Bayesian Networks in Educational Testing

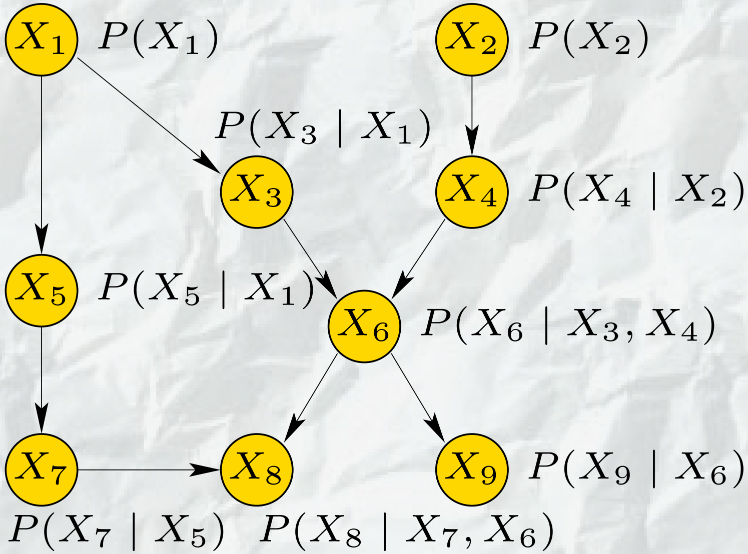
Jiří Vomlel

Institute of Information Theory and Automation
Academy of Science of the Czech Republic
Prague

This presentation is available at:
`http://www.utia.cas.cz/vomlel/`

Contents:

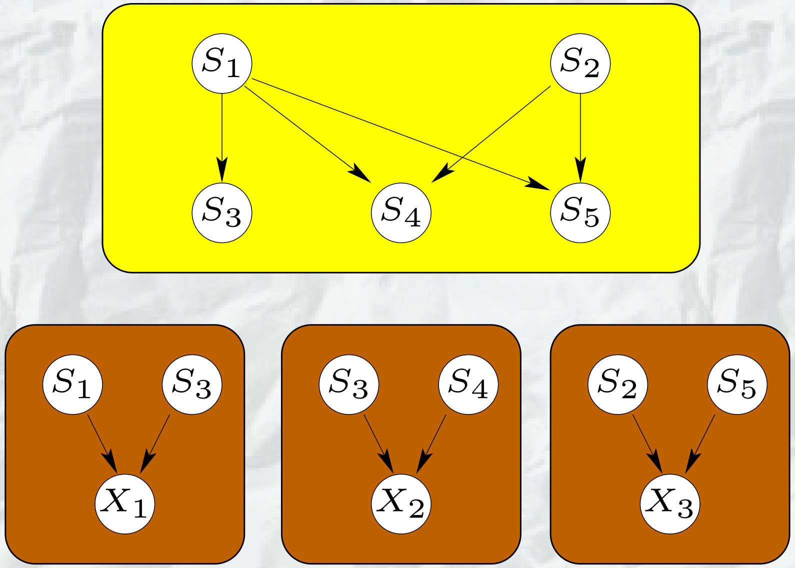
Bayesian network



$$
\begin{aligned}
P(X_1, \ldots, X_9) \; = & \\
= \quad & P(X_9|X_8, \ldots, X_1) \cdot P(X_8|X_7, \ldots, X_1) \cdot \ldots \cdot P(X_2|X_1) \cdot P(X_1) \\
= \quad & P(X_9|X_6) \cdot P(X_8|X_7, X_6) \cdot P(X_7|X_5) \cdot P(X_6|X_4, X_3) \\
& \cdot P(X_5|X_1) \cdot P(X_4|X_2) \cdot P(X_3|X_1) \cdot P(X_2) \cdot P(X_1)
\end{aligned}
$$

Student and evidence models

(R. Almond and R. Mislevy, 1999)



Sudents solving problems with fractions

- A group of university students from Aalborg University prepared paper tests that were given to students at Brønderslev High School.
- Four elementary skills, four operational skills, and abilities to apply operational skills to complex tasks were tested.
- 149 students solved the test.
- The tests were analyzed in detail so that it was possible for most of students to decide whether they have or have not the tested skills.
- Several different models were learned using the PC algorithm
- Models were compared on how well they predict skills using the leave-one-out method.

Examples of tasks - operations with fractions

T_1: $\left(\frac{3}{4} \cdot \frac{5}{6}\right) - \frac{1}{8}$ = $\frac{15}{24} - \frac{1}{8} = \frac{5}{8} - \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$

T_2: $\frac{1}{6} + \frac{1}{12}$ = $\frac{2}{12} + \frac{1}{12} = \frac{3}{12} = \frac{1}{4}$

T_3: $\frac{1}{4} \cdot 1\frac{1}{2}$ = $\frac{1}{4} \cdot \frac{3}{2} = \frac{3}{8}$

T_4: $\left(\frac{1}{2} \cdot \frac{1}{2}\right) \cdot \left(\frac{1}{3} + \frac{1}{3}\right)$ = $\frac{1}{4} \cdot \frac{2}{3} = \frac{2}{12} = \frac{1}{6}$.

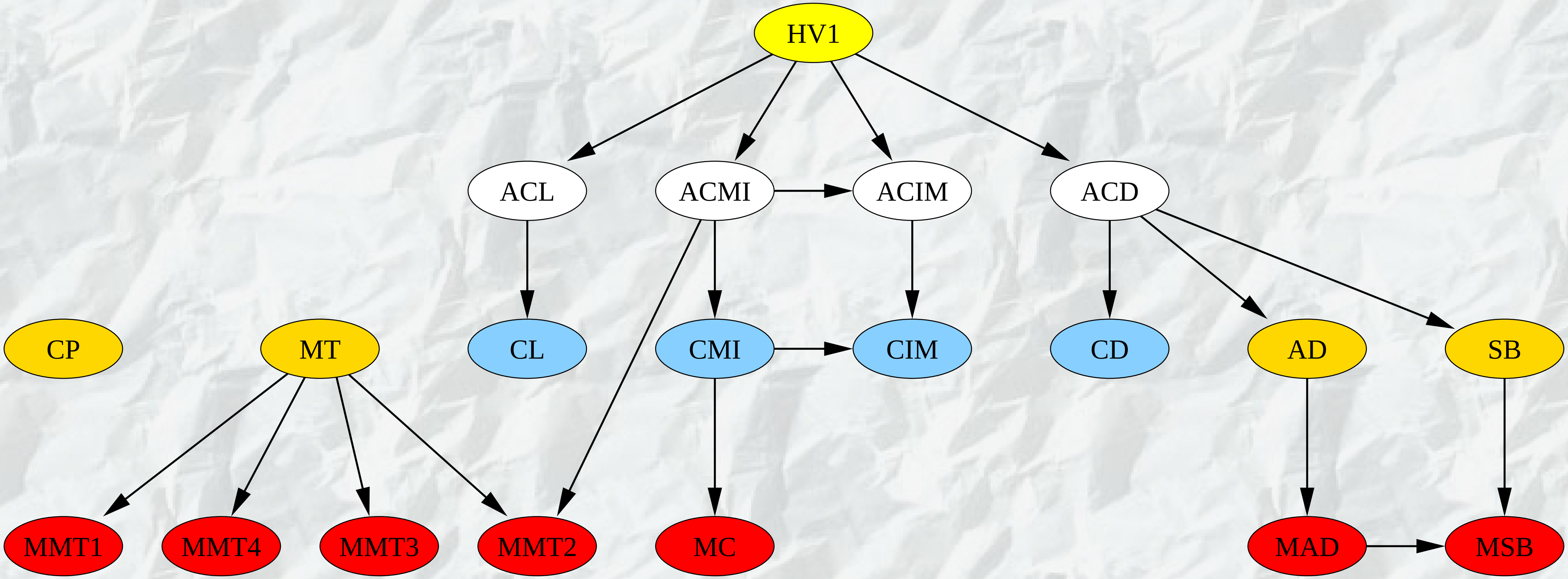
Elementary and operational skills

CP	Comparison (common numerator or denominator)	$\frac{1}{2} > \frac{1}{3}, \quad \frac{2}{3} > \frac{1}{3}$
AD	Addition (comm. denom.)	$\frac{1}{7} + \frac{2}{7} = \frac{1+2}{7} = \frac{3}{7}$
SB	Subtract. (comm. denom.)	$\frac{2}{5} - \frac{1}{5} = \frac{2-1}{5} = \frac{1}{5}$
MT	Multiplication	$\frac{1}{2} \cdot \frac{3}{5} = \frac{3}{10}$
CD	Common denominator	$\left(\frac{1}{2}, \frac{2}{3}\right) = \left(\frac{3}{6}, \frac{4}{6}\right)$
CL	Cancelling out	$\frac{4}{6} = \frac{2 \cdot 2}{2 \cdot 3} = \frac{2}{3}$
CIM	Conv. to mixed numbers	$\frac{7}{2} = \frac{3 \cdot 2+1}{2} = 3\frac{1}{2}$
CMI	Conv. to improp. fractions	$3\frac{1}{2} = \frac{3 \cdot 2+1}{2} = \frac{7}{2}$

Misconceptions

Label	Description	Occurrence
MAD	$\frac{a}{b} + \frac{c}{d} = \frac{a+c}{b+d}$	14.8%
MSB	$\frac{a}{b} - \frac{c}{d} = \frac{a-c}{b-d}$	9.4%
MMT1	$\frac{a}{b} \cdot \frac{c}{b} = \frac{a \cdot c}{b}$	14.1%
MMT2	$\frac{a}{b} \cdot \frac{c}{b} = \frac{a+c}{b \cdot b}$	8.1%
MMT3	$\frac{a}{b} \cdot \frac{c}{d} = \frac{a \cdot d}{b \cdot c}$	15.4%
MMT4	$\frac{a}{b} \cdot \frac{c}{d} = \frac{a \cdot c}{b+d}$	8.1%
MC	$a\frac{b}{c} = \frac{a \cdot b}{c}$	4.0%

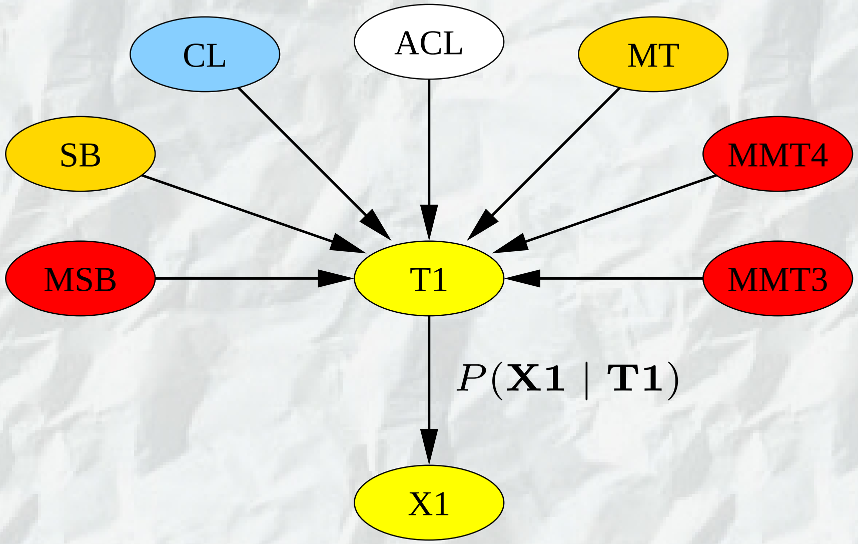
Student model



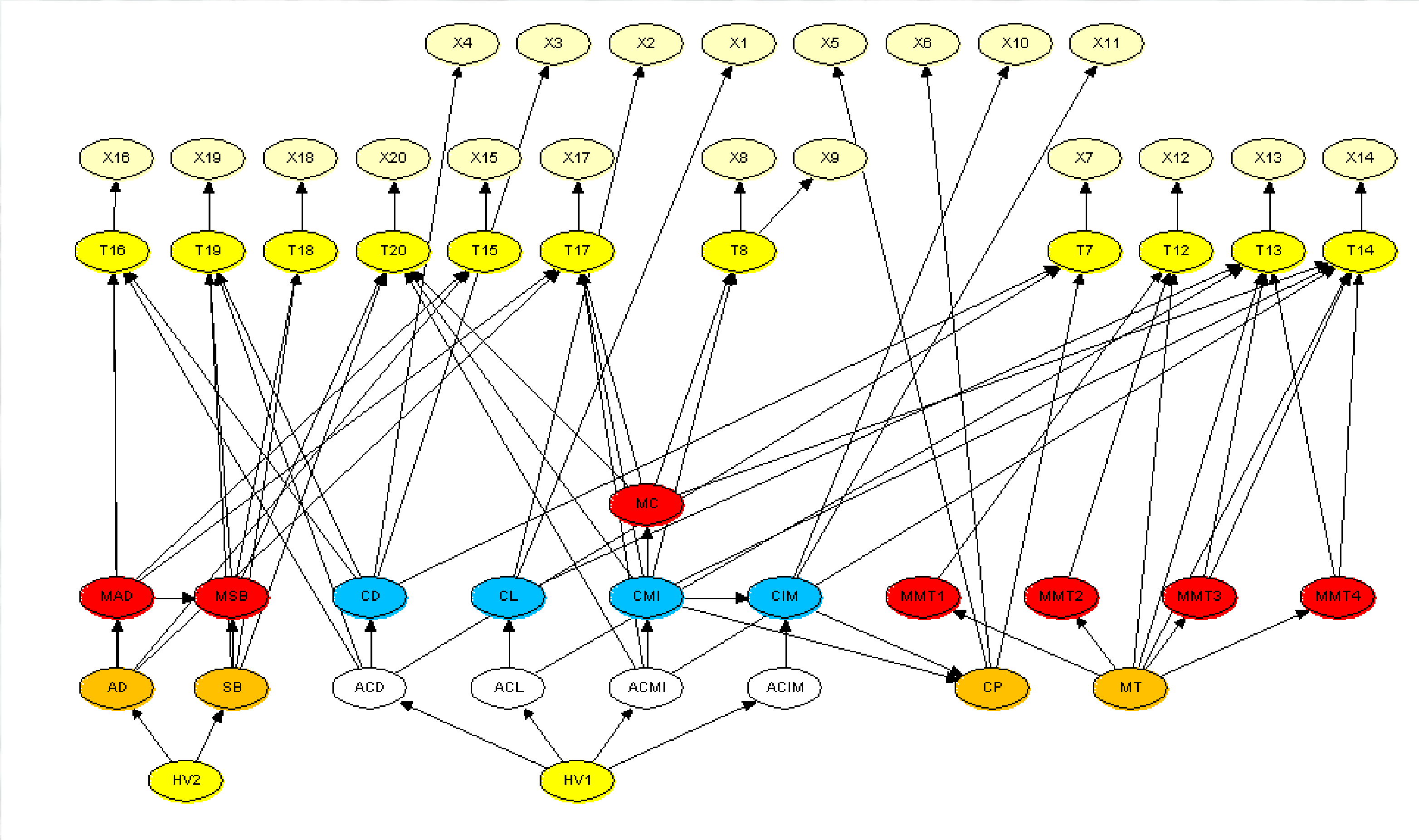
Evidence model for task $T1$

$$\left(\frac{3}{4}\cdot\frac{5}{6}\right)-\frac{1}{8}=\frac{15}{24}-\frac{1}{8}=\frac{5}{8}-\frac{1}{8}=\frac{4}{8}=\frac{1}{2}$$

$$T1 \quad \Leftrightarrow \quad MT\ \&\ CL\ \&\ ACL\ \&\ SB\ \&\ \neg MMT3\ \&\ \neg MMT4\ \&\ \neg MSB$$



The overal model



See the model in Hugin:

http://www.hugin.com

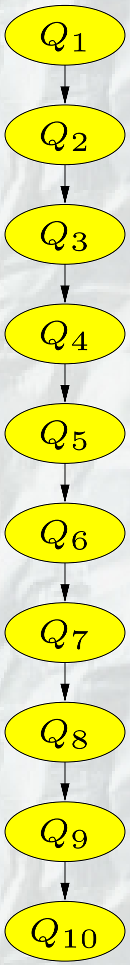
Tested models

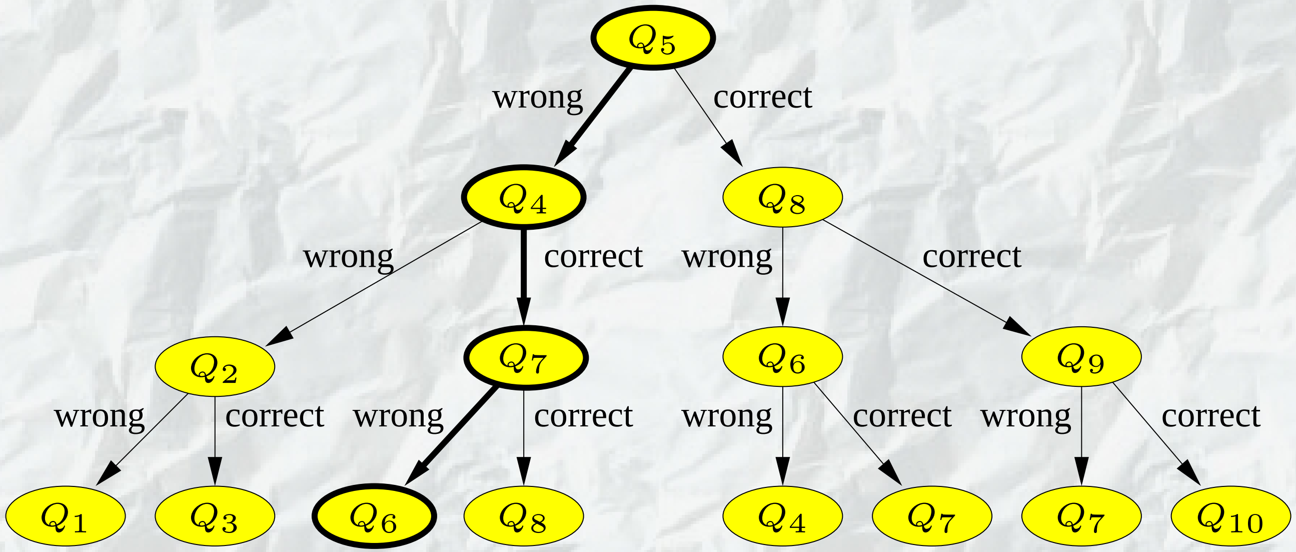
(a)	two hidden variables with several restrictions on presence or absence of edges,
(b)	**two hidden variables and only few restrictions,**
(c)	one hidden variable and few restrictions,
(d)	no hidden variable and few restrictions,
(e)	no hidden variable and only obvious logical constraints as restrictions,
(f)	no hidden variable and independent skills,
(g)	Naïve Bayes model with one hidden variable (with two states) being parent of all skills,
(h)	as above, but the hidden variable has three states,

Comparison of different models using t values

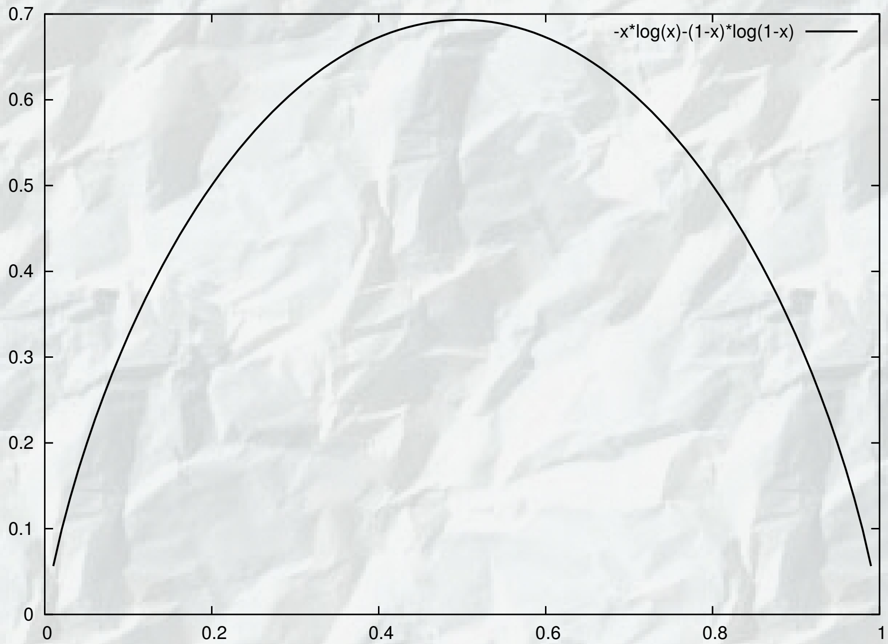
	vs. (b)	vs. (c)	vs. (d)	vs. (e)	vs. (f)	vs. (g)	vs. (h)	total
(a)	$\mathbf{-3.293}$	$\mathbf{-2.650}$	$\mathbf{-2.567}$	0.308	$\mathbf{2.323}$	-1.318	$\mathbf{-2.192}$	$\mathbf{-3}$
(b)		0.983	0.578	$\mathbf{3.579}$	$\mathbf{5.296}$	1.734	0.759	$\mathbf{+3}$
(c)			-0.265	$\mathbf{2.612}$	$\mathbf{4.488}$	1.154	-0.004	$\mathbf{+3}$
(d)				$\mathbf{2.839}$	$\mathbf{4.487}$	1.327	0.091	$\mathbf{+3}$
(e)					$\mathbf{3.837}$	$\mathbf{-1.977}$	$\mathbf{-2.739}$	$\mathbf{-4}$
(f)						$\mathbf{-4.071}$	$\mathbf{-4.728}$	$\mathbf{-7}$
(g)							-1.641	$\mathbf{+2}$
(h)								$\mathbf{+3}$

Fixed Test vs. Adaptive Test





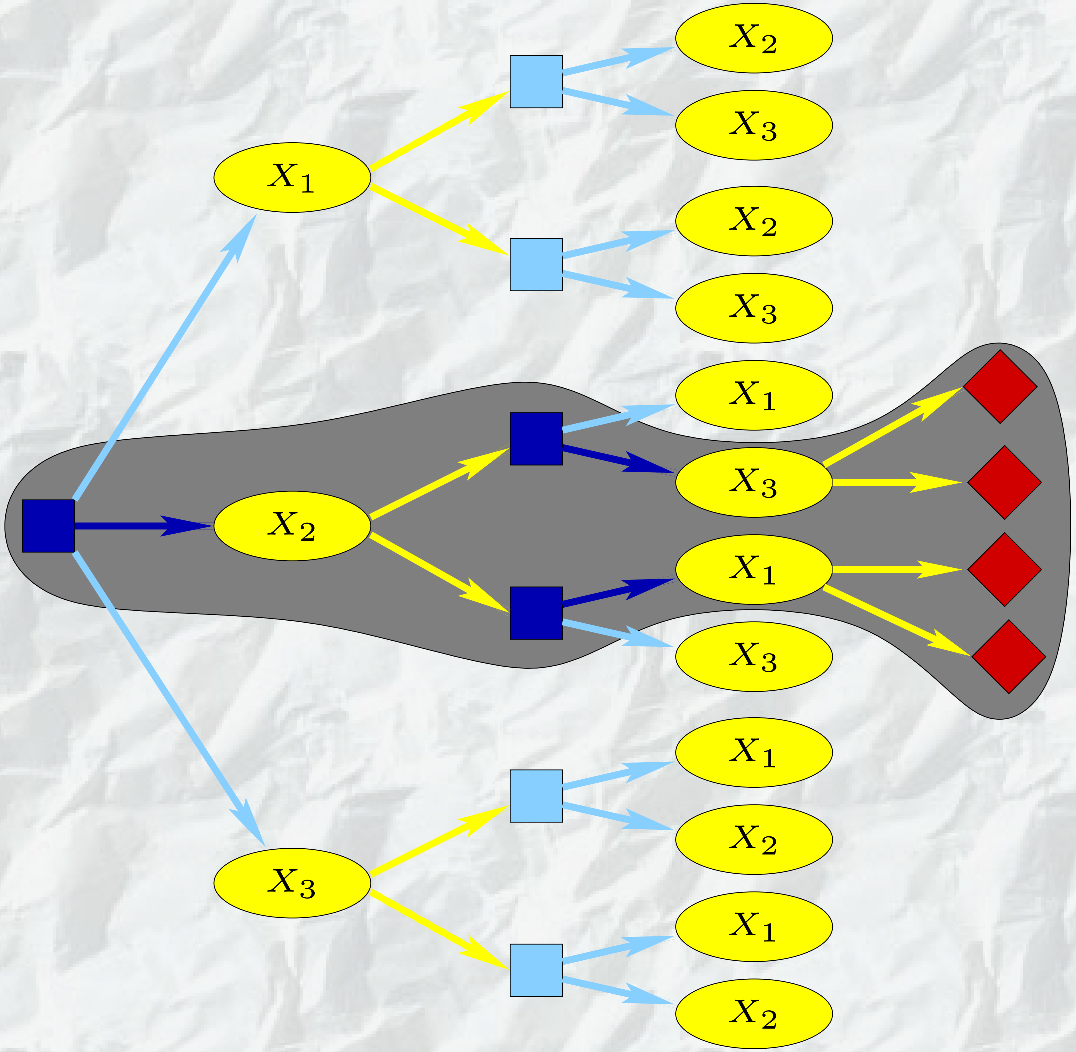
Entropy as an information criteria



Entropy of probability distribution $P(\mathbf{S})$ on skills $\mathbf{S}$ is defined as

$$H\left(P(\mathbf{S})\right) \;=\; -\sum_{\mathbf{s}} P(\mathbf{S}=\mathbf{s}) \cdot \log P(\mathbf{S}=\mathbf{s})$$

“The lower the entropy the more we know about a student.”



Entropy in node n

$$H(\mathbf{e}_n) \quad = \quad H(P(\mathbf{S} \mid \mathbf{e}_n))$$

Expected entropy at the end of test $\mathbf{t}$

$$E_H(\mathbf{t}) \quad = \quad \sum_{\ell \in \mathcal{L}(\mathbf{t})} P(\mathbf{e}_\ell) \cdot H(\mathbf{e}_\ell)$$

$\mathcal{T}$ ... the set of all possible tests (e.g. of a given length)

A test $\mathbf{t}^\star$ is **optimal** iff

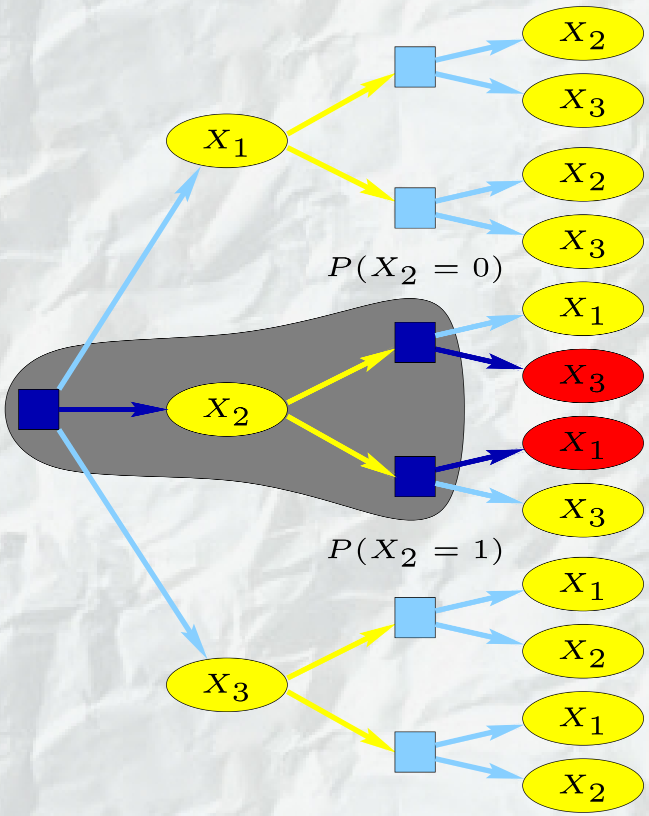
$$\mathbf{t}^\star \quad = \quad \arg\min_{\mathbf{t} \in \mathcal{T}} E_H(\mathbf{t}) \quad .$$

A **myopically optimal** test $\mathbf{t}$ is a test where each question $X^\star$ of $\mathbf{t}$ minimizes the expected value of entropy after the question is answered:

$$X^\star = \arg\min_{X \in \mathcal{X}} E_H(\mathbf{t}_{\downarrow X}) \ ,$$

i.e. it works as if the test finished after the selected question $X^\star$.

Myopic construction of a fixed test



$$\begin{aligned} e\_list &= \{\{X_2 = 0\}, \{X_2 = 1\}\} \\ counts[3] &= P(X_2 = 0) \;=\; 0.7 \\ counts[1] &= P(X_2 = 1) \;=\; 0.3 \end{aligned}$$

$e\_list := [\emptyset]$;

$test := [\,]$;

for $i := 1$ to $|\mathcal{X}|$ do $counts[i] := 0$;

for $position := 1$ to $test\_lenght$ do

  $new\_e\_list := [\,]$;

  for all $\mathbf{e} \in e\_list$ do

    $i := most\_informative\_X(\mathbf{e})$;

    $counts[i] := counts[i] + P(\mathbf{e})$;

    for all $x_i \in \mathbb{X}_i$ do

      $append(new\_e\_list, \{\mathbf{e} \cup \{X_i = x_i\}\})$;

  $e\_list := new\_e\_list$;

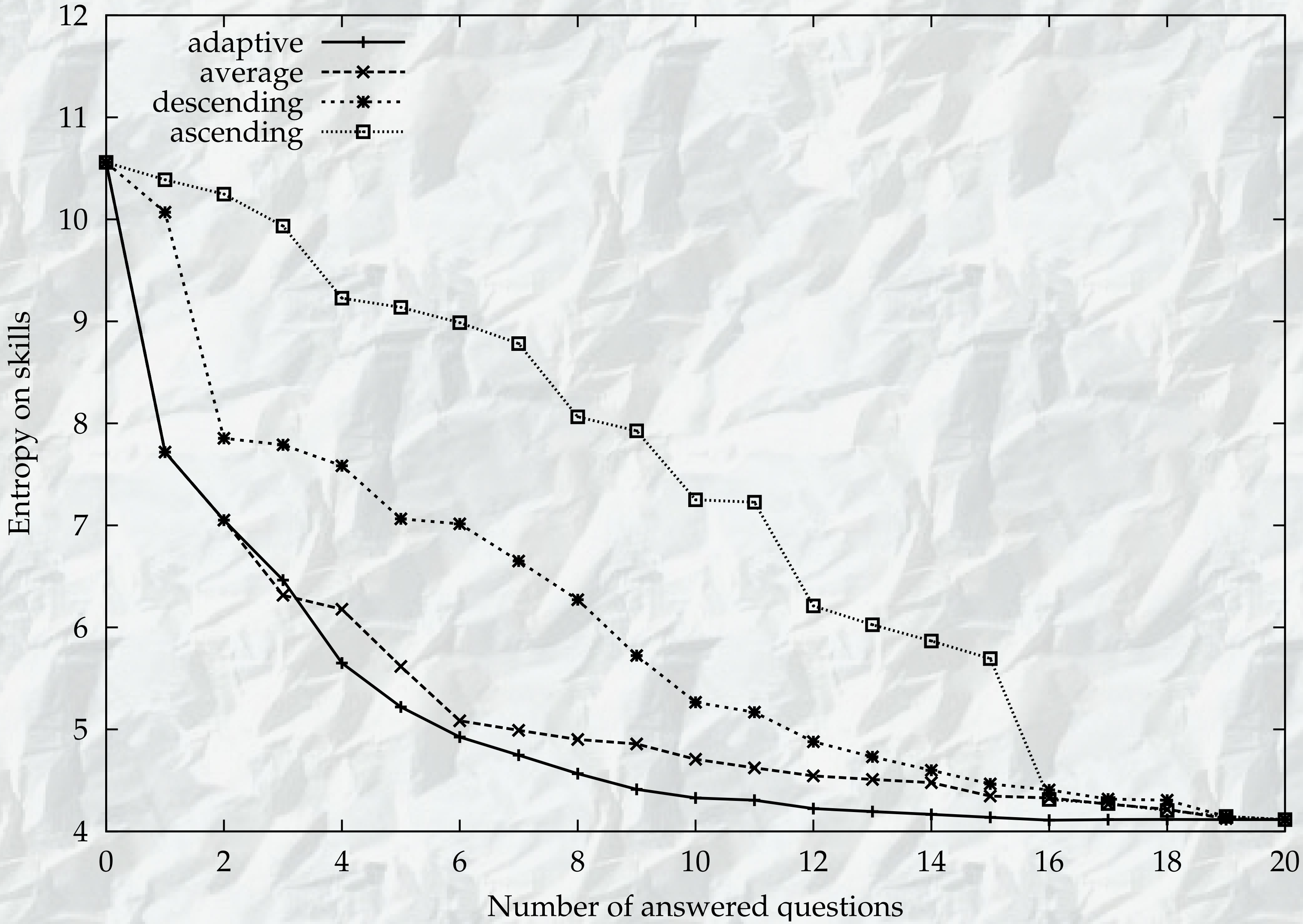
  $i^\star := \arg\max_i \; counts[i]$;

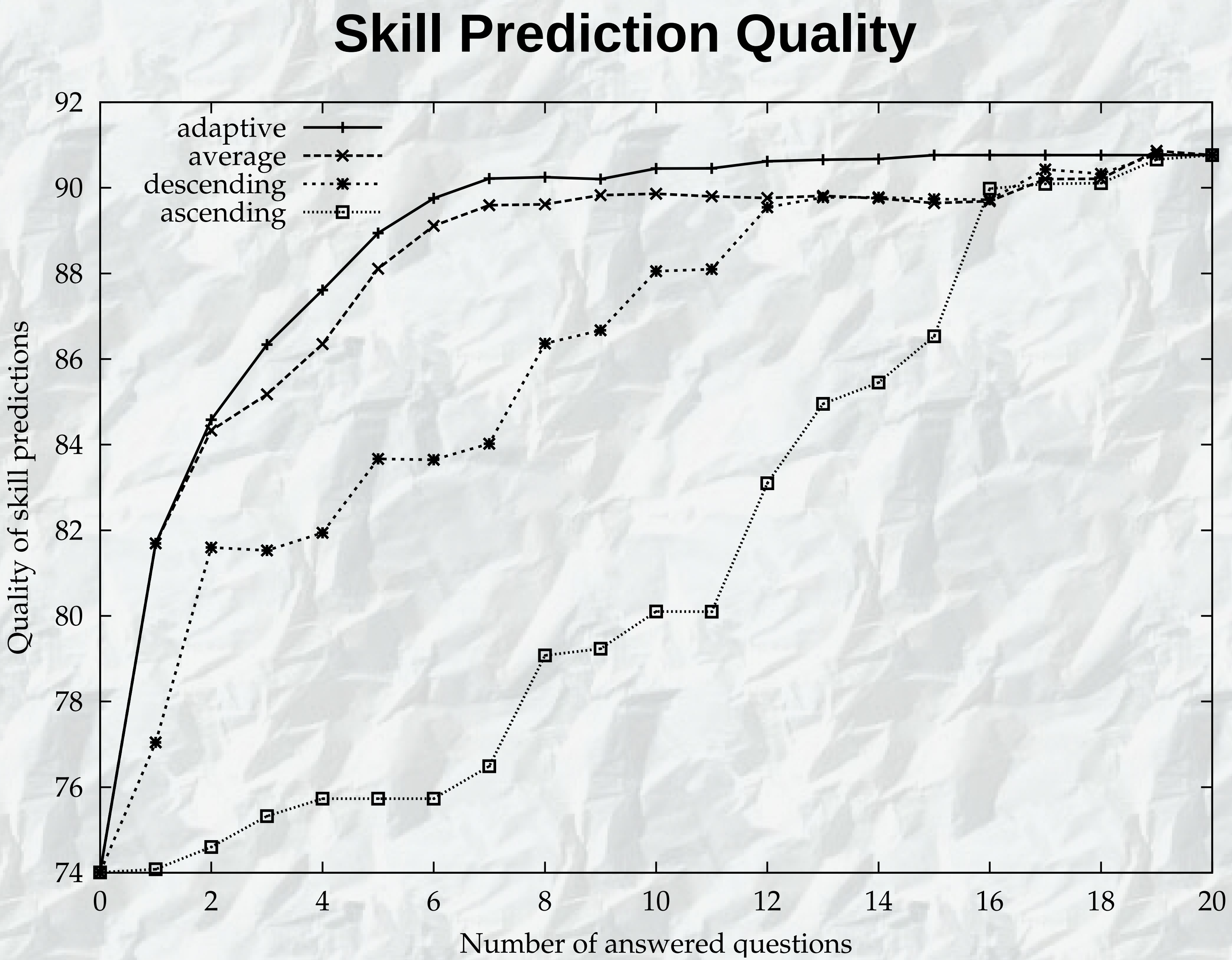
  $append(test, X_{i^\star})$;

  $counts[i^\star] := 0$;

$return(test)$;

Entropy of the probability distributions on the skills





Skill Prediction Quality
adaptive
average
descending
ascending
Quality of skill predictions
Number of answered questions
92
90
88
86
84
82
80
78
76
74
0
2
4
6
8
10
12
14
16
18
20

Conclusions

- Empirical evidence shows that educational testing can benefit from application of Bayesian networks. Adaptive tests may substantially **reduce the number of questions** that are necessary to be asked.
- **Method for** the design of a **fixed test** provided good results on tested data. It may be regarded as a good cheap alternative to computerized adaptive tests when they are not suitable.
- One theoretical problem related to application of Bayesian networks to educational testing is **efficient inference exploiting deterministic relations** in the model. This problem is a topic of our current research.