Bayesian Networks in Educational Testing

Jiří Vomlel

Laboratory for Intelligent Systems Prague
University of Economics

This presentation is available at:
`http://www.utia.cas.cz/vomlel/slides/lisp2002.pdf`

Contents:

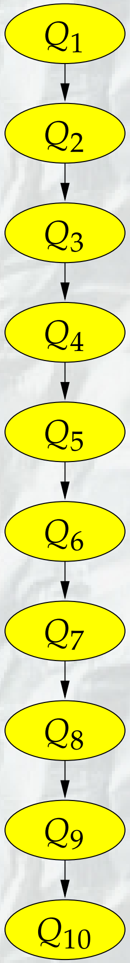
- Educational testing is a “big business”.
- What is a fixed test and an adaptive test?
- An example: a test of basic operations with fractions.
- Optimal and myopically optimal tests.
- Construction of a myopically optimal fixed test.
- Results of experiments.
- Ane example showing that modeling dependence between skills is important.
- Conclusions.

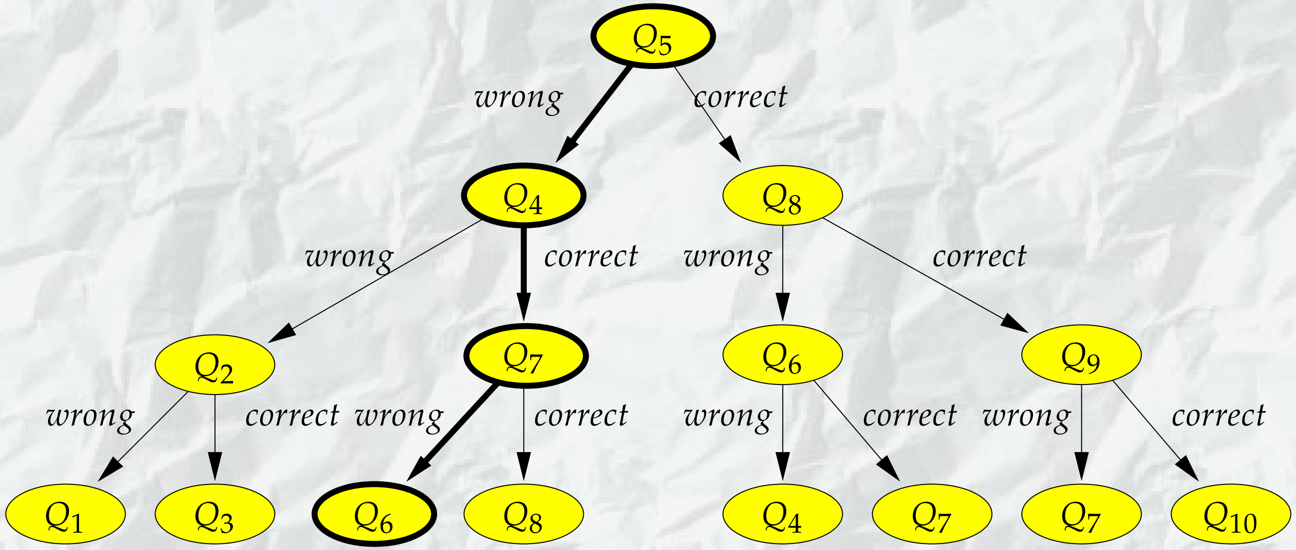
Educational Testing Service (ETS)

- Educational Testing Service is the world's largest private educational testing organization with 2,300 regular employees.
- Volumes for ETS's Largest Exams in 2000-2001:

3,185,000	SAT I Reasoning Test and SAT II: Subject Area Tests (the SAT test is the standard college admission test in US)
2,293,000	PSAT: Preliminary SAT/National Merit Scholarship Qualifying Test
1,421,000	AP: Advanced Placement Program
801,000	The Praxis Series: Professional Assessments for Beginning Teachers and Pre-Professional Skills Tests
787,000	TOEFL: Test of English as a Foreign Language
449,000	GRE: Graduate Record Examinations General Test
	etc.

Fixed Test vs. Adaptive Test





Computerized Adaptive Testing (CAT)

Objective:	**An optimal test for each examinee**
Two basic steps:	(1) examinee's knowledge level is estimated
	(2) questions appropriate for the level are selected.

R. Almond and R. Mislevy from **ETS** proposed to use graphical models in CAT.

- one **student model** (relations between skills, abilities, etc.)
- several **evidence models**, one for each task or question.

CAT for basic operations with fractions

Examples of tasks:

$$
\begin{aligned}
T_1: \quad & \left(\frac{3}{4} \cdot \frac{5}{6}\right) - \frac{1}{8} & = \quad & \frac{15}{24} - \frac{1}{8} = \frac{5}{8} - \frac{1}{8} = \frac{4}{8} = \frac{1}{2} \\
T_2: \quad & \frac{1}{6} + \frac{1}{12} & = \quad & \frac{2}{12} + \frac{1}{12} = \frac{3}{12} = \frac{1}{4} \\
T_3: \quad & \frac{1}{4} \cdot 1\frac{1}{2} & = \quad & \frac{1}{4} \cdot \frac{3}{2} = \frac{3}{8} \\
T_4: \quad & \left(\frac{1}{2} \cdot \frac{1}{2}\right) \cdot \left(\frac{1}{3} + \frac{1}{3}\right) & = \quad & \frac{1}{4} \cdot \frac{2}{3} = \frac{2}{12} = \frac{1}{6} \quad .
\end{aligned}
$$

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	$3\frac{1}{2} = \frac{3 \cdot 2 + 1}{2} = \frac{7}{2}$

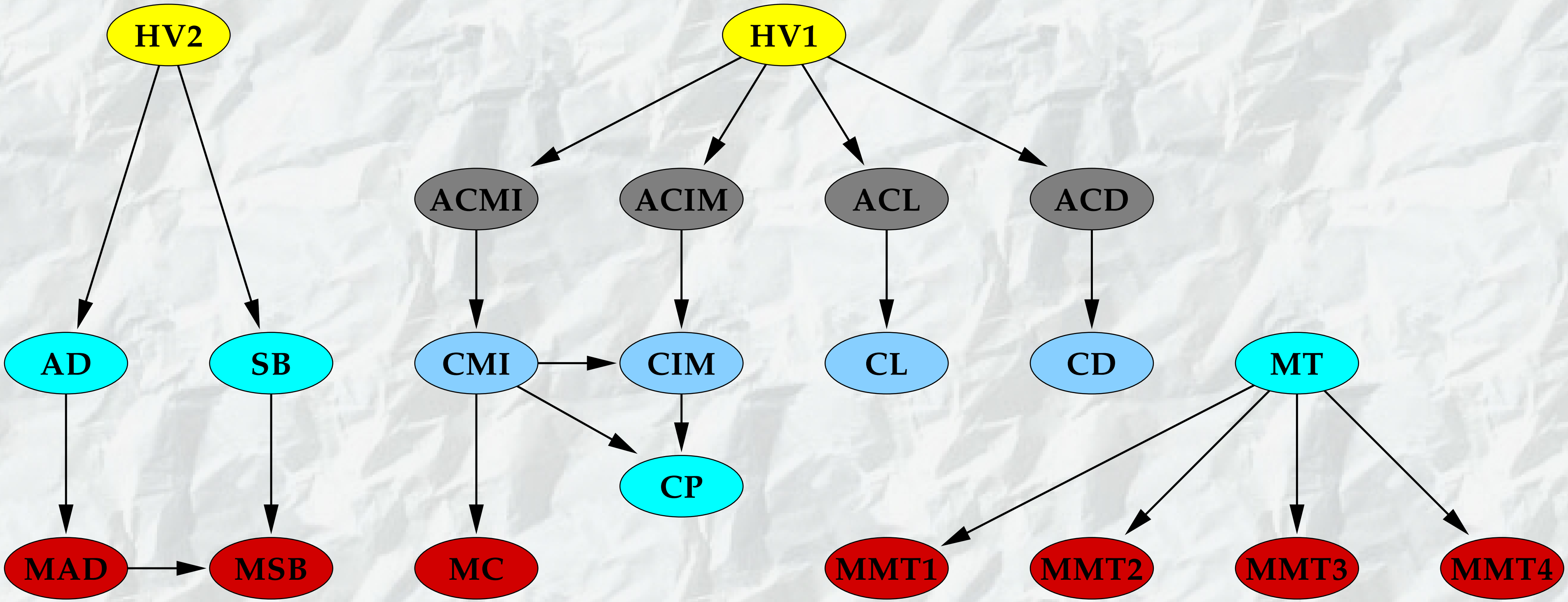
Misconceptions

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%

Process that lead to the student model

- decision on what skills will be tested, preparation of paper tests
- paper tests given to students at Brønderslev high school, 149 students did the test.
- analysis of results, finding misconceptions, summarizing results into a data file,
- learning a Bayesian network model using the PC-algorithm and the EM-algorithm,
- attempts to explain some relations between skills and misconceptions using hidden variables,
- a new learning phase with hidden variables included, certain edges required to be part of the learned model.

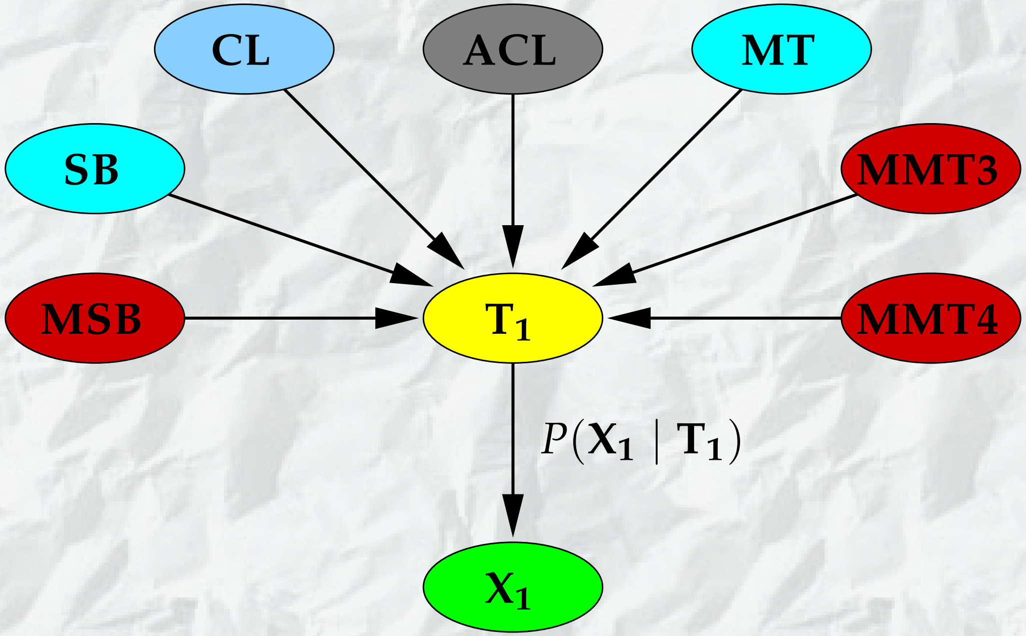
Student model



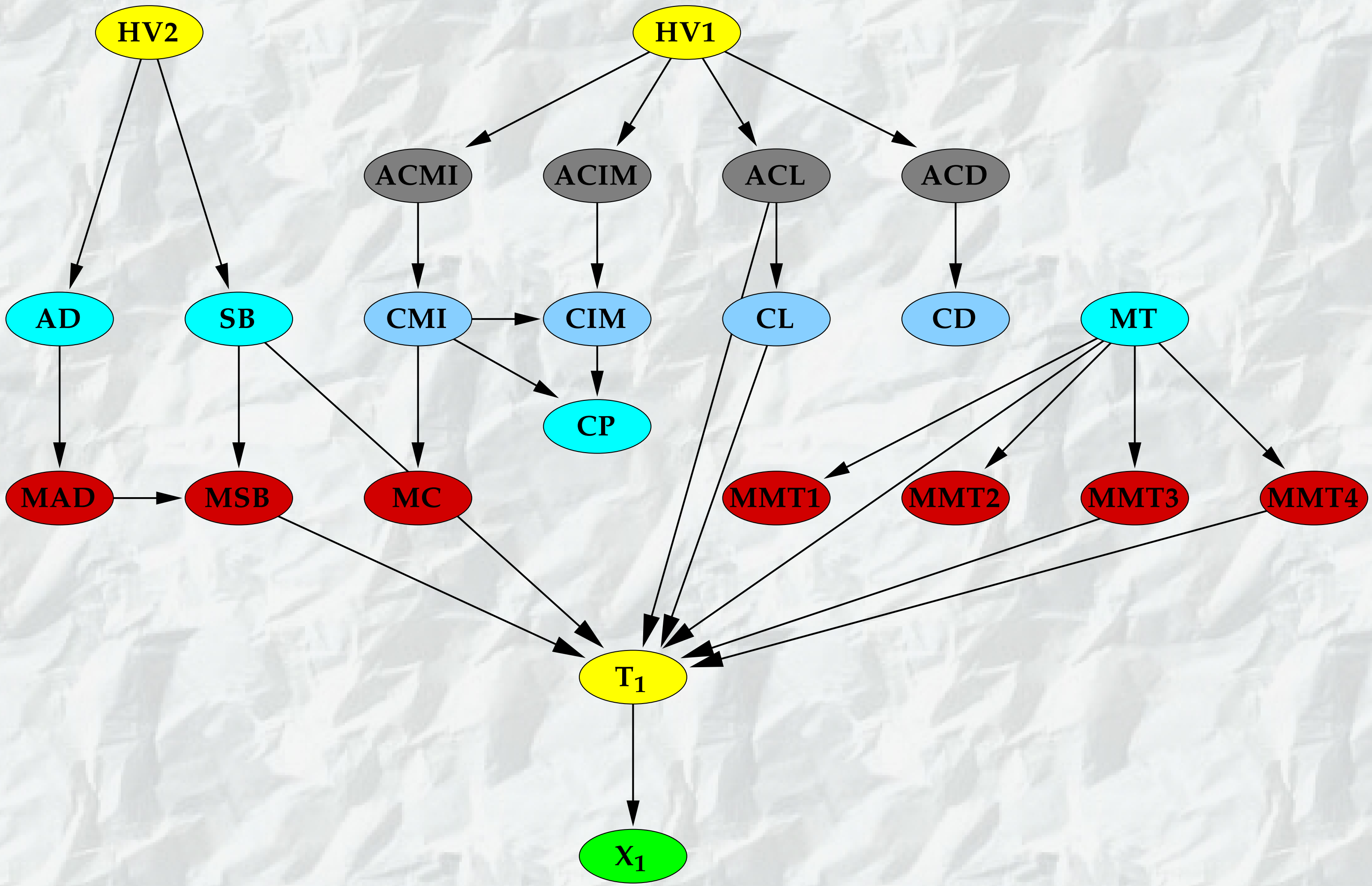
Evidence model for task T_1

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

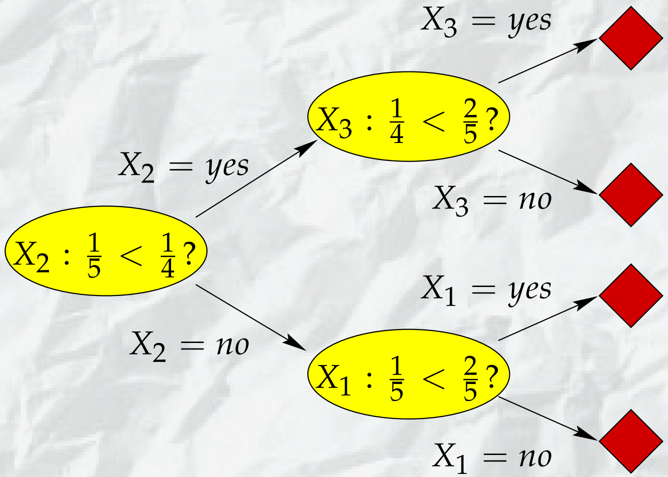
$$T_1 \quad \Leftrightarrow \quad MT \ \&\ CL \ \&\ ACL \ \&\ SB \ \&\ \neg MMT3 \ \&\ \neg MMT4 \ \&\ \neg MSB$$



Student + Evidence model



Example of an adaptive test

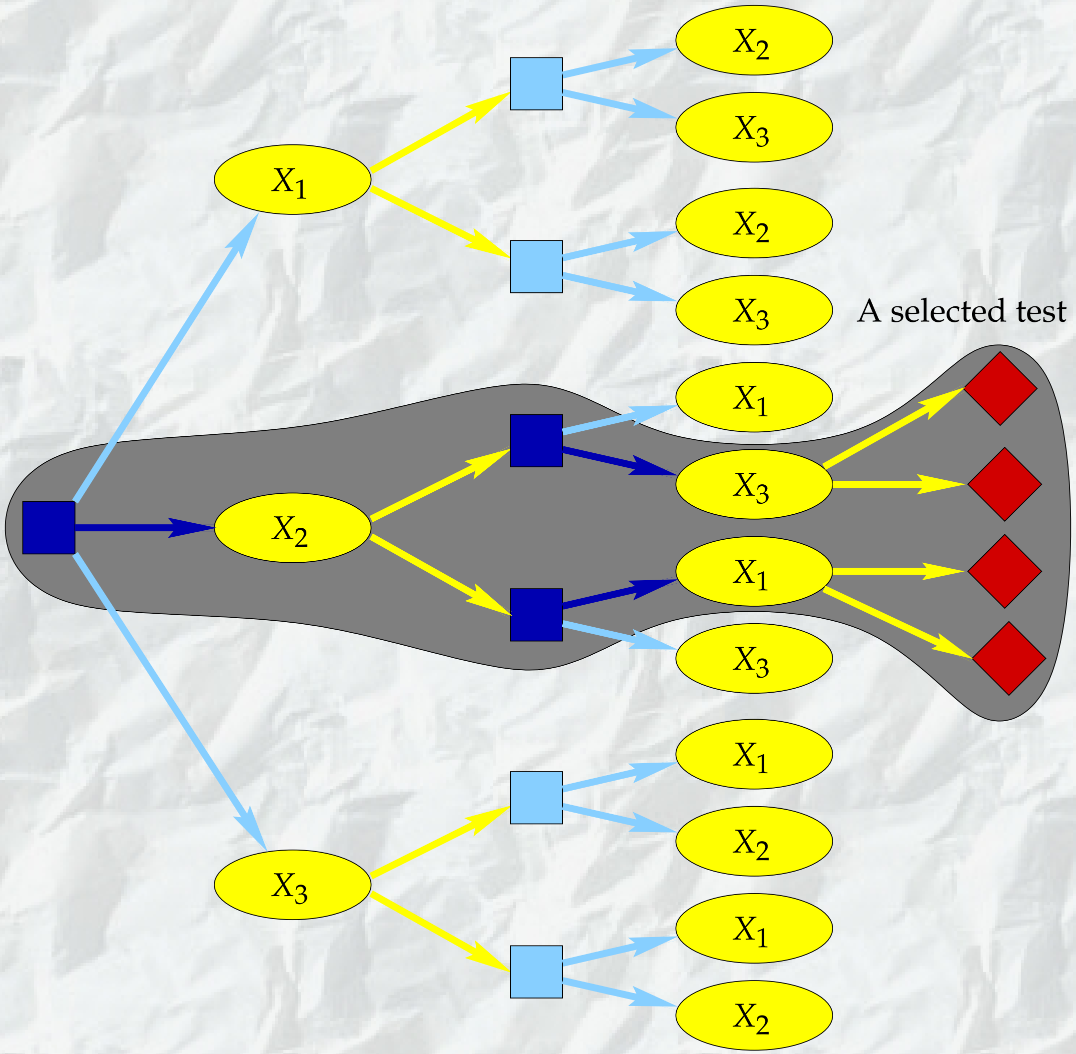


Entropy of a probability distribution $P(S_i)$

$$H\left(P(S_i)\right) = -\sum_{s_i \in \mathbb{S}_i} P(S_i = s_i) \cdot \log P(S_i = s_i)$$

Total entropy in a node n: $H(\mathbf{e}_n) = \sum_{S_i \in \mathcal{S}} H(P(S_i \mid \mathbf{e}_n))$.

Expected entropy at the end of a test $\mathbf{t}$ is
$EH(\mathbf{t}) = \sum_{\ell \in \mathcal{L}(\mathbf{t})} P(\mathbf{e}_\ell) \cdot H(\mathbf{e}_\ell)$.



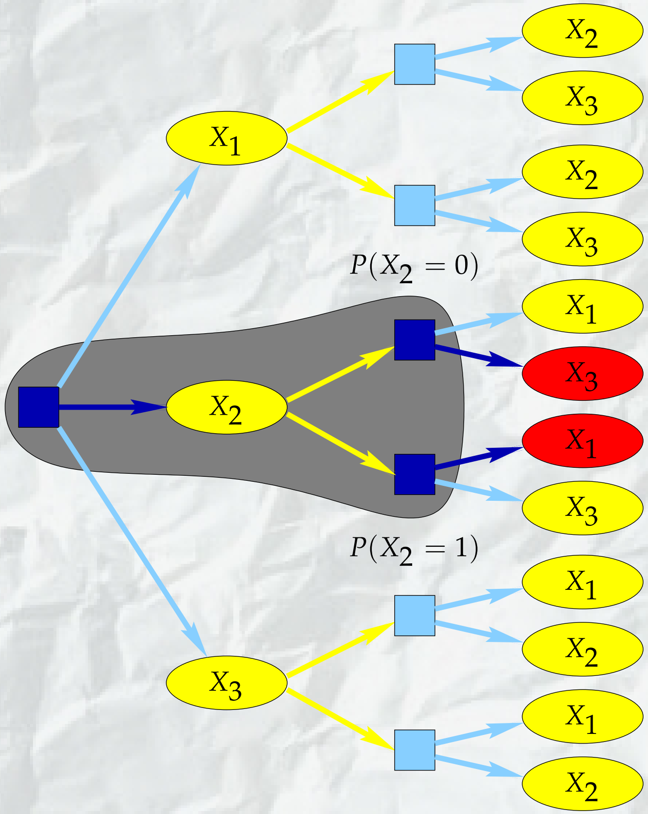
Let $\mathcal{T}$ be the set of all possible tests. A test $\mathbf{t}^\star$ is **optimal** iff

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

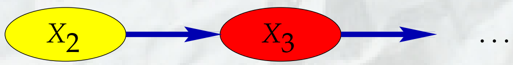
A **myopically optimal** test $\mathbf{t}$ is a test where each question $X^\star$ of $\mathbf{t}$ minimizes the expected value of entropy after the question is answered:

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

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



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



Myopic construction of a fixed test

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

$test := [\ ];$

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

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

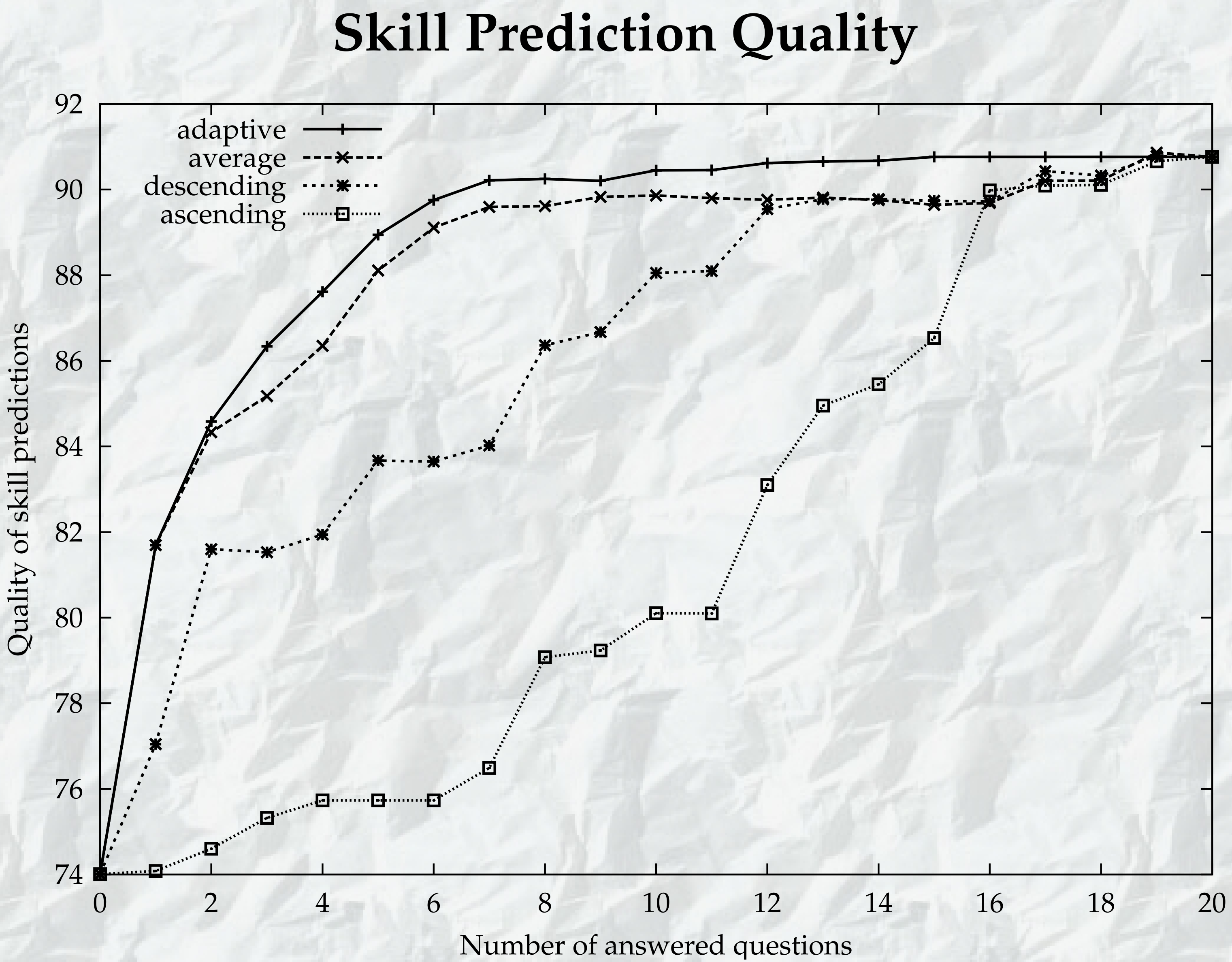
  $e\_list := new\_e\_list;$

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

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

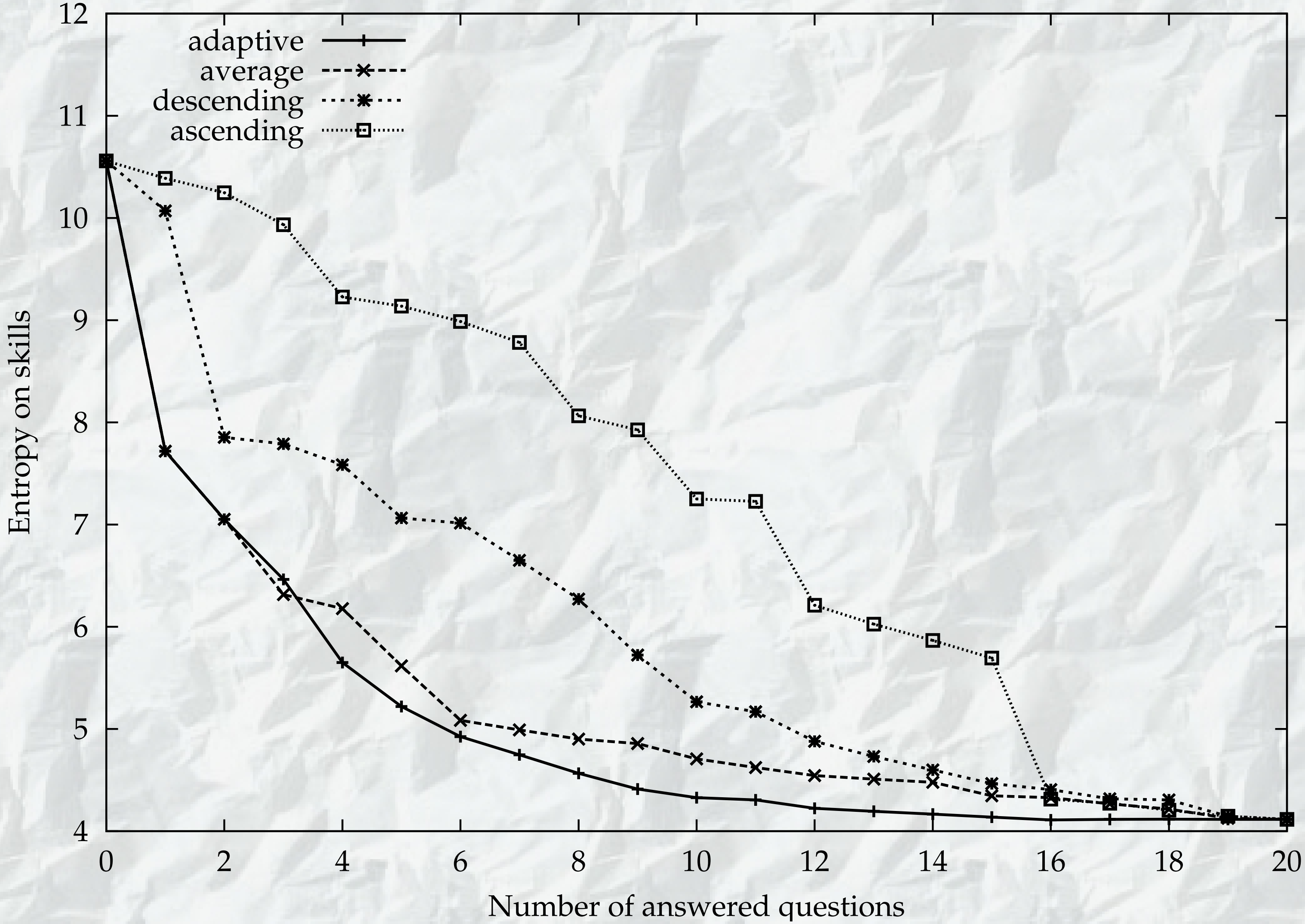
  $counts[i^\star] := 0;$

$return(test);$

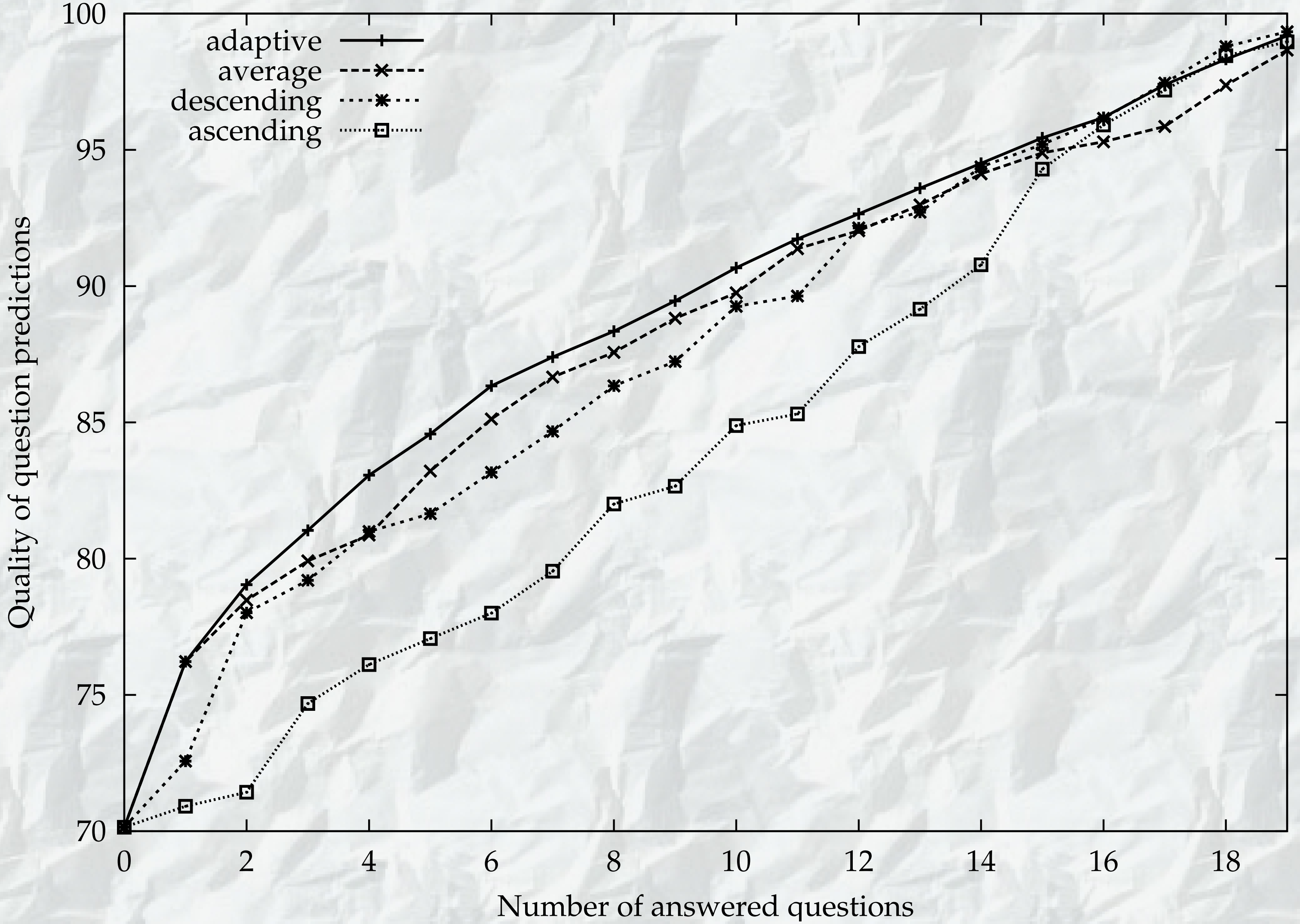


Skill Prediction Quality
adaptive
average
descending
ascending
Quality of skill predictions
Number of answered questions
92
90
88
86
84
82
80
78
76
74
0
2
4
6
8
10
12
14
16
18
20

Total entropy of probability of skills



Question Prediction Quality



An example of a simple diagnostic task

Diagnosis of the absence or the presence of three skills

$$S_1, S_2, S_3$$

by use of a bank of three questions

$$X_{1,2}, X_{1,3}, X_{2,3} \quad .$$

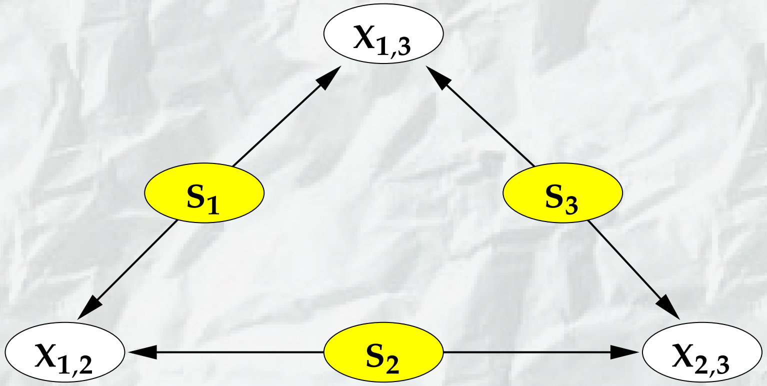
such that

$$P(X_{i,j} = 1 | S_i = s_i, S_j = s_j) \quad = \quad \begin{cases} 1 & \text{if } (s_i, s_j) = (1,1) \\ 0 & \text{otherwise.} \end{cases}$$

Assume answers to all questions from the item bank are wrong, i.e.

$$X_{1,2} = 0, \;\; X_{1,3} = 0, \;\; X_{2,3} = 0 \quad .$$

Reasoning assuming skill independency



All skills are independent

$$P(S_1, S_2, S_3) \quad = \quad P(S_1) \cdot P(S_2) \cdot P(S_3)$$

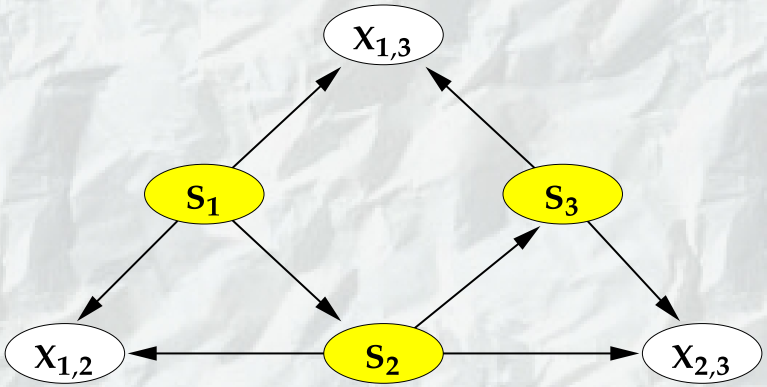
and $P(S_i), i = 1, 2, 3$ are uniform.

Then the probabilities for $j = 1, 2, 3$ are:

$$P(S_j = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) = 0.75 \quad ,$$

i.e. we can not decide which skills are present and which are missing.

Modeling dependence between skills



with deterministic hierarchy

$$S_1 \Rightarrow S_2,\ S_2 \Rightarrow S_3$$

$$\begin{aligned}
P(S_1 = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) &= 1 \\
P(S_2 = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) &= 1 \\
P(S_3 = 0 \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) &= 0.5
\end{aligned}$$

Observe, that for $i = 1, 2, 3$

$$P(S_i \mid X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0) = P(S_i \mid X_{2,3} = 0)\ \text{, i.e.}$$

$X_{2,3} = 0$ gives the **same information** as $X_{1,2} = 0, X_{1,3} = 0, X_{2,3} = 0$.

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.
- The **new 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 was addressed in our UAI 2002 paper.

... and this is the END.

It's time to have a beer.

... or are there any questions?