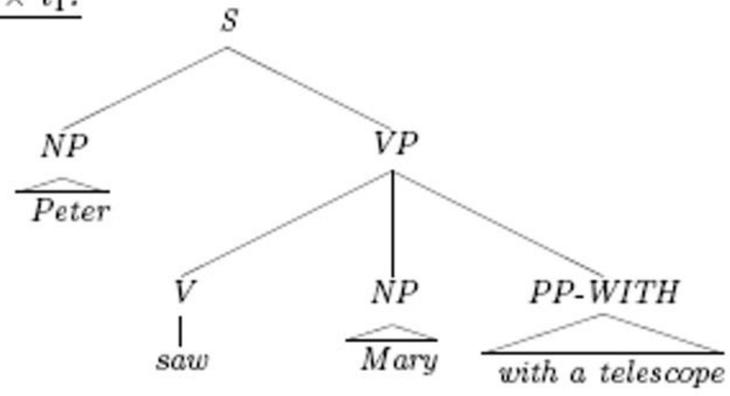
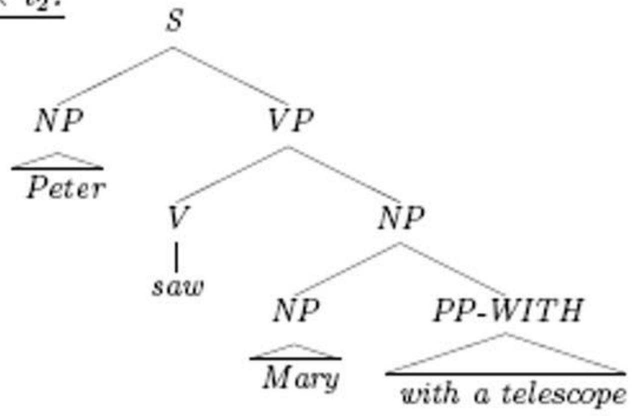
Statistical Parsing: Part 2

An Example

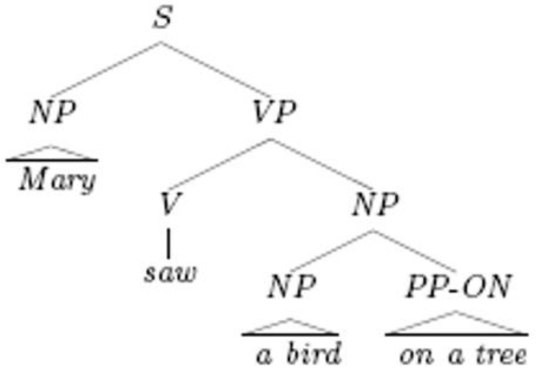
1 $100 \times t_1$:



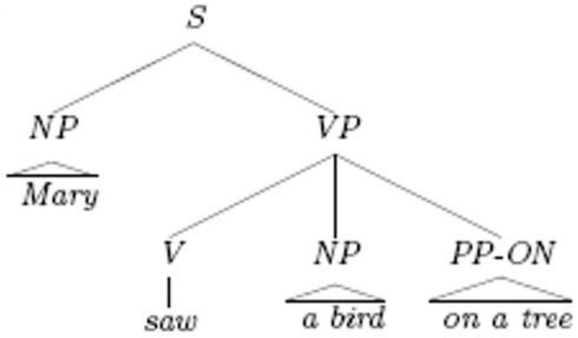
2 $5 \times t_2$:



3 $100 \times t_3$:

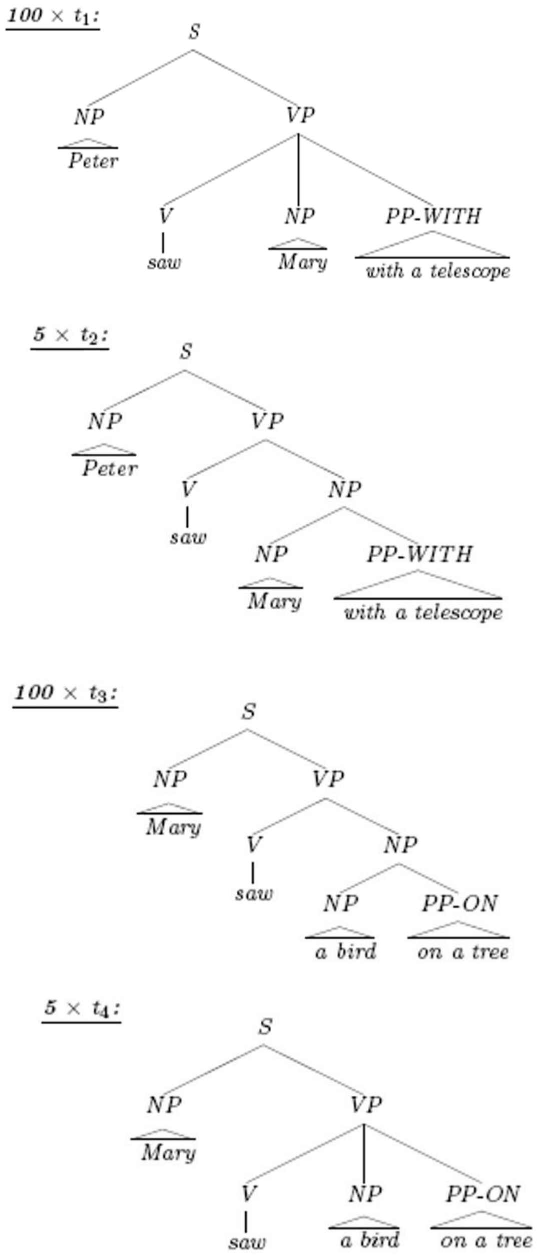


4 $5 \times t_4$:



Refer Workbook

Parameter Estimation: Maximum Likelihood Estimates

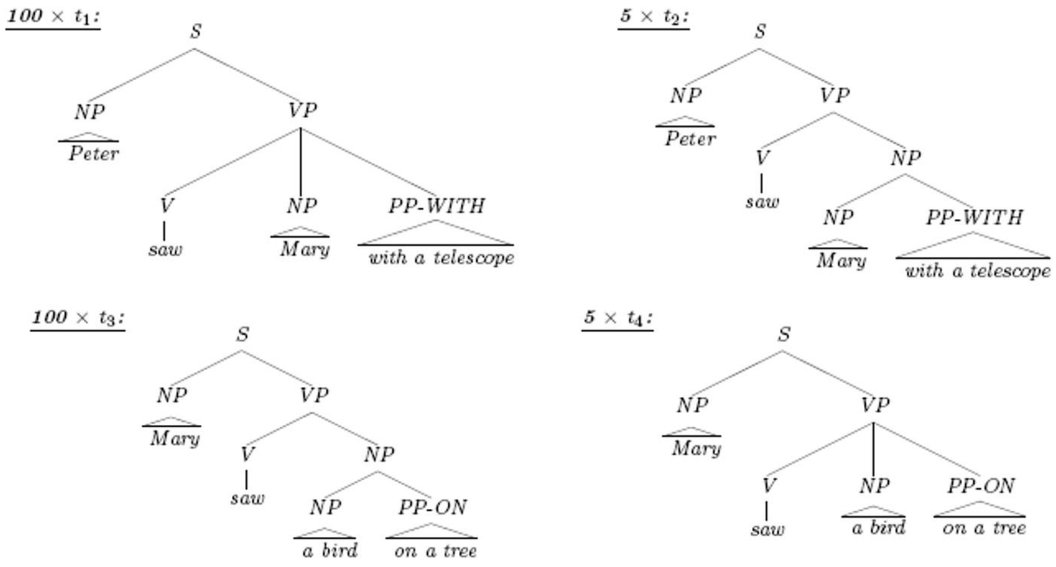


CFG rule	Rule frequency	Rule probability
S $\longrightarrow$ NP VP	100 + 5 + 100 + 5	$\frac{210}{210} = 1.000$
VP $\longrightarrow$ V NP PP-WITH	100	$\frac{100}{210} \approx 0.476$
VP $\longrightarrow$ V NP PP-ON	5	$\frac{5}{210} \approx 0.024$
VP $\longrightarrow$ V NP	5 + 100	$\frac{105}{210} = 0.500$
NP $\longrightarrow$ Peter	100 + 5	$\frac{105}{525} = 0.200$
NP $\longrightarrow$ Mary	100 + 5 + 100 + 5	$\frac{210}{525} = 0.400$
NP $\longrightarrow$ a bird	100 + 5	$\frac{105}{525} = 0.200$
NP $\longrightarrow$ NP PP-WITH	5	$\frac{5}{525} \approx 0.010$
NP $\longrightarrow$ NP PP-ON	100	$\frac{100}{525} \approx 0.190$
PP-WITH $\longrightarrow$ with a telescope	100 + 5	$\frac{105}{105} = 1.000$
PP-ON $\longrightarrow$ on a tree	100 + 5	$\frac{105}{105} = 1.000$
V $\longrightarrow$ saw	100 + 5 + 100 + 5	$\frac{210}{210} = 1.000$

Refer Workbook

Using the PCFG for disambiguation

CFG rule	Rule frequency	Rule probability
S $\longrightarrow$ NP VP	100 + 5 + 100 + 5	$\frac{210}{210} = 1.000$
VP $\longrightarrow$ V NP PP-WITH	100	$\frac{100}{210} \approx 0.476$
VP $\longrightarrow$ V NP PP-ON	5	$\frac{5}{210} \approx 0.024$
VP $\longrightarrow$ V NP	5 + 100	$\frac{105}{210} = 0.500$
NP $\longrightarrow$ Peter	100 + 5	$\frac{105}{525} = 0.200$
NP $\longrightarrow$ Mary	100 + 5 + 100 + 5	$\frac{210}{525} = 0.400$
NP $\longrightarrow$ a bird	100 + 5	$\frac{105}{525} = 0.200$
NP $\longrightarrow$ NP PP-WITH	5	$\frac{5}{525} \approx 0.010$
NP $\longrightarrow$ NP PP-ON	100	$\frac{100}{525} \approx 0.190$
PP-WITH $\longrightarrow$ with a telescope	100 + 5	$\frac{105}{105} = 1.000$
PP-ON $\longrightarrow$ on a tree	100 + 5	$\frac{105}{105} = 1.000$
V $\longrightarrow$ saw	100 + 5 + 100 + 5	$\frac{210}{210} = 1.000$



"Peter saw Mary with a telescope"	$p(\textbf{VP} \longrightarrow \textbf{V NP PP-WITH}) > p(\textbf{VP} \longrightarrow \textbf{V NP}) \cdot p(\textbf{NP} \longrightarrow \textbf{NP PP-WITH})$ $0.476 > 0.500 \cdot 0.010$
"Mary saw a bird on a tree"	$p(\textbf{VP} \longrightarrow \textbf{V NP PP-ON}) < p(\textbf{VP} \longrightarrow \textbf{V NP}) \cdot p(\textbf{NP} \longrightarrow \textbf{NP PP-ON})$ $0.024 < 0.500 \cdot 0.190.$

Try this : Mary saw a bird on a tree with a telescope

Refer Workbook

Incomplete data problem in PCFG parameter estimation

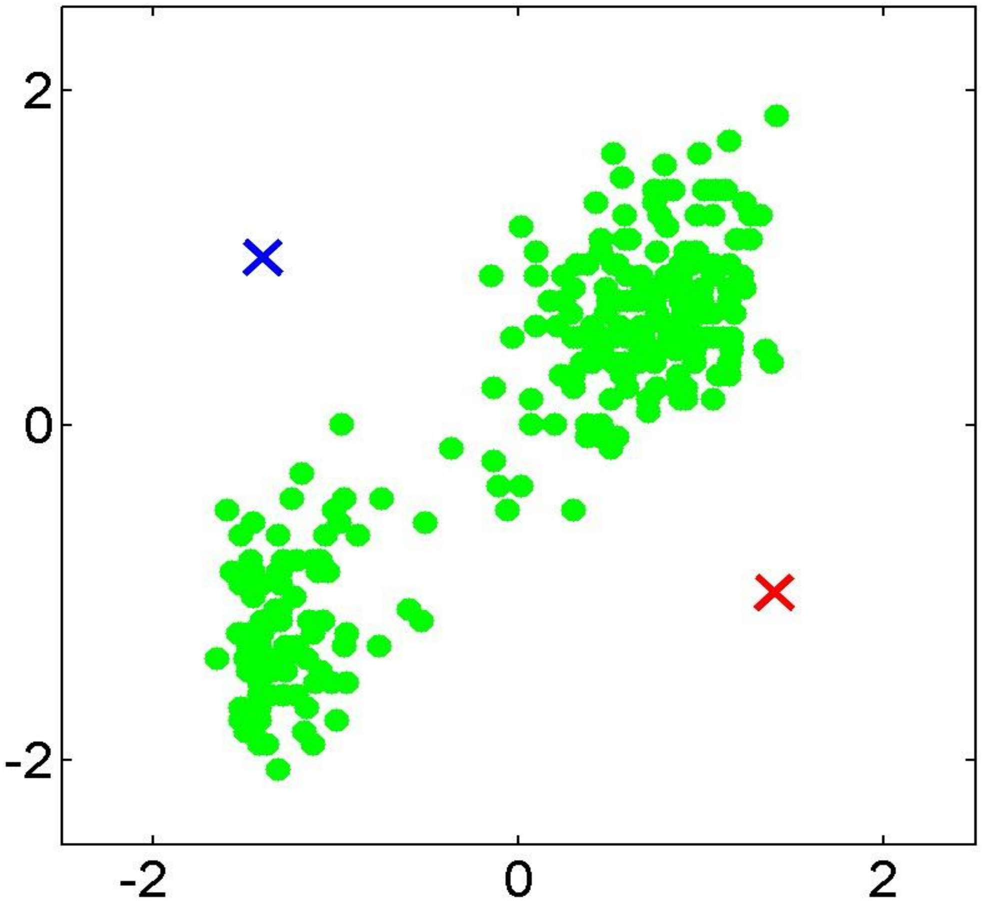
$f(y)$	y
5	y_1
10	$y2$

$$y_1 = \text{“}\textit{Mary saw a bird on a tree}\text{”}$$
$$y_2 = \text{“}\textit{a bird on a tree saw a worm}\text{”}$$

$$
\begin{aligned}
S &\longrightarrow NP\ VP \\
VP &\longrightarrow V\ NP \\
VP &\longrightarrow V\ NP\ PP \\
NP &\longrightarrow NP\ PP \\
NP &\longrightarrow Mary \\
NP &\longrightarrow a\ bird \\
NP &\longrightarrow a\ worm \\
PP &\longrightarrow on\ a\ tree \\
V &\longrightarrow saw
\end{aligned}
$$

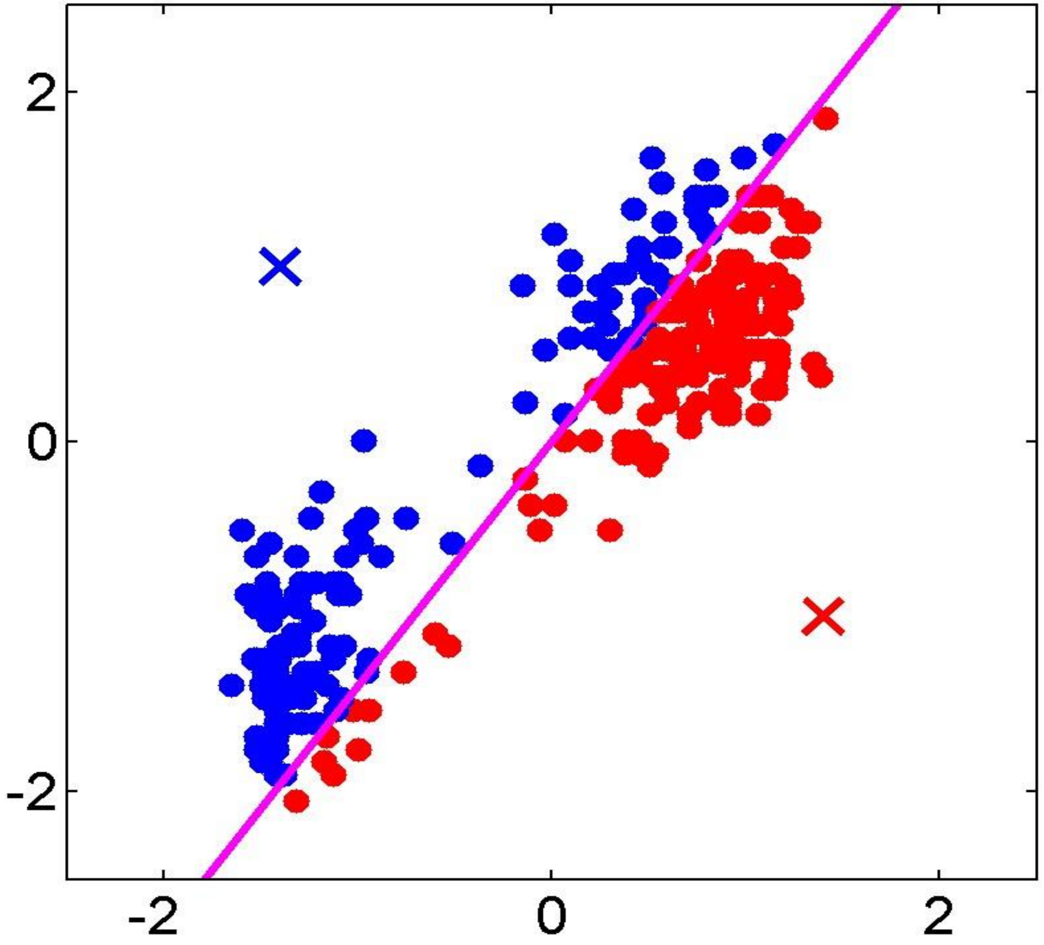
Refer Workbook

K-means clustering: M_0



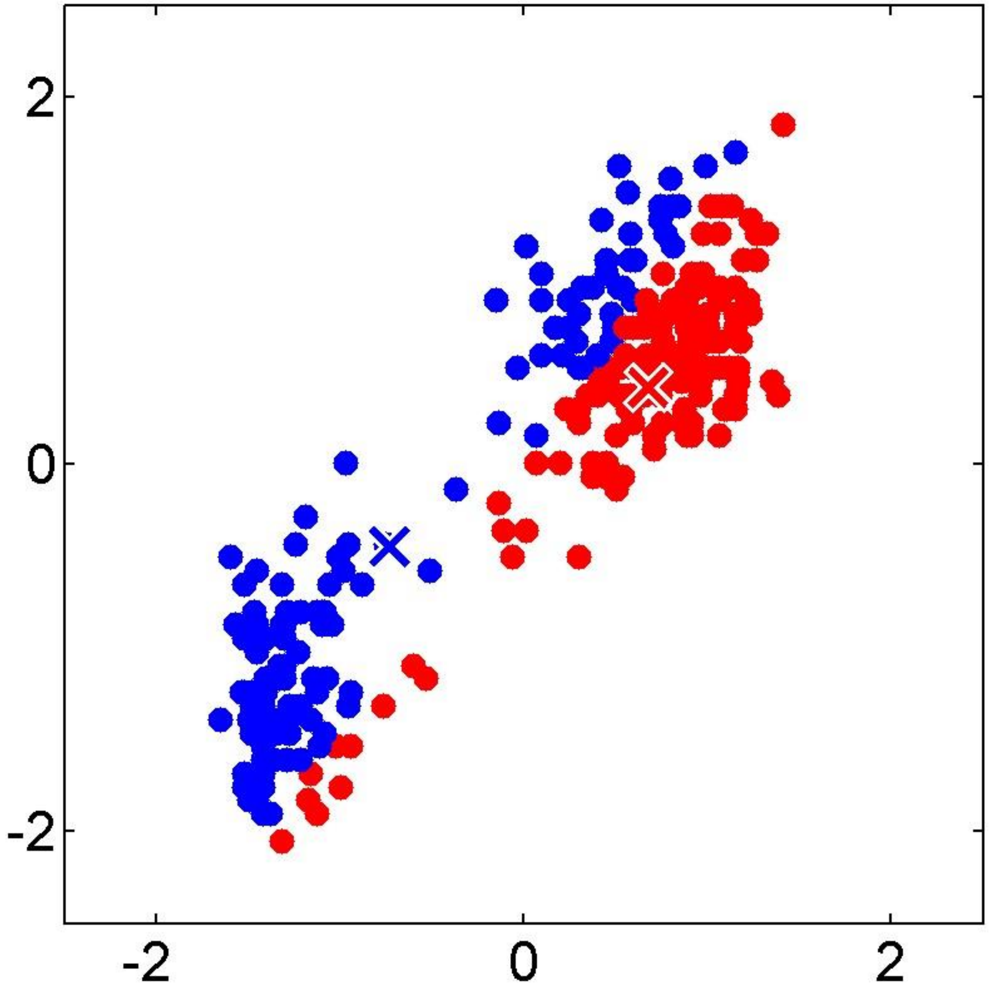
Ack: Chris Bishop

K-means clustering: E_1



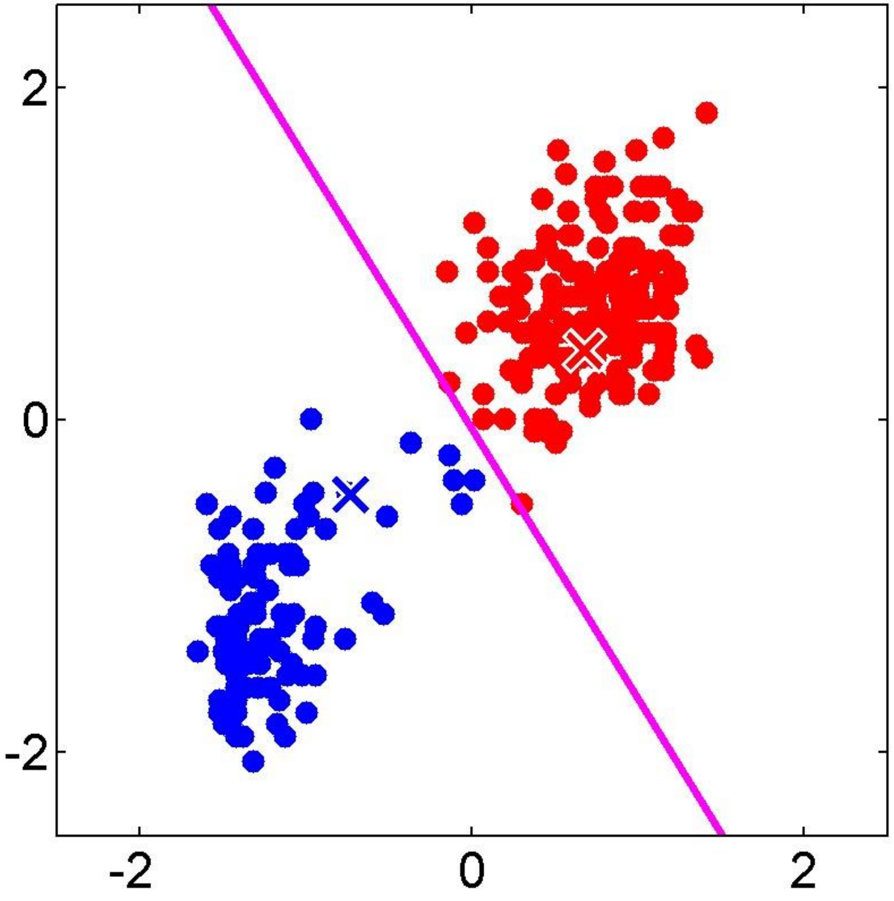
Ack: Chris Bishop

K-means clustering: M_1



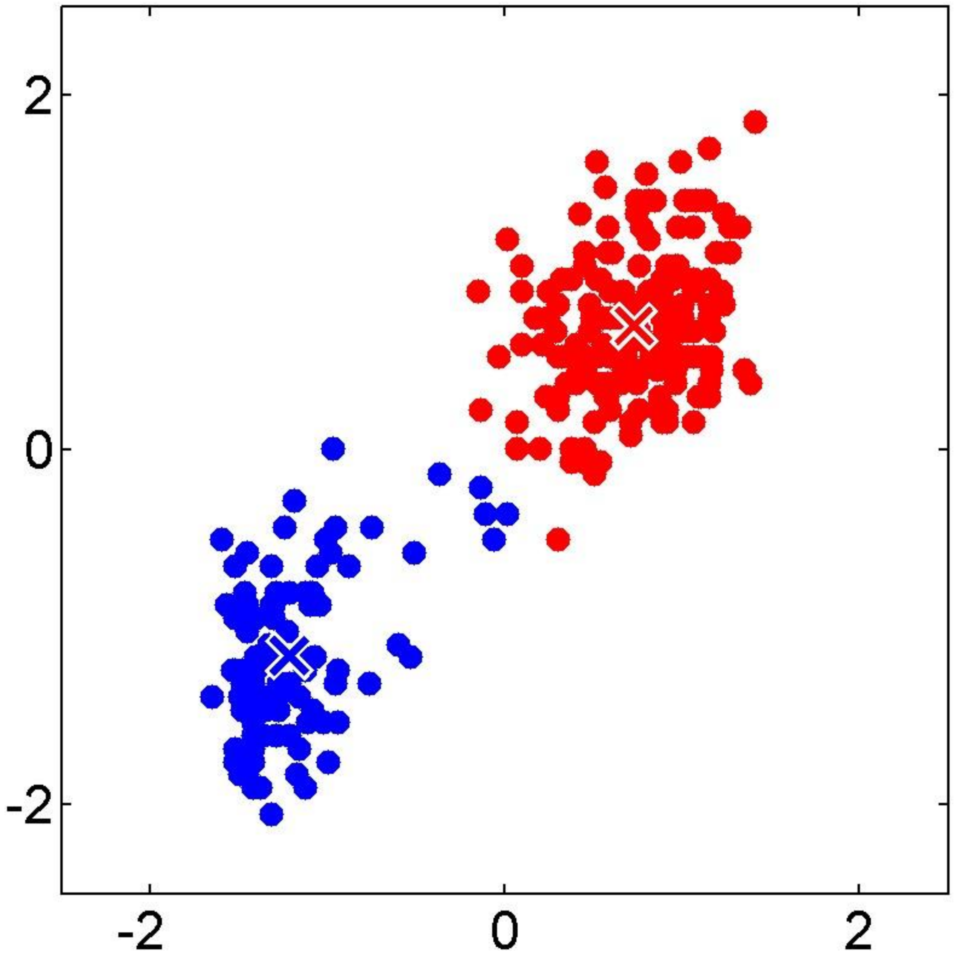
Ack: Chris Bishop

K-means clustering: E_2



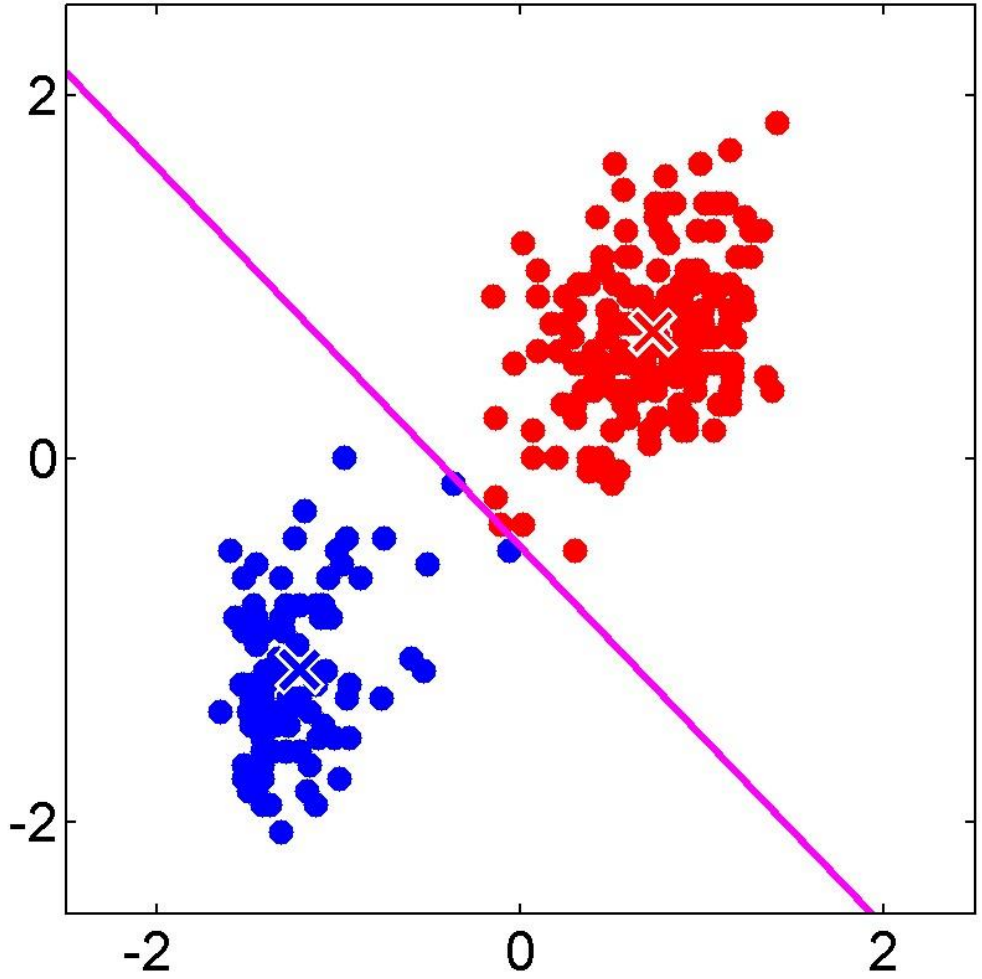
Ack: Chris Bishop

K-means clustering: M_2



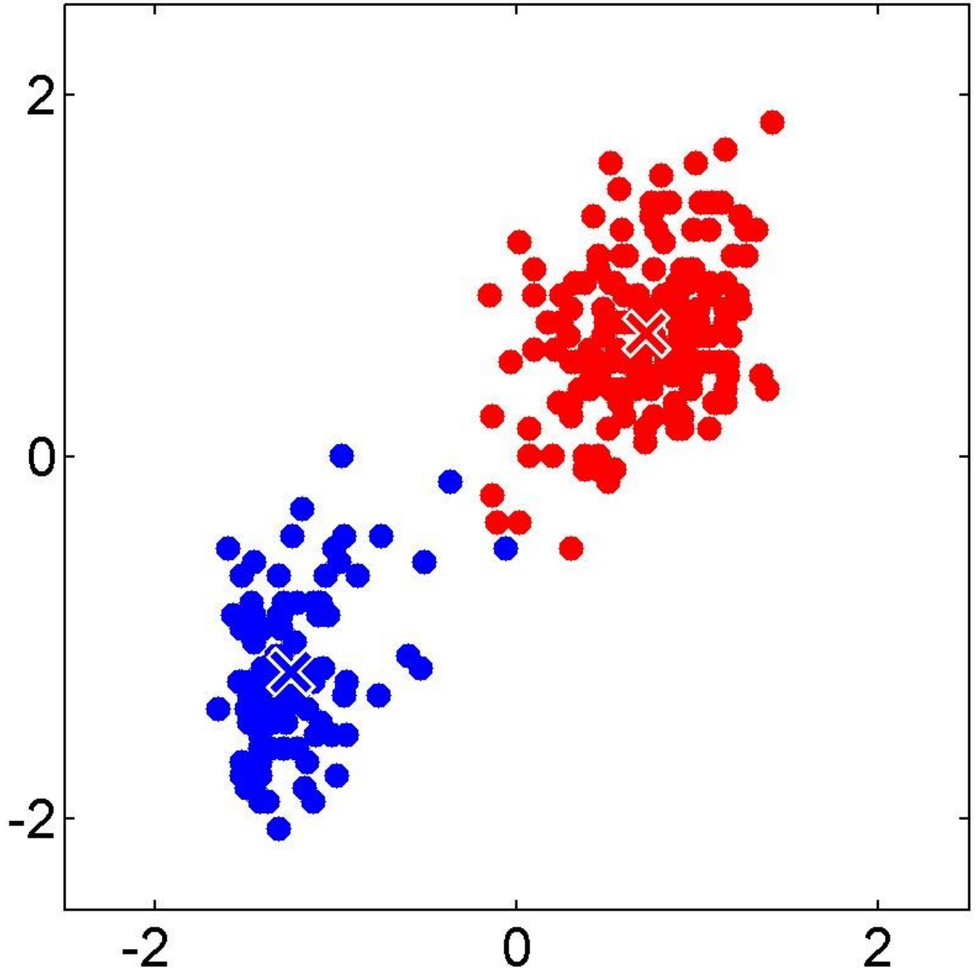
Ack: Chris Bishop

K-means clustering: E_3



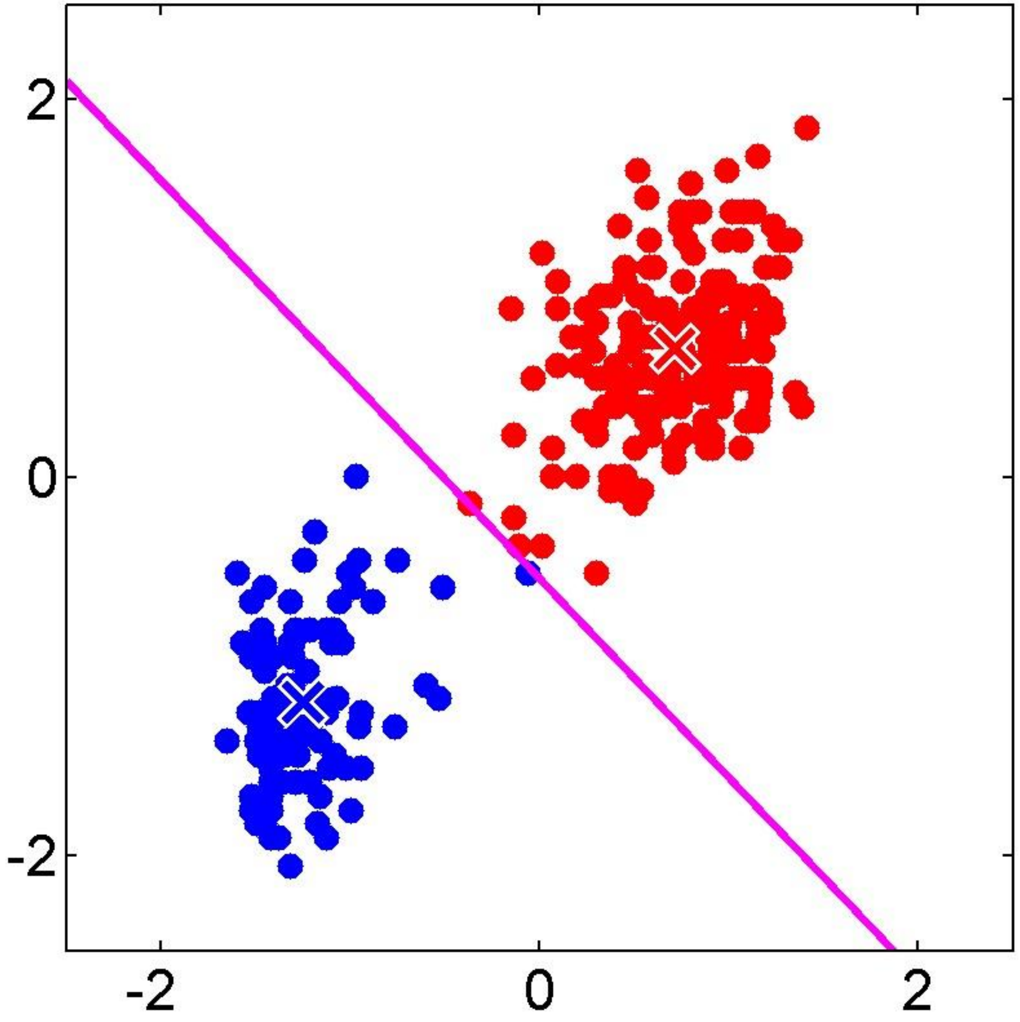
Ack: Chris Bishop

K-means clustering: M_3



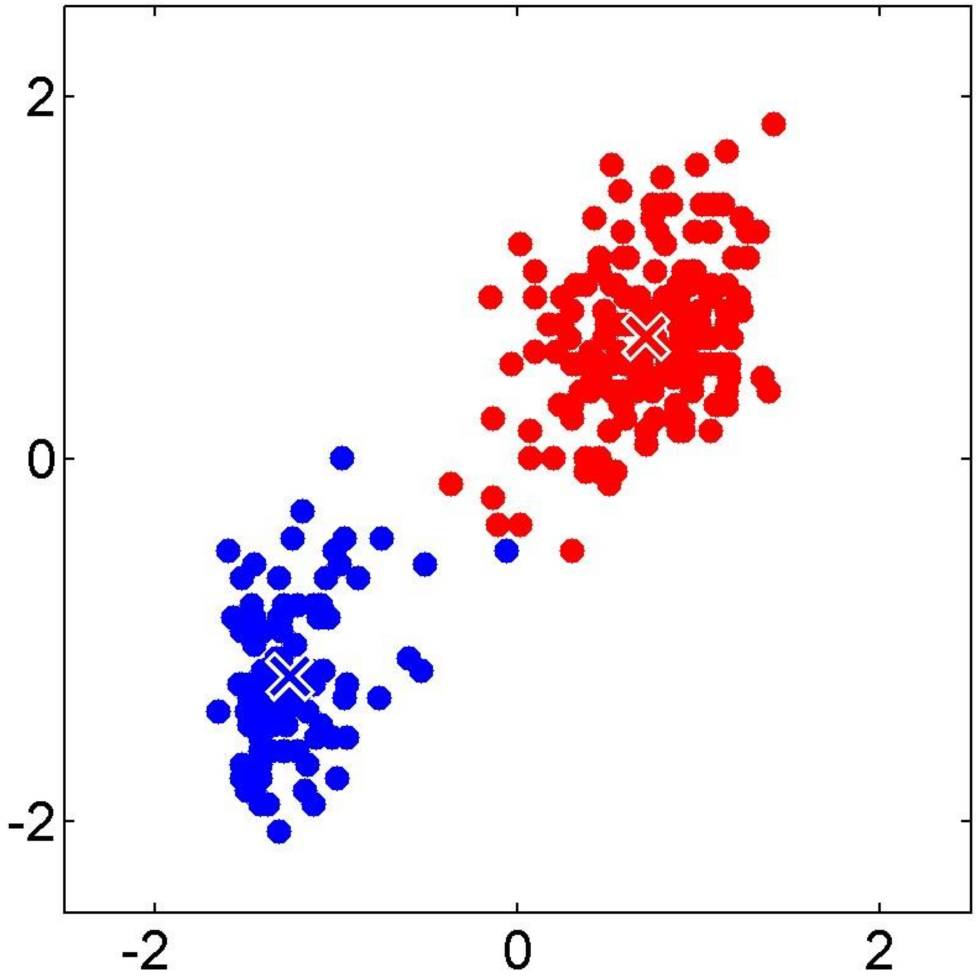
Ack: Chris Bishop

K-means clustering: E_4



Ack: Chris Bishop

K-means clustering: M_4



Ack: Chris Bishop

Responsibilities

- *Responsibilities* assign data points to clusters $r_{nk} \in \{0, 1\}$

  such that $\sum_k r_{nk} = 1$

  Example: 5 data points and 3 clusters

$$(r_{nk}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Ack: Chris Bishop

K-means Cost Function

data

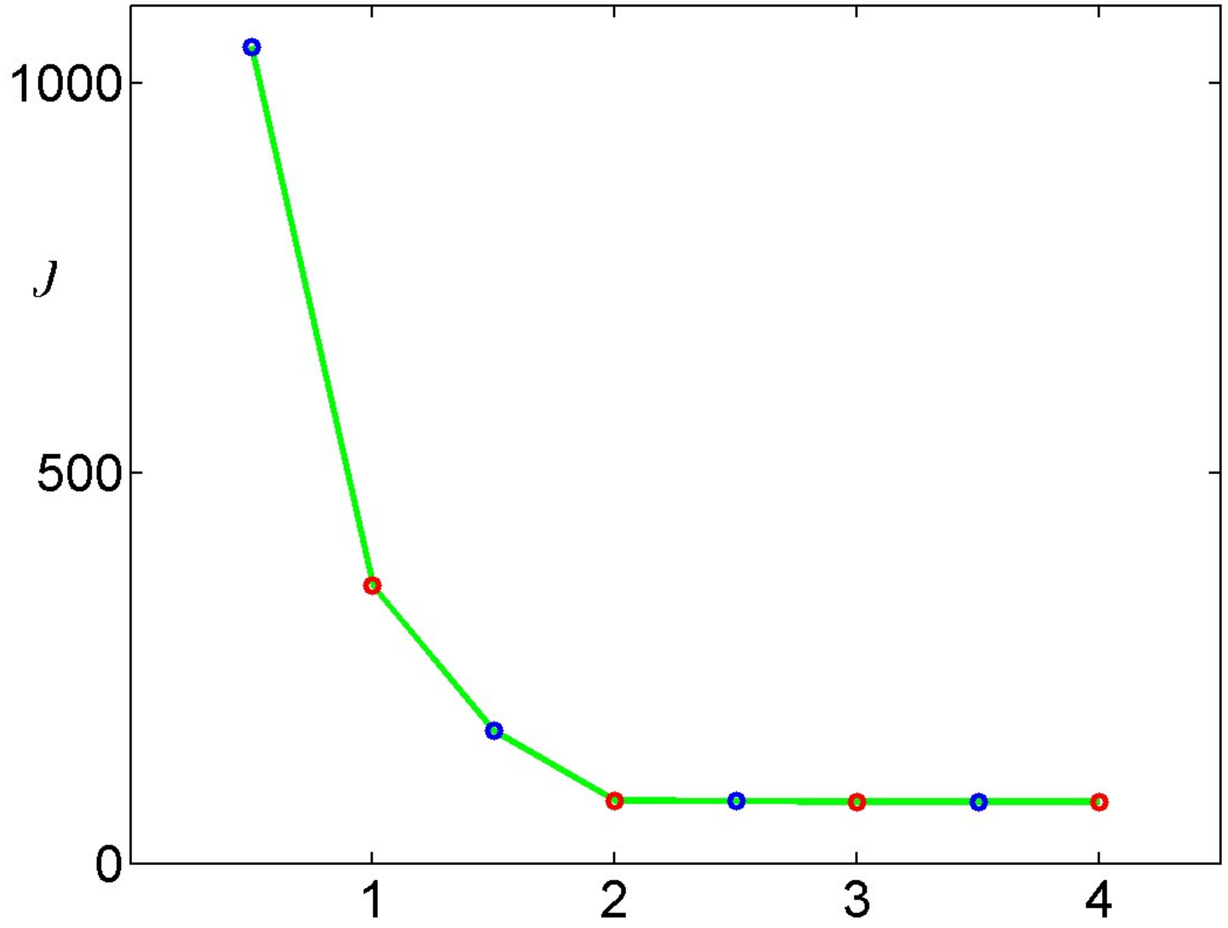
$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

responsibilities

prototypes

Ack: Chris Bishop

Minimizing the Cost Function



Ack: Chris Bishop

Minimizing the Cost Function

- E-step: minimize J w.r.t. r_{nk}
  - assigns each data point to nearest prototype
- M-step: minimize J w.r.t $\boldsymbol{\mu}_k$
  - gives

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{kn} \mathbf{x}_n}{\sum_n r_{kn}}$$

  - each prototype set to the mean of points in that cluster
- Convergence guaranteed since there is a finite number of possible settings for the responsibilities

Ack: Chris Bishop

Observation

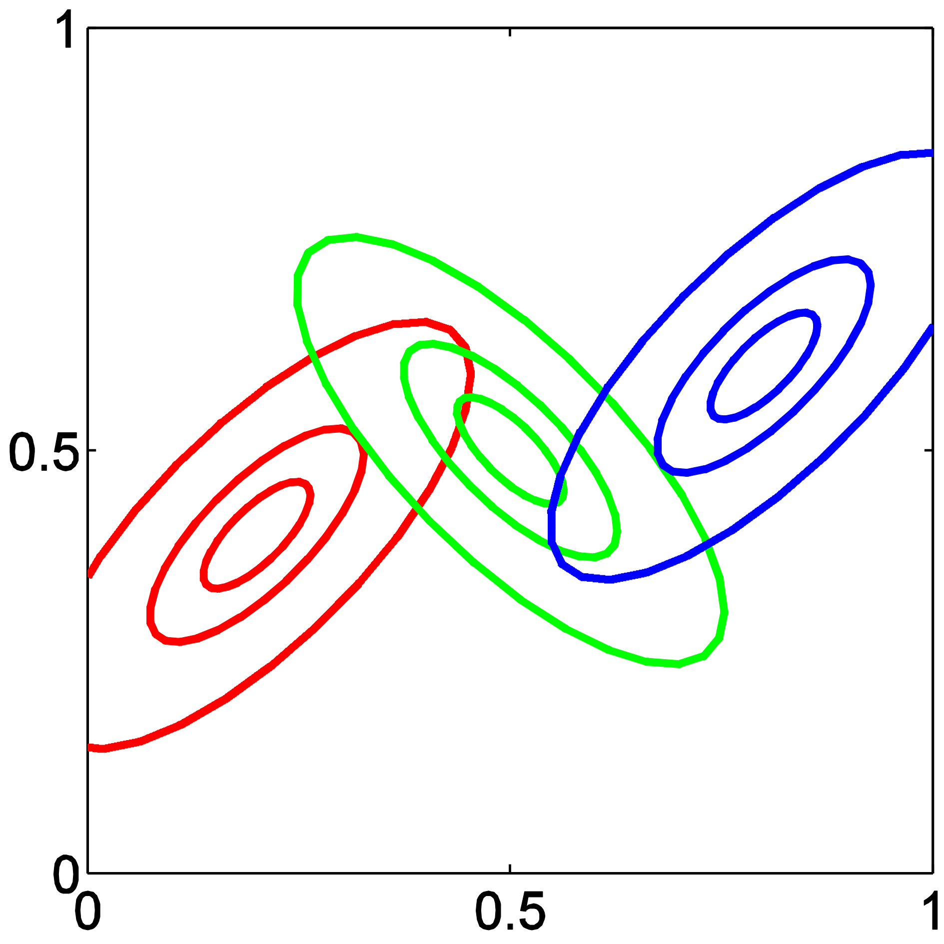
Limitation: Hard assignments of data points to clusters – small shift of a data point can flip it to a different cluster

Solution: replace 'hard' clustering of K-means with 'soft' probabilistic assignments

Represent the probability distribution of the data as a *Gaussian Mixture Model*

Ack: Chris Bishop

Gaussian Mixture Model



Ack: Chris Bishop

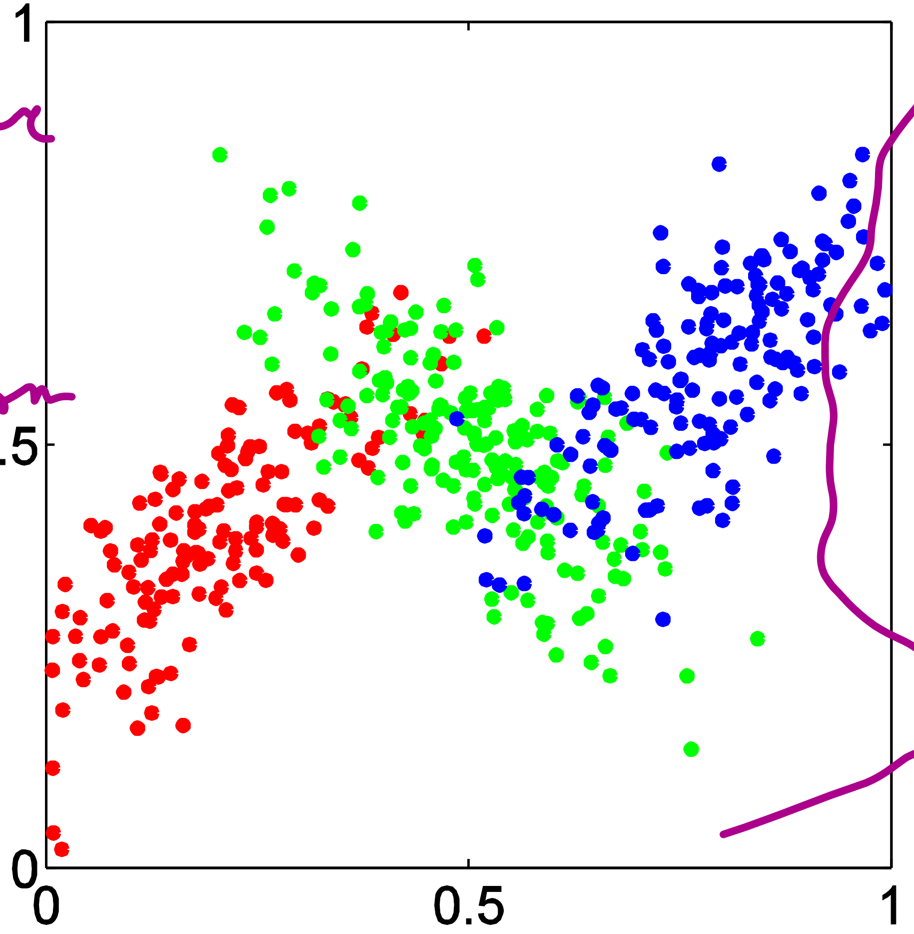
Synthetic Data Set: Complete Data

Parse Tree

Sentence

Symptom

Disease



Disease

Gen. model

Symptom

Inference

(μ, Σ)

θ

Generative model

data points

inference

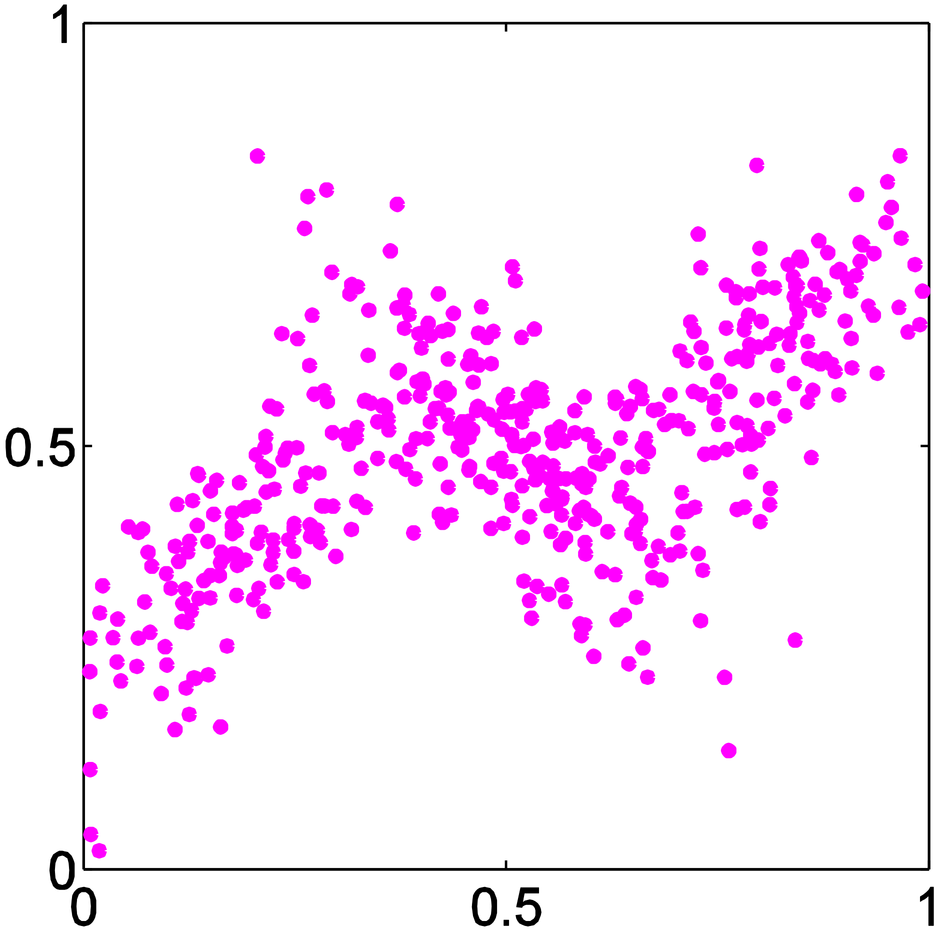
Ack: Chris Bishop

Fitting the Gaussian Mixture

- We wish to invert this process – given the data set, find the corresponding parameters:
  - mixing coefficients
  - means
  - covariances
- If we knew which component generated each data point, the maximum likelihood solution would involve fitting each component to the corresponding cluster
- Problem: the data set is unlabelled
- We shall refer to the labels as *latent* (= hidden) variables

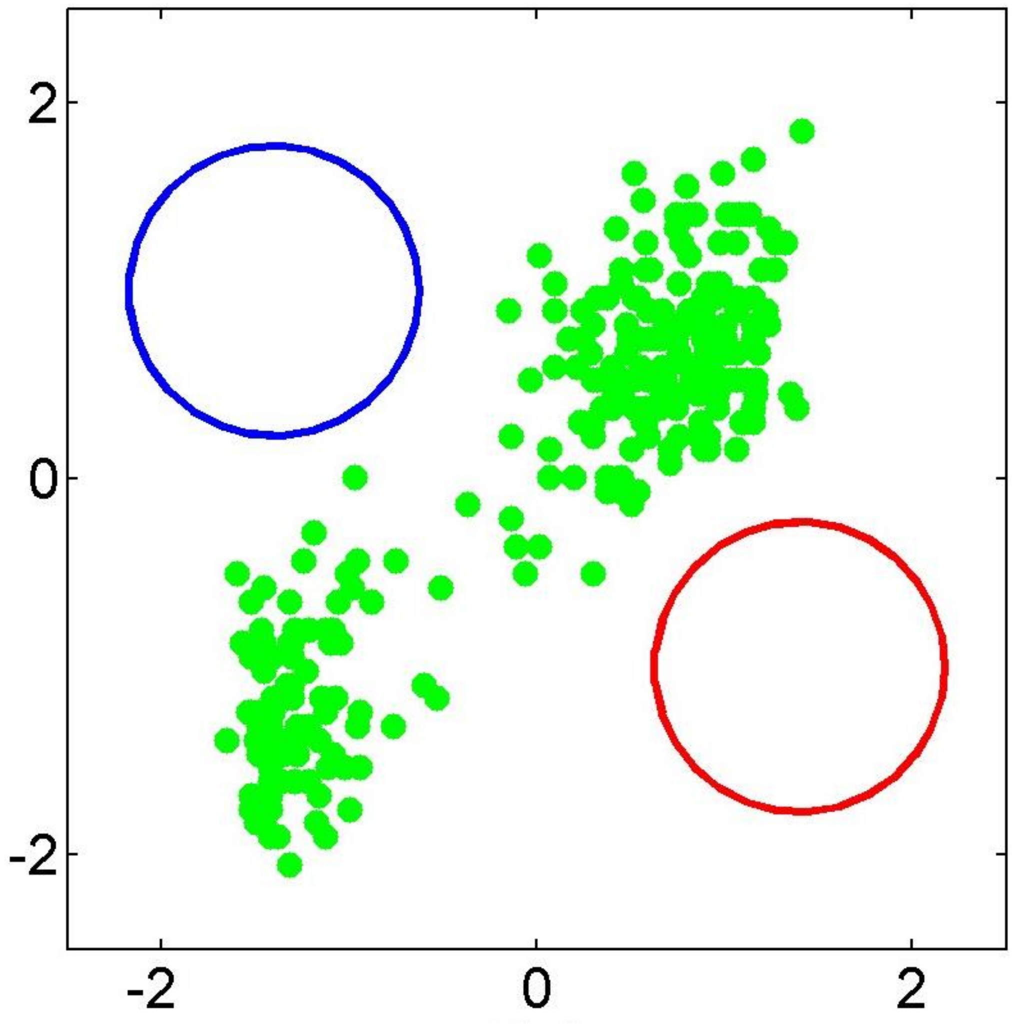
Ack: Chris Bishop

Synthetic Data Set Without Labels



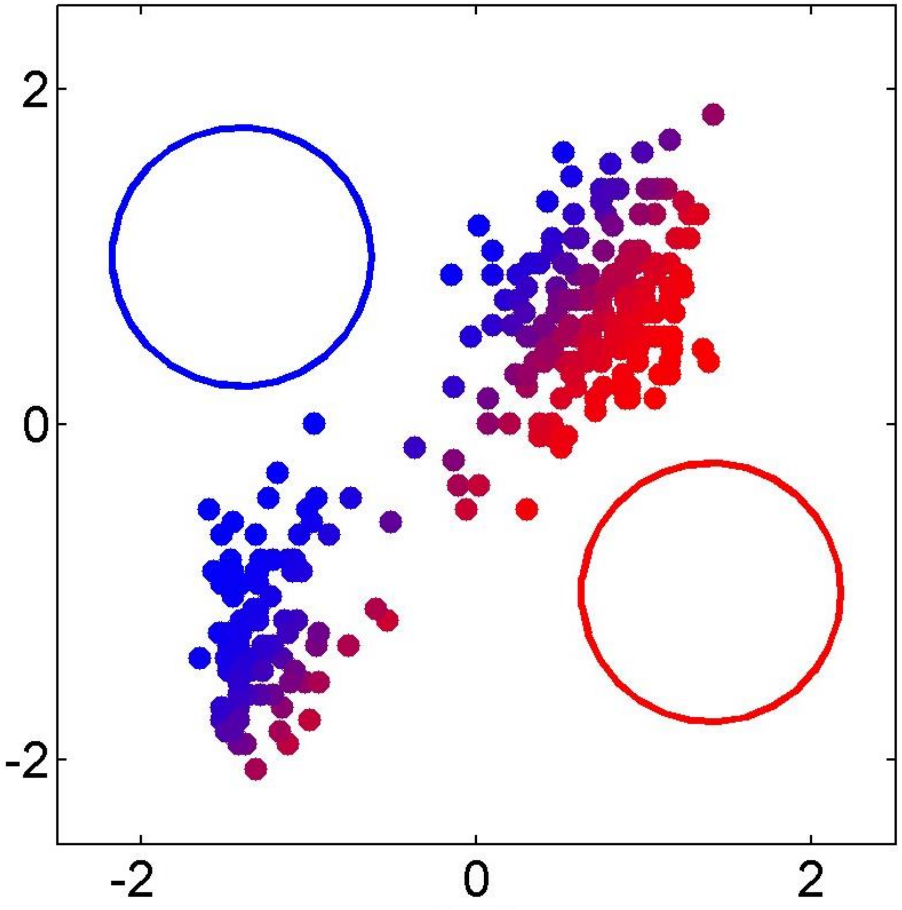
Ack: Chris Bishop

Synthetic Data: Incomplete Data (M_0)



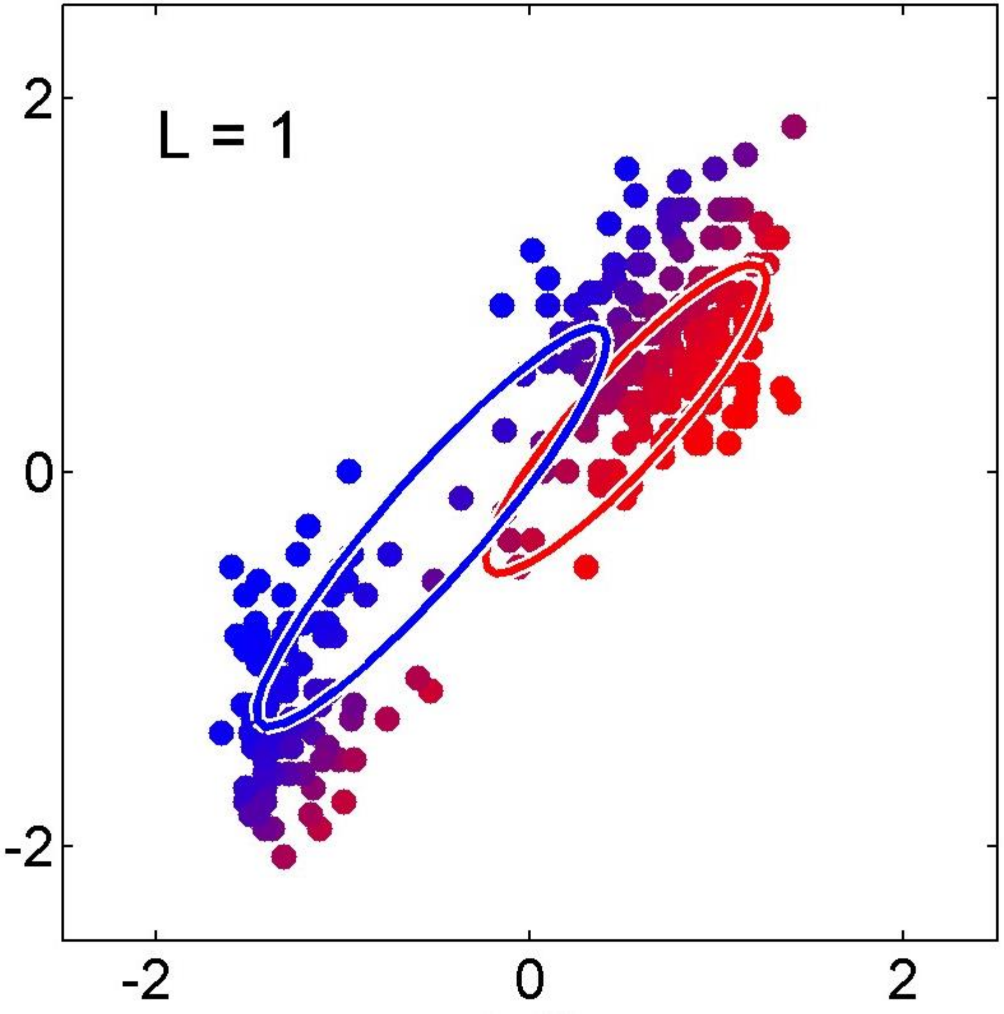
Ack: Chris Bishop

GMM: M_0+E_1



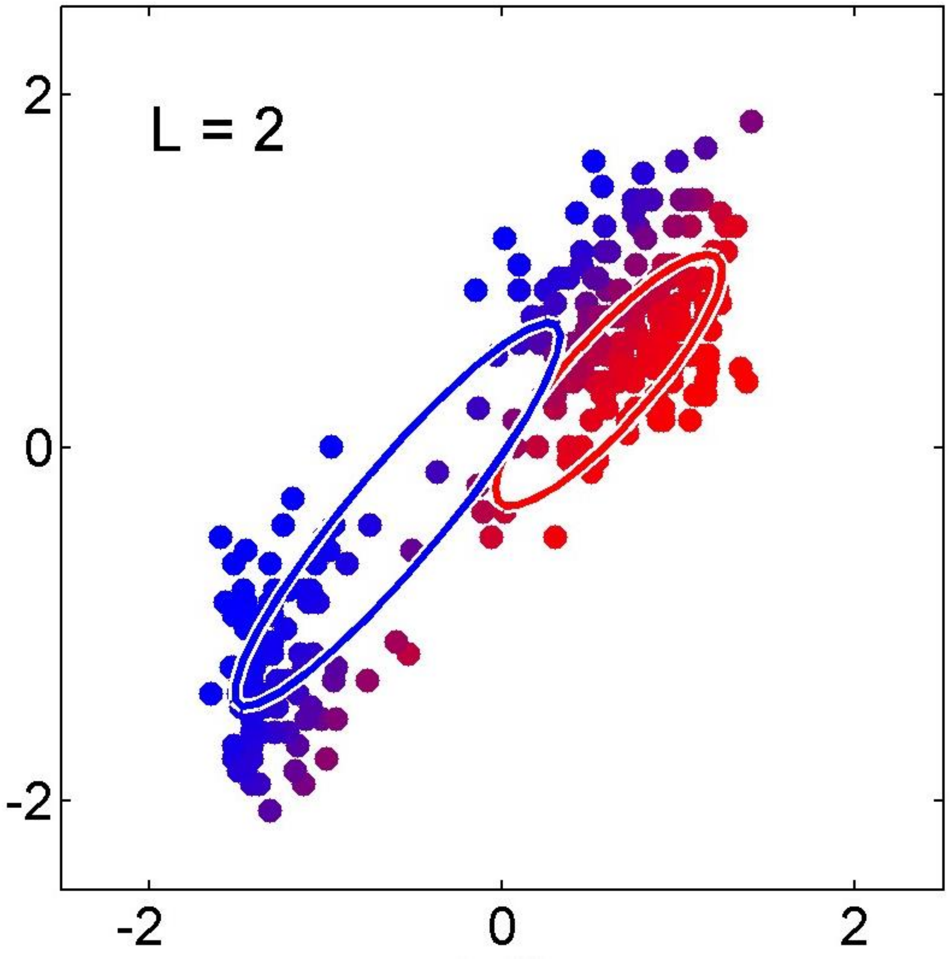
Ack: Chris Bishop

GMM: M_1



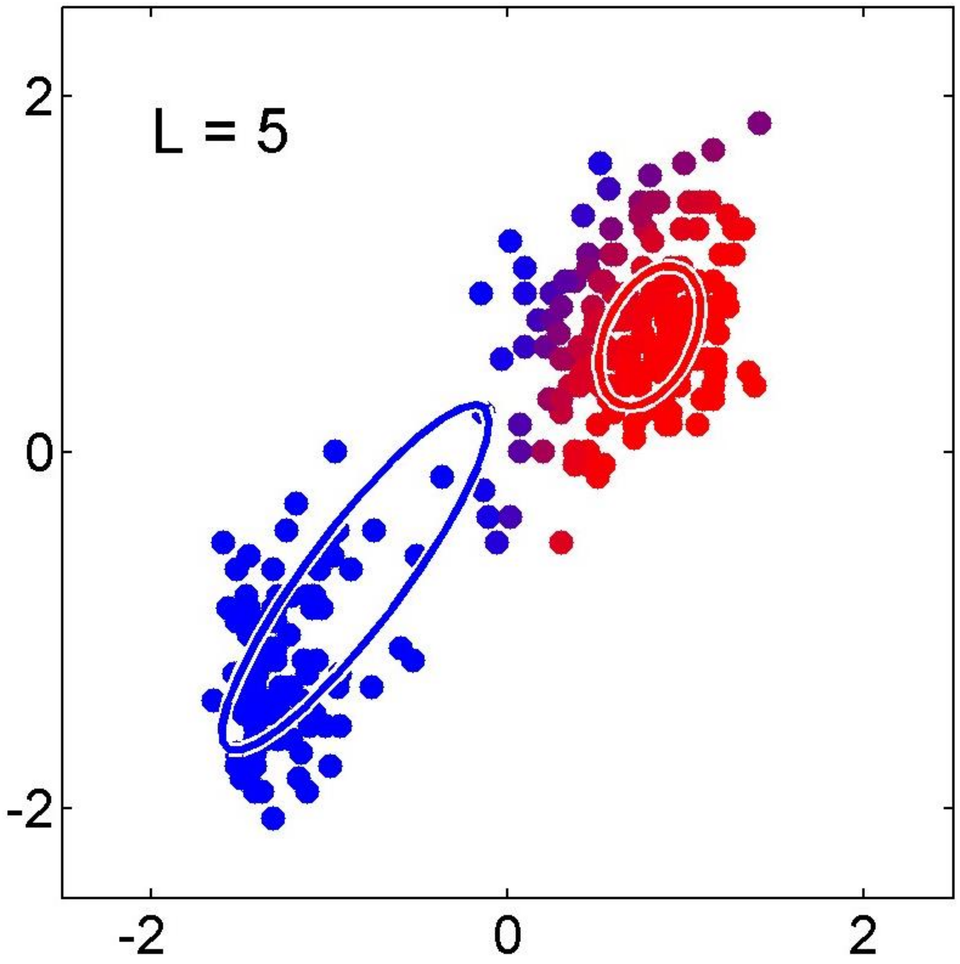
Ack: Chris Bishop

GMM: M_1+E_2



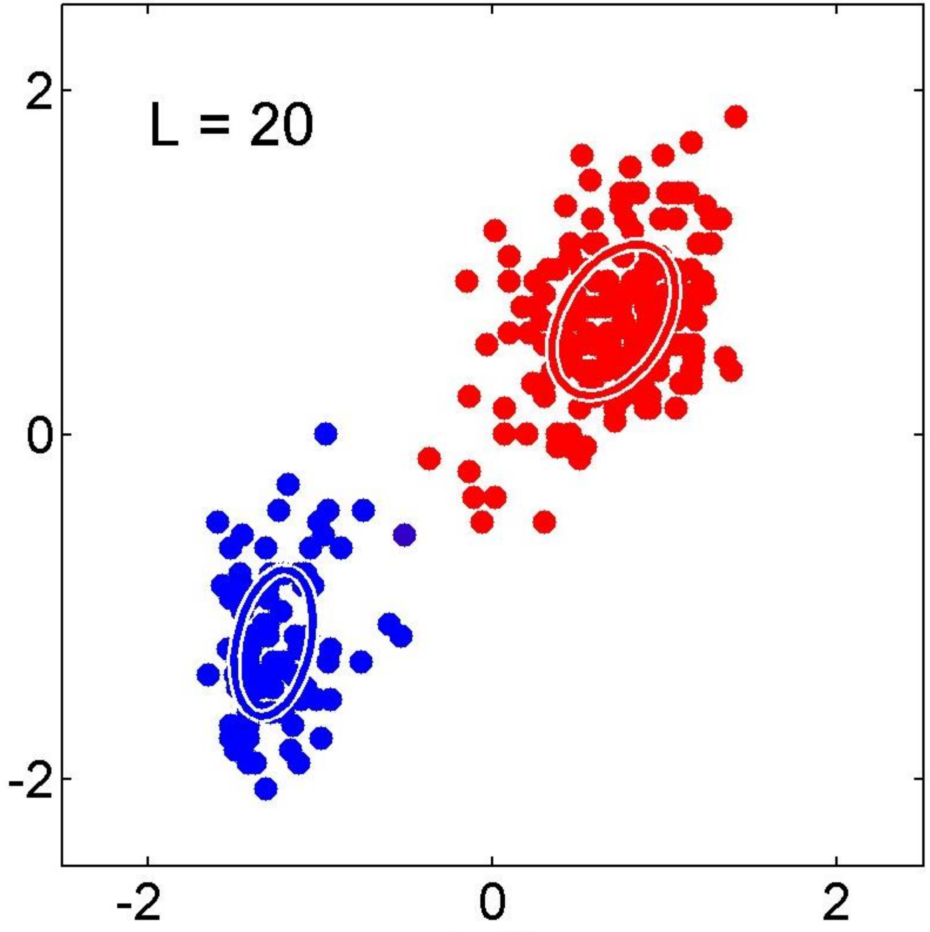
Ack: Chris Bishop

GMM: M_2+E_3



Ack: Chris Bishop

GMM: M_3+E_4



Ack: Chris Bishop

Example 2: Complete data

$$\boxed{\theta_A^{k}(1-\theta)^{n-k}}$$

$$\left(\frac{k}{n}\right)$$

a Maximum likelihood

		Coin A	Coin B
B	H T T T H H T H T H		5 H, 5 T
A	H H H H T H H H H H	9 H, 1 T	
A	H T H H H H H T H H	8 H, 2 T	
B	H T H T T T H H T T		4 H, 6 T
A	T H H H T H H H T H	7 H, 3 T	
		24 H, 6 T	9 H, 11 T

5 sets, 10 tosses per set

$$\hat{\theta}_A = \frac{24}{24+6} = 0.80$$

$$\hat{\theta}_B = \frac{9}{9+11} = 0.45$$

Refer: shared paper

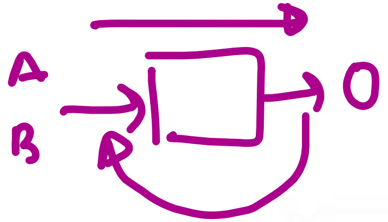
Example 2: Incomplete Data

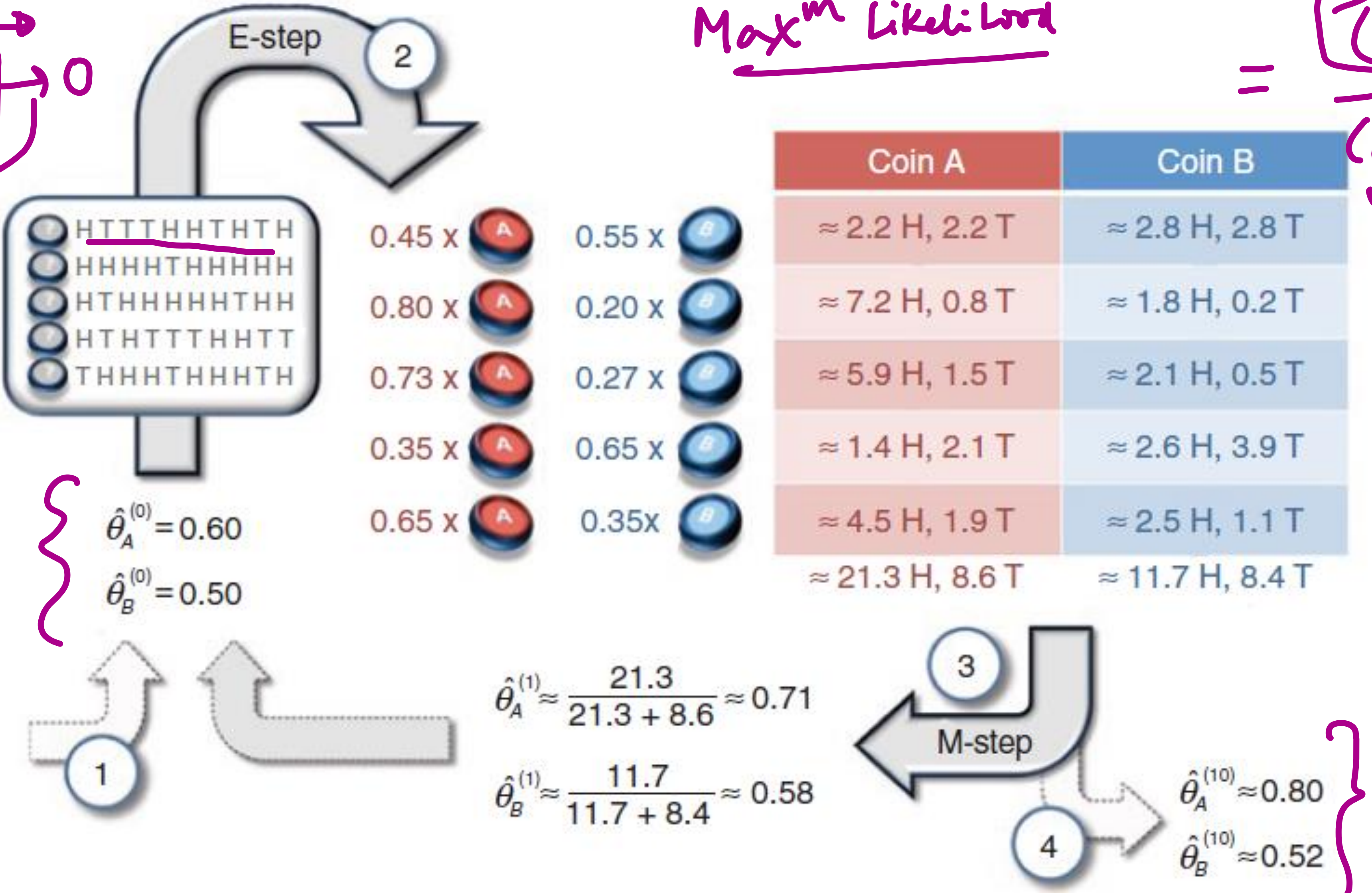
$$P(A|O) = \frac{P(O|A).P(A)}{P(O|A).P(A) + P(O|B).P(B)}$$

Max^m Likelihood

$$= \frac{(0.6)^5(0.4)^5}{(0.6)^5(0.4)^5 + (0.5)^5}$$

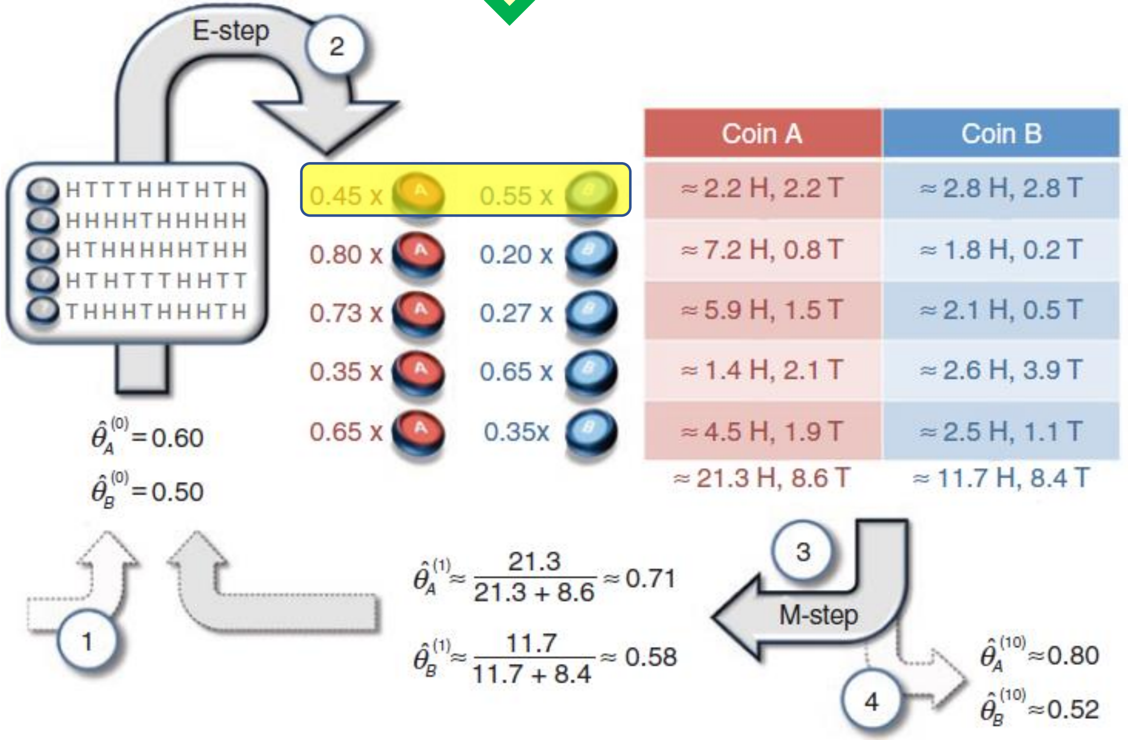
$= 0.45$





Refer: shared paper

Example 2: Doubt posted in feedback



$O = \langle HTTTTHHTHTH \rangle$

Posterior calculation. 5 heads, 5 tails, used generative model.

$$P(A|O) = \frac{P(O|A).P(A)}{P(O)} = \frac{P(O|A).P(A)}{P(O|A).P(A) + P(O|B).P(B)}$$

$$= \frac{(0.6)^5 (0.4)^5 . 0.5}{(0.6)^5(0.4)^5 . 0.5 + (0.5)^5 (0.5)^5 . 0.5}$$ uniform prior.

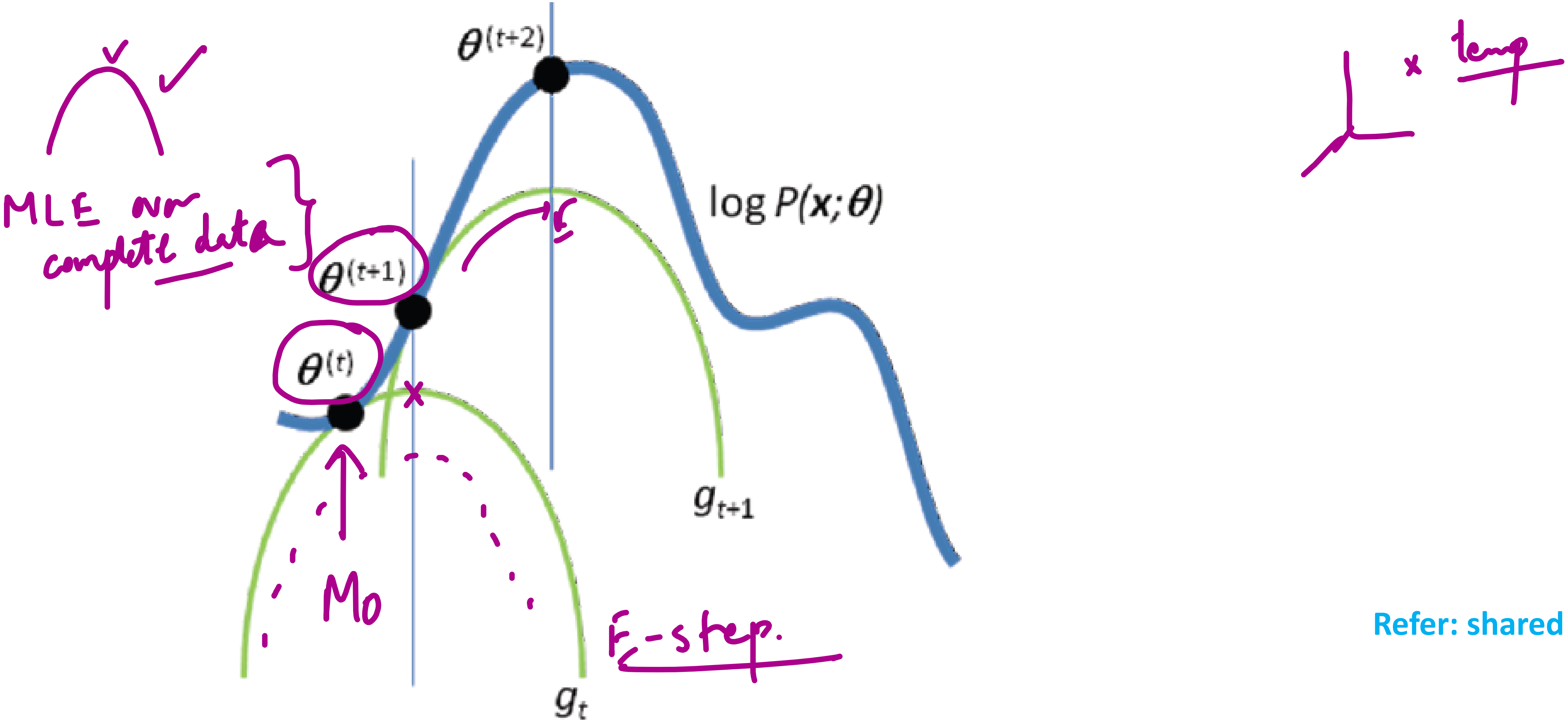
$= 0.45.$

Similarly $P(B|O) = \frac{P(O|B).P(B)}{P(O|A).P(A) + P(O|B).P(B)} = 0.55$

Refer: shared paper

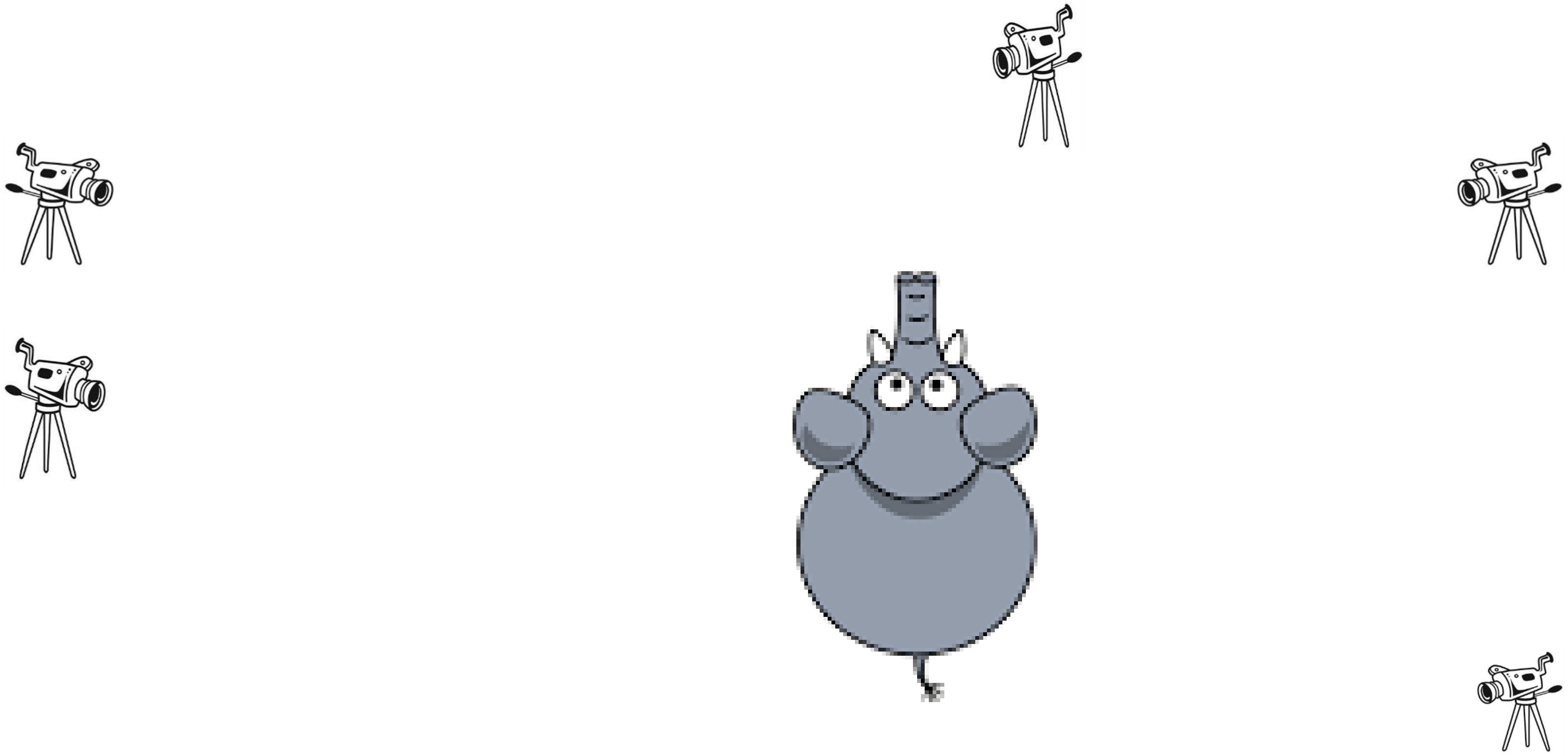
Goal of E step is to estimate posteriors like P(A|O).
Estimating P(O|A) is the preparation for E step. This involves using the generative storyline.

EM intuition : lower bound maximization



Refer: shared paper

EM intuition



EM over PCFGs : an example

$f(y)$	y
5	y_1
10	$y2$

y_1 = "Mary saw a bird on a tree"
y_2 = "a bird on a tree saw a worm"

M_0 step

$$
\begin{array}{llll}
S & \longrightarrow & NP\ VP & (1.00) \\
VP & \longrightarrow & V\ NP & (0.50) \\
VP & \longrightarrow & V\ NP\ PP & (0.50) \\
NP & \longrightarrow & NP\ PP & (0.25) \\
NP & \longrightarrow & Mary & (0.25) \\
NP & \longrightarrow & a\ bird & (0.25) \\
NP & \longrightarrow & a\ worm & (0.25) \\
PP & \longrightarrow & on\ a\ tree & (1.00) \\
V & \longrightarrow & saw & (1.00)
\end{array}
$$

Refer Workbook

EM over PCFGs : an example

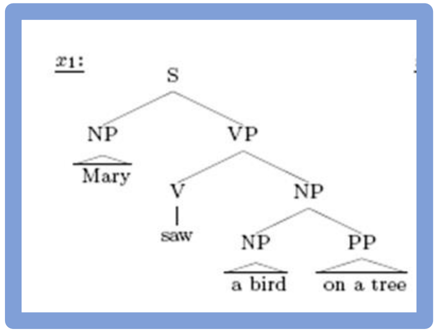
Diseases. $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

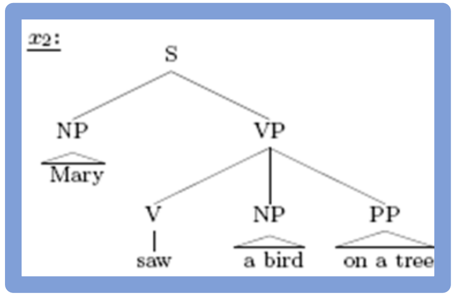
obs/Symp. y_1

Parse trees → NOISY CHANNEL → Sentence

Inferring

y_1 = *"Mary saw a bird on a tree"*





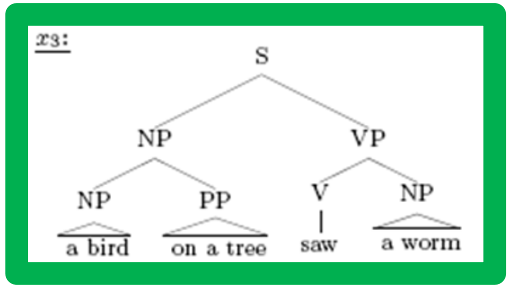
$p(A|O)$

$$P(x_1|y_1) = \frac{\boxed{P(y_1|x_1)}\,P(x_1)}{P(y_1|x_1)\cdot P(x_1) + p(y|x_2)P}$$

COLOURING

$P(x_2|y_1)$

y_2 = *"a bird on a tree saw a worm"*

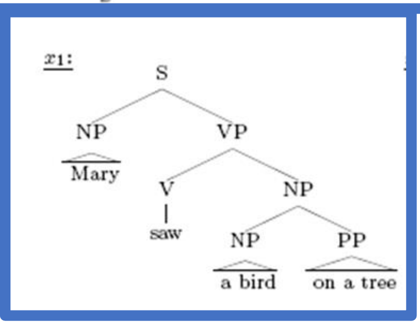


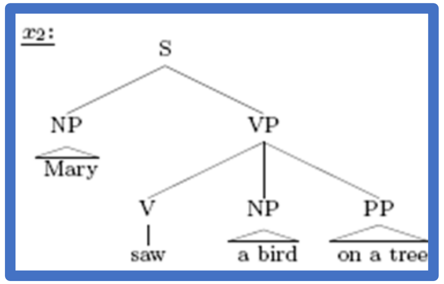
Refer Workbook

EM over PCFGs : an example (preparation for E step)

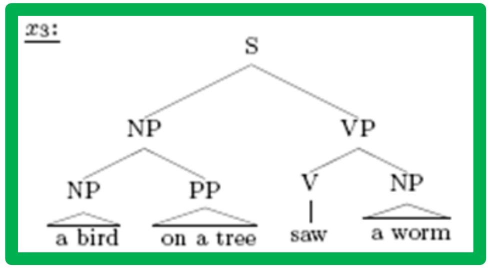
Preparatory :

y_1 = "*Mary saw a bird on a tree*"





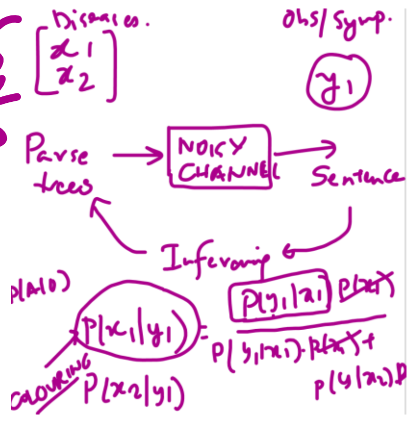
y_2 = "*a bird on a tree saw a worm*"



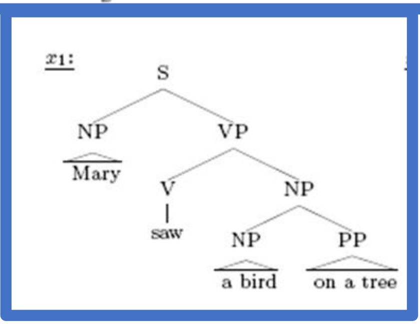
$$
\begin{aligned}
p_0(x_1) &= p(\text{S} \longrightarrow \text{NP VP}) \cdot p\left(\frac{\text{NP}}{\text{Mary}}\right) \cdot p(\text{VP} \longrightarrow \text{V NP}) \cdot p(\text{V} \longrightarrow \text{saw}) \cdot p\left(\frac{\text{NP}}{\text{a bird}}\right) \cdot p\left(\frac{\text{PP}}{\text{on a tree}}\right) \\
&= 1.00 \cdot 0.25 \cdot 0.50 \cdot 1.00 \cdot 0.25 \cdot 0.25 \cdot 1.00 \\
&= \underline{0.0078125} \quad P(y_1|x_1) \\
p_0(y_1) &= p_0(x_1) + p_0(x_2) \\
&= 0.0078125 + 0.0312500 \quad p(y_1|x_2) \\
&= 0.0390625
\end{aligned}
$$

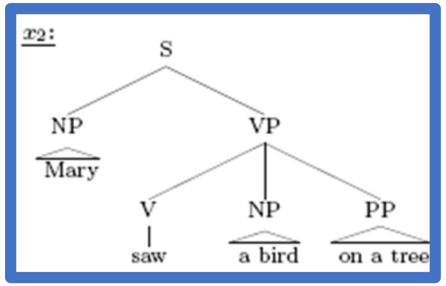
$$
\left\{
\begin{aligned}
P(x_1|y_1) &= 1/5 \\
P(x_2|y_1) &= 4/5 \\
P(x_3|y_2) &= 1
\end{aligned}
\right.
$$

$p_0(x)$	x		$p_0(y)$	y
$P(y_1\|x_1)$ → 0.0078125	x_1 ✓	1/5	0.0390625	y_1
$P(y_1\|x_2)$ → 0.0312500	x_2 ✓		0.0078125	y_2
$P(y_2\|x_3)$ → 0.0078125	x_3 ✓			

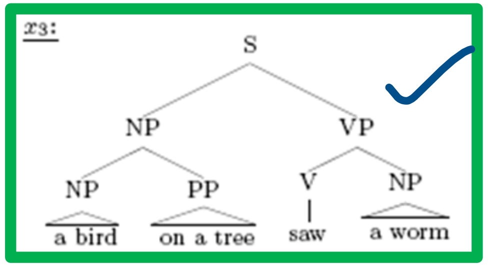


y_1 = "Mary saw a bird on a tree"





y_2 = "a bird on a tree saw a worm"



$f(y)$	y
5	y_1
10	$y2$

y_1 = "Mary saw a bird on a tree"
y_2 = "a bird on a tree saw a worm"

$p_0(x)$	x
0.0078125	x_1
0.0312500	x_2
0.0078125	x_3

$p_0(y)$	y
0.0390625	y_1
0.0078125	y_2

$f_{T_q}(x)$	x
1	x_1
4	x_2
10	x_3

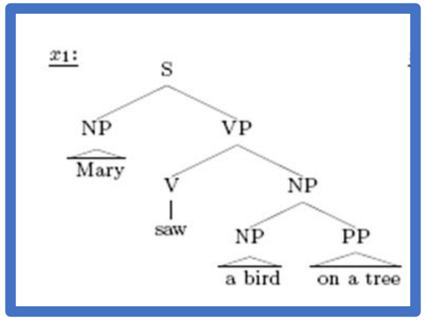
$$\begin{aligned} f_{T_q}(x_1) &= f(\text{yield}(x_1)) \cdot q(x_1|\text{yield}(x_1)) \\ &= f(y_1) \cdot q(x_1|y_1) \\ &= f(y_1) \cdot \frac{q(x_1)}{q(y_1)} \\ &= 5 \cdot \frac{0.0078125}{0.0390625} \\ &= 1 \end{aligned}$$

Refer Workbook

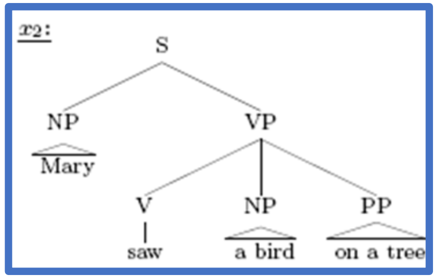
M step : estimate the PCFG parameters

$f_{T_q}(x)$	x
1	x_1
4	x_2
10	x_3

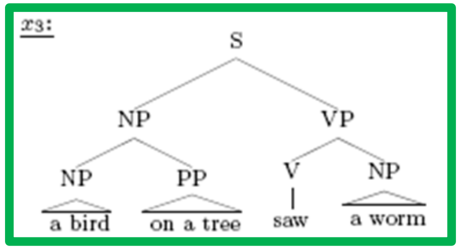
1 instance of



4 instances of



10 instances of



$$
\begin{array}{lcll}
\text{S} & \longrightarrow & \text{NP VP} & (1.000) \\
\text{VP} & \longrightarrow & \text{V NP} & (0.733 \approx \frac{1+10}{15}) \\
\text{VP} & \longrightarrow & \text{V NP PP} & (0.267 \approx \frac{4}{15}) \\
\text{NP} & \longrightarrow & \text{NP PP} & (0.268 \approx \frac{1+10}{41}) \\
\text{NP} & \longrightarrow & \text{Mary} & (0.122 \approx \frac{1+4}{41}) \\
\text{NP} & \longrightarrow & \text{a bird} & (0.366 \approx \frac{1+4+10}{41}) \\
\text{NP} & \longrightarrow & \text{a worm} & (0.244 \approx \frac{10}{41}) \\
\text{PP} & \longrightarrow & \text{on a tree} & (1.000) \\
\text{V} & \longrightarrow & \text{saw} & (1.000)
\end{array}
$$

Refer Workbook

M step: comparison of results of M_1 and M_0 steps

After M_1 step

$$
\begin{array}{llll}
\mathrm{S} & \longrightarrow & \mathrm{NP\ VP} & (1.000) \\
\mathrm{VP} & \longrightarrow & \mathrm{V\ NP} & (0.733 \approx \frac{1+10}{15}) \\
\mathrm{VP} & \longrightarrow & \mathrm{V\ NP\ PP} & (0.267 \approx \frac{4}{15}) \\
\mathrm{NP} & \longrightarrow & \mathrm{NP\ PP} & (0.268 \approx \frac{1+10}{41}) \\
\mathrm{NP} & \longrightarrow & \mathrm{Mary} & (0.122 \approx \frac{1+4}{41}) \\
\mathrm{NP} & \longrightarrow & \mathrm{a\ bird} & (0.366 \approx \frac{1+4+10}{41}) \\
\mathrm{NP} & \longrightarrow & \mathrm{a\ worm} & (0.244 \approx \frac{10}{41}) \\
\mathrm{PP} & \longrightarrow & \mathrm{on\ a\ tree} & (1.000) \\
\mathrm{V} & \longrightarrow & \mathrm{saw} & (1.000)
\end{array}
$$

After M_0 step

$$
\begin{array}{llll}
S & \longrightarrow & NP\ VP & (1.00) \\
VP & \longrightarrow & V\ NP & (0.50) \\
VP & \longrightarrow & V\ NP\ PP & (0.50) \\
NP & \longrightarrow & NP\ PP & (0.25) \\
NP & \longrightarrow & Mary & (0.25) \\
NP & \longrightarrow & a\ bird & (0.25) \\
NP & \longrightarrow & a\ worm & (0.25) \\
PP & \longrightarrow & on\ a\ tree & (1.00) \\
V & \longrightarrow & saw & (1.00)
\end{array}
$$

Refer Workbook

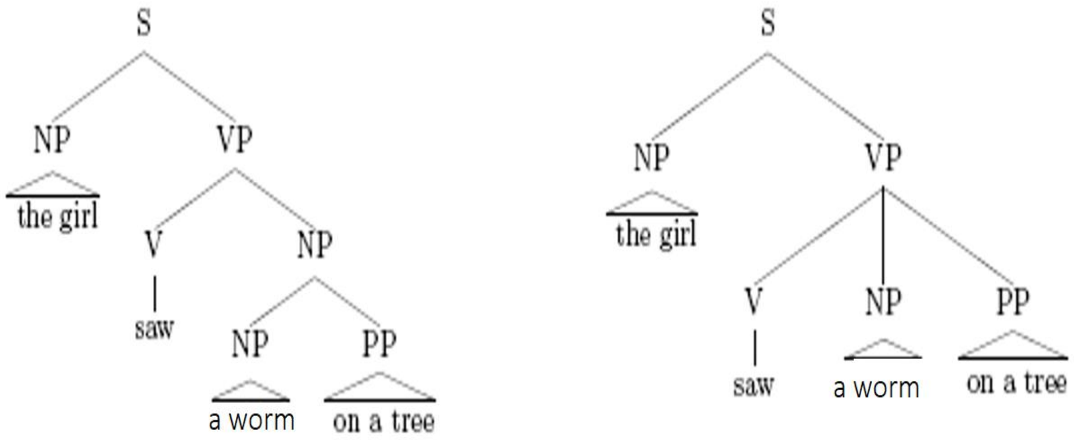
Finally …

CFG rule	p_0	p_1	p_2	p_3	…	p_{18}
S $\longrightarrow$ NP VP	1.000	1.000	1.000	1.000		1.000
VP $\longrightarrow$ V NP	0.500	0.733	0.807	0.850		0.967
VP $\longrightarrow$ V NP PP	0.500	0.267	0.193	0.150		0.033
NP $\longrightarrow$ NP PP	0.250	0.268	0.287	0.298		0.326
NP $\longrightarrow$ Mary	0.250	0.122	0.118	0.117		0.112
NP $\longrightarrow$ a bird	0.250	0.366	0.357	0.351		0.337
NP $\longrightarrow$ a worm	0.250	0.244	0.238	0.234		0.225
PP $\longrightarrow$ on a tree	1.000	1.000	1.000	1.000		1.000
V $\longrightarrow$ saw	1.000	1.000	1.000	1.000		1.000

Refer Workbook

Using the PCFG for disambiguation

CFG rule	p_0	p_1	p_2	p_3	...	p_{18}
S $\longrightarrow$ NP VP	1.000	1.000	1.000	1.000		1.000
VP $\longrightarrow$ V NP	0.500	0.733	0.807	0.850		0.967
VP $\longrightarrow$ V NP PP	0.500	0.267	0.193	0.150		0.033
NP $\longrightarrow$ NP PP	0.250	0.268	0.287	0.298		0.326
NP $\longrightarrow$ Mary	0.250	0.122	0.118	0.117		0.112
NP $\longrightarrow$ a bird	0.250	0.366	0.357	0.351		0.337
NP $\longrightarrow$ a worm	0.250	0.244	0.238	0.234		0.225
PP $\longrightarrow$ on a tree	1.000	1.000	1.000	1.000		1.000
V $\longrightarrow$ saw	1.000	1.000	1.000	1.000		1.000



$$\boxed{p(\textbf{VP} \longrightarrow \textbf{V NP}) \cdot p(\textbf{NP} \longrightarrow \textbf{NP PP}) > p(\textbf{VP} \longrightarrow \textbf{V NP PP})}$$

The parse on the left is preferred if:

$$p(\text{VP} \longrightarrow \text{V NP}) \cdot p(\text{NP} \longrightarrow \text{NP PP}) > p(\text{VP} \longrightarrow \text{V NP PP})$$

p	$p(\text{VP} \longrightarrow \text{V NP}) \cdot p(\text{NP} \longrightarrow \text{NP PP})$	$p(\text{VP} \longrightarrow \text{V NP PP})$
p_0	$0.500 \cdot 0.250 = 0.125$	0.500
p_1	$0.733 \cdot 0.268 = 0.196$	0.267
p_2	$0.807 \cdot 0.287 = \mathbf{0.232}$	0.193
p_3	$0.850 \cdot 0.298 = \mathbf{0.253}$	0.150
$\vdots$		
p_{18}	$0.967 \cdot 0.326 = \mathbf{0.315}$	0.033

Refer Workbook

PCFGs : The Big Picture

Sentence → **Statistical Parser (Parsing cum disambiguation)** → Most probable parse(s)

↑

Identifying CFG rules, grammar transformation etc.

Corpus annotation

Learning parameters (PCFG rule probabilities) by EM or MLE

Refer Workbook

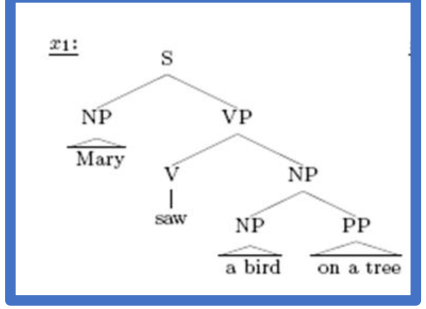
Perpetual Motion Machine?

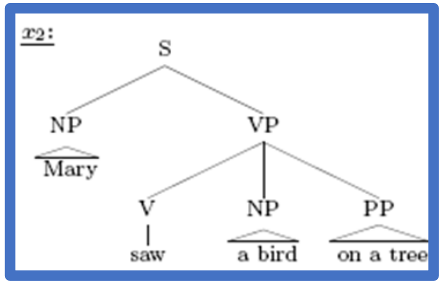
- Is this magic? Where does the extra knowledge come from?

Perpetual motion wheels from a drawing by Leonardo da Vinci

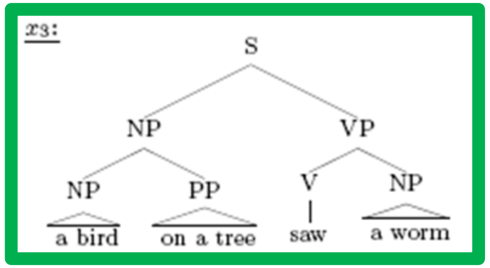
Ack:Wikipedia

y_1 = "*Mary saw a bird on a tree*"



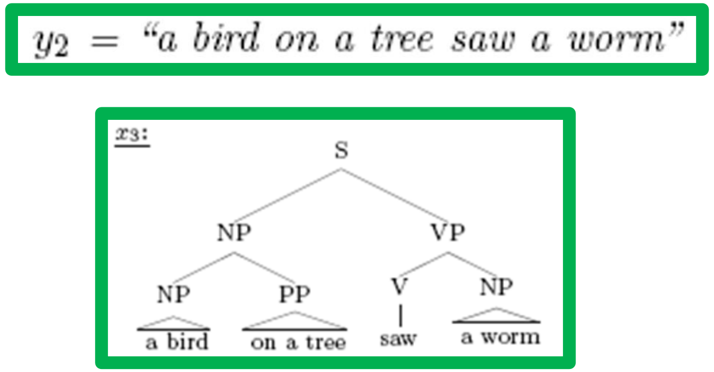


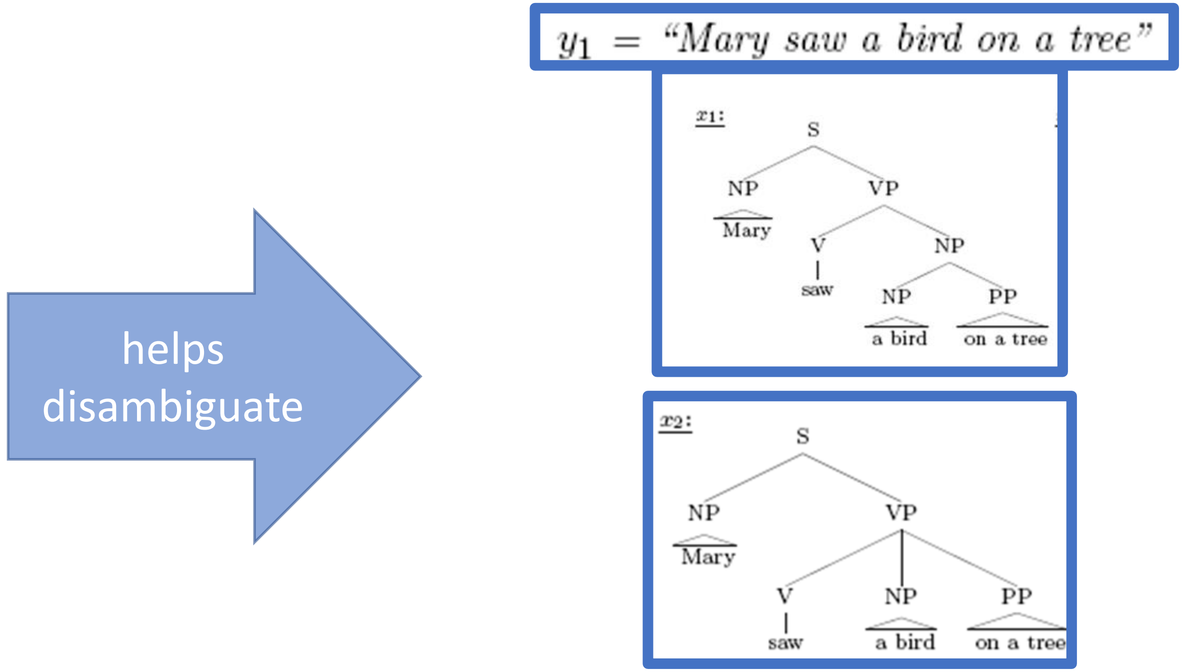
y_2 = "*a bird on a tree saw a worm*"



Perpetual Motion Machine?

- Is this magic? Where does the extra knowledge come from?





Ack:Wikipedia

The Unifying Picture

	K-means	EM	Biased Coins	PCFG
Source				
Observations				
Generative storyline (preparation for E step)				
Parameters Estimated				
Why it works				

Reference

A Tutorial on the Expectation-Maximization Algorithm Including Maximum-Likelihood Estimation and EM Training of Probabilistic Context-Free Grammars

By Detlef Prescher