FROM PPROXIMATE MEMBERSHIP TO DATA INTEGRATION

Adrien Vieilleribière

Groupe de Travail MoVe

25/03/2010

FROM APPROXIMATE MEMBERSHIP TO APPROXIMATE DATA TRANSFORMATIONS

- We consider documents as words and trees and study how to compare them with some regular schemas. We first show that we can decide **“approximate membership”** (i.e. decide if an instance is close to the given schema) in time independent of the size of the input document. Then, we extend the situation to **“approximate data exchange”** where we wonder whether the input can be transformed by a given transducer to an instance close to a given target schema. We finally study “**approximate structural consistency**” where we decide the existence of a transducer satisfying the approximate data exchange conditions.
- We show that these three decision problems are testable: they can be solved in time independent of the size of the input structure. Moreover, the given statistical representation of the input can be efficiently approximated by sampling and the obtained testers are robust to noise.

MOTIVATIONS

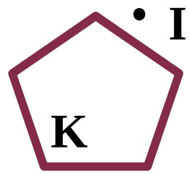
- Manage Huge and Imperfect Data
  - Robustness to noise.
  - Decisions based on statistics.
- Approximation and Complexity
- Interoperability
  - Exchange data between different Schemas.
  - Automatic mappings

OUTLINE

I Preliminaries

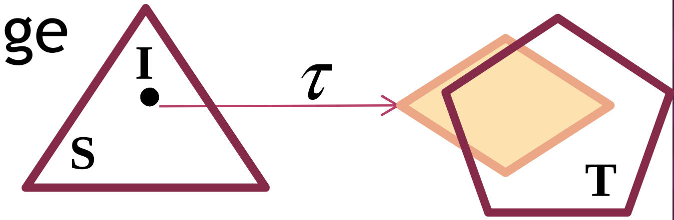
- Distance
- Statistical Embeddings

II Approximate Membership



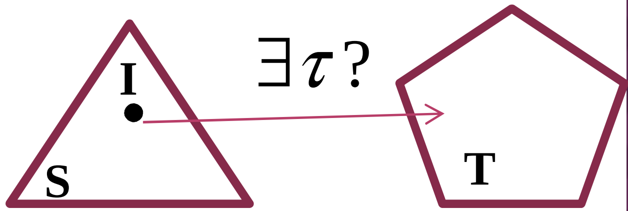
- Word Case
- Tree Case

III Approximate data exchange



- Transformations
- Consistency
- Typechecking

IV Approximate Structural Consistency



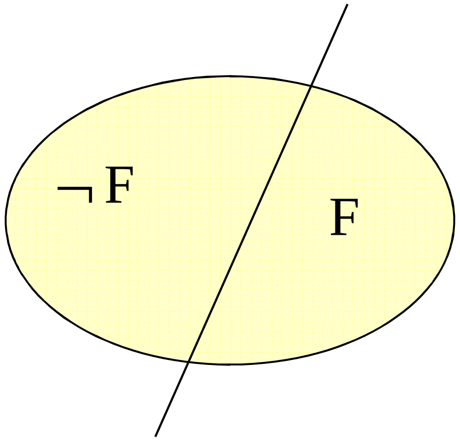
- Base mappings

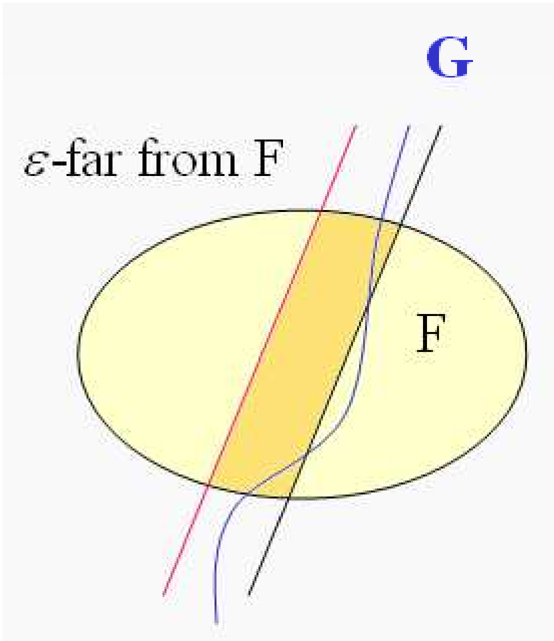
I PRELIMINARIES

- Testers
- Distance between Structures
- Statistical Embedding
- Soundness & Robustness

APPROXIMATION : INTUITION

- Satisfiability : T |= F_{DTD}
- Approximate Satisfiability : T $|=^{\varepsilon}$ F_{DTD}





TESTABLE PROPERTY

Let F be a property on a class K of structures U

An **ε-tester** for F is a (probabilistic) algorithm A such that:

- If U |= F, A accepts
- If U is ε-far from F, A rejects (with high probability)

A property F is **testable** if there exists a probabilistic algorithm A s.t.

- For all ε it is an **ε-tester** for F
- Time(A) independent of n=|U|.

R. Rubinfeld, M. Sudan, Robust characterizations of polynomials, 1994
O. Goldreich, S. Goldwasser and D. Ron, Property Testing and its connection to Learning and Approximation 1996.

Links with sampling, linear time corrector.

WORD EDIT DISTANCE WITH MOVES

- Edit Distance : Insertions, Deletions, Modifications
- Moves

0111000011110011001

0111011110000011001

$$dist\ (W\ , L\ ) = \underset{W\ '\in L}{Min}\ \{dist\ (W\ ,W\ ')\}$$

$$dist\ (L\ , L\ ') = \underset{W\ \in L\ ,W\ '\in L\ '}{Min}\ \{dist\ (W\ ,W\ ')\}$$

WORD EMBEDDING (UNIFORM STATISTIC)

$For\ \Sigma = \{0,1\}:$

$$ustat_2(w) = \frac{1}{n-k+1}(\#n_1, \#n_2, \ldots, \#n_{2^k})$$

$\#n_1$ number of "00...0"
$\#n_2$ number of "00...1"
.. ..

.. ..
$\#n_{2^k}$ number of "11...1"

***W*=001010101110**

$$ustat_2(w) = \frac{1}{11}(1,4,4,2)$$

$Dimension\ |\Sigma|^k$

Unitary

length *n :* n-k+1 blocks of length k
For k=2, n=12, 11 blocks

SOUDNESS

ε-close strings have close statistics

Simple edit: $$|u.stat(w) - u.stat(w')| \le \frac{2k}{n-k+1} \le \frac{2}{n.\varepsilon}$$

Move:
w=A.B.C.D, w'=A.C.B.D $$|u.stat(w) - u.stat(w')| \le \frac{2.3(k-1)}{n-k+1} \le \frac{6}{n.\varepsilon}$$

Hence, for ε^2.n operations, $$|u.stat(w) - u.stat(w')| \le 6.\varepsilon$$

$$\boxed{dist(w,w') \le \varepsilon^2.n \Rightarrow |u.stat(w) - u.stat(w')| \le 6.\varepsilon}$$

Robustness of **u.stat** is Harder!
Auxiliary distributions and key lemmas.

SOUNDNESS AND ROBUSTNESS

Soundness: ε-close strings have close statistics

$$dist(w,w') \leq \varepsilon.n$$

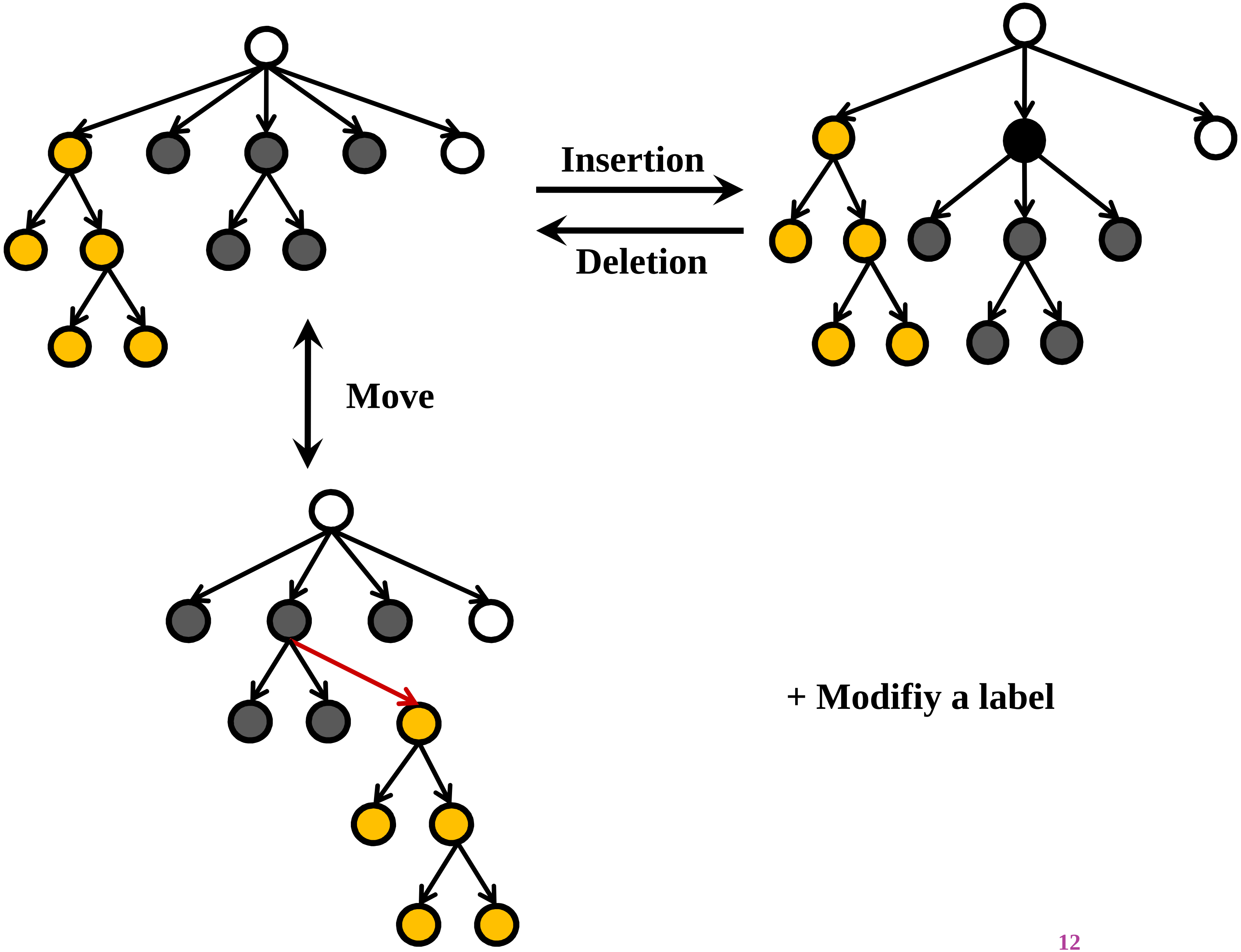
Robustness: ε-far strings have far statistics

$$dist(w,w') \geq \varepsilon.n$$

Approximate Satisfiability and Equivalence

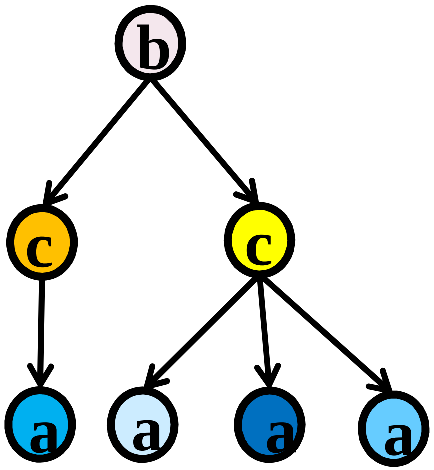
E. Fischer, F. Magniez, M. de Rougemont, LICS06.

TREE EDIT DISTANCE WITH MOVES

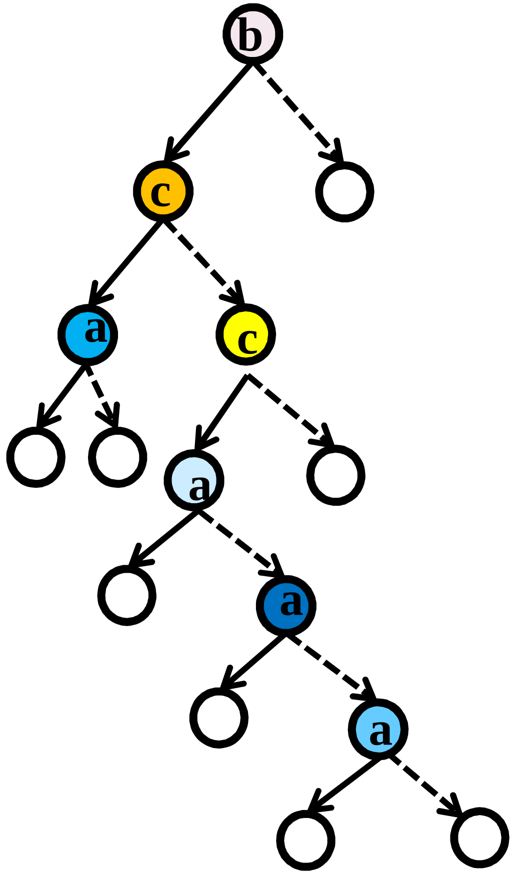


+ Modifiy a label

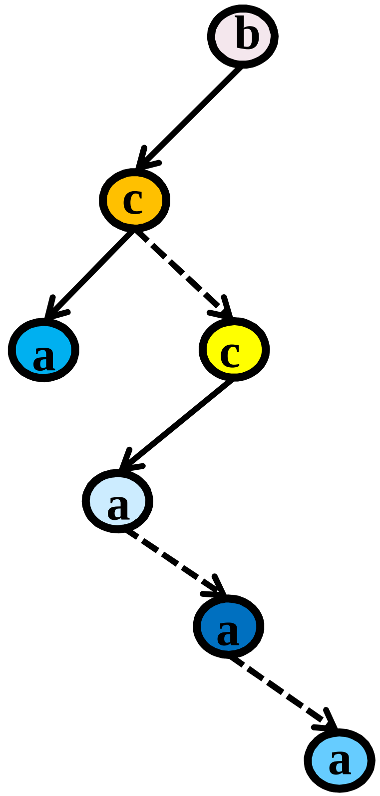
RABIN ENCODING



Unranked Tree



Binary Tree



2-extended tree

TREE EMBEDDING (USTAT)

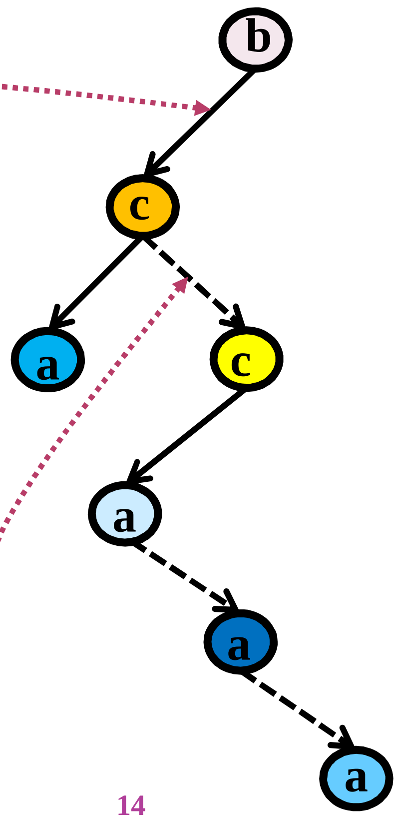
$$\begin{array}{c} 2^{k-1}\,\textit{lignes} \\ (\textit{types}) \end{array} \left\{ \overbrace{\begin{pmatrix} \textit{Unitary} \\ \textit{Matrix} \end{pmatrix}}^{\begin{array}{c} |\Sigma|^k\ \textit{colones} \\ (\textit{labels}) \end{array}} \right.$$

Frequency of the segments of length k

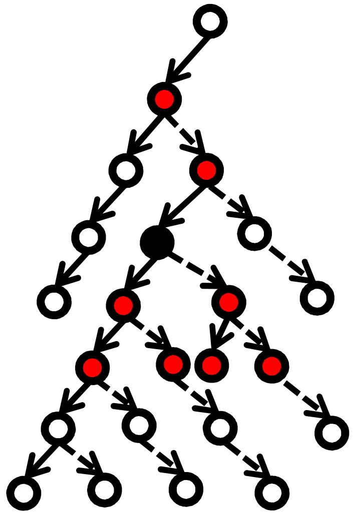
In practice : Sparse Matrices

Example for k=2

$$u.stat_2(t) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{ccccccccc} aa & ab & ac & ba & bb & bc & ca & cb & cc \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \frac{1}{6} & \frac{2}{6} & 0 & 0 \\ \frac{2}{6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{6} \end{pmatrix} \end{array}$$



SOUNDNESS



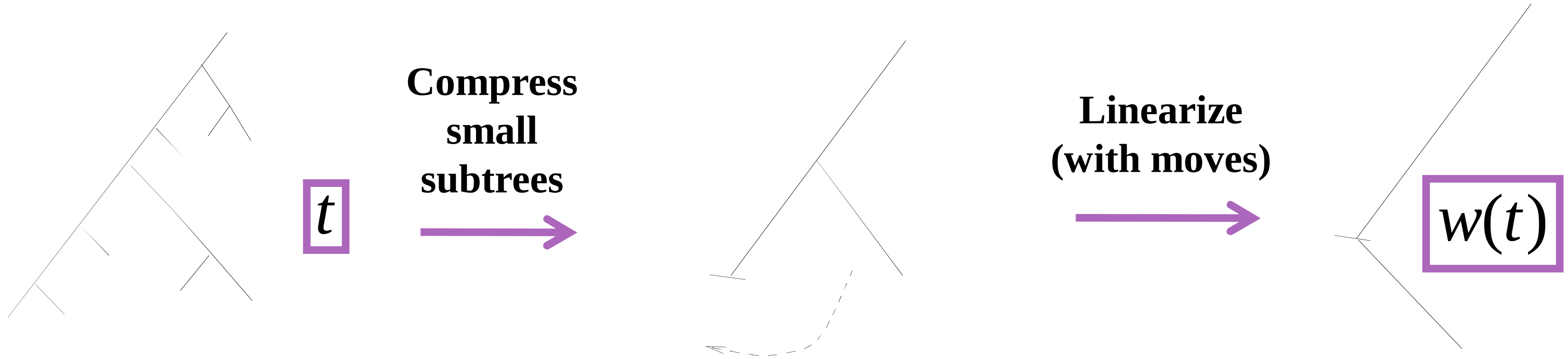
$$1\,\text{operation} \;\rightarrow\; \left|u.stat_k(t) - u.stat_k(t')\right| \leq \frac{3 \times k \times \sum_{i=1}^{k-1} 2^i}{n} = \frac{3k\left(2^k - 1\right)}{n}$$

Hence, for $\dfrac{\varepsilon^2}{2^{1/\varepsilon} - 1}$ operations,

$$\boxed{dist(t,t') \leq \frac{\varepsilon^2}{2^{1/\varepsilon} - 1}.n \Rightarrow \left|u.stat(t) - u.stat(t')\right| \leq 3.\varepsilon}$$

ROBUSTNESS

Main idea : Use robustness on words to prove robustness on trees



- There a c≤1 such that for any couple of trees (t,t') of same size

$$\boxed{\left|ustat\left(w(t)\right)-ustat\left(w(t')\right)\right| \le c\left|ustat\left(t\right)-ustat\left(t'\right)\right|}$$

One segment of length k in w(t) → $\le 2^k$ segments on ustat(t)

- There is g and c' such that for any couple of trees (t,t') of same size

$$\boxed{if\ \left(\left|ustat\left(t\right)-ustat\left(t'\right)\right| \le g\left(\varepsilon\right)\right)\ then\ \left(dist\left(t,t'\right) \le c'.\varepsilon\right)}$$

Less than ε.n forks

Sequence of operations on w(t) → Sequence on t

SOUNDNESS AND ROBUSTNESS

Soundness: ε-close trees have close statistics

$$dist(I, I') \leq \varepsilon . |I|$$

Robustness: ε-far trees have far statistics

$$dist(I, I') \geq \varepsilon . |I|$$

Transformation de mots, d'arbres, et de statistiques
Thèse A.Vieilleribière 2008

OUTLINE

I Preliminaries

II Approximate Membership

III Approximate data exchange

IV Approximate Structural Consistency

II APPROXIMATE MEMBERSHIP

- Main problem
- Key ideas
- Sampling
- Language embedding
- Testers

APPROXIMATE MEMBERSHIP : DEFINITION

Given

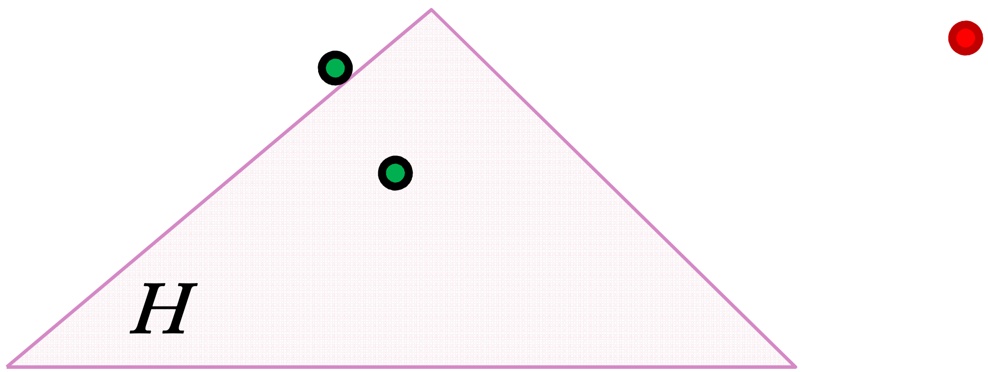
- an instance I (word or tree),
- a schema K (regular expression or DTD) and
- an approximation parameter ε,

decide if I is ε-close to K.

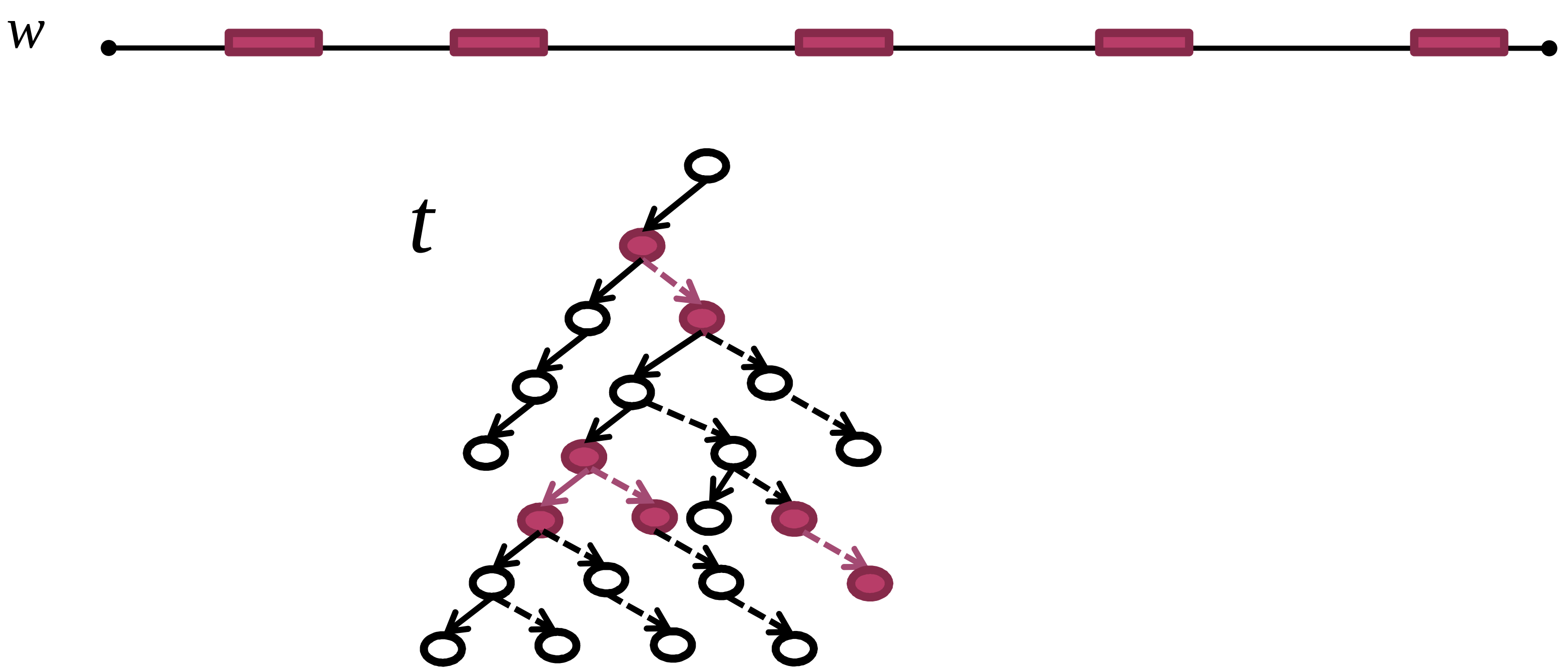
$$\exists I' \text{ s.t.} \begin{cases} dist(I, I') \leq \varepsilon \\ I' \in K \end{cases} \quad ?$$

APPROXIMATE MEMBERSHIP : SOLUTION

- Sample I : $u.\hat{stat}_{1/\varepsilon}(I)$
- Construct H : embedding of K
- Check the geometrical distance
- Accept if the distance is small, reject otherwise



SAMPLING



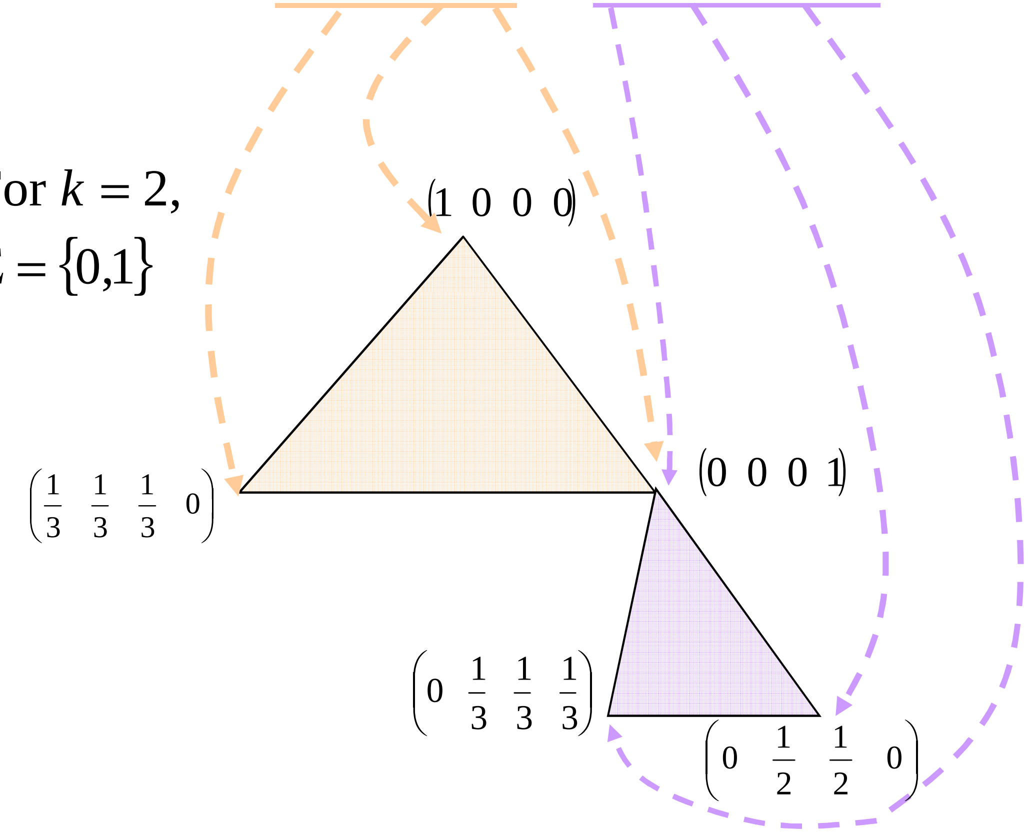
$\widehat{u.stat}(I)[u] = \frac{1}{N}\sum_{i=1\ldots N} X_i[u]$: average on uniform samples

$\widehat{u.stat}(I) \approx u.stat(I)$ [Chernoff]

STATISTICS OF A (WORD) SCHEMA

For $k = 2$,
$\Sigma = \{0,1\}$

r = (010)*0*1* + 1*(01)*(110)*



H={$u.stat(w)$: w in r } is approximated by a union of polytopes.

2 polytopes for r.

$$x : \lambda_1 s_1 + \lambda_2 s_2 + \lambda_3 s_3$$

$$\lambda_i \geq 0, \sum_i \lambda_i = 1$$

STATISTICS OF A DTD

root : a*b
a : c.d
c : a.f + g
b : e*

For $k = 2$,
$\Sigma = \{a, b, c, d, e\}$

$$\tau_1 = a(., \underline{a})$$

t_1	0 1
aa	0 1

$$\tau_2 = e(., \underline{e})$$

t_2	0 1
ee	0 1

$$\tau_3 = a(c(\underline{a}(., f), d), .)$$

t_3	0	1
ac	1/4	0
af	0	1/4
ca	1/4	0
cd	0	1/4

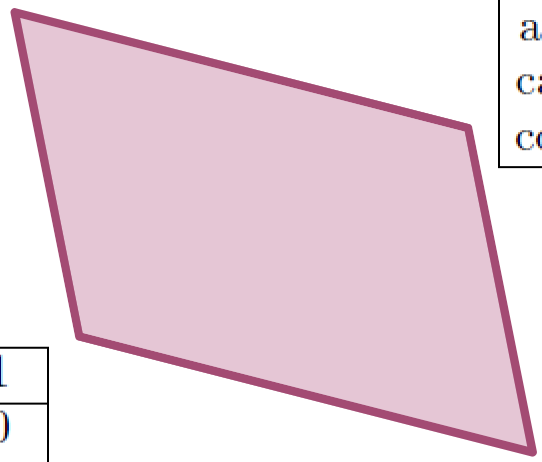
$$\tau_4 = a(c(g, d), \underline{a})$$

t_4	0	1
aa	1/4	0
ac	0	1/4
cd	1/4	0
cf	0	1/4

STATISTICS OF A DTD

t_1	0	1
aa	0	1

t_3	0	1
ac	1/4	0
af	0	1/4
ca	1/4	0
cd	0	1/4



t_4	0	1
aa	1/4	0
ac	0	1/4
cd	1/4	0
cf	0	1/4

t_2	0	1
ee	0	1

$$H = \left\{ \lambda_1 \cdot t_1 + \lambda_2 \cdot t_2 + \lambda_3 \cdot t_3 + \lambda_4 \cdot t_4 \;\middle|\; \sum_i \lambda_i = 1 \right\}$$

TESTER (ε-Membership)

1. **Sample t (in the Rabin encoding) with** $N \in O\left(\frac{2|\Sigma_S|^{2/\varepsilon} \ln(|\Sigma_S|)}{\varepsilon^5}\right)$ **samples and let** $ustat_k^*(t)$ **be the estimation of** $ustat_k(t)$
2. **Enumerate all the possible polytope** H_T **and**

   Accept if one is ε-close to $ustat_k^*(t)$

   else Reject

- Adapted to the approximation required ε
- Constant time decision with respect to $|I|$

OUTLINE

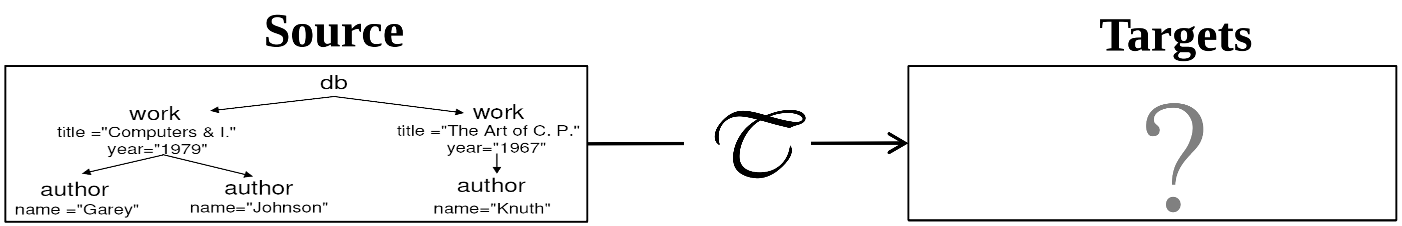
I Preliminaries

II Approximate Membership

III Approximate data exchange

IV Approximate Structural Consistency

DATA EXCHANGE



<!ELEMENT db (work*)>
<!ELEMENT work (author*)>
<!ATTLIST work title CDATA
#REQUIRED year CDATA>
<!ELEMENT author (EMPTY)>
<!ATTLIST author name CDATA #REQUIRED>

<!ELEMENT bib (livre*)>
<!ELEMENT livre (auteur+, titre , annee)>
<!ELEMENT auteur #PCDATA>
<!ELEMENT titre #PCDATA>
<!ELEMENT annee #PCDATA>

- Data Exchange setting: (K_S, τ, K_T)
  - [Fagin et al. 2002, Arenas, Libkin 2005]

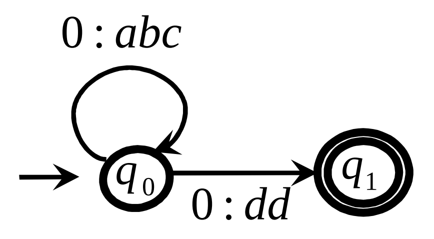
- **Source-Consistency:** Given a source structure I in K_S, is there a target J in K_T s.t. (I,J) in τ ?

- **Typechecking:** Decide if for all I in K_S and all J s.t. (I,J) in τ, J is in K_T.

FINITE STATE MACHINE

$$\tau = \left(\Sigma_s, \Sigma_T, Q, i, F, \delta\right)$$

$$\delta \subseteq Q \times \Sigma_S \times \left(\Sigma_T\right)^* \times Q$$



$$\delta : \begin{cases} (q_0,0) \rightarrow a(b(c(q_0)) \\ (q_0,0) \rightarrow d(d(q_1)) \end{cases}$$

$$\tau(00) = \{abcabc, abcdd\}$$

XSL-TRANSDUCER

W. Martens, F. Neven (PODS04)

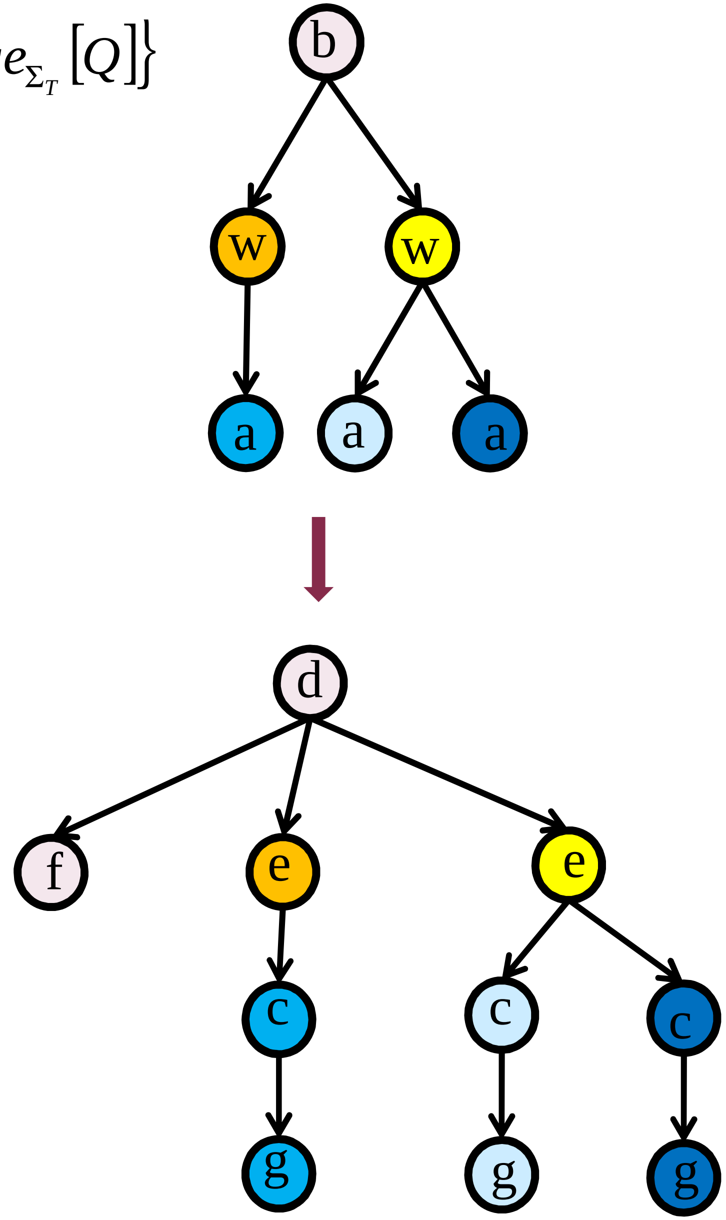
$$\tau = (\Sigma_s, \Sigma_T, Q, q_0, \delta) \quad \delta : \{(q,a) \to Hedge_{\Sigma_T}[Q]\}$$

Linear rules

$(q_0, b) \to d(f\ q_0)$

$(q_0, w) \to e(q_0)$

$(q_0, a) \to c(g(q_0))$

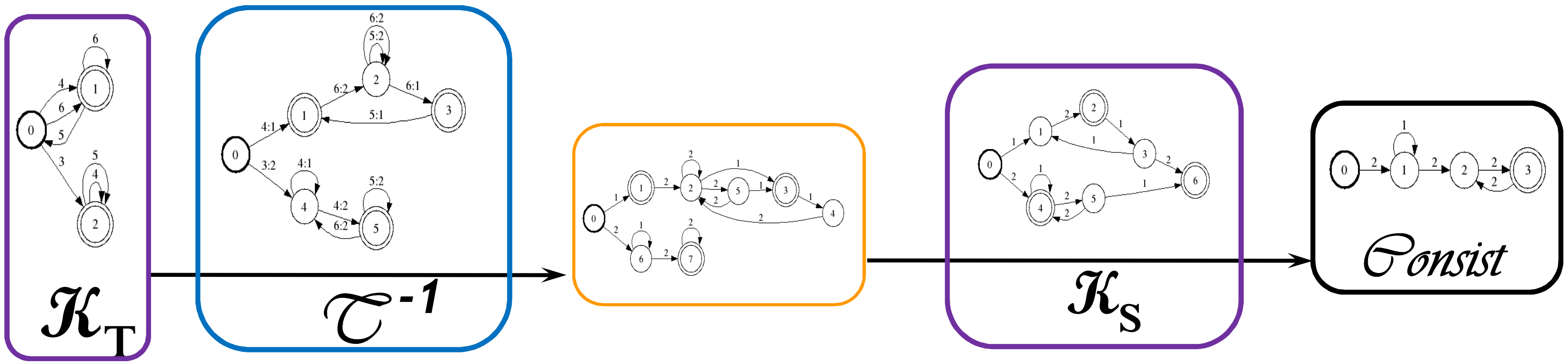


CONSISTENCY

- For a FSM or a transducer, the inverse image of a regular language is regular

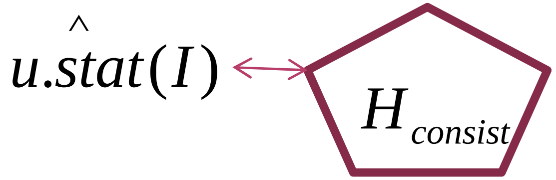
- Regular languages closed by intersection

Consistent inputs form a regular language

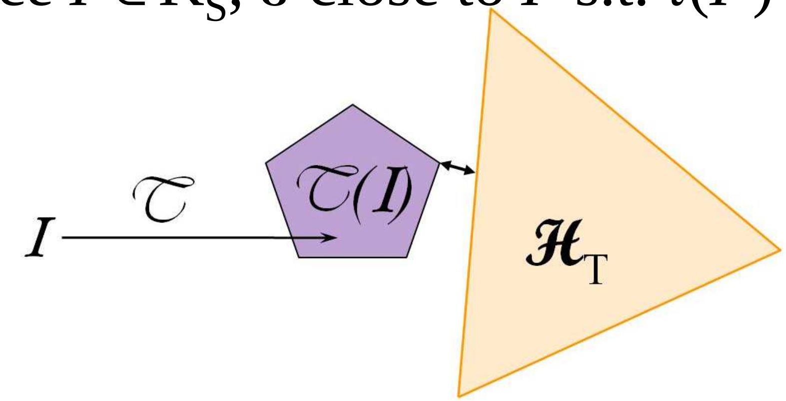


APPROXIMATE CONSISTENCY

Simple ε-Source-Consistency: Given a source structure I, is there a source $I' \in K_S$, ε-close to I s.t. $\tau(I') \in K_T$?



Generalized ε-Source-Consistency: Given a source structure I, is there a source $I' \in K_S$, ε-close to I s.t. $\tau(I')$ *is* ε-close to K_T ?



Constant time decision

M. de Rougemont, A. Vieilleribière
Approximate Data Exchange, ICDT07.

TYPECHECKING

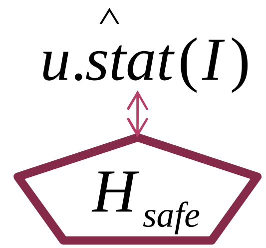
- For a FSM or a transducer, the inverse image of a regular language is regular
- Regular languages closed by intersection and complement

$$K_{safe} = K_{consist} \cap \overline{\left(\tau^{-1}\left(\overline{K_T}\right)\right)}$$

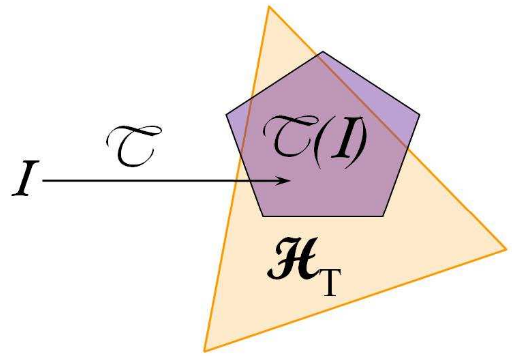
Safe inputs form a regular language

APPROXIMATE TYPECHECKING

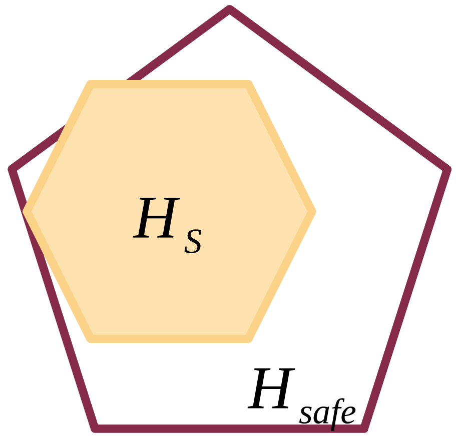
Simple ε-source-typechecking:
Given a source structure *I*, is there a source $I' \in K_S$, ε-close to *I* s.t. $\tau(I') \subseteq K_T$?



Generalized ε-source-typechecking :
Given a source structure *I*, is there a source $I' \in K_S$, ε-close to *I* s.t. $\tau(I')$ *is* always ε-close to K_T ?



ε-Typechecking :
for any source structure $I \in K_S$,
is there a source $I' \in K_S$, ε-close to *I* s.t. $\tau(I')$ *is* always ε-close to K_T ?



OUTLINE

IV APPROXIMATE STRUCTURAL CONSISTENCY

- Main Problem
- Transformations
- Base mappings
- Key points

APPROXIMATE STRUCTURAL CONSISTENCY : DEFINITION

- Fix a source schema S, a target schema T and parameters $k = 1/\varepsilon$ (the precision), α (the ratio) and m (number of states).

- Given an input (word or tree of size n) in S, we decide if there is a transducer τ with m states, and I' ε-close to I such that
  - $\tau(I')$ is ε-close to T and
  - $\alpha.n \leq |\tau(I')| \leq n/\alpha$

APPROXIMATE STRUCTURAL CONSISTENCY : MAIN DEFINITIONS

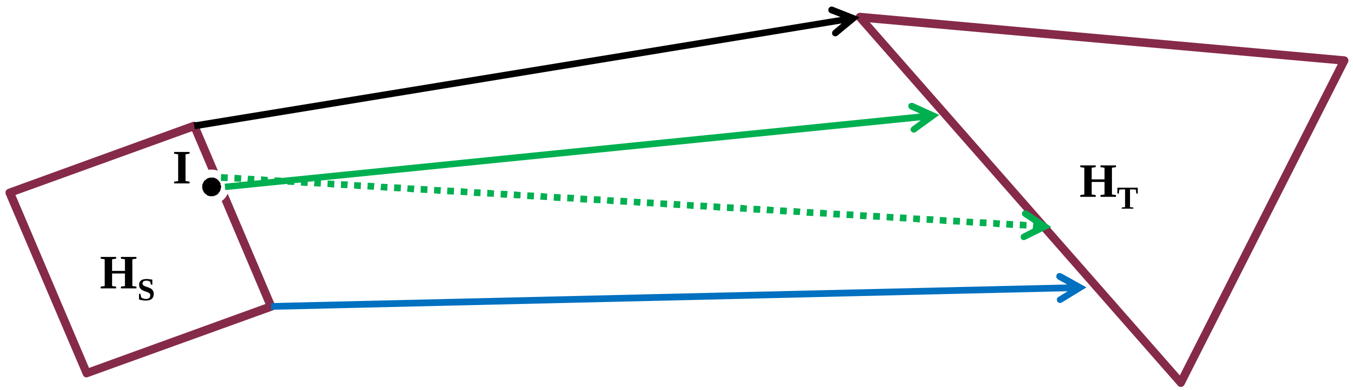
- A (partial) **base mapping** μ between *S* and *T* is a partial function $H_i^S \rightarrow H_j^T$ only defined on some summits of the polytope
  i.e. $\mu(s_i) = t_j$ for $s_i \in$ C and $t_j \in D$.

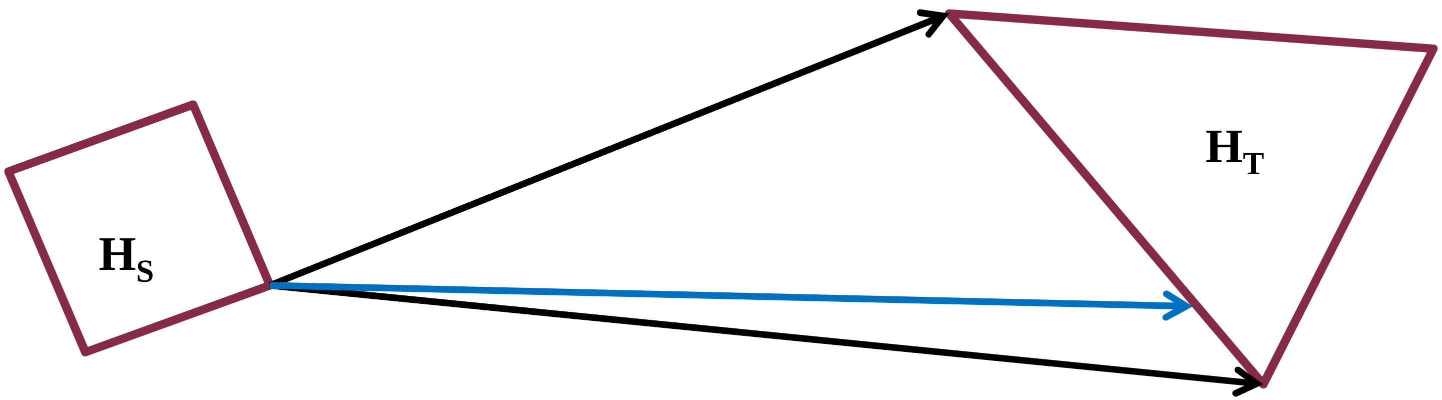
- A 1-state transducer τ between *S* and *T* is **compatible with μ**, if

$$\tau(u_i) = v_j \qquad s_i = ustat_k(u_i)$$

$$\mu(s_i) = t_j \qquad t_j = ustat_k(v_j)$$

BASE MAPPING : INTUITION



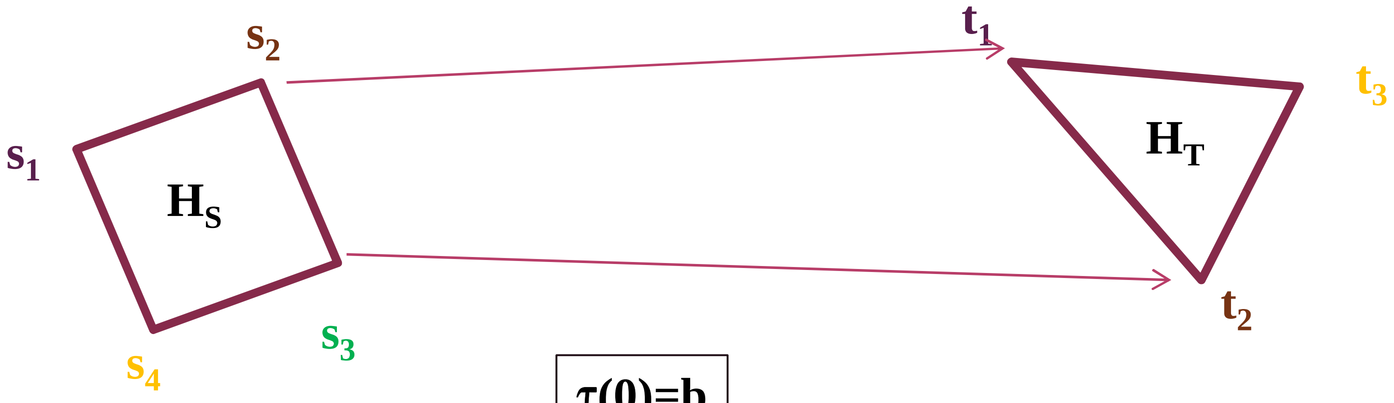


BASE MAPPING : EXAMPLE

$S=\underbrace{(001)^*}_{u_1}.\underbrace{(01)^*}_{u_2}.\underbrace{1^*}_{u_3}.\underbrace{(011)^*}_{u_4}$

$T=\underbrace{(ab)^*}_{v_1}.\underbrace{a^*}_{v_2}.\underbrace{(aabc)^*}_{v_3}$

$w_1 = 0101111=(01)^2 1^3$
$w = 000111$



$\boxed{\begin{array}{l}\tau(0)=b\\ \tau(1)=a\end{array}}$

τ is μ compatible

$\tau(u_1)= \tau(01)=ba$
$\tau(u_2)= \tau(1)=a$
$ustat_2(\tau(u_1)^*)=ustat_2((ba)^*)=ustat_2((ab)^*)=ustat_2((v_1)^*)$
$ustat_2(\tau(u_2)^*)=ustat_2((v_2)^*)$

A.S.C. : KEY IDEAS

- Estimate ustat$_k$(I) $\rightarrow$ Set of polytopes (possible decompositions)

- There is a transducer satisfying A.S.C.
  $\leftrightarrow$ there are some mappings between embeddings
  - one state transducer $\leftrightarrow$ a (partial) base mapping
  - *m* states transducer $\leftrightarrow$ *m* distinct base mappings

- There are (compatible) base mapping between embedding
  $\leftrightarrow$ a specific linear program has a solution

CONCLUSION

- We have **testers** for
  - **approximate membership**
  - **approximate data exchange**
  - **approximate structural consistency**

- Decisions are only based on
  - a **sampling** of the input
  - **statistical representations** of schemas

- They work **in time independent on the size on the input** (word or tree)

- They are **insensitive to noise**

QUESTIONS ?

Adrien Vieilleribière: vieille@lri.fr