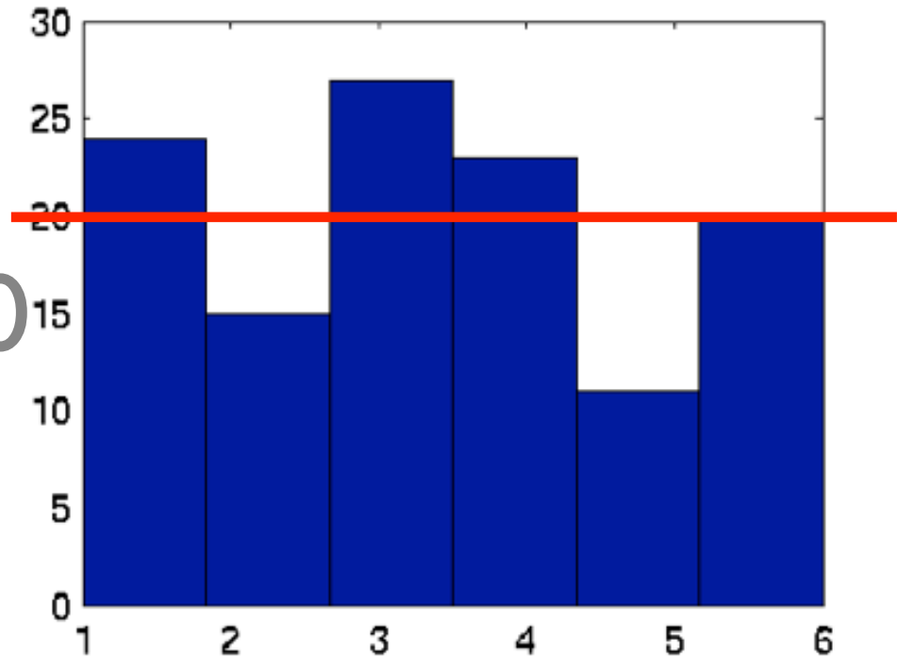
24



60



120



Exponential Families

- Density function

$$p(x;\theta) = \exp\left(\langle \phi(x), \theta \rangle - g(\theta)\right)$$

$$\text{where } g(\theta) = \log \sum_{x'} \exp\left(\langle \phi(x'), \theta \rangle\right)$$

- Log partition function generates cumulants

$$\partial_\theta g(\theta) = \mathbf{E}\left[\phi(x)\right]$$

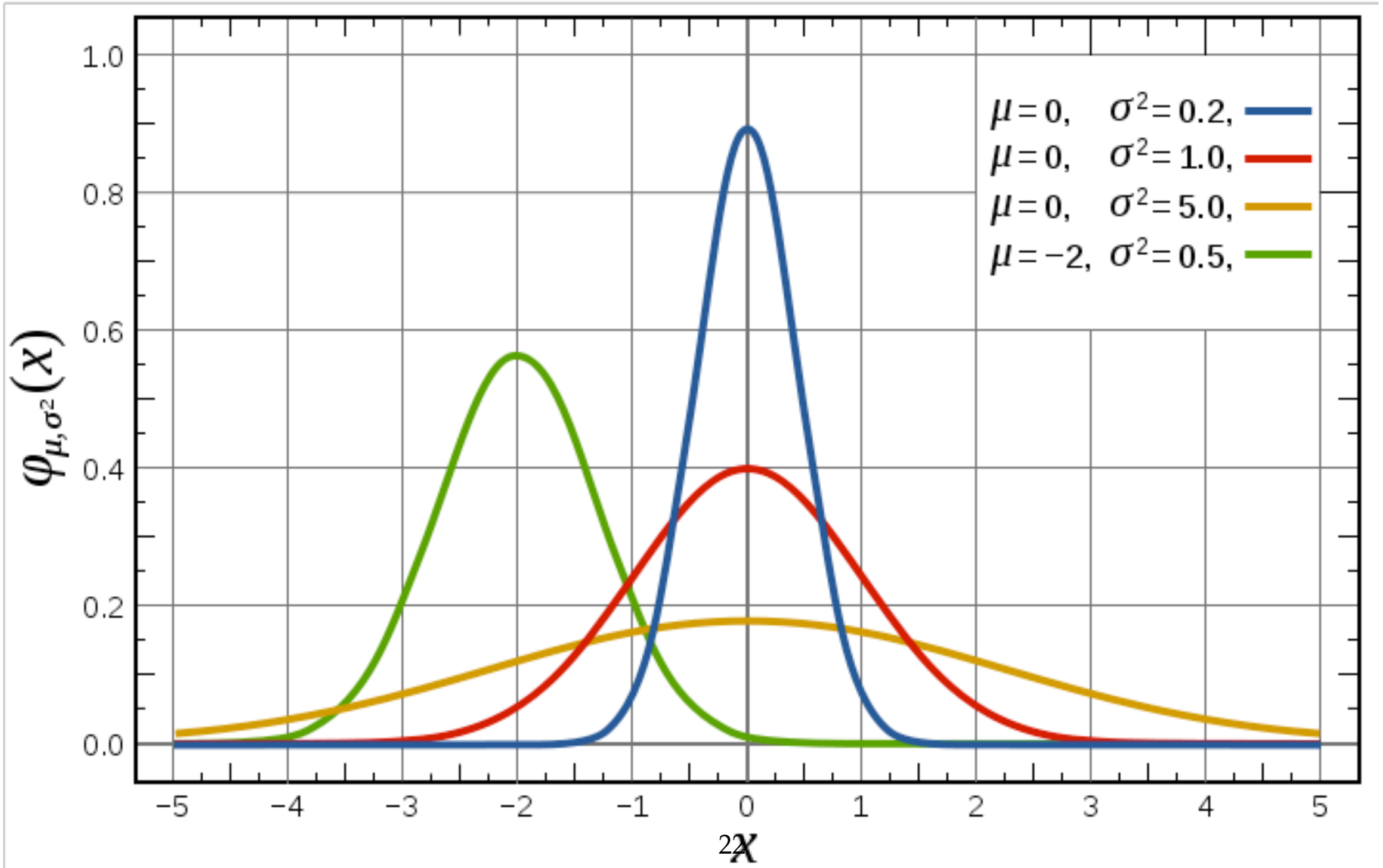
$$\partial_\theta^2 g(\theta) = \mathrm{Var}\left[\phi(x)\right]$$

- g is convex (second derivative is p.s.d.)

Examples

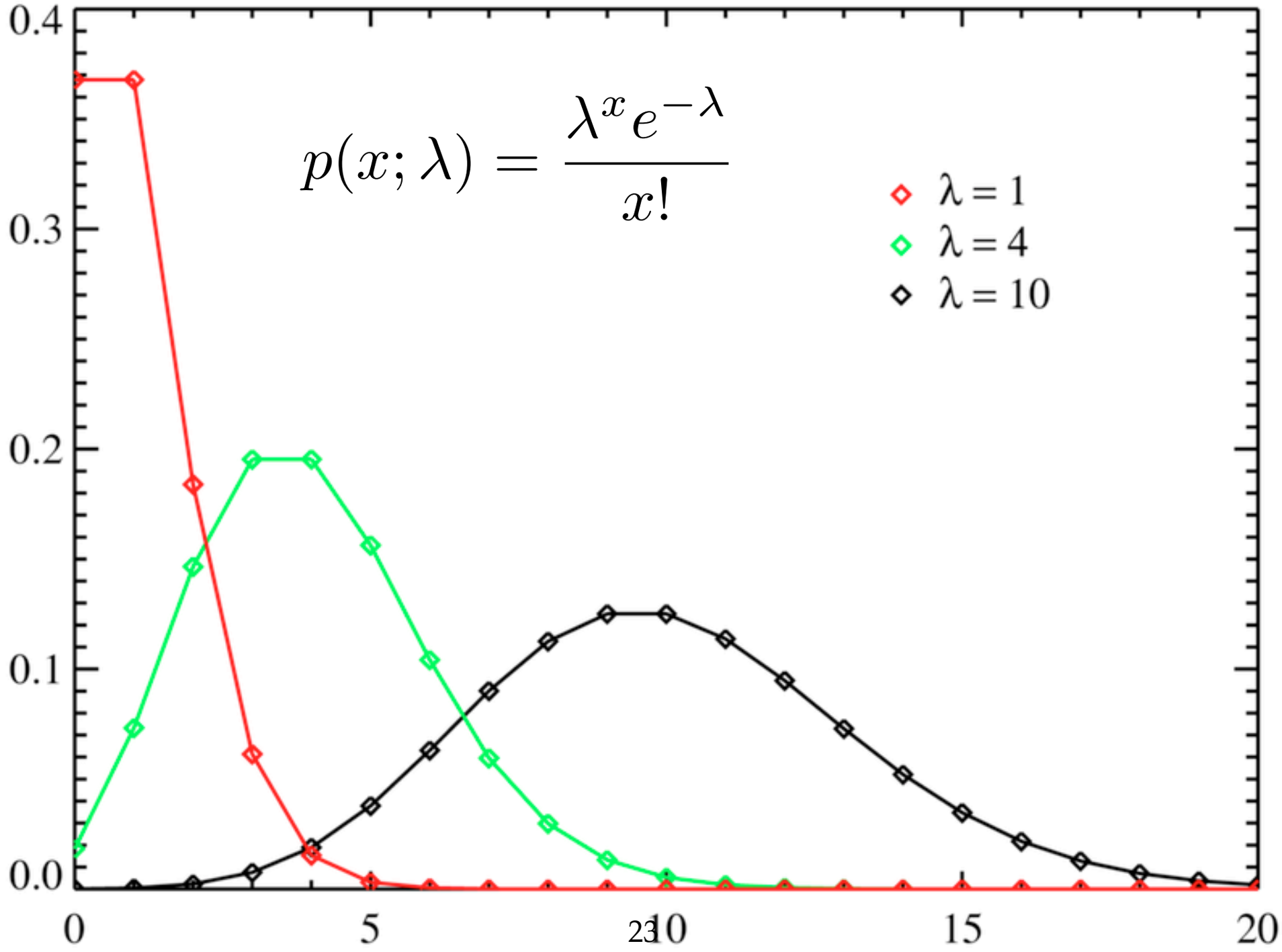
- Binomial Distribution $\phi(x) = x$
- Discrete Distribution $\phi(x) = e_x$
  (e_x is unit vector for x)
- Gaussian $\phi(x) = \left(x, \frac{1}{2}xx^\top\right)$
- Poisson (counting measure 1/x!) $\phi(x) = x$
- Dirichlet, Beta, Gamma, Wishart, ...

Normal Distribution

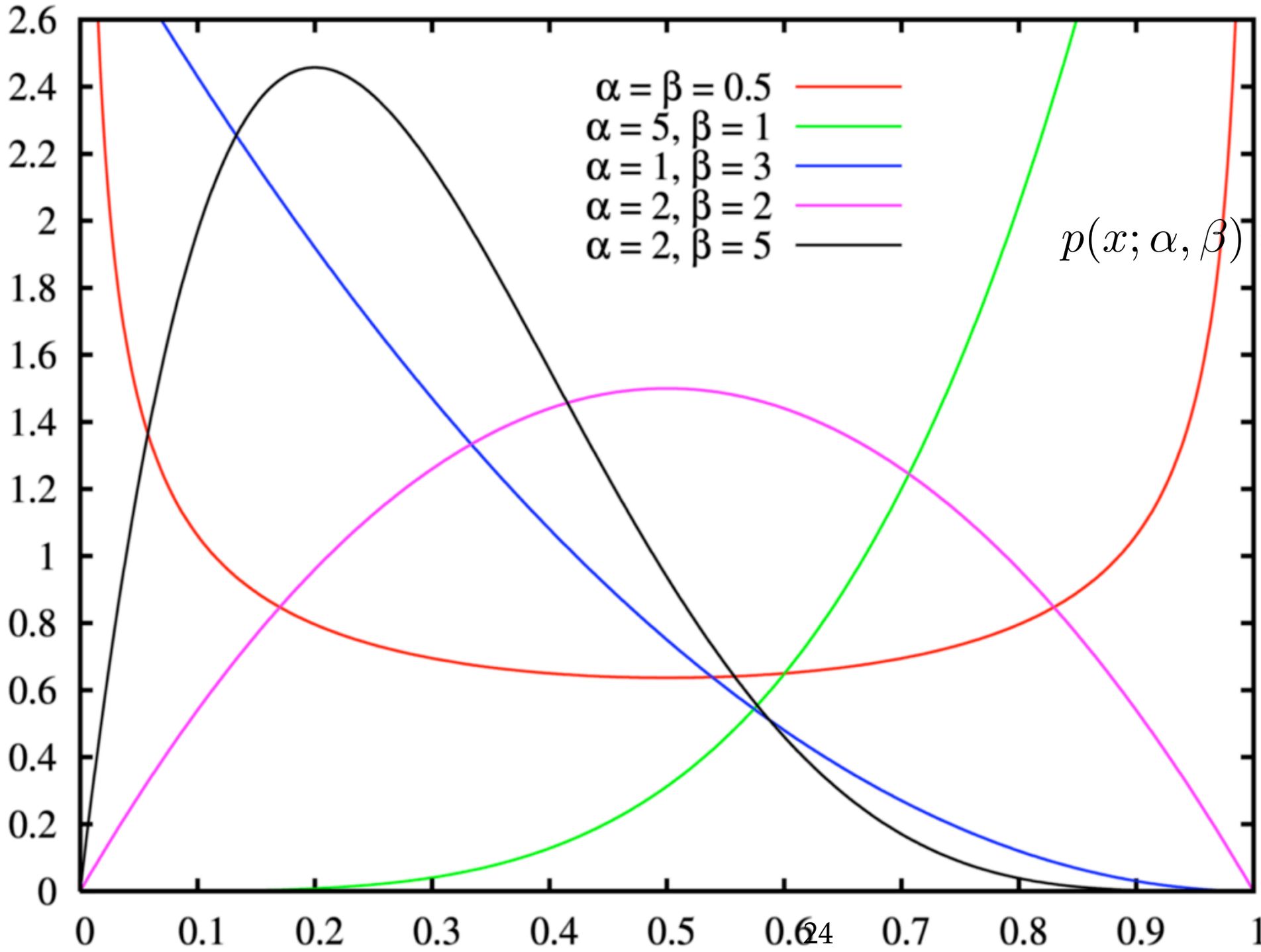


Poisson Distribution

$$p(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

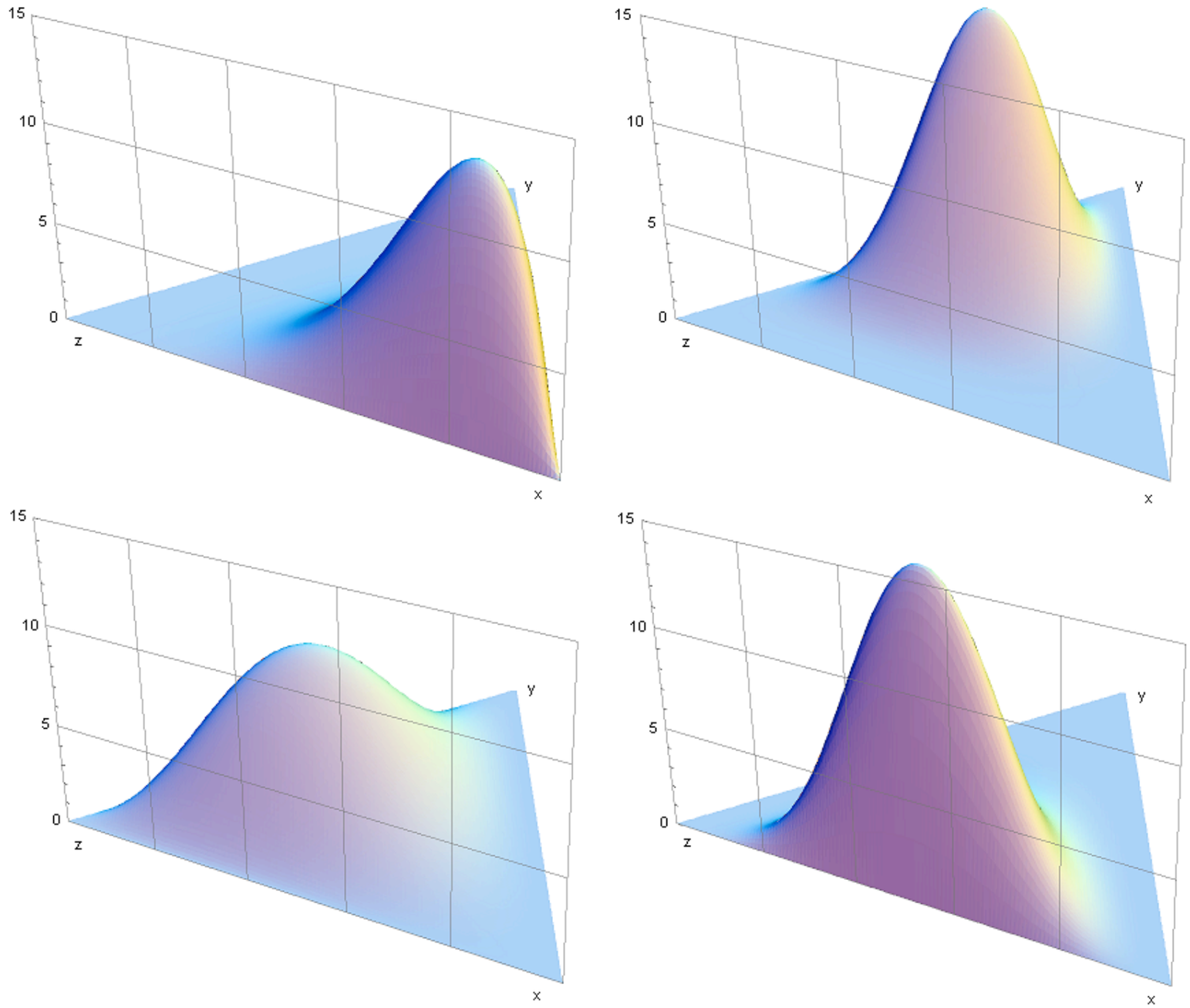


Beta Distribution



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

Dirichlet Distribution



... this looks like a distribution over distributions ...

Maximum Likelihood

- Negative log-likelihood

$$-\log p(X;\theta) = \sum_{i=1}^{n} g(\theta) - \langle \phi(x_i), \theta \rangle$$

- Taking derivatives

$$-\partial_\theta \log p(X;\theta) = m \left[\mathbf{E}[\phi(x)] - \frac{1}{m} \sum_{i=1}^{n} \phi(x_i) \right]$$

We pick the parameter such that the distribution matches the empirical average.

Priors

- Unless we have lots of data estimates are weak
- Usually we have an idea of what to expect

$$p(\theta|X) \propto p(X|\theta) \cdot p(\theta)$$

- Solution 1: Conjugate prior
  Add 'fake' observations (Laplace smoothing)

$$\frac{1}{n}\sum_{i=1}^{n}\phi(x_i) \longrightarrow \frac{1}{n+m}\sum_{i=1}^{n}\phi(x_i) + \frac{m}{n+m}\mu_0$$

- Solution 2: Parameter Smoothing

$$p(\theta) \propto \exp(-\lambda \|\theta\|_1) \text{ or } p(\theta) \propto \exp(-\lambda \|\theta\|_2^2)$$

Conjugate Prior in action

$$m_i = m \cdot [\mu_0]_i$$

- Discrete Distribution

$$p(x = i) = \frac{n_i}{n} \longrightarrow p(x = i) = \frac{n_i + m_i}{n + m}$$

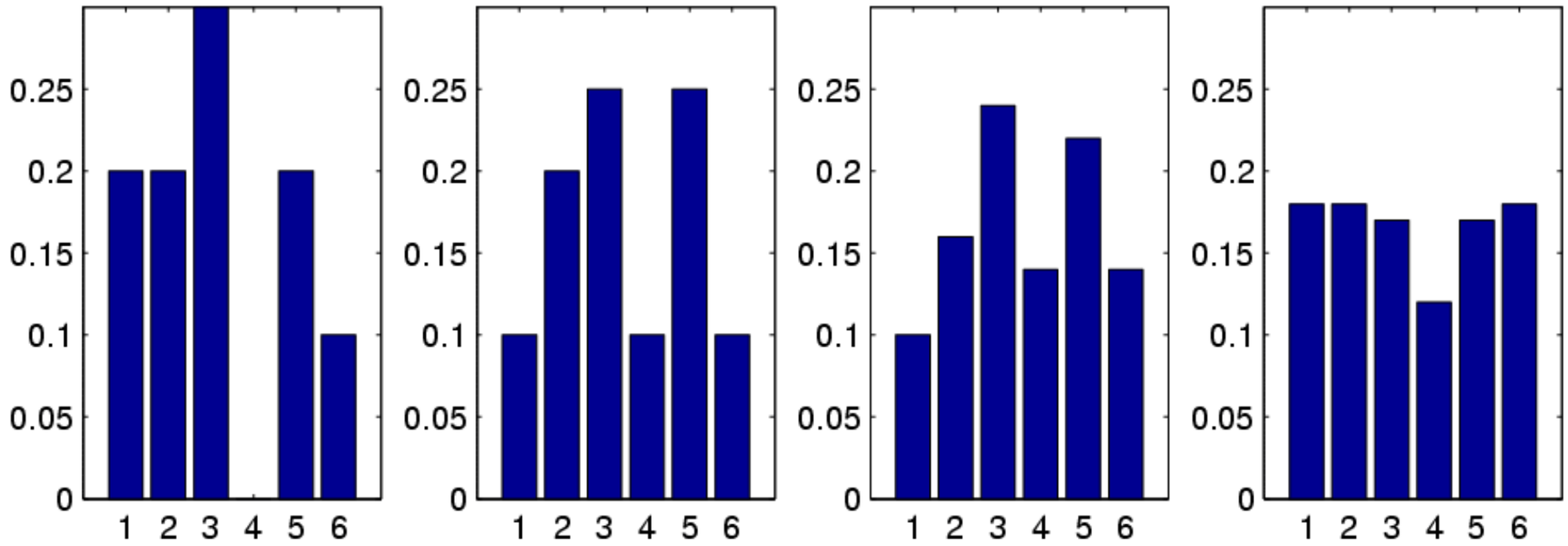
- Tossing a dice

Outcome	1	2	3	4	5	6
Counts	3	6	2	1	4	4
MLE	0.15	0.30	0.10	0.05	0.20	0.20
MAP ($m_0 = 6$)	0.15	0.27	0.12	0.08	0.19	0.19
MAP ($m_0 = 100$)	0.16	0.19	0.16	0.15	0.17	0.17

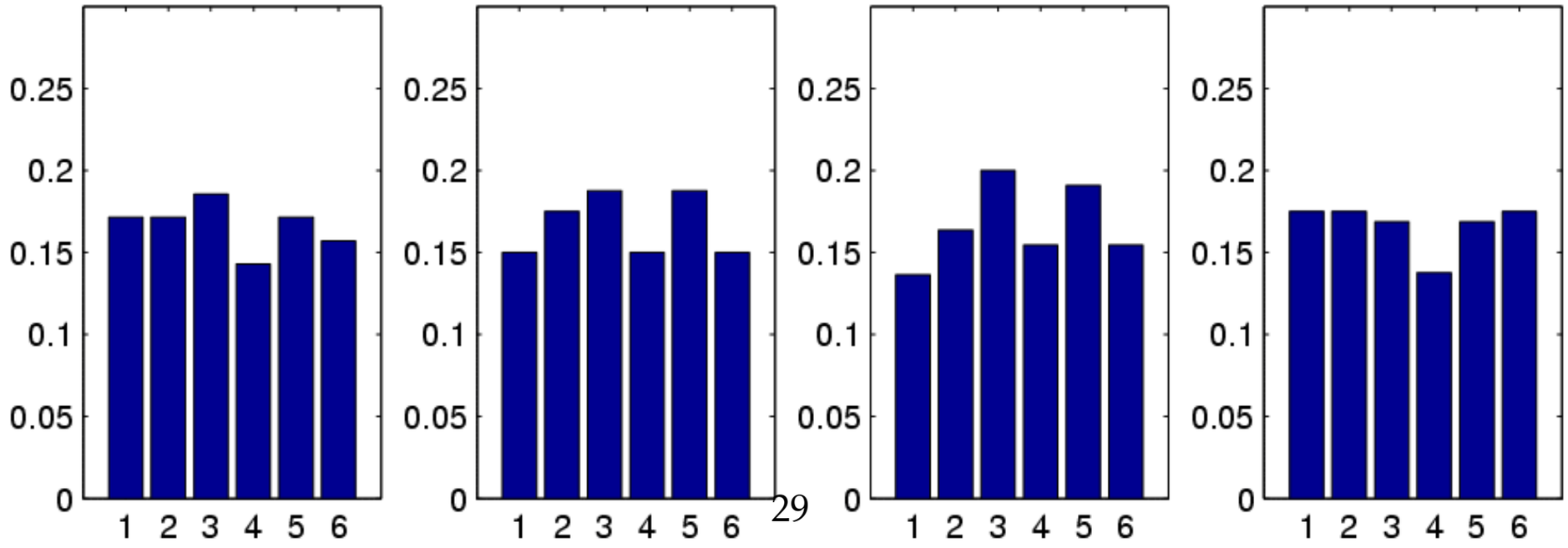
- Rule of thumb
  need 10 data points (or prior) per parameter

Honest dice

MLE

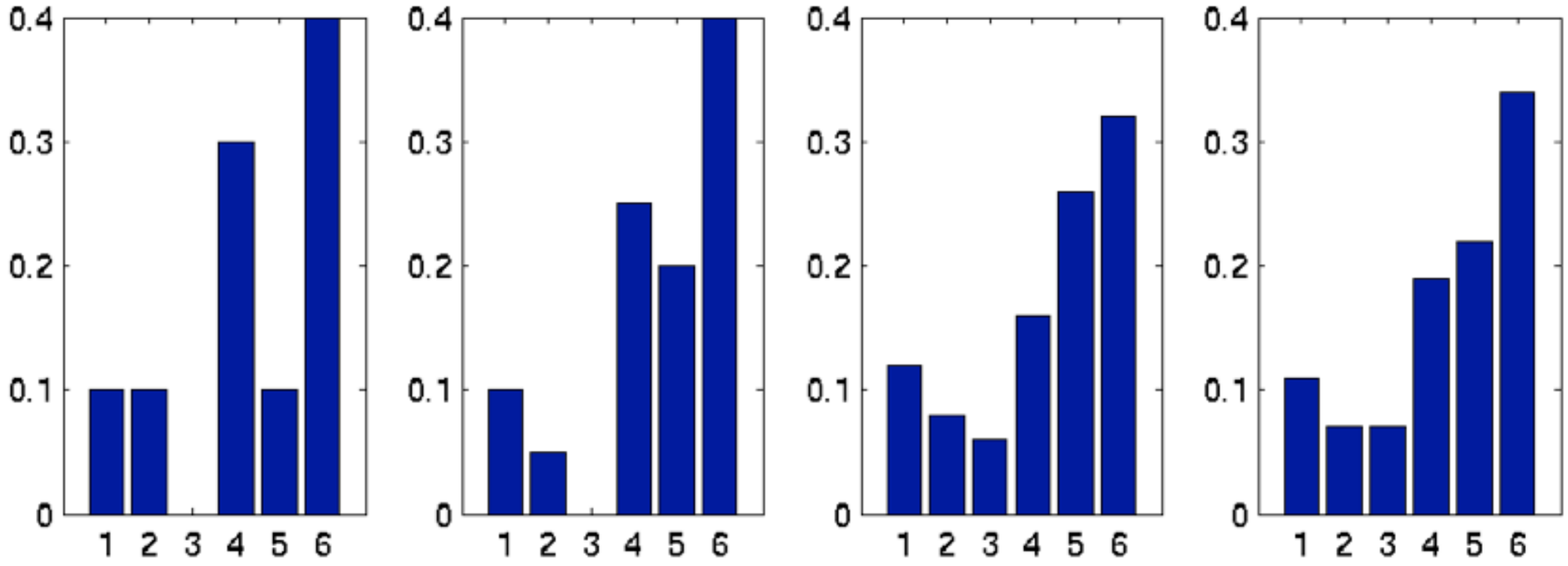


MAP

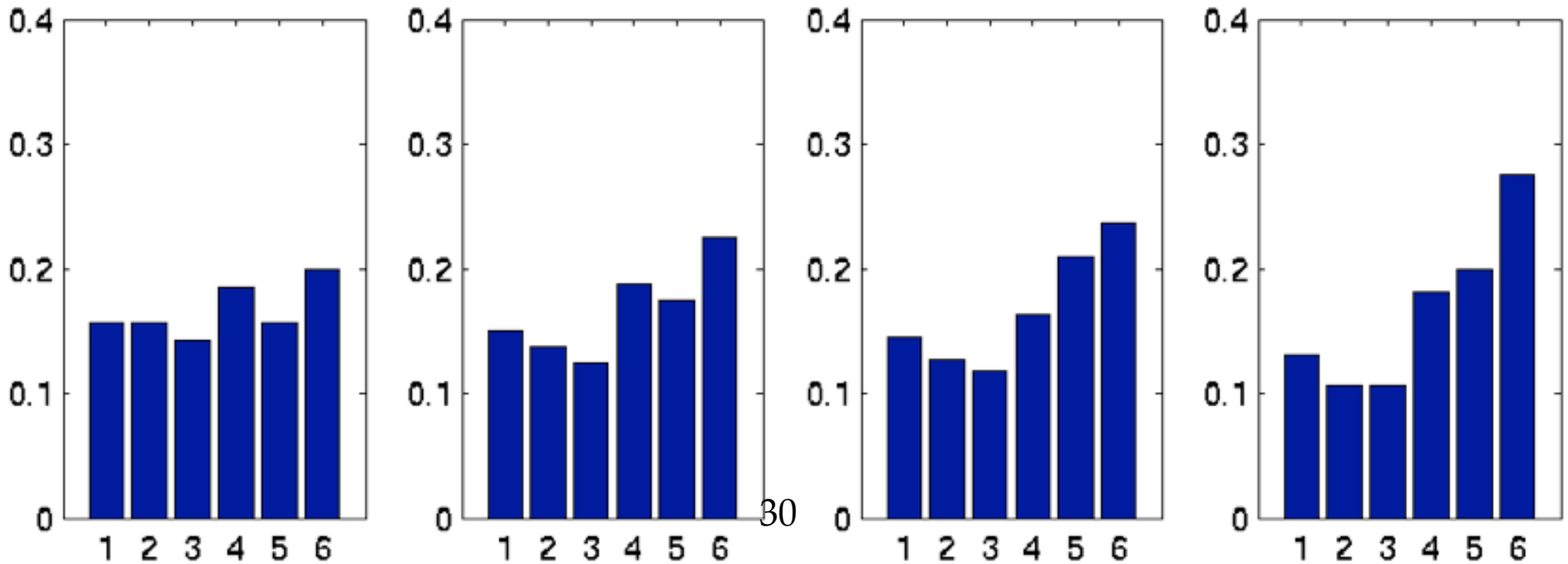


Tainted dice

MLE

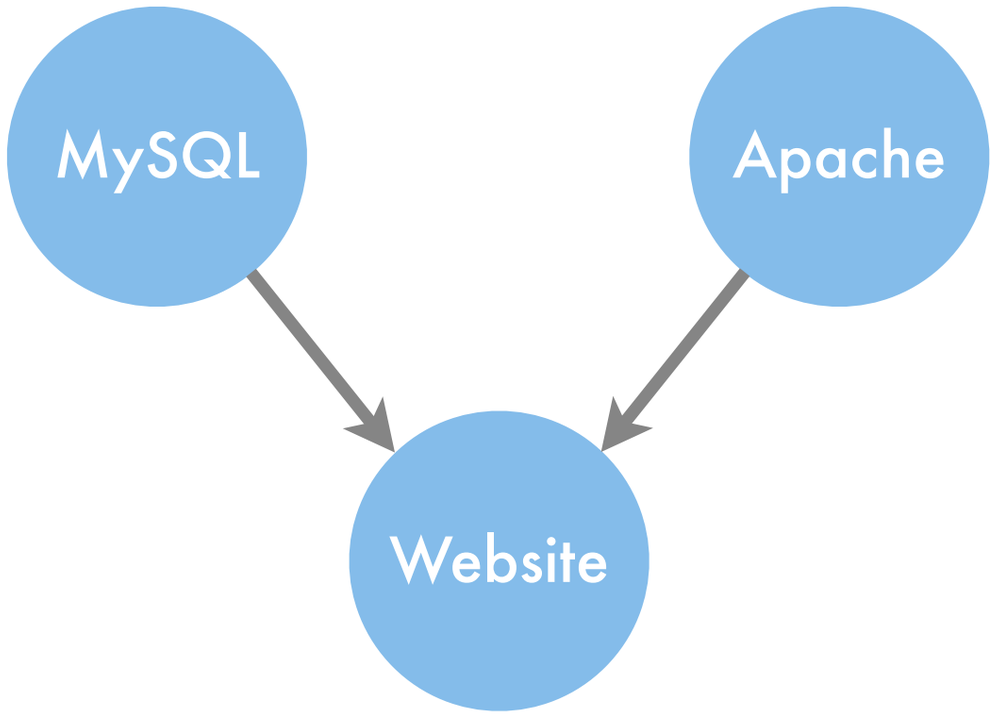


MAP



Directed Graphical Models

… some Yahoo service



- Joint distribution (assume a and m are independent)

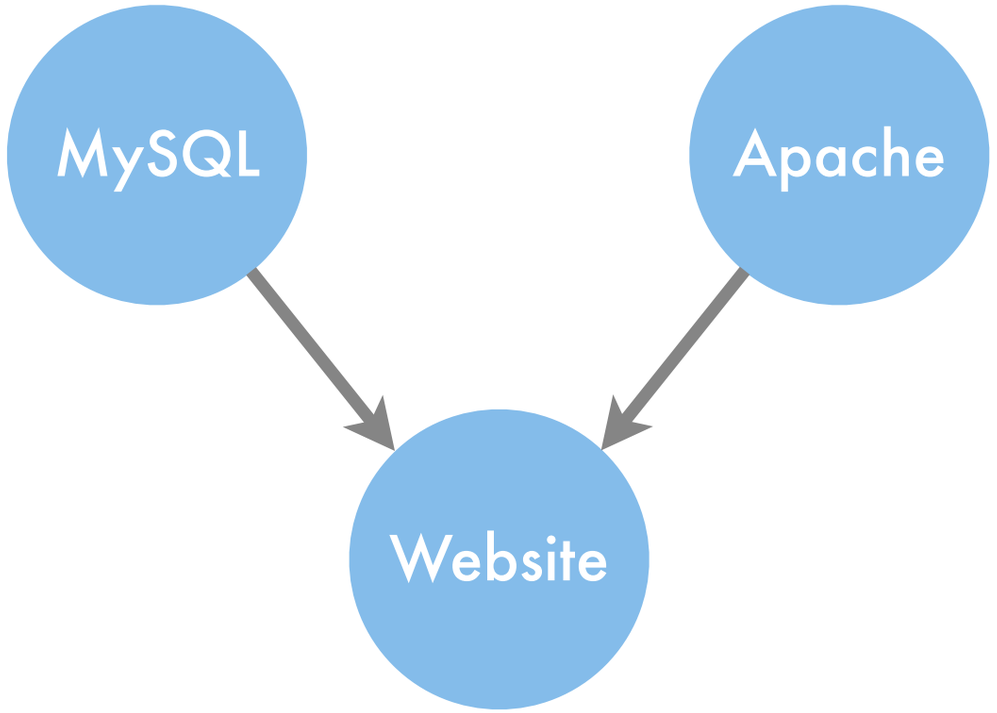
$$p(m, a, w) = p(w|m, a)p(m)p(a)$$

- Explaining away

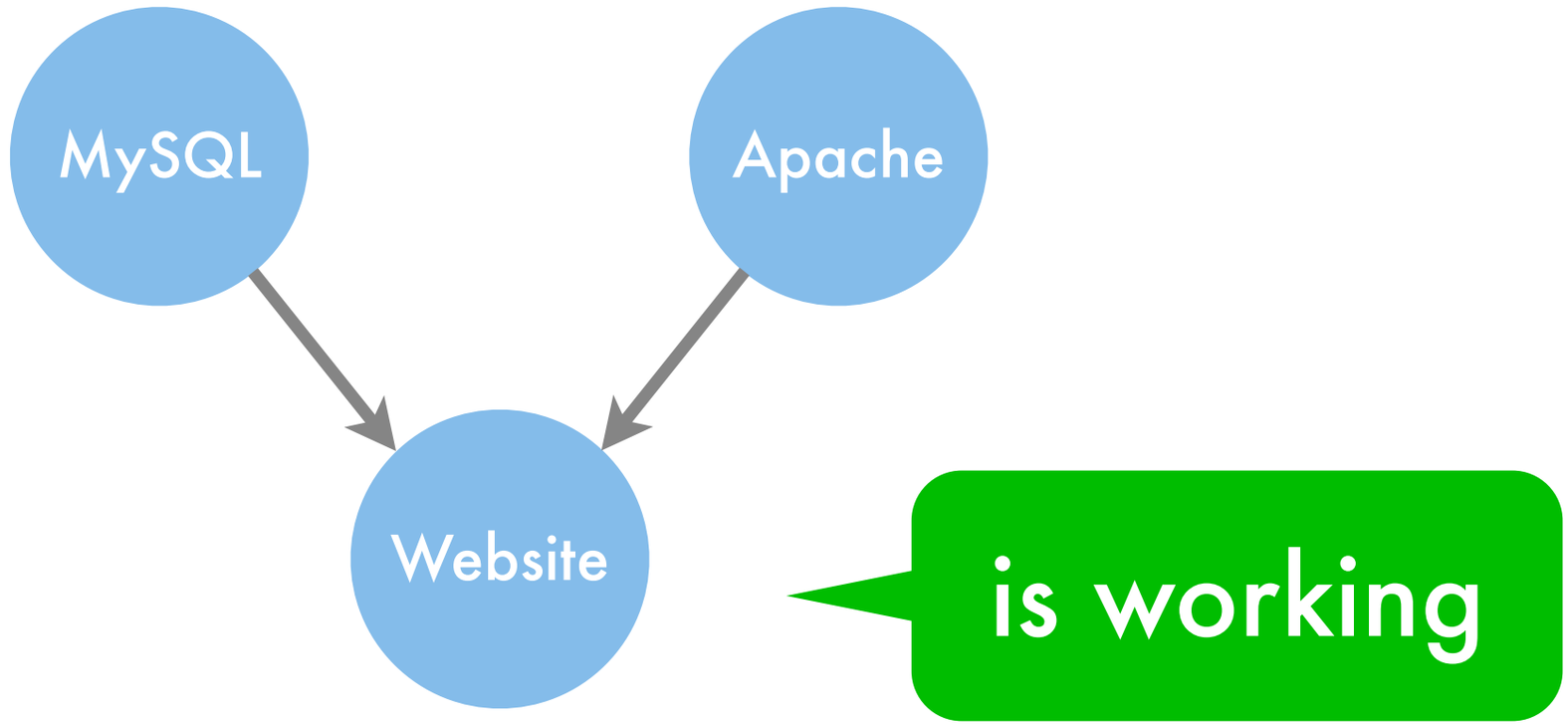
$$p(m, a|w) = \frac{p(w|m, a)p(m)p(a)}{\sum_{m',a'} p(w|m', a')p(m')p(a')}$$

a and m are dependent conditioned on w

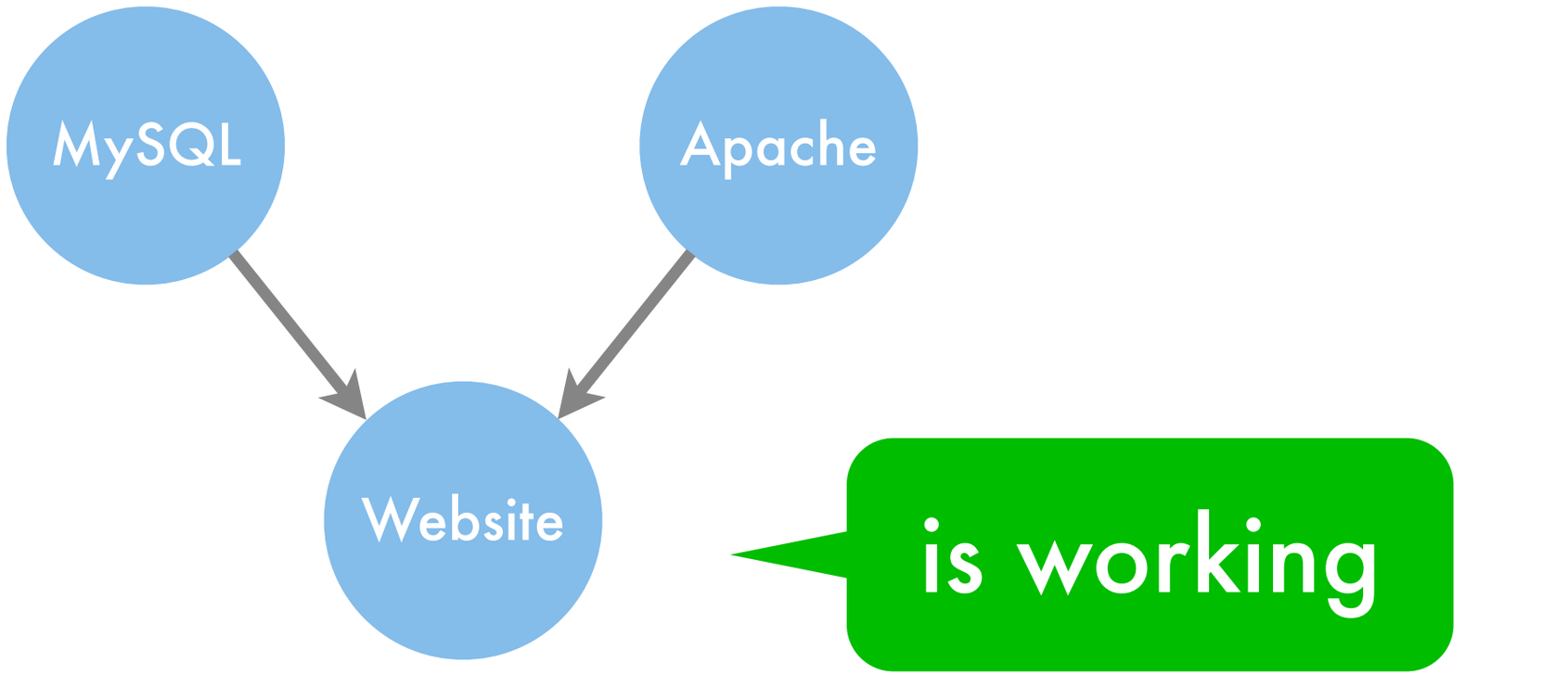
… some Yahoo service



… some Yahoo service

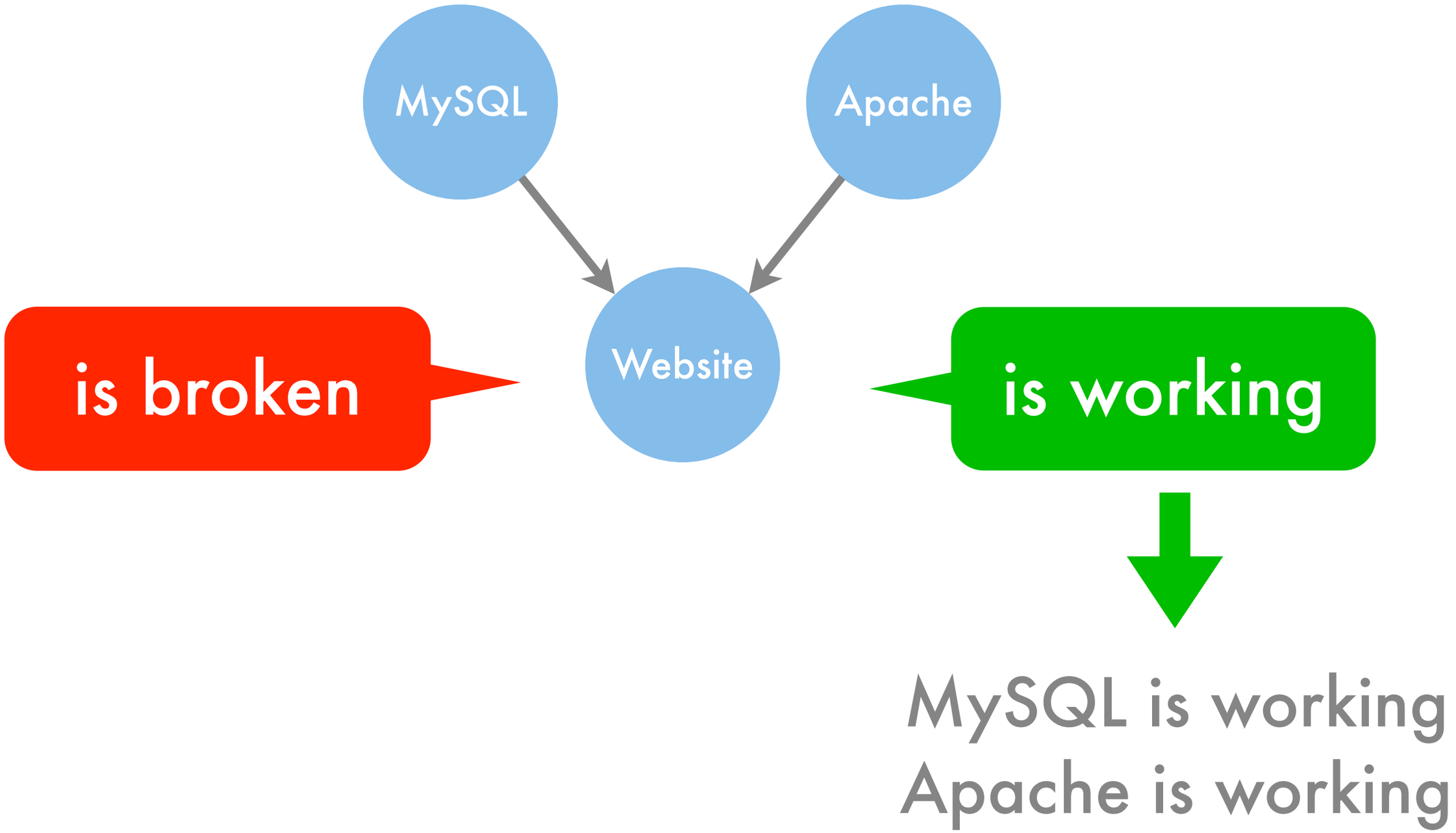


… some Yahoo service



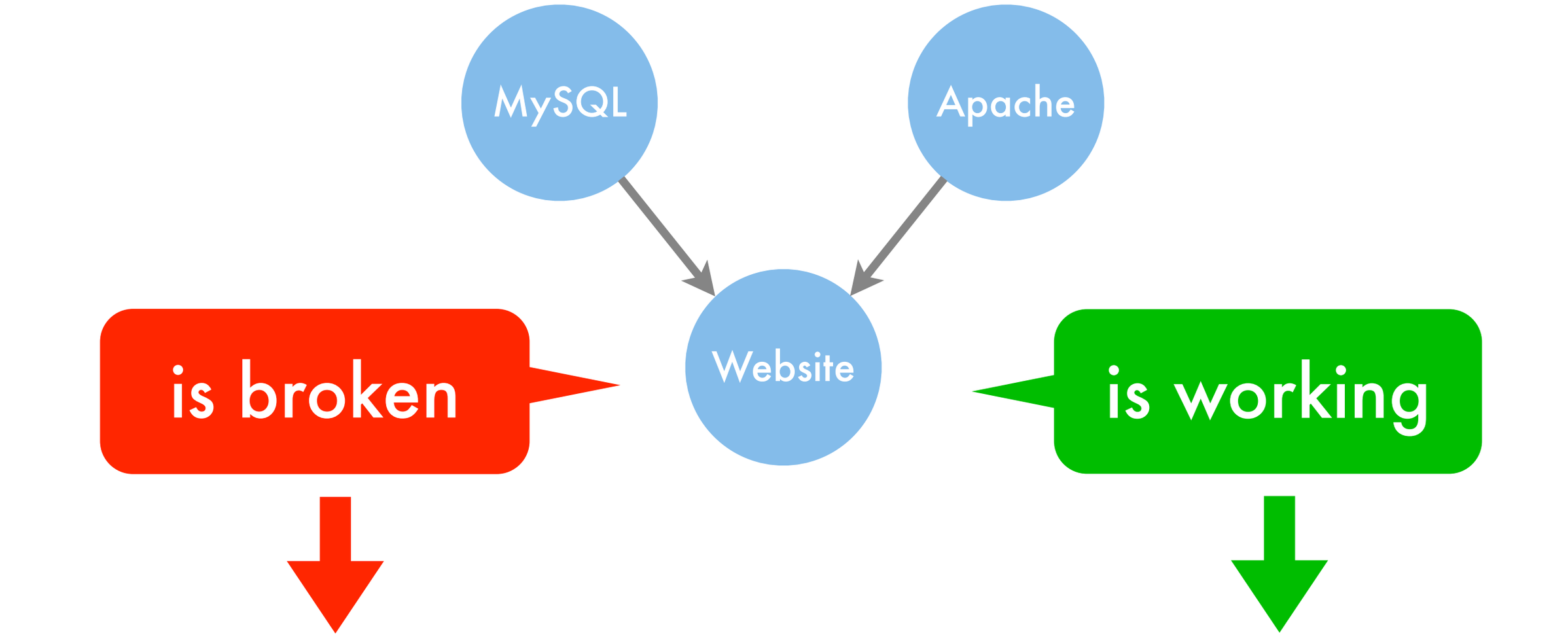
MySQL is working
Apache is working

… some Yahoo service

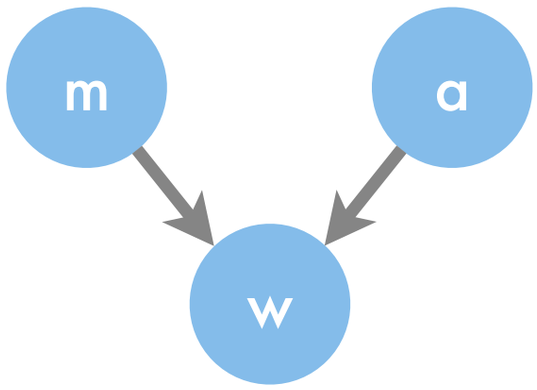


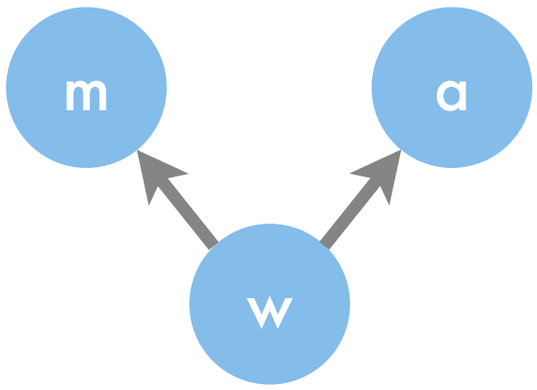
MySQL is working
Apache is working

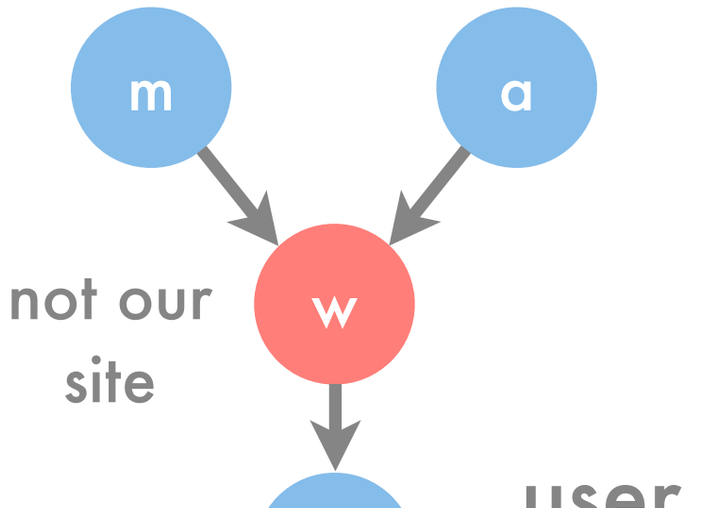
… some Yahoo service



Directed graphical model

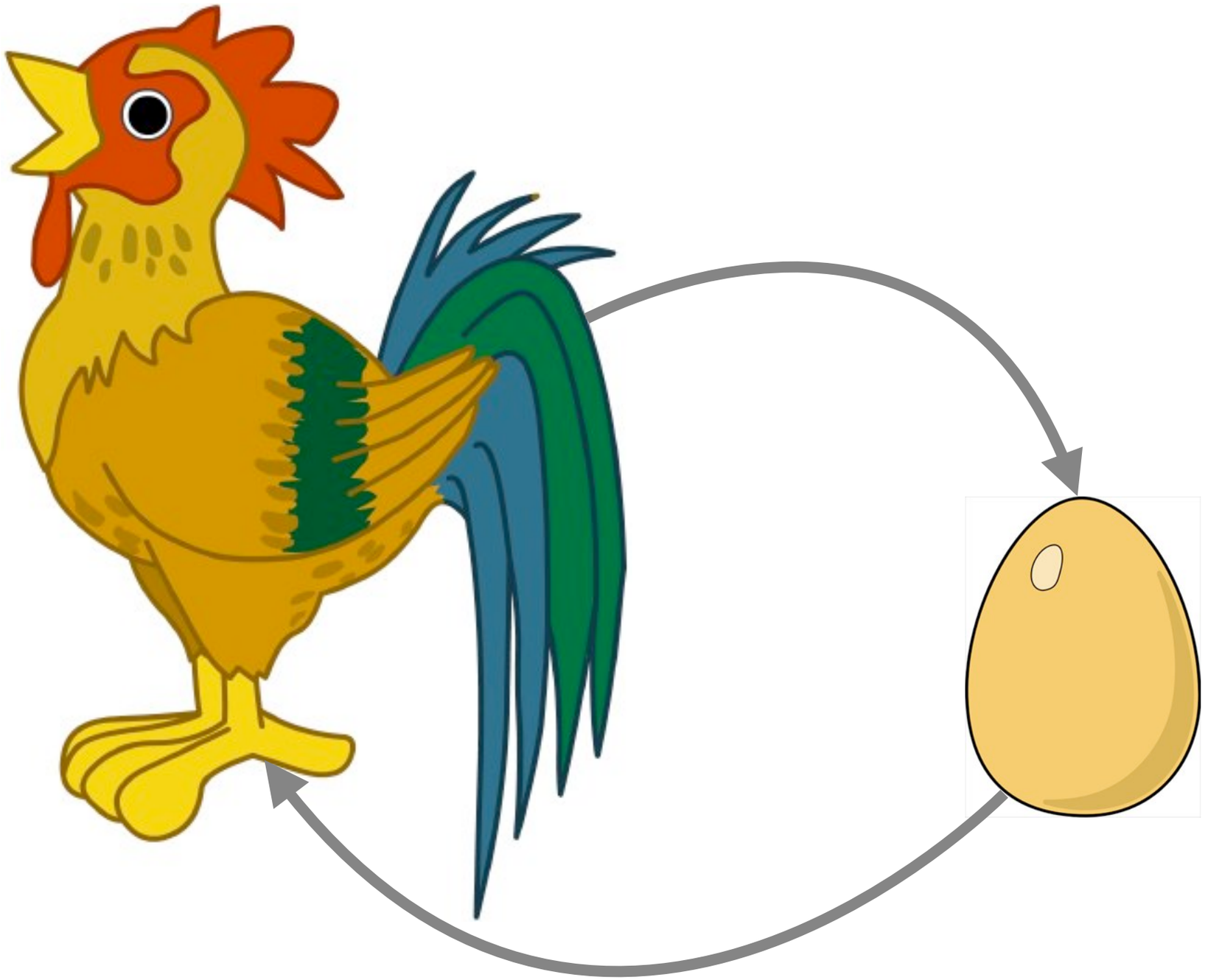




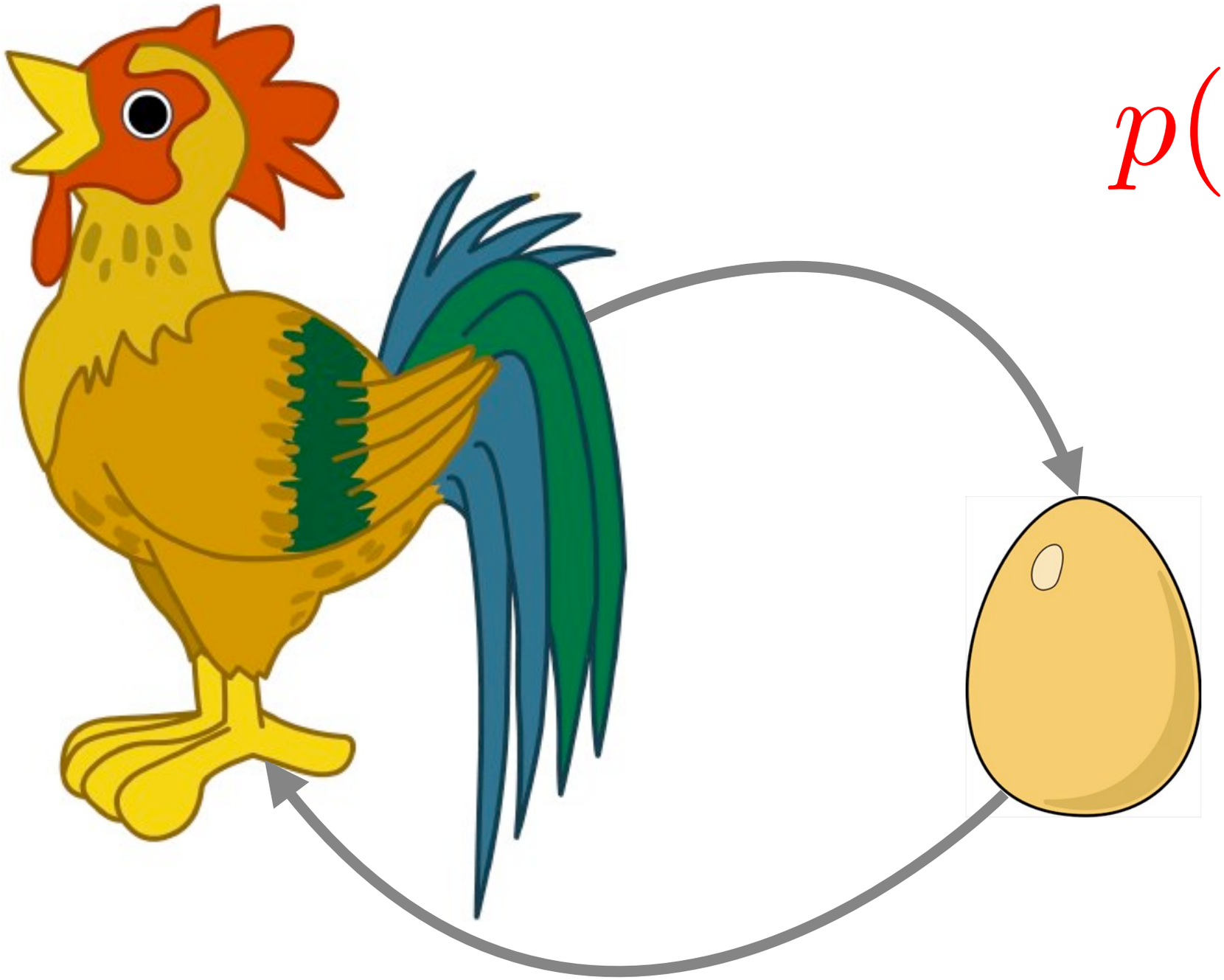


- Easier estimation
  - 7 parameters for full joint distribution
  - 4+1+1 for factorizing distribution
- Causal relations
- Inference for unobserved variables

No loops allowed

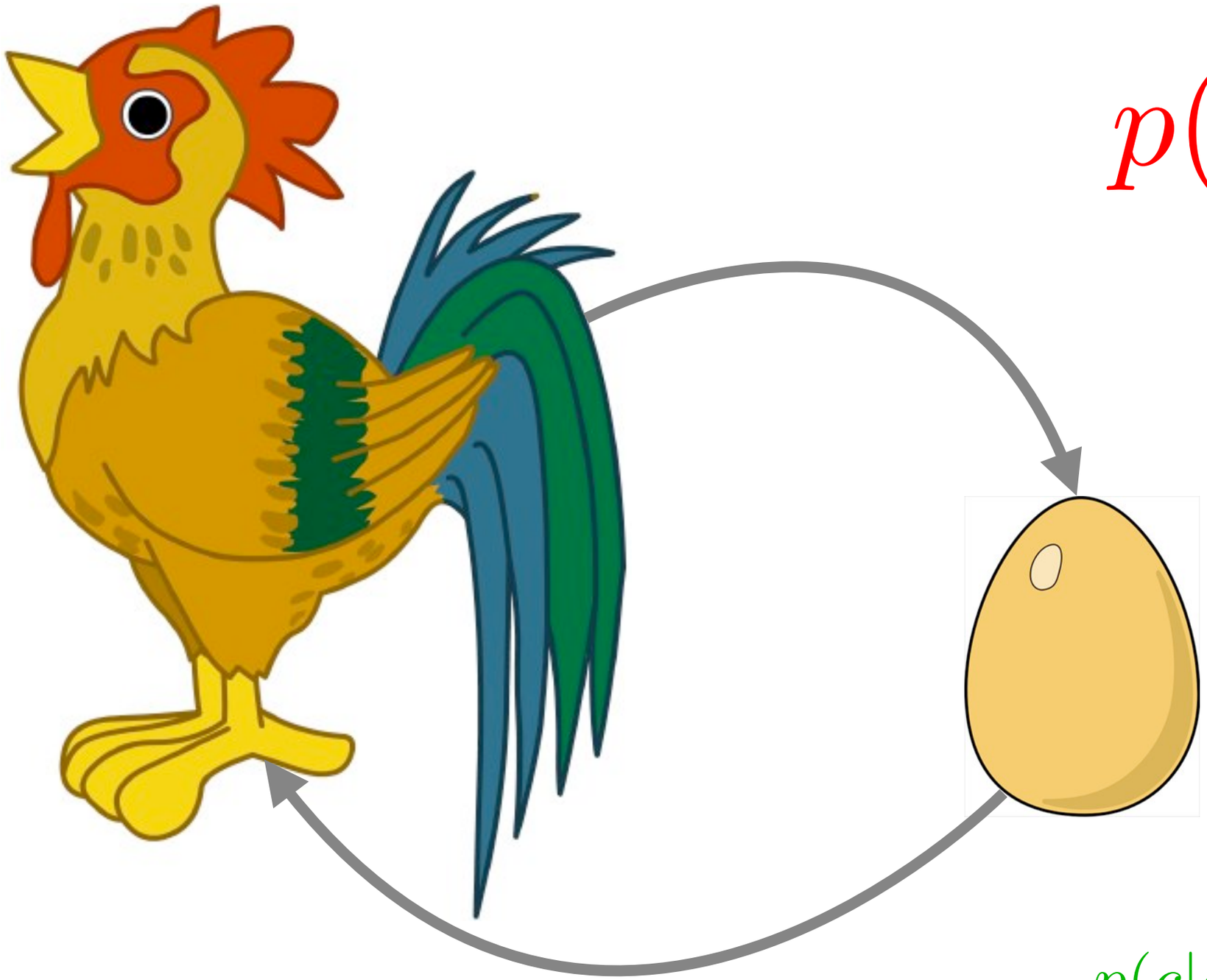


No loops allowed



$$p(c|e)p(e|c)$$

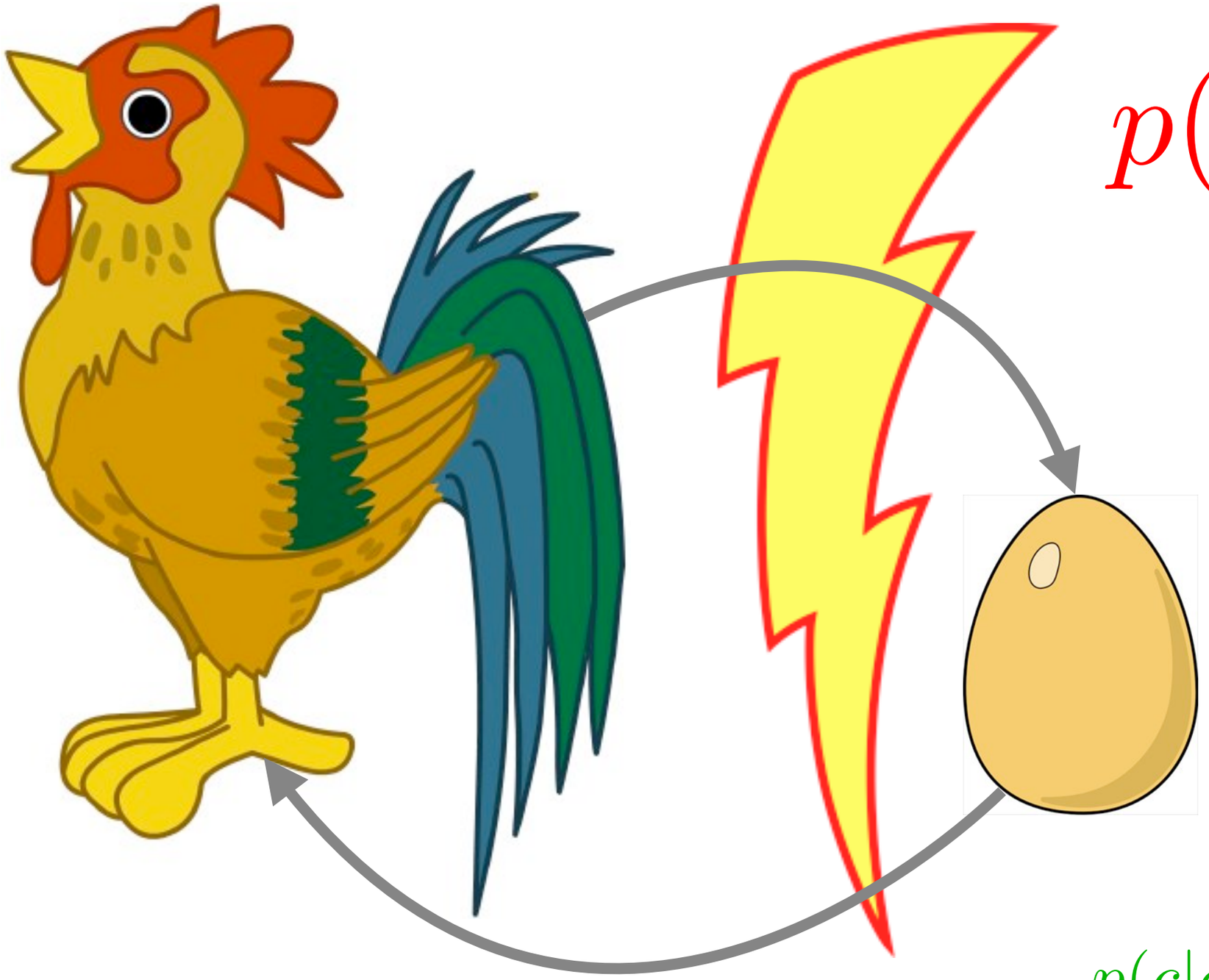
No loops allowed



$p(c|e)p(e|c)$

$p(c|e)p(e)$ or $p(e|c)p(c)$

No loops allowed



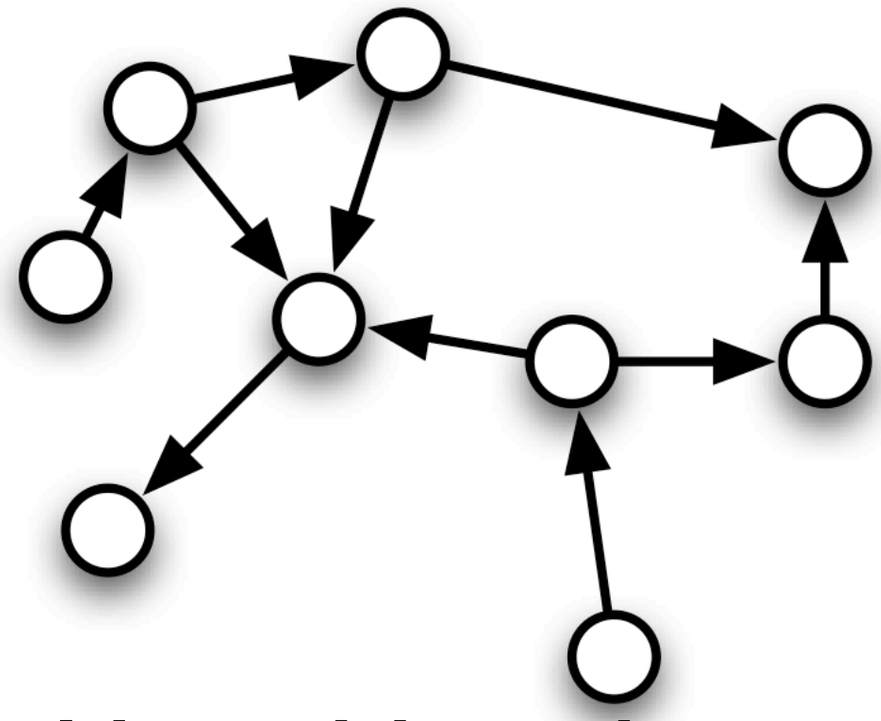
$$p(c|e)p(e|c)$$

$$p(c|e)p(e) \text{ or } p(e|c)p(c)$$

Directed Graphical Model

- Joint probability distribution

$$p(x) = \prod_i p(x_i | x_{\text{parents(i)}})$$



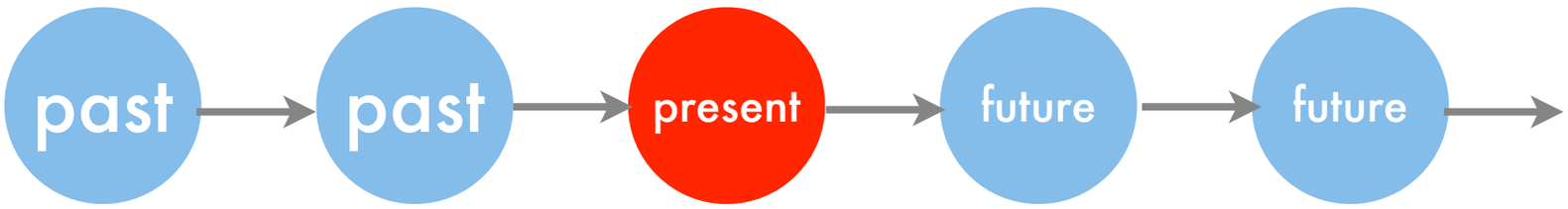
- Parameter estimation
  - If x is fully observed the likelihood breaks up

$$\log p(x|\theta) = \sum_i \log p(x_i | x_{\text{parents(i)}}, \theta)$$

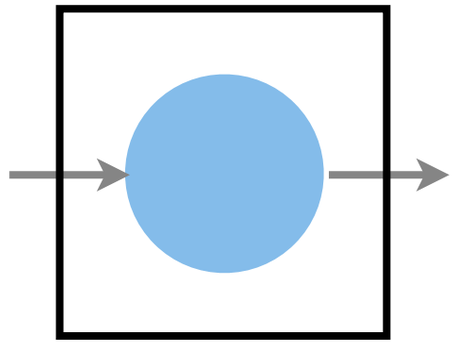
  - If x is partially observed things get interesting maximization, EM, variational, sampling ...

Chains

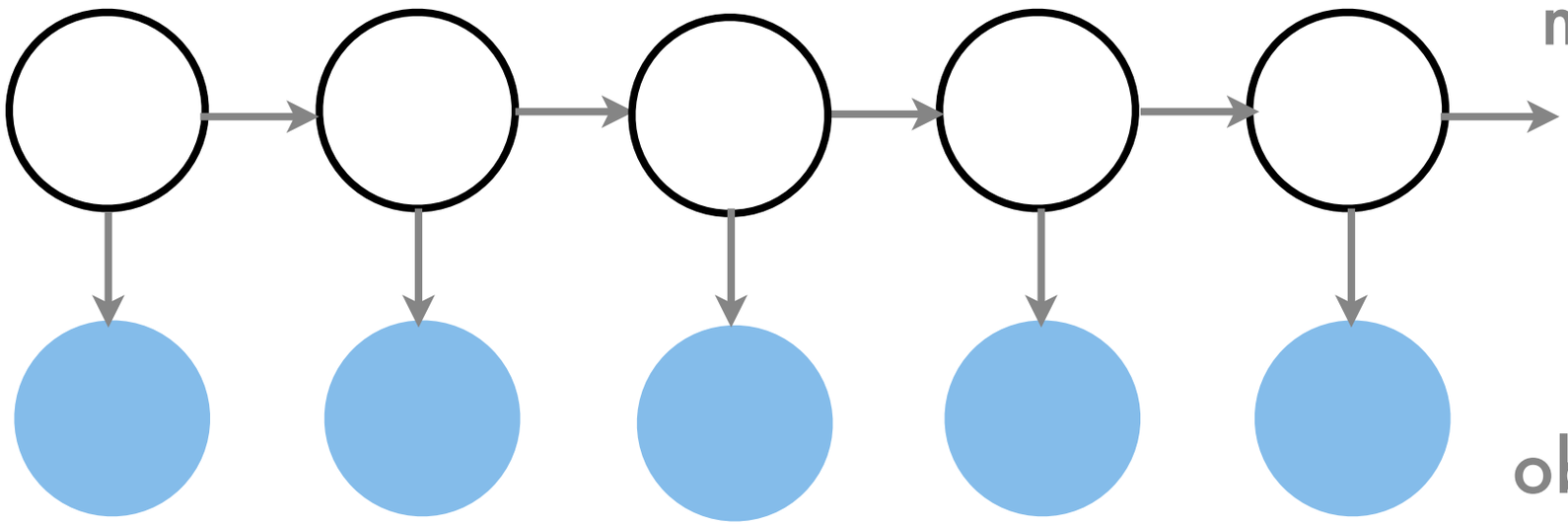
Markov Chain



Plate

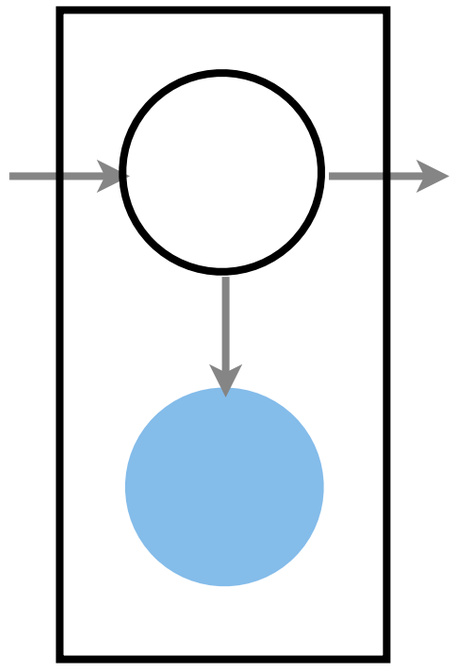


Hidden Markov Chain



user's mindset

observed user action



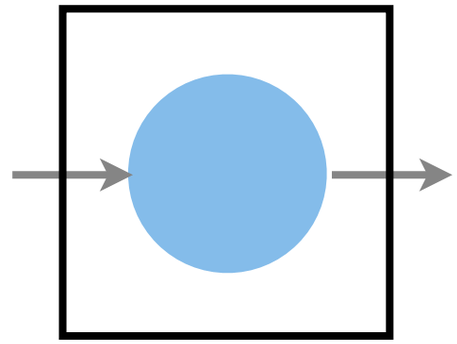
user model for traversal through search results

Chains

Markov Chain

$$p(x;\theta) = p(x_0;\theta) \prod_{i=1}^{n-1} p(x_{i+1}|x_i;\theta)$$

Plate

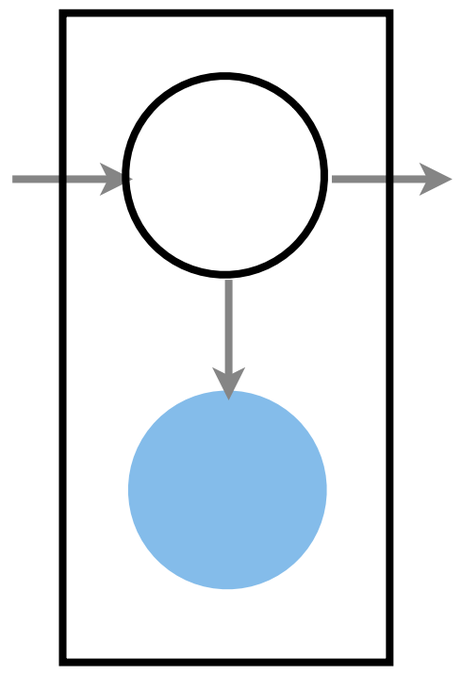


Hidden Markov Chain

$$p(x,y;\theta) = p(x_0;\theta) \prod_{i=1}^{n-1} p(x_{i+1}|x_i;\theta) \prod_{i=1}^{n} p(y_i|x_i)$$

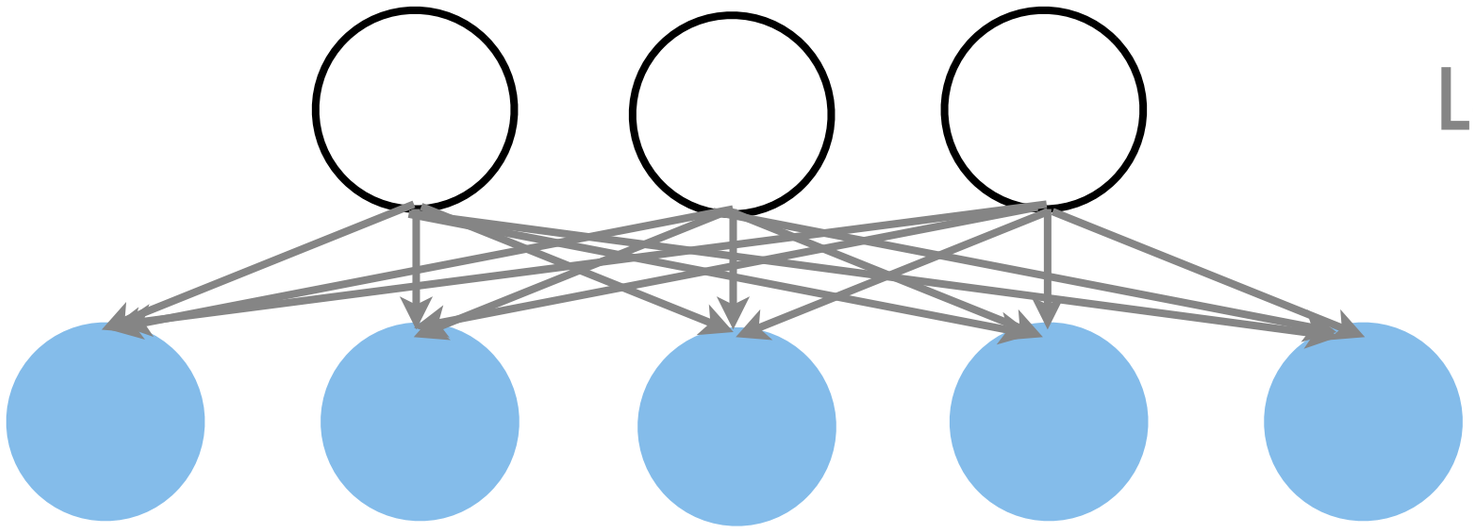
user's mindset

observed user action



user model for traversal through search results

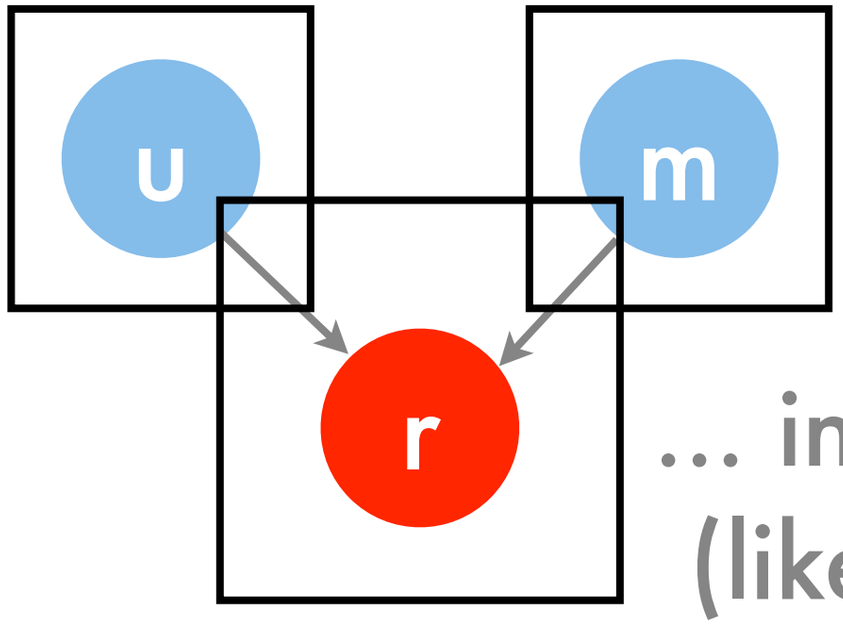
Factor Graphs



- Observed effects
  Click behavior, queries, watched news, emails
- Latent factors
  User profile, news content, hot keywords, social connectivity graph, events

Recommender Systems

news,
SearchMonkey
answers
social
ranking
OMG
personals



... intersecting plates ...
(like nested for loops)

- Users u
- Movies m
- Ratings r (but only for a subset of users)

Challenges

- How to design models
  - Common (engineering) sense
  - Computational tractability
- Dependency analysis
  - Bayes ball (not in this lecture)
- Inference
  - Easy for fully observed situations
  - Many algorithms if not fully observed
  - Dynamic programming / message passing

Challenges

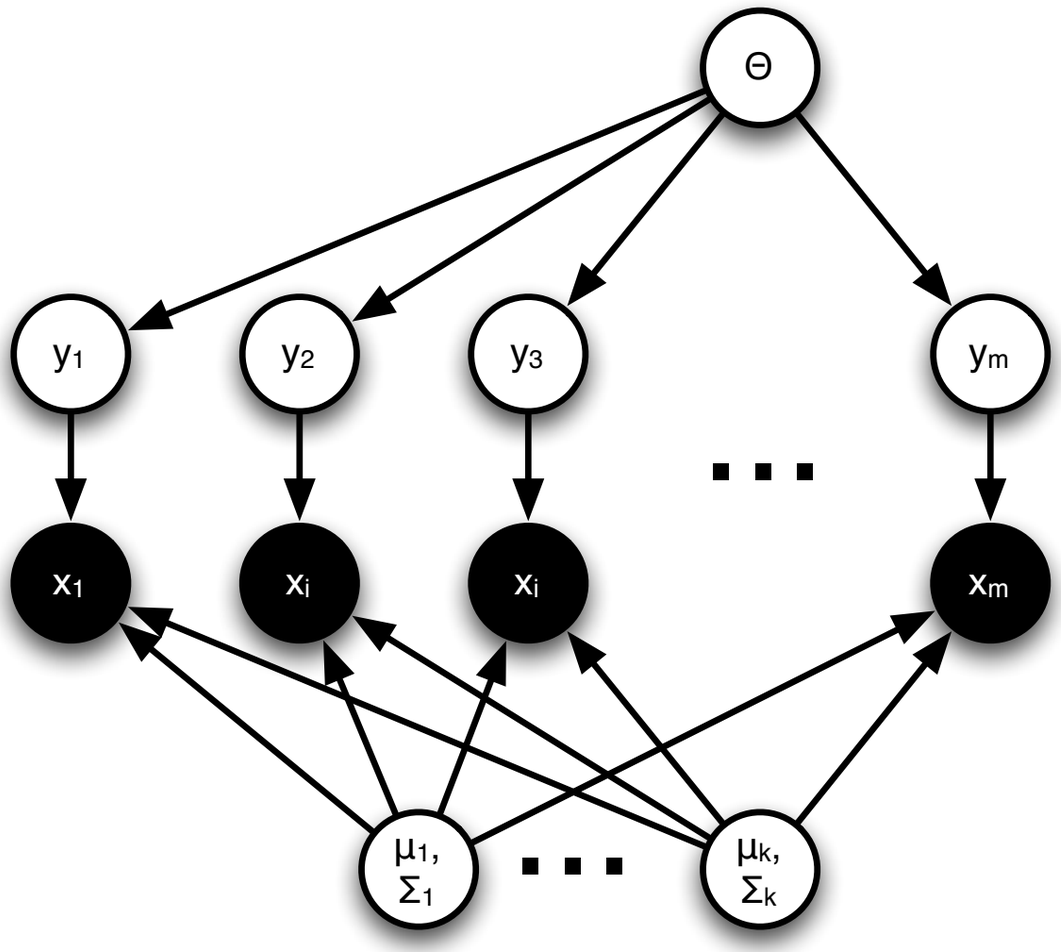
- How to design models
  - Common (engineering) sense
  - Computational tractability
- Dependency analysis
  - Bayes ball (not in this lecture)
- Inference
  - Easy for fully observed situations
  - Many algorithms if not fully observed
  - Dynamic programming / message passing

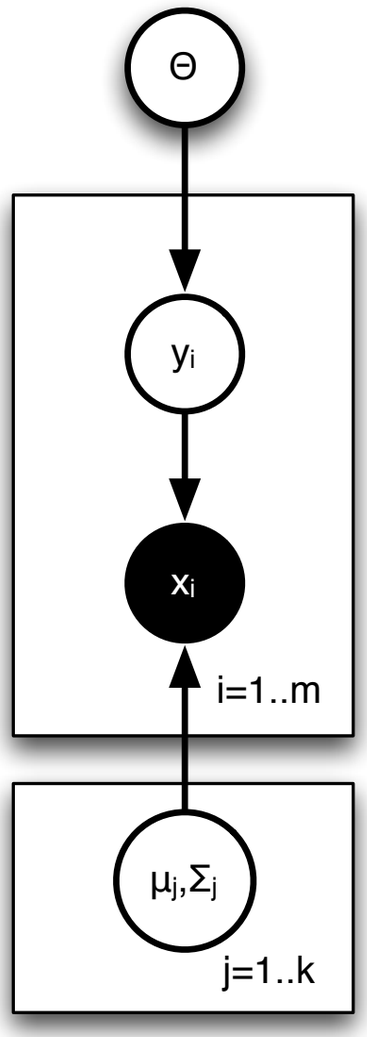
your job

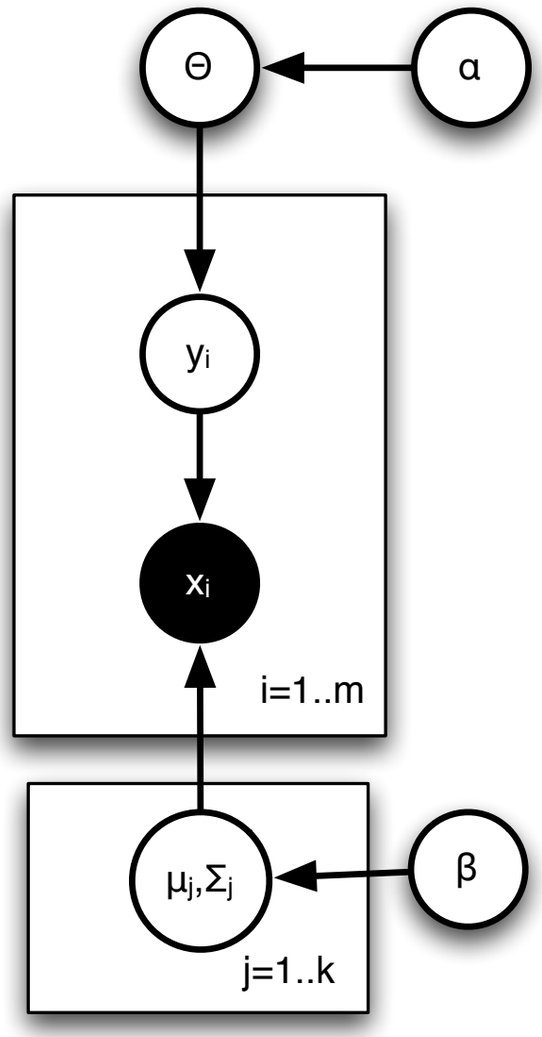
my job

Mixture Models - Clustering

Basic Idea







$$p(X, Y|\theta, \sigma, \mu) = \prod_{i=1}^{n} p(x_i|y_i, \sigma, \mu)p(y_i|\theta)$$

What can we cluster?

mails

text

news

urls

products

users

queries

locations

spammers

ads

events

abuse

Mixture of Gaussians

- Draw cluster label y from discrete distribution
- Draw data x from Gaussian for given cluster y

- Prior for discrete distribution - Dirichlet
- Prior for Gaussians - Gauss-Wishart distribution

- Problem: we don't know y
  - If we knew the parameters we could get y
  - If we knew y we could get the parameters

k-means

- Fixed uniform variance for all Gaussians
- Fixed uniform distribution over clusters
- Initialize centers with random subset of points
- Find most likely cluster y for x

$$y_i = \operatorname*{argmax}_y p(x_i|y, \sigma, \mu)$$

- Find most likely center for given cluster

$$\mu_y = \frac{1}{n_y} \sum_i \{y_i = y\} x_i$$

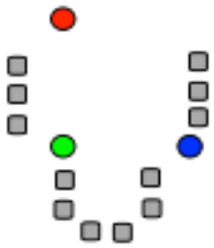
- Repeat until converged

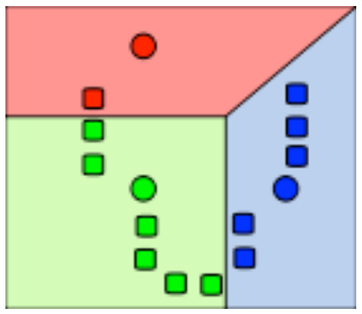
k-means

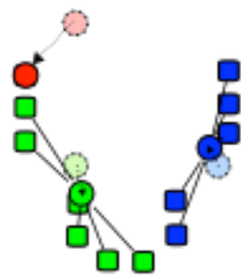
- Pro
  - simple algorithm
  - can be implemented by MapReduce passes
- Con
  - no proper probabilistic representation
  - can get stuck easily in local minima

k-means

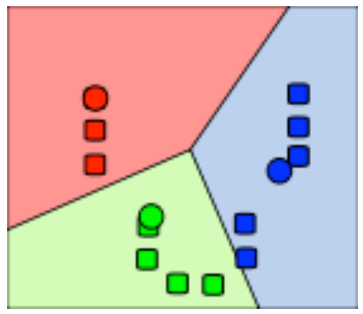
partitioning







partitioning



initialization

update

Expectation Maximization

Expectation Maximization

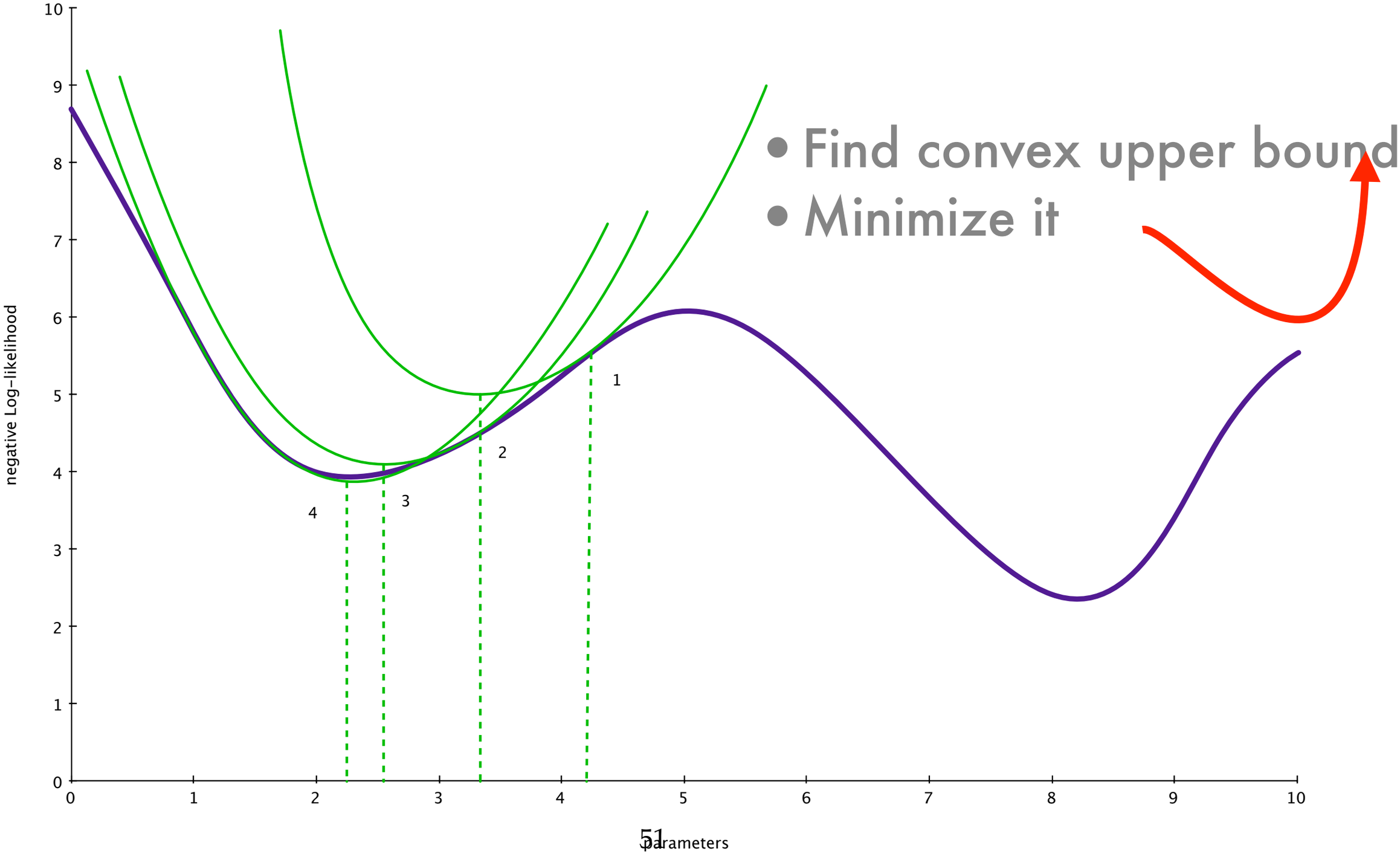
- Optimization Problem

$$\underset{\theta,\mu,\sigma}{\text{maximize}}\, p(X|\theta,\sigma,\mu) = \underset{\theta,\mu,\sigma}{\text{maximize}} \sum_{Y} \prod_{i=1}^{n} p(x_i|y_i,\sigma,\mu)p(y_i|\theta)$$

  This problem is nonconvex and difficult to solve

- Key idea
  If we knew p(y|x) we could estimate the remaining parameters easily and vice versa

Nonconvex Minimization



- Find convex upper bound
- Minimize it

Expectation Maximization

- Variational Bound

$$\log p(x;\theta) \geq \log p(x;\theta) - D(q(y)\|p(y|x;\theta))$$

$$= \int dq(y)\,[\log p(x;\theta) + \log p(y|x;\theta) - \log q(y)]$$

$$= \int dq(y) \log p(x,y;\theta) - \int dq(y) \log q(y)$$

This inequality is tight for p(y|x) = q(y)

- Expectation step

$$q(y) = p(y|x;\theta)$$

- Maximization step

$$\theta^* = \operatorname*{argmax}_\theta \int dq(y) \log p(x,y;\theta)$$

Expectation Maximization

- Variational Bound

$$\log p(x;\theta) \geq \log p(x;\theta) - D(q(y)\|p(y|x;\theta))$$

$$= \int dq(y)\,[\log p(x;\theta) + \log p(y|x;\theta) - \log q(y)]$$

$$= \int dq(y) \log p(x,y;\theta) - \int dq(y) \log q(y)$$

This inequality is tight for p(y|x) = q(y)

- Expectation step

$$q(y) = p(y|x;\theta)$$

find bound

- Maximization step

$$\theta^* = \operatorname*{argmax}_\theta \int dq(y) \log p(x,y;\theta)$$

Expectation Maximization

- Variational Bound

$$\log p(x;\theta) \geq \log p(x;\theta) - D(q(y)\|p(y|x;\theta))$$

$$= \int dq(y)\,[\log p(x;\theta) + \log p(y|x;\theta) - \log q(y)]$$

$$= \int dq(y) \log p(x,y;\theta) - \int dq(y) \log q(y)$$

This inequality is tight for p(y|x) = q(y)

- Expectation step

$$q(y) = p(y|x;\theta)$$

find bound

- Maximization step

$$\theta^* = \operatorname*{argmax}_\theta \int dq(y) \log p(x,y;\theta)$$

maximize it

Expectation Step

- Factorizing distribution

$$q(Y) = \prod_i q_i(y)$$

- E-Step

$$q_i(y) \propto p(x_i|y_i, \mu, \sigma)p(y_i|\theta) \text{ hence}$$

$$m_{iy} := \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_y|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x_i - \mu_y)\Sigma_y^{-1}(x_i - \mu_y)\right] p(y)$$

$$q_i(y) = \frac{m_{iy}}{\sum_{y'} m_{iy'}}$$

Maximization Step

- Log-likelihood

$$\log p(X, Y|\theta, \mu, \sigma) = \sum_{i=1}^{n} \log p(x_i|y_i, \mu, \sigma) + \log p(y_i|\theta)$$

- Cluster distribution (weighted Gaussian MLE)

$$n_y = \sum_i q_i(y)$$

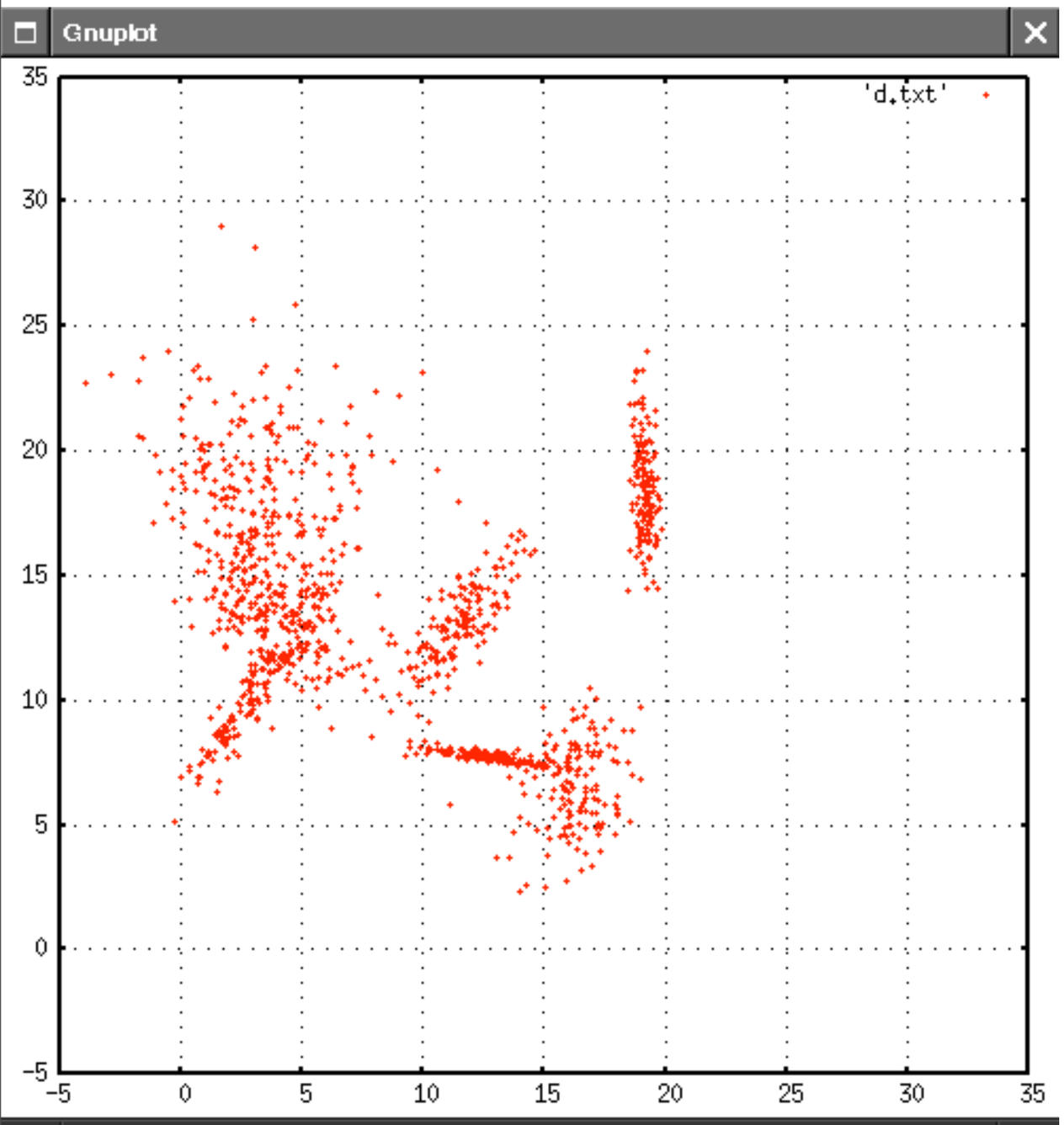
$$\mu_y = \frac{1}{n_y} \sum_{i=1}^{n} q_i(y) x_i$$

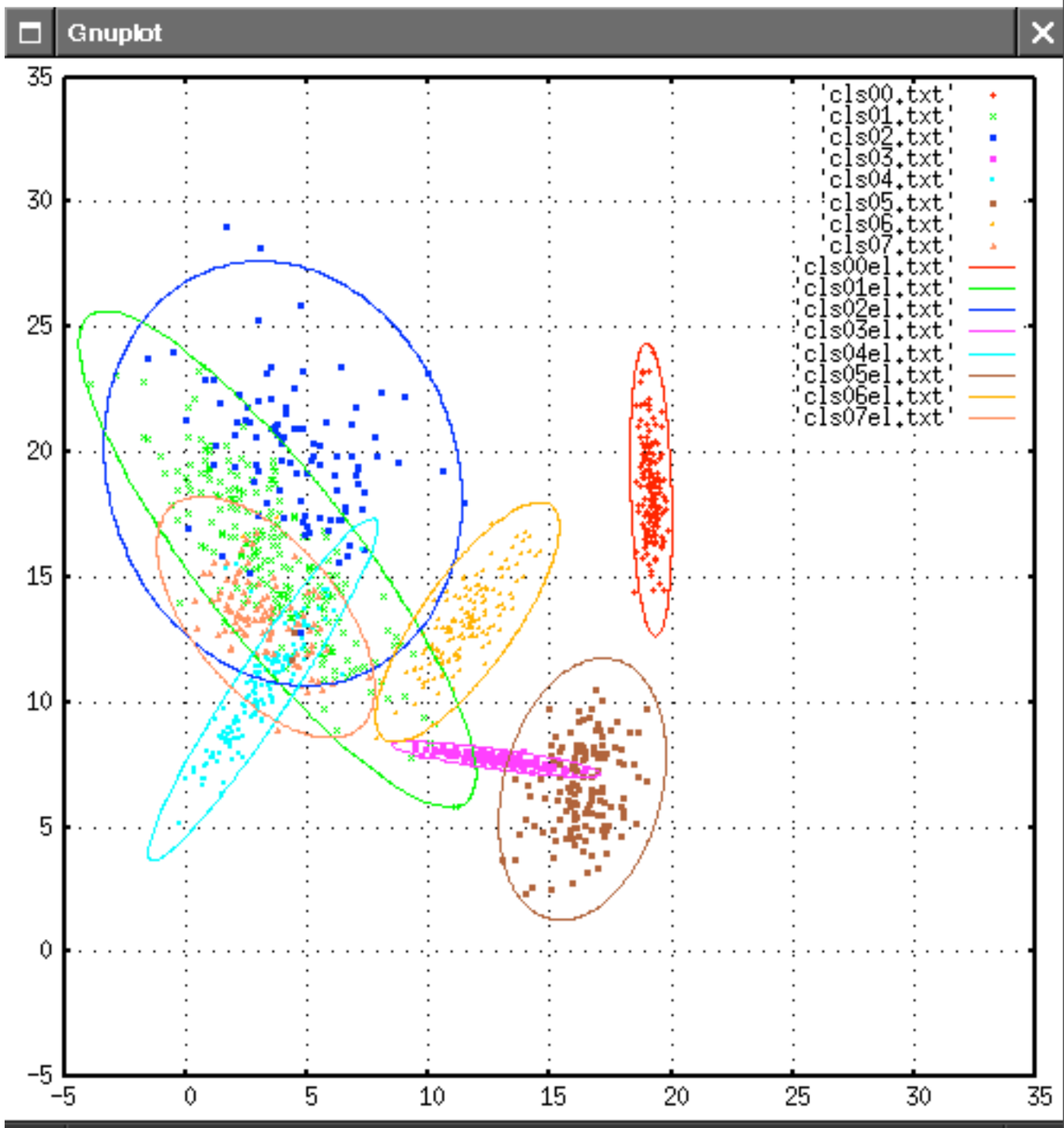
$$\Sigma_y = \frac{1}{n_y} \sum_{i=1}^{n} q_i(y) x_i x_i^\top - \mu_y \mu_y^\top$$

- Cluster probabilities

$$\theta^* = \operatorname*{argmax}_\theta \sum_{i=1}^{n} \sum_y q_i(y) \log p(y_i|\theta) \text{ hence } p(y|\theta) = \frac{n_y}{n}$$

EM Clustering in action





Problem

Estimates may diverge
(infinite variance, zero probability, tiny clusters)

Solution

- Use priors for μ, σ, θ
  - Dirichlet distribution for cluster probabilities
  - Gauss-Wishart for Gaussian
- Cluster distribution

$$n_y = n_0 + \sum_i q_i(y)$$

$$\mu_y = \frac{1}{n_y} \sum_{i=1}^{n} q_i(y) x_i$$

$$\Sigma_y = \frac{1}{n_y} \sum_{i=1}^{n} q_i(y) x_i x_i^\top + \frac{n_0}{n_y} \mathbf{1} - \mu_y \mu_y^\top$$

- Cluster probabilities

$$p(y|\theta) = \frac{n_y}{n + k \cdot n_0}$$

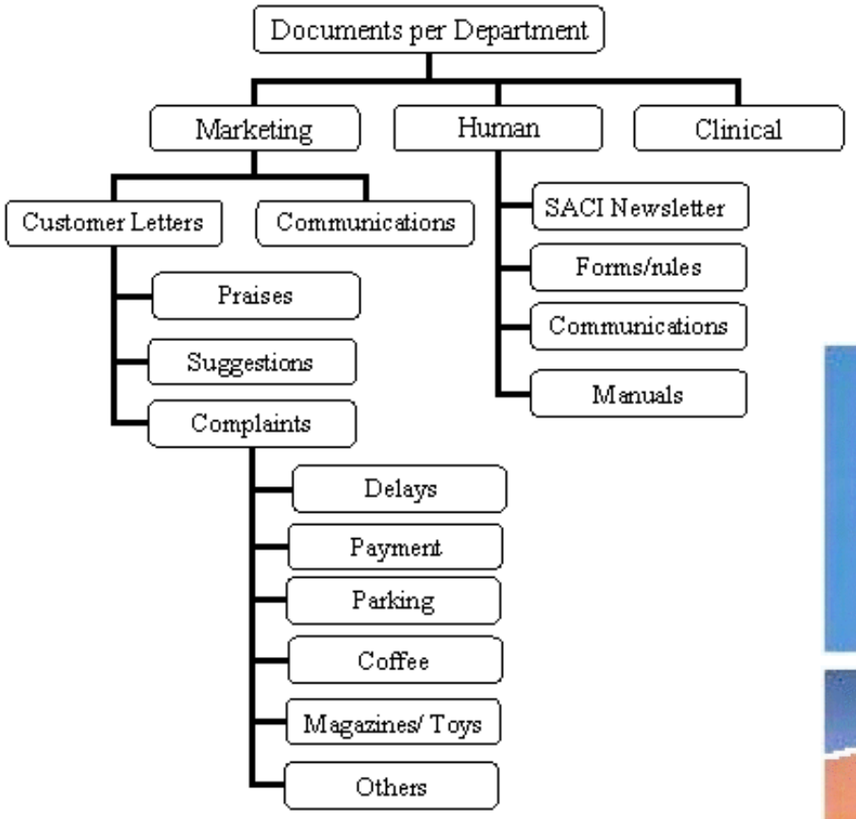
Variational Approximation

- Lower bound on log-likelihood

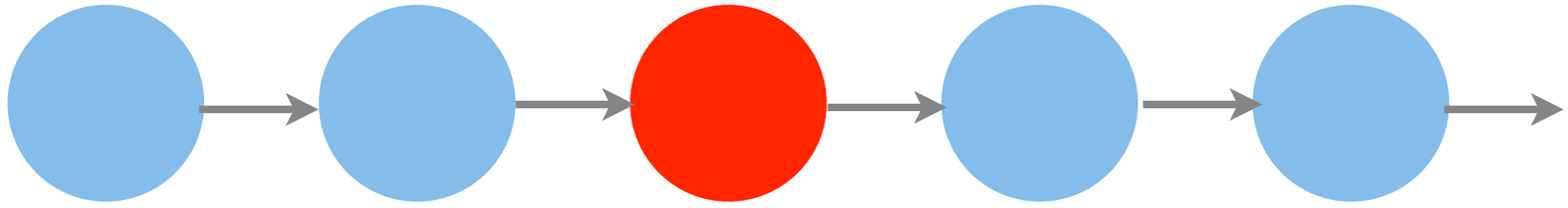
$$\log p(x;\theta) \geq \int dq(y) \log p(x,y;\theta) - \int dq(y) \log q(y)$$

- Key insight - inequality holds for any q
  - Find q within subset Q to tighten inequality
  - Find parameters to maximize for fixed q
- Inference for graphical models where joint probability computation is infeasible

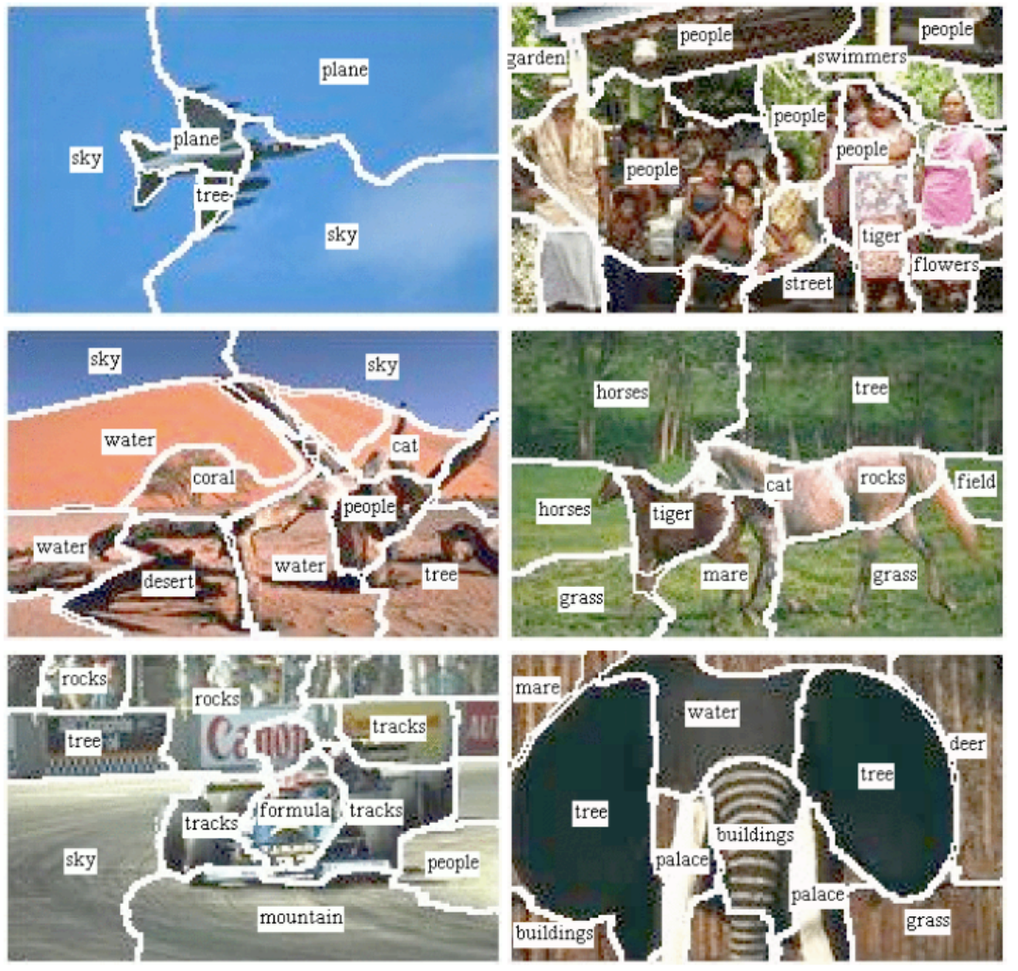
Beyond mixtures



taxonomies



chains

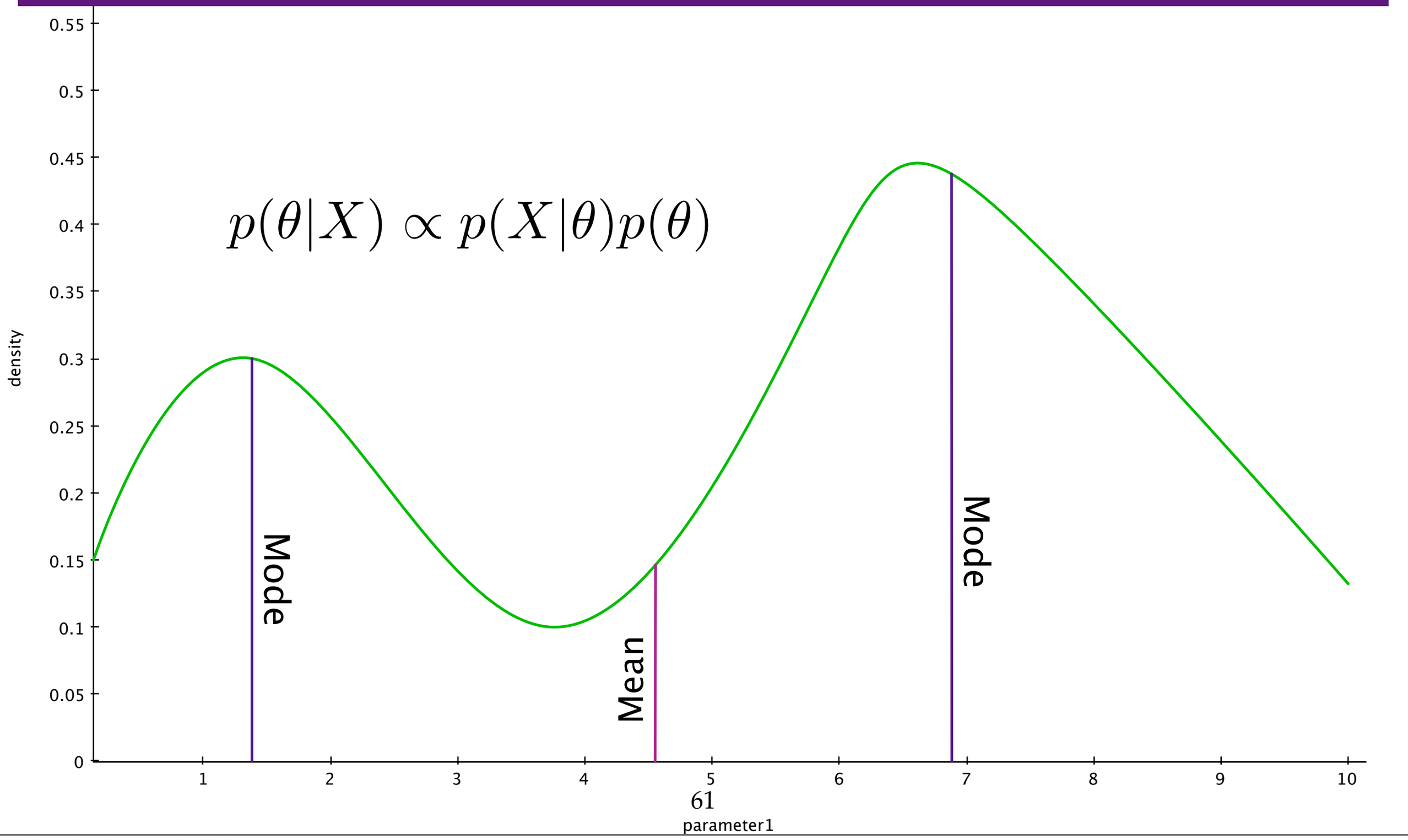


topics

Sampling

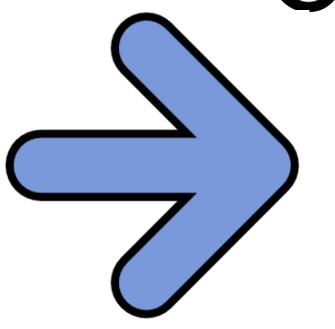
Is maximization (always) good?

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$

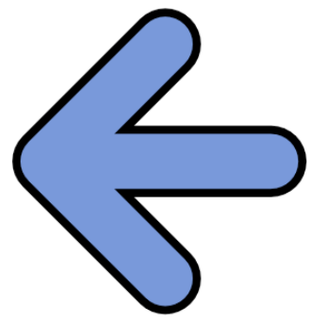


Sampling

- Key idea
  - Want accurate distribution of the posterior
  - Sample from posterior distribution rather than maximizing it
- Problem - direct sampling is usually intractable
- Solutions
  - Markov Chain Monte Carlo (complicated)
  - Gibbs Sampling (somewhat simpler)

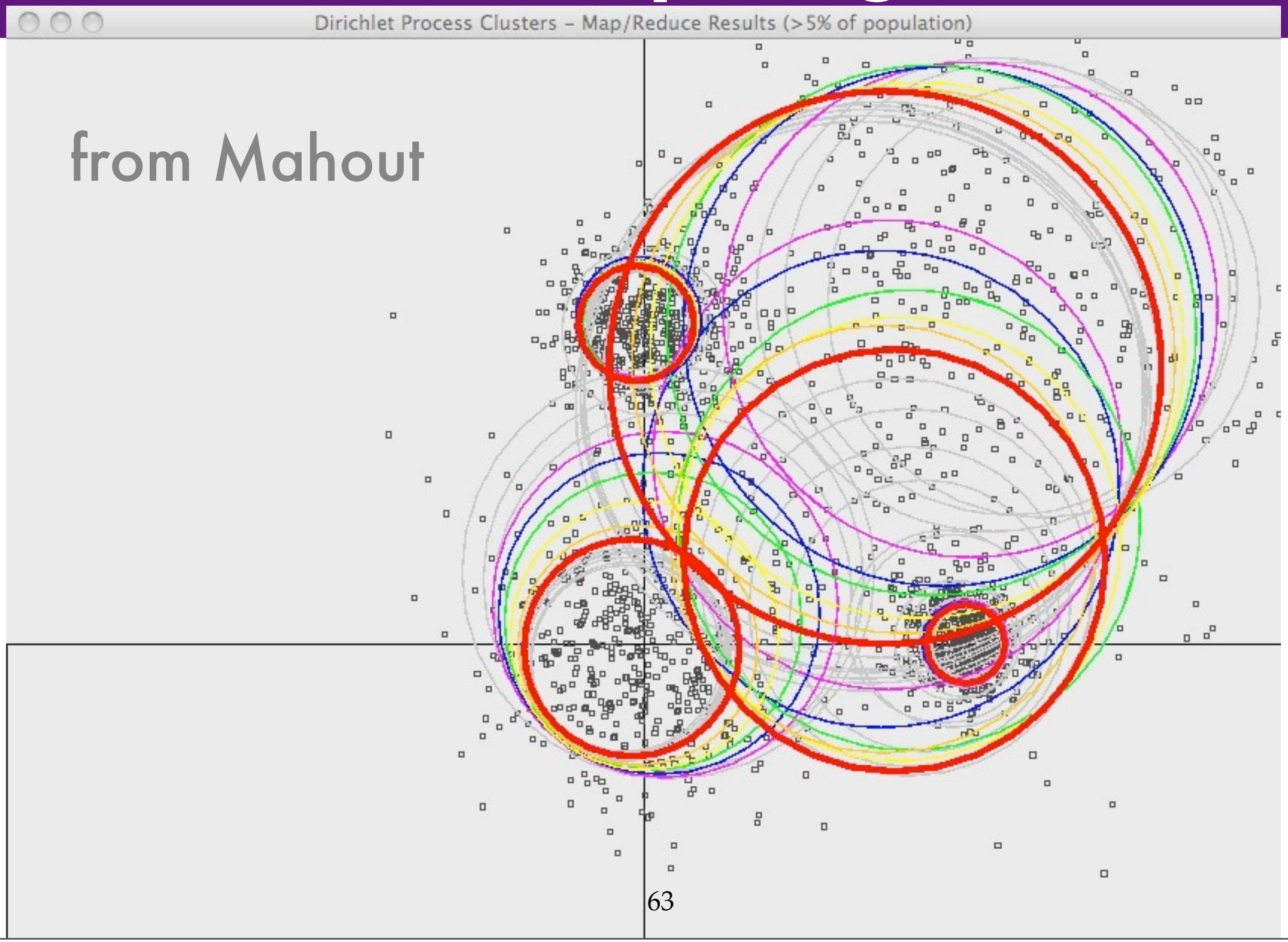


$$x \sim p(x|x') \text{ and then } x' \sim p(x'|x)$$



Sampling

from Mahout



Outlook

applications, applications

- Markov Chains
- Dynamic Programming introduction
- Hidden Markov Models
- Kalman Filter (briefly)
- Topic Models

Further reading
http://alex.smola.org/drafts/thebook.pdf