5th SIPTA school
on imprecise probability

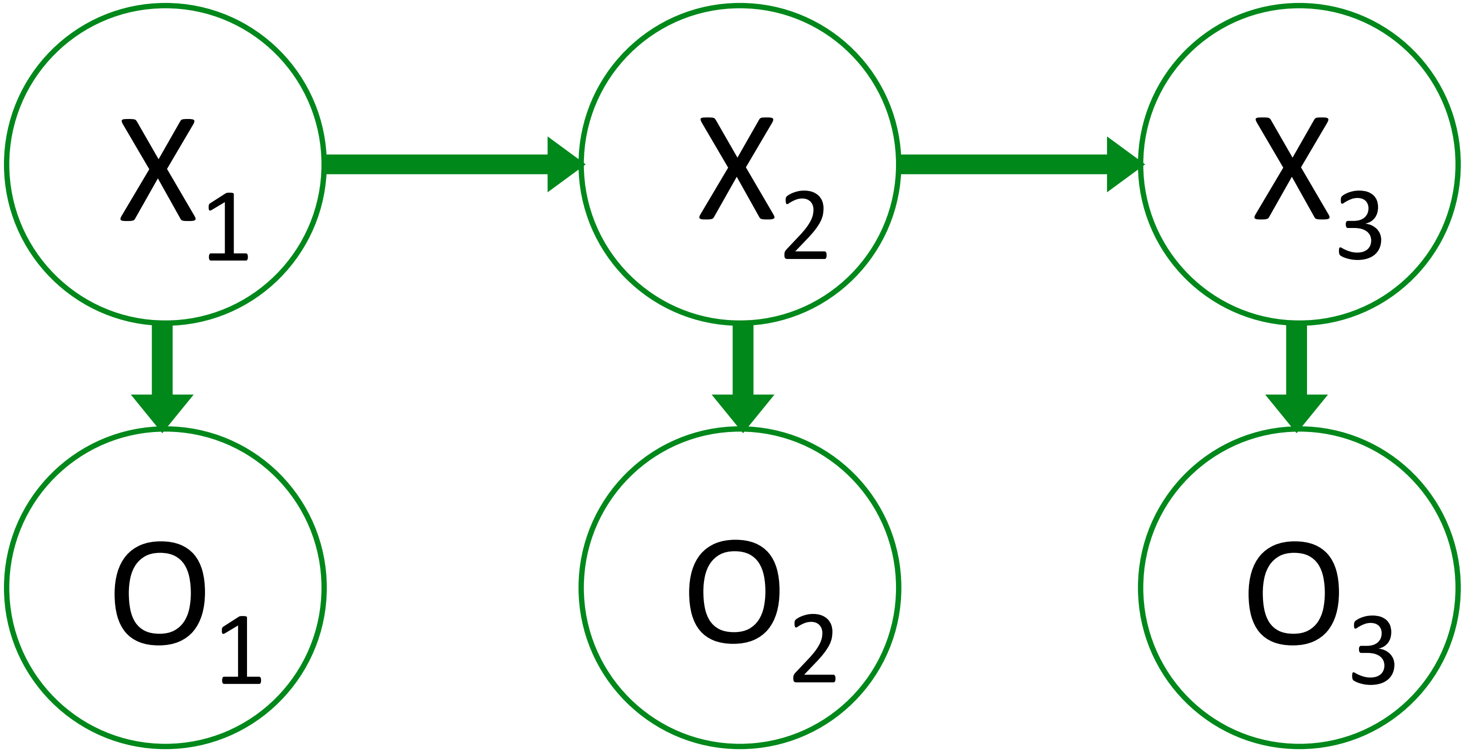
16-20 July 2012, Pescara (Italy)

Robustly correcting mistakes made by OCR software

Jasper De Bock
University of Ghent (Belgium)
jasper.debock@ugent.be

(imprecise) state sequence estimation

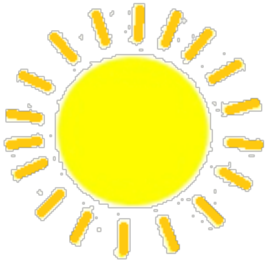
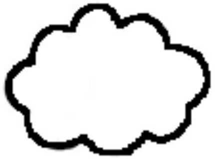
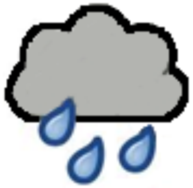
A sequence of hidden state variables

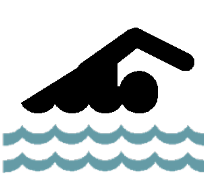
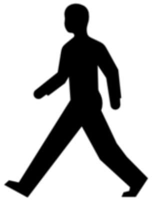
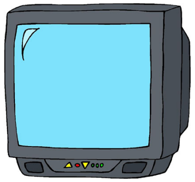


A sequence of observable output variables

(imprecise) state sequence estimation

A sequence of hidden state variables

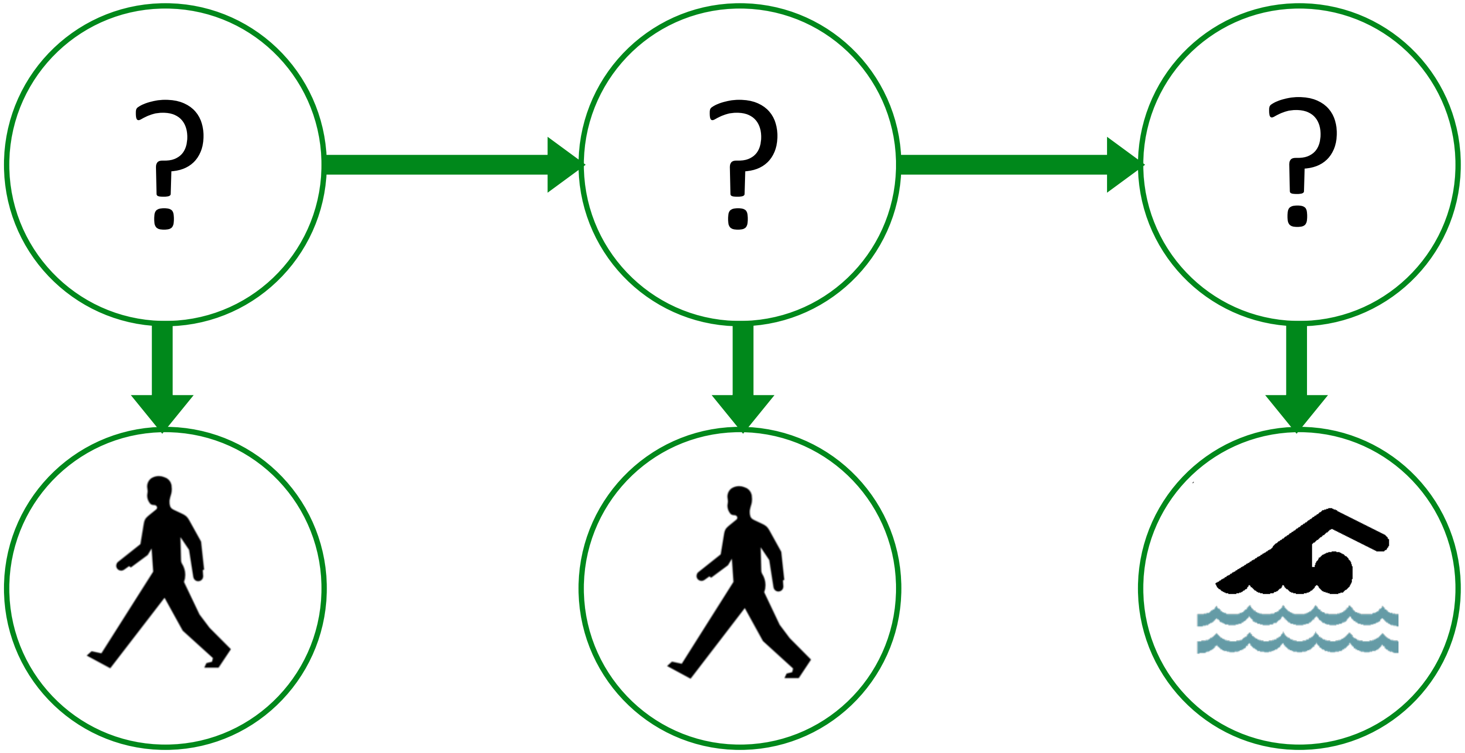
X =  or  or 

O =  or  or 

A sequence of observable output variables

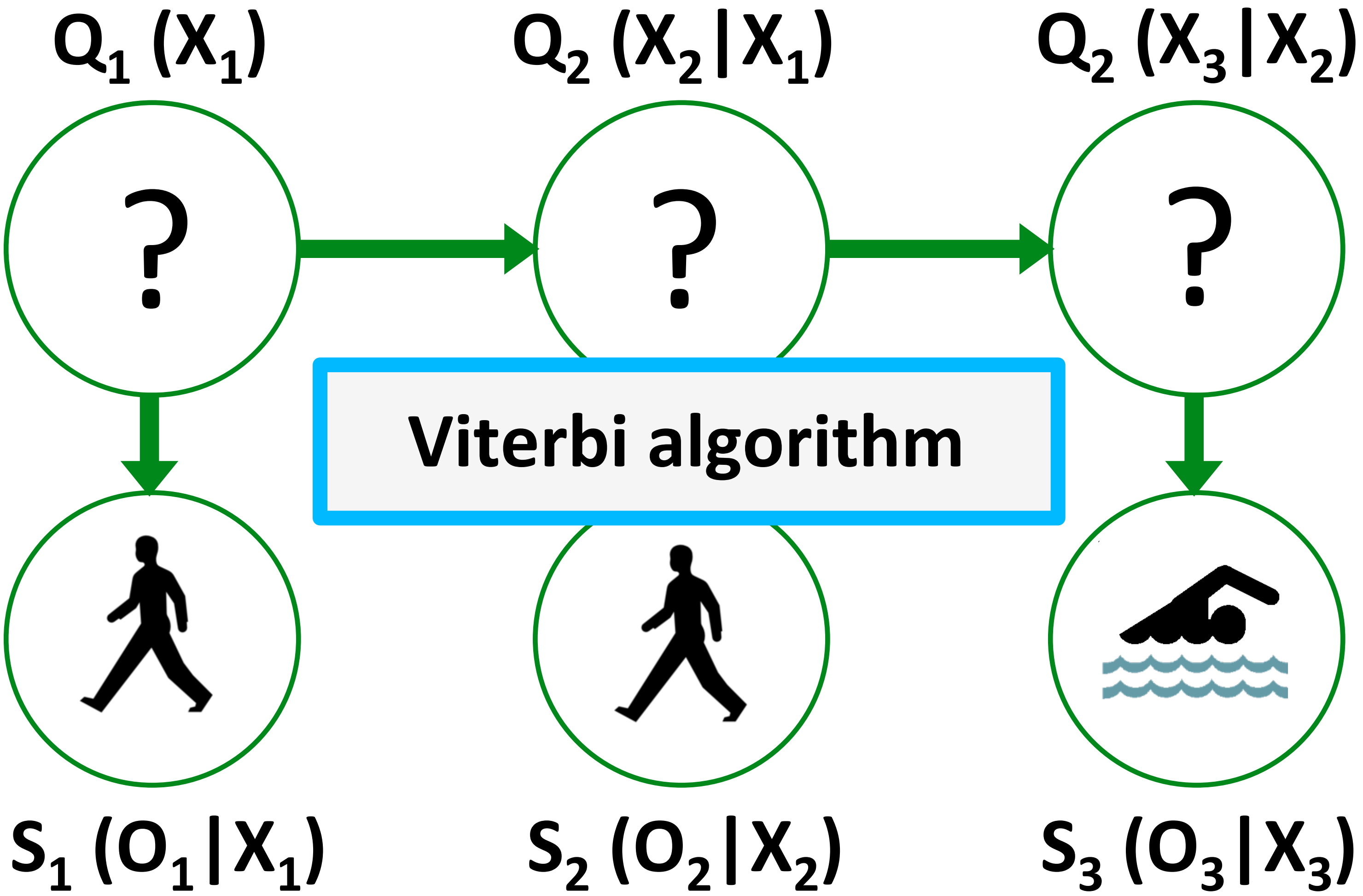
(imprecise) state sequence estimation

A sequence of hidden state variables



A sequence of observable output variables

(imprecise) state sequence estimation



$Q_1\ (X_1)$

$Q_2\ (X_2|X_1)$

$Q_2\ (X_3|X_2)$

?

?

?

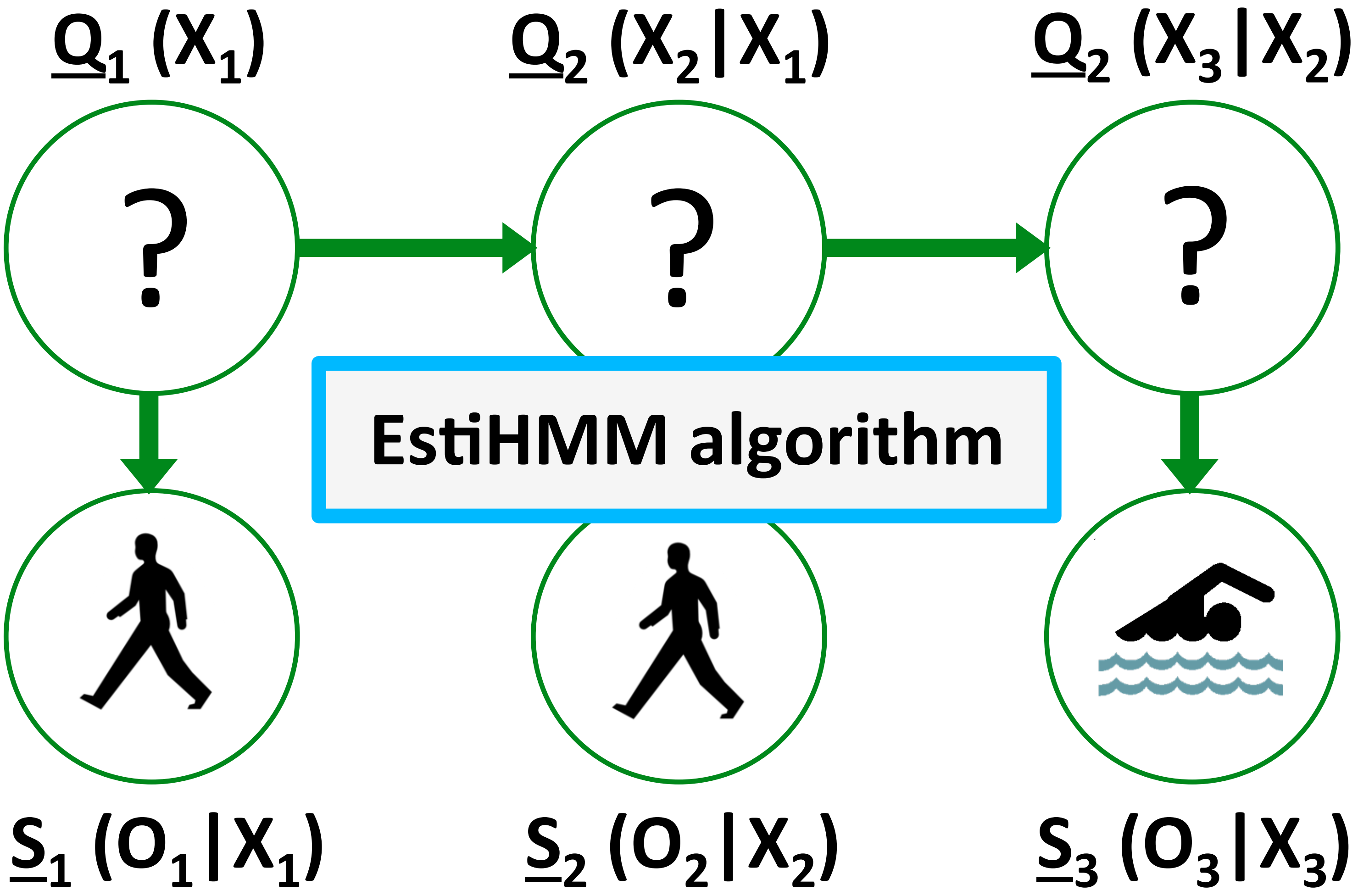
Viterbi algorithm

$S_1\ (O_1|X_1)$

$S_2\ (O_2|X_2)$

$S_3\ (O_3|X_3)$

(imprecise) state sequence estimation



$\underline{Q}_1(X_1)$ $\underline{Q}_2(X_2|X_1)$ $\underline{Q}_2(X_3|X_2)$

?

?

?

EstiHMM algorithm

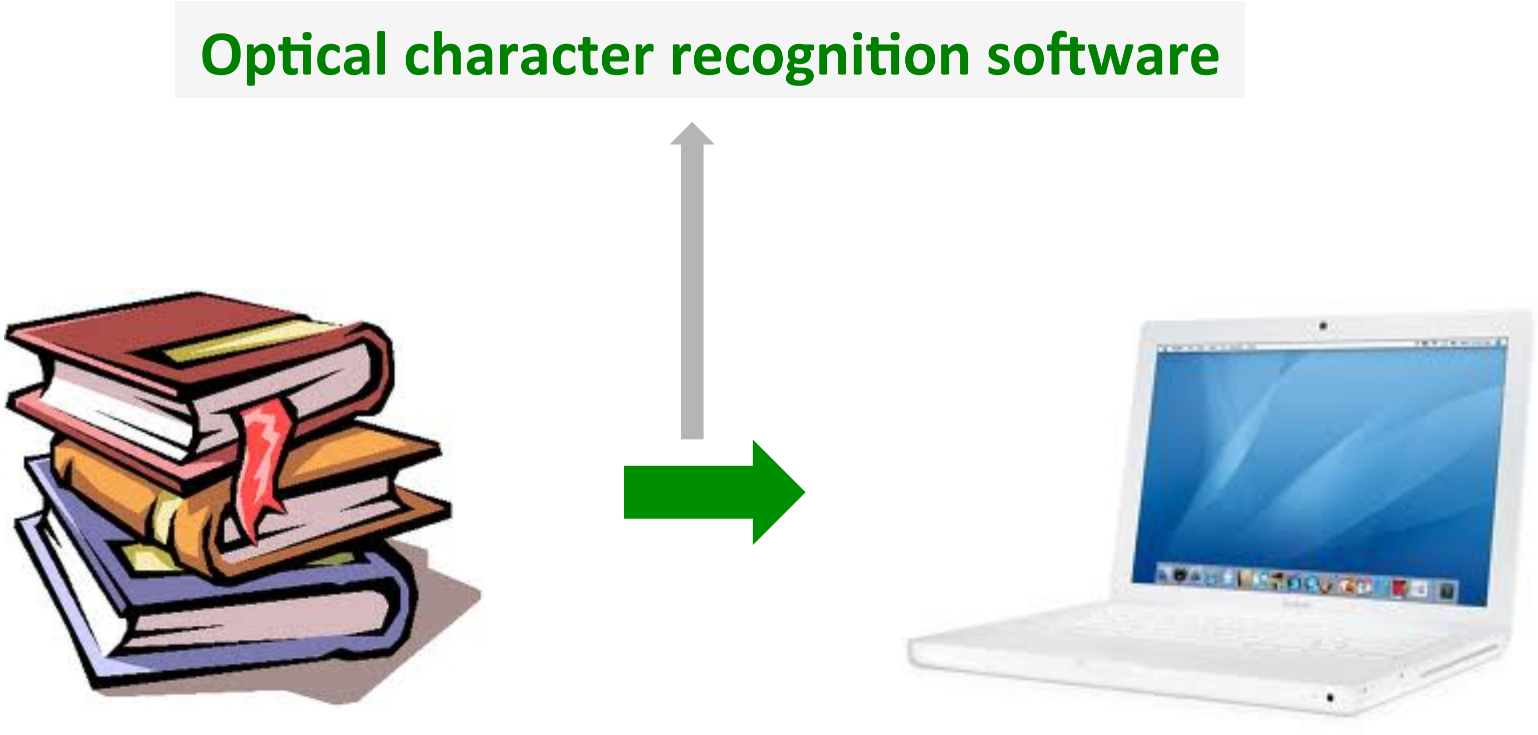
$\underline{S}_1(O_1|X_1)$ $\underline{S}_2(O_2|X_2)$ $\underline{S}_3(O_3|X_3)$

APPLICATIONS ?

Applications of state sequence estimation

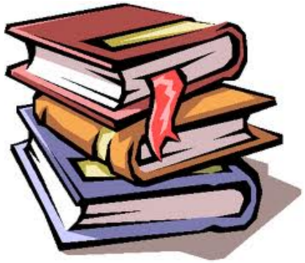
- Speech recognition
- Bio-informatics
  - Finding CpG-islands
  - Locating introns and exons
- Grammatical tagging
- **OCR postprocessing**
- …

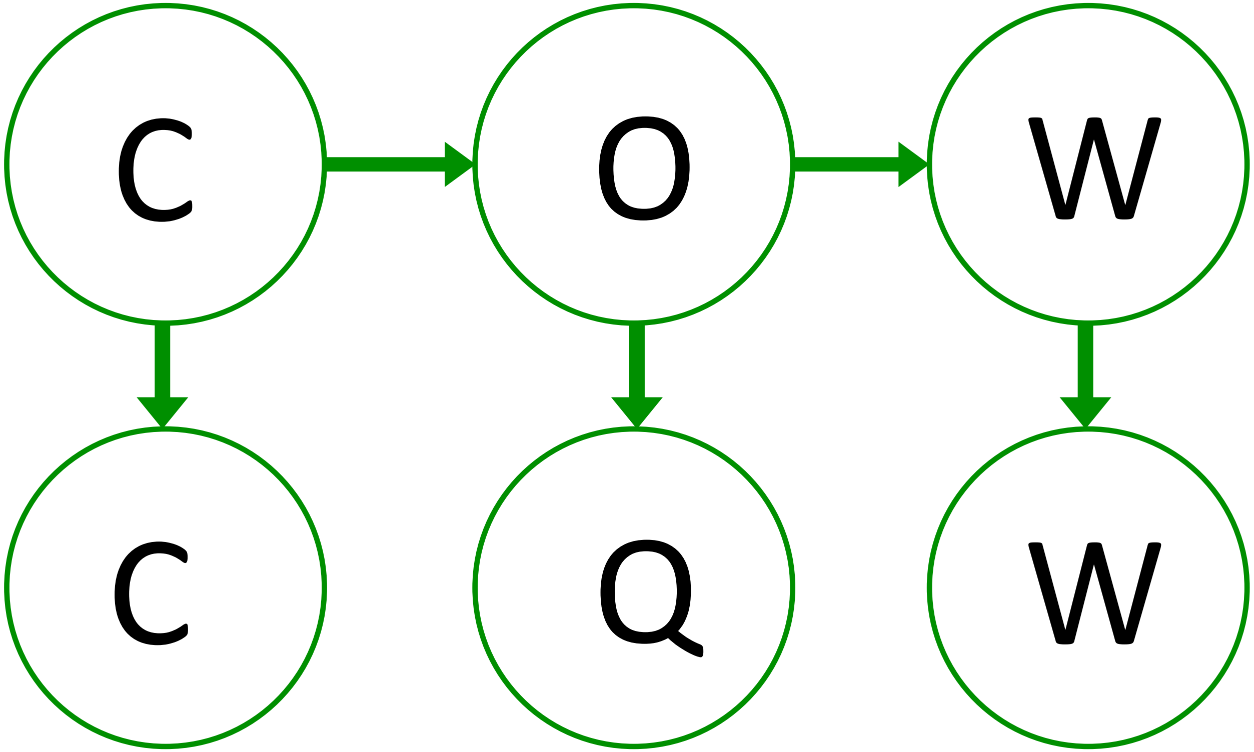
OCR postprocessing



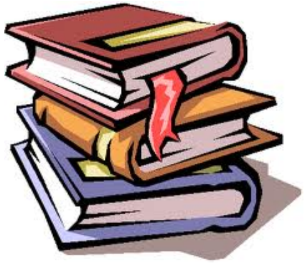
Optical character recognition software

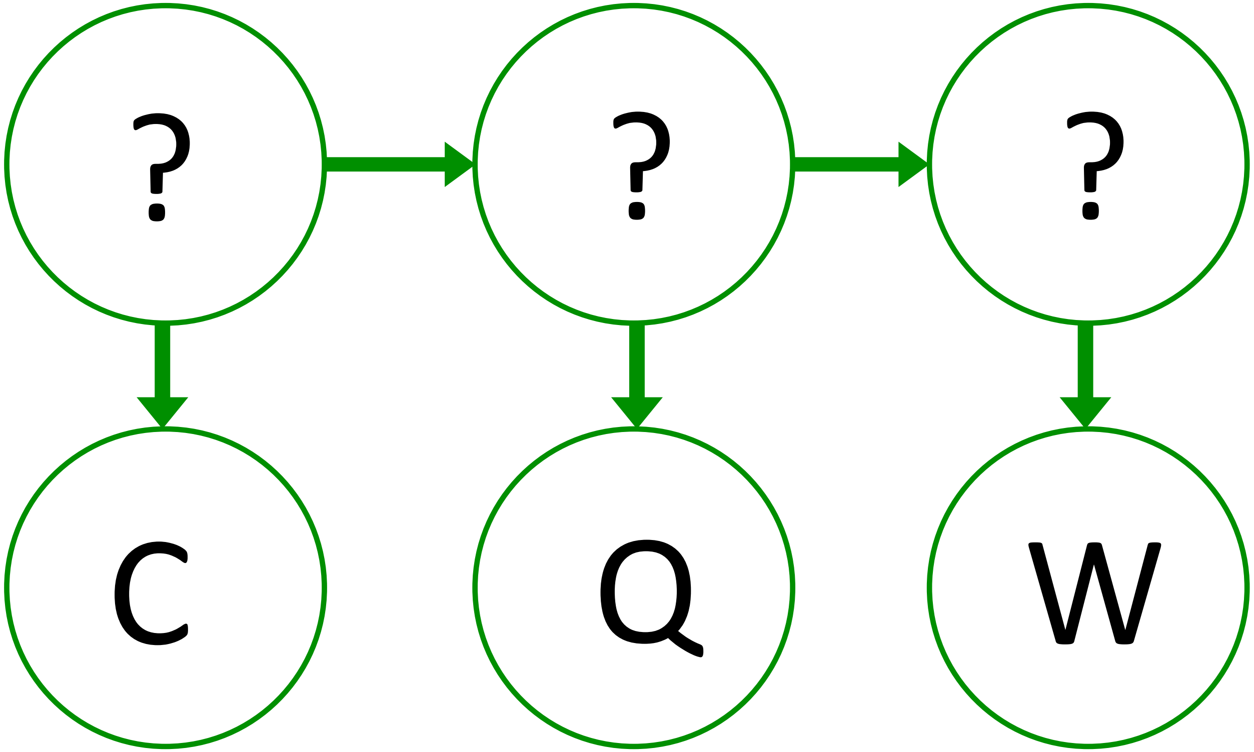
OCR postprocessing





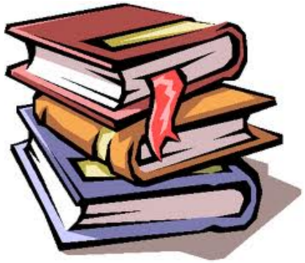
OCR postprocessing

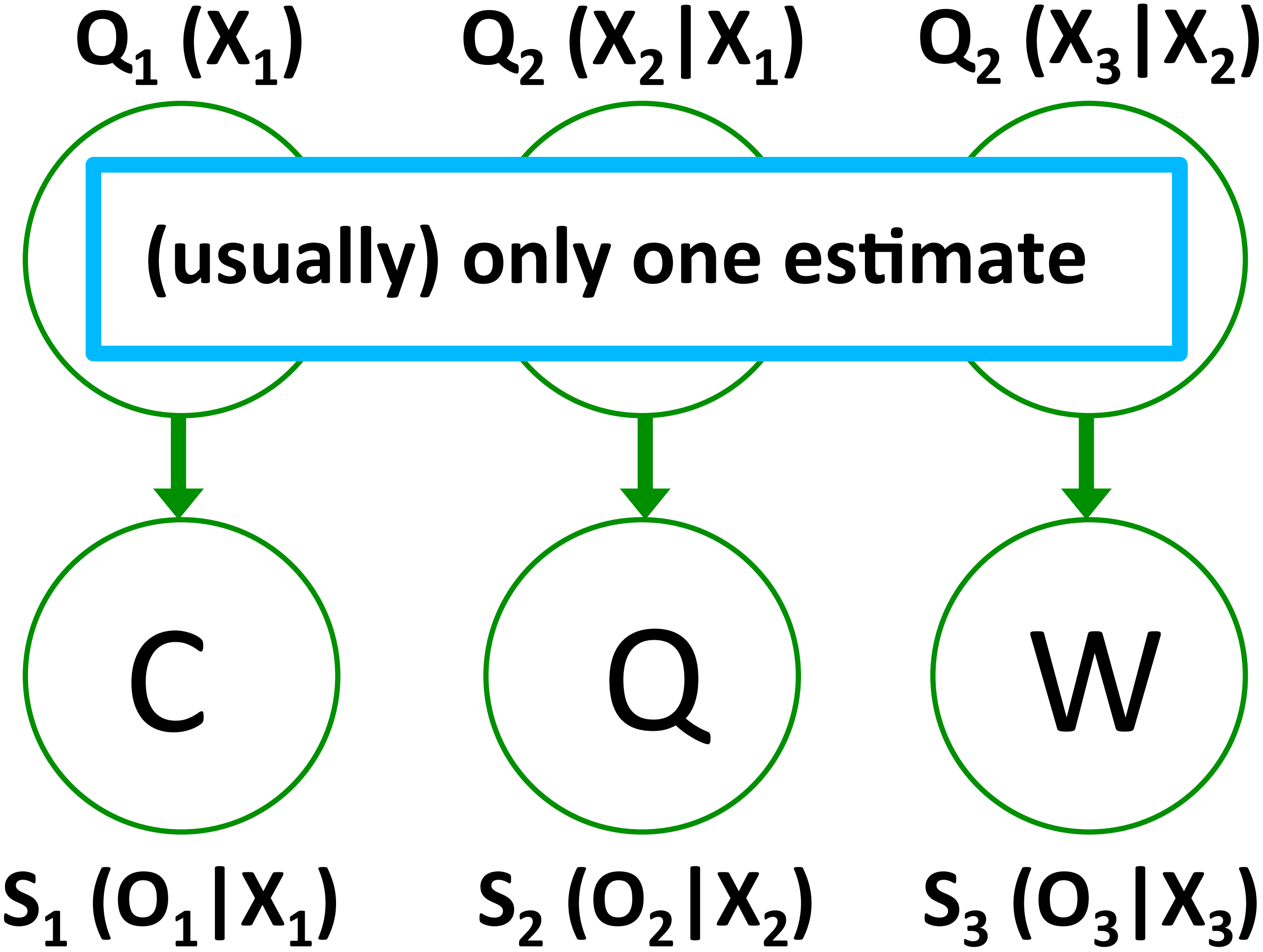




OCR postprocessing

Viterbi





$Q_1\ (X_1)$

$Q_2\ (X_2|X_1)$

$Q_2\ (X_3|X_2)$

(usually) only one estimate

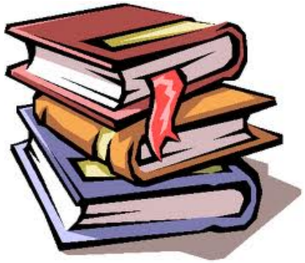
$S_1\ (O_1|X_1)$

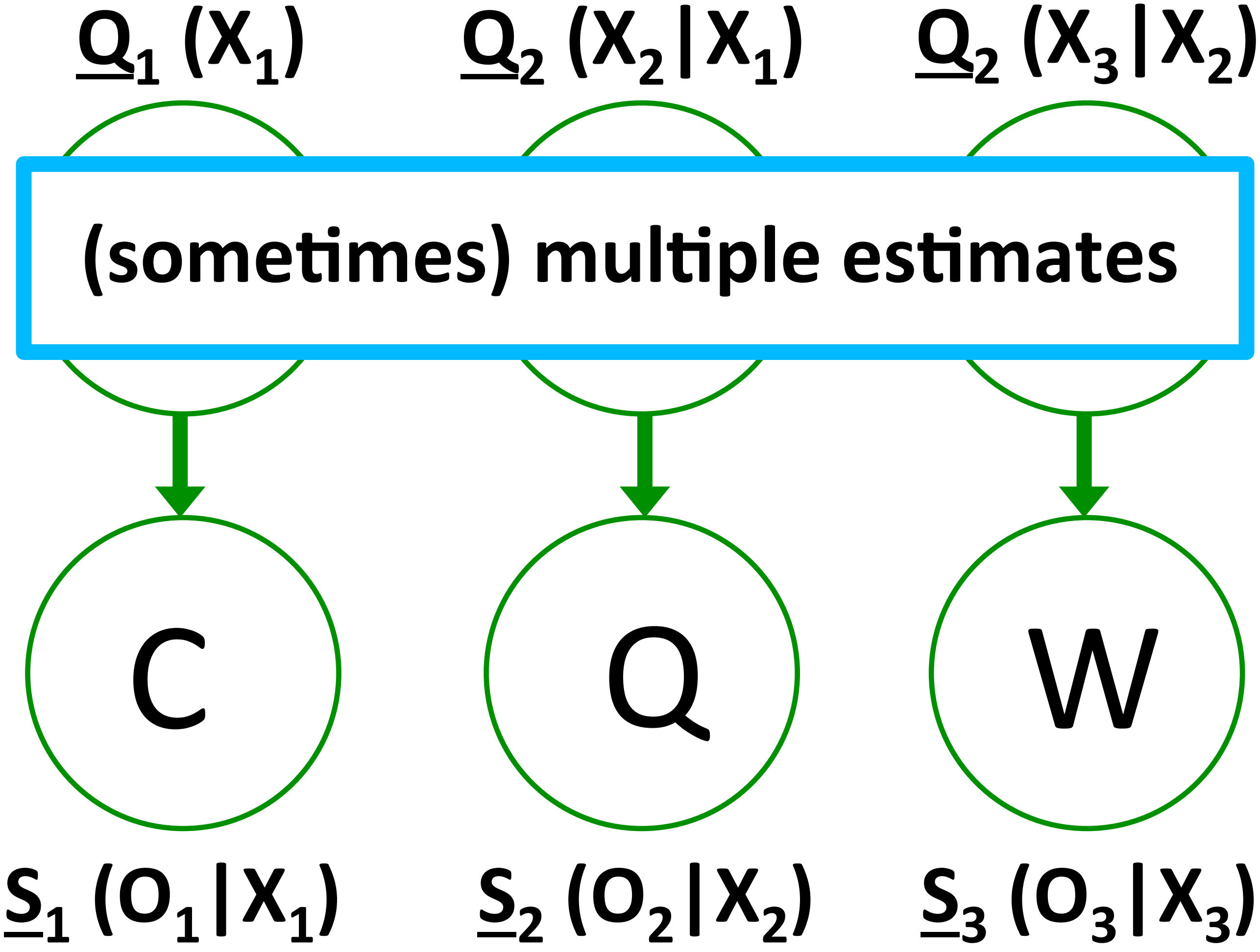
$S_2\ (O_2|X_2)$

$S_3\ (O_3|X_3)$

OCR postprocessing

EstiHMM





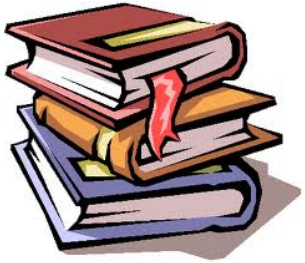
$\underline{Q}_1\ (X_1)$ $\underline{Q}_2\ (X_2|X_1)$ $\underline{Q}_2\ (X_3|X_2)$

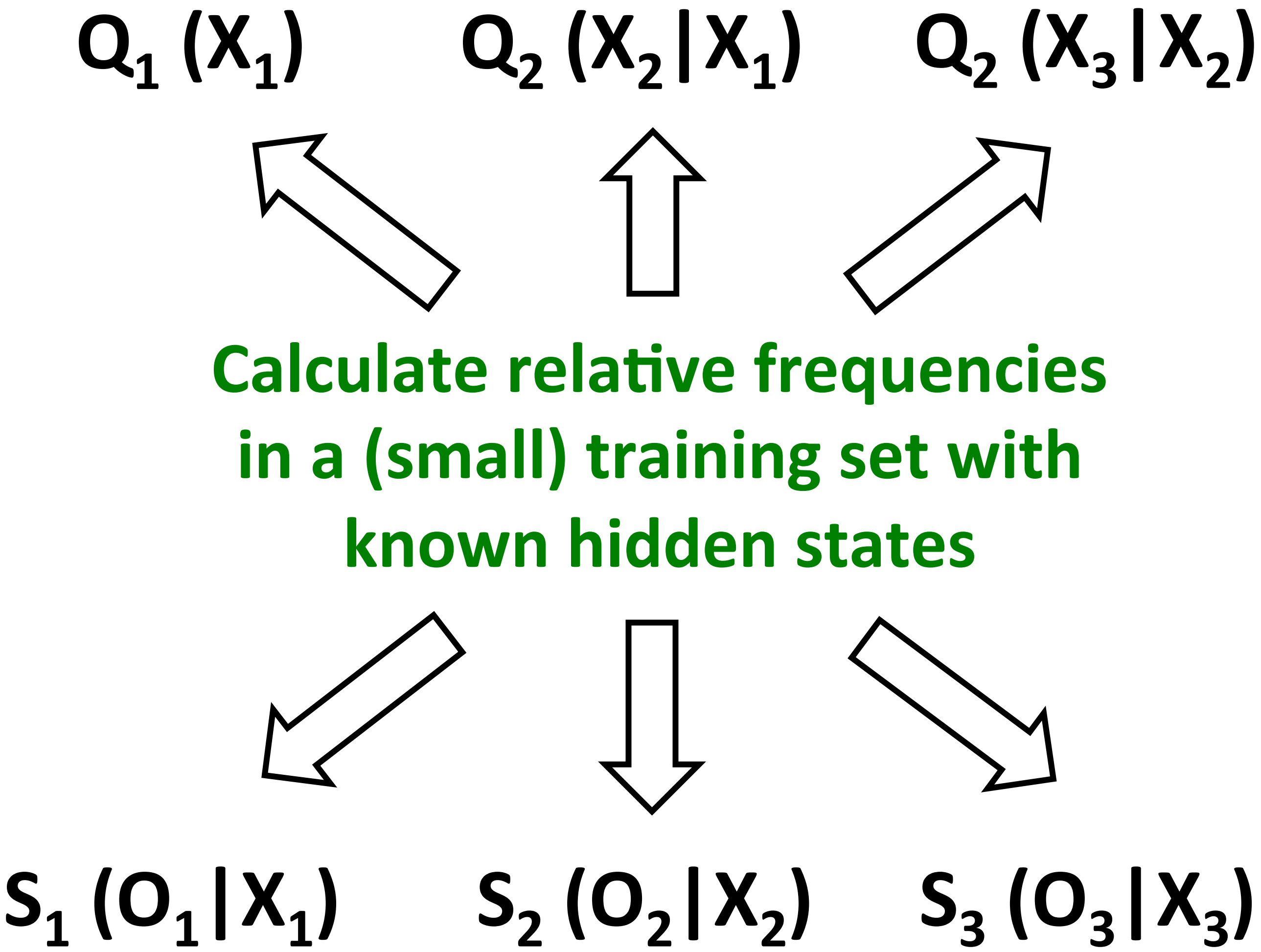
(sometimes) multiple estimates

$\underline{S}_1\ (O_1|X_1)$ $\underline{S}_2\ (O_2|X_2)$ $\underline{S}_3\ (O_3|X_3)$

OCR postprocessing

Viterbi





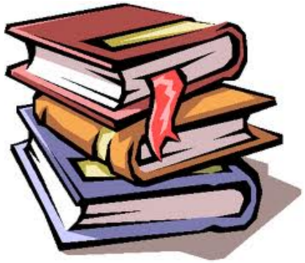
$Q_1\ (X_1)$ $Q_2\ (X_2|X_1)$ $Q_2\ (X_3|X_2)$

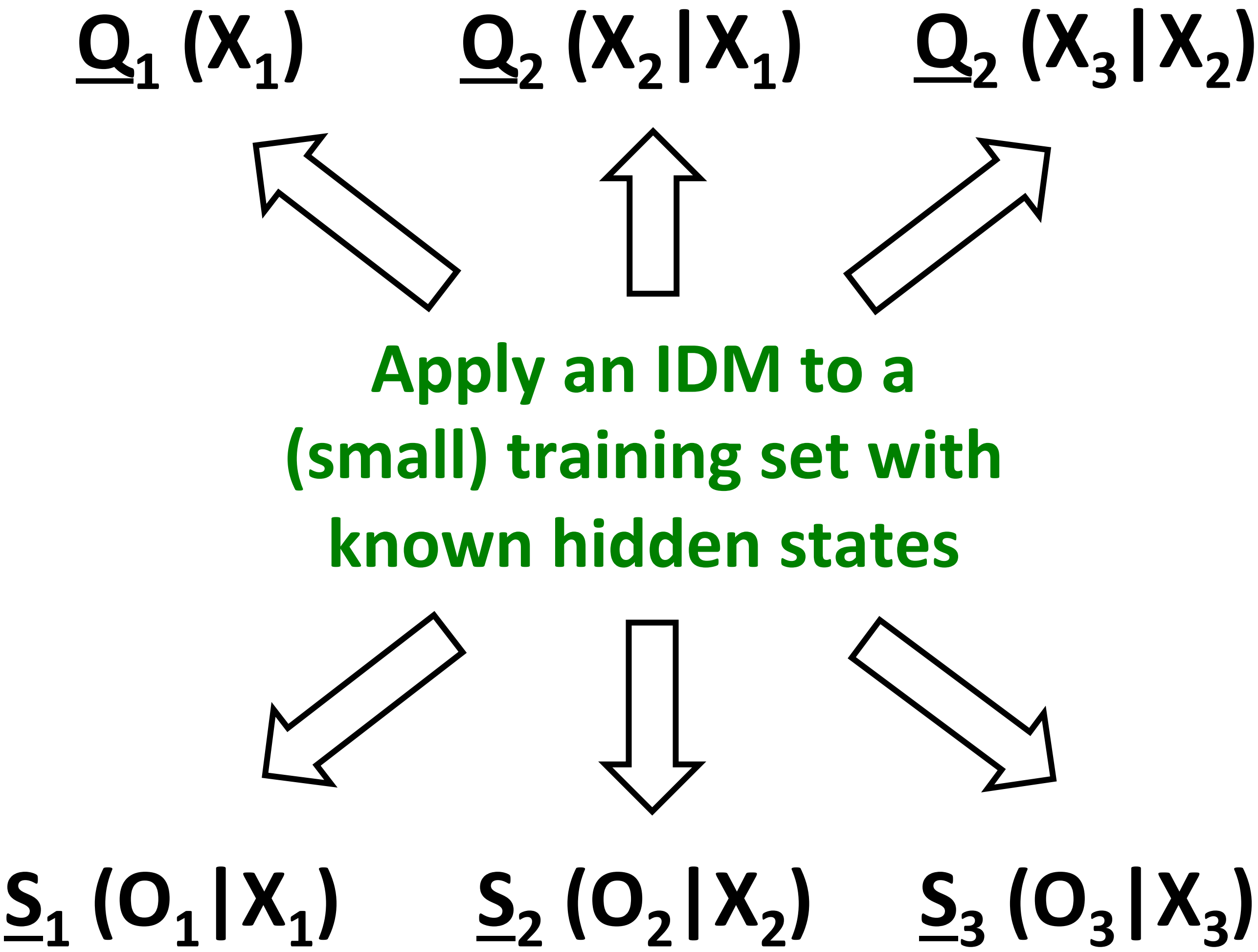
Calculate relative frequencies in a (small) training set with known hidden states

$S_1\ (O_1|X_1)$ $S_2\ (O_2|X_2)$ $S_3\ (O_3|X_3)$

OCR postprocessing

EstiHMM





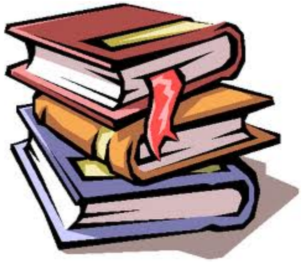
$\underline{Q}_1 (X_1)$ $\underline{Q}_2 (X_2|X_1)$ $\underline{Q}_2 (X_3|X_2)$

Apply an IDM to a (small) training set with known hidden states

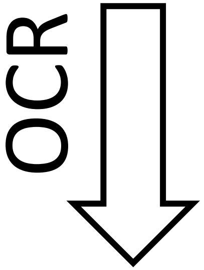
$\underline{S}_1 (O_1|X_1)$ $\underline{S}_2 (O_2|X_2)$ $\underline{S}_3 (O_3|X_3)$

OCR postprocessing

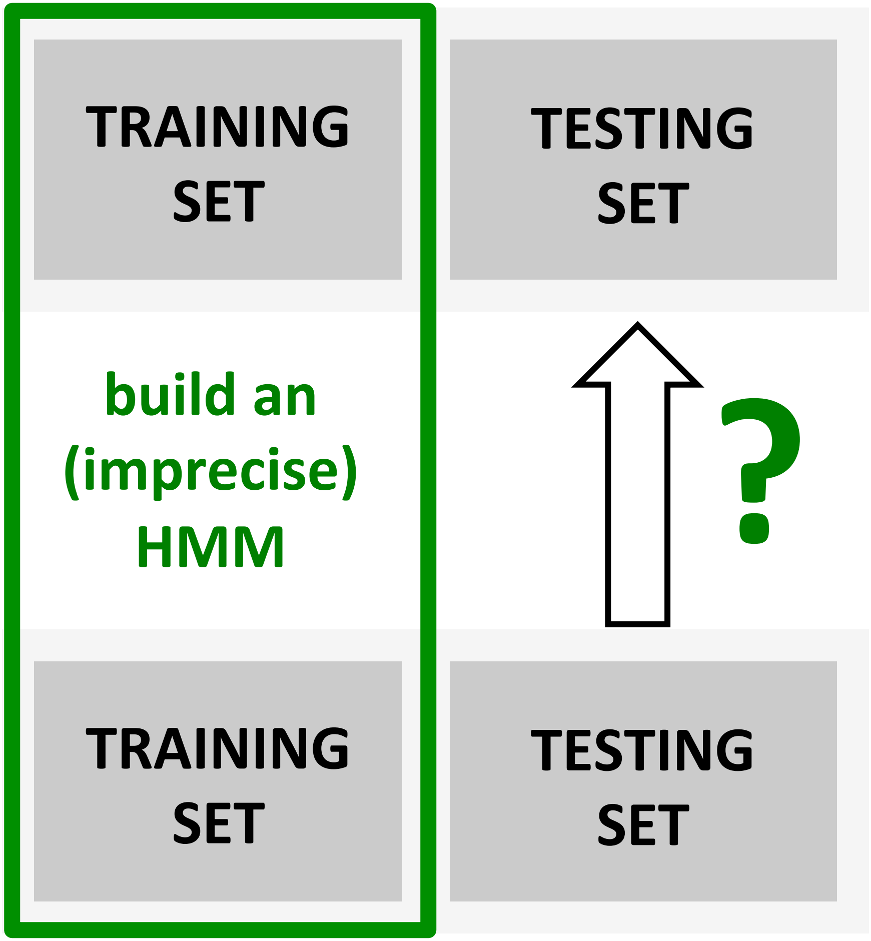
La Divina Commedia



ORIGINAL WORDS IN THE BOOK



CORRESPONDING WORDS IN TEXT DOCUMENT



TRAINING SET

TESTING SET

build an (imprecise) HMM

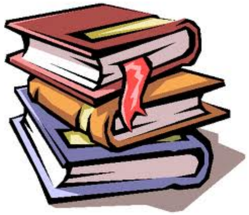
?

TRAINING SET

TESTING SET

OCR postprocessing

La Divina Commedia



original

VITA

correctly read

digital

VITA

Solution Viterbi

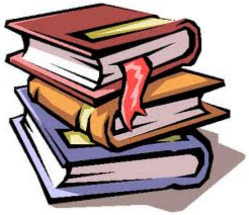
VITA

Solution(s) EstiHMM-algoritme

VITA

OCR postprocessing

La Divina Commedia



original

CON

incorrectly read

digital

CCN

Solution Viterbi

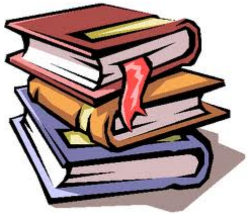
CON

Solution(s) EstiHMM-algoritme

CON

OCR postprocessing

La Divina Commedia



original

EH

correctly read

digital

EH

Solution Viterbi

EN

Solution(s) EstiHMM-algoritme

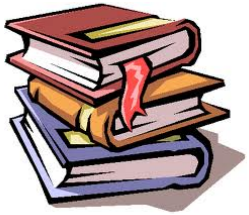
CH

EH

EN

OCR postprocessing

La Divina Commedia



original

IO

incorrectly read

digital

ZO

Solution Viterbi

LO

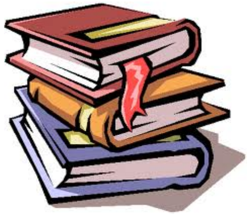
Solution(s) EstiHMM-algoritme

LO

IO

OCR postprocessing

La Divina Commedia



original

CHE

incorrectly read

digital

CNE

Solution Viterbi

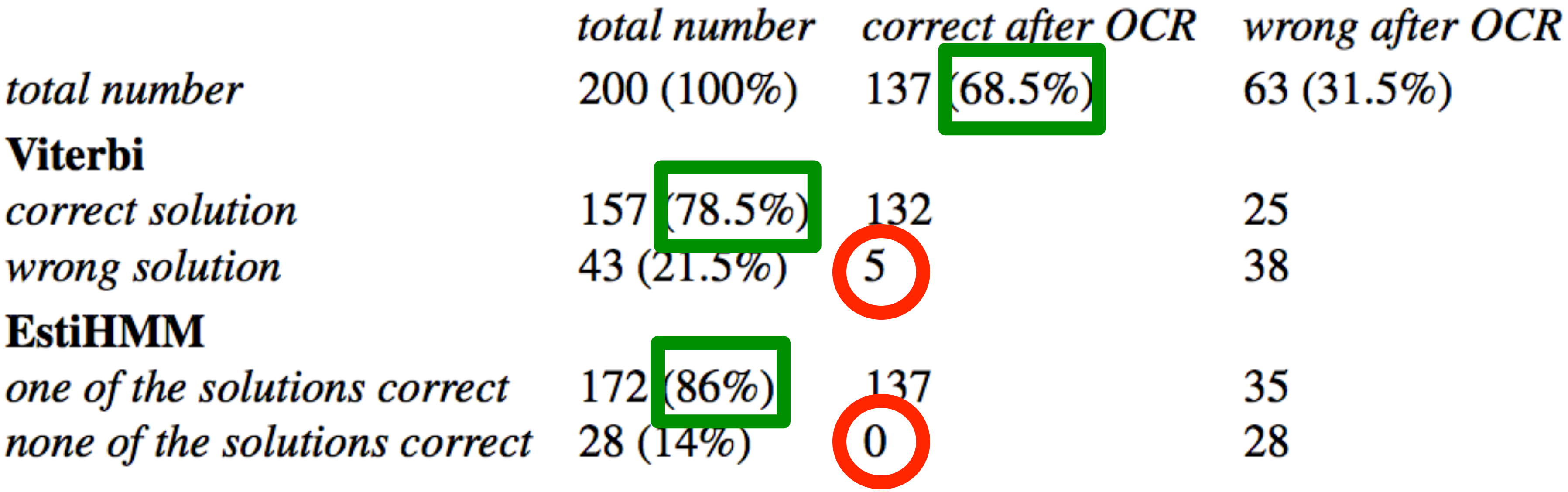
ONE

Solution(s) EstiHMM-algoritme

CBE CHE
CNE CZE
ONE

OCR postprocessing

La Divina Commedia

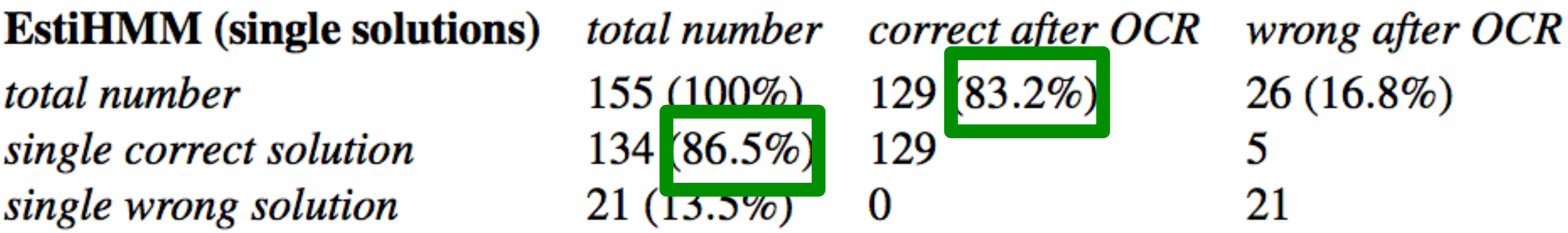


	total number	*correct after OCR*	*wrong after OCR*
total number	200 (100%)	137 (68.5%)	63 (31.5%)
Viterbi			
correct solution	157 (78.5%)	132	25
wrong solution	43 (21.5%)	5	38
EstiHMM			
one of the solutions correct	172 (86%)	137	35
none of the solutions correct	28 (14%)	0	28

- **Both algorithms are able to detect and correct errors**
- The EstiHMM algorithm (in this case) does not introduce errors in words that were already correct
- **EstiHMM** sometimes **returns multiple solutions** and therefore (of course) includes the correct solution more often

OCR postprocessing

La Divina Commedia

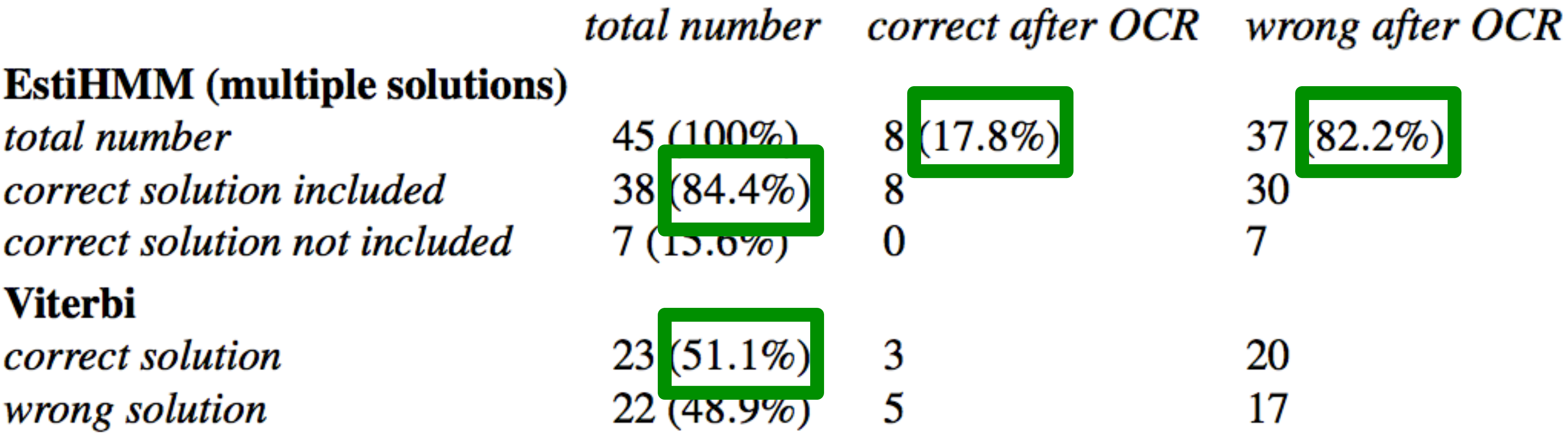


EstiHMM (single solutions)	*total number*	*correct after OCR*	*wrong after OCR*
total number	155 (100%)	129 (83.2%)	26 (16.8%)
single correct solution	134 (86.5%)	129	5
single wrong solution	21 (13.5%)	0	21

- If the EstiHMM algorithm gives a **single solution**, it will be **identical to the solution given by the Viterbi algorithm**
- EstiHMM giving **a single solution serves as an indication** that
  - the word we are applying it to does not contain **errors**
  - the result returned by the **Viterbi algorithm is correct**

OCR postprocessing

La Divina Commedia



	total number	*correct after OCR*	*wrong after OCR*
EstiHMM (multiple solutions)			
total number	45 (100%)	8 (17.8%)	37 (82.2%)
correct solution included	38 (84.4%)	8	30
correct solution not included	7 (15.6%)	0	7
Viterbi			
correct solution	23 (51.1%)	3	20
wrong solution	22 (48.9%)	5	17

- EstiHMM giving **multiple solutions serves as an indication** that
  - the word we are applying it to does indeed contain **errors**
  - the result returned by the **Viterbi algorithm is less reliable**
- EstiHMM can be used to **robustify the precise result** given by the Viterbi algorithm

HOW CAN THIS BE USEFUL ?

How can undecisiveness be useful?

- As a method of picking out the hard problems, which you then try to solve with more expensive or time-consuming methods **(solve easy cases automatically and use experts only for the difficult ones!)**
- If not deciding is a useful choice too, because making a wrong decision is dangerous or expensive **(choosing between specific and general medication)**

Thanks for your attention!