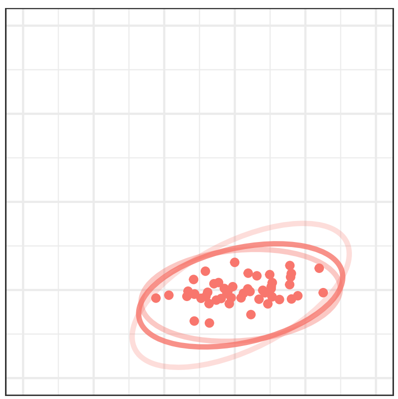
Unsupervised language acquisition

Vowel inventories and the lexicon



```
whats.that
the.dog.gie
yeah
wheres.the.doggie
…
```

Roger Levy

9.19: Computational Psycholinguistics

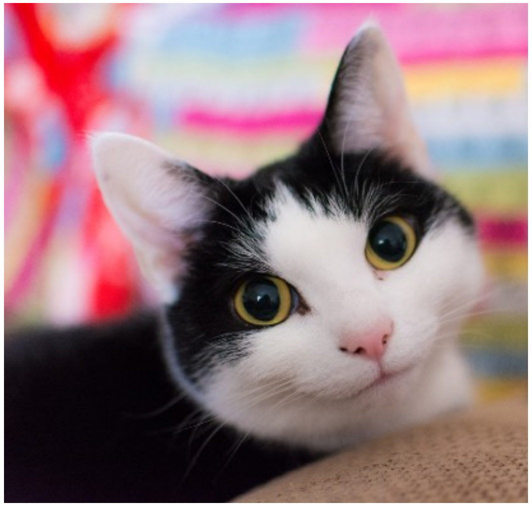
24 & 29 November 2021

First language acquisition

(Image credit RikRokYu Castro)

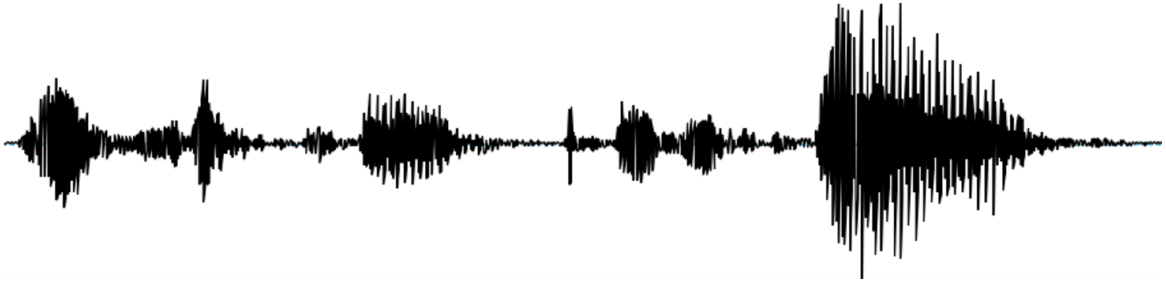
First language acquisition

(Image credit RikRokYu Castro)

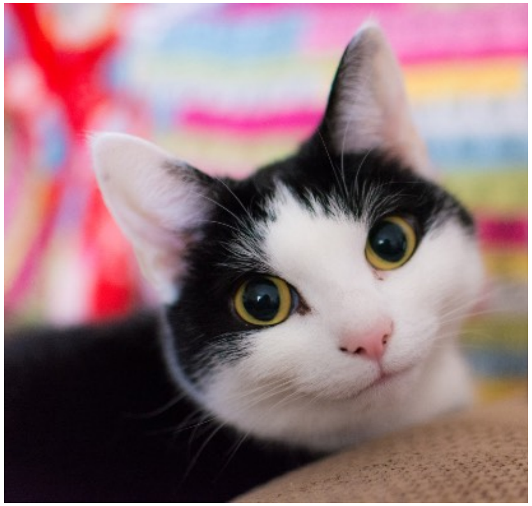


(Image credit Peter Taylor)

First language acquisition

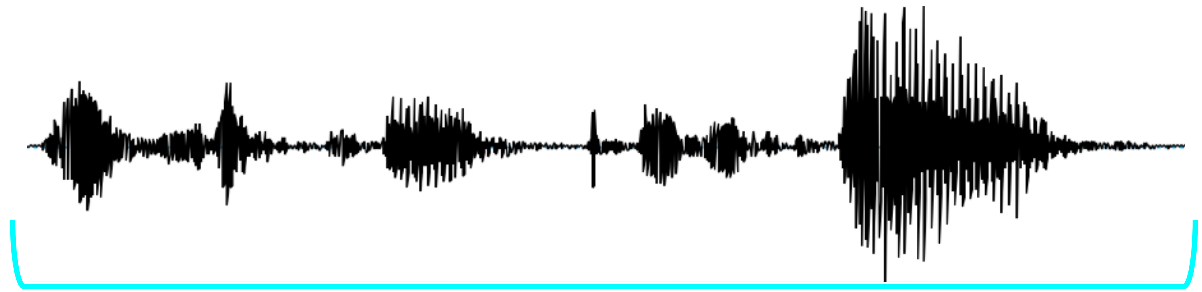


(Image credit RikRokYu Castro)

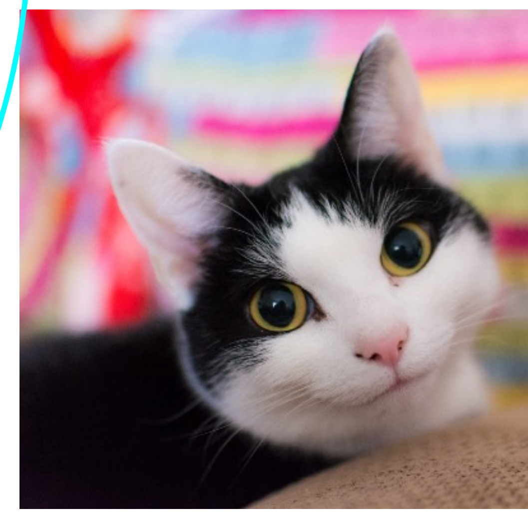


(Image credit Peter Taylor)

First language acquisition

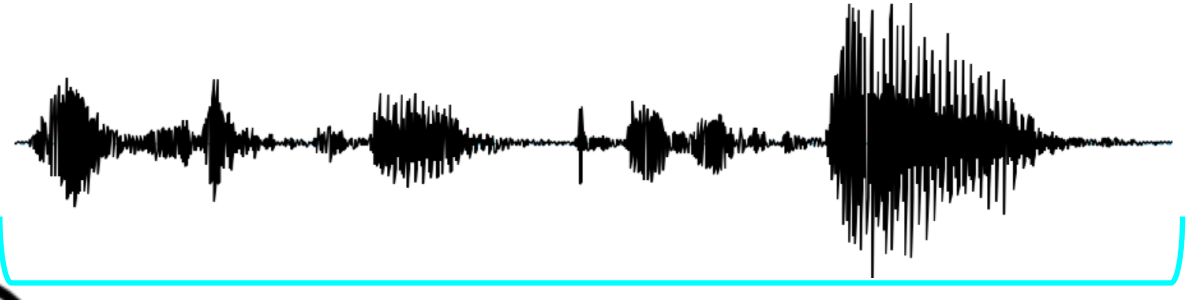


(Image credit RikRokYu Castro)

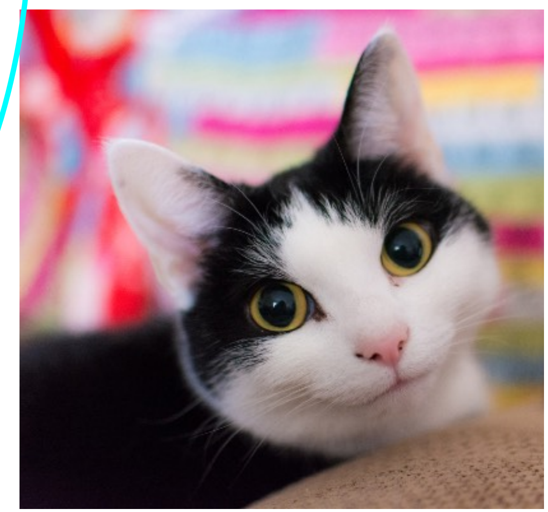


(Image credit Peter Taylor)

First language acquisition

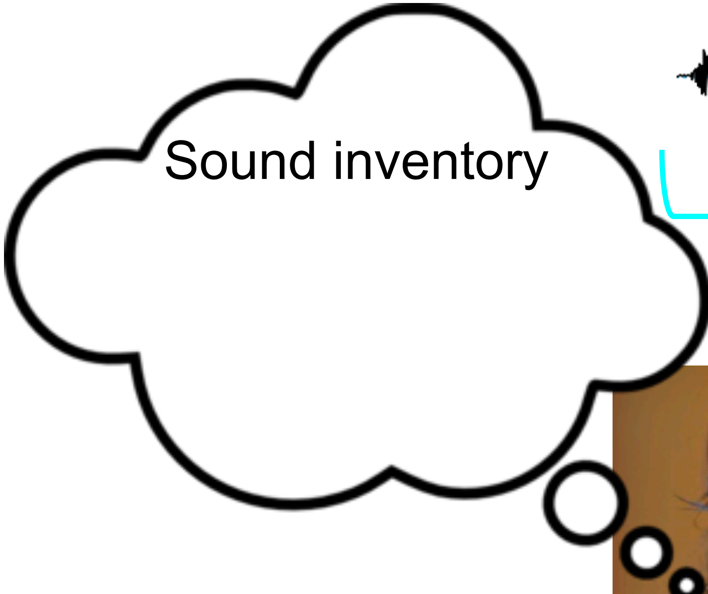


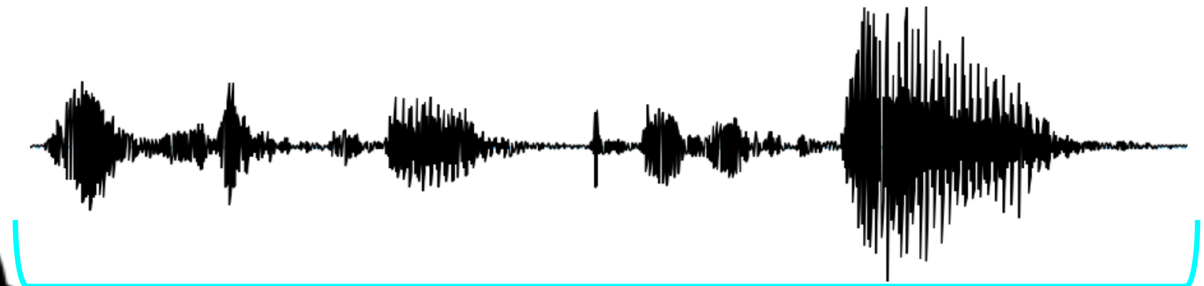
(Image credit RikRokYu Castro)



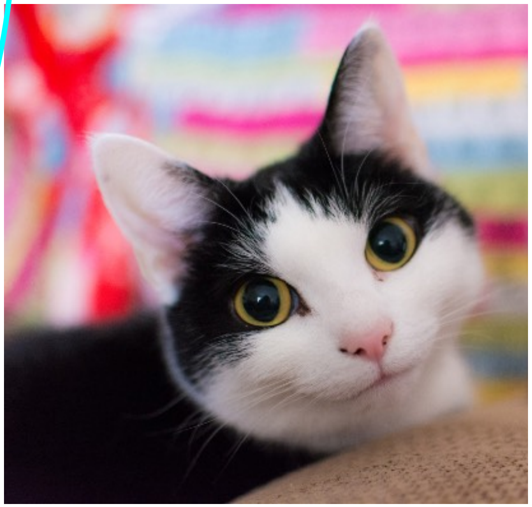
(Image credit Peter Taylor)

First language acquisition

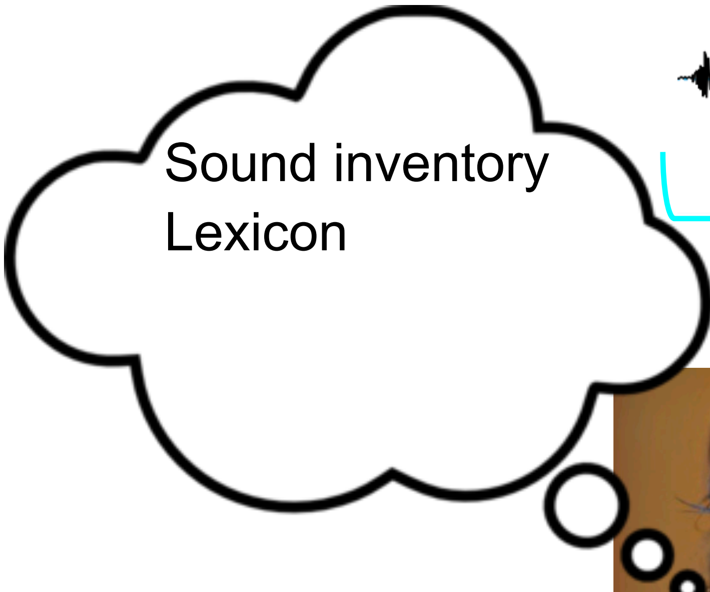


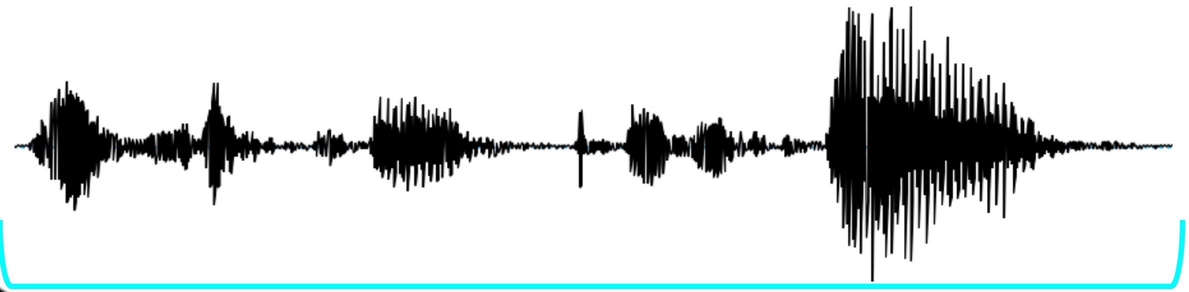



(Image credit RikRokYu Castro)

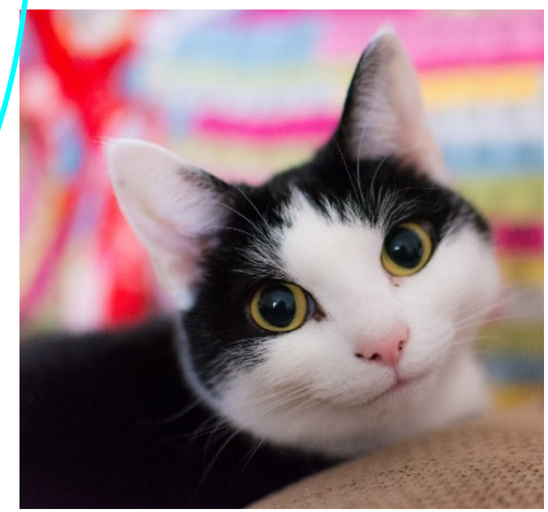

(Image credit Peter Taylor)

First language acquisition



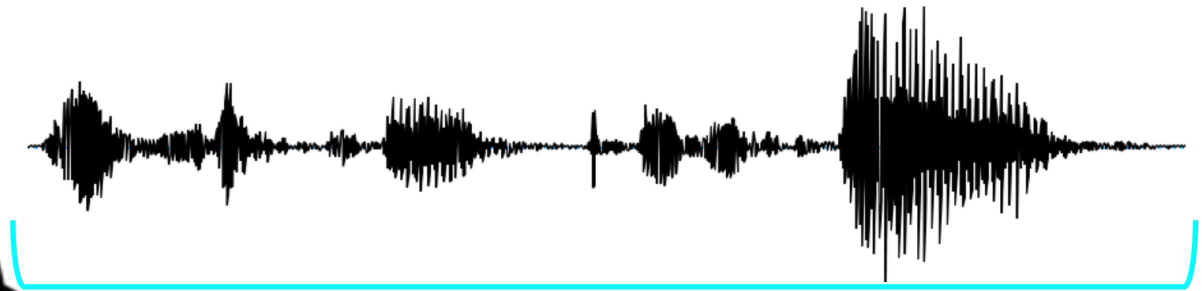


(Image credit RikRokYu Castro)



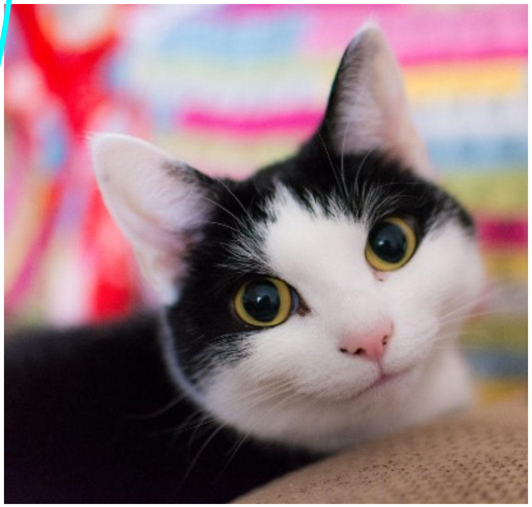
(Image credit Peter Taylor)

First language acquisition



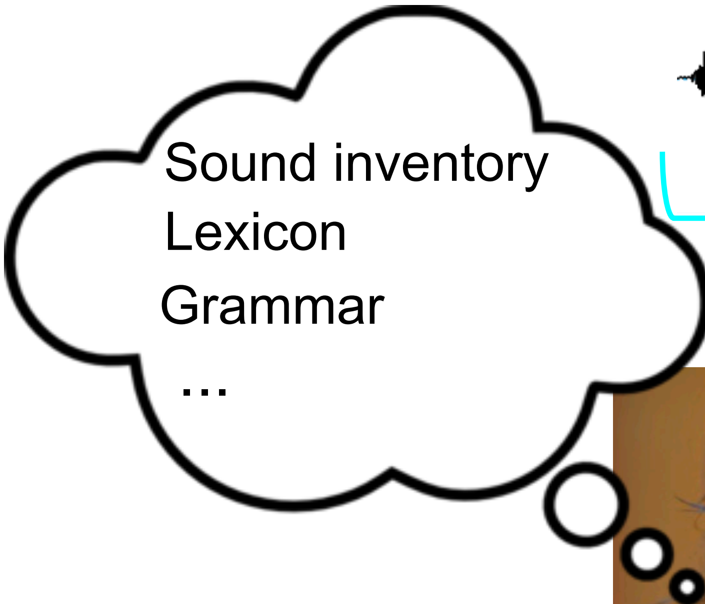
Sound inventory
Lexicon
Grammar
...

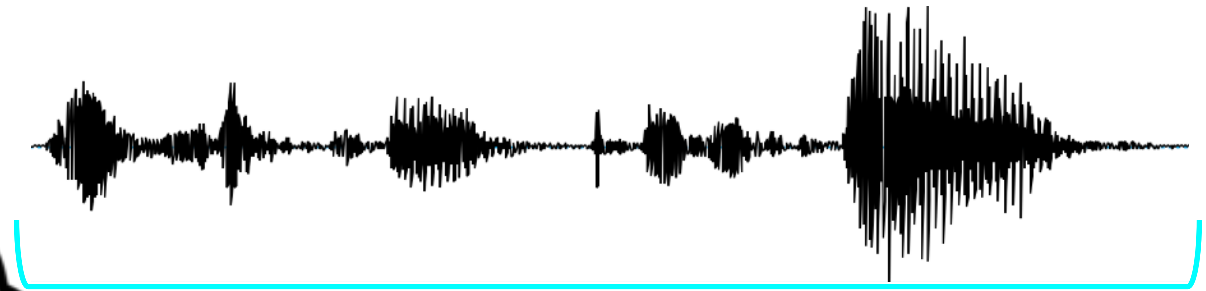
(Image credit RikRokYu Castro)



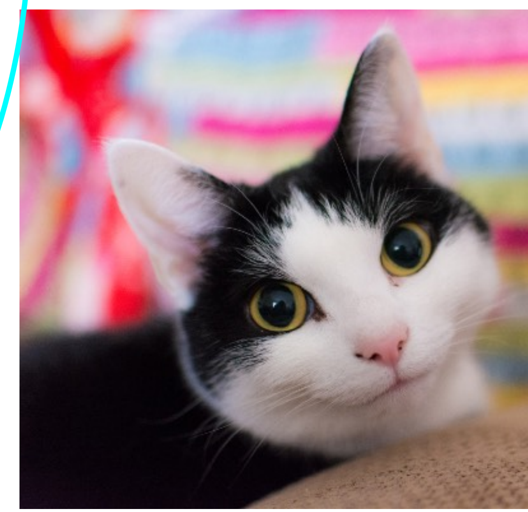
(Image credit Peter Taylor)

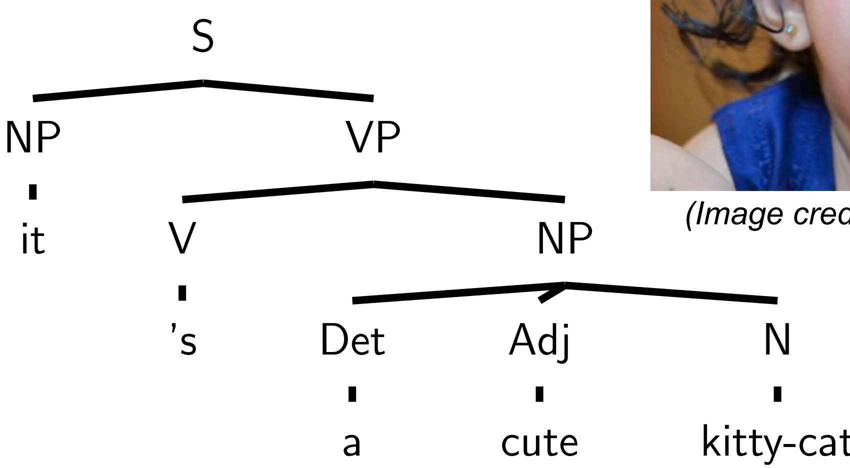
First language acquisition





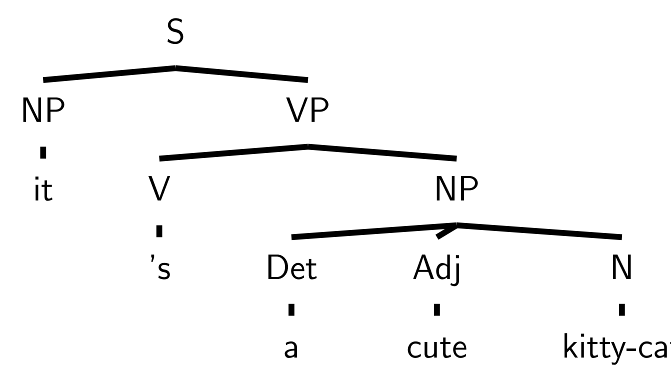

(Image credit RikRokYu Castro)


(Image credit Peter Taylor)



Language acquisition: unsupervised learning

How do inductive bias and positive* linguistic input data alone deliver the generalizations underlying native-speaker linguistic competence and performance?

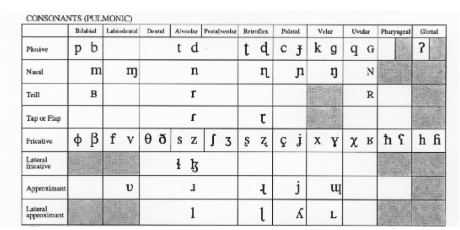


*No negative evidence!

Language learning as hierarchical Bayesian inference

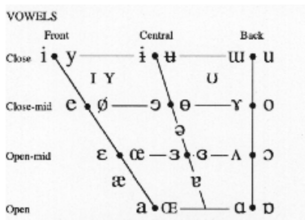
Language learning as hierarchical Bayesian inference

$$P(\text{phonemes}|\text{speech input}) \propto P(\text{speech input}|\text{phonemes})P(\text{phonemes})$$



CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			



Language learning as hierarchical Bayesian inference

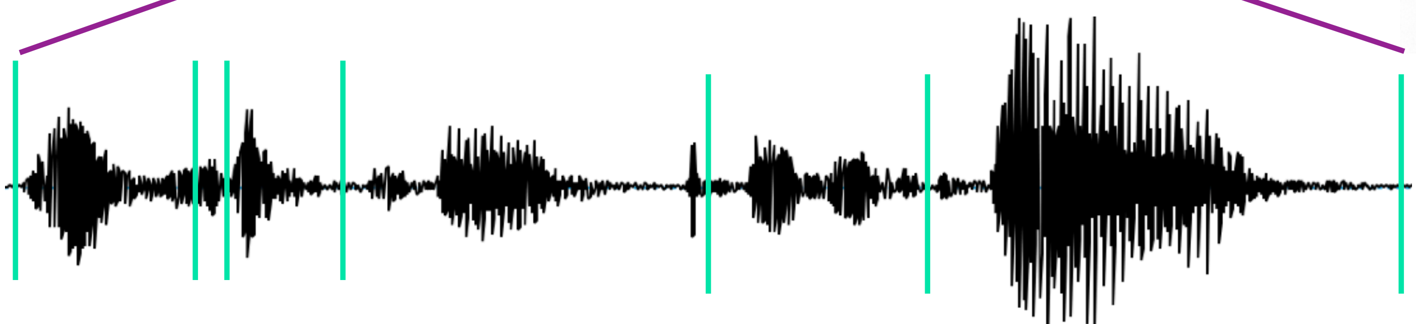
a	**dog**
the	**kitty-cat**
it	**barks**
is	**today**
cute	**near**
tall	**ever**
...	**...**

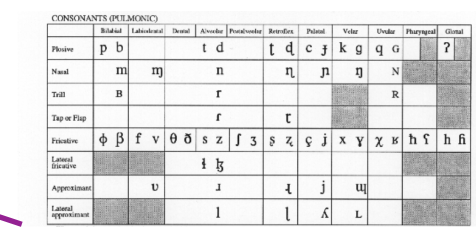
$$P(\text{words}|\text{phonemes}) \propto P(\text{phonemes}|\text{words})P(\text{words})$$

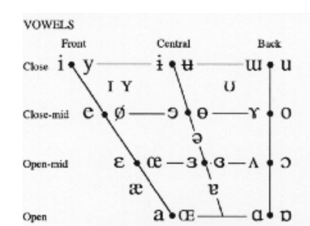
it's a cute kitty-cat

$$P(\text{phonemes}|\text{speech input}) \propto P(\text{speech input}|\text{phonemes})P(\text{phonemes})$$

ɪt|s|ə|kjut|kɪɾi|kæt|

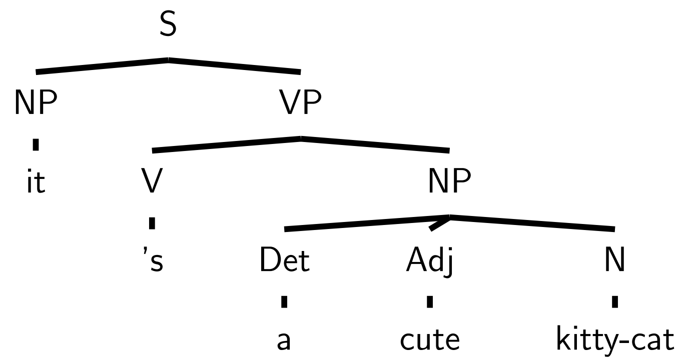






Language learning as hierarchical Bayesian inference

$$P(\text{grammar}|\text{words}) \propto P(\text{words}|\text{grammar})P(\text{grammar})$$




```
S  → NP VP
NP → Det N
NP → Det Adj N
VP → V NP
...
```

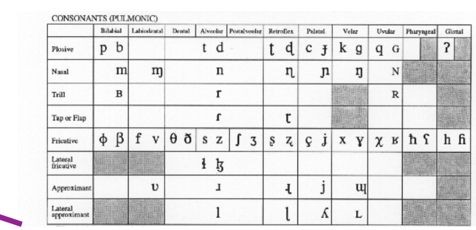
$$P(\text{words}|\text{phonemes}) \propto P(\text{phonemes}|\text{words})P(\text{words})$$

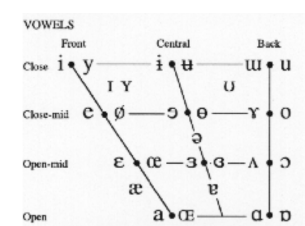
a	dog
the	kitty-cat
it	barks
is	today
cute	near
tall	ever
...	...

it's a cute kitty-cat

$$P(\text{phonemes}|\text{speech input}) \propto P(\text{speech input}|\text{phonemes})P(\text{phonemes})$$

ɪt|s|ə|kjut|kɪɾi|kæt|

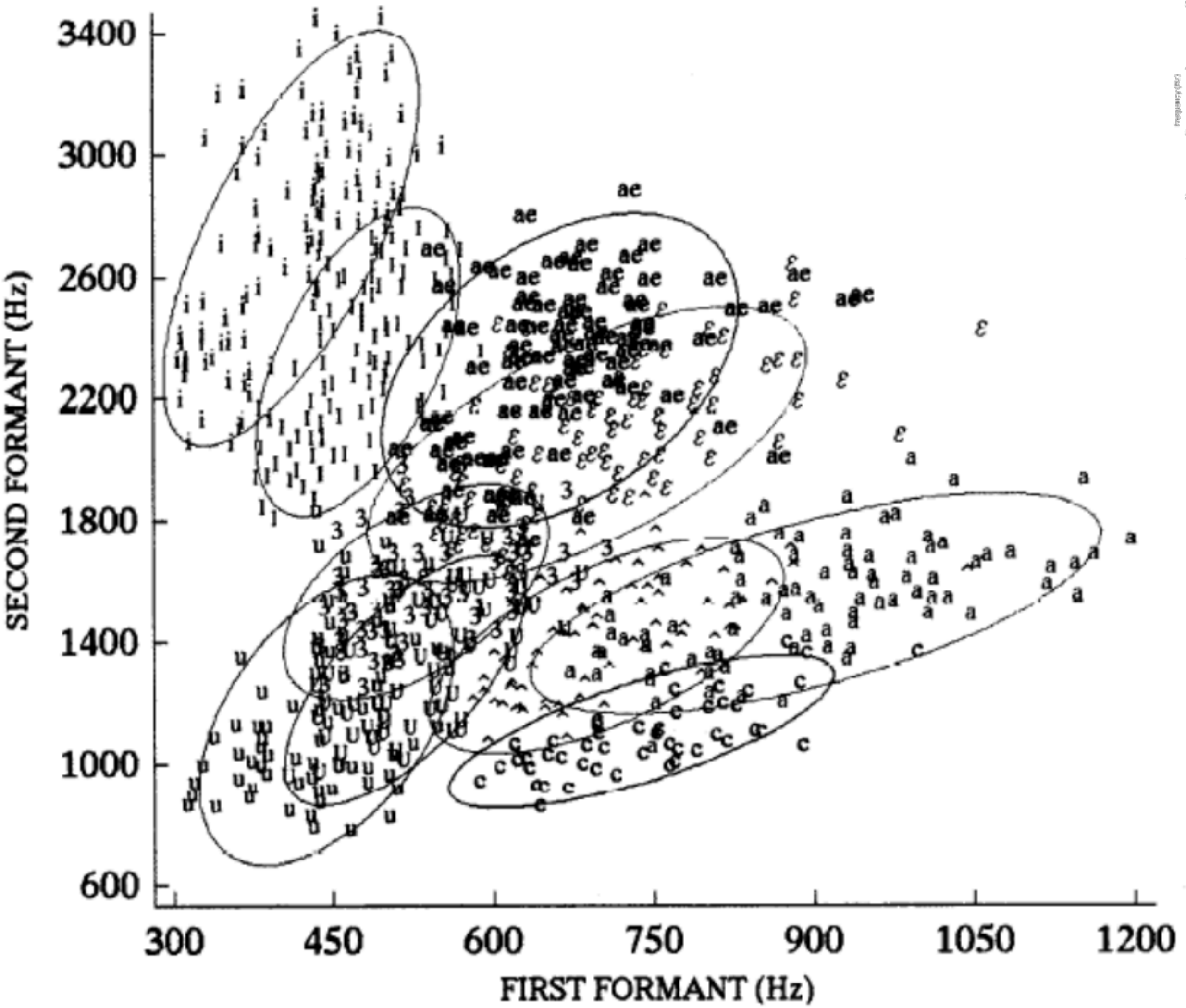


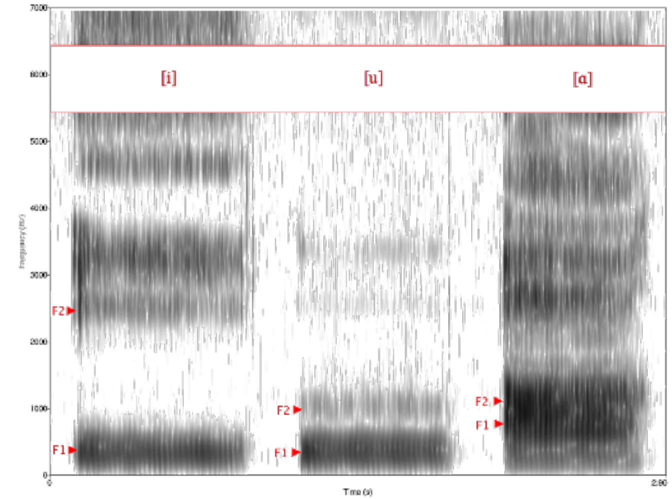


Today's agenda

- We'll illustrate unsupervised learning by examining two central parts of the language acquisition problem:
  - Learning phonetic categories
  - Segmenting input into words
- In the process, we'll cover some key Bayesian methods for unsupervised learning
  - Conjugate priors
  - Gibbs Sampling

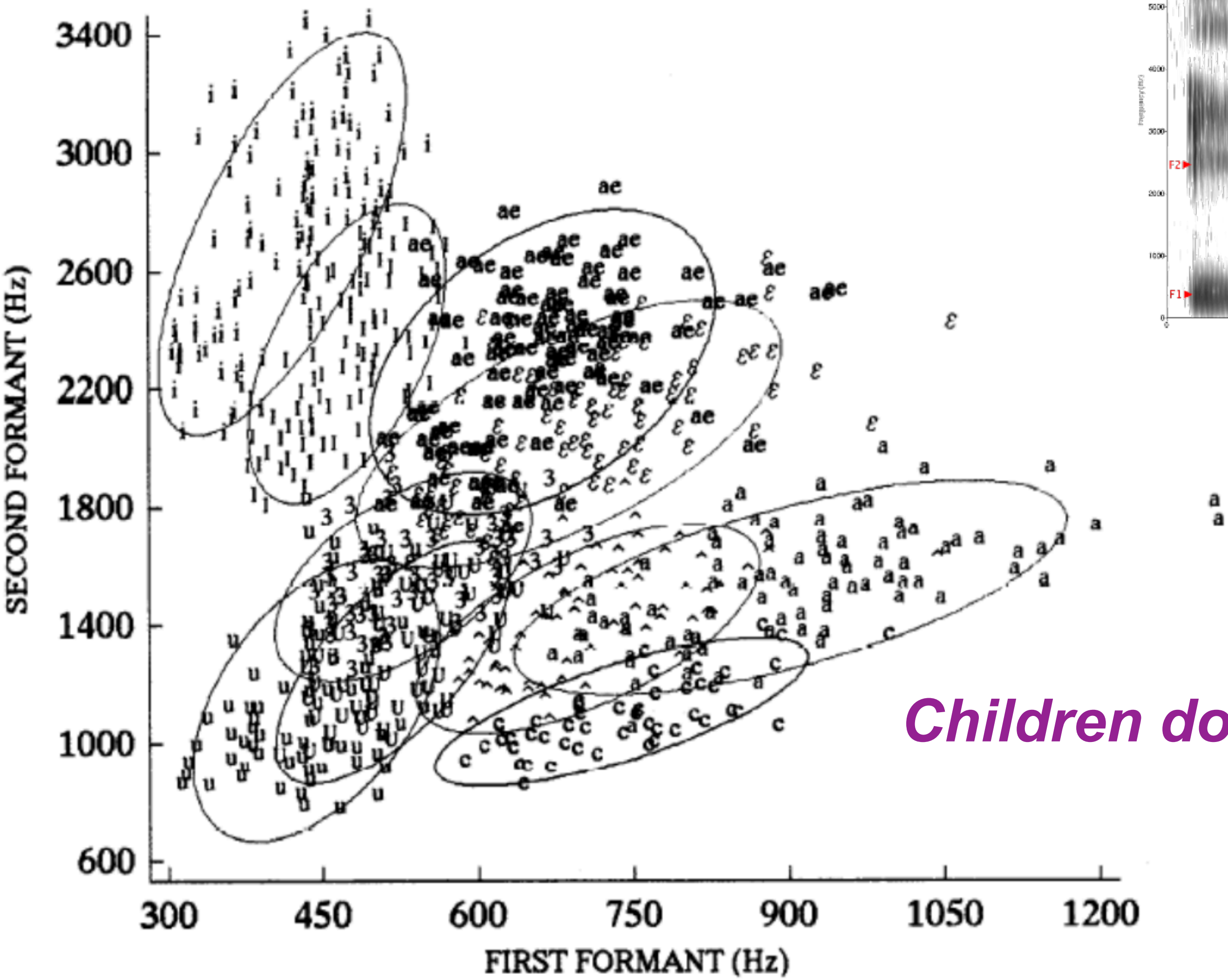
English vowel inventory, in formant space

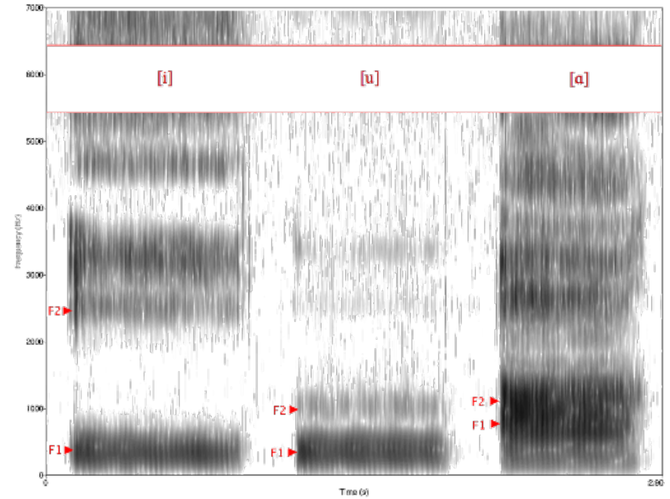




(Peterson & Barney, 1952)

English vowel inventory, in formant space

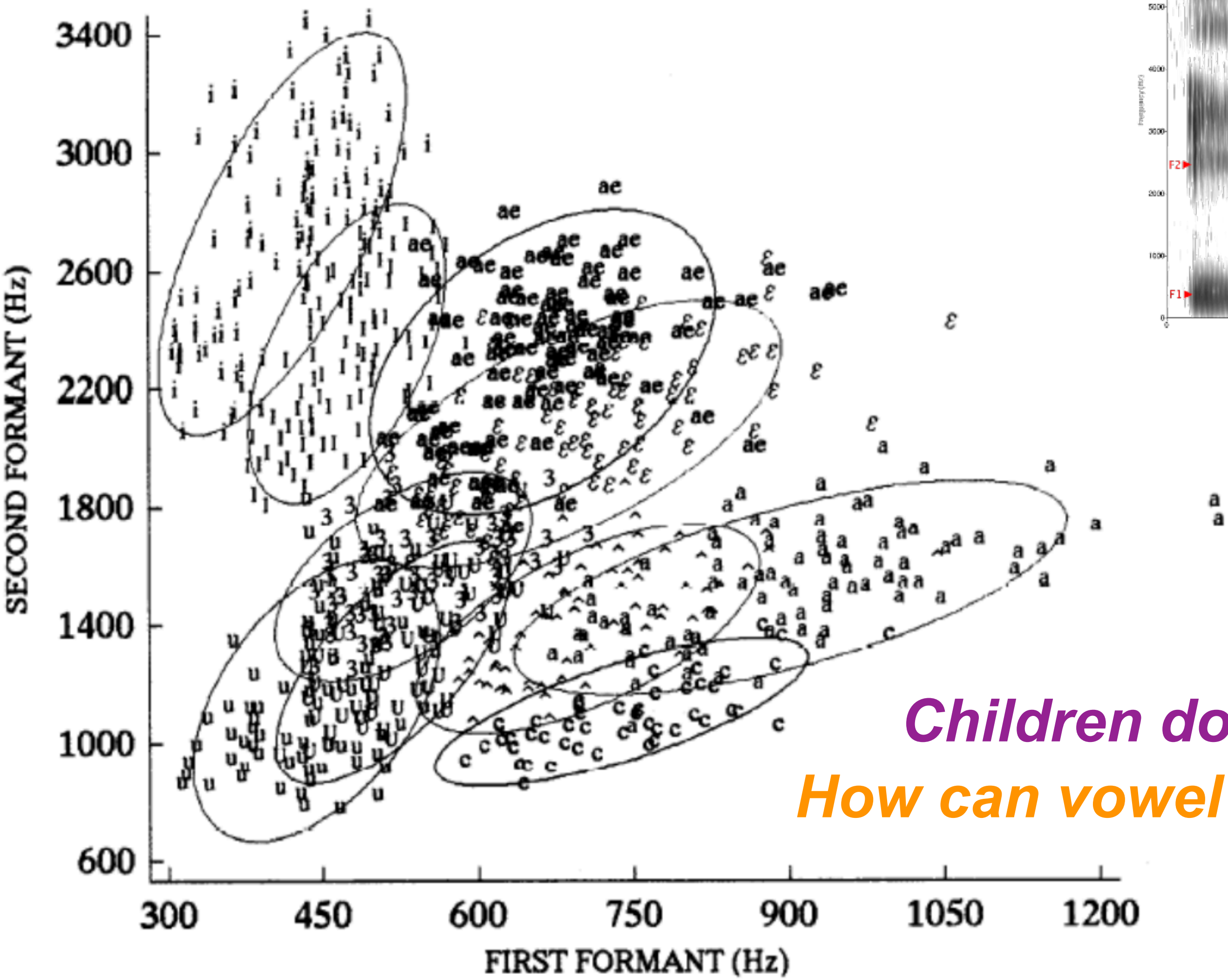


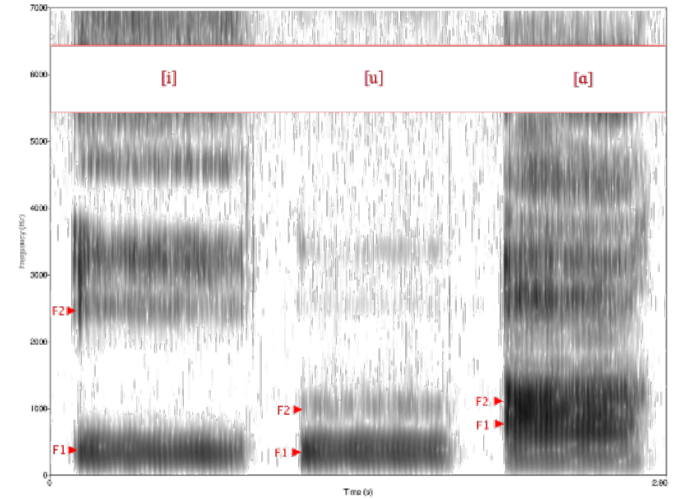


Children do not get the labels!

(Peterson & Barney, 1952)

English vowel inventory, in formant space



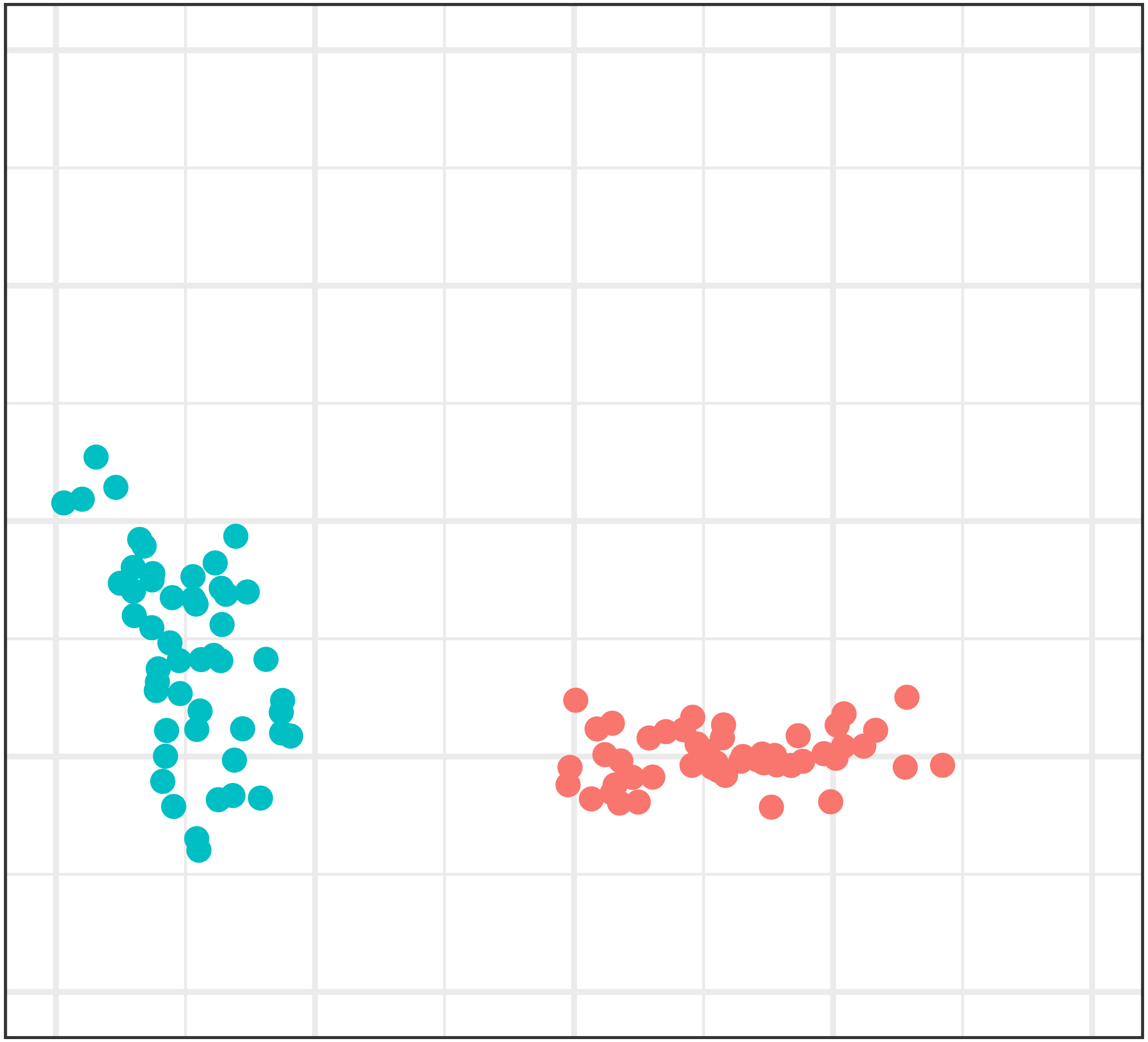


Children do not get the labels!

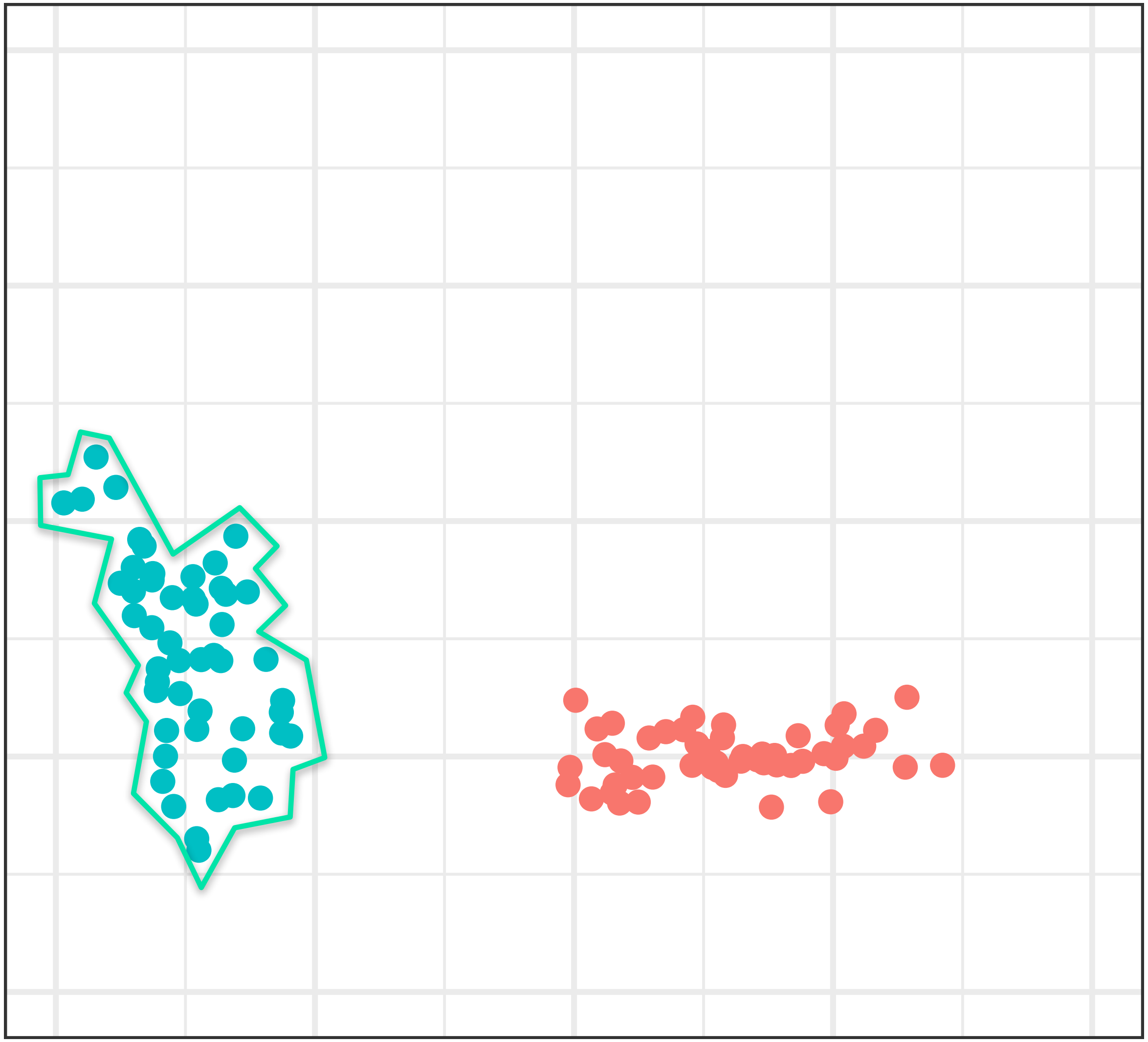
How can vowel forms be learned?

(Peterson & Barney, 1952)

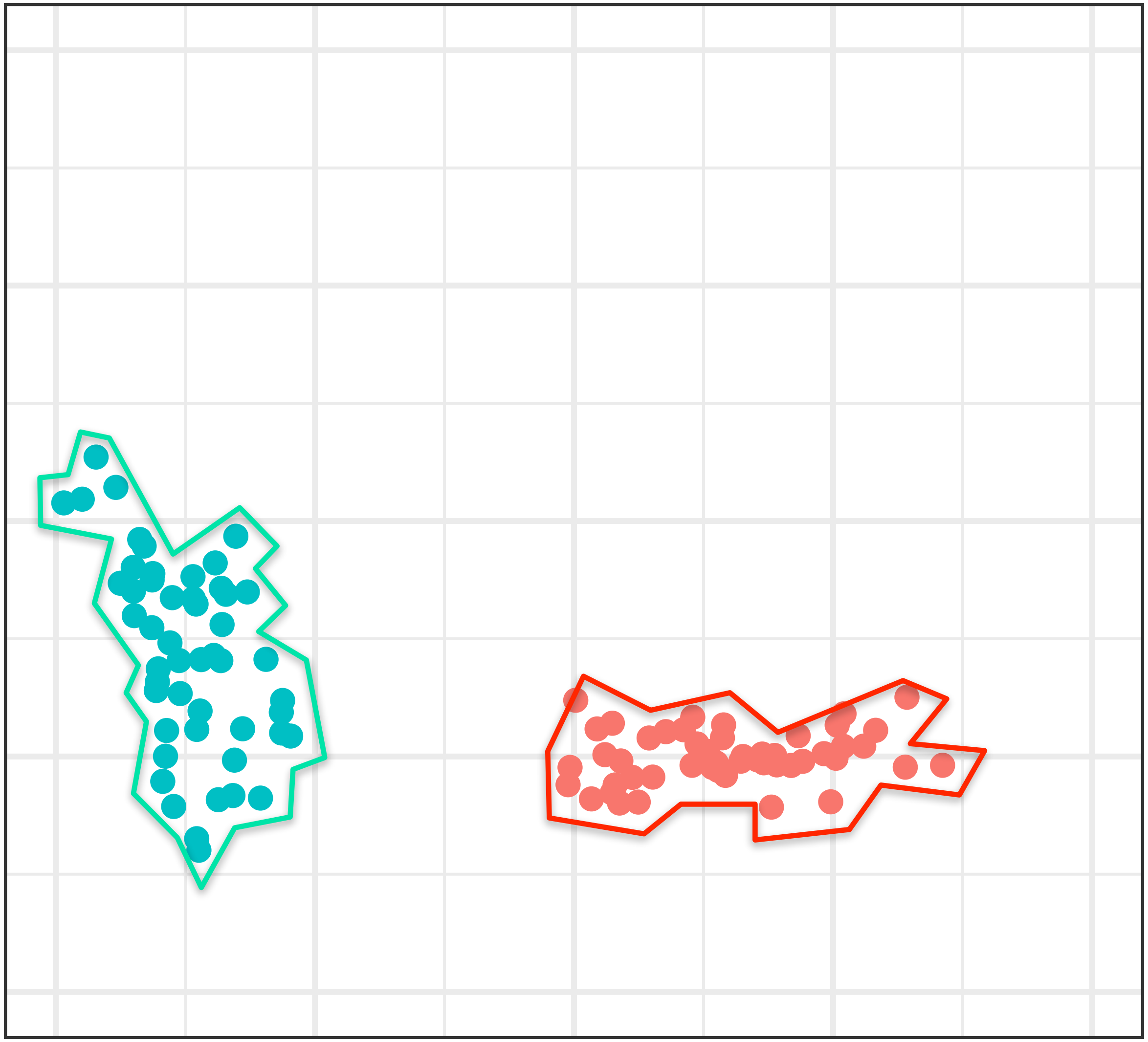
Category learning in continuous spaces



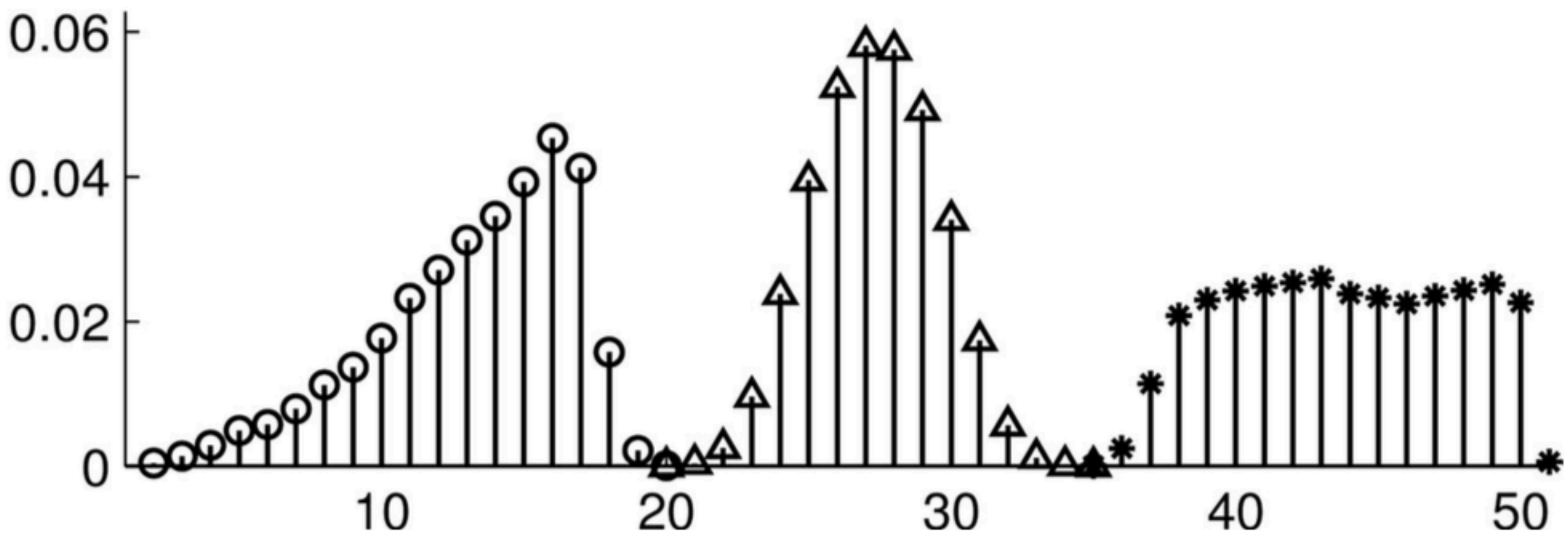
Category learning in continuous spaces



Category learning in continuous spaces



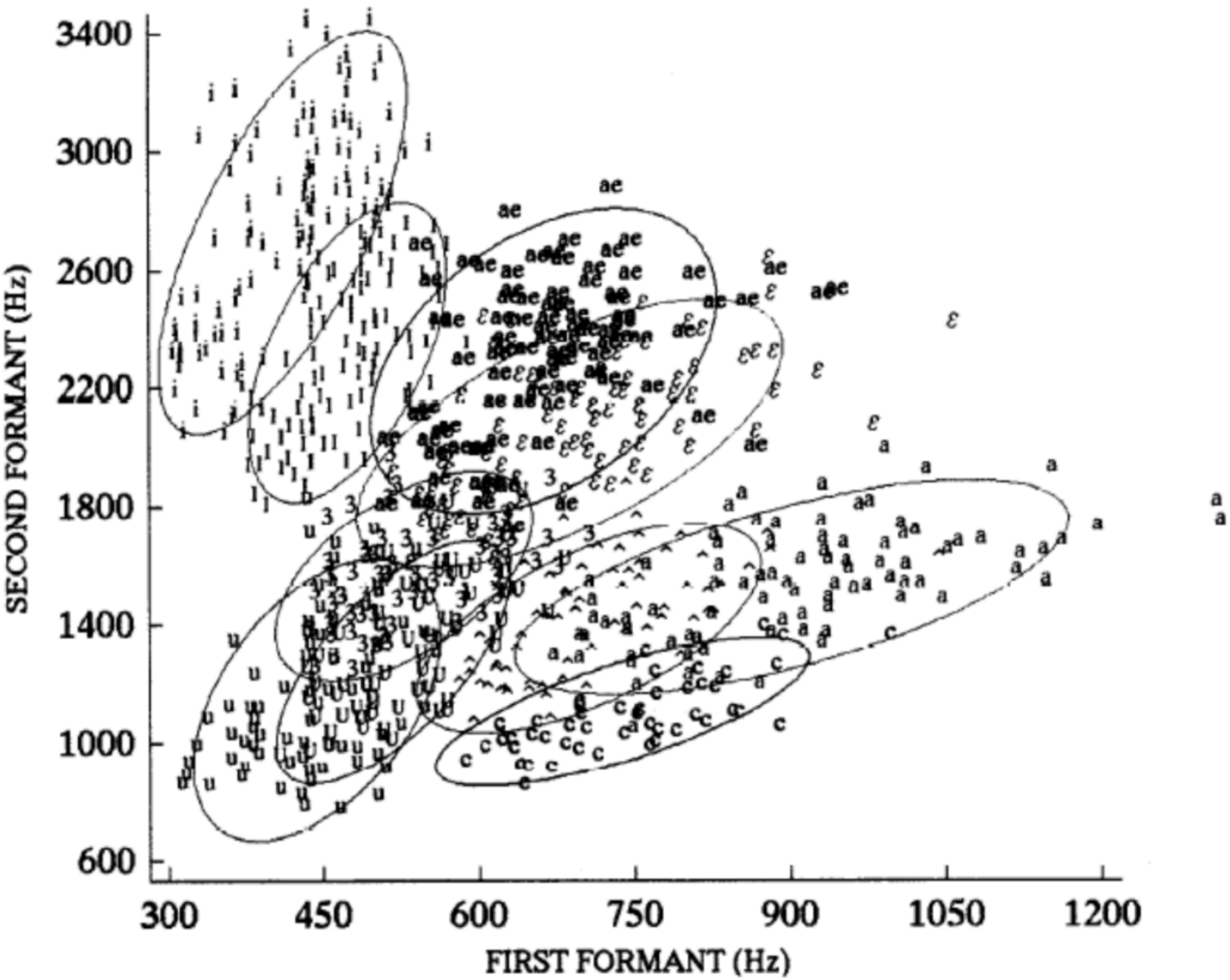
One approach to mixture estimation



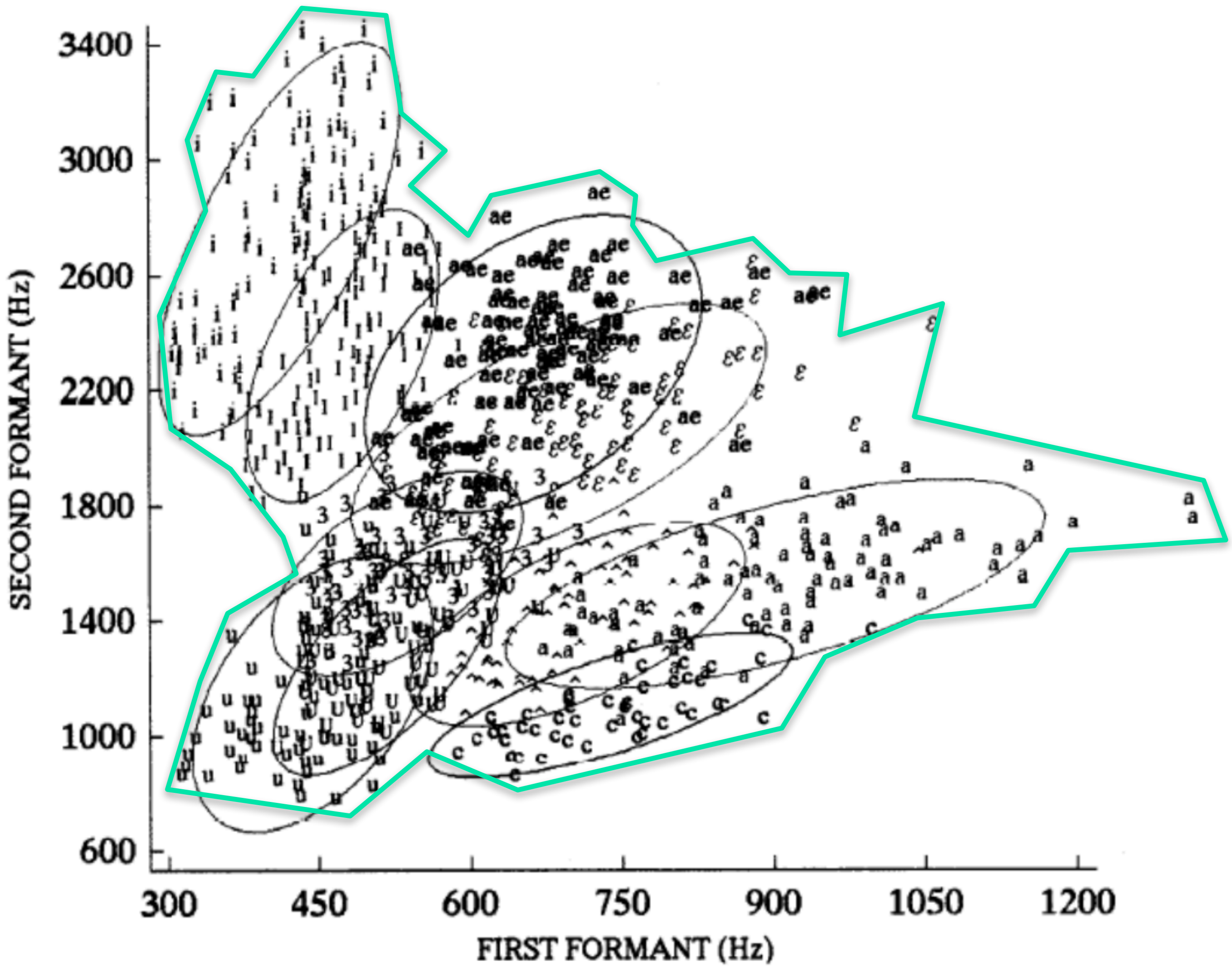
"Topographic online mixture estimation" — highly nonparametric

(Vallabha et al., 2007)

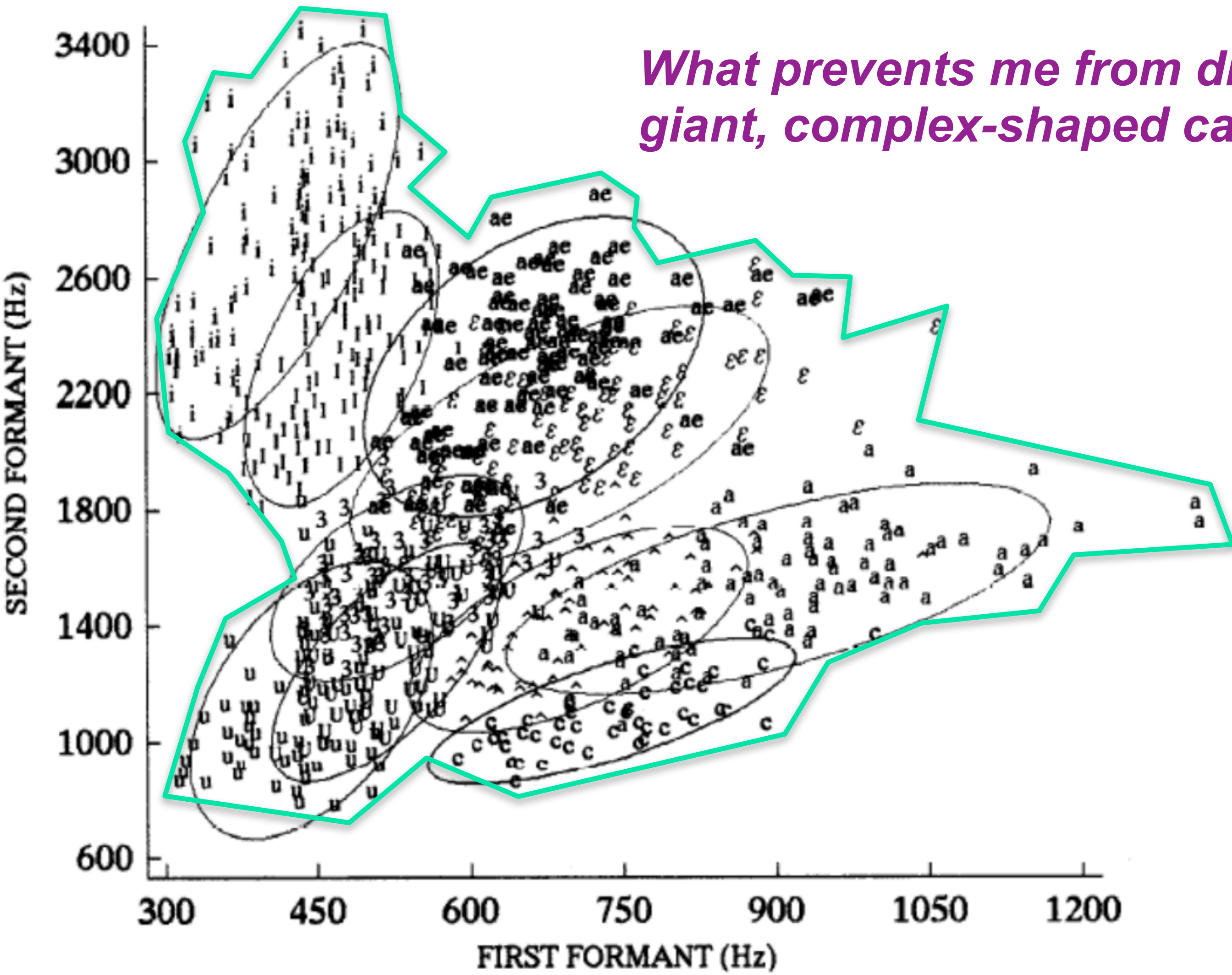
Limitations



Limitations



Limitations



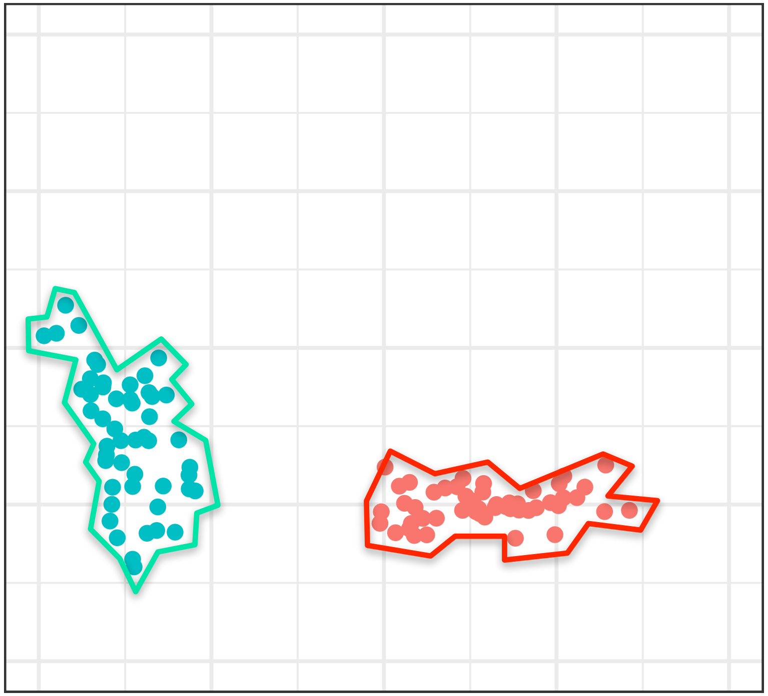
What prevents me from drawing one giant, complex-shaped category?

Parametric categories

- One possibility: a *strong* inductive bias for category shapes
- Classic approach: data must be a mixture of *Gaussian* categories

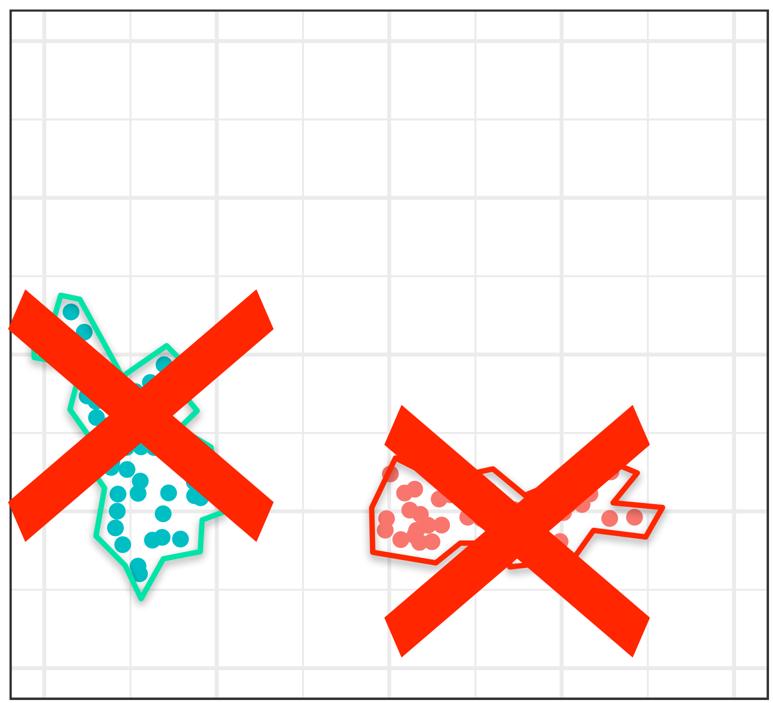
Parametric categories

- One possibility: a *strong* inductive bias for category shapes
- Classic approach: data must be a mixture of *Gaussian* categories



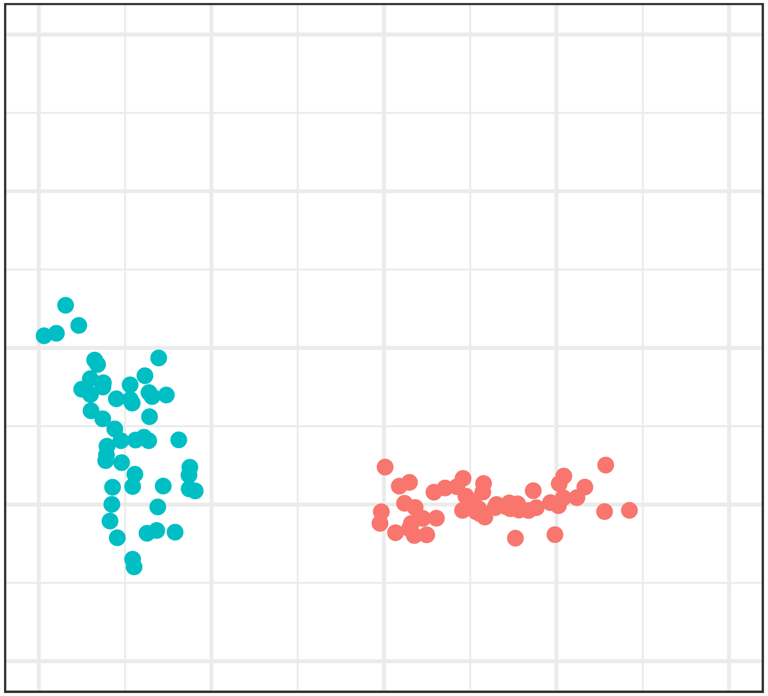
Parametric categories

- One possibility: a *strong* inductive bias for category shapes
- Classic approach: data must be a mixture of *Gaussian* categories



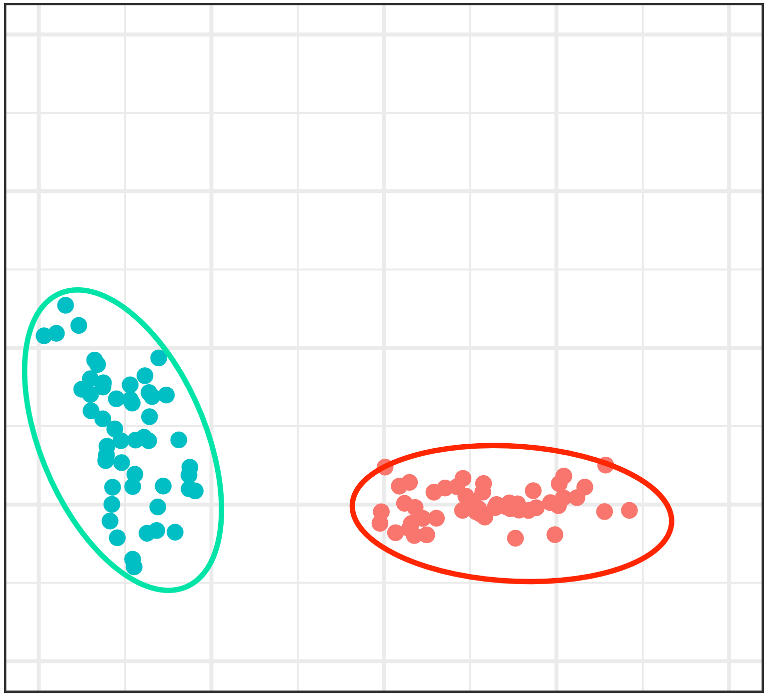
Parametric categories

- One possibility: a *strong* inductive bias for category shapes
- Classic approach: data must be a mixture of *Gaussian* categories

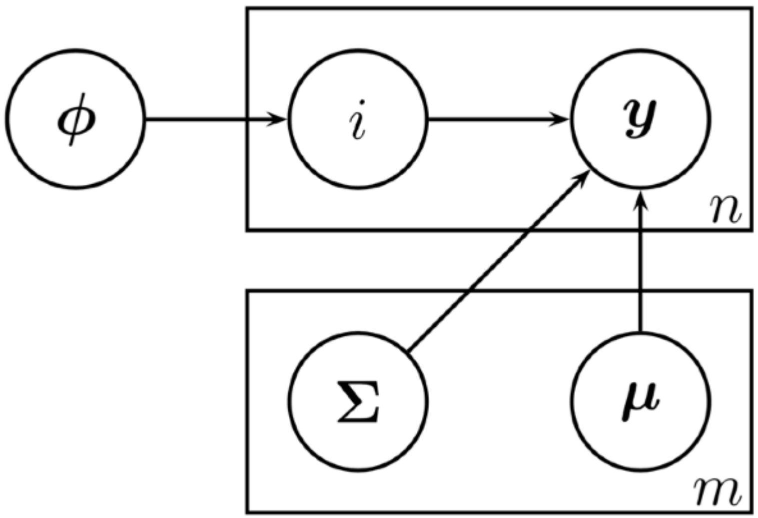


Parametric categories

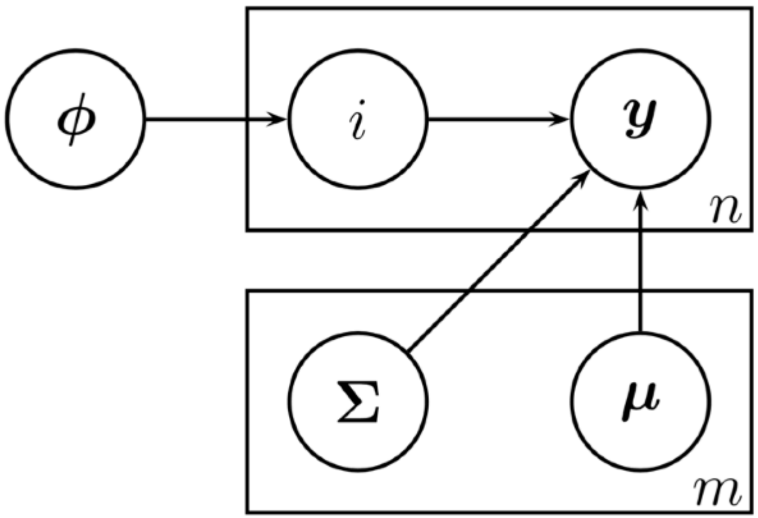
- One possibility: a *strong* inductive bias for category shapes
- Classic approach: data must be a mixture of *Gaussian* categories



Mixture of Gaussians

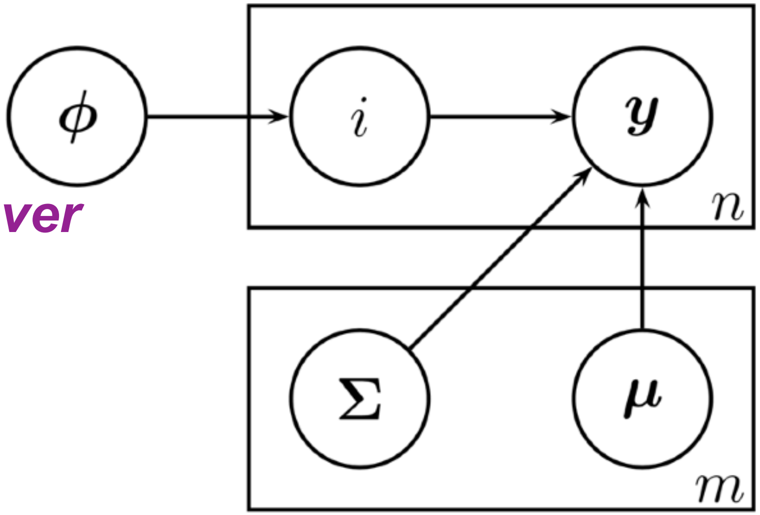


Mixture of Gaussians



m Gaussian categories, each with its own mean & covariance matrix

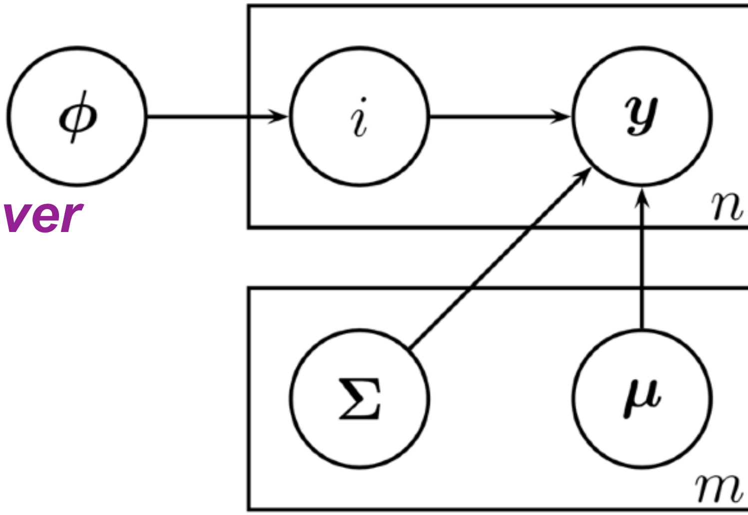
Mixture of Gaussians



Multinomial over m categories

m Gaussian categories, each with its own mean & covariance matrix

Mixture of Gaussians

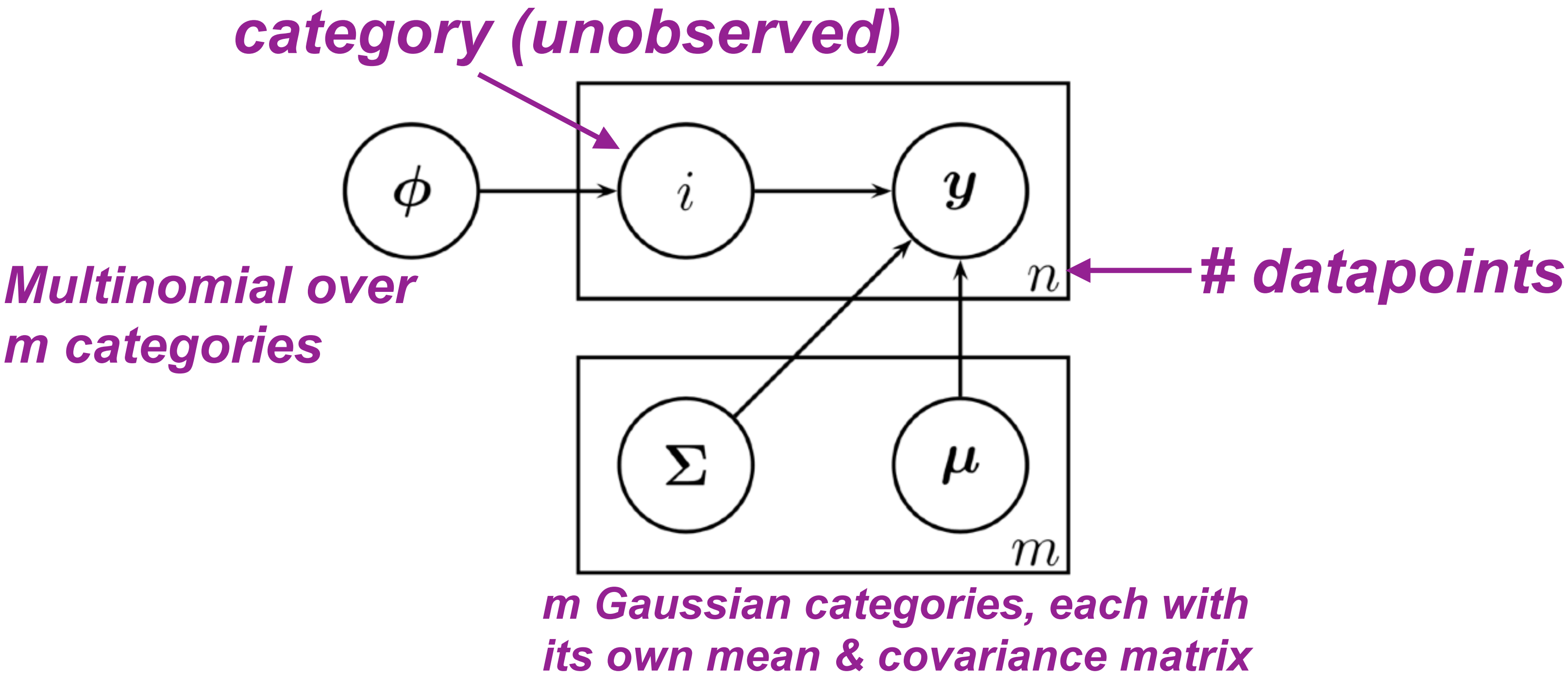


Multinomial over m categories

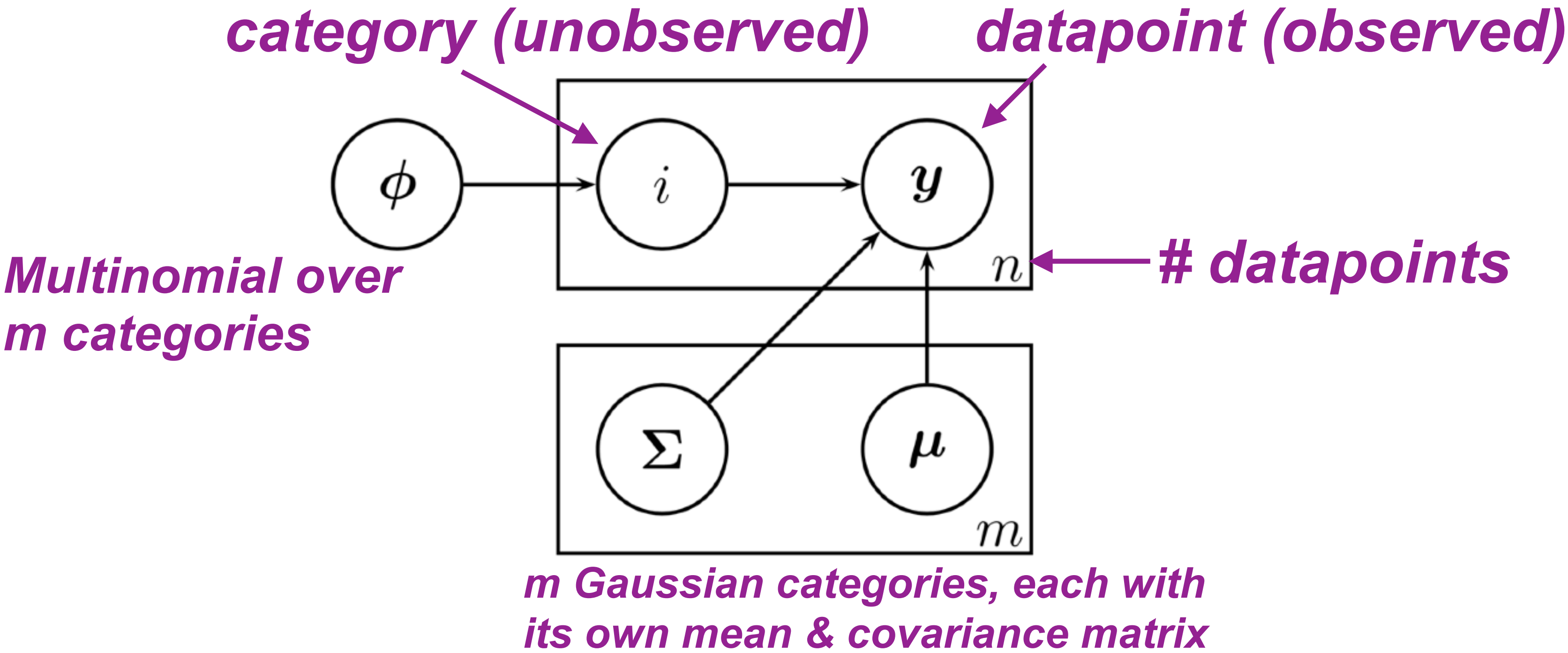
 # datapoints

m Gaussian categories, each with its own mean & covariance matrix

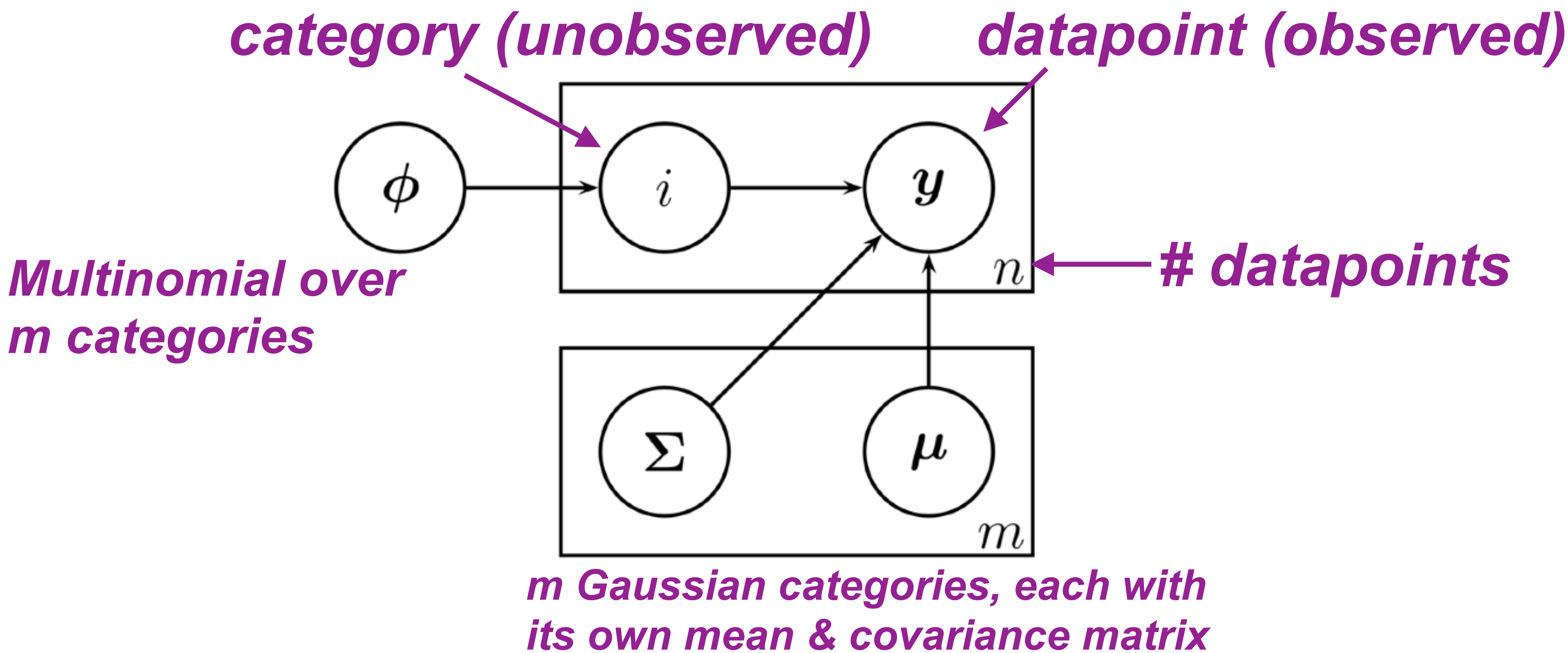
Mixture of Gaussians



Mixture of Gaussians



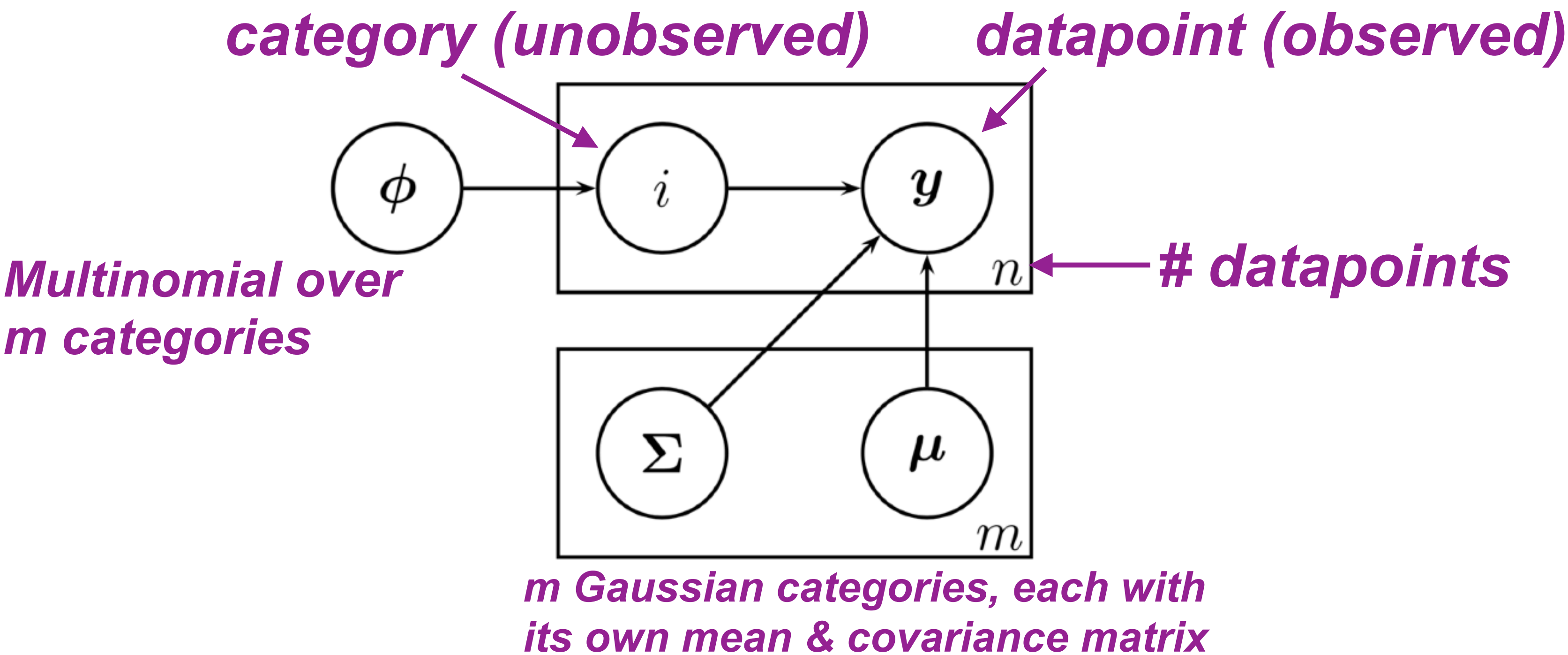
Mixture of Gaussians



- Standard inference techniques we might think about:
  - Maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma, \mu, \phi} P(\mathbf{y} | \Sigma, \mu, \phi)$$

Mixture of Gaussians



- Standard inference techniques we might think about:
  - Maximum likelihood

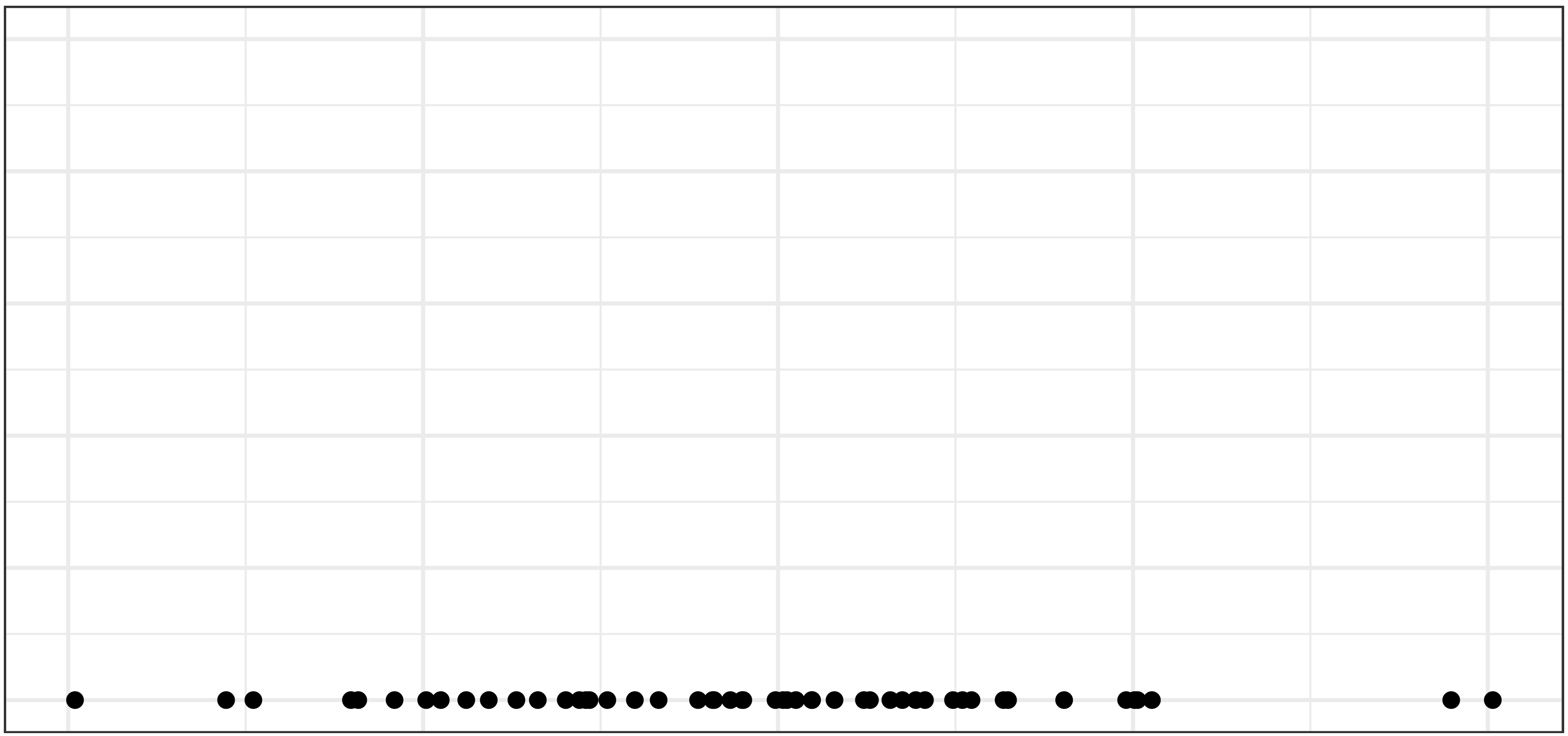
$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma,\mu,\phi)$$

  - Bayesian inference

$$P(\Sigma,\mu,\phi|\mathbf{y}) \propto P(\mathbf{y}|\Sigma,\mu,\phi)P(\Sigma,\mu,\phi)$$

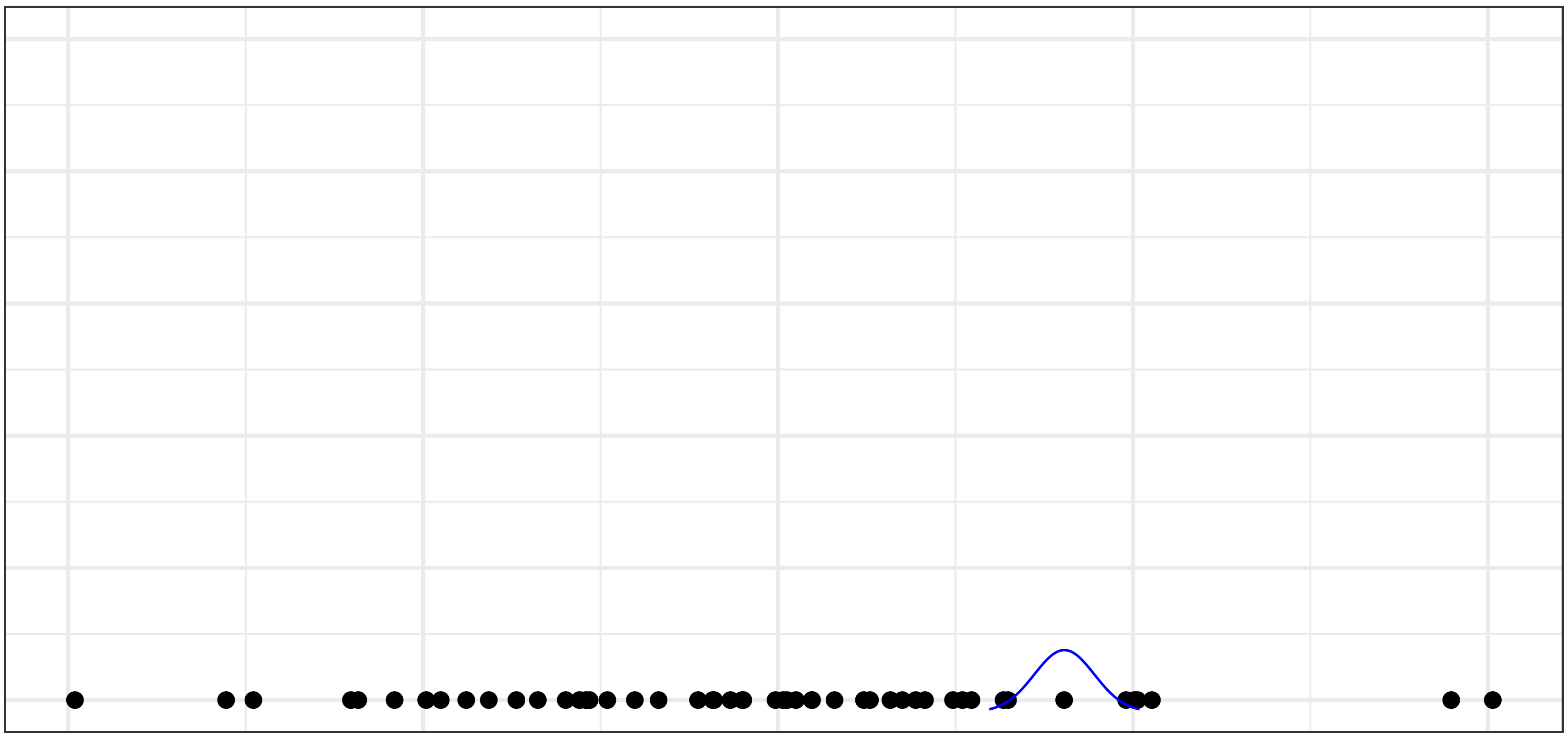
Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma, \mu, \phi)$$



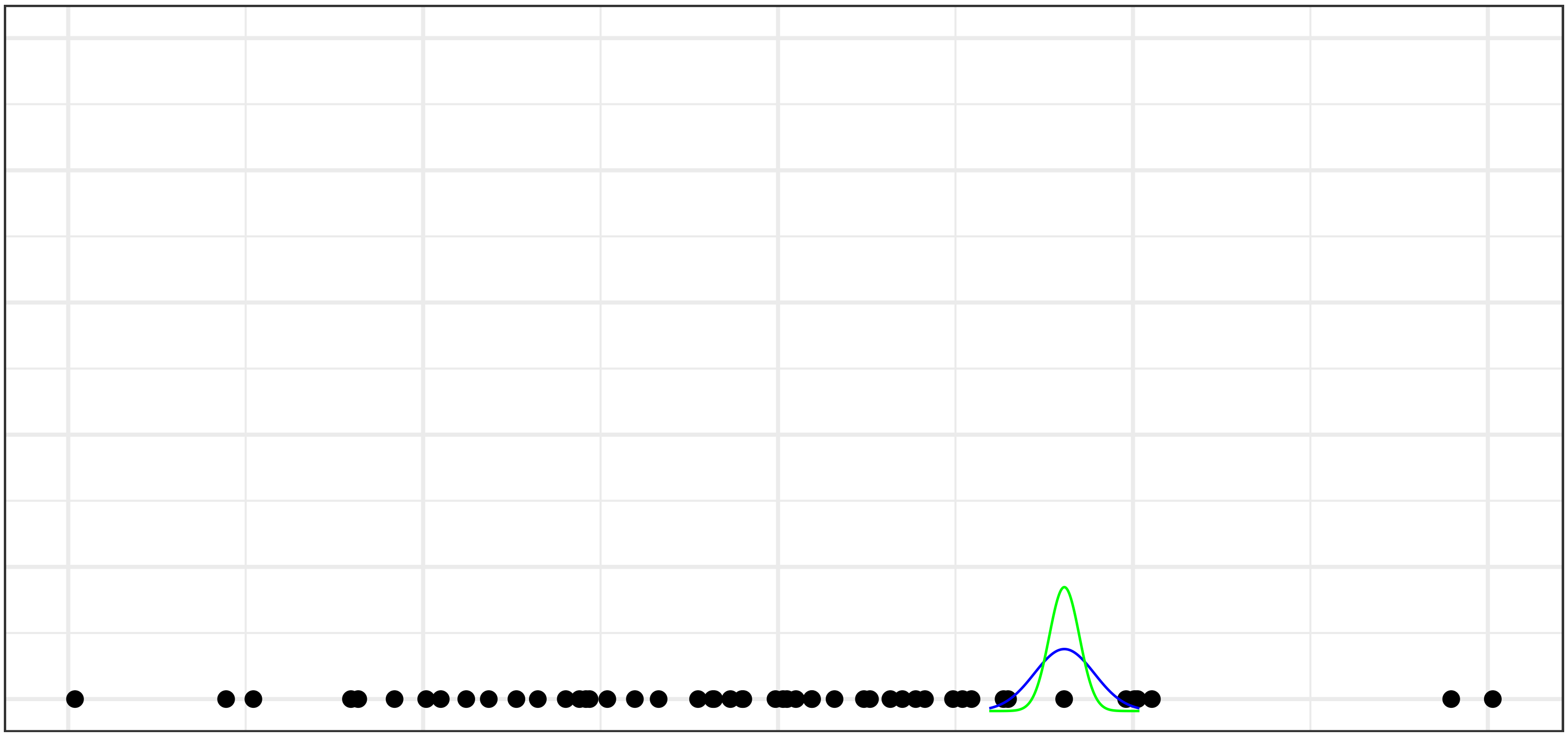
Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y} | \Sigma, \mu, \phi)$$



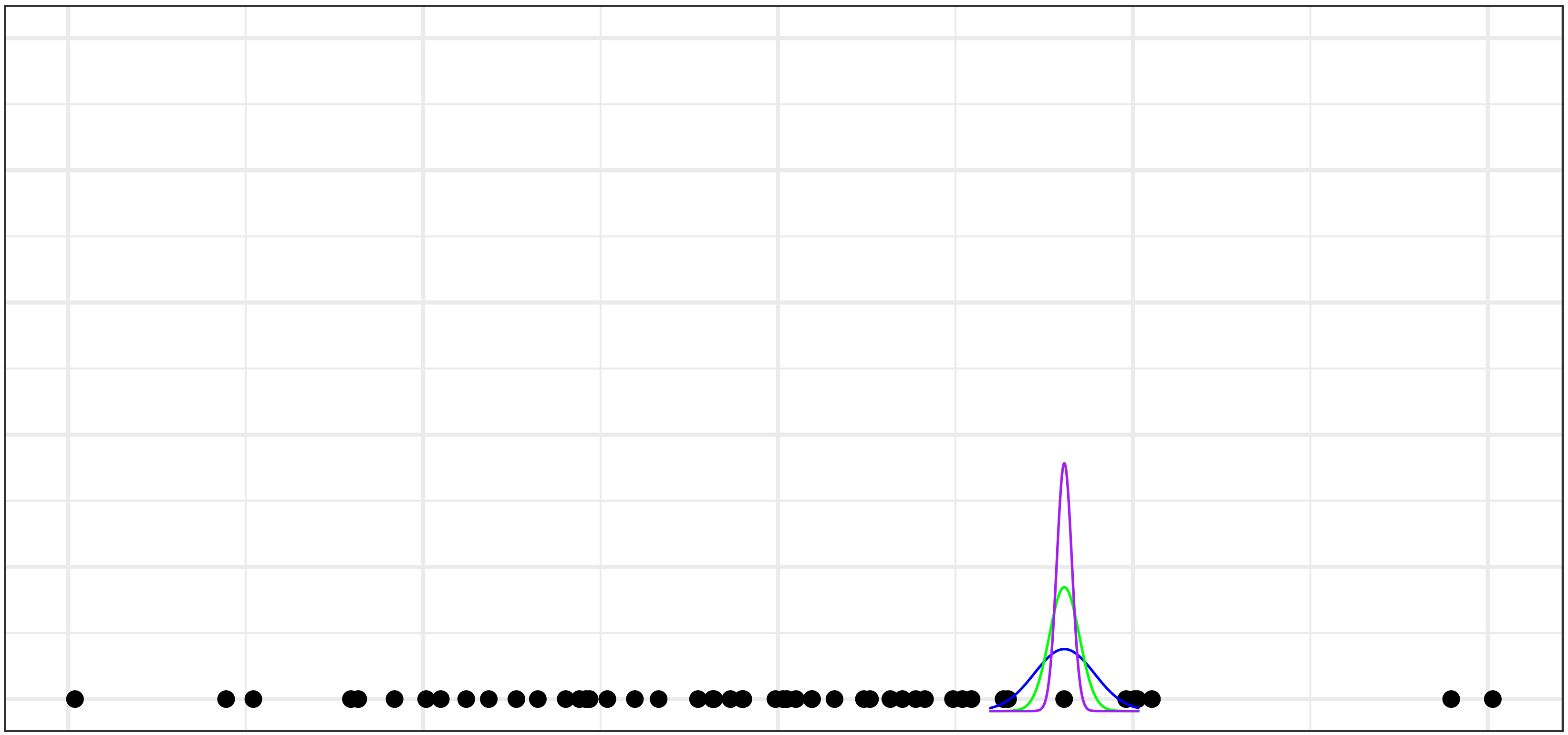
Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma, \mu, \phi)$$



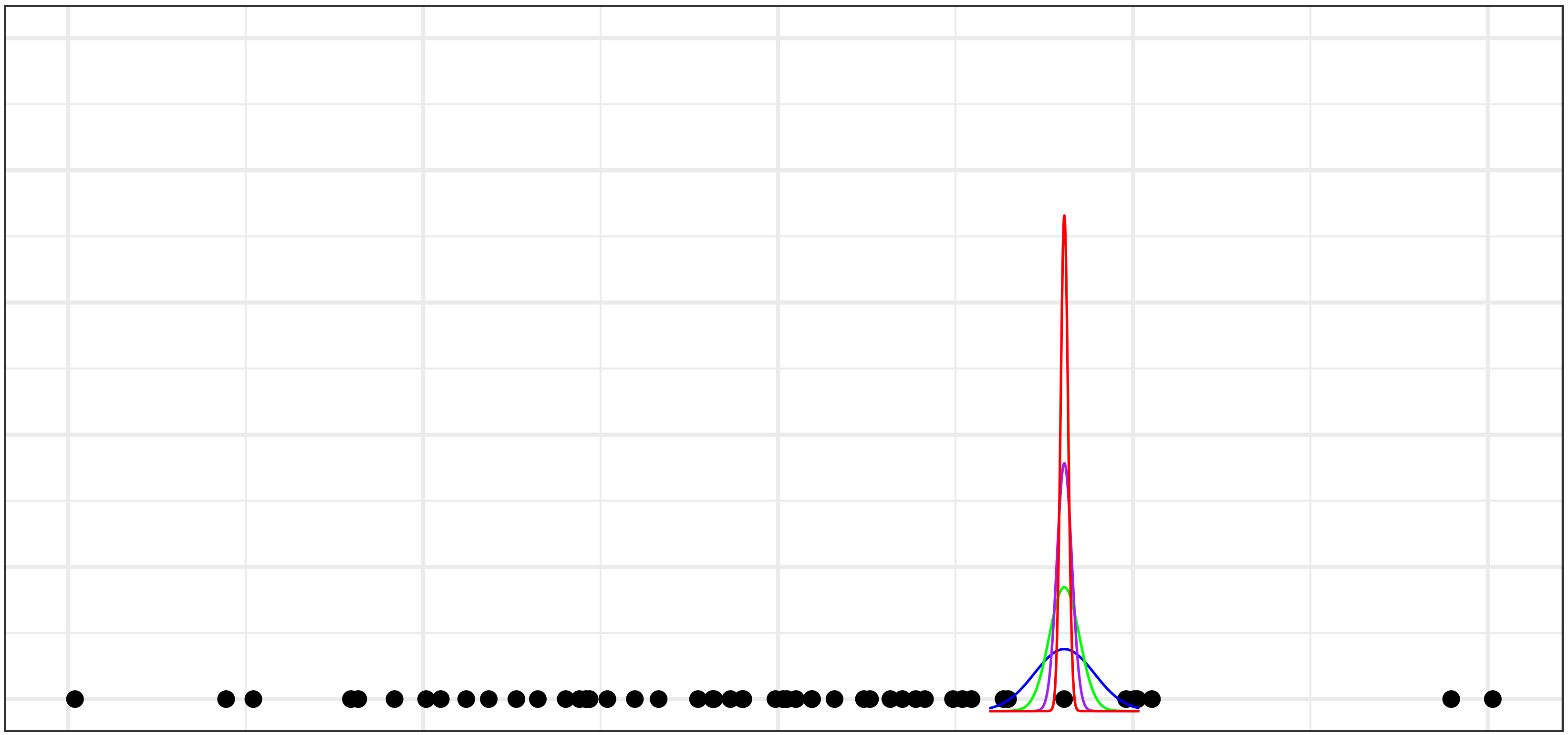
Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \mathop{\mathrm{arg\,max}}_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma, \mu, \phi)$$



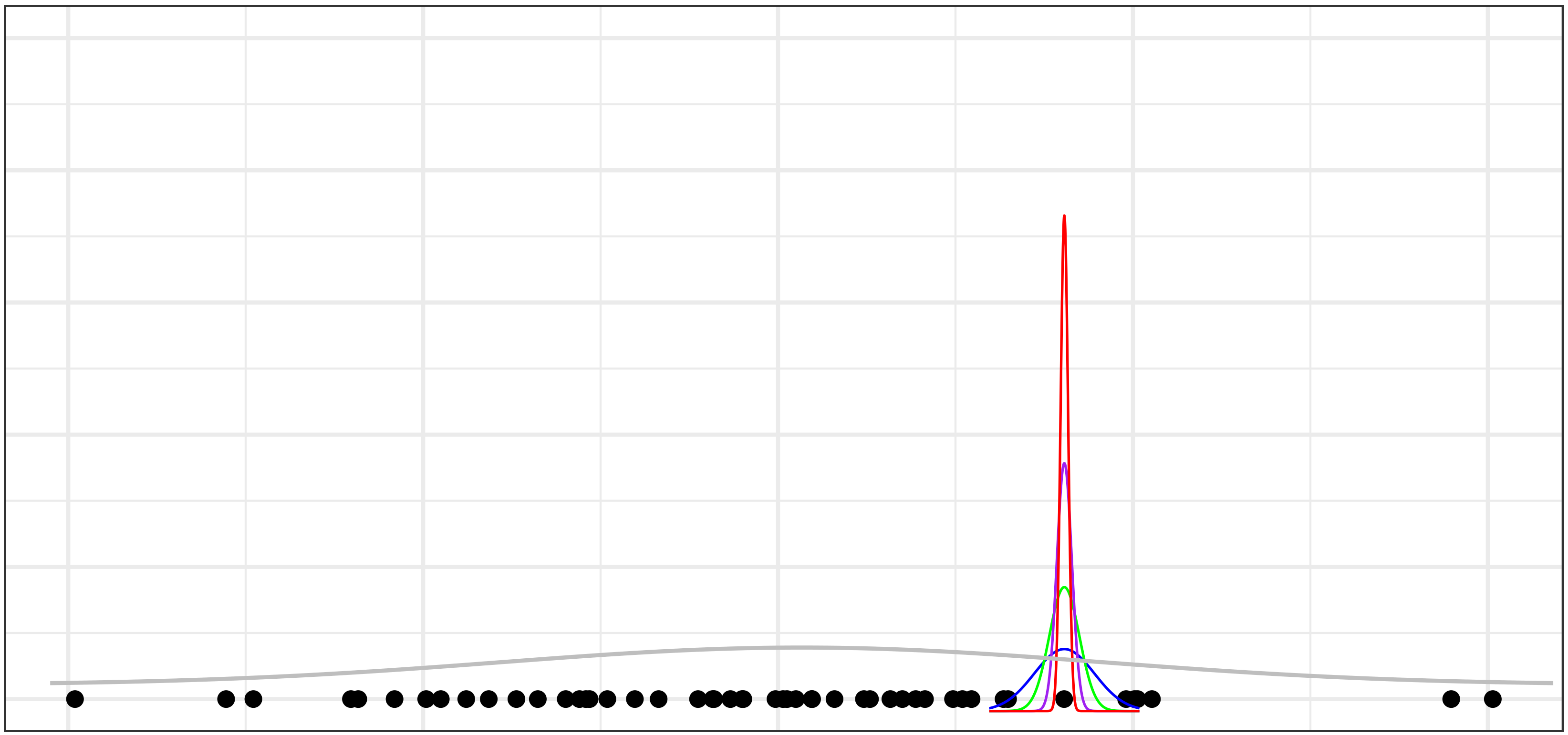
Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma, \mu, \phi)$$



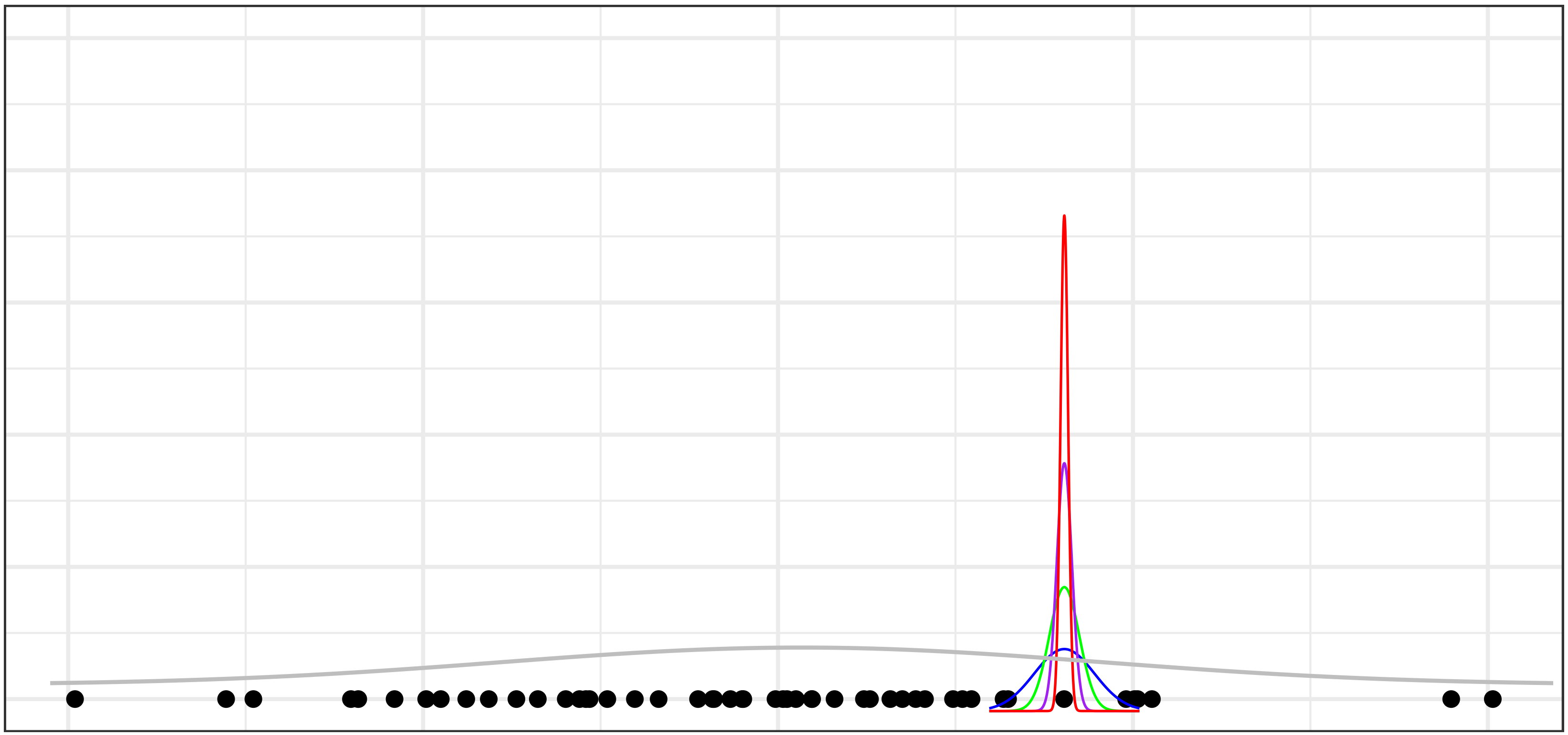
Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma,\mu,\phi)$$



Major problem for maximum likelihood

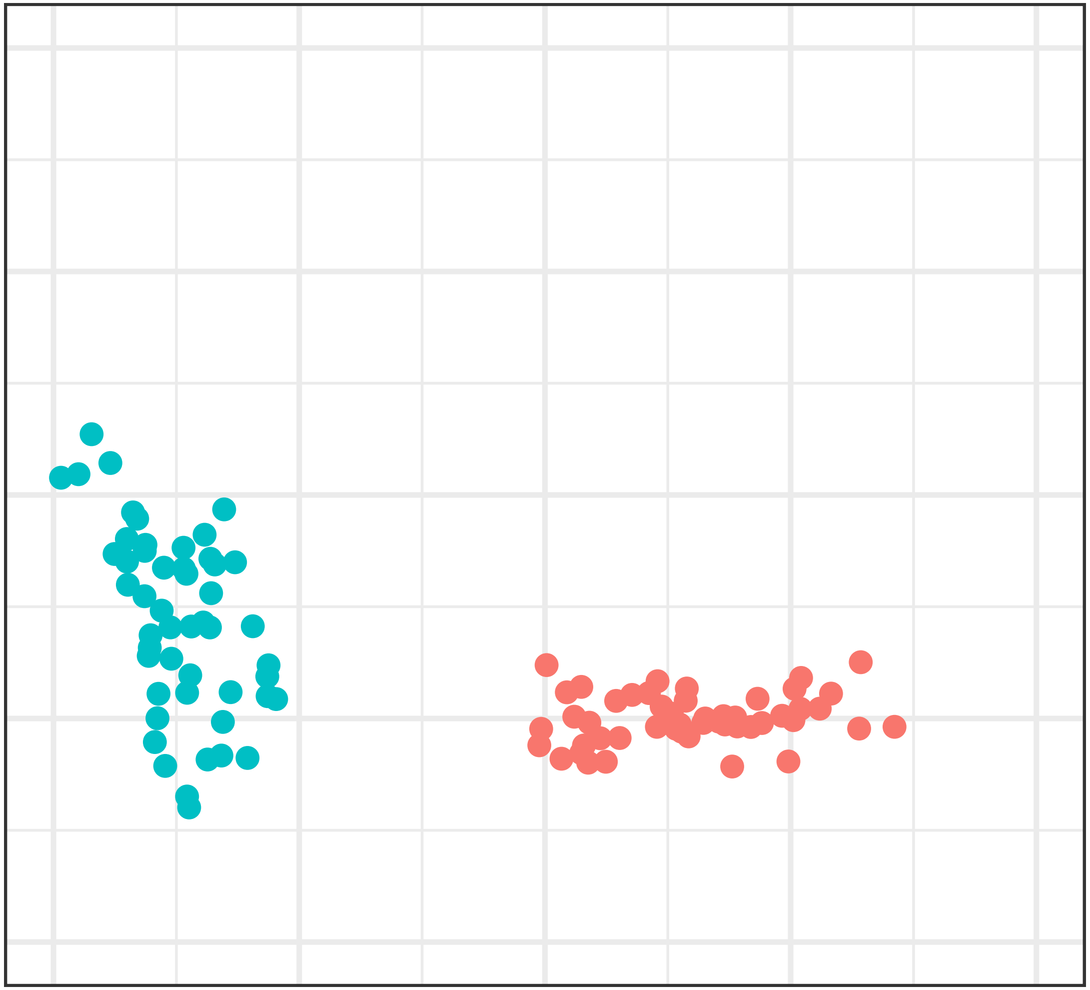
$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma, \mu, \phi)$$



An arbitrarily narrow Gaussian can assign arbitrarily high likelihood!

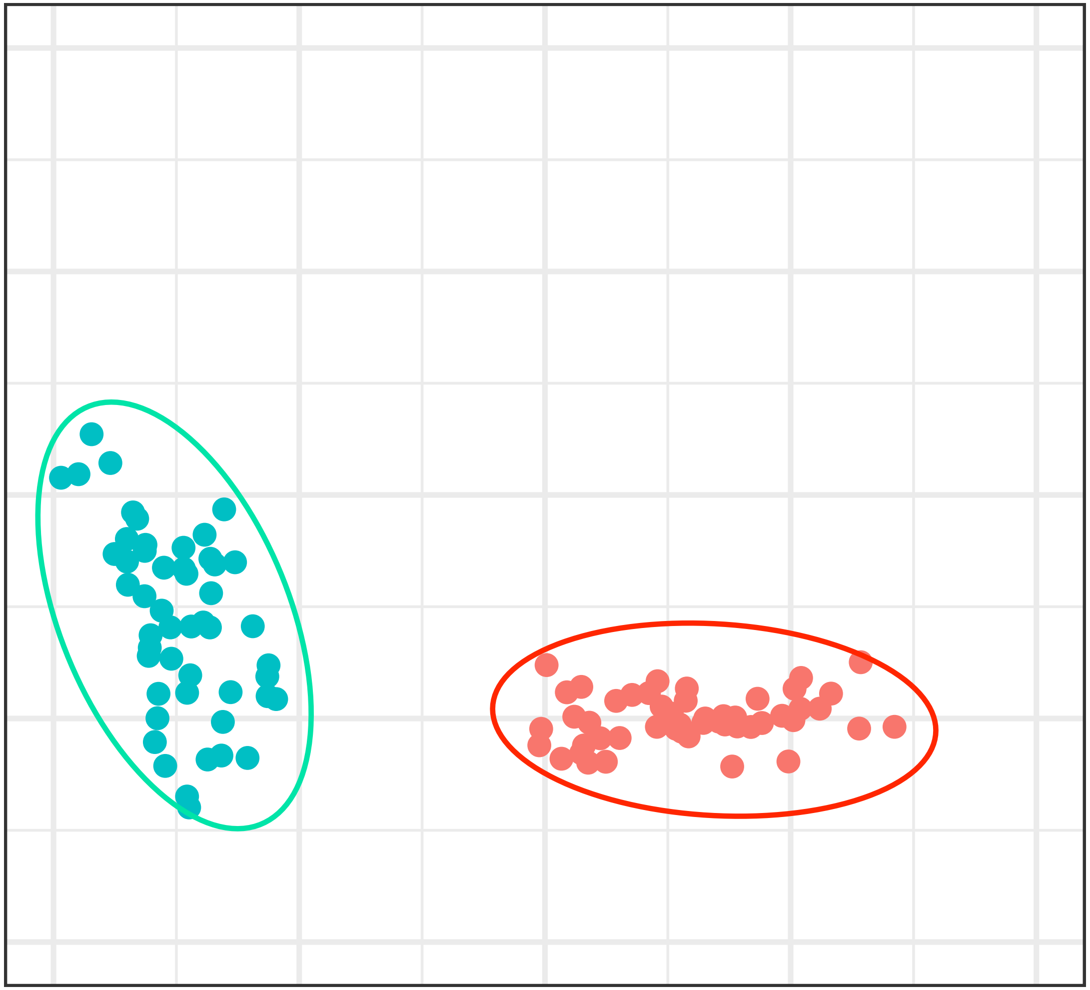
Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma, \mu, \phi)$$



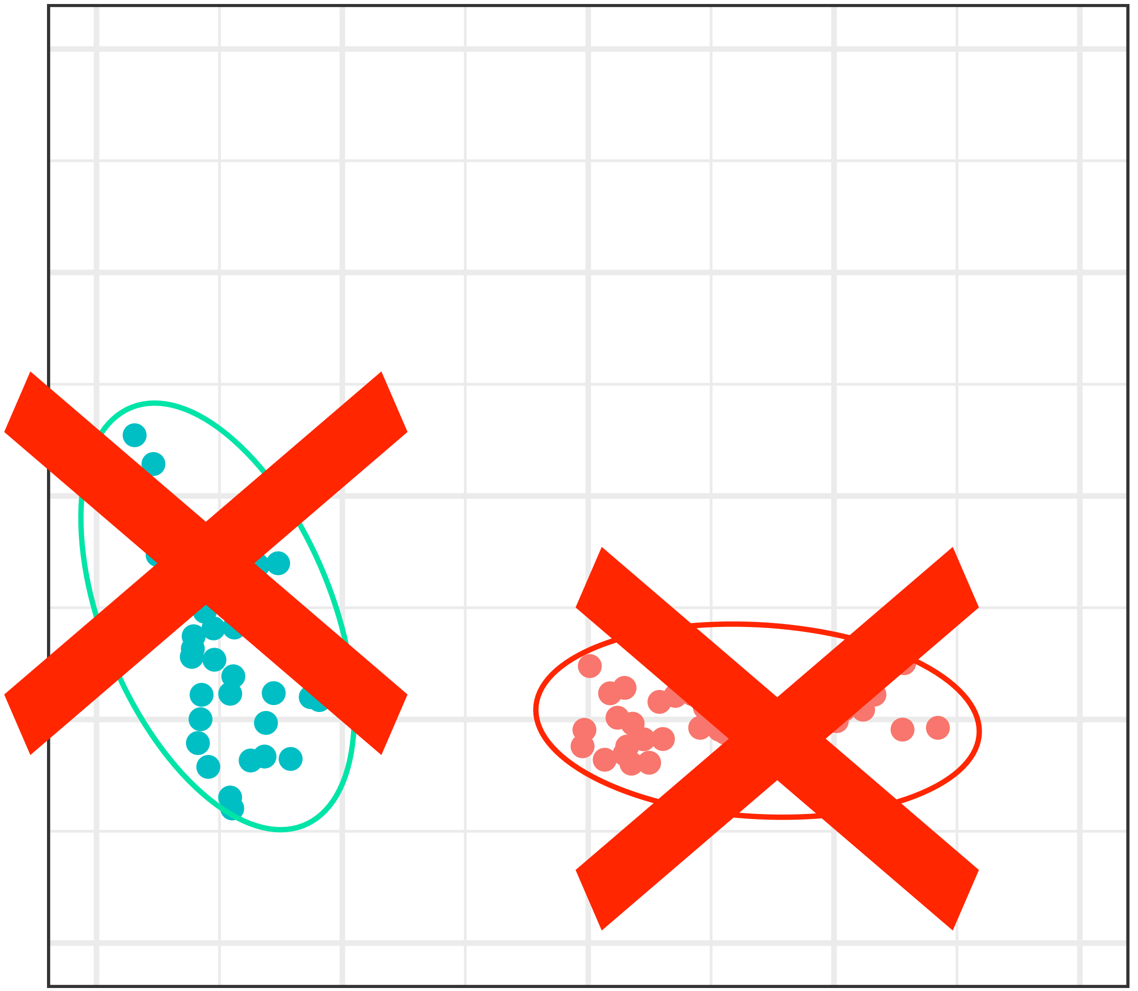
Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma,\mu,\phi)$$



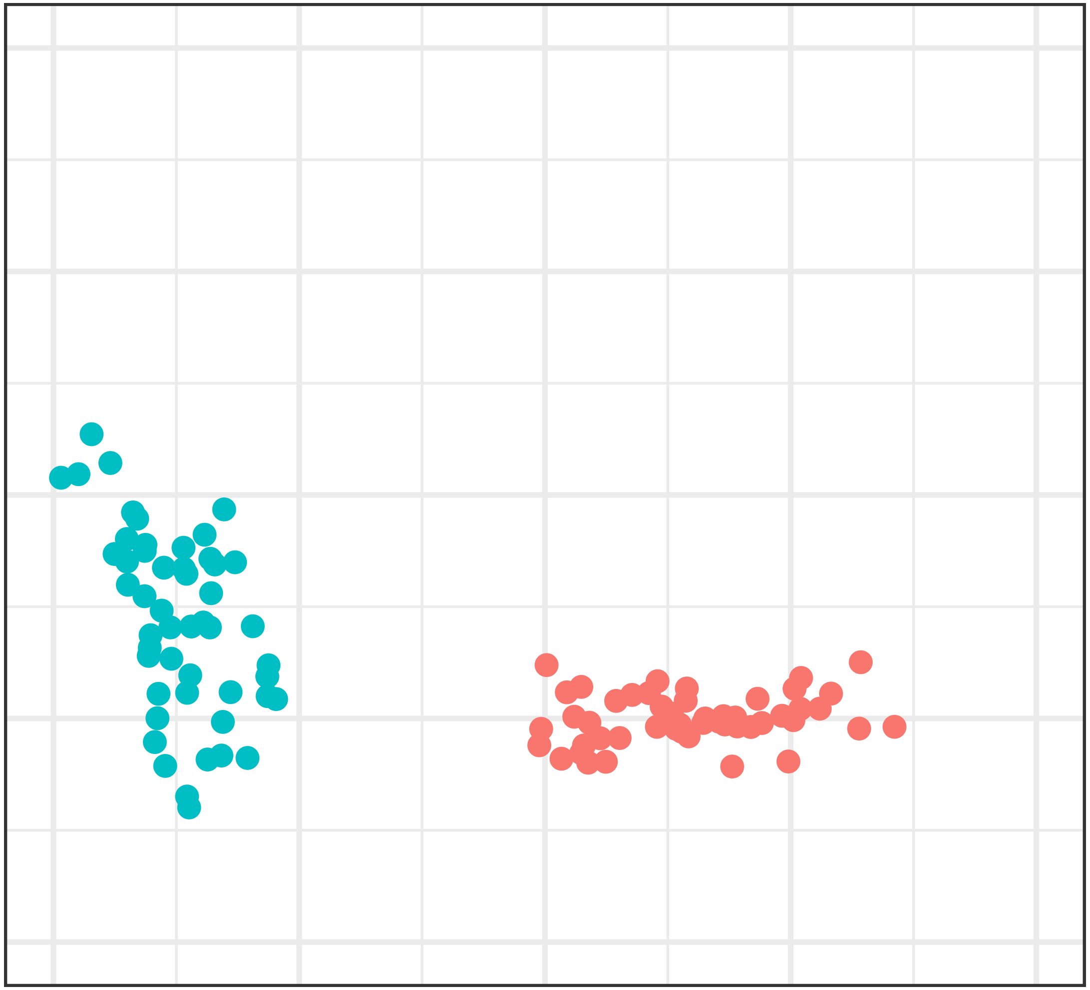
Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma, \mu, \phi)$$



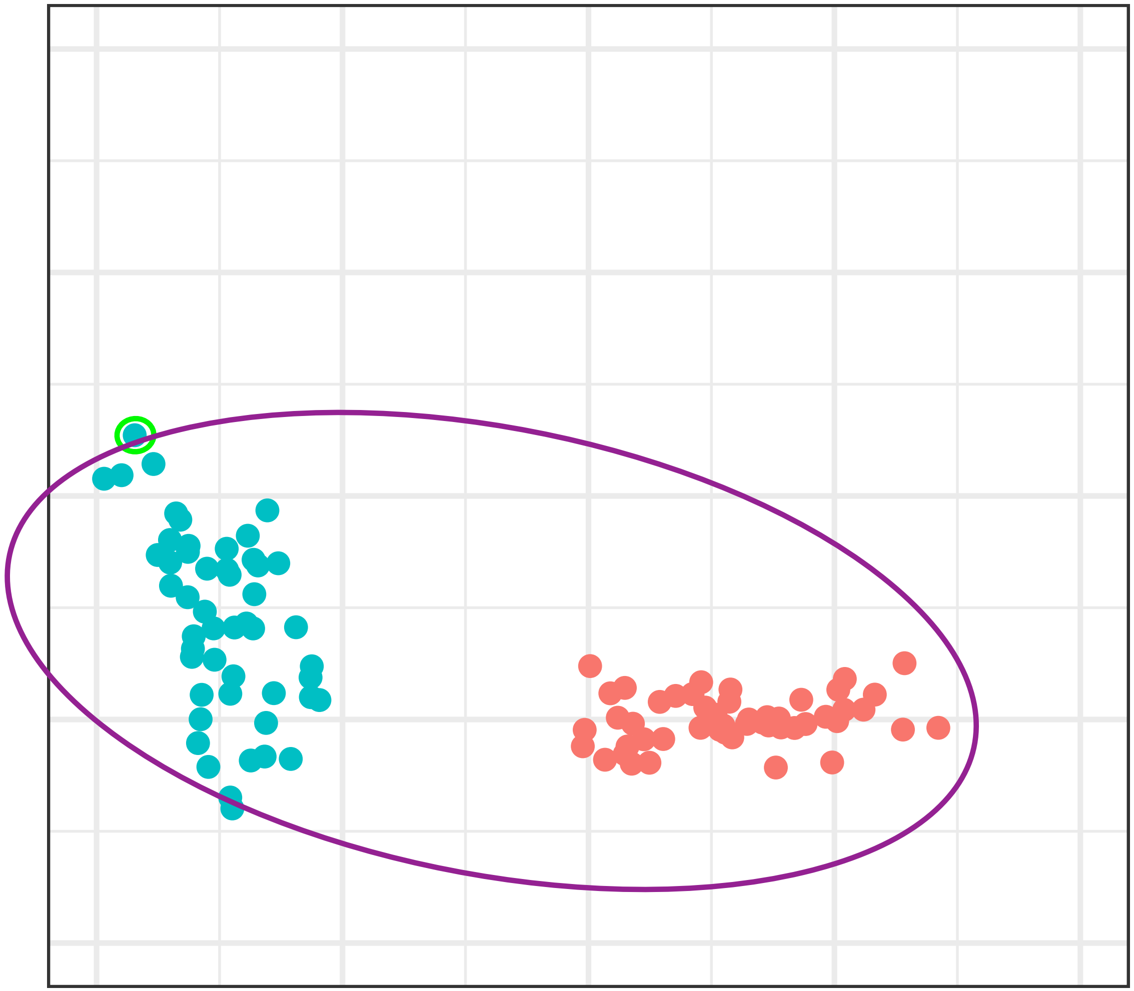
Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y} | \Sigma, \mu, \phi)$$



Major problem for maximum likelihood

$$\langle \widehat{\Sigma}, \widehat{\mu}, \widehat{\phi} \rangle = \arg\max_{\Sigma,\mu,\phi} P(\mathbf{y}|\Sigma, \mu, \phi)$$

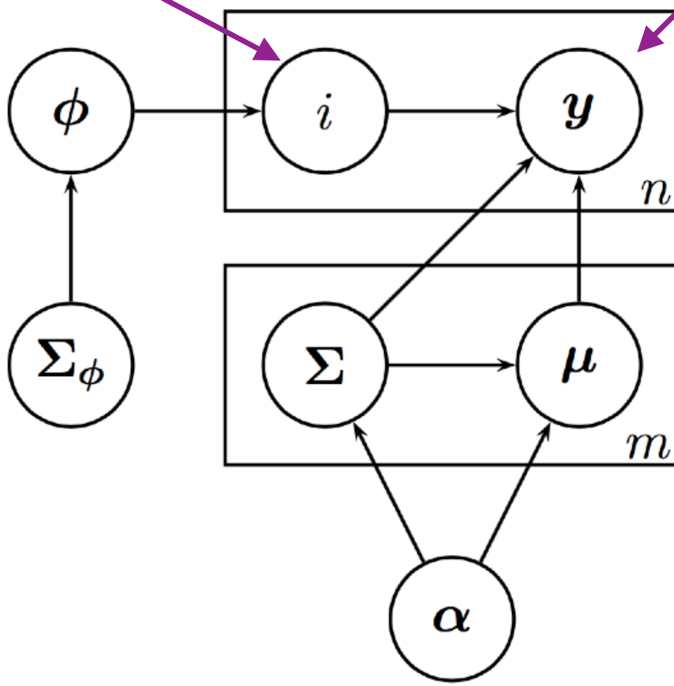


Bayesian mixture of Gaussians

category (unobserved)

datapoint (observed)

Multinomial over m categories



 # datapoints

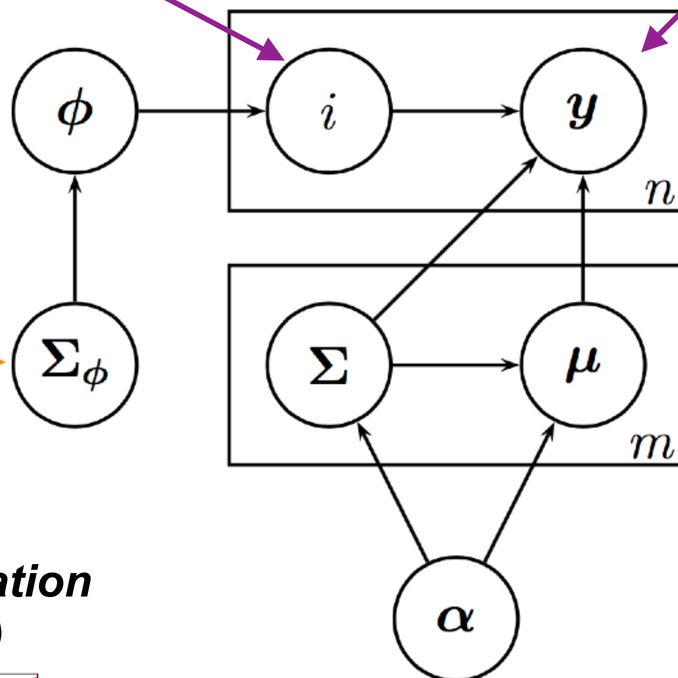
m Gaussian categories, each with its own mean & covariance matrix

Bayesian mixture of Gaussians

category (unobserved)

datapoint (observed)

Multinomial over m categories

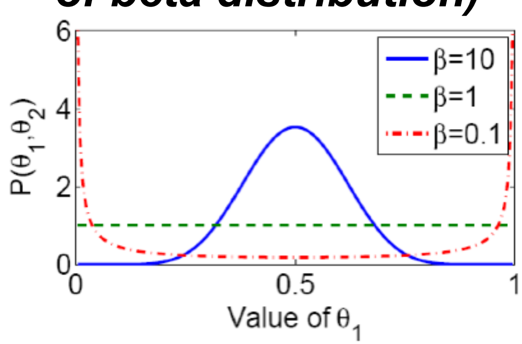


 # datapoints

Prior over category probabilities

m Gaussian categories, each with its own mean & covariance matrix

Dirichlet (generalization of beta distribution)

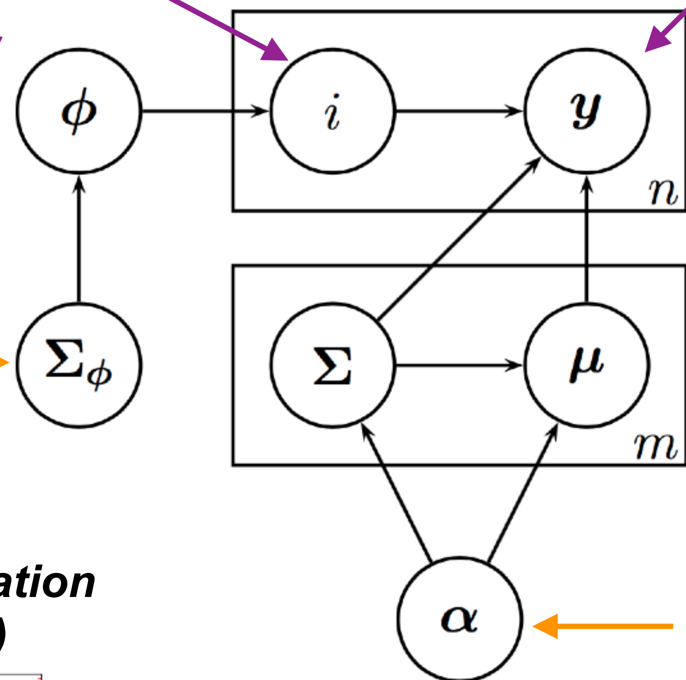


Bayesian mixture of Gaussians

category (unobserved)

datapoint (observed)

Multinomial over m categories

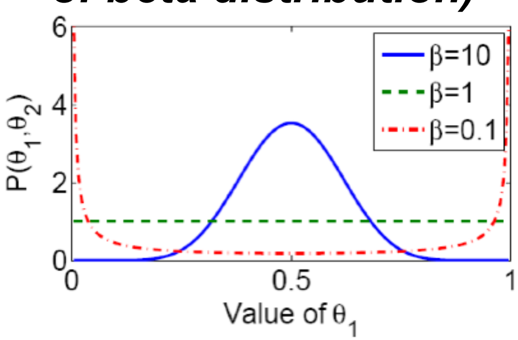


 # datapoints

Prior over category probabilities

m Gaussian categories, each with its own mean & covariance matrix

Dirichlet (generalization of beta distribution)



Prior over Gaussian category size and shapes: DON'T EXPECT TINY CATEGORIES!

Conjugate priors

Conjugate priors

- A CONJUGATE PRIOR for param(s) θ of a model is a family F of probability distributions $P(\theta)$ such that conditioning on data D keeps the posterior $P(\theta | D)$ within F

Conjugate priors

- A CONJUGATE PRIOR for param(s) θ of a model is a family F of probability distributions $P(\theta)$ such that conditioning on data D keeps the posterior $P(\theta | D)$ within F
- **Example:** the BETA DISTRIBUTION is conjugate to binomial:

(Image credit Micah Sittig; CC BY)

Binomial $P(r \text{ heads} \,|\, n \text{ flips}, p) = \binom{n}{r} p^r (1-p)^{n-r}$

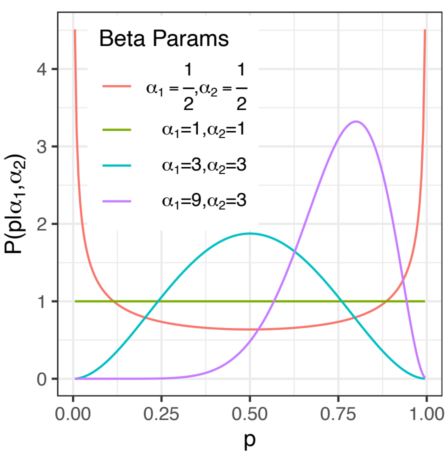
Conjugate priors

- A CONJUGATE PRIOR for param(s) θ of a model is a family F of probability distributions $P(\theta)$ such that conditioning on data D keeps the posterior $P(\theta \mid D)$ within F
- **Example:** the BETA DISTRIBUTION is conjugate to binomial:

(Image credit Micah Sittig; CC BY)

Binomial $P(r \text{ heads} \mid n \text{ flips}, p) = \binom{n}{r} p^r (1-p)^{n-r}$

Beta $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1} \quad [\alpha_1, \alpha_2 > 0]$



Conjugate priors

- A CONJUGATE PRIOR for param(s) θ of a model is a family F of probability distributions $P(\theta)$ such that conditioning on data D keeps the posterior $P(\theta \mid D)$ within F
- **Example:** the BETA DISTRIBUTION is conjugate to binomial:

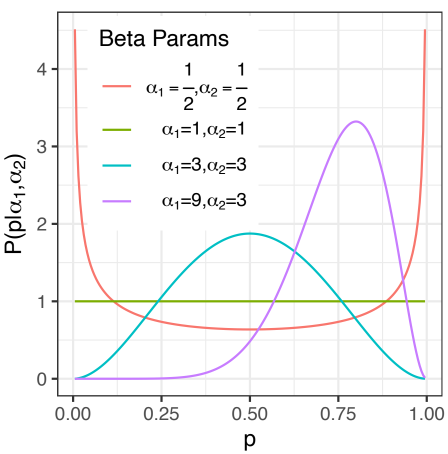
(Image credit Micah Sittig; CC BY)

Binomial $P(r \text{ heads} \mid n \text{ flips}, p) = \binom{n}{r} p^r (1-p)^{n-r}$

Beta $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1} \quad [\alpha_1, \alpha_2 > 0]$

$$P(p \mid \alpha_1, \alpha_2, D) \propto P(D \mid p, \alpha_1, \alpha_2) P(p \mid \alpha_1, \alpha_2)$$

n flips, r heads



Conjugate priors

- A CONJUGATE PRIOR for param(s) θ of a model is a family F of probability distributions $P(\theta)$ such that conditioning on data D keeps the posterior $P(\theta \mid D)$ within F
- **Example:** the BETA DISTRIBUTION is conjugate to binomial:

(Image credit Micah Sittig; CC BY)

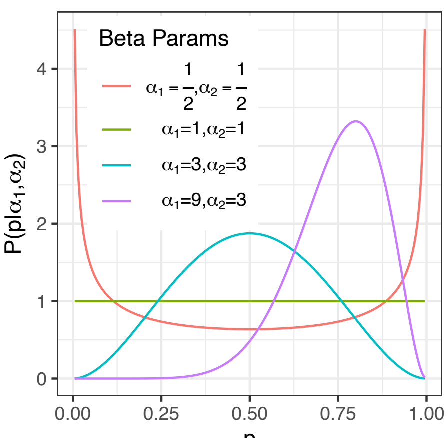
Binomial $P(r \text{ heads} \mid n \text{ flips}, p) = \binom{n}{r} p^r (1-p)^{n-r}$

Beta $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1} \quad [\alpha_1, \alpha_2 > 0]$

$$P(p \mid \alpha_1, \alpha_2, D) \propto P(D \mid p, \alpha_1, \alpha_2) P(p \mid \alpha_1, \alpha_2)$$

n flips, r heads

$$\propto \binom{n}{r} p^r (1-p)^{n-r} p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$$



Conjugate priors

- A CONJUGATE PRIOR for param(s) θ of a model is a family F of probability distributions $P(\theta)$ such that conditioning on data D keeps the posterior $P(\theta | D)$ within F
- **Example:** the BETA DISTRIBUTION is conjugate to binomial:

(Image credit Micah Sittig; CC BY)

Binomial $P(r \text{ heads} \,|\, n \text{ flips}, p) = \binom{n}{r} p^r (1-p)^{n-r}$

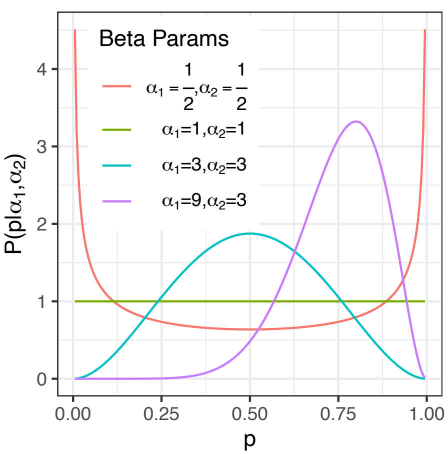
Beta $P(p \,|\, \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1} \quad [\alpha_1, \alpha_2 > 0]$

$$P(p \,|\, \alpha_1, \alpha_2, D) \propto P(D \,|\, p, \alpha_1, \alpha_2) P(p \,|\, \alpha_1, \alpha_2)$$

n flips, r heads

$$\propto \binom{n}{r} p^r (1-p)^{n-r} p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$$

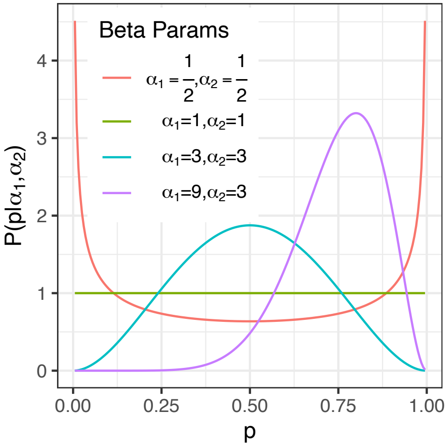
$$\propto \binom{n}{r} p^{\alpha_1 - 1 + r}(1-p)^{\alpha_2 - 1 + n - r}$$



Conjugate priors

- A CONJUGATE PRIOR for param(s) θ of a model is a family F of probability distributions $P(\theta)$ such that conditioning on data D keeps the posterior $P(\theta \mid D)$ within F
- **Example:** the BETA DISTRIBUTION is conjugate to binomial:

(Image credit Micah Sittig; CC BY)



Binomial $P(r \text{ heads} \mid n \text{ flips}, p) = \binom{n}{r} p^r (1-p)^{n-r}$

Beta $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1} \quad [\alpha_1, \alpha_2 > 0]$

$$P(p \mid \alpha_1, \alpha_2, D) \propto P(D \mid p, \alpha_1, \alpha_2) P(p \mid \alpha_1, \alpha_2)$$

n flips, r heads

$$\propto \binom{n}{r} p^r (1-p)^{n-r} p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$$

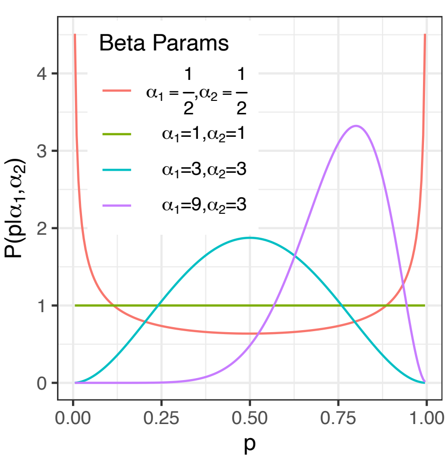
$$\propto \binom{n}{r} p^{\alpha_1 - 1 + r}(1-p)^{\alpha_2 - 1 + n - r}$$

$$\propto p^{\alpha_1 + r - 1}(1-p)^{\alpha_2 + n - r - 1}$$

Conjugate priors

- A CONJUGATE PRIOR for param(s) θ of a model is a family F of probability distributions $P(\theta)$ such that conditioning on data D keeps the posterior $P(\theta | D)$ within F
- **Example:** the BETA DISTRIBUTION is conjugate to binomial:

(Image credit Micah Sittig; CC BY)



Binomial $P(r \text{ heads} \mid n \text{ flips}, p) = \binom{n}{r} p^r (1-p)^{n-r}$

Beta $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1} \quad [\alpha_1, \alpha_2 > 0]$

$$P(p \mid \alpha_1, \alpha_2, D) \propto P(D \mid p, \alpha_1, \alpha_2) P(p \mid \alpha_1, \alpha_2)$$

n flips, *r* heads

$$\propto \binom{n}{r} p^r (1-p)^{n-r} p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$$

$$\propto \binom{n}{r} p^{\alpha_1 - 1 + r}(1-p)^{\alpha_2 - 1 + n - r}$$

Conjugate!

$$\propto p^{\alpha_1 + r - 1}(1-p)^{\alpha_2 + n - r - 1}$$

Conjugate priors in action

(Image credit Micah Sittig; CC BY)

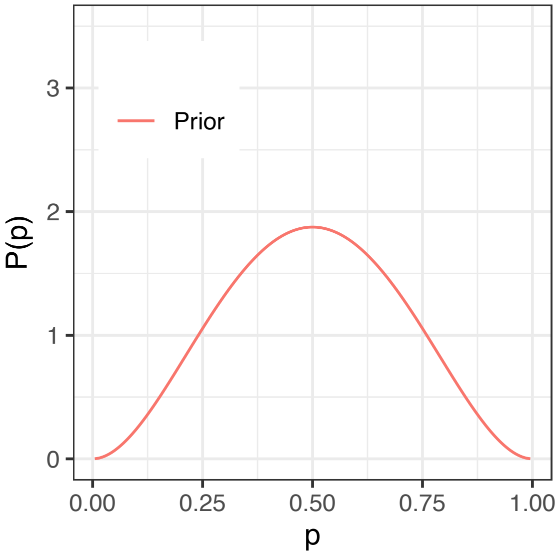
Prior: $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$ $\quad [\alpha_1, \alpha_2 > 0]$

Conjugate priors in action

(Image credit Micah Sittig; CC BY)

Prior: $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$ $\quad [\alpha_1, \alpha_2 > 0]$

Mild belief that the coin is fair: $\alpha_1 = 3, \alpha_2 = 3$



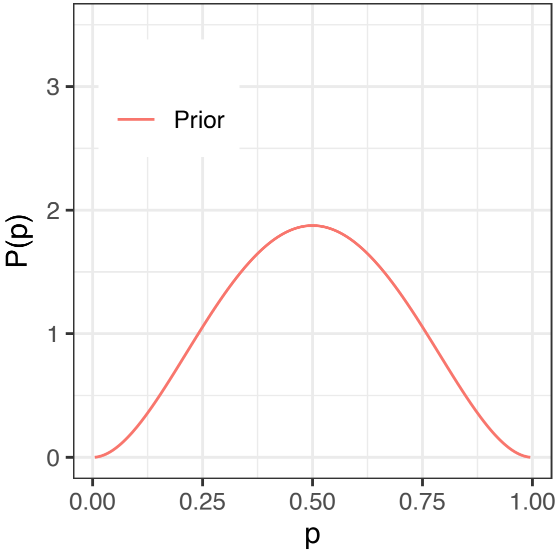
Conjugate priors in action

(Image credit Micah Sittig; CC BY)

Prior: $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$ $\quad [\alpha_1, \alpha_2 > 0]$

Mild belief that the coin is fair: $\alpha_1 = 3, \alpha_2 = 3$

Data: **`HTHHHTHHHT`**



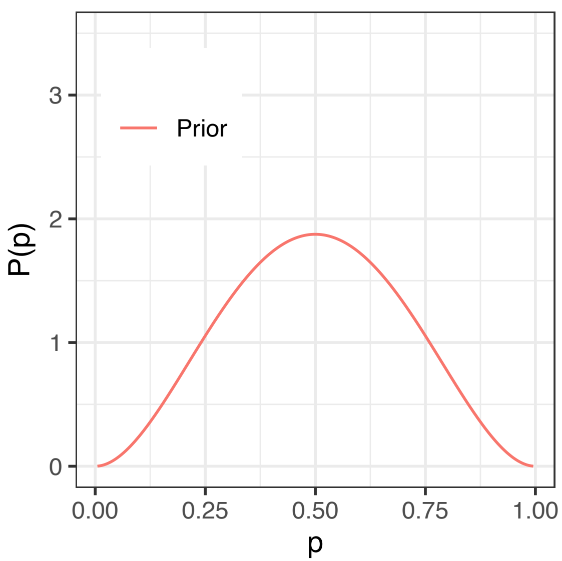
Conjugate priors in action

(Image credit Micah Sittig; CC BY)

Prior: $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$ $[\alpha_1, \alpha_2 > 0]$

Mild belief that the coin is fair: $\alpha_1 = 3, \alpha_2 = 3$

Data: **`HTHHHTHHHT`** $n = 10$ flips, $r = 7$ heads



Conjugate priors in action

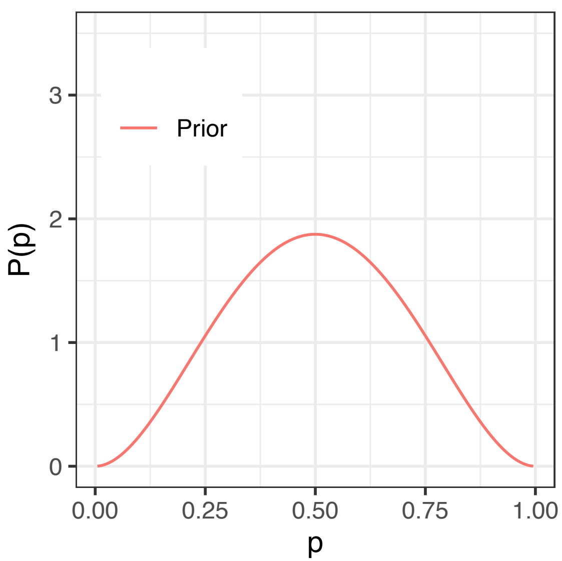
(Image credit Micah Sittig; CC BY)

Prior: $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$ $[\alpha_1, \alpha_2 > 0]$

Mild belief that the coin is fair: $\alpha_1 = 3, \alpha_2 = 3$

Data: **`HTHHHTHHHT`** $n = 10$ flips, $r = 7$ heads

Posterior: $P(p \mid \alpha_1, \alpha_2, D) \propto p^{\overbrace{\alpha_1 + r}^{\alpha_1'} - 1}(1-p)^{\overbrace{\alpha_2 + n - r}^{\alpha_2'} - 1}$



Conjugate priors in action

(Image credit Micah Sittig; CC BY)

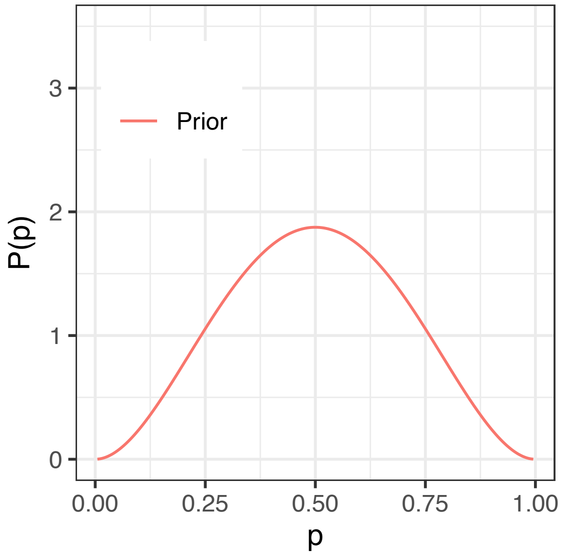
Prior: $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$ $\quad [\alpha_1, \alpha_2 > 0]$

Mild belief that the coin is fair: $\alpha_1 = 3, \alpha_2 = 3$

Data: **`HTHHHTHHHT`** $n = 10$ flips, $r = 7$ heads

Posterior: $P(p \mid \alpha_1, \alpha_2, D) \propto p^{\overbrace{\alpha_1 + r}^{\alpha_1'} - 1}(1-p)^{\overbrace{\alpha_2 + n - r}^{\alpha_2'} - 1}$

$$P(p \mid \alpha_1, \alpha_2, D) \propto p^{\alpha_1' - 1}(1-p)^{\alpha_2' - 1}$$



Conjugate priors in action

(Image credit Micah Sittig; CC BY)

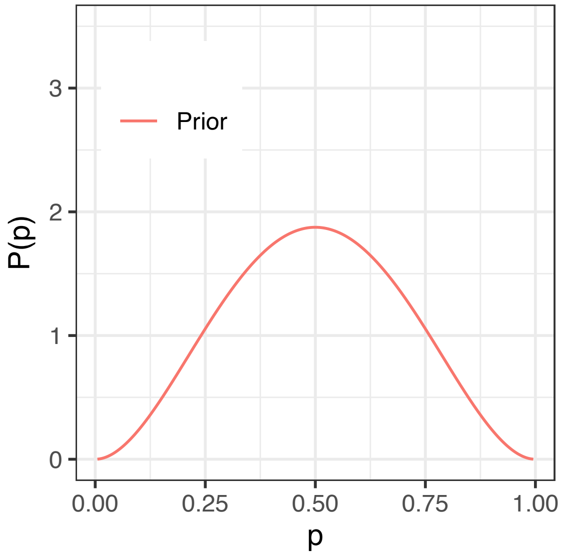
Prior: $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$ $\quad [\alpha_1, \alpha_2 > 0]$

Mild belief that the coin is fair: $\alpha_1 = 3, \alpha_2 = 3$

Data: **HTHHHTHHHT** $\quad n = 10$ flips, $r = 7$ heads

Posterior: $P(p \mid \alpha_1, \alpha_2, D) \propto p^{\overbrace{\alpha_1 + r}^{\alpha_1'} - 1}(1-p)^{\overbrace{\alpha_2 + n - r}^{\alpha_2'} - 1}$

$\alpha_1' = 10, \alpha_2' = 6 \qquad P(p \mid \alpha_1, \alpha_2, D) \propto p^{\alpha_1' - 1}(1-p)^{\alpha_2' - 1}$



Conjugate priors in action


(Image credit Micah Sittig; CC BY)

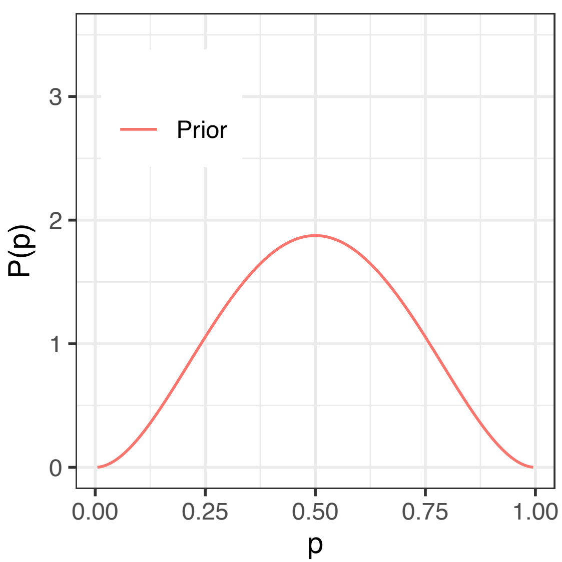
Prior: $P(p \mid \alpha_1, \alpha_2) \propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$ $[\alpha_1, \alpha_2 > 0]$

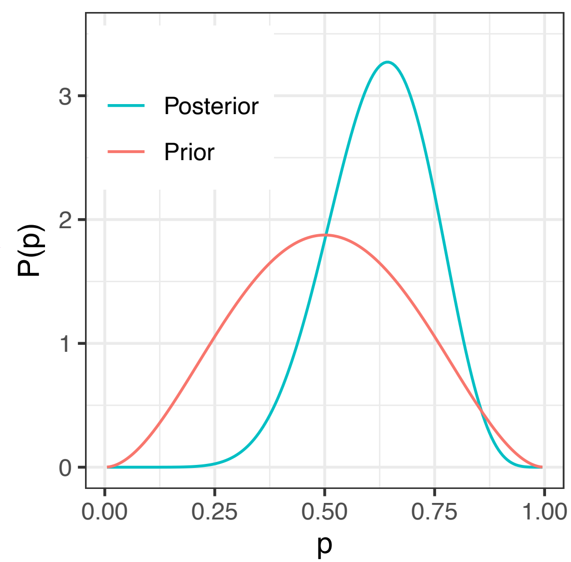
Mild belief that the coin is fair: $\alpha_1 = 3, \alpha_2 = 3$

Data: **`HTHHHTHHHT`** $n = 10$ flips, $r = 7$ heads

Posterior: $P(p \mid \alpha_1, \alpha_2, D) \propto p^{\overbrace{\alpha_1 + r}^{\alpha_1'} - 1}(1-p)^{\overbrace{\alpha_2 + n - r}^{\alpha_2'} - 1}$

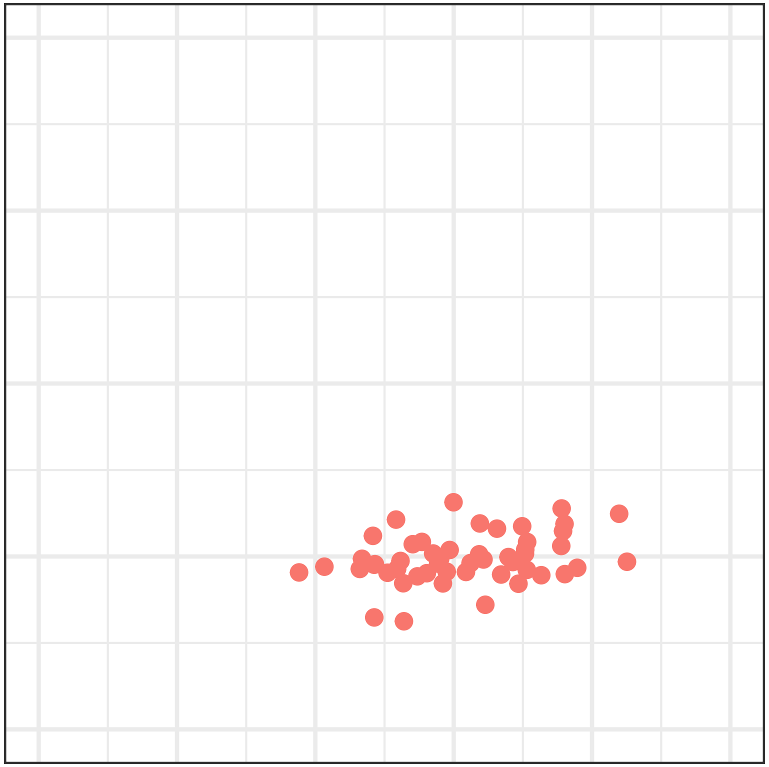
$\alpha_1' = 10, \alpha_2' = 6$ $P(p \mid \alpha_1, \alpha_2, D) \propto p^{\alpha_1' - 1}(1-p)^{\alpha_2' - 1}$





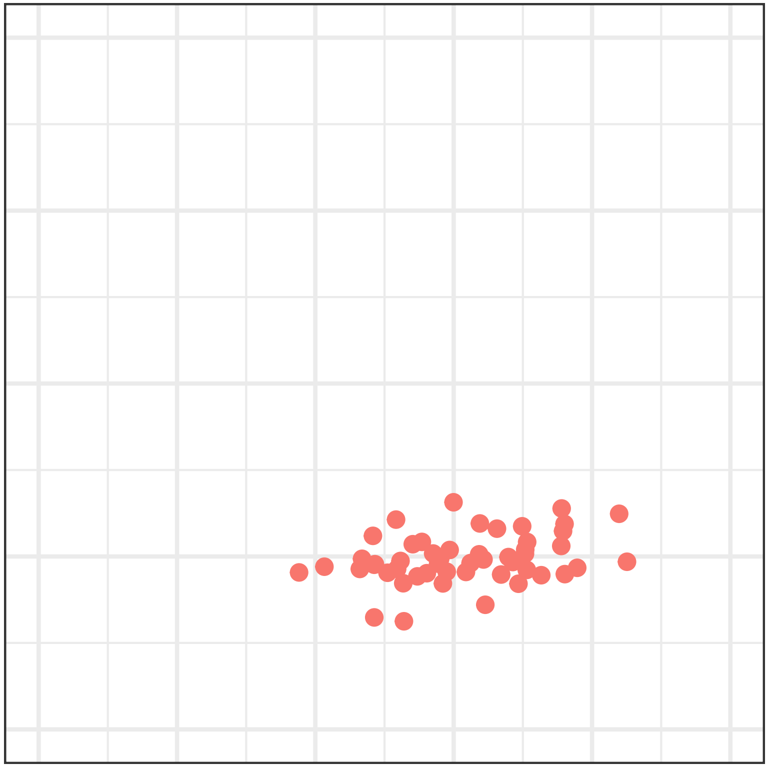
Conjugate prior for Gaussians

$$P(\mu, \Sigma \mid D) = ???$$



Conjugate prior for Gaussians

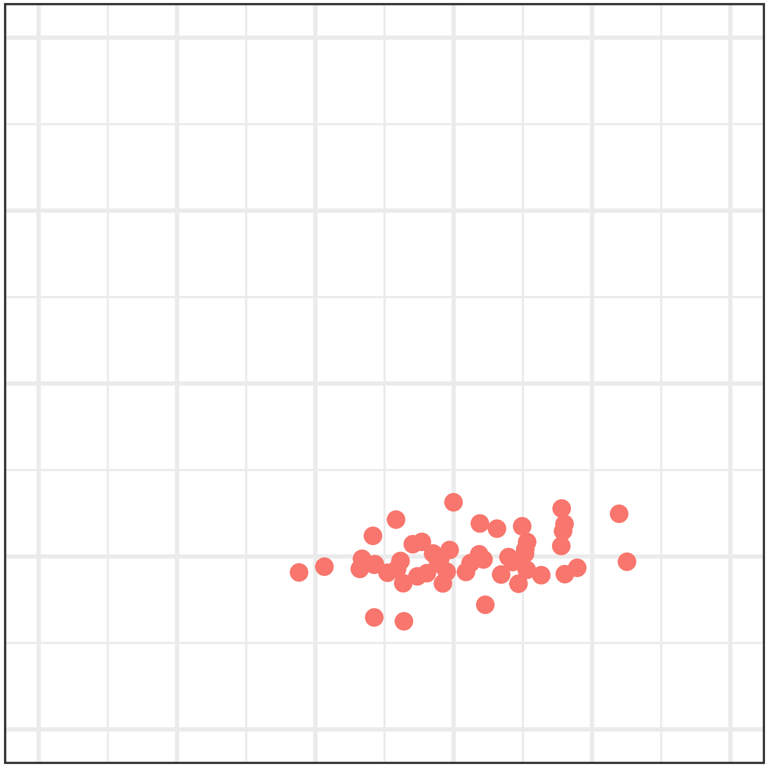
$$P(\mu, \Sigma \mid D) = ???$$



$$\mu \mid \Sigma \sim N(\mu_0, \Sigma)$$

Conjugate prior for Gaussians

$$P(\mu, \Sigma \mid D) = ???$$



$$\mu \mid \Sigma \sim N(\mu_0, \Sigma)$$

$$\Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

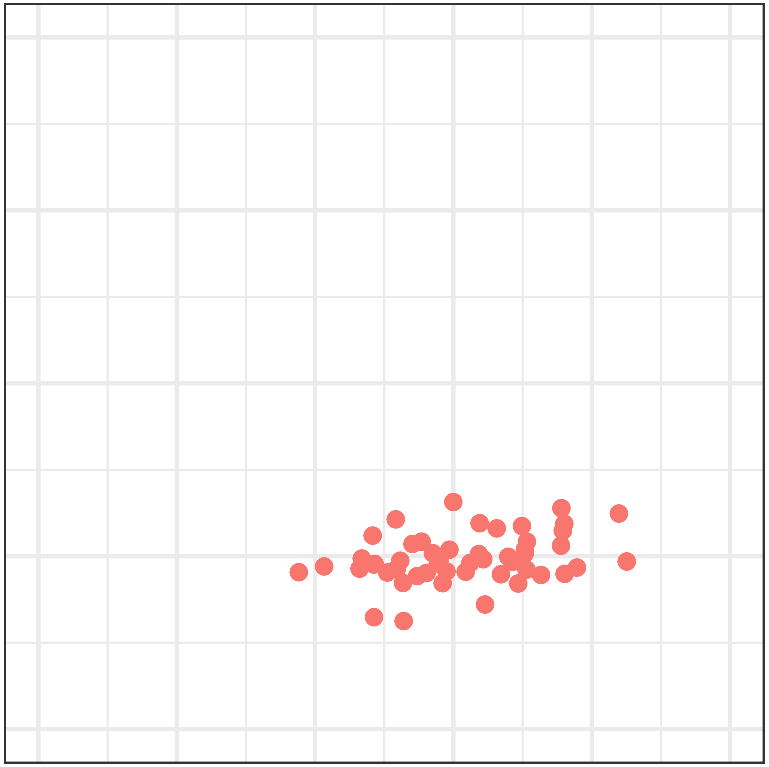
Conjugate prior for Gaussians

$$P(\mu, \Sigma \mid D) = ???$$

$$\mu \mid \Sigma \sim N(\mu_0, \Sigma)$$

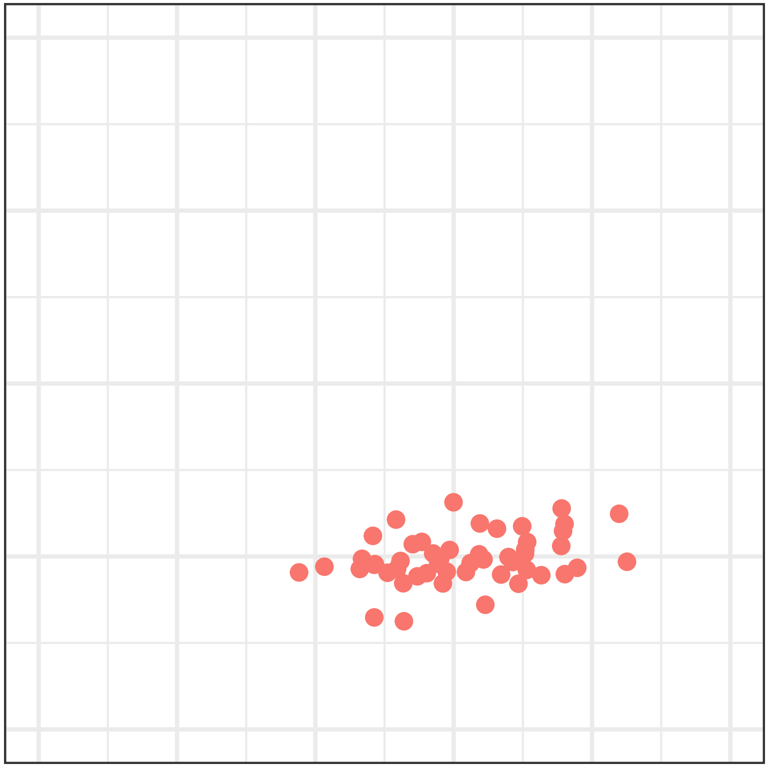
Positive, semi-definite
scale matrix

$$\Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$



Conjugate prior for Gaussians

$$P(\mu, \Sigma \,|\, D) = ???$$



Positive, semi-definite
scale matrix

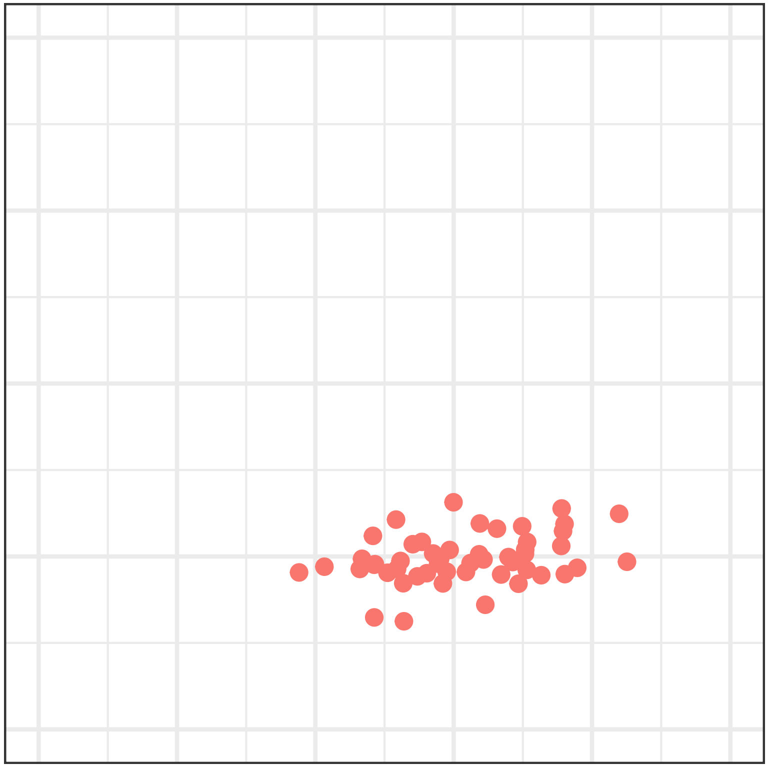
$$\mu \,|\, \Sigma \sim N(\mu_0, \Sigma)$$

$$\Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

Degrees of freedom (functions sort of like # pseudo-observations)

Conjugate prior for Gaussians

$$P(\mu, \Sigma \mid D) = ???$$



$$\mu \mid \Sigma \sim N(\mu_0, \Sigma)$$

Positive, semi-definite **scale matrix**

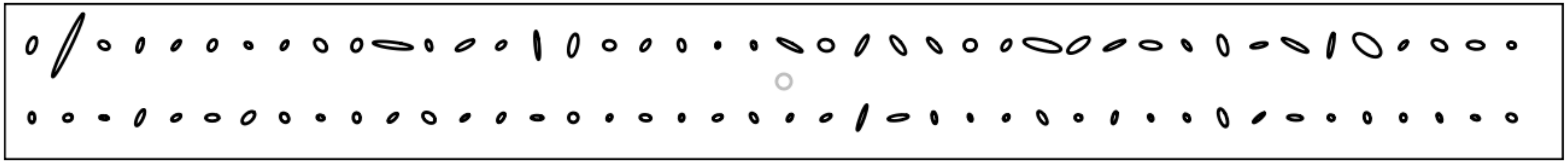
$$\Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

Degrees of freedom (functions sort of like # pseudo-observations)

Samples from Inverse-Wishart(I, ν):

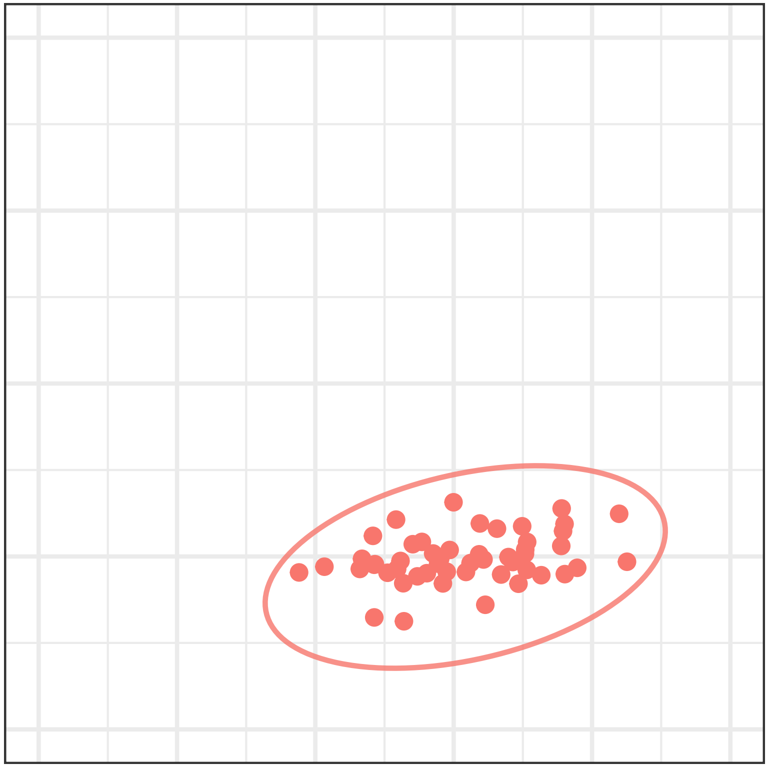
$\nu = 2$

$\nu = 5$



Conjugate prior for Gaussians

$$P(\mu, \Sigma \mid D) = ???$$



Positive, semi-definite **scale matrix**

$$\mu \mid \Sigma \sim N(\mu_0, \Sigma)$$

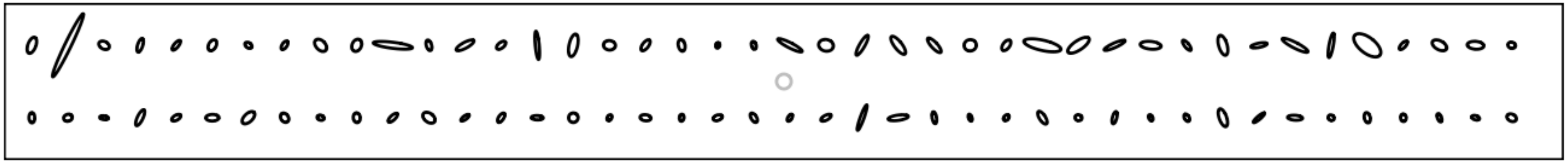
$$\Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

Degrees of freedom (functions sort of like # pseudo-observations)

Samples from Inverse-Wishart(I, ν):

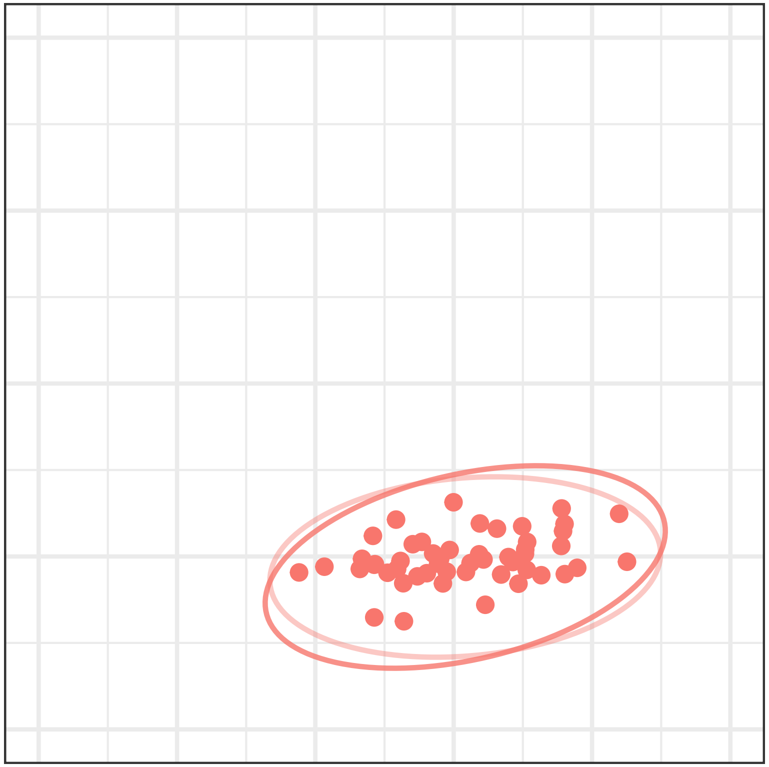
$\nu = 2$

$\nu = 5$



Conjugate prior for Gaussians

$$P(\mu, \Sigma \mid D) = ???$$



$$\mu \mid \Sigma \sim N(\mu_0, \Sigma)$$

Positive, semi-definite
scale matrix

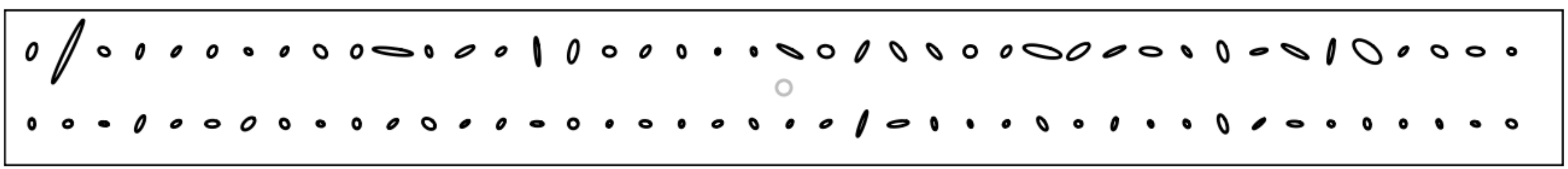
$$\Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

Degrees of freedom (functions sort of like # pseudo-observations)

Samples from Inverse-Wishart(I, ν):

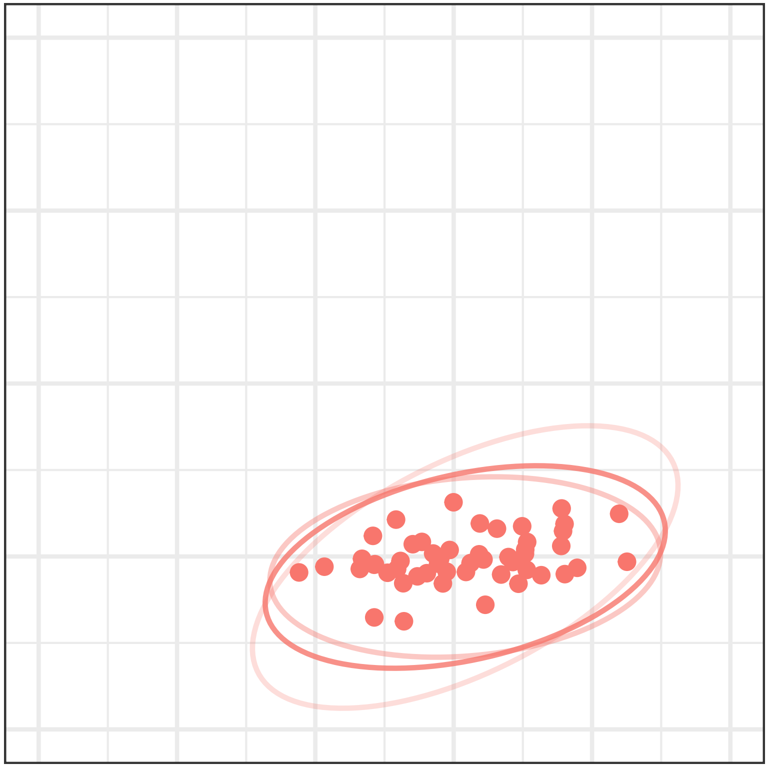
$\nu = 2$

$\nu = 5$



Conjugate prior for Gaussians

$$P(\mu, \Sigma \mid D) = ???$$



$$\mu \mid \Sigma \sim N(\mu_0, \Sigma)$$

Positive, semi-definite
scale matrix

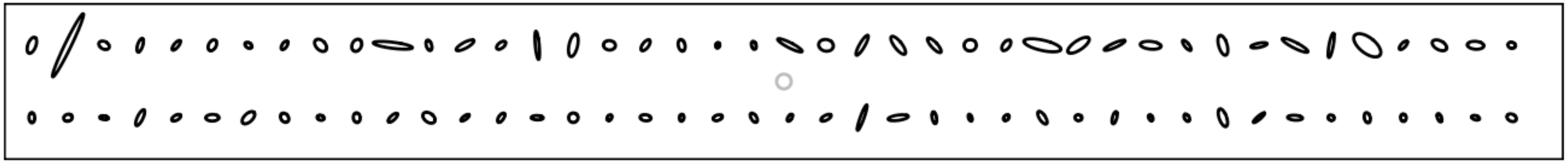
$$\Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

Degrees of freedom (functions sort of like # pseudo-observations)

Samples from Inverse-Wishart(I, ν):

$\nu = 2$

$\nu = 5$



Motivating Gibbs sampling

Motivating Gibbs sampling

- With conjugacy, we can get exact posteriors given data D

$$\mu \,|\, \Sigma \sim N(\mu_0, \Sigma) \qquad \Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

Motivating Gibbs sampling

- With conjugacy, we can get exact posteriors given data D

$$\mu \,|\, \Sigma \sim N(\mu_0, \Sigma) \qquad \Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

- We can also compute marginal data likelihoods

$P(D \,|\, \mu_0, S, \nu)$ [detailed formulae not shown]

Motivating Gibbs sampling

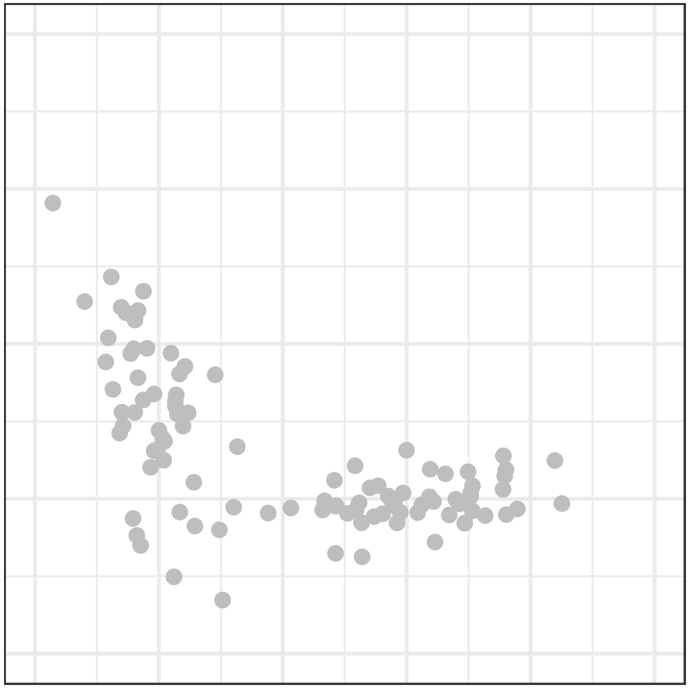
- With conjugacy, we can get exact posteriors given data D

$$\mu \,|\, \Sigma \sim N(\mu_0, \Sigma) \qquad \Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

- We can also compute marginal data likelihoods

$$P(D \,|\, \mu_0, S, \nu) \qquad \text{[detailed formulae not shown]}$$

- **Problem:** what if we don't know category identities?



Motivating Gibbs sampling

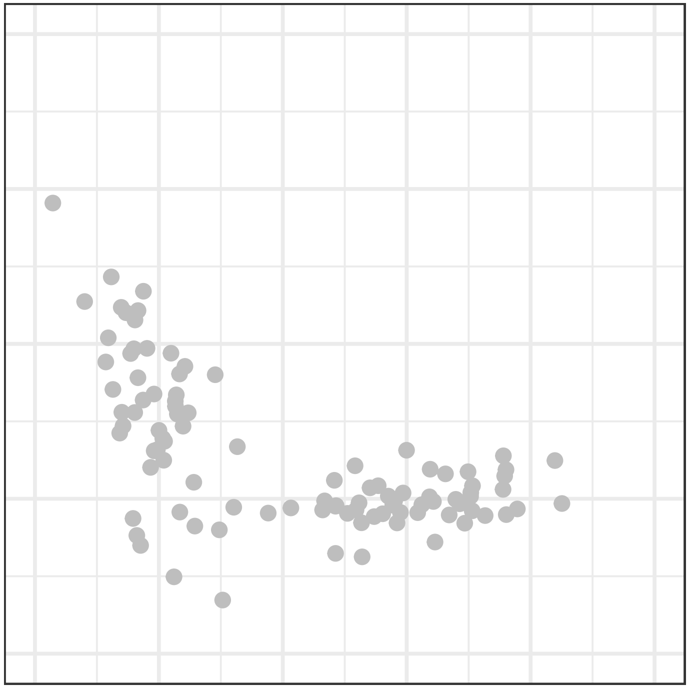
- With conjugacy, we can get exact posteriors given data D

$$\mu \mid \Sigma \sim N(\mu_0, \Sigma) \qquad \Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

- We can also compute marginal data likelihoods

$$P(D \mid \mu_0, S, \nu) \qquad \text{[detailed formulae not shown]}$$

- **Problem:** what if we don't know category identities?



...if we knew **category parameters** $\{\mu_i, \Sigma_i\}$, we could compute **category membership** probability for each observation...

Motivating Gibbs sampling

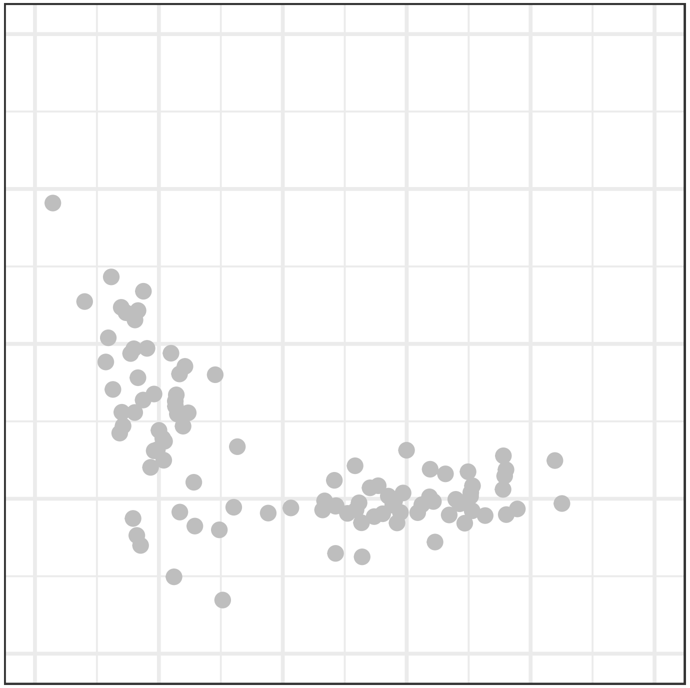
- With conjugacy, we can get exact posteriors given data D

$$\mu \mid \Sigma \sim N(\mu_0, \Sigma) \qquad \Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

- We can also compute marginal data likelihoods

$$P(D \mid \mu_0, S, \nu)$$ [detailed formulae not shown]

- **Problem:** what if we don't know category identities?



...if we knew **category parameters** $\{\mu_i, \Sigma_i\}$, we could compute **category membership** probability for each observation...

...if we knew **category memberships**, we could put posteriors on **category parameters** $\{\mu_i, \Sigma_i\}$...

Motivating Gibbs sampling

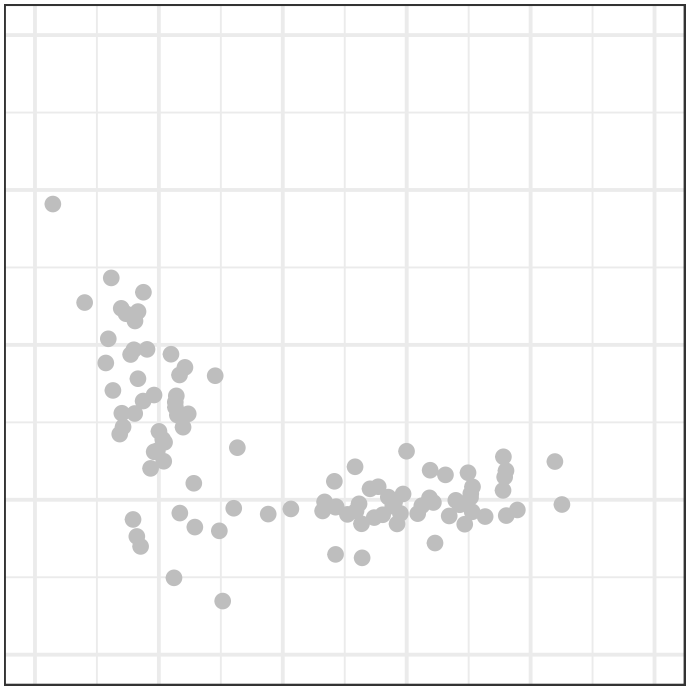
- With conjugacy, we can get exact posteriors given data D

$$\mu \,|\, \Sigma \sim N(\mu_0, \Sigma) \qquad \Sigma \sim \text{Inverse-Wishart}(S^{-1}, \nu)$$

- We can also compute marginal data likelihoods

$$P(D \,|\, \mu_0, S, \nu)$$ [detailed formulae not shown]

- **Problem:** what if we don't know category identities?

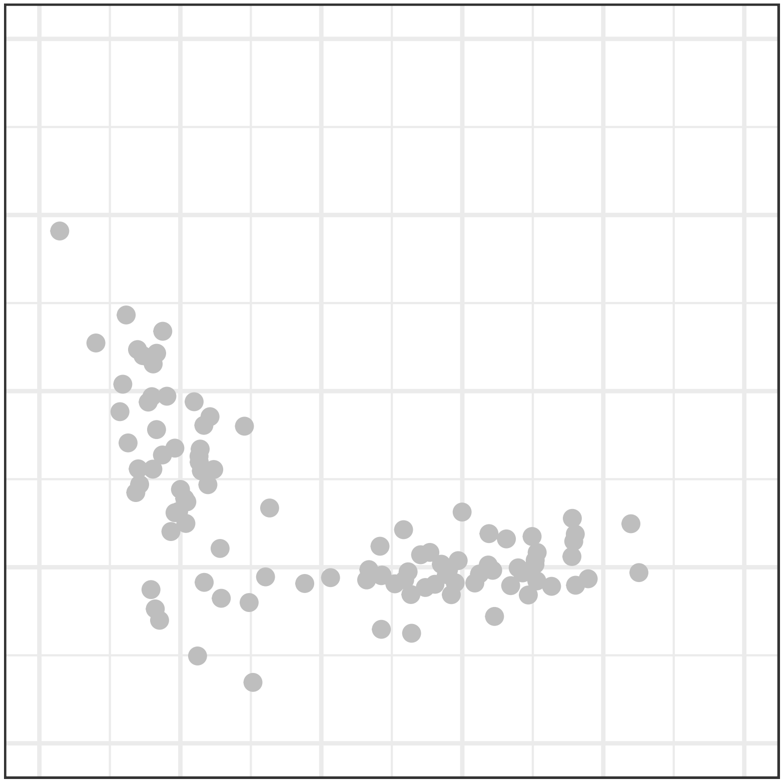


...if we knew **category parameters** $\{\mu_i, \Sigma_i\}$, we could compute **category membership** probability for each observation...

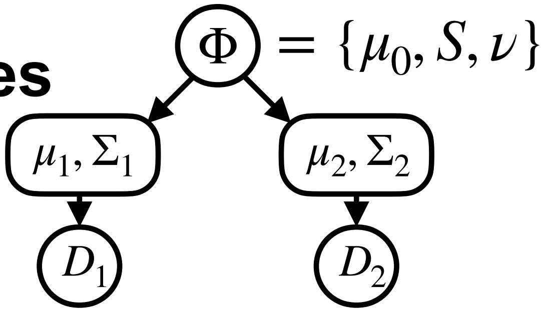
...if we knew **category memberships**, we could put posteriors on **category parameters** $\{\mu_i, \Sigma_i\}$...

...but we don't have either!

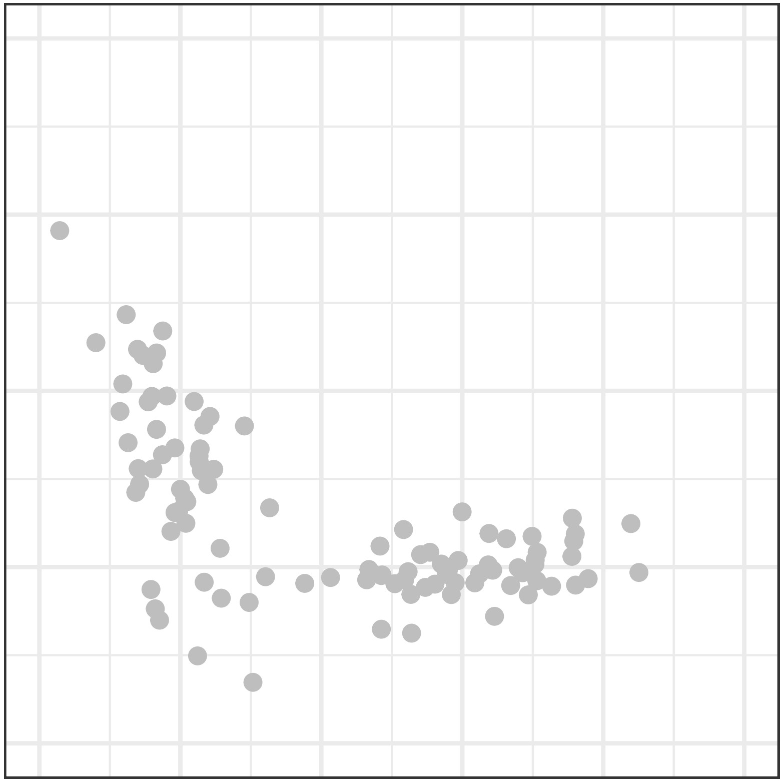
Motivating Gibbs sampling



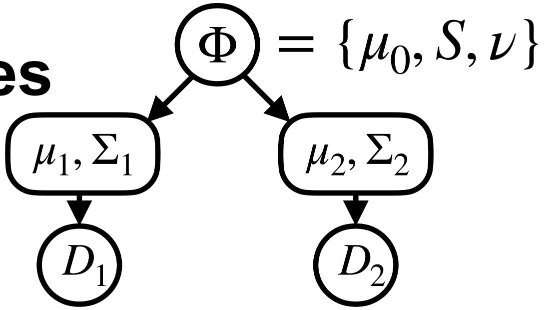
Clustering into multiple categories



Motivating Gibbs sampling



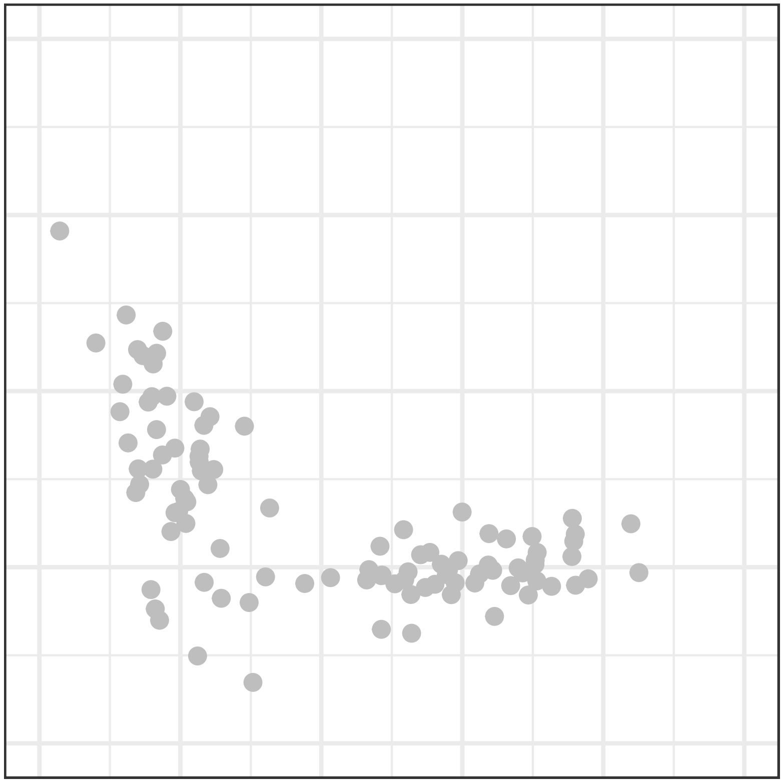
Clustering into multiple categories



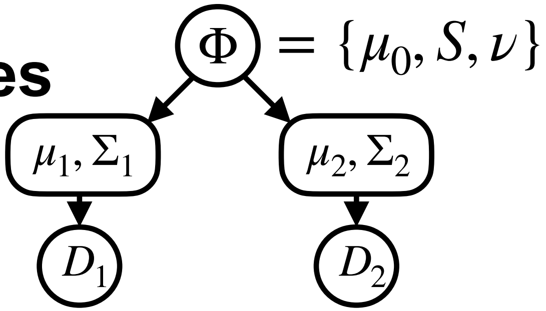
- We can score observation–category assignments Z:

  $P(Z \mid \Phi, D) \propto P(D \mid Z, \Phi) P(Z \mid \Phi)$

Motivating Gibbs sampling



Clustering into multiple categories

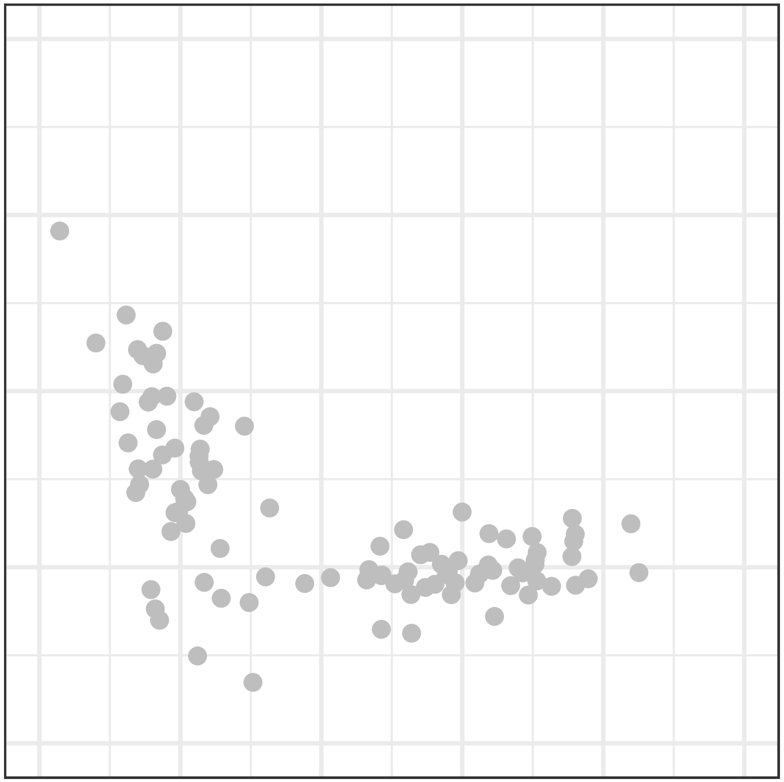


- We can score observation–category assignments Z:

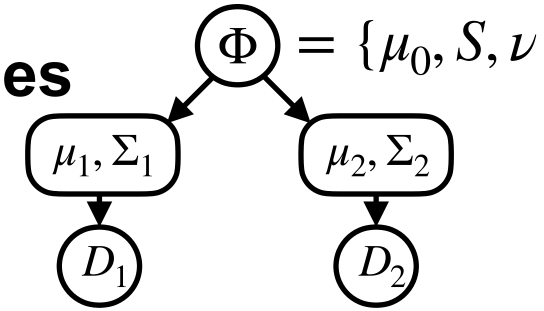
$$P(Z \mid \Phi, D) \propto P(D \mid Z, \Phi) P(Z \mid \Phi)$$

$D_i \equiv$ observations assigned by Z to category i

Motivating Gibbs sampling



Clustering into multiple categories



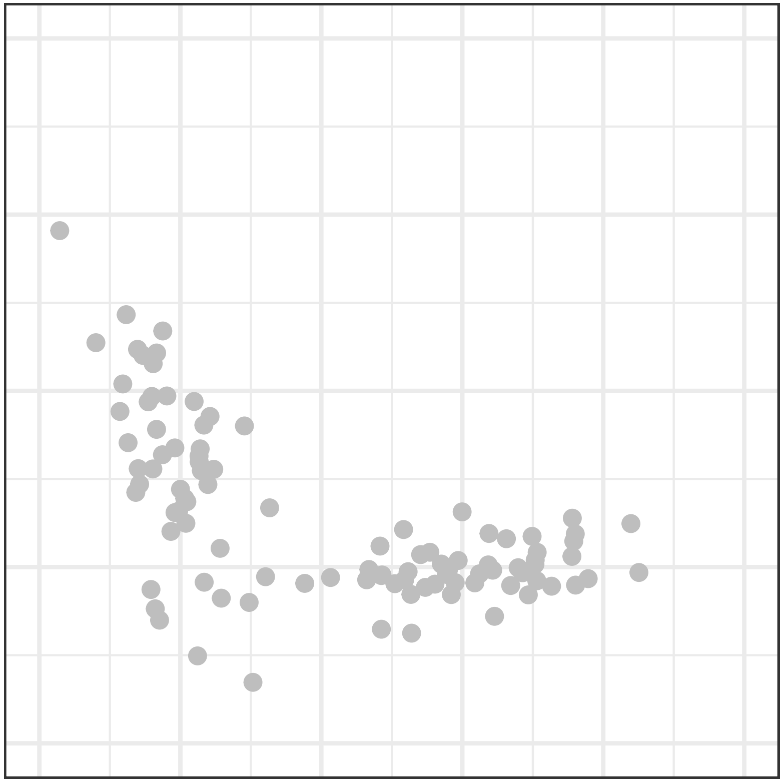
- We can score observation–category assignments Z:

$$P(Z \mid \Phi, D) \propto P(D \mid Z, \Phi) P(Z \mid \Phi)$$

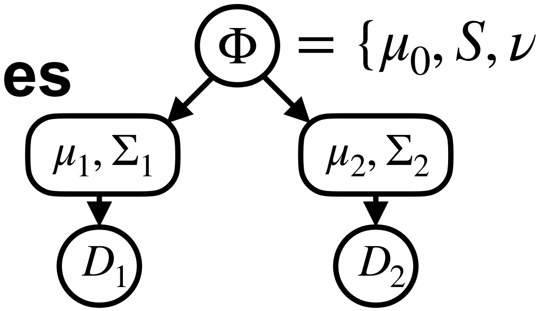
$$P(Z \mid \Phi, D) \propto P(D_1 \mid \Phi) P(D_2 \mid \Phi) P(Z \mid \Phi)$$

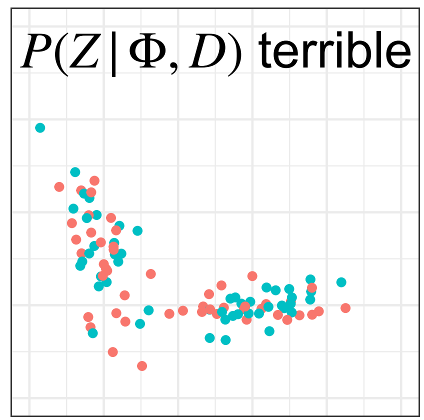
$D_i \equiv$ observations assigned by Z to category i

Motivating Gibbs sampling



Clustering into multiple categories





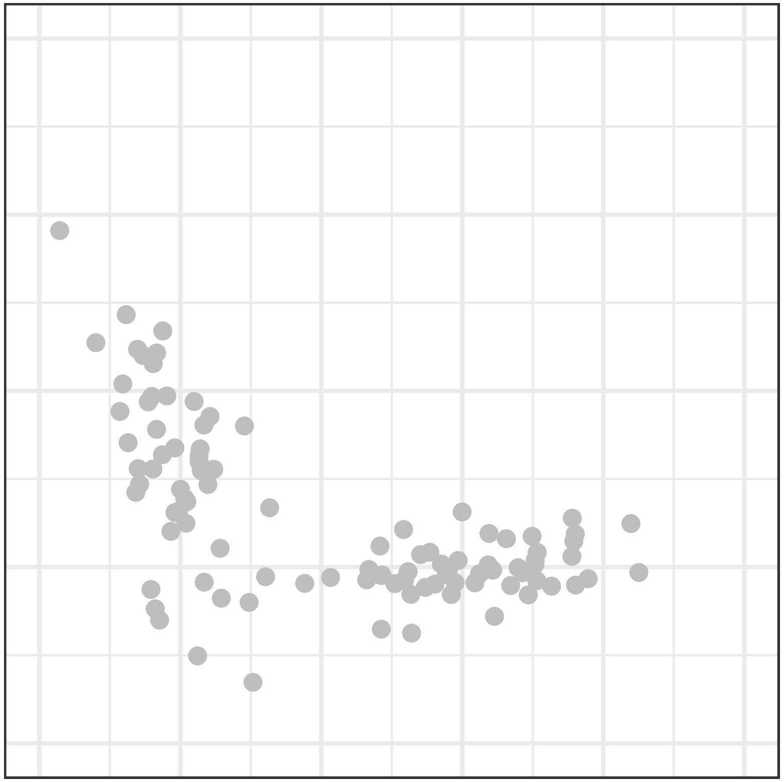
- We can score observation–category assignments Z:

$$P(Z \mid \Phi, D) \propto P(D \mid Z, \Phi) P(Z \mid \Phi)$$

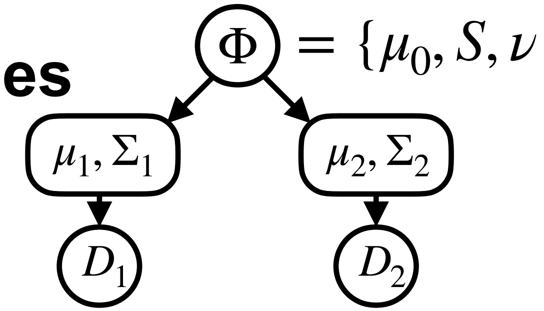
$$P(Z \mid \Phi, D) \propto P(D_1 \mid \Phi) P(D_2 \mid \Phi) P(Z \mid \Phi)$$

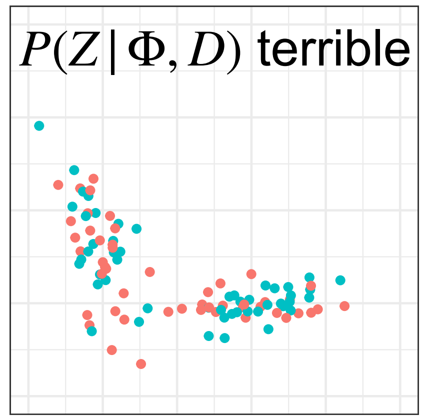
$D_i \equiv$ observations assigned by Z to category i

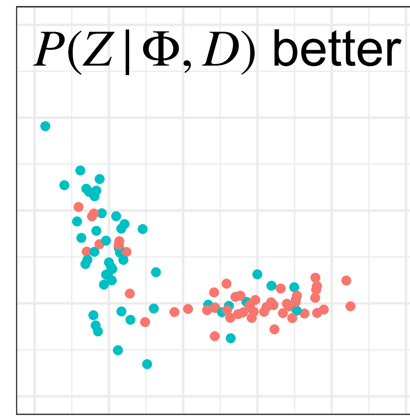
Motivating Gibbs sampling



Clustering into multiple categories







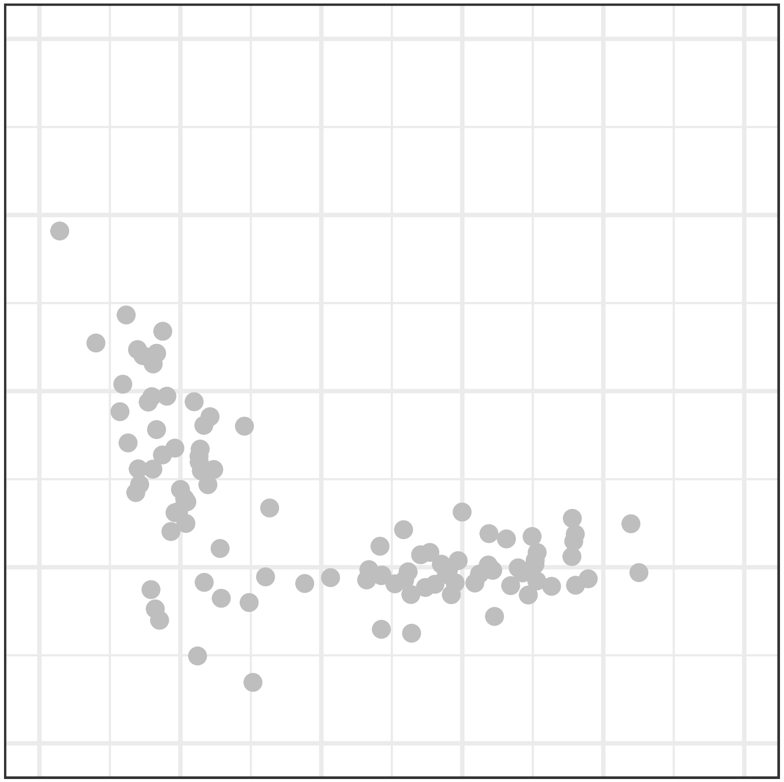
- We can score observation–category assignments Z:

$$P(Z \mid \Phi, D) \propto P(D \mid Z, \Phi) P(Z \mid \Phi)$$

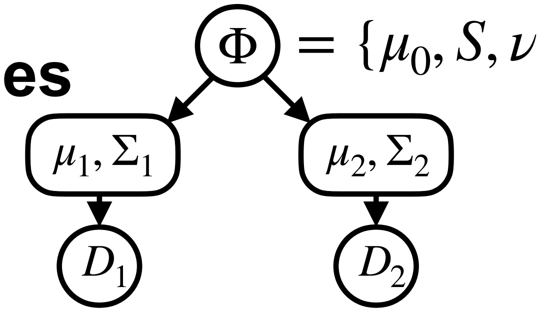
$$P(Z \mid \Phi, D) \propto P(D_1 \mid \Phi) P(D_2 \mid \Phi) P(Z \mid \Phi)$$

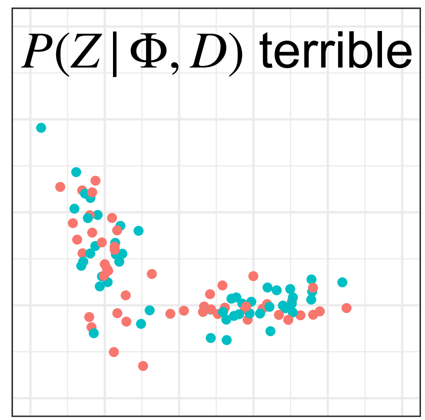
$D_i \equiv$ observations assigned by Z to category i

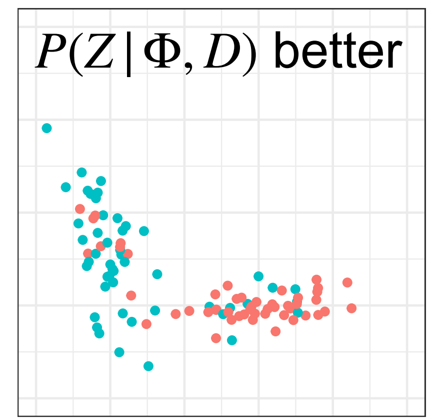
Motivating Gibbs sampling



Clustering into multiple categories







- We can score observation–category assignments Z:

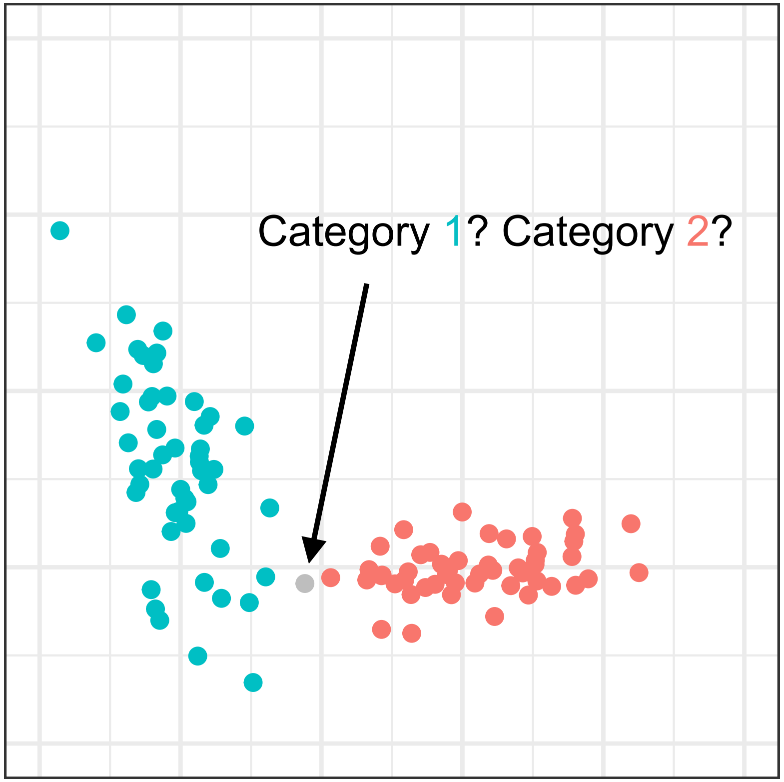
$$P(Z \mid \Phi, D) \propto P(D \mid Z, \Phi) P(Z \mid \Phi)$$

$$P(Z \mid \Phi, D) \propto P(D_1 \mid \Phi) P(D_2 \mid \Phi) P(Z \mid \Phi)$$

$D_i \equiv$ observations assigned by Z to category i

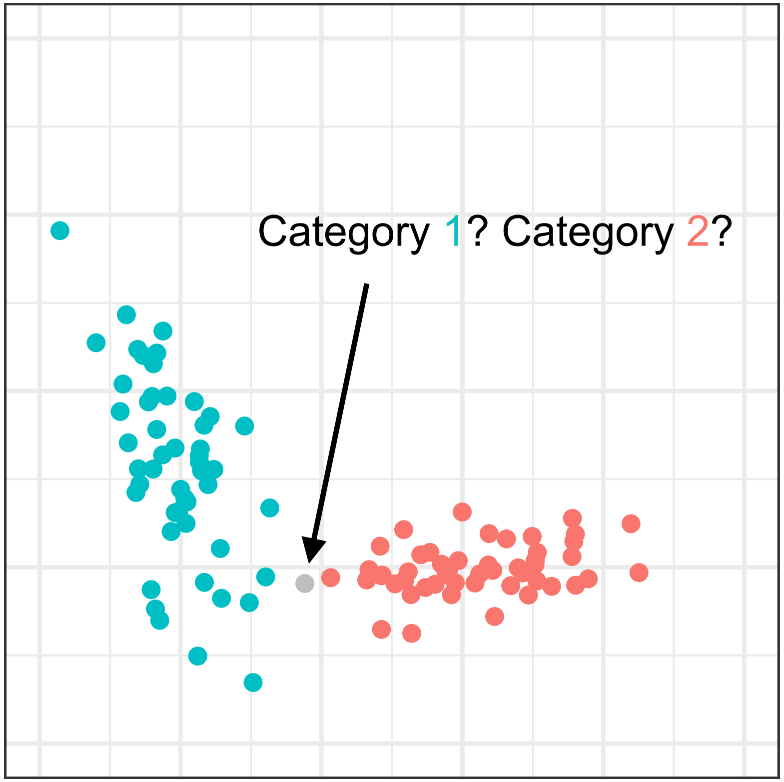
- **Problem:** number of assignments is exponential in data set size (N observations, k classes $\rightarrow k^N$ assignments)

Motivating Gibbs sampling



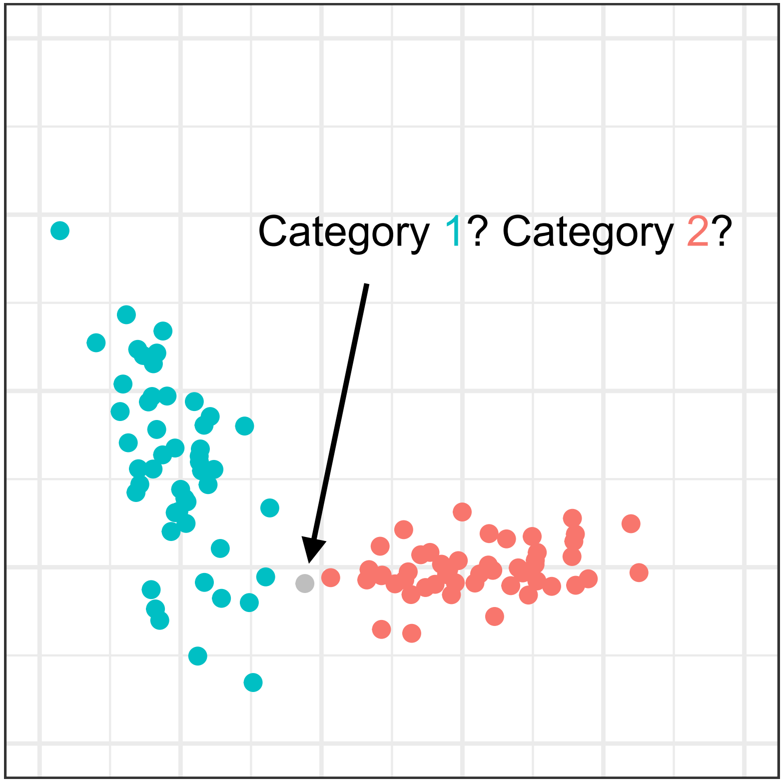
Motivating Gibbs sampling

- **Idea:** suppose we knew category assignments for *all but one* of our observations d_j



Motivating Gibbs sampling

- **Idea:** suppose we knew category assignments for *all but one* of our observations d_j
- Conditional on its assignment z_j given other assignments Z_{-j} easily

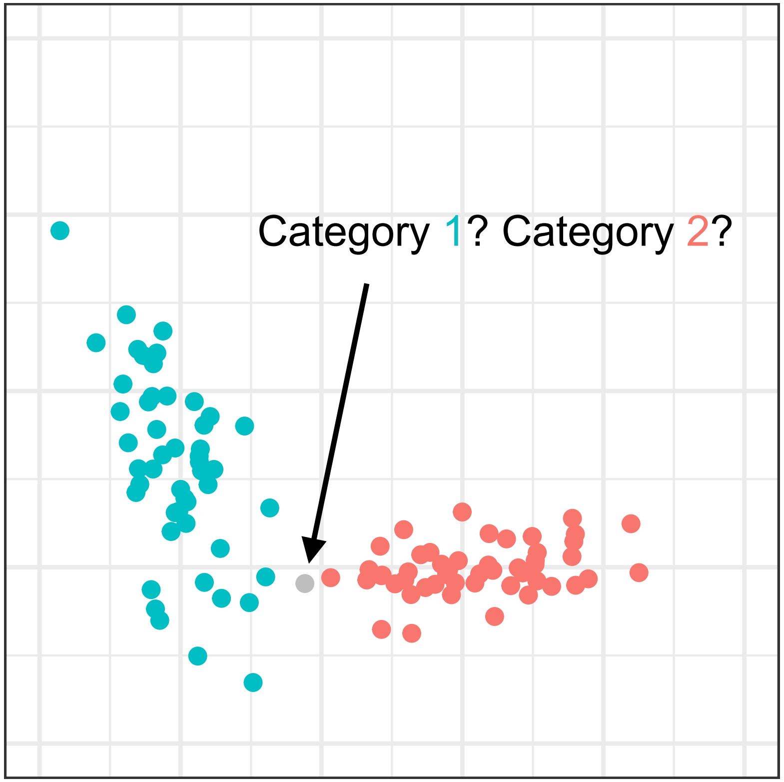


A complete assignment Z

$$P(z_j = i \mid D, Z_{-j}, \Phi) = \frac{P(\overbrace{z_j = i, Z_{-j}}^{\text{A complete assignment } Z} \mid D, \Phi)}{P(Z_{-j} \mid D, \Phi)}$$

Motivating Gibbs sampling

- **Idea:** suppose we knew category assignments for *all but one* of our observations d_j
- Conditional on its assignment z_j given other assignments Z_{-j} easily



$$P(z_j = i \mid D, Z_{-j}, \Phi) = \frac{P(\overbrace{z_j = i, Z_{-j}}^{\text{A complete assignment } Z} \mid D, \Phi)}{P(Z_{-j} \mid D, \Phi)}$$

- **Gibbs sampling approach:** iterate through all observations in your dataset; at each iteration, "forget" the observation's category assignment and resample from $P(z_j = i \mid D, Z_{-j}, \Phi)$

Gibbs sampling, in general

Gibbs sampling, in general

- Suppose there exists a joint distribution over a collection of random variables $X = x_1, \ldots, x_n$

Gibbs sampling, in general

- Suppose there exists a joint distribution over a collection of random variables $X = x_1, \ldots, x_n$
- At time $t = 0$, initialize to $x_1^{(0)}, \ldots, x_n^{(0)}$

Gibbs sampling, in general

- Suppose there exists a joint distribution over a collection of random variables $X = x_1, \ldots, x_n$
- At time $t = 0$, initialize to $x_1^{(0)}, \ldots, x_n^{(0)}$
- At time $t + 1$, for $i \in 1, \ldots, n$, do:

$$x_i^{(t+1)} \sim P(x_i \mid x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_n^{(t)}) \quad *$$

Gibbs sampling, in general

- Suppose there exists a joint distribution over a collection of random variables $X = x_1, \ldots, x_n$
- At time $t = 0$, initialize to $x_1^{(0)}, \ldots, x_n^{(0)}$
- At time $t + 1$, for $i \in 1, \ldots, n$, do:

$$x_i^{(t+1)} \sim P(x_i \,|\, x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_n^{(t)}) \;*$$

- This stochastic update of $X^{(t)}$ to $X^{(t+1)}$ constitutes a MARKOV CHAIN on X that over time converges to the joint distribution $P(x_1, \ldots, x_n)$†

†*Provided this Markov Chain has a stationary distribution and is unique; Robert & Casella, 2004*

Gibbs sampling, in general

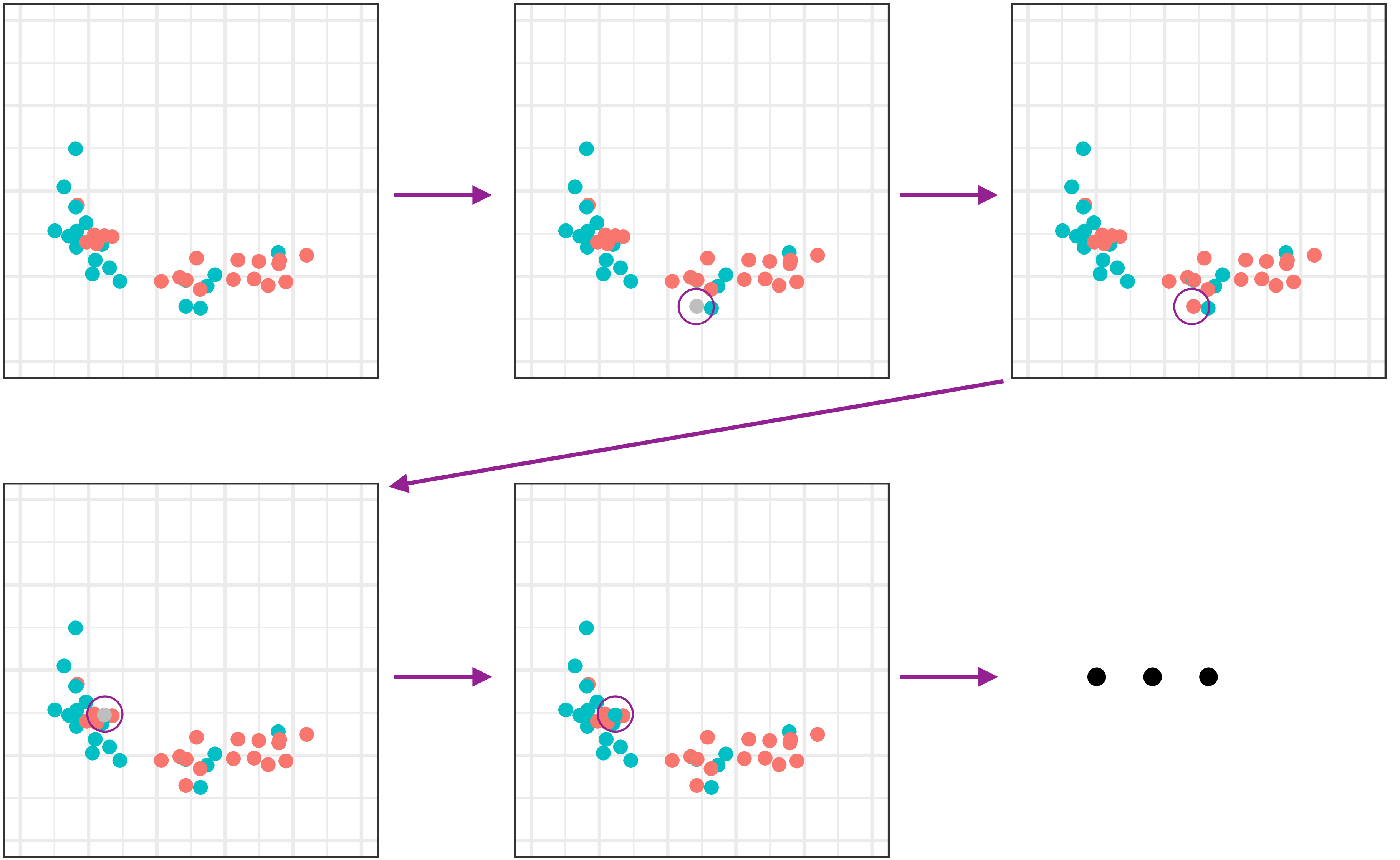
- Suppose there exists a joint distribution over a collection of random variables $X = x_1, \ldots, x_n$
- At time $t = 0$, initialize to $x_1^{(0)}, \ldots, x_n^{(0)}$
- At time $t+1$, for $i \in 1, \ldots, n$, do:

$$x_i^{(t+1)} \sim P(x_i \,|\, x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_n^{(t)}) \quad *$$

- This stochastic update of $X^{(t)}$ to $X^{(t+1)}$ constitutes a MARKOV CHAIN on X that over time converges to the joint distribution $P(x_1, \ldots, x_n)$†
- This is useful in cases (such as the one we've just seen) where working with the full joint distribution is hard, but working with the conditional distribution * is easier

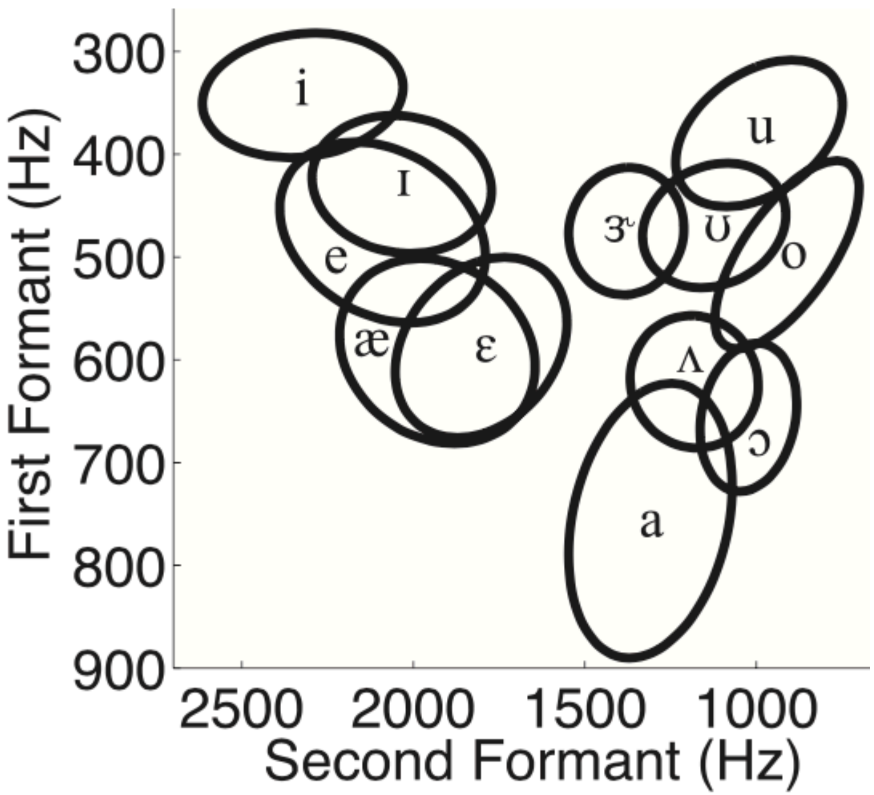
†*Provided this Markov Chain has a stationary distribution and is unique; Robert & Casella, 2004*

Gibbs sampling in action

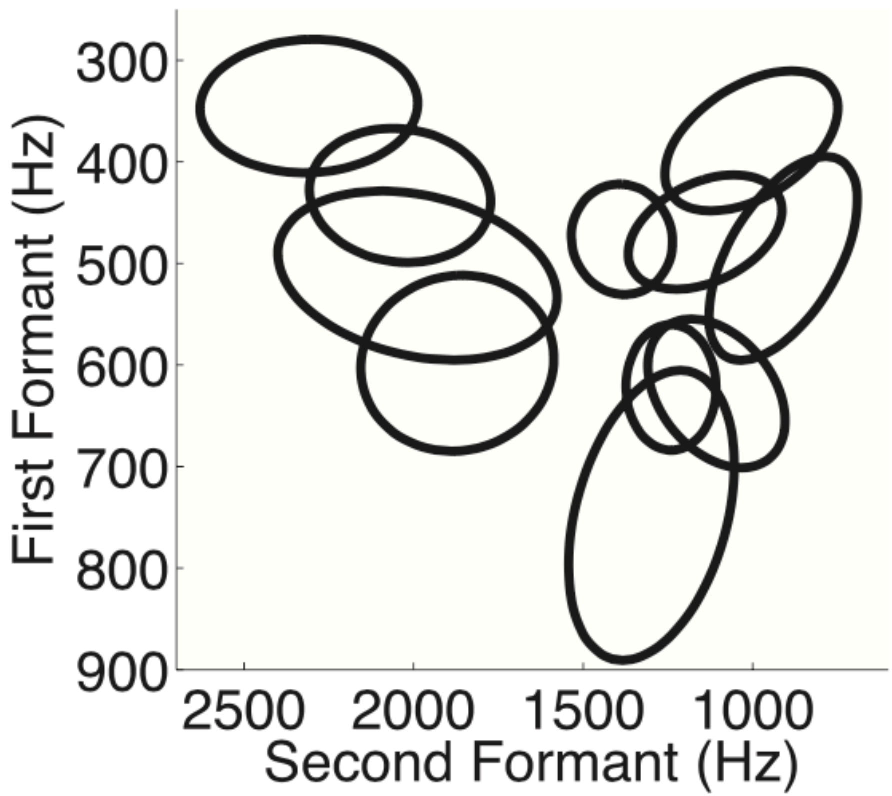


Results on real vowel data

Empirical distribution



Learned distribution



(Feldman et al., 2013)

Limitations I haven't gotten to yet

- I haven't told you about learning the *number of categories*
  - General idea: trade off pressure for *fewer categories* with *better fit to data*
- The learned category set still confuses similar vowels
- We haven't taken into account key sources of information used by linguists (and probably human learners!) to judge whether two tokens are in the same phonetic category: linguistic and extra-linguistic context! *(Feldman et al., 2013)*

	æ	ɛ
h_nd	hænd	
m_t	mæt	mɛt
fr_nd		frɛnd

Limitations I haven't gotten to yet

- I haven't told you about learning the *number of categories*
  - General idea: trade off pressure for *fewer categories* with *better fit to data*
- The learned category set still confuses similar vowels
- We haven't taken into account key sources of information used by linguists (and probably human learners!) to judge whether two tokens are in the same phonetic category: linguistic and extra-linguistic context! *(Feldman et al., 2013)*

	æ	ɛ
h_nd	hænd	
m_t	mæt	mɛt
fr_nd		frɛnd

Limitations I haven't gotten to yet

- I haven't told you about learning the *number of categories*
  - General idea: trade off pressure for *fewer categories* with *better fit to data*
- The learned category set still confuses similar vowels
- We haven't taken into account key sources of information used by linguists (and probably human learners!) to judge whether two tokens are in the same phonetic category: linguistic and extra-linguistic context! *(Feldman et al., 2013)*

	æ	ɛ
h_nd	hænd	
m_t	mæt	mɛt
fr_nd		frɛnd

Early word acquisition as statistical learning

Statistical Learning by 8-Month-Old Infants

Jenny R. Saffran, Richard N. Aslin, Elissa L. Newport

Learners rely on a combination of experience-independent and experience-dependent mechanisms to extract information from the environment. Language acquisition involves both types of mechanisms, but most theorists emphasize the relative importance of experience-independent mechanisms. The present study shows that a fundamental task of language acquisition, segmentation of words from fluent speech, can be accomplished by 8-month-old infants based solely on the statistical relationships between neighboring speech sounds. Moreover, this word segmentation was based on statistical learning from only 2 minutes of exposure, suggesting that infants have access to a powerful mechanism for the computation of statistical properties of the language input.

During early development, the speed and accuracy with which an organism extracts environmental information can be extremely important for its survival. Some species have evolved highly constrained neural mechanisms to ensure that environmental information is properly interpreted, even in the absence of experience with the environment (*1*). Other species are dependent on a period of interaction with the environment that clarifies the information to which attention should be directed and the consequences of behaviors guided by that information (*2*). Depending on the developmental status and the task facing a particular organism, both experience-independent and experience-dependent mechanisms may be involved in the extraction of information and the control of behavior.

In the domain of language acquisition, two facts have supported the interpretation that experience-independent mechanisms are both necessary and dominant. First, highly complex forms of language production develop extremely rapidly (*3*). Second, the language input available to the young child is both incomplete and sparsely represented compared to the child's eventual linguistic abilities (*4*). Thus, most theories of language acquisition have emphasized the critical role played by experience-independent internal structures over the role of experience-dependent factors (*5*).

It is undeniable that experience-dependent mechanisms are also required for the acquisition of language. Many aspects of a particular natural language must be acquired from listening experience. For example, acquiring the specific words and phonological structure of a language requires exposure to a significant corpus of language input. Moreover, long before infants begin to produce their native language, they acquire information about its sound properties (*6*). Nevertheless, given the daunting task of acquiring linguistic information from listening experience during early development, few theorists have entertained the hypothesis that learning plays a primary role in the acquisition of more complicated aspects of language, favoring instead experience-independent mechanisms (*7*). Young humans are generally viewed as poor learners, suggesting that innate factors are primarily responsible for the acquisition of language.

Here we investigate the nature of the

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.

(Saffran et al., 1996; Aslin et al., 1998)

Early word acquisition as statistical learning

Statistical Learning by 8-Month-Old Infants

Jenny R. Saffran, Richard N. Aslin, Elissa L. Newport

Learners rely on a combination of experience-independent and experience-dependent mechanisms to extract information from the environment. Language acquisition involves both types of mechanisms, but most theorists emphasize the relative importance of experience-independent mechanisms. The present study shows that a fundamental task of language acquisition, segmentation of words from fluent speech, can be accomplished by 8-month-old infants based solely on the statistical relationships between neighboring speech sounds. Moreover, this word segmentation was based on statistical learning from only 2 minutes of exposure, suggesting that infants have access to a powerful mechanism for the computation of statistical properties of the language input.

During early development, the speed and accuracy with which an organism extracts environmental information can be extremely important for its survival. Some species have evolved highly constrained neural mechanisms to ensure that environmental information is properly interpreted, even in the absence of experience with the environment (*1*). Other species are dependent on a period of interaction with the environment that clarifies the information to which attention should be directed and the consequences of behaviors guided by that information (*2*). Depending on the developmental status and the task facing a particular organism, both experience-independent and experience-dependent mechanisms may be involved in the extraction of information and the control of behavior.

In the domain of language acquisition, two facts have supported the interpretation that experience-independent mechanisms are both necessary and dominant. First, highly complex forms of language production develop extremely rapidly (*3*). Second, the language input available to the young child is both incomplete and sparsely represented compared to the child's eventual linguistic abilities (*4*). Thus, most theories of language acquisition have emphasized the critical role played by experience-independent internal structures over the role of experience-dependent factors (*5*).

It is undeniable that experience-dependent mechanisms are also required for the acquisition of language. Many aspects of a particular natural language must be acquired from listening experience. For example, acquiring the specific words and phonological structure of a language requires exposure to a significant corpus of language input. Moreover, long before infants begin to produce their native language, they acquire information about its sound properties (*6*). Nevertheless, given the daunting task of acquiring linguistic information from listening experience during early development, few theorists have entertained the hypothesis that learning plays a primary role in the acquisition of more complicated aspects of language, favoring instead experience-independent mechanisms (*7*). Young humans are generally viewed as poor learners, suggesting that innate factors are primarily responsible for the acquisition of language.

Here we investigate the nature of the

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.

Stimulus from Saffran et al. 1996 type experiment

(Saffran et al., 1996; Aslin et al., 1998)

Early word acquisition as statistical learning

Statistical Learning by 8-Month-Old Infants

Jenny R. Saffran, Richard N. Aslin, Elissa L. Newport

Learners rely on a combination of experience-independent and experience-dependent mechanisms to extract information from the environment. Language acquisition involves both types of mechanisms, but most theorists emphasize the relative importance of experience-independent mechanisms. The present study shows that a fundamental task of language acquisition, segmentation of words from fluent speech, can be accomplished by 8-month-old infants based solely on the statistical relationships between neighboring speech sounds. Moreover, this word segmentation was based on statistical learning from only 2 minutes of exposure, suggesting that infants have access to a powerful mechanism for the computation of statistical properties of the language input.

During early development, the speed and accuracy with which an organism extracts environmental information can be extremely important for its survival. Some species have evolved highly constrained neural mechanisms to ensure that environmental information is properly interpreted, even in the absence of experience with the environment (*1*). Other species are dependent on a period of interaction with the environment that clarifies the information to which attention should be directed and the consequences of behaviors guided by that information (*2*). Depending on the developmental status and the task facing a particular organism, both experience-independent and experience-dependent mechanisms may be involved in the extraction of information and the control of behavior.

In the domain of language acquisition, two facts have supported the interpretation that experience-independent mechanisms are both necessary and dominant. First, highly complex forms of language production develop extremely rapidly (*3*). Second, the language input available to the young child is both incomplete and sparsely represented compared to the child's eventual linguistic abilities (*4*). Thus, most theories of language acquisition have emphasized the critical role played by experience-independent internal structures over the role of experience-dependent factors (*5*).

It is undeniable that experience-dependent mechanisms are also required for the acquisition of language. Many aspects of a particular natural language must be acquired from listening experience. For example, acquiring the specific words and phonological structure of a language requires exposure to a significant corpus of language input. Moreover, long before infants begin to produce their native language, they acquire information about its sound properties (*6*). Nevertheless, given the daunting task of acquiring linguistic information from listening experience during early development, few theorists have entertained the hypothesis that learning plays a primary role in the acquisition of more complicated aspects of language, favoring instead experience-independent mechanisms (*7*). Young humans are generally viewed as poor learners, suggesting that innate factors are primarily responsible for the acquisition of language.

Here we investigate the nature of the

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.

Stimulus from Saffran et al. 1996 type experiment

(Saffran et al., 1996; Aslin et al., 1998)

Early word acquisition as statistical learning

Statistical Learning by 8-Month-Old Infants

Jenny R. Saffran, Richard N. Aslin, Elissa L. Newport

Learners rely on a combination of experience-independent and experience-dependent mechanisms to extract information from the environment. Language acquisition involves both types of mechanisms, but most theorists emphasize the relative importance of experience-independent mechanisms. The present study shows that a fundamental task of language acquisition, segmentation of words from fluent speech, can be accomplished by 8-month-old infants based solely on the statistical relationships between neighboring speech sounds. Moreover, this word segmentation was based on statistical learning from only 2 minutes of exposure, suggesting that infants have access to a powerful mechanism for the computation of statistical properties of the language input.

During early development, the speed and accuracy with which an organism extracts environmental information can be extremely important for its survival. Some species have evolved highly constrained neural mechanisms to ensure that environmental information is properly interpreted, even in the absence of experience with the environment (*1*). Other species are dependent on a period of interaction with the environment that clarifies the information to which attention should be directed and the consequences of behaviors guided by that information (*2*). Depending on the developmental status and the task facing a particular organism, both experience-independent and experience-dependent mechanisms may be involved in the extraction of information and the control of behavior.

In the domain of language acquisition, two facts have supported the interpretation that experience-independent mechanisms are both necessary and dominant. First, highly complex forms of language production develop extremely rapidly (*3*). Second, the language input available to the young child is both incomplete and sparsely represented compared to the child's eventual linguistic abilities (*4*). Thus, most theories of language acquisition have emphasized the critical role played by experience-independent internal structures over the role of experience-dependent factors (*5*).

It is undeniable that experience-dependent mechanisms are also required for the acquisition of language. Many aspects of a particular natural language must be acquired from listening experience. For example, acquiring the specific words and phonological structure of a language requires exposure to a significant corpus of language input. Moreover, long before infants begin to produce their native language, they acquire information about its sound properties (*6*). Nevertheless, given the daunting task of acquiring linguistic information from listening experience during early development, few theorists have entertained the hypothesis that learning plays a primary role in the acquisition of more complicated aspects of language, favoring instead experience-independent mechanisms (*7*). Young humans are generally viewed as poor learners, suggesting that innate factors are primarily responsible for the acquisition of language.

Here we investigate the nature of the

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.

SCIENCE • VOL. 274 • 13 DECEMBER 1996

Stimulus from Saffran et al. 1996 type experiment

What do you think the words are?

(Saffran et al., 1996; Aslin et al., 1998)

Early word acquisition as statistical learning

Statistical Learning by 8-Month-Old Infants

Jenny R. Saffran, Richard N. Aslin, Elissa L. Newport

Learners rely on a combination of experience-independent and experience-dependent mechanisms to extract information from the environment. Language acquisition involves both types of mechanisms, but most theorists emphasize the relative importance of experience-independent mechanisms. The present study shows that a fundamental task of language acquisition, segmentation of words from fluent speech, can be accomplished by 8-month-old infants based solely on the statistical relationships between neighboring speech sounds. Moreover, this word segmentation was based on statistical learning from only 2 minutes of exposure, suggesting that infants have access to a powerful mechanism for the computation of statistical properties of the language input.

During early development, the speed and accuracy with which an organism extracts environmental information can be extremely important for its survival. Some species have evolved highly constrained neural mechanisms to ensure that environmental information is properly interpreted, even in the absence of experience with the environment (*1*). Other species are dependent on a period of interaction with the environment that clarifies the information to which attention should be directed and the consequences of behaviors guided by that information (*2*). Depending on the developmental status and the task facing a particular organism, both experience-independent and experience-dependent mechanisms may be involved in the extraction of information and the control of behavior.

In the domain of language acquisition, two facts have supported the interpretation that experience-independent mechanisms are both necessary and dominant. First, highly complex forms of language production develop extremely rapidly (*3*). Second, the language input available to the young child is both incomplete and sparsely represented compared to the child's eventual linguistic abilities (*4*). Thus, most theories of language acquisition have emphasized the critical role played by experience-independent internal structures over the role of experience-dependent factors (*5*).

It is undeniable that experience-dependent mechanisms are also required for the acquisition of language. Many aspects of a particular natural language must be acquired from listening experience. For example, acquiring the specific words and phonological structure of a language requires exposure to a significant corpus of language input. Moreover, long before infants begin to produce their native language, they acquire information about its sound properties (*6*). Nevertheless, given the daunting task of acquiring linguistic information from listening experience during early development, few theorists have entertained the hypothesis that learning plays a primary role in the acquisition of more complicated aspects of language, favoring instead experience-independent mechanisms (*7*). Young humans are generally viewed as poor learners, suggesting that innate factors are primarily responsible for the acquisition of language.

Here we investigate the nature of the

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.

SCIENCE • VOL. 274 • 13 DECEMBER 1996

Stimulus from Saffran et al. 1996 type experiment

What do you think the words are?

pigola
golatu

daropi
tudaro

(Saffran et al., 1996; Aslin et al., 1998)

Early word acquisition as statistical learning

Statistical Learning by 8-Month-Old Infants

Jenny R. Saffran, Richard N. Aslin, Elissa L. Newport

Learners rely on a combination of experience-independent and experience-dependent mechanisms to extract information from the environment. Language acquisition involves both types of mechanisms, but most theorists emphasize the relative importance of experience-independent mechanisms. The present study shows that a fundamental task of language acquisition, segmentation of words from fluent speech, can be accomplished by 8-month-old infants based solely on the statistical relationships between neighboring speech sounds. Moreover, this word segmentation was based on statistical learning from only 2 minutes of exposure, suggesting that infants have access to a powerful mechanism for the computation of statistical properties of the language input.

During early development, the speed and accuracy with which an organism extracts environmental information can be extremely important for its survival. Some species have evolved highly constrained neural mechanisms to ensure that environmental information is properly interpreted, even in the absence of experience with the environment (1). Other species are dependent on a period of interaction with the environment that clarifies the information to which attention should be directed and the consequences of behaviors guided by that information (2). Depending on the developmental status and the task facing a particular organism, both experience-independent and experience-dependent mechanisms may be involved in the extraction of information and the control of behavior.

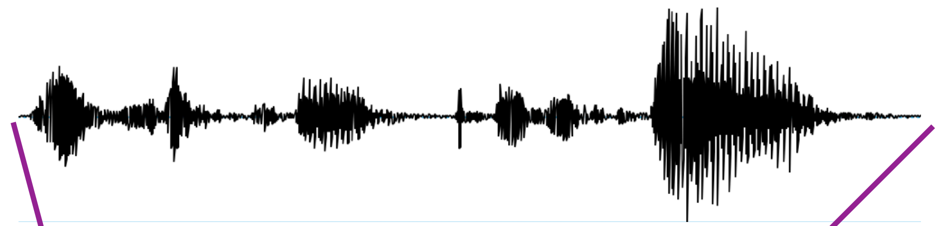
In the domain of language acquisition, two facts have supported the interpretation that experience-independent mechanisms are both necessary and dominant. First, highly complex forms of language production develop extremely rapidly (3). Second, the language input available to the young child is both incomplete and sparsely represented compared to the child's eventual linguistic abilities (4). Thus, most theories of language acquisition have emphasized the critical role played by experience-independent internal structures over the role of experience-dependent factors (5).

It is undeniable that experience-dependent mechanisms are also required for the acquisition of language. Many aspects of a particular natural language must be acquired from listening experience. For example, acquiring the specific words and phonological structure of a language requires exposure to a significant corpus of language input. Moreover, long before infants begin to produce their native language, they acquire information about its sound properties (6). Nevertheless, given the daunting task of acquiring linguistic information from listening experience during early development, few theorists have entertained the hypothesis that learning plays a primary role in the acquisition of more complicated aspects of language, favoring instead experience-independent mechanisms (7). Young humans are generally viewed as poor learners, suggesting that innate factors are primarily responsible for the acquisition of language.

Here we investigate the nature of the

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.

SCIENCE • VOL. 274 • 13 DECEMBER 1996



it's a cute kitty-cat

Stimulus from Saffran et al. 1996 type experiment

What do you think the words are?

pigola **daropi**
golatu **tudaro**

(Saffran et al., 1996; Aslin et al., 1998)

Early word acquisition as statistical learning

Statistical Learning by 8-Month-Old Infants

Jenny R. Saffran, Richard N. Aslin, Elissa L. Newport

Learners rely on a combination of experience-independent and experience-dependent mechanisms to extract information from the environment. Language acquisition involves both types of mechanisms, but most theorists emphasize the relative importance of experience-independent mechanisms. The present study shows that a fundamental task of language acquisition, segmentation of words from fluent speech, can be accomplished by 8-month-old infants based solely on the statistical relationships between neighboring speech sounds. Moreover, this word segmentation was based on statistical learning from only 2 minutes of exposure, suggesting that infants have access to a powerful mechanism for the computation of statistical properties of the language input.

During early development, the speed and accuracy with which an organism extracts environmental information can be extremely important for its survival. Some species have evolved highly constrained neural mechanisms to ensure that environmental information is properly interpreted, even in the absence of experience with the environment (1). Other species are dependent on a period of interaction with the environment that clarifies the information to which attention should be directed and the consequences of behaviors guided by that information (2). Depending on the developmental status and the task facing a particular organism, both experience-independent and experience-dependent mechanisms may be involved in the extraction of information and the control of behavior.

In the domain of language acquisition, two facts have supported the interpretation that experience-independent mechanisms are both necessary and dominant. First, highly complex forms of language production develop extremely rapidly (3). Second, the language input available to the young child is both incomplete and sparsely represented compared to the child's eventual linguistic abilities (4). Thus, most theories of language acquisition have emphasized the critical role played by experience-independent internal structures over the role of experience-dependent factors (5).

It is undeniable that experience-dependent mechanisms are also required for the acquisition of language. Many aspects of a particular natural language must be acquired from listening experience. For example, acquiring the specific words and phonological structure of a language requires exposure to a significant corpus of language input. Moreover, long before infants begin to produce their native language, they acquire information about its sound properties (6). Nevertheless, given the daunting task of acquiring linguistic information from listening experience during early development, few theorists have entertained the hypothesis that learning plays a primary role in the acquisition of more complicated aspects of language, favoring instead experience-independent mechanisms (7). Young humans are generally viewed as poor learners, suggesting that innate factors are primarily responsible for the acquisition of language.

Here we investigate the nature of the

Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA.

Stimulus from Saffran et al. 1996 type experiment

What do you think the words are?

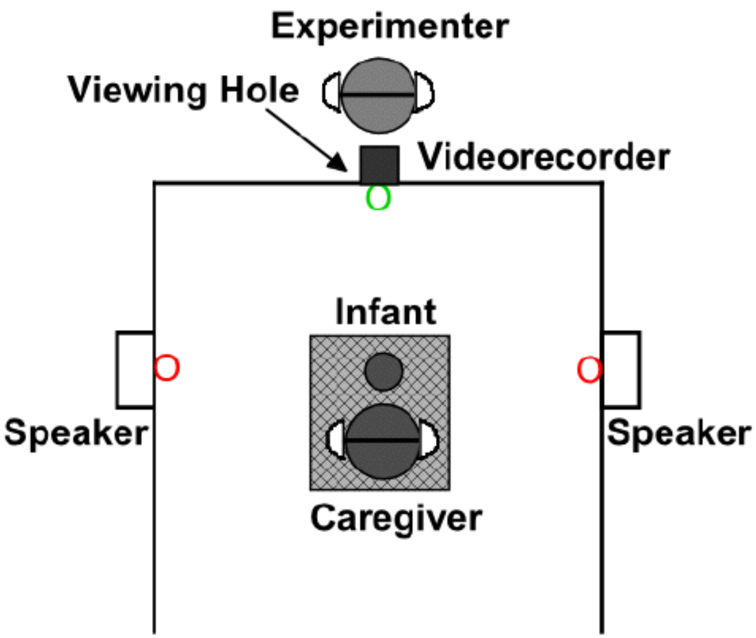
pigola **daropi**

golatu **tudaro**

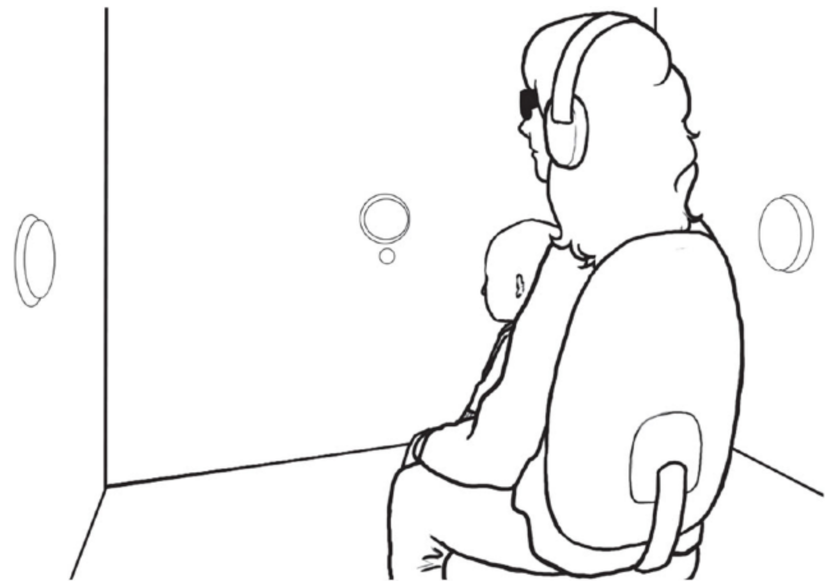
(Saffran et al., 1996; Aslin et al., 1998)

https://www.youtube.com/watch?v=EFIxiflDk_o, starting ~5:34

Head turn preference procedure



(Figure from Tincoff & Jusczyk, 1996)



(Figure from Gervain & Werker, 2013)

The information in transition probabilities

aɪ si ə da gi
pɛt ðə kɪ ti
si ðə bal
aɪ ləv ju
du ju si ðə da gi

(Saffran et al., 1996; Aslin et al., 1998)

The information in transition probabilities

aɪ si ə da gi
pɛt ðə kɪ ti
si ðə bal
aɪ ləv ju
du ju si ðə da gi

S_{i-1}		S_i	$P(S_{i-1}\|S_i)$
aɪ	→	si	0.5
	→	ləv	0.5
si	→	ə	0.33
	→	ðə	0.67
ə	→	da	1.0
da	→	gi	1.0
gi	→	<eos>	1.0
pɛt	→	ðə	1.0
ðə	→	kɪ	0.33
	→	bal	0.33
	→	da	0.33
kɪ	→	ti	1.0
ti	→	<eos>	1.0
bal	→	<eos>	1.0
ləv	→	ju	1.0
ju	→	<eos>	0.5
	→	si	0.5
du	→	ju	1.0

(Saffran et al., 1996; Aslin et al., 1998)

The information in transition probabilities

aɪ | si | ə da gi
pɛt ðə | kɪ ti
si | ðə | bal
aɪ | ləv ju
du ju | si | ðə | da gi

S_{i-1}		S_i	$P(S_{i-1} \mid S_i)$
aɪ	→	si	0.5
	→	ləv	0.5
si	→	ə	0.33
	→	ðə	0.67
ə	→	da	1.0
da	→	gi	1.0
gi	→	<eos>	1.0
pɛt	→	ðə	1.0
ðə	→	kɪ	0.33
	→	bal	0.33
	→	da	0.33
kɪ	→	ti	1.0
ti	→	<eos>	1.0
bal	→	<eos>	1.0
ləv	→	ju	1.0
ju	→	<eos>	0.5
	→	si	0.5
du	→	ju	1.0

(Saffran et al., 1996; Aslin et al., 1998)

The information in transition probabilities

aɪ | si | ə da gi
pɛt ðə | kɪ ti
si | ðə | bal
aɪ | ləv ju
du ju | si | ðə | da gi

I see adoggie
pet the kitty
see the ball
I love you
doyou see the doggie

S_{i-1}	S_i	$P(S_{i-1} \mid S_i)$
aɪ	→ si	0.5
	→ ləv	0.5
si	→ ə	0.33
	→ ðə	0.67
ə	→ da	1.0
da	→ gi	1.0
gi	→ <eos>	1.0
pɛt	→ ðə	1.0
ðə	→ kɪ	0.33
	→ bal	0.33
	→ da	0.33
kɪ	→ ti	1.0
ti	→ <eos>	1.0
bal	→ <eos>	1.0
ləv	→ ju	1.0
ju	→ <eos>	0.5
	→ si	0.5
du	→ ju	1.0

(Saffran et al., 1996; Aslin et al., 1998)

Mutual information among syllables

$$\text{MI}(X,Y\,|\,I) = \sum_{x,y} P(X,Y\,|\,I)\log\frac{P(X,Y\,|\,I)}{P(X\,|\,I)P(Y\,|\,I)}$$

[$\text{MI}(X,Y\,|\,I) \geq 0$ always]

Mutual information among syllables

- ***Mutual information*** between two random variables quantifies their strength of association:

$$\text{MI}(X,Y\,|\,I) = \sum_{x,y} P(X,Y\,|\,I)\log\frac{P(X,Y\,|\,I)}{P(X\,|\,I)P(Y\,|\,I)} \qquad [\text{MI}(X,Y\,|\,I) \geq 0 \text{ always}]$$

Mutual information among syllables

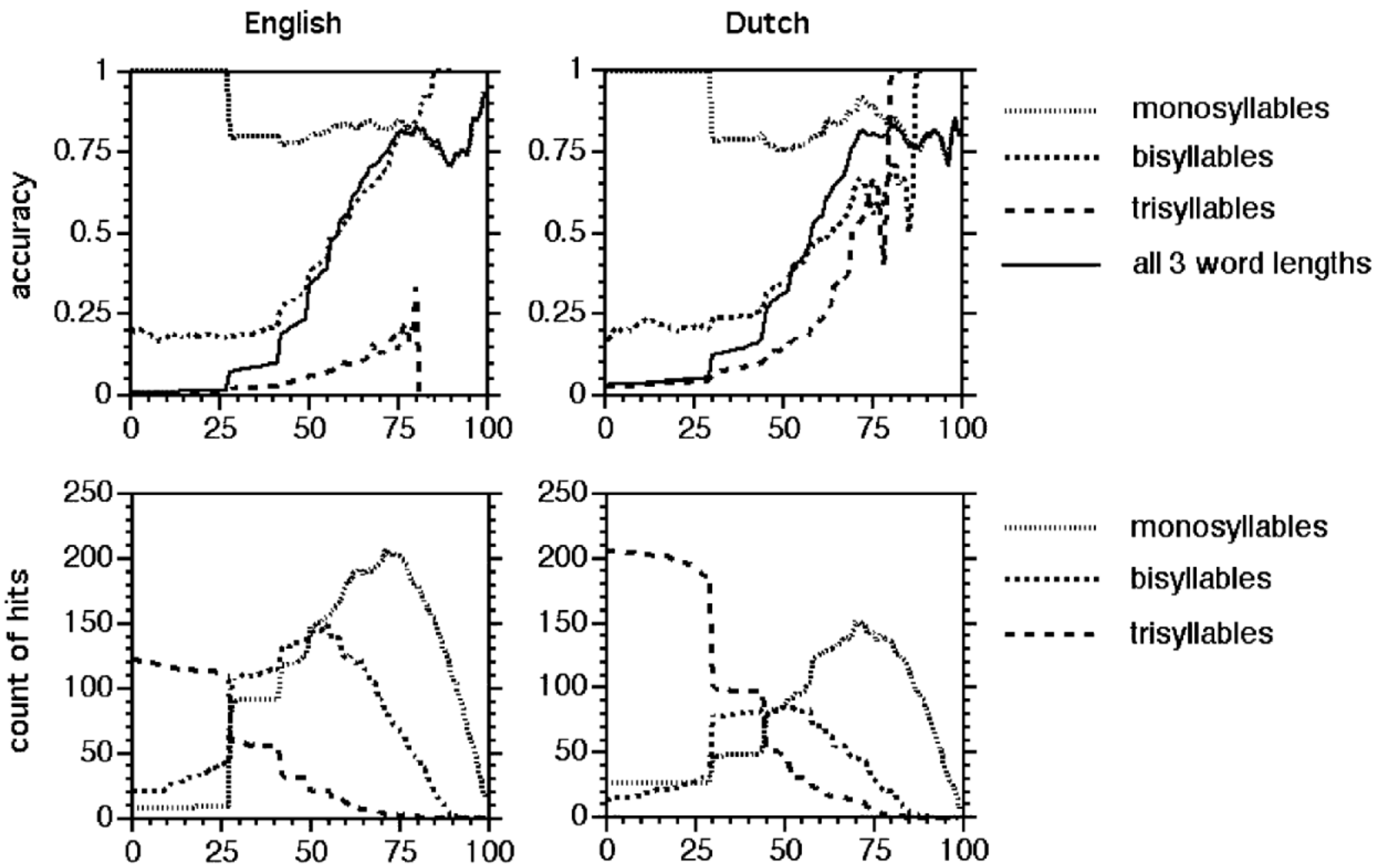
- ***Mutual information*** between two random variables quantifies their strength of association:

$$\text{MI}(X,Y\,|\,I)=\sum_{x,y}P(X,Y\,|\,I)\log\frac{P(X,Y\,|\,I)}{P(X\,|\,I)P(Y\,|\,I)}\qquad[\text{MI}(X,Y\,|\,I)\geq 0\text{ always}]$$

- **P*ointwise mutual information*** evaluates at a single value of $X=x,Y=y$

$$\text{pMI}(x,y\,|\,I)=\log\frac{P(x,y\,|\,I)}{P(x\,|\,I)P(y\,|\,I)}\qquad[\text{can be positive or negative}]$$

Word segmentation with pMI



$$\text{accuracy} = \frac{\text{hits}}{\text{hits} + \text{false alarms}}$$

(Swingley, 1995)

A generative model for word segmentation

a	0.015	dog	0.005
the	0.045	kitty-cat	0.00005
it	0.005	barks	0.0001
is	0.075	today	0.01
cute	0.0005	near	0.015
tall	0.003	ever	0.02
...		...	

A generative model for word segmentation

- Assume a ***probabilistic lexicon***

`a`	`0.015`	`dog`	`0.005`
`the`	`0.045`	`kitty-cat`	`0.00005`
`it`	`0.005`	`barks`	`0.0001`
`is`	`0.075`	`today`	`0.01`
`cute`	`0.0005`	`near`	`0.015`
`tall`	`0.003`	`ever`	`0.02`
`...`		`...`	

A generative model for word segmentation

- Assume a ***probabilistic lexicon***

a	0.015	dog	0.005
the	0.045	kitty-cat	0.00005
it	0.005	barks	0.0001
is	0.075	today	0.01
cute	0.0005	near	0.015
tall	0.003	ever	0.02
...		...	

- Assume a corpus comes into the world through sampling from that lexicon as a ***unigram distribution***

it is a cute kitty-cat
ɪt ɪs ə kjut kɪɾikæt

A generative model for word segmentation

- Assume a ***probabilistic lexicon***

a	0.015	dog	0.005
the	0.045	kitty-cat	0.00005
it	0.005	barks	0.0001
is	0.075	today	0.01
cute	0.0005	near	0.015
tall	0.003	ever	0.02
...		...	

- Assume a corpus comes into the world through sampling from that lexicon as a ***unigram distribution***

it is a cute kitty-cat
ɪt ɪs ə kjut kɪɾikæt

- Word boundaries are implicit in the generative process, but ***unobserved for the learner***

ɪtɪsəkjutkɪɾikæt

A generative model for word segmentation

ɪt|ɪs|ə|kjut|kɪɾi|kæt|

A generative model for word segmentation

- **Goal of learner:** infer segmentation of corpus into **words**

ɪt|ɪs|ə|kjut|kɪɾi|kæt|

A generative model for word segmentation

- **Goal of learner:** infer segmentation of corpus into **words**

ɪt|ɪs|ə|kjut|kɪɾi|kæt|

- Given a segmentation into words, we can consider candidate probabilistic lexica

a	0.015	dog	0.005
the	0.045	kitty-cat	0.00005
it	0.005	barks	0.0001
is	0.075	today	0.01
cute	0.0005	near	0.015
tall	0.003	ever	0.02
...		...	

a	0.02	dog	0.002
the	0.04	kitty-cat	0.0001
it	0.01	barks	0.00005
is	0.08	today	0.005
cute	0.001	near	0.01
tall	0.002	ever	0.025
...		...	

...

A generative model for word segmentation

- **Goal of learner:** infer segmentation of corpus into **words**

ɪt|ɪs|ə|kjut|kɪɾi|kæt|

- Given a segmentation into words, we can consider candidate probabilistic lexica

a	0.015	dog	0.005
the	0.045	kitty-cat	0.00005
it	0.005	barks	0.0001
is	0.075	today	0.01
cute	0.0005	near	0.015
tall	0.003	ever	0.02
...		...	

a	0.02	dog	0.002
the	0.04	kitty-cat	0.0001
it	0.01	barks	0.00005
is	0.08	today	0.005
cute	0.001	near	0.01
tall	0.002	ever	0.025
...		...	

...

- We can consider our standard inference candidates:
  - Maximum likelihood

$$\langle \widehat{\text{Lexicon}} \rangle = \arg\max_{\text{Lexicon}} P(\text{Corpus}) \qquad \langle \widehat{\text{Words}} \rangle = \arg\max_{\text{Words}} P(\text{Corpus}|\text{Words}, \widehat{\text{Lexicon}})$$

A generative model for word segmentation

- **Goal of learner:** infer segmentation of corpus into **words**

ɪt|ɪs|ə|kjut|kɪɾi|kæt|

- Given a segmentation into words, we can consider candidate probabilistic lexica

a	0.015	dog	0.005
the	0.045	kitty-cat	0.00005
it	0.005	barks	0.0001
is	0.075	today	0.01
cute	0.0005	near	0.015
tall	0.003	ever	0.02
...		...	

a	0.02	dog	0.002
the	0.04	kitty-cat	0.0001
it	0.01	barks	0.00005
is	0.08	today	0.005
cute	0.001	near	0.01
tall	0.002	ever	0.025
...		...	

...

- We can consider our standard inference candidates:
  - Maximum likelihood

$$\langle \widehat{\text{Lexicon}} \rangle = \arg\max_{\text{Lexicon}} P(\text{Corpus}) \qquad \langle \widehat{\text{Words}} \rangle = \arg\max_{\text{Words}} P(\text{Corpus}|\text{Words}, \widehat{\text{Lexicon}})$$

  - Bayesian inference

$$\begin{aligned} P(\text{Lexicon}, \text{Words}|\text{Corpus}) &\propto P(\text{Corpus}|\text{Lexicon}, \text{Words}) P(\text{Lexicon}, \text{Words}) \\ &\propto P(\text{Corpus}|\text{Words}) P(\text{Words}|\text{Lexicon}) P(\text{Lexicon}) \end{aligned}$$

A problem for maximum likelihood

Unsegmented		Segmented
aɪ si ə da gi	→	aɪ|si|ə da gi
pɛt ðə kɪ ti		pɛt ðə|kɪ ti
si ðə bal		si|ðə|bal
aɪ ləv ju		aɪ|ləv ju
du ju si ðə da gi		du ju|si|ðə|da gi

a	0.015	dog	0.005
the	0.045	kitty-cat	0.00005
it	0.005	barks	0.0001
is	0.075	today	0.01
cute	0.0005	near	0.015
tall	0.003	ever	0.02
...		...	

$$\langle \widehat{\text{Lexicon}} \rangle = \arg\max_{\text{Lexicon}} P(\text{Corpus}) \qquad \langle \widehat{\text{Words}} \rangle = \arg\max_{\text{Words}} P(\text{Corpus}|\text{Words}, \widehat{\text{Lexicon}})$$

or

$$\langle \widehat{\text{Lexicon}}, \widehat{\text{Words}} \rangle = \arg\max_{\text{Lexicon},\text{Words}} P(\text{Corpus}|\text{Lexicon}, \text{Words})$$

This will not work...why???

Bayesian inference for word segmentation

$$\begin{aligned} P(\text{Lexicon}, \text{Words}|\text{Corpus}) &\propto P(\text{Corpus}|\text{Lexicon}, \text{Words})P(\text{Lexicon}, \text{Words}) \\ &\propto P(\text{Corpus}|\text{Words})P(\text{Words}|\text{Lexicon})P(\text{Lexicon}) \end{aligned}$$

$$P(\text{Corpus}|\text{Words}) = \begin{cases} 1 & \text{The word sequence matches the corpus} \\ 0 & \text{Otherwise} \end{cases}$$

$P(\text{Words}|\text{Lexicon})$ is easy—it's a unigram model

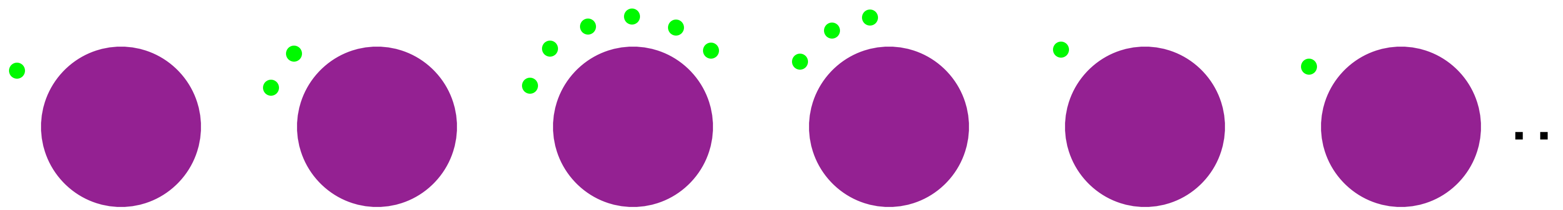
But $P(\text{Lexicon})$ is hard!

Instead, we will **integrate out** the lexicon, so we are doing:

$$P(\text{Words}|\text{Corpus}) \propto \int_{\text{Lexicon}} P(\text{Corpus}|\text{Lexicon}, \text{Words})P(\text{Lexicon}, \text{Words})$$

leaving the lexicon implicit and focusing on inferring likely segmentations of the corpus.

The Chinese Restaurant process

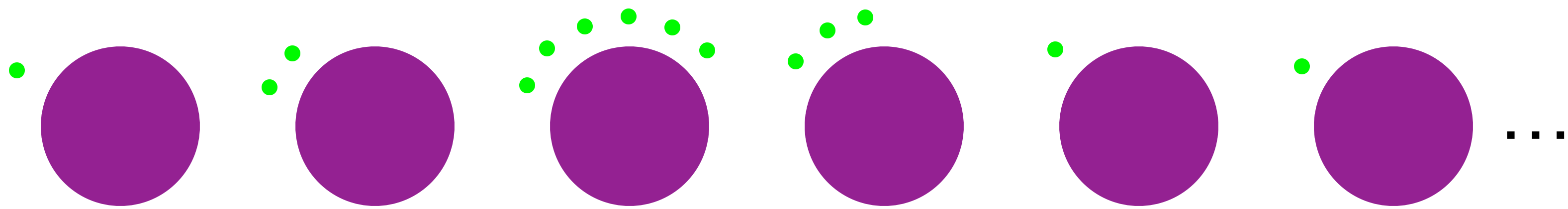


$n_1 = 1 \quad n_2 = 2 \quad n_3 = 6 \quad n_4 = 3 \quad n_5 = 1 \quad n_6 = 1 \quad \ldots$

(Ferguson, 1973; Aldous, 1985; Pitman, 1995)

The Chinese Restaurant process

- The metaphor: an unbounded Chinese restaurant
  - An unbounded number of tables
  - Each table can seat an unbounded number of customers

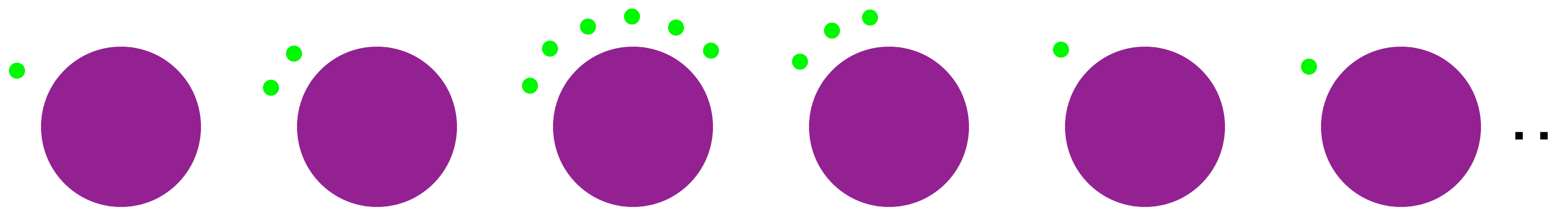


$n_1 = 1 \quad n_2 = 2 \quad n_3 = 6 \quad n_4 = 3 \quad n_5 = 1 \quad n_6 = 1$

(Ferguson, 1973; Aldous, 1985; Pitman, 1995)

The Chinese Restaurant process

- The metaphor: an unbounded Chinese restaurant
  - An unbounded number of tables
  - Each table can seat an unbounded number of customers



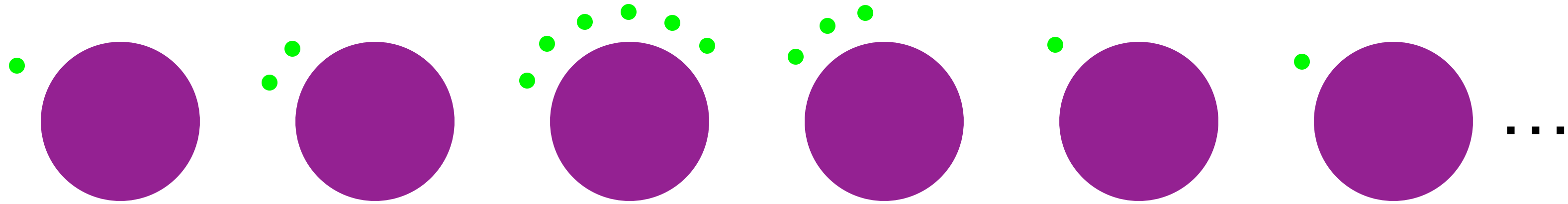
$n_1 = 1$ $n_2 = 2$ $n_3 = 6$ $n_4 = 3$ $n_5 = 1$ $n_6 = 1$

- For our learning problem, the tables are **categories**, and the customers at a table are **instances of that category**

(Ferguson, 1973; Aldous, 1985; Pitman, 1995)

The Chinese Restaurant process

- The metaphor: an unbounded Chinese restaurant
  - An unbounded number of tables
  - Each table can seat an unbounded number of customers



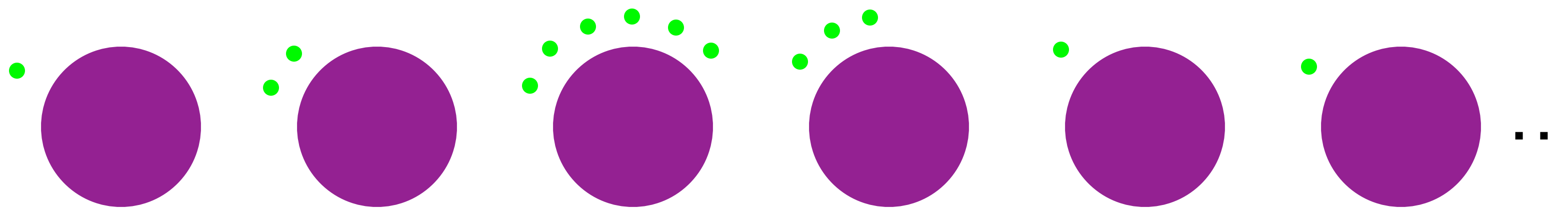
$n_1 = 1$ $\quad n_2 = 2$ $\quad n_3 = 6$ $\quad n_4 = 3$ $\quad n_5 = 1$ $\quad n_6 = 1$

- For our learning problem, the tables are **categories**, and the customers at a table are **instances of that category**
- At any time, $n_k :=$ # number of instance of category k

(Ferguson, 1973; Aldous, 1985; Pitman, 1995)

The Chinese Restaurant process

- The metaphor: an unbounded Chinese restaurant
  - An unbounded number of tables
  - Each table can seat an unbounded number of customers



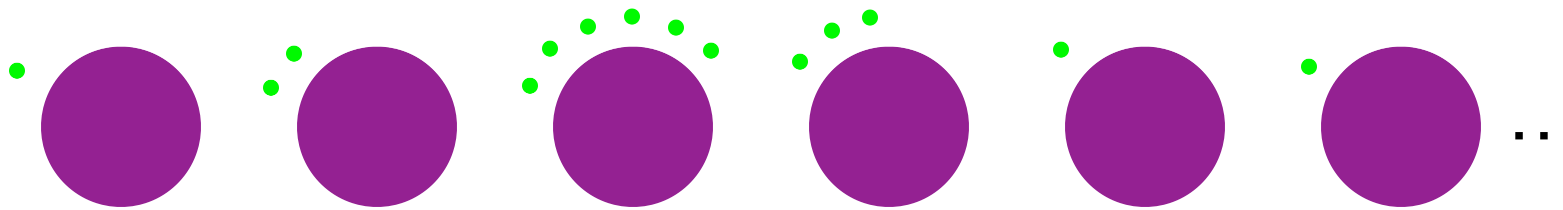
$n_1 = 1$ $n_2 = 2$ $n_3 = 6$ $n_4 = 3$ $n_5 = 1$ $n_6 = 1$

- For our learning problem, the tables are **categories**, and the customers at a table are **instances of that category**
- At any time, $n_k :=$ # number of instance of category k
- Call the category of the i-th customer z_i

(Ferguson, 1973; Aldous, 1985; Pitman, 1995)

The Chinese Restaurant process

- The metaphor: an unbounded Chinese restaurant
  - An unbounded number of tables
  - Each table can seat an unbounded number of customers



$n_1 = 1$ $n_2 = 2$ $n_3 = 6$ $n_4 = 3$ $n_5 = 1$ $n_6 = 1$

- For our learning problem, the tables are **categories**, and the customers at a table are **instances of that category**
- At any time, $n_k :=$ # number of instance of category k
- Call the category of the i-th customer z_i
- Probability of the *next* instance's category:

$$P(z_i | z_{1 \dots i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{for \textbf{old} categories with at least one instance} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{that } z_i \text{ will be a \textbf{new} category} \end{cases}$$

(Ferguson, 1973; Aldous, 1985; Pitman, 1995)

The base distribution for word forms

The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies

The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0

The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w

Probability of word ending (≡geometric distr. parameter)

The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w

Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process

The base distribution for word forms

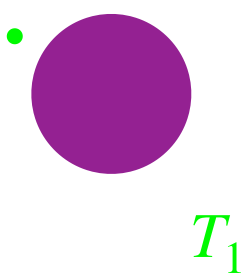
- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w

Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



The base distribution for word forms

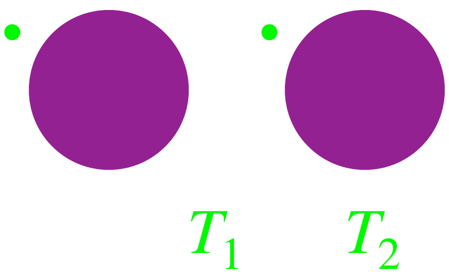
- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w

Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



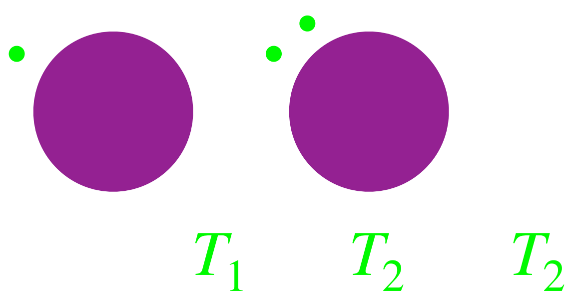
The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



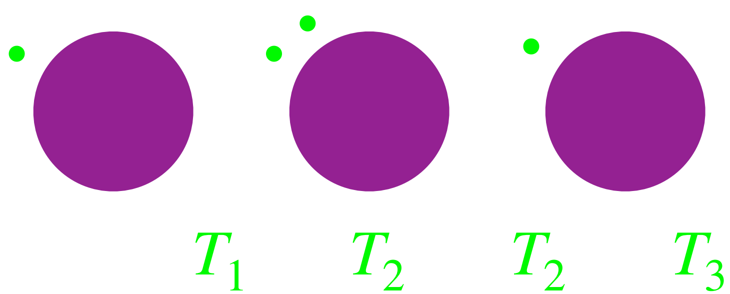
The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



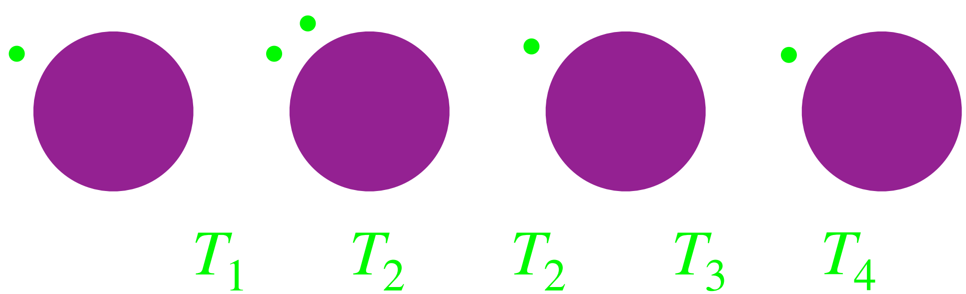
The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



The base distribution for word forms

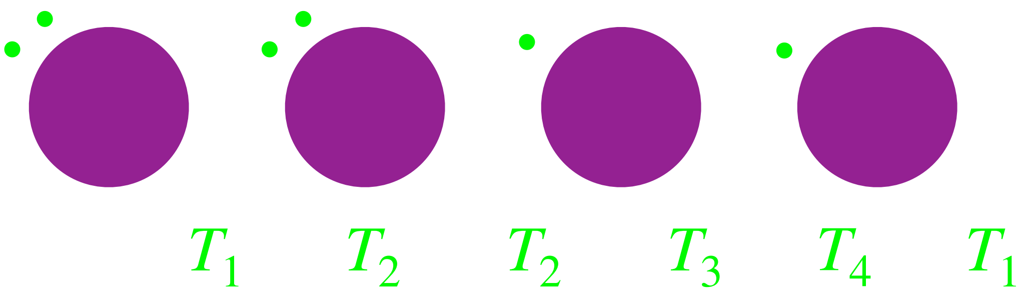
- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w

Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



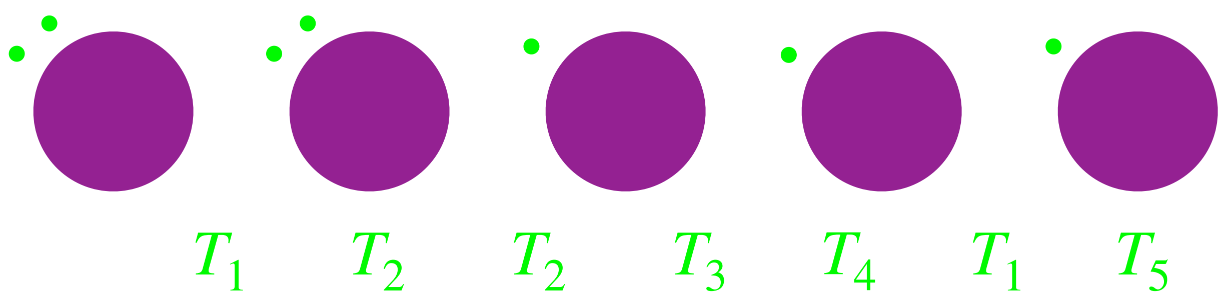
The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



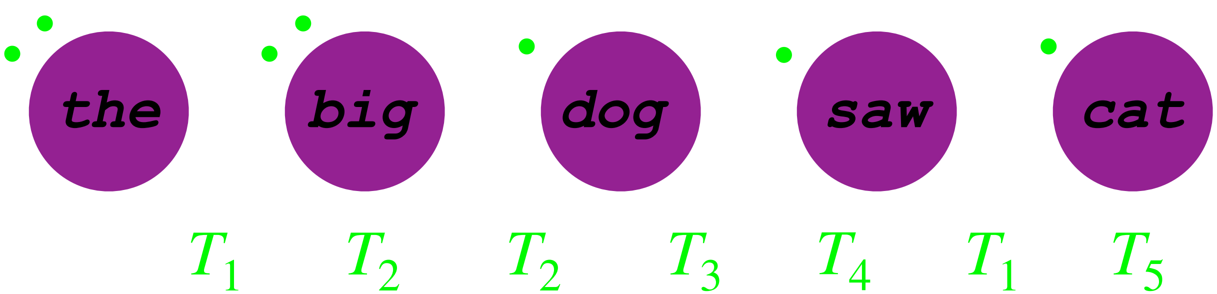
The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



The base distribution for word forms

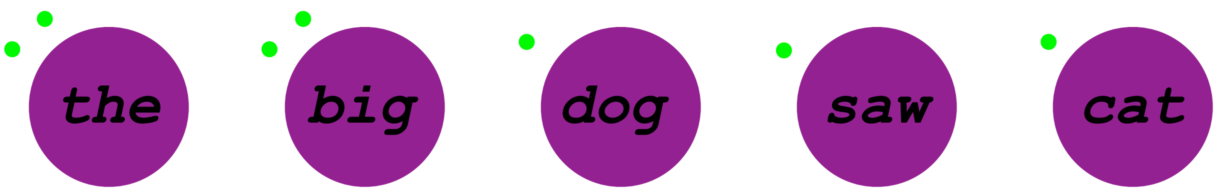
- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w

Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



the big big dog saw the cat

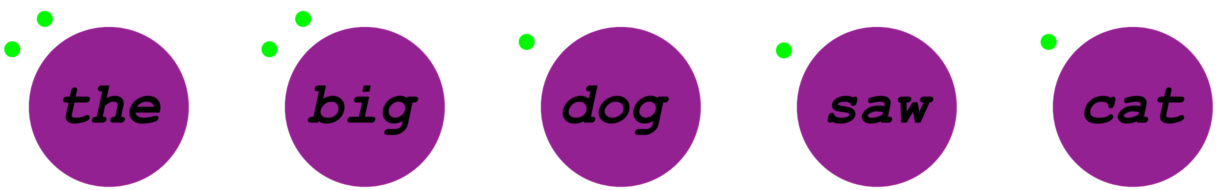
The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



the big big dog saw the cat

- These two stages comprise a DIRICHLET PROCESS

$$w_i \mid G \sim G$$

$$G \mid \alpha_0, P_0 \sim \mathrm{DP}(\alpha_0, P_0)$$

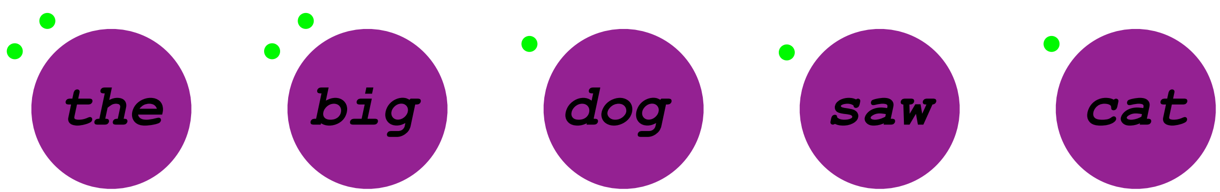
The base distribution for word forms

- Chinese Restaurant Process→distr. of word frequencies
- For distribution over word forms, a BASE MEASURE P_0
- Simplest such model: word lengths geometrically distributed, word forms of a given length uniformly distr.

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Characters of w Probability of word ending (≡geometric distr. parameter)

- Yields a two-stage view of the generative process



the big big dog saw the cat

- These two stages comprise a DIRICHLET PROCESS

Nonparametric Bayesian model

$$w_i \,|\, G \sim G$$

$$G \,|\, \alpha_0, P_0 \sim \mathrm{DP}(\alpha_0, P_0)$$

Unigram model word segmentation model

- The complete story of how a corpus comes into being…
- First, a probability distribution over corpus length *N*:

$$P(N) = (1 - p_{C_\#})^N p_{C_\#}$$

- Next, a probability distribution for the type identity of each new word:

$$P(z_i = k | z_{1\ldots i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{if } k \text{ is an old word type} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{if } k \text{ is a new word type} \end{cases}$$

- Finally, a probability distribution over the phonological form of each word type

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Inference via Gibbs Sampling

Inference via Gibbs Sampling

- Recall: Gibbs Sampling is a Markov-Chain Monte Carlo method for sampling from a joint distr. on $X = x_1, \ldots, x_n$

Inference via Gibbs Sampling

- Recall: Gibbs Sampling is a Markov-Chain Monte Carlo method for sampling from a joint distr. on $X = x_1, \ldots, x_n$
- Let each $\{x_i\}$ be a binary indicator of the presence/absence of a word boundary at each possible position

```
aɪ.si.ə.da.gi
pɛt.ðə.kɪ.ti
si.ðə.bal
aɪ.ləv.ju
du ju.si.ðə.da.gi
```

Inference via Gibbs Sampling

- Recall: Gibbs Sampling is a Markov-Chain Monte Carlo method for sampling from a joint distr. on $X = x_1, \ldots, x_n$
- Let each $\{x_i\}$ be a binary indicator of the presence/absence of a word boundary at each possible position

```
aɪ|si|ə da gi
pɛt ðə|kɪ ti
si|ðə|bal
aɪ|ləv ju
du ju|si|ðə|da gi
```

Inference via Gibbs Sampling

- Recall: Gibbs Sampling is a Markov-Chain Monte Carlo method for sampling from a joint distr. on $X = x_1, \ldots, x_n$
- Let each $\{x_i\}$ be a binary indicator of the presence/ absence of a word boundary at each possible position

```
aɪ|si|ə da gi
pɛt|ðə|kɪ ti
si|ðə|bal
aɪ|ləv ju
du ju|si|ðə|da gi
```

Inference via Gibbs Sampling

- Recall: Gibbs Sampling is a Markov-Chain Monte Carlo method for sampling from a joint distr. on $X = x_1, \ldots, x_n$
- Let each $\{x_i\}$ be a binary indicator of the presence/absence of a word boundary at each possible position

aɪ|si|ə da gi
pɛt|ðə|kɪ ti
si|ðə|bal
aɪ|ləv ju
du ju|si|ðə|da gi

$$\frac{P(x_i = 1 \,|\, X_{-i})}{P(x_i = 0 \,|\, X_{-i})} = \ldots$$

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

$$P(z_i = k | z_{1 \ldots i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{if } k \text{ is an old word type} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{if } k \text{ is a new word type} \end{cases}$$

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Inference via Gibbs Sampling

- Recall: Gibbs Sampling is a Markov-Chain Monte Carlo method for sampling from a joint distr. on $X = x_1, \ldots, x_n$
- Let each $\{x_i\}$ be a binary indicator of the presence/ absence of a word boundary at each possible position

aɪ|si|ə da gi
pɛt|ðə|kɪ ti
si|ðə|bal
aɪ|ləv ju
du ju|si|ðə|da gi

$$\frac{P(x_i = 1 \mid X_{-i})}{P(x_i = 0 \mid X_{-i})} = \ldots$$

Inserting this word boundary...

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

$$P(z_i = k | z_{1\ldots i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{if } k \text{ is an old word type} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{if } k \text{ is a new word type} \end{cases}$$

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Inference via Gibbs Sampling

- Recall: Gibbs Sampling is a Markov-Chain Monte Carlo method for sampling from a joint distr. on $X = x_1, \ldots, x_n$
- Let each $\{x_i\}$ be a binary indicator of the presence/absence of a word boundary at each possible position

```
aɪ|si|ə da gi
pɛt|ðə|kɪ ti
si|ðə|bal
aɪ|ləv ju
du ju|si|ðə|da gi
```

$$\frac{P(x_i = 1 \,|\, X_{-i})}{P(x_i = 0 \,|\, X_{-i})} = \ldots$$

Inserting this word boundary...

- Increases corpus length

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

$$P(z_i = k | z_{1 \ldots i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{if } k \text{ is an old word type} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{if } k \text{ is a new word type} \end{cases}$$

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Inference via Gibbs Sampling

- Recall: Gibbs Sampling is a Markov-Chain Monte Carlo method for sampling from a joint distr. on $X = x_1, \ldots, x_n$
- Let each $\{x_i\}$ be a binary indicator of the presence/absence of a word boundary at each possible position

```
aɪ|si|ə da gi
pɛt|ðə|kɪ ti
si|ðə|bal
aɪ|ləv ju
du ju|si|ðə|da gi
```

$$\frac{P(x_i = 1 \,|\, X_{-i})}{P(x_i = 0 \,|\, X_{-i})} = \ldots$$

Inserting this word boundary...

- Increases corpus length
- removes **pɛtðə** from lexicon and adds **pɛt**

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

$$P(z_i = k | z_{1\ldots i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{if } k \text{ is an old word type} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{if } k \text{ is a new word type} \end{cases}$$

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Inference via Gibbs Sampling

- Recall: Gibbs Sampling is a Markov-Chain Monte Carlo method for sampling from a joint distr. on $X = x_1, \ldots, x_n$
- Let each $\{x_i\}$ be a binary indicator of the presence/absence of a word boundary at each possible position

aɪ|si|ə da gi
pɛt|ðə|kɪ ti
si|ðə|bal
aɪ|ləv ju
du ju|si|ðə|da gi

$$\frac{P(x_i = 1 \mid X_{-i})}{P(x_i = 0 \mid X_{-i})} = \ldots$$

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

$$P(z_i = k \mid z_{1 \ldots i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{if } k \text{ is an old word type} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{if } k \text{ is a new word type} \end{cases}$$

$$P_0(w = x_1 \ldots x_n) \propto p_\#(1 - p_\#)^n$$

Inserting this word boundary...

- Increases corpus length
- removes **pɛtðə** from lexicon and adds **pɛt**
- adds a token of **ðə**

Quantitative results

Precision: $\frac{\#\text{ correct}}{\#\text{ hypothesized by model}}$

Recall: $\frac{\#\text{ correct}}{\#\text{ present in ground truth}}$

	Boundaries		Word tokens		Lexical items	
	Prec	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**
Venkataraman (2001)	80.6	**84.8**	**67.7**	**70.2**	52.9	51.3
Brent (1999)	80.3	84.3	67.0	69.4	53.6	51.3
GGJ unigram model	**92.4**	62.2	61.9	47.6	**57.0**	**57.5**

- This model is very precise in the word boundaries it proposes, and does best in lexicon recovery
- But it seems to undersegment...

Example learned segmentations

```
youwant to see thebook
look theres aboy with his hat
and adoggie
you wantto lookatthis
lookatthis
havea drink
okay now
whatsthis
whatsthat
whatisit
look canyou take itout
...
```

(Goldwater, Griffiths, & Johnson, 2006, 2009)

Example learned segmentations

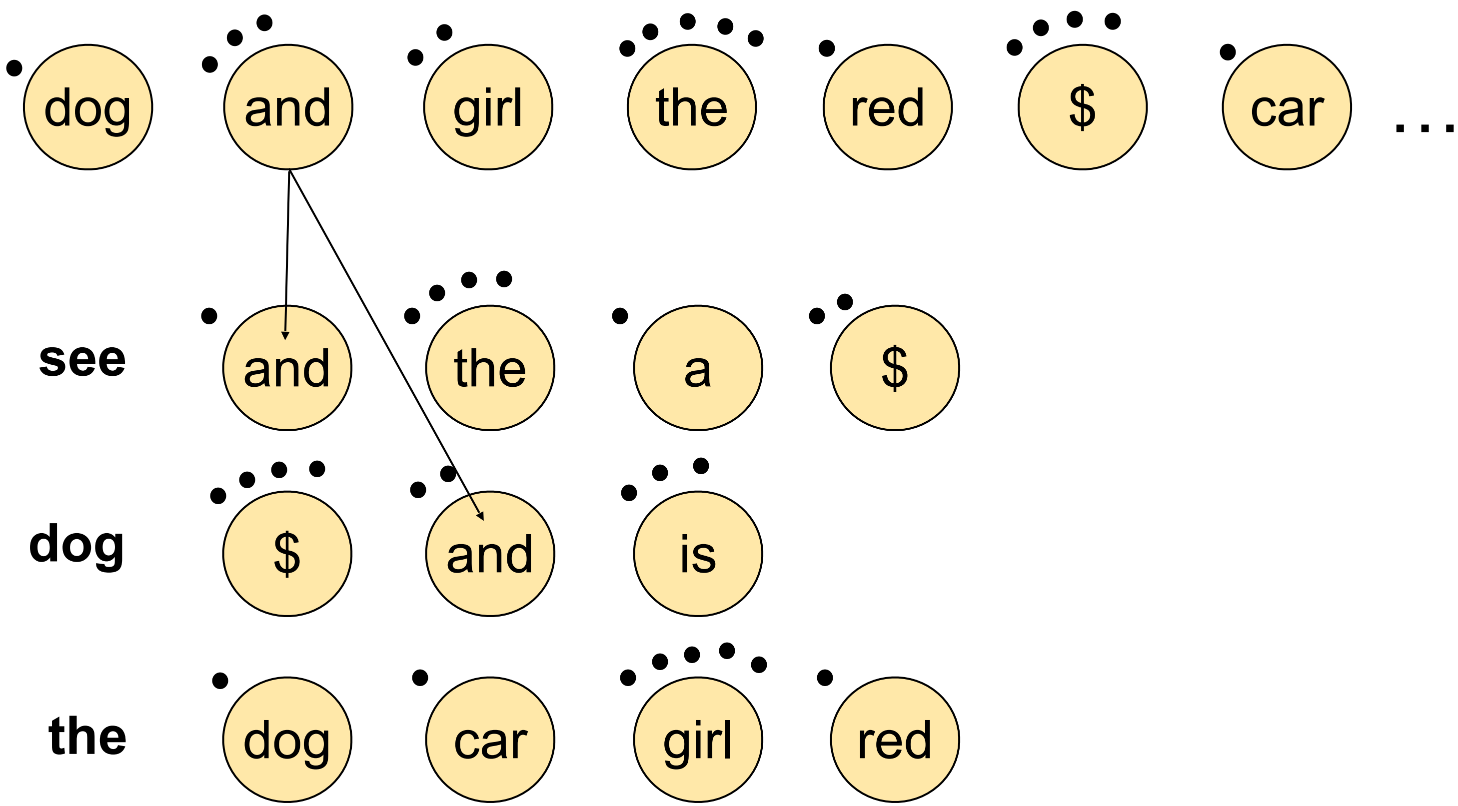
```
youwant to see thebook
look theres aboy with his hat
and adoggie
you wantto lookatthis
lookatthis
havea drink
okay now
whatsthis
whatsthat
whatisit
look canyou take itout
...
```

- ...but some of the under-segmentations seem plausible!

(Goldwater, Griffiths, & Johnson, 2006, 2009)

Bigram model with a hierarchical Dirichlet Process

1. Generate G, a distribution over words, using $\mathrm{DP}(\alpha_0, P_0)$.
2. For each word in the data, generate a distribution over the words that follow it, using $\mathrm{DP}(\alpha_1, G)$.



[Goldwater, Griffiths, & Johnson, 2006, 2009; visualization courtesy of Sharon Goldwater]

Bigram model quantitative performance

	Boundaries		Word tokens		Lexical items	
	Prec	Rec	Prec	Rec	Prec	Rec
Venkataraman (2001)	80.6	**84.8**	67.7	**70.2**	52.9	51.3
Brent (1999)	80.3	84.3	67.0	69.4	53.6	51.3
GGJ unigram model	**92.4**	62.2	61.9	47.6	57.0	**57.5**
GGJ bigram model	90.3	80.8	**75.2**	69.6	**63.5**	**55.2**

Summary: unsupervised word segmentation

- Local transition statistics of phonemes & syllables provide rich cues for word boundaries, even without meaning
- Unsupervised word segmentation as latent variable (word boundary) inference in a simple generative model
- Fitting via corpus maximum likelihood intrinsically problematic: need a bias toward "sensible looking words"
- Bayesian inference naturally offers a principled trade-off
- Non-parametric Bayesian models allow unsupervised learning with unbounded numbers of categories—learning new words or structures as new data are encountered

A new model

- Assume words $\mathbf{w} = \{w_1 \ldots w_n\}$ are generated as follows:

$$\begin{aligned} w_i \mid G &\sim G \\ G \mid \alpha_0, P_0 &\sim \mathrm{DP}(\alpha_0, P_0) \end{aligned}$$

  - G: is analogous to θ in BHMM, but with infinite dimension. As with θ, we integrate it out.
  - $\mathrm{DP}(\alpha_0, P_0)$: a Dirichlet process with concentration parameter α_0 and base distribution P_0.

[slide courtesy of Sharon Goldwater]

The Dirichlet distribution

- First: recap on Dirichlet distribution (generalization of beta distribution) in order to get to the Dirichlet process
- The *k*-class Dirichlet distribution is a probability distribution over *k*-class multinomial distributions with parameters α_i

$$\mathcal{D}(\pi_1, \ldots, \pi_k) \stackrel{\text{def}}{=} \frac{1}{Z} \pi_1^{\alpha_1 - 1} \pi_2^{\alpha_2 - 1} \ldots \pi_k^{\alpha_k - 1}$$

- The normalizing constant *Z* is

$$Z = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_k)}{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}$$

- Symmetric Dirichlet: all α_i are set to the same value α

[slide courtesy of Sharon Goldwater]

The Dirichlet process

- Our generative model for word segmentation assumes data (words!) arise from clusters. Defines
  - A distribution over the number and size of the clusters.
  - A distribution P_0 over the parameters describing the distribution of data in each cluster.
- Clusters = frequencies of different words.
- Cluster parameters = identities of different words.

The Chinese restaurant process

- In the DP, the number of items in each cluster is defined by the Chinese restaurant process:
  - Restaurant has an infinite number of tables, with infinite seating capacity.
  - The table chosen by the ith customer, z_i, depends on the seating arrangement of the previous $i-1$ customers :

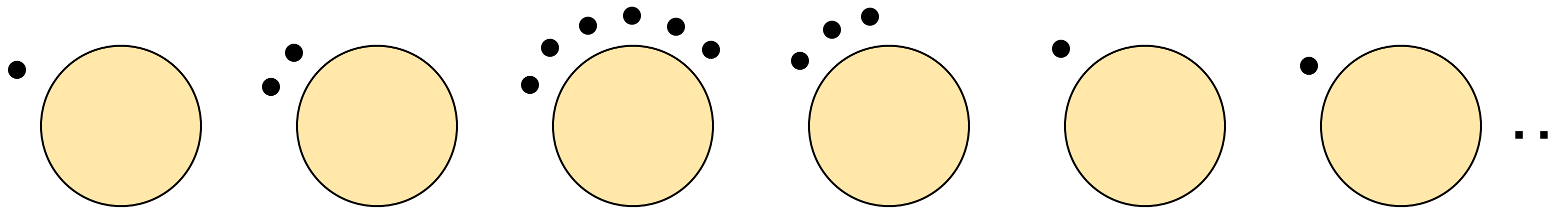
$$P(z_i = k | \mathbf{z}_{-i}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{if the } k\text{-th table is occupied} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{if } k \text{ is the next unoccupied table} \end{cases}$$

- CRP produces a power-law distribution over cluster sizes.

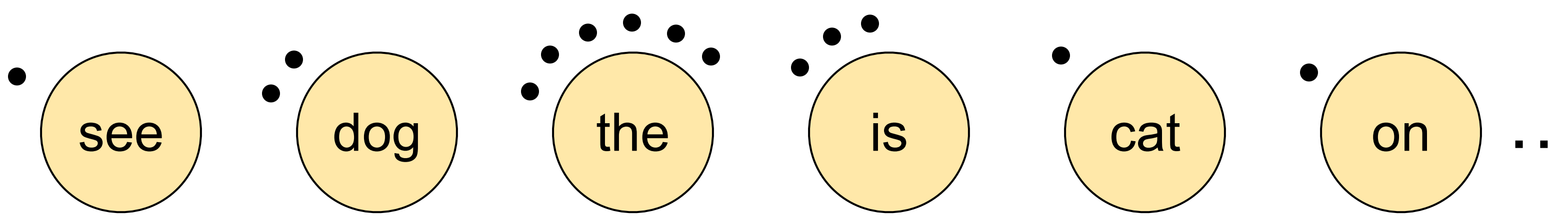
[slide courtesy of Sharon Goldwater]

The two-stage restaurant

1. Assign data points to clusters (tables).

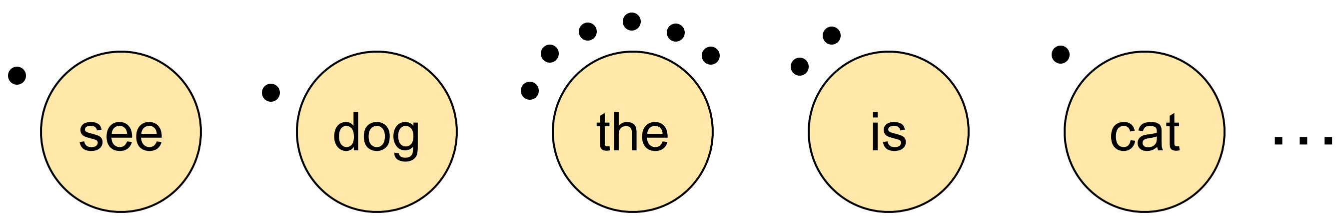


2. Sample labels for tables using P_0.



Alternative view

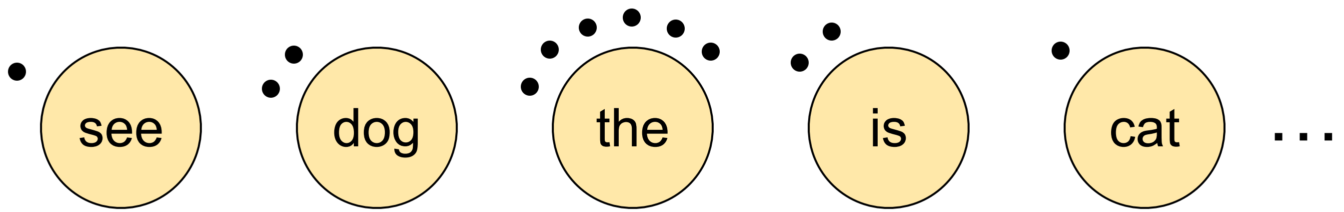
- Equivalently, words are generated sequentially using a cache model: previously generated words are more likely to be generated again.



… the

Alternative view

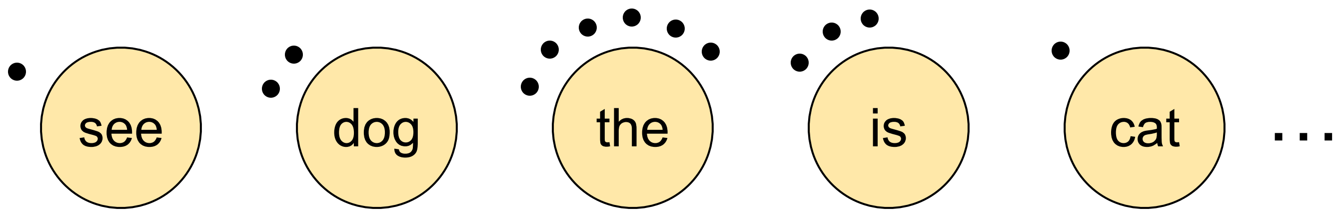
- Equivalently, words are generated sequentially using a cache model: previously generated words are more likely to be generated again.



… the dog

Alternative view

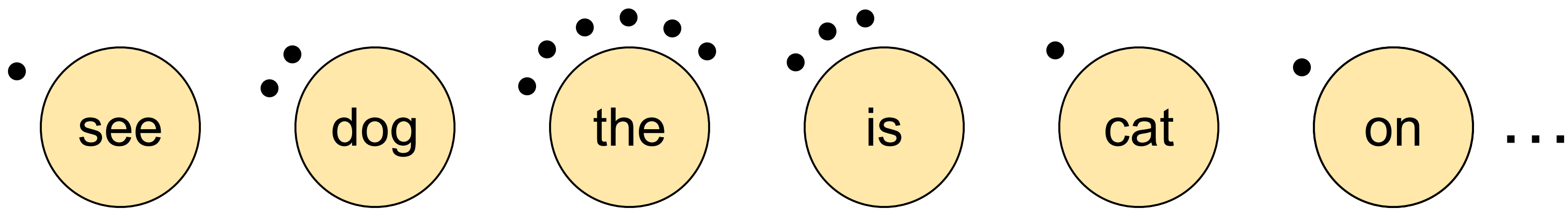
- Equivalently, words are generated sequentially using a cache model: previously generated words are more likely to be generated again.



… the dog is

Alternative view

- Equivalently, words are generated sequentially using a cache model: previously generated words are more likely to be generated again.



… the dog is on

Unigram model

- DP model yields the following distribution over words:

$$P(w_i = w \mid \mathbf{w}_{-i}) = \frac{n_w + \alpha P_0(w)}{i - 1 + \alpha}$$

with $P_0(w = x_1 \dots x_m) = \prod_{i=1}^{m} P(x_i)$ for characters $x_1 \dots x_m$.

  - P_0 favors shorter lexical items.
  - Words are not independent, but are exchangeable: a unigram model: $P(w_1, w_2, w_3, w_4) = P(w_2, w_4, w_1, w_3)$

- Input corpus contains utterance boundaries. We assume a geometric distribution on utterance lengths.

[slide courtesy of Sharon Goldwater]

Unigram model, in more detail

- How a corpus comes into being…
- First, a probability distribution over corpus length *N*:

$$P(N = n) = (1 - p_{C_\#})^n p_{C_\#}$$

- Next, a probability distribution for the type identity of each new word:

$$P(z_i = k | z_{1 \ldots i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha_0} & \text{if } k \text{ is an old word type} \\ \frac{\alpha_0}{i-1+\alpha_0} & \text{if } k \text{ is a new word type} \end{cases}$$

- Finally, a probability distribution over the phonological form of each word type

$$P(form(k) = w_k) = \frac{1}{V^{length(w_k)+1}}$$

Advantages of DP language models

- Solutions are sparse, yet grow as data size grows.
  - Smaller values α_0 of lead to fewer lexical items generated by $P_{0.}$
- Models lexical items separately from frequencies.
  - Different choices for P_0 can infer different kinds of linguistic structure.
- Amenable to standard search procedures (e.g., Gibbs sampling).

[slide courtesy of Sharon Goldwater]

Gibbs sampling

- Compare pairs of hypotheses differing by a single word boundary:

```
whats.that
the.doggie
yeah
wheres.the.doggie
…
```

```
whats.that
the.dog.gie
yeah
wheres.the.doggie
…
```

- Calculate the probabilities of the words that differ, given current analysis of all other words.
- Sample a hypothesis according to the ratio of probabilities.

[slide courtesy of Sharon Goldwater]

Experiments

- Input: same corpus as Brent (1999), Venkataraman (2001).
  - 9790 utterances of transcribed child-directed speech.
  - Example input:

```
youwanttoseethebook
looktheresaboywithhishat
andadoggie
youwanttolookatthis
...
```

- Using different values of α_0, evaluate on a single sample after 20k iterations.

Example results

```
youwant to see thebook
look theres aboy with his hat
and adoggie
you wantto lookatthis
lookatthis
havea drink
okay now
whatsthis
whatsthat
whatisit
look canyou take itout
...
```

Quantitative evaluation

- Proposed boundaries are more accurate than other models, but fewer proposals are made.
- Result: lower accuracy on words.

	Boundaries		Word tokens	
	Prec	**Rec**	**Prec**	**Rec**
Venk. (2001)	80.6	**84.8**	**67.7**	**70.2**
Brent (1999)	80.3	84.3	67.0	69.4
DP model	**92.4**	62.2	61.9	47.6

Precision: #correct / #found

Recall: #found / #true

[slide courtesy of Sharon Goldwater]

What happened?

- DP model assumes (falsely) that words have the same probability regardless of context.

$$P(\texttt{that}) = .024 \qquad P(\texttt{that}|\texttt{whats}) = .46 \qquad P(\texttt{that}|\texttt{to}) = .0019$$

- Positing collocations allows the model to capture word-to-word dependencies.

[slide courtesy of Sharon Goldwater]

What about other unigram models?

- Venkataraman's system is based on MLE, so we know results are due to constraints imposed by search.
- Brent's search algorithm also yields non-optimal solution.
  - Our solution has higher probability under his model than his own solution does.
  - On randomly permuted corpus, our system achieves 96% accuracy; Brent gets 81%.
- Any (reasonable) unigram model will undersegment.
- Bottom line: previous results were accidental properties of search, not systematic properties of models.

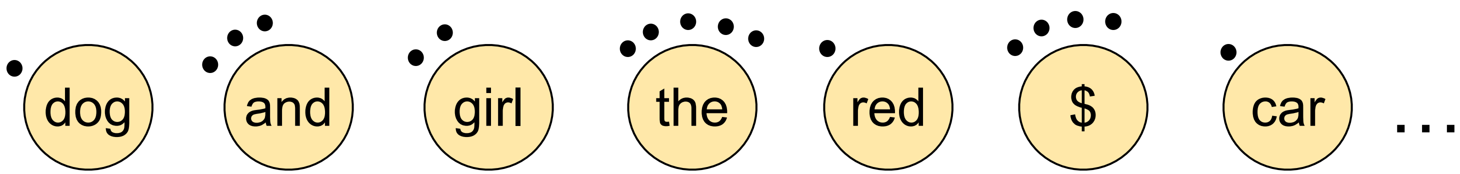
[slide courtesy of Sharon Goldwater]

Improving the model

- By incorporating context (using a bigram model), perhaps we can improve segmentation…

Hierachical Dirichlet process

1. Generate G, a distribution over words, using $\mathrm{DP}(\alpha_0, P_0)$.



[slide courtesy of Sharon Goldwater]

Example results

```
you want to see the book
look theres a boy with his hat
and a doggie
you want to lookat this
lookat this
have a drink
okay now
whats this
whats that
whatis it
look canyou take it out
...
```

[slide courtesy of Sharon Goldwater]

Quantitative evaluation

- With appropriate choice of α and β,
  - Boundary precision nearly as good as unigram, recall much better.
  - F-score (avg. of prec, rec) on all three measures outperforms all previously published models.

	Boundaries		**Word tokens**		**Lexicon**	**Prec**
	Prec	**Rec**	**Prec**	**Rec**	**Rec**	
DP (unigram)	**92.4**	62.2	61.9	47.6	57.0	**57.5**
Venk. (bigram)	81.7	82.5	68.1	68.6	54.5	57.0
HDP (bigram)	89.9	**83.8**	**75.7**	**72.1**	**63.1**	50.3

[slide courtesy of Sharon Goldwater]

Summary: word segmentation

- Our approach to word segmentation using infinite Bayesian models allowed us to
  - Incorporate sensible priors to avoid trivial solutions (à la MLE).
  - Examine the effects of modeling assumptions without limitations from search.
  - Demonstrate the importance of context for word segmentation.
  - Achieve the best published results on this corpus.

[slide courtesy of Sharon Goldwater]