Challenges for Efficient Query Evaluation on Structured Probabilistic Data

SUM2016

SEPTEMBER 23, 2016, NICE

Antoine Amarilli, Silviu Maniu, Mikaël Monet

A probabilistic database

R	
a	d
f	e
d	a
b	e

S	
d	e
f	c
a	e
c	e

Q		
c	e	f

Amarilli Antoine, Maniu Silviu, Monet Mikaël

A probabilistic database

R	
a	d
f	e
d	a
b	e

S	
d	e
f	c
a	e
c	e

R		
a	d	0.2
f	e	0.7
d	a	0.13
b	e	0.81

S		
d	e	0.005
f	c	0.9
a	e	0.7
c	e	0.23

Q		
c	e	f

Q			
c	e	f	0.66

Amarilli Antoine, Maniu Silviu, Monet Mikaël

A probabilistic database

R	
a	d
f	e
d	a
b	e

S	
d	e
f	c
a	e
c	e

R		
a	d	0.2
f	e	0.7
d	a	0.13
b	e	0.81

S		
d	e	0.005
f	c	0.9
a	e	0.7
c	e	0.23

Q		
c	e	f

Q			
c	e	f	0.66

TID model

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Probability of a possible world

R		
a	d	0.2
f	e	0.7
d	a	0.13
b	e	0.81

S		
d	e	0.005
f	c	0.9
a	e	0.7
c	e	0.23

A possible

World I

R	
f	e
d	a

S	
c	e

Q			
c	e	f	0.66

Q		
c	e	f

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Probability of a possible world

R		
a	d	0.2
f	e	0.7
d	a	0.13
b	e	0.81

S		
d	e	0.005
f	c	0.9
a	e	0.7
c	e	0.23

A possible

World I

R	
f	e
d	a

S	
c	e

Q			
c	e	f	0.66

Q		
c	e	f

Probability Pr(I) of this possible world =
0.7*0.13*0.23*0.66

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Probability of a possible world

R		
a	d	0.2
f	e	0.7
d	a	0.13
b	e	0.81

S		
d	e	0.005
f	c	0.9
a	e	0.7
c	e	0.23

A possible

World I

R	
f	e
d	a

S	
c	e

Q			
c	e	f	0.66

Q		
c	e	f

Probability Pr(I) of this possible world =
0.7*0.13*0.23*0.66
(1-0.2)(1-0.81)*(1-0.005)*(1-0.9)*(1-0.7).

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Probabilistic query evaluation (PQE)

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Probabilistic query evaluation (PQE)

- Focus on Boolean queries (yes/no)

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Probabilistic query evaluation (PQE)

- Focus on Boolean queries (yes/no)
- Probability of a query $\mathfrak{Q}$ on probabilistic instance $\mathfrak{T}$:

$$\mathrm{P}(\mathfrak{Q}) = \sum_{I \subseteq \mathfrak{T},\, \mathfrak{Q} \models I} \Pr(I)$$

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Probabilistic query evaluation (PQE)

Amarilli Antoine, Maniu Silviu, Monet Mikaël

- Focus on Boolean queries (yes/no)
- Probability of a query $\mathfrak{Q}$ on probabilistic instance $\mathfrak{T}$:

$$\mathrm{P}(\mathfrak{Q}) = \sum_{I \subseteq \mathfrak{T},\, \mathfrak{Q} \models I} \Pr(I)$$

- Problem: in general #P-hard

1) Approximate probability computation

Amarilli Antoine, Maniu Silviu, Monet Mikaël

1) Approximate probability computation

- Monte-Carlo sampling

Amarilli Antoine, Maniu Silviu, Monet Mikaël

1) Approximate probability computation

- Monte-Carlo sampling
- Inconvenient: running time quadratic in desired precision.

Amarilli Antoine, Maniu Silviu, Monet Mikaël

1) Approximate probability computation

Amarilli Antoine, Maniu Silviu, Monet Mikaël

- Monte-Carlo sampling

- Inconvenient: running time quadratic in desired precision.

  ⇒ Not adequate for low probabilities.

2) Restricting the class of queries

- [Dalvi and Suciu 2012] show the following dichotomy for any UCQ Q :

2) Restricting the class of queries

- [Dalvi and Suciu 2012] show the following dichotomy for any UCQ Q :
- Either PQE is #P-hard on all intances

2) Restricting the class of queries

- [Dalvi and Suciu 2012] show the following dichotomy for any UCQ Q :
- Either PQE is #P-hard on all intances
- Either PQE is PTIME on all instances

2) Restricting the class of queries

- [Dalvi and Suciu 2012] show the following dichotomy for any UCQ Q :
- Either PQE is #P-hard on all intances
- Either PQE is PTIME on all instances

- Simple conjunctive query $\exists x,y\ R(x),T(x,y),S(y)$ is already #P-hard !

2) Restricting the class of queries

- [Dalvi and Suciu 2012] show the following dichotomy for any UCQ Q :
- Either PQE is #P-hard on all intances
- Either PQE is PTIME on all instances

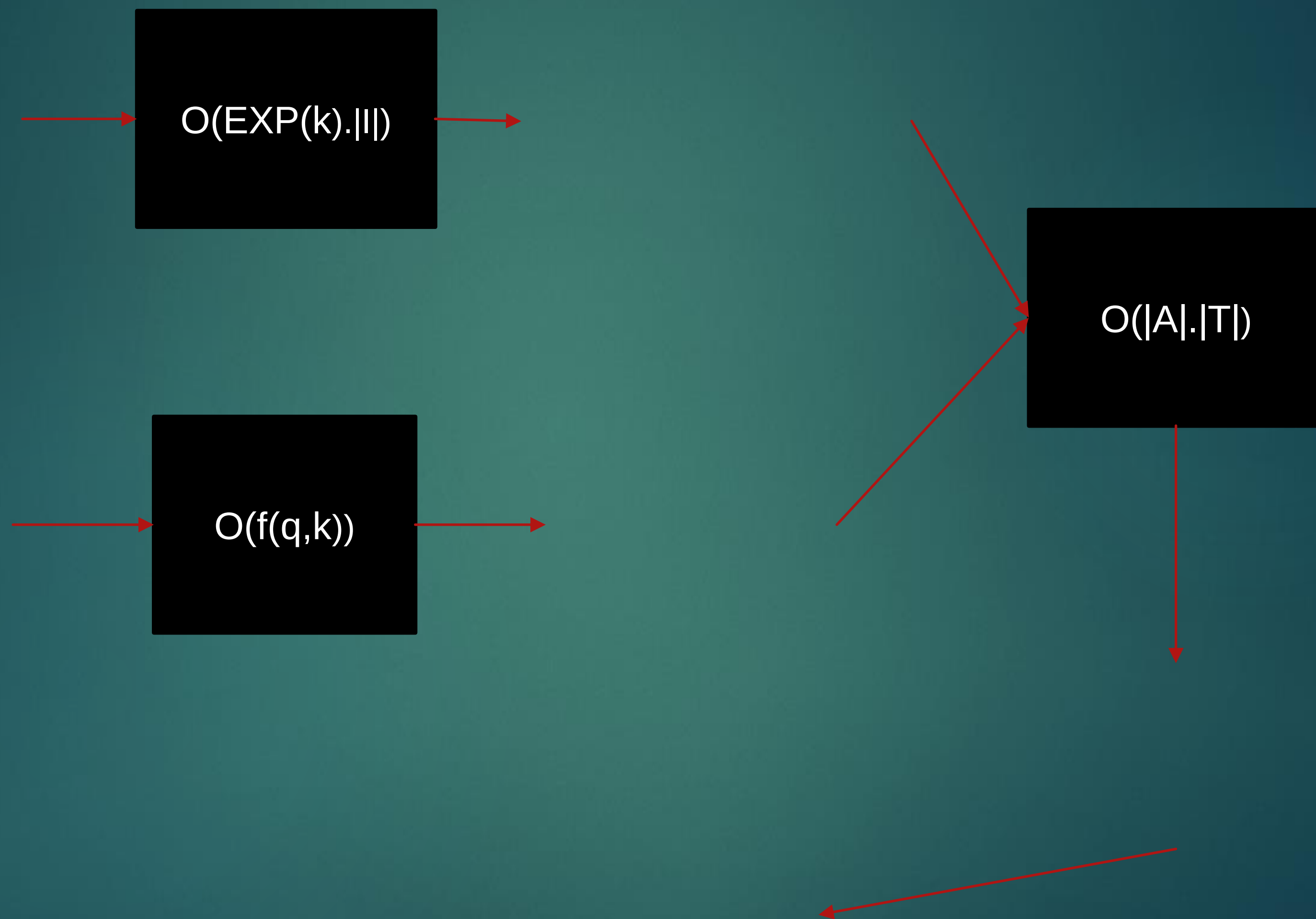
- Simple conjunctive query $\exists x,y\ R(x),T(x,y),S(y)$ is already #P-hard !

- Criterion is to crisp

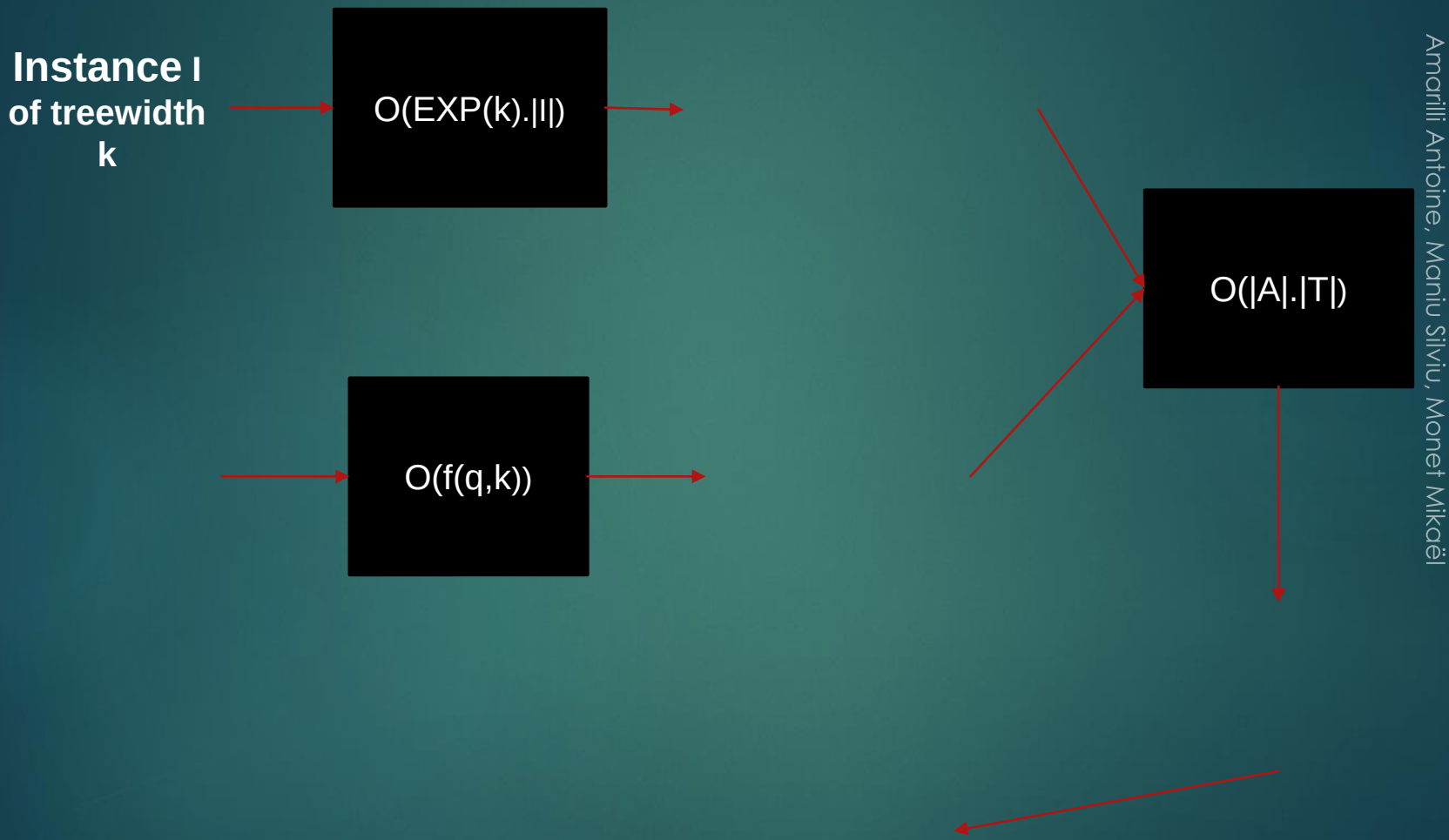
3) Restricting the shape of the instances

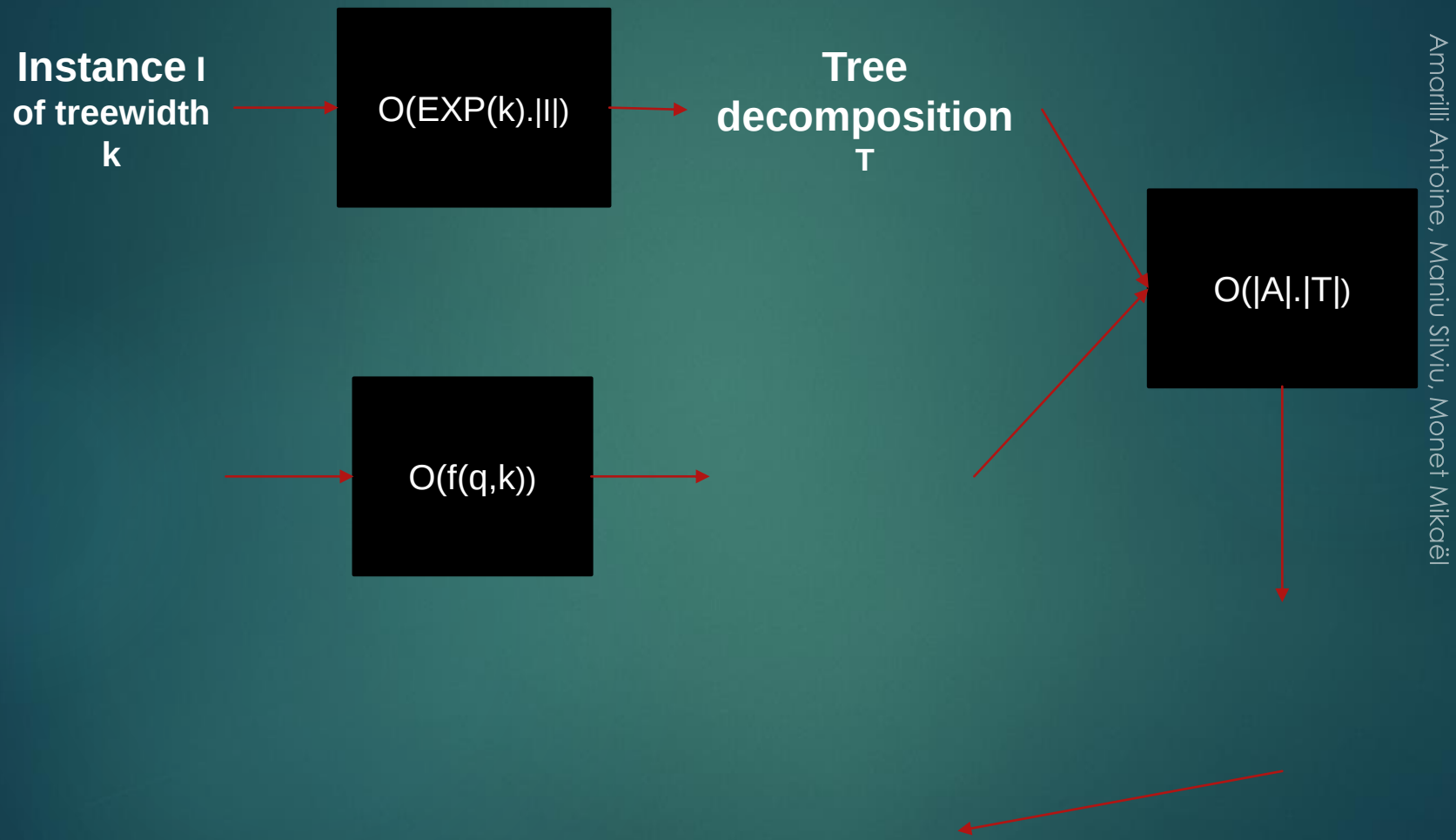
- Bound the ***treewidth*** of instances by a constant.

- Treewidth: mesure used to tell how far a graph is from being a tree

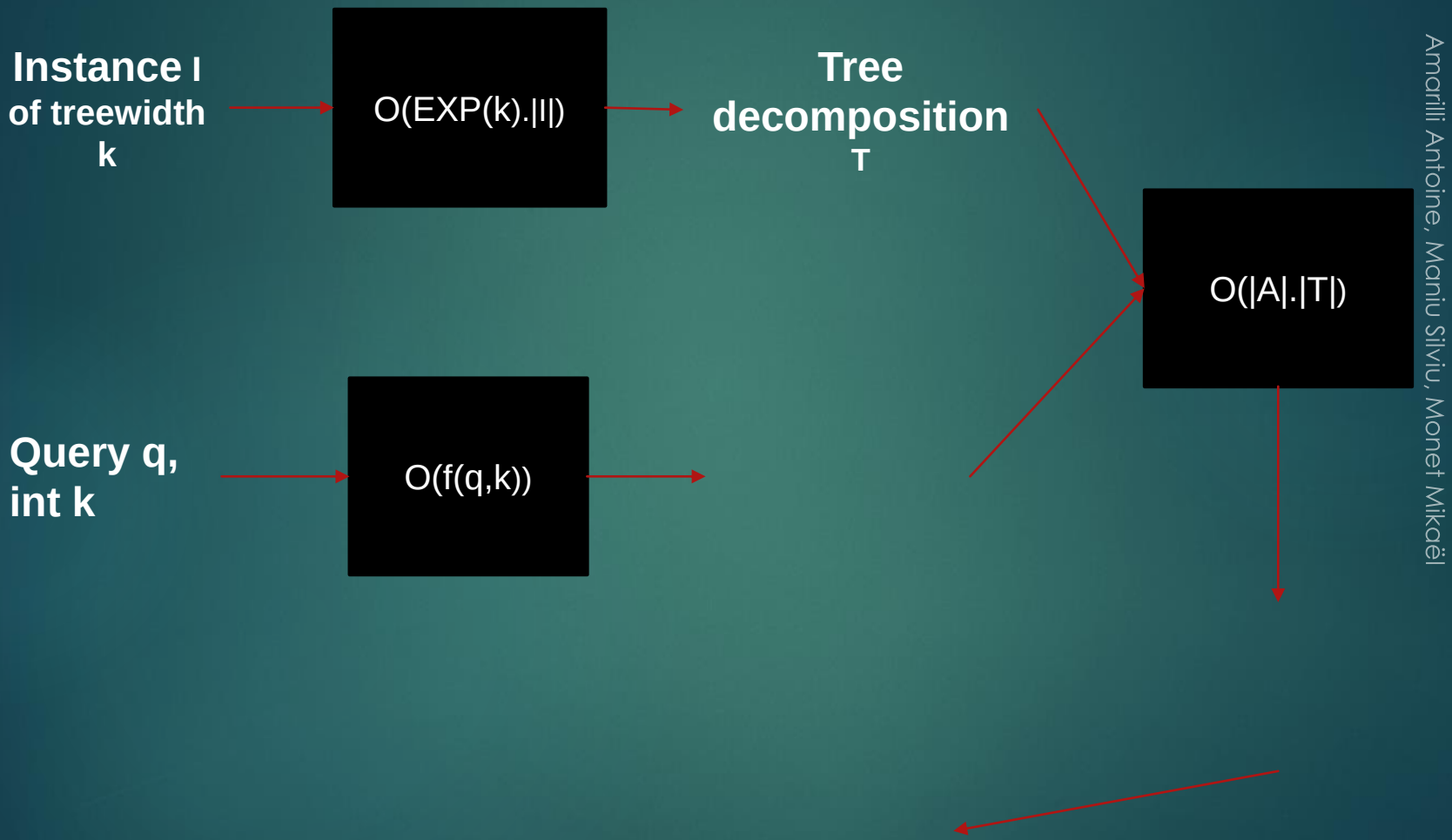


Amarilli Antoine, Maniu Silviu, Monet Mikaël

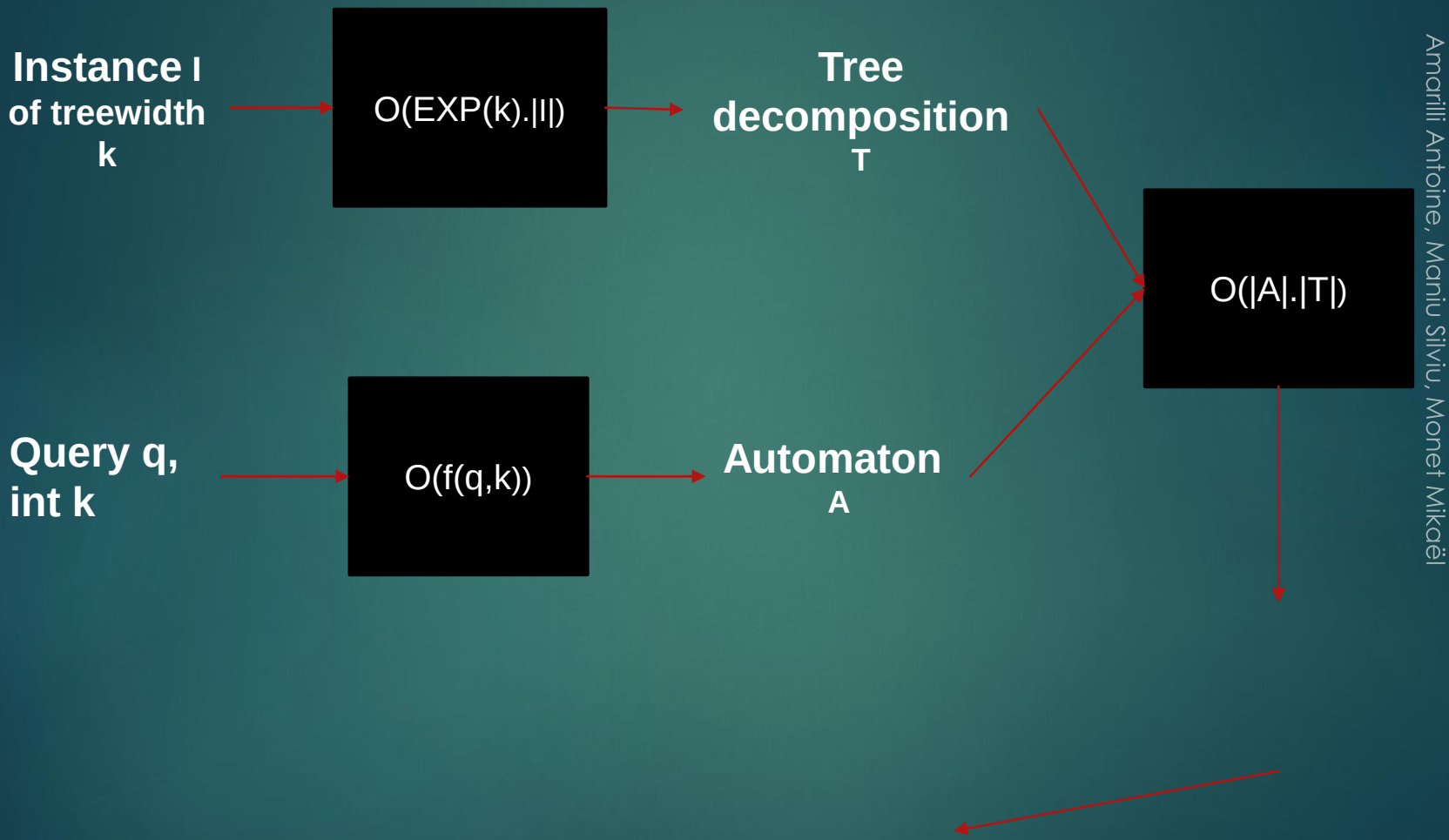

Instance I
of treewidth
k
O(EXP(k).|I|)
O(f(q,k))
O(|A|.|T|)



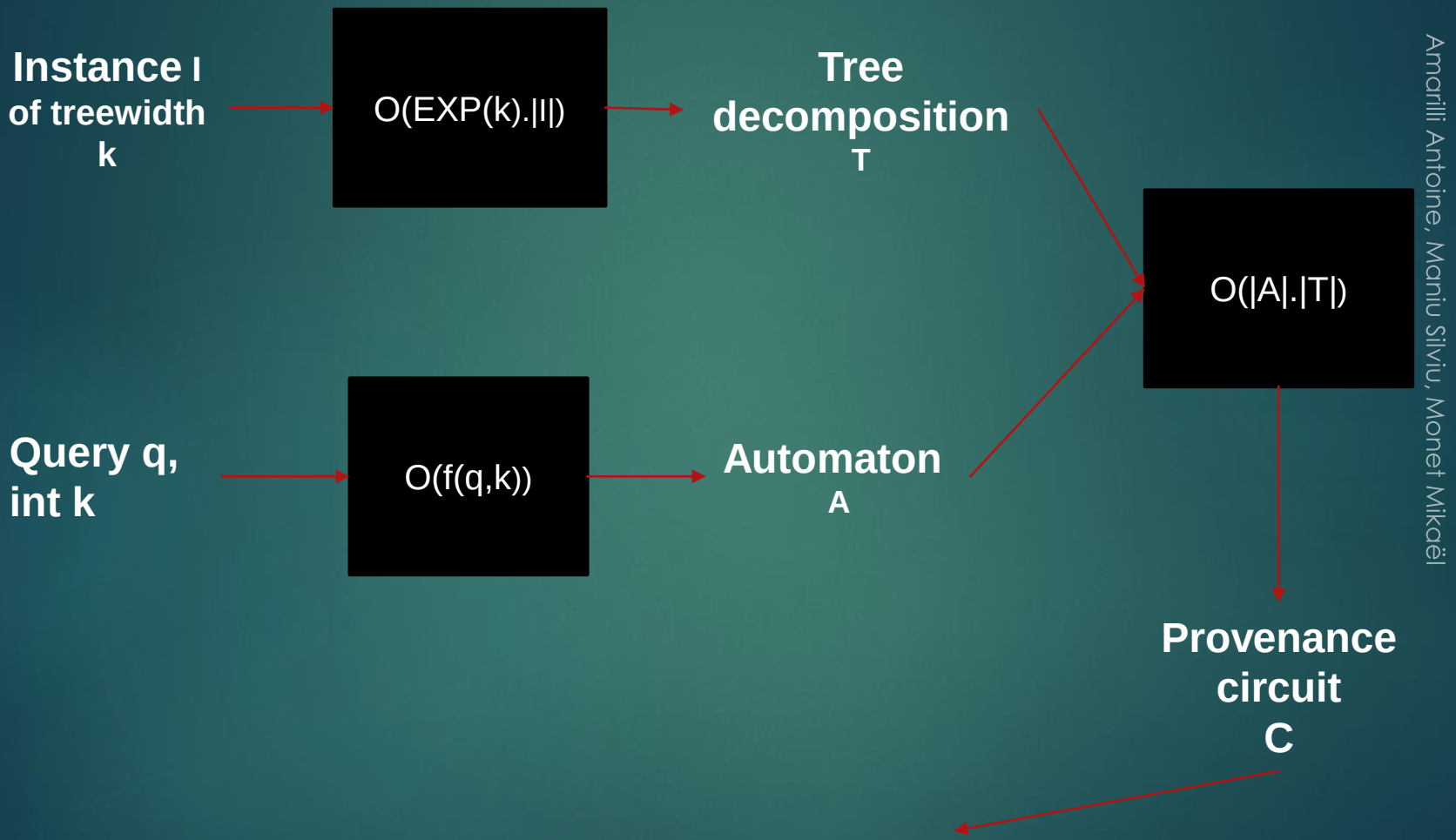
Amarilli Antoine, Maniu Silviu, Monet Mikaël



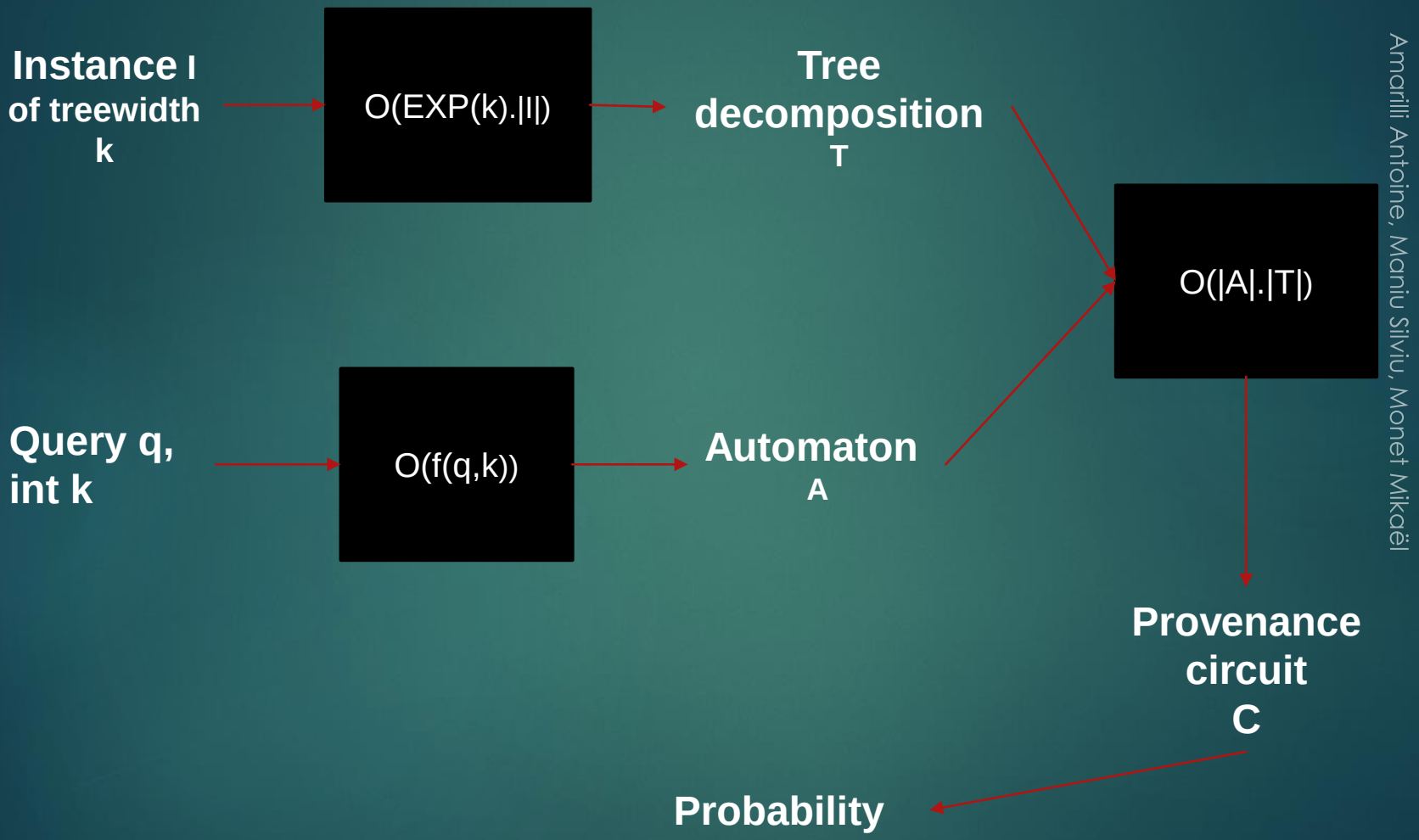
Amarilli Antoine, Maniu Silviu, Monet Mikaël



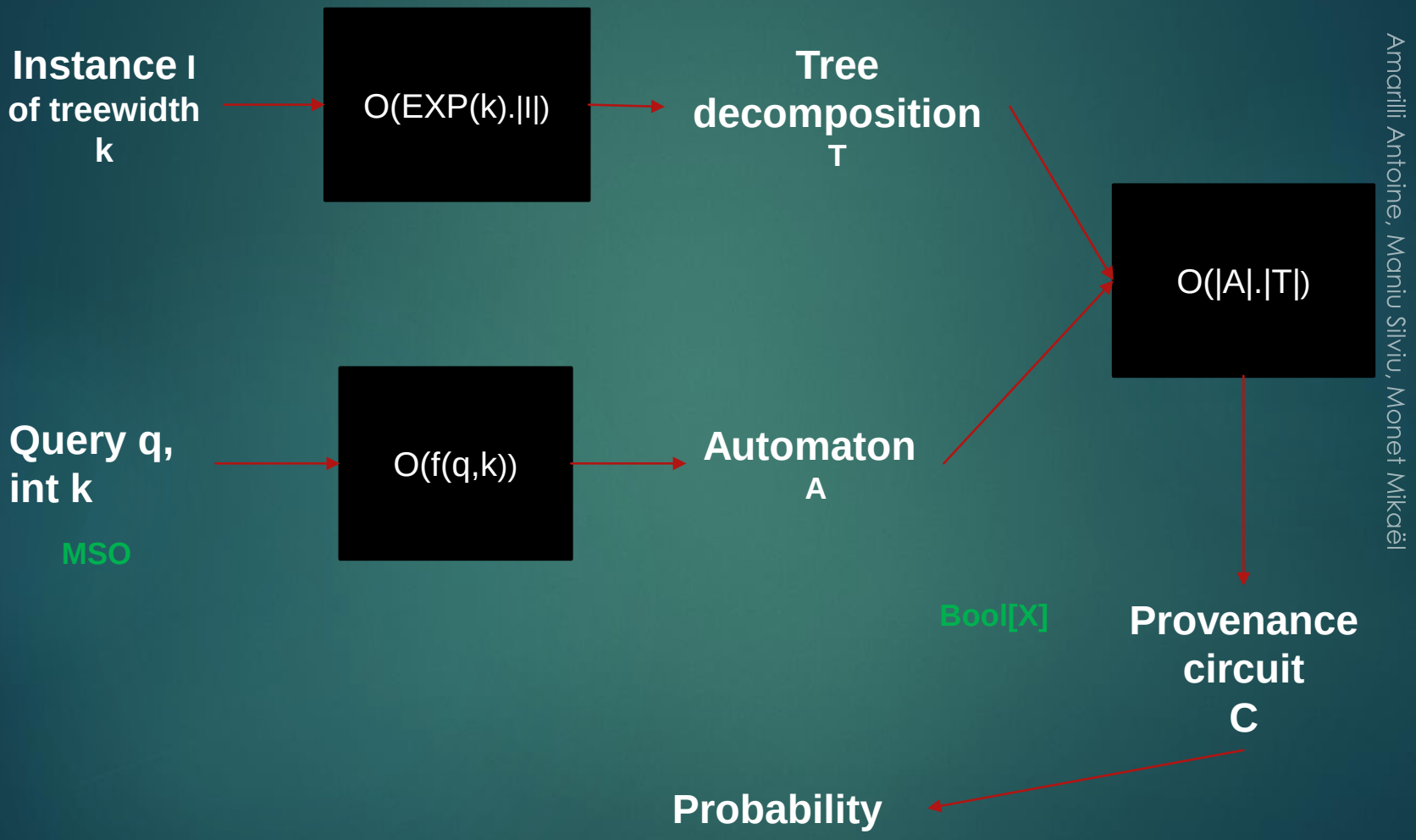
Amarilli Antoine, Maniu Silviu, Monet Mikaël



Amarilli Antoine, Maniu Silviu, Monet Mikaël



Amarilli Antoine, Maniu Silviu, Monet Mikaël

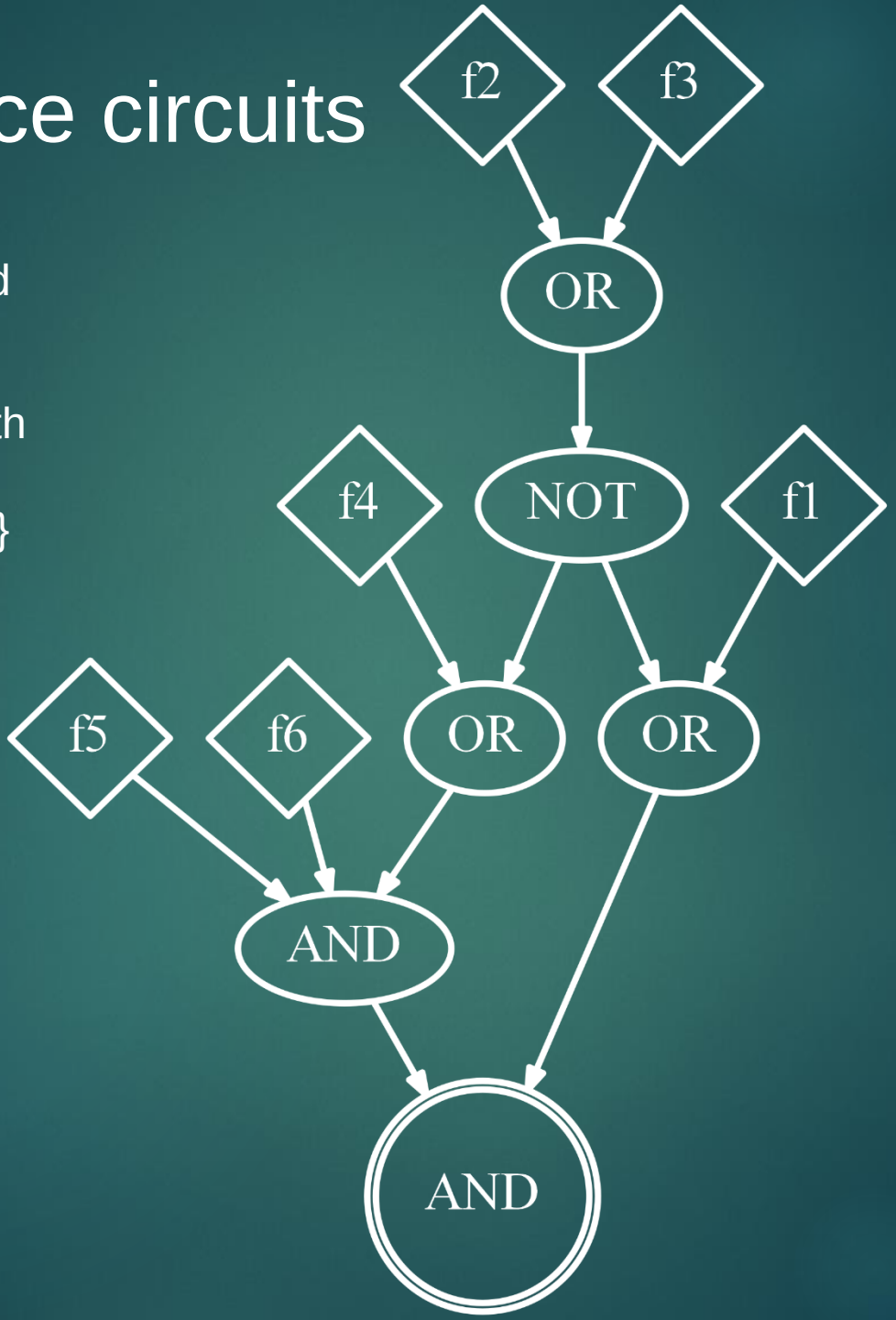


Amarilli Antoine, Maniu Silviu, Monet Mikaël

Provenance circuits

Some query q, fixed

Some instance I with facts
{f1, f2, f3, f4, f5, f6}

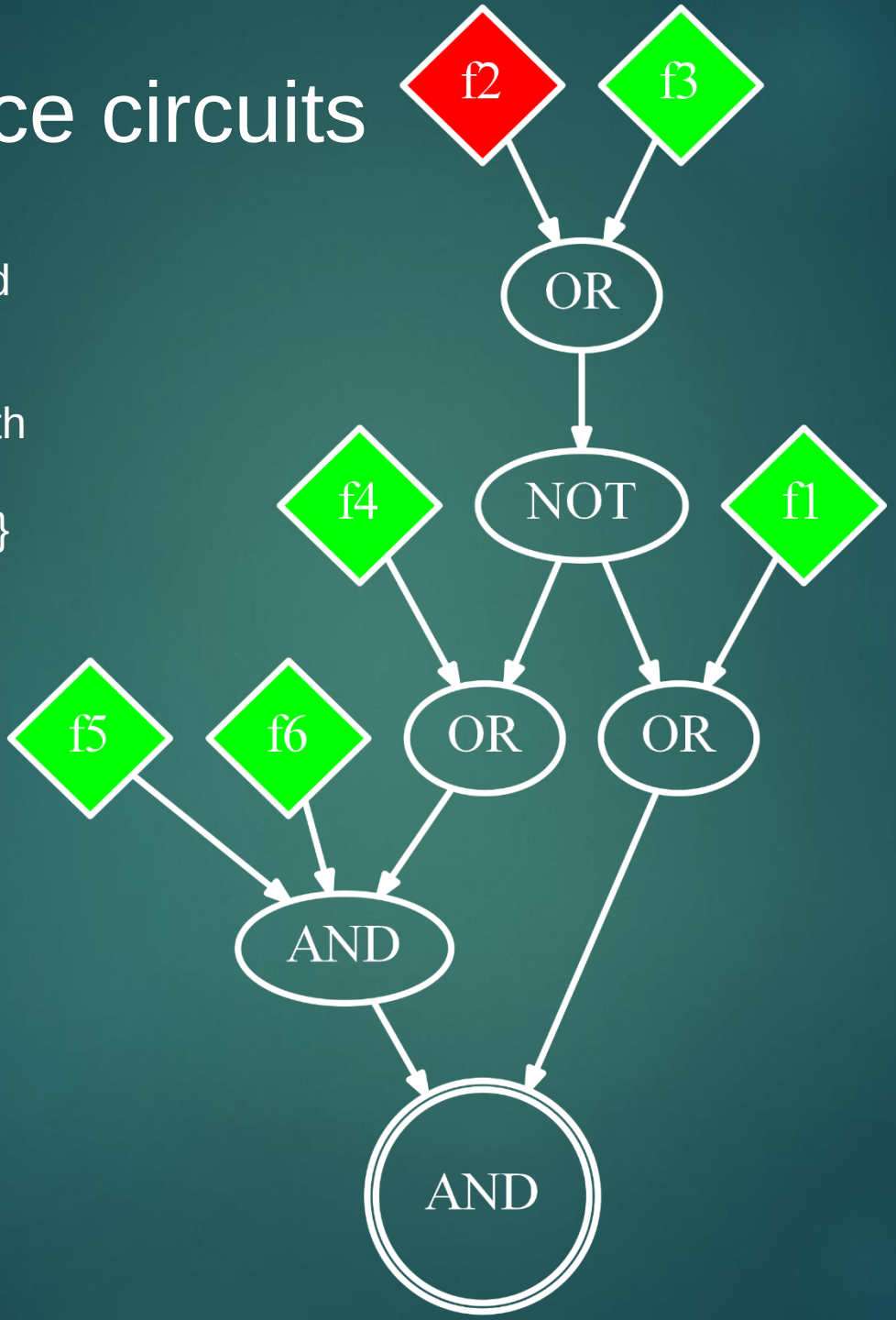


Amarilli Antoine, Maniu Silviu, Monet Mikaël

Provenance circuits

Some query q, fixed

Some instance I with facts
{f1, f2, f3, f4, f5, f6}

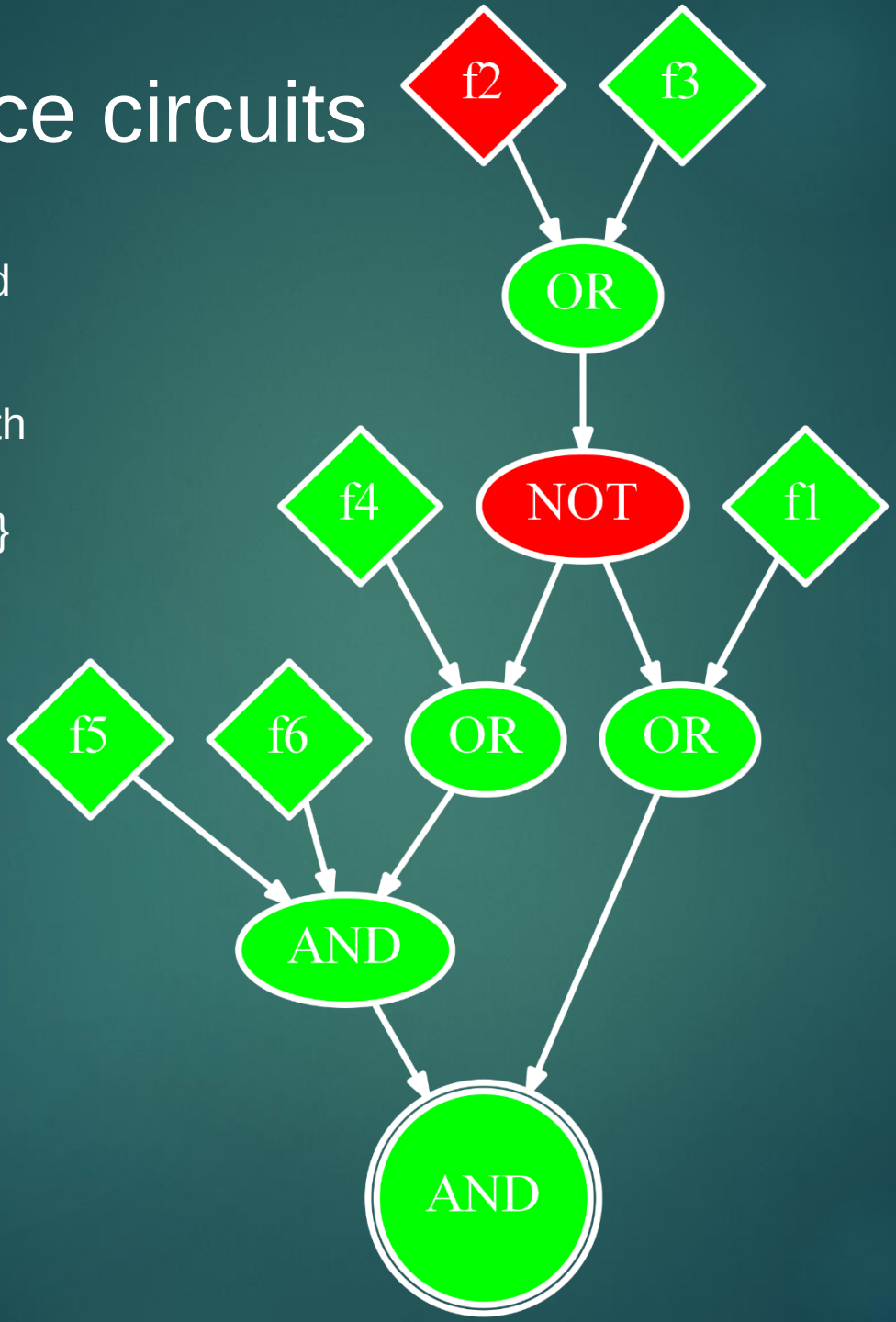


Amarilli Antoine, Maniu Silviu, Monet Mikaël

Provenance circuits

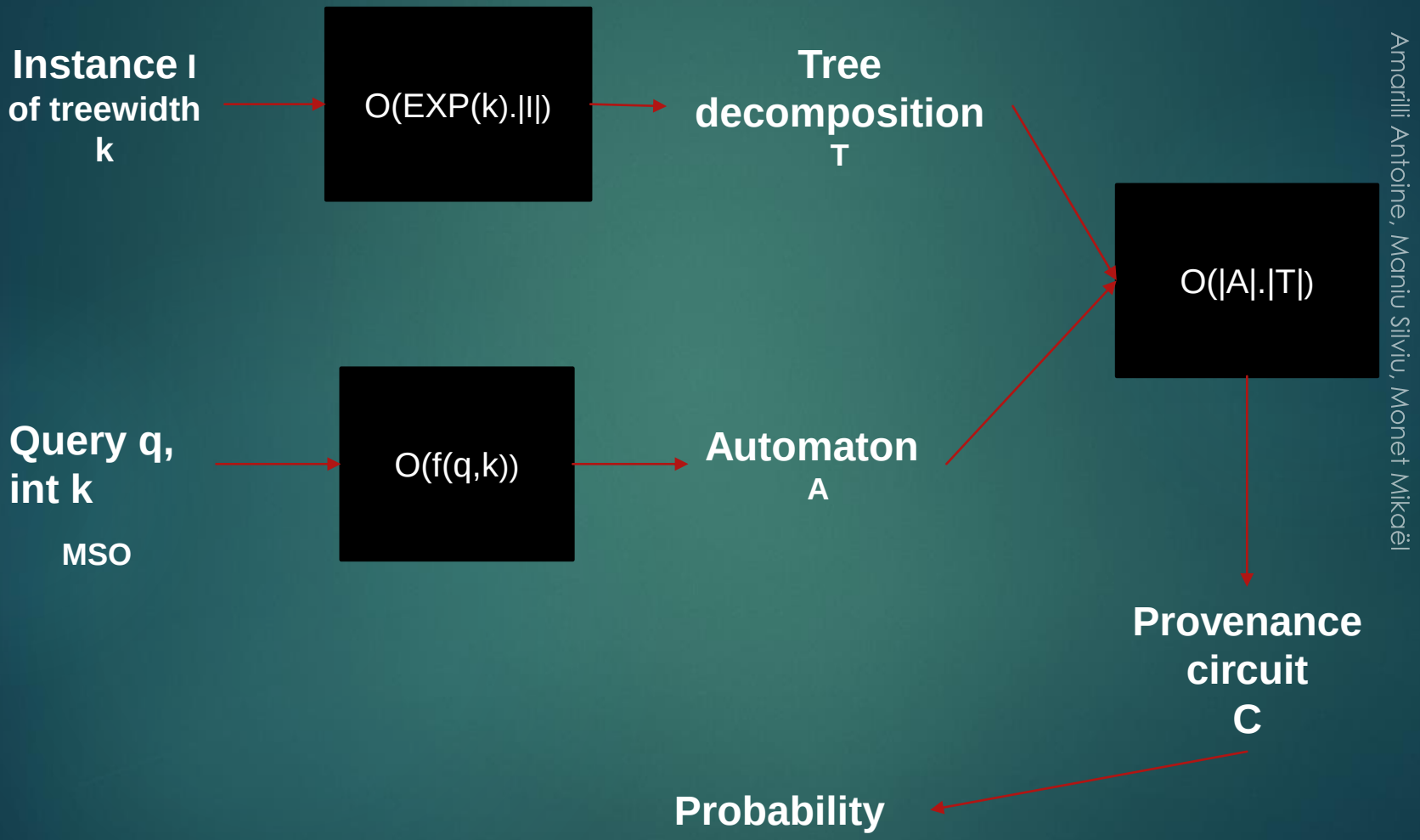
Some query q, fixed

Some instance I with facts
{f1, f2, f3, f4, f5, f6}



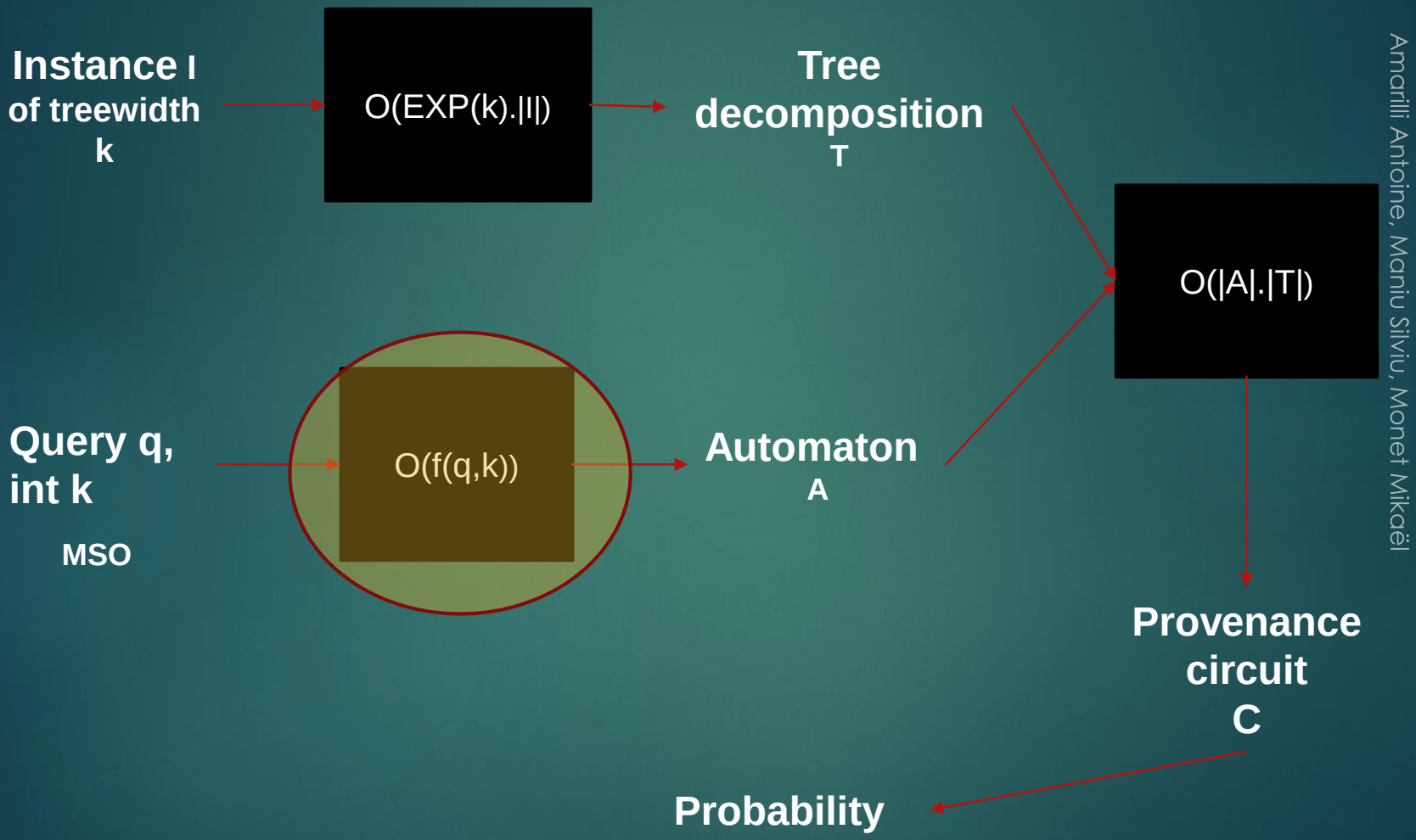
Amarilli Antoine, Maniu Silviu, Monet Mikaël

Problems



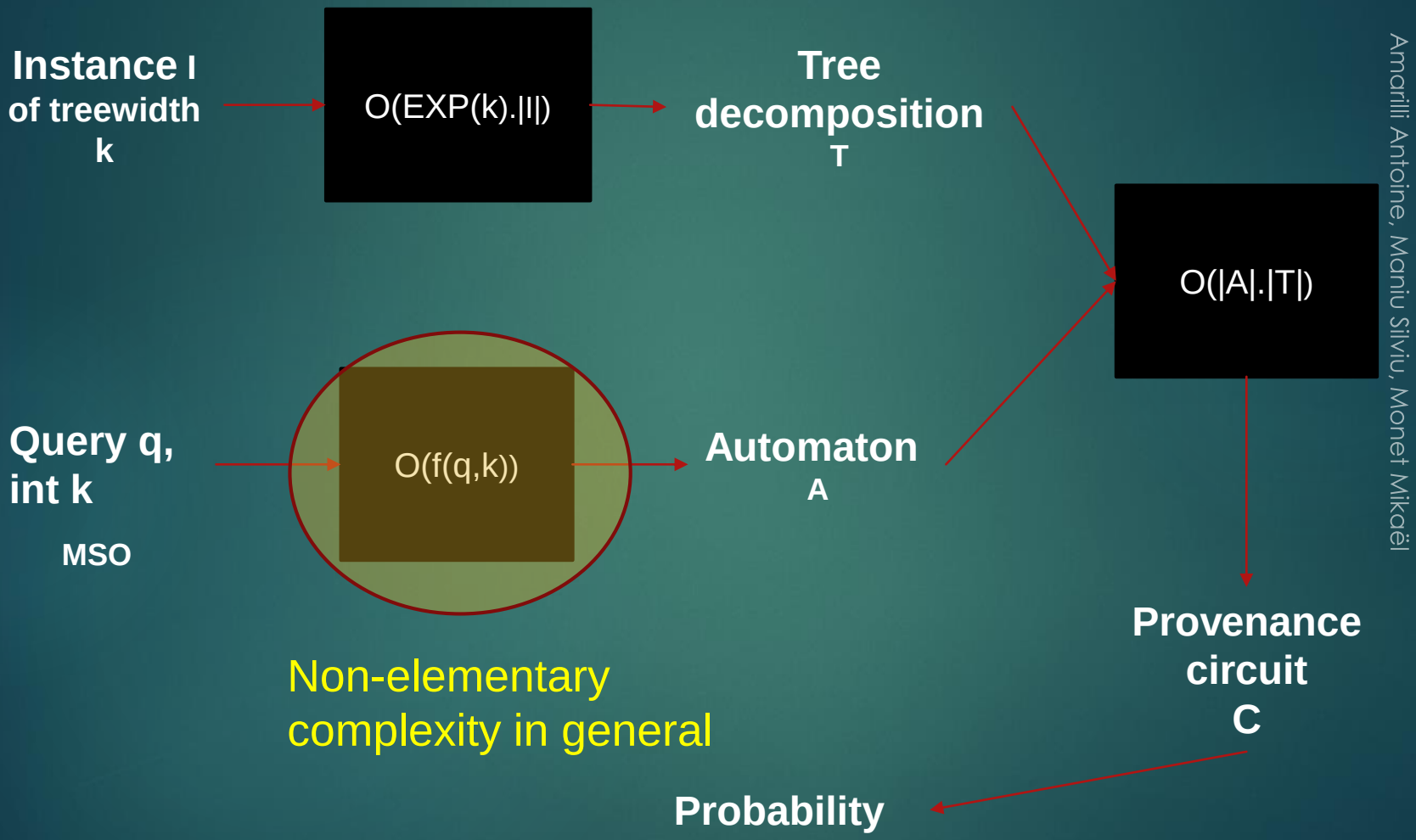
Amarilli Antoine, Maniu Silviu, Monet Mikaël

Problems



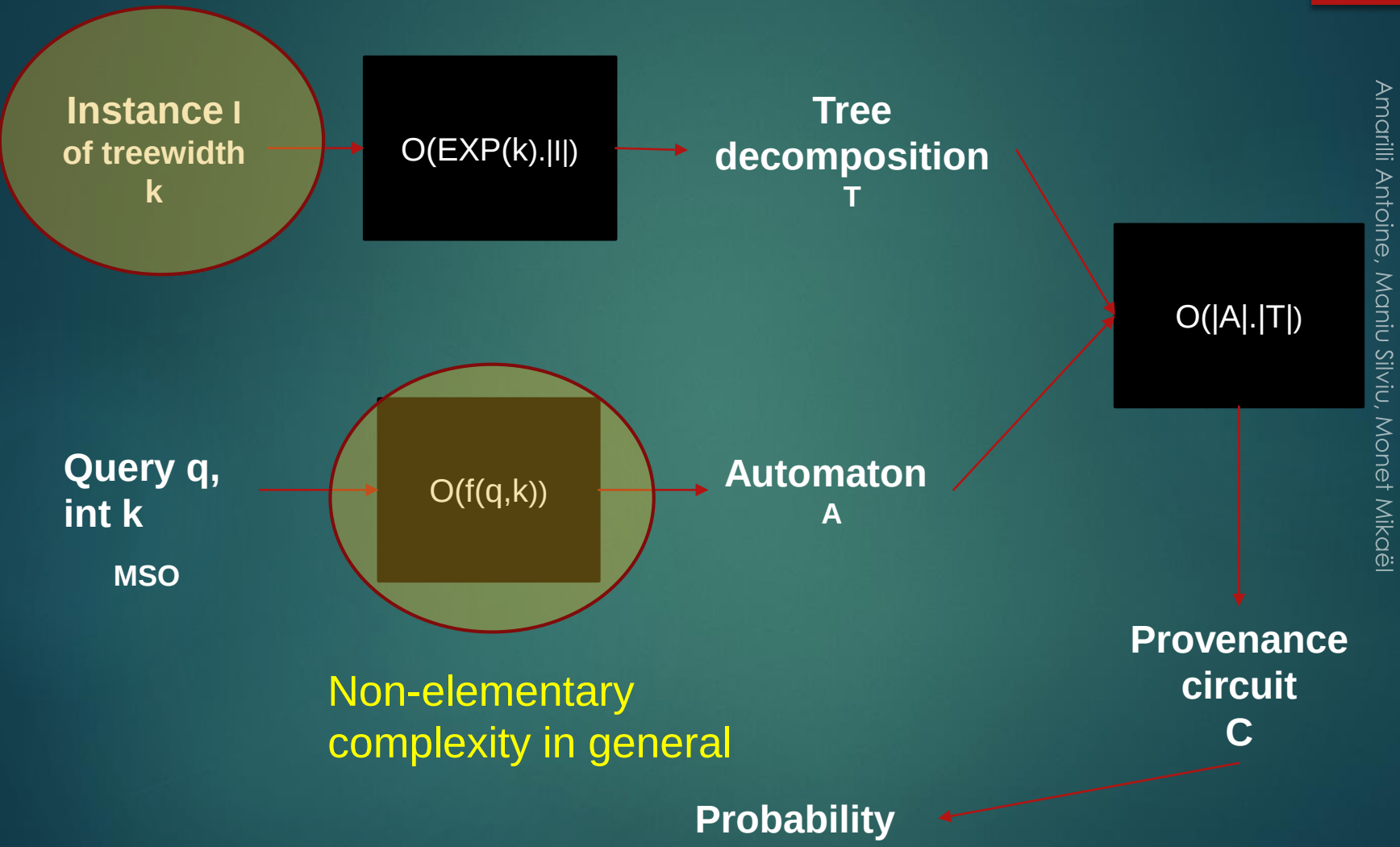
Amarilli Antoine, Maniu Silviu, Monet Mikaël

Problems



Amarilli Antoine, Maniu Silviu, Monet Mikaël

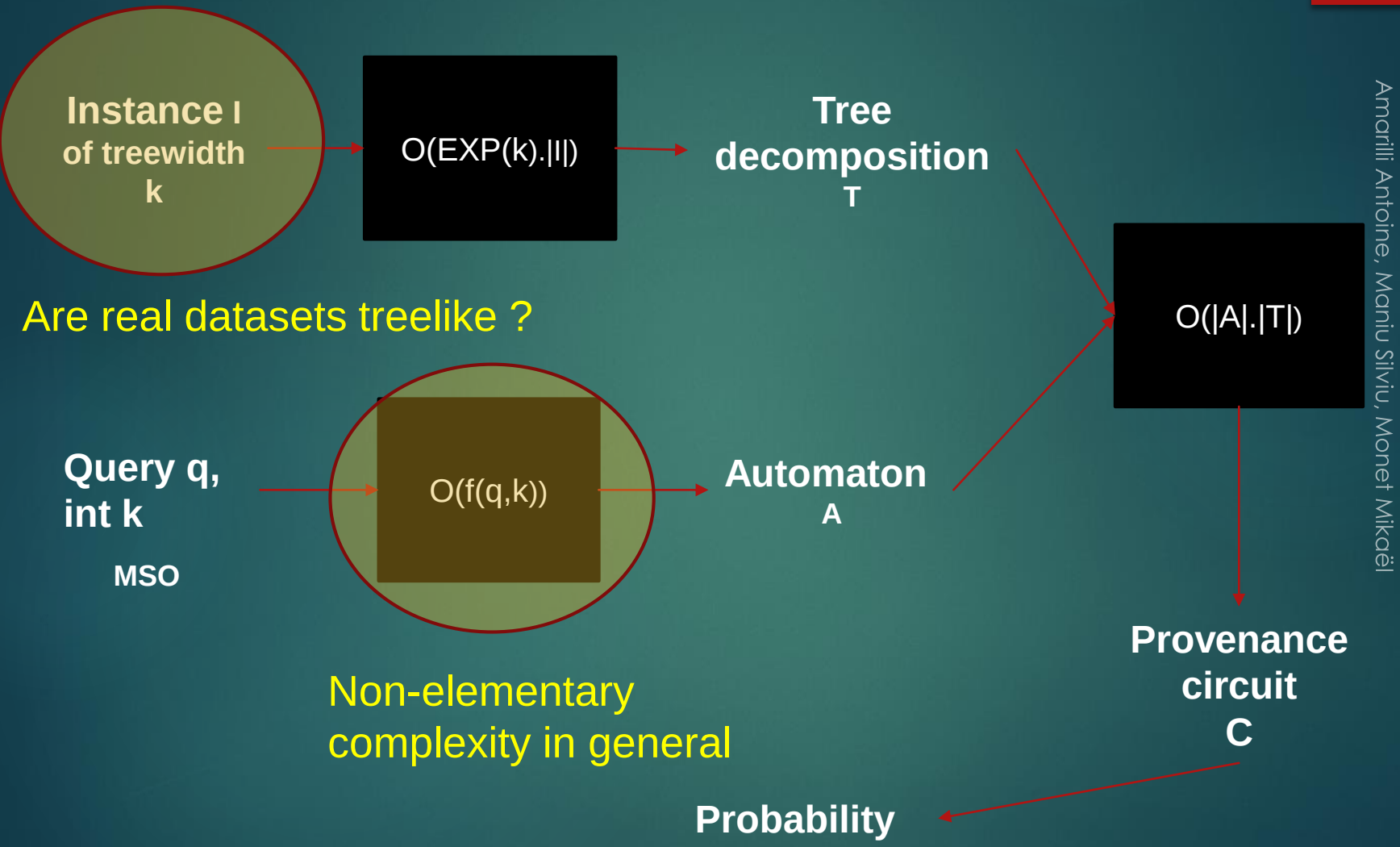
Problems



Non-elementary
complexity in general

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Problems



Amarilli Antoine, Maniu Silviu, Monet Mikaël

Current work

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Current work

- From bottom-up tree automtata to alternating two-way automata

Current work

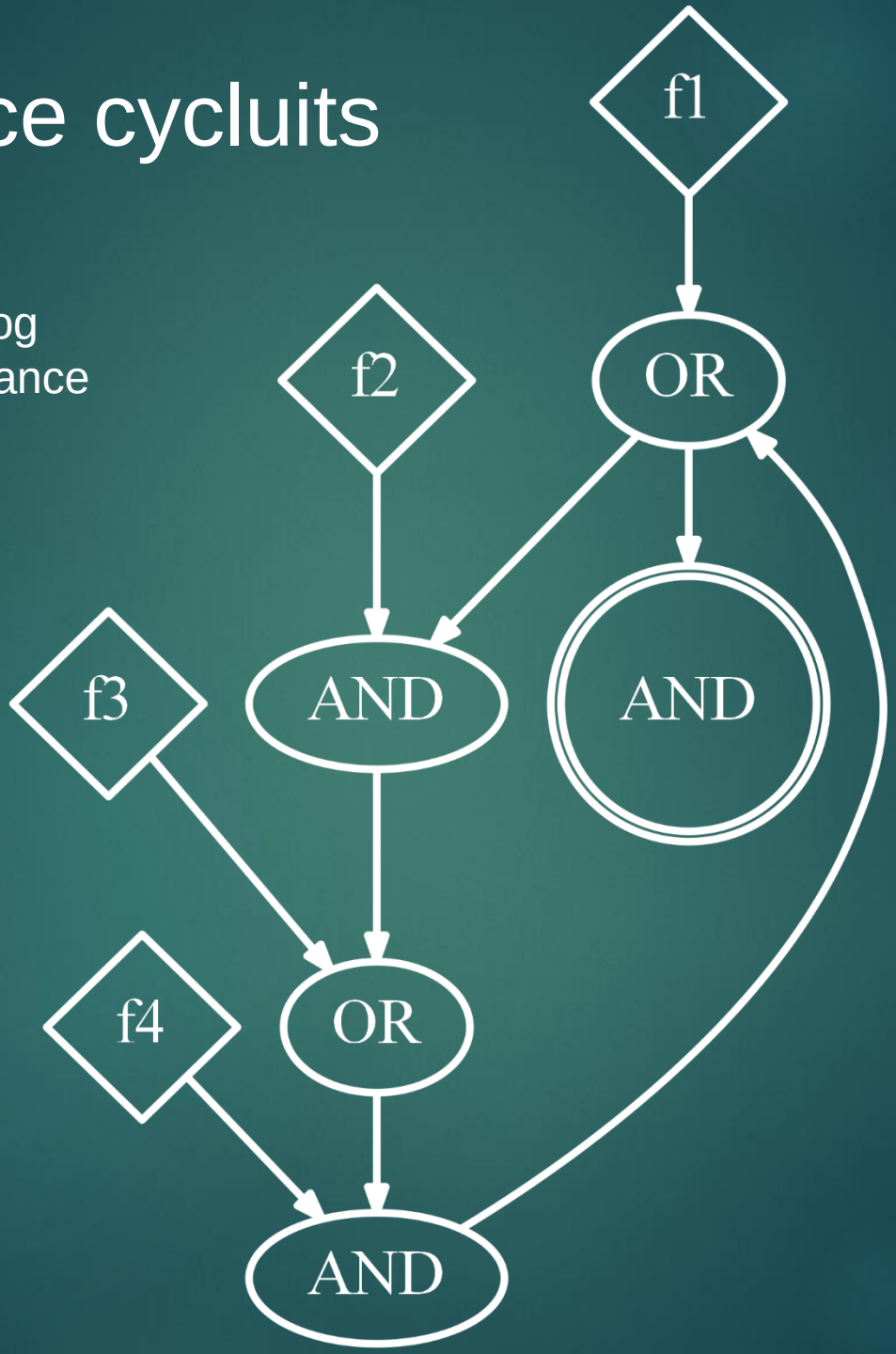
- From bottom-up tree automtata to alternating two-way automata
- Introduce Intensionally-Clique-Guarded Datalog (ICG-Datalog) parameterized by body-size

Current work

- From bottom-up tree automtata to alternating two-way automata
- Introduce Intensionally-Clique-Guarded Datalog (ICG-Datalog) parameterized by body-size
- Provenance as a cyclic circuit ! (cycluit)

Provenance cycluits

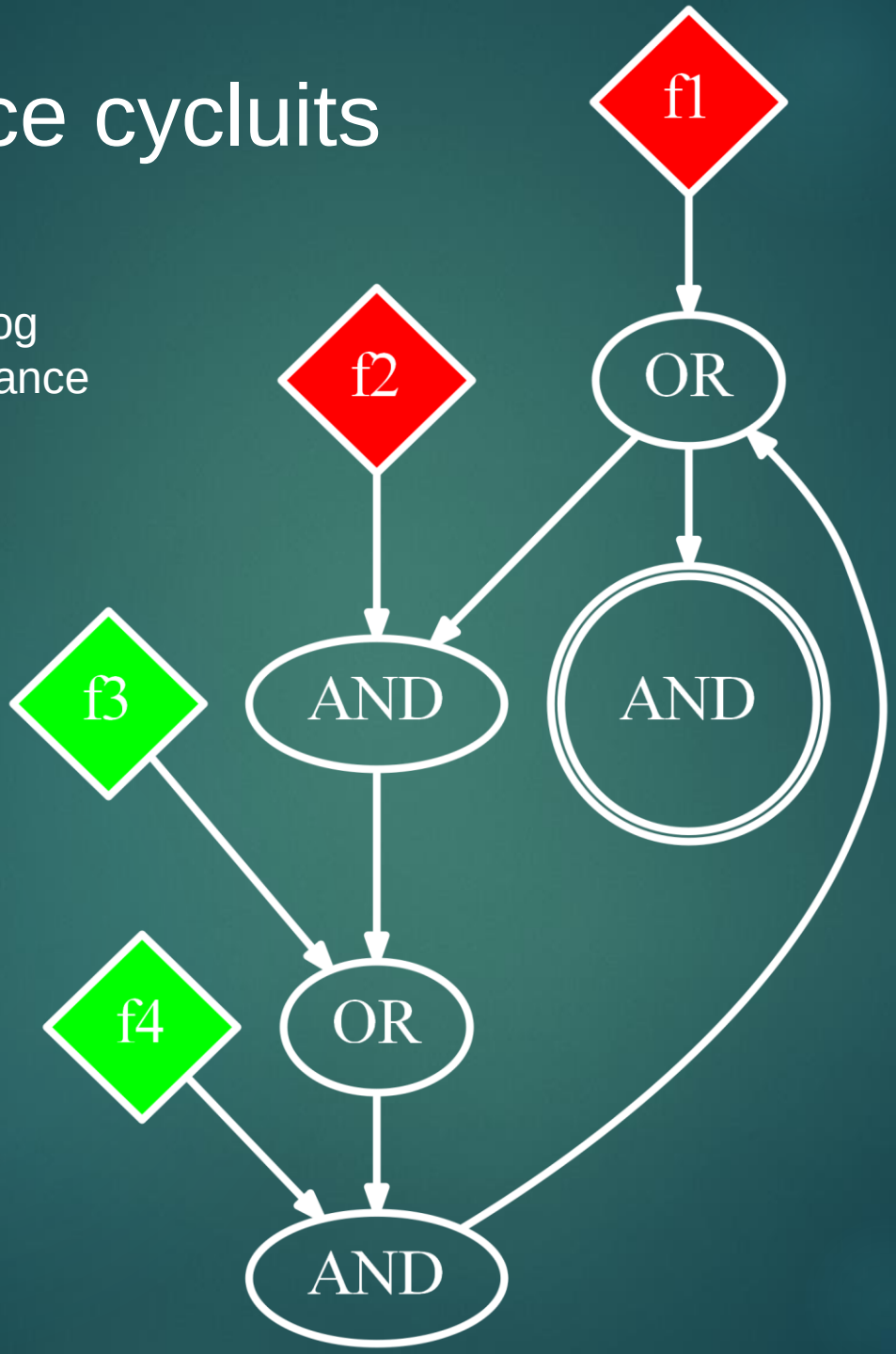
Some ICG-Datalog program, some instance I with facts {f1, f2, f3, f4}



Amarilli Antoine, Maniu Silviu, Monet Mikaël

Provenance cycluits

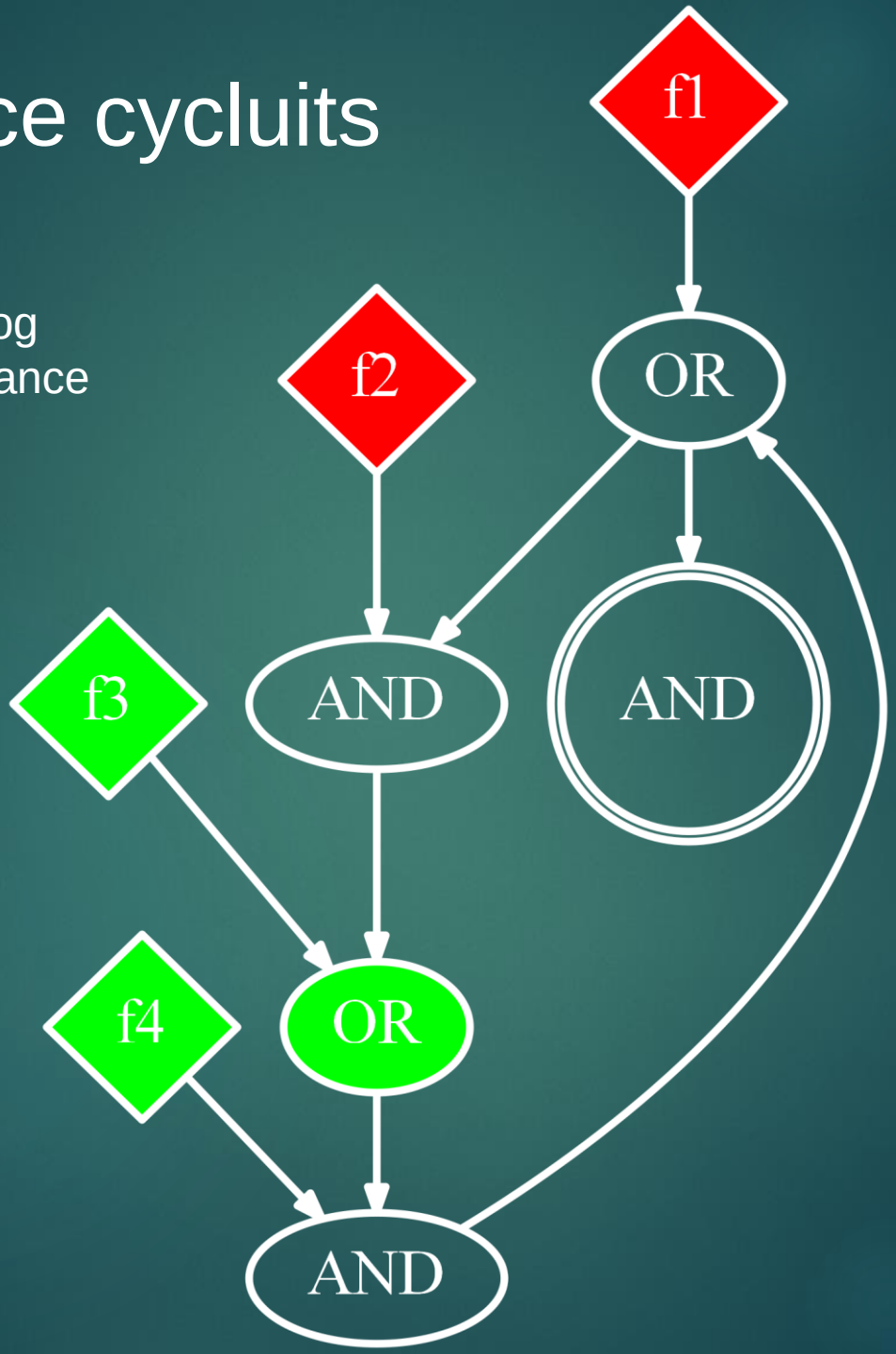
Some ICG-Datalog program, some instance I with facts {f1, f2, f3, f4}



Amarilli Antoine, Maniu Silviu, Monet Mikaël

Provenance cycluits

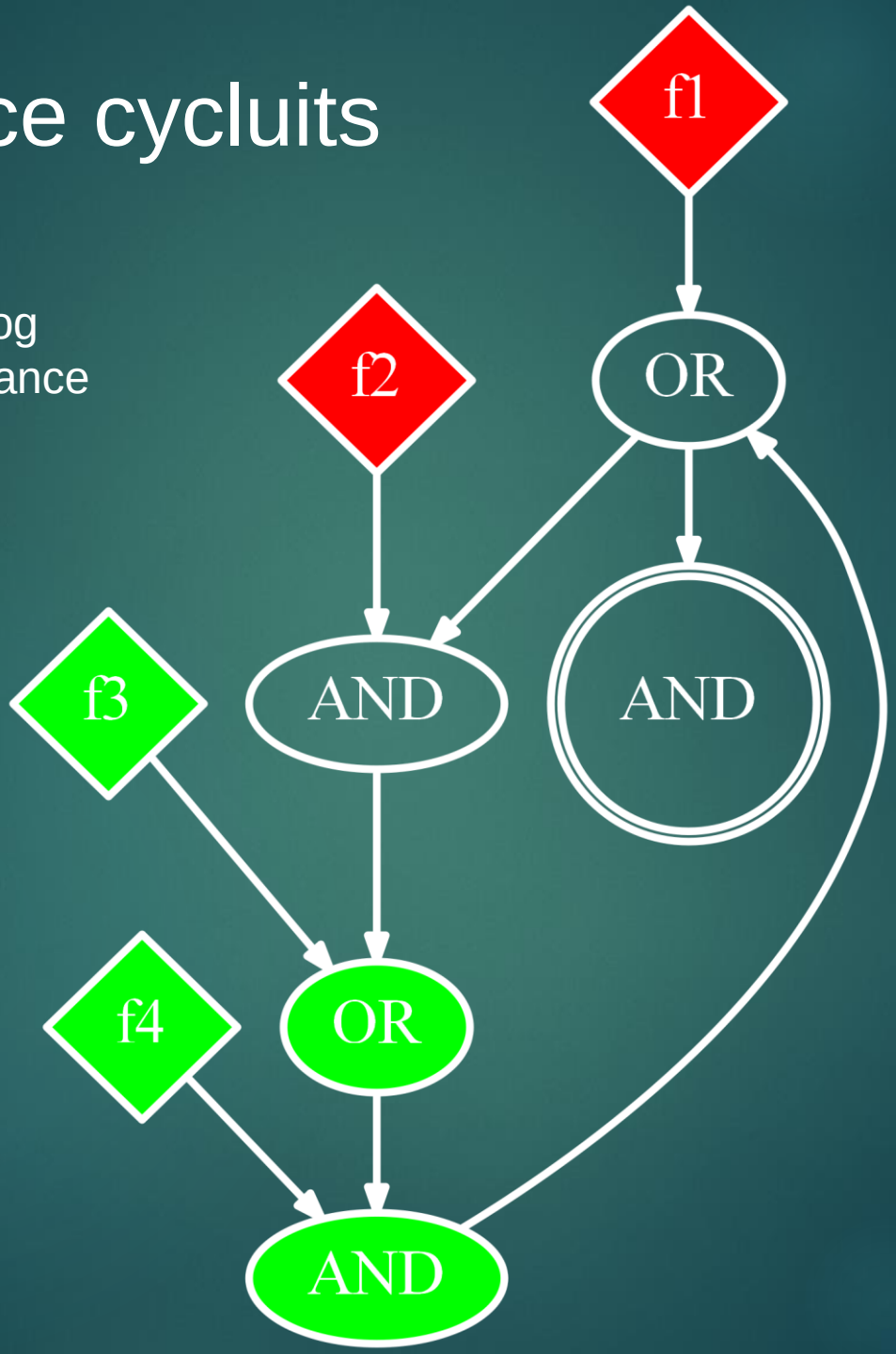
Some ICG-Datalog program, some instance I with facts {f1, f2, f3, f4}



Amarilli Antoine, Maniu Silviu, Monet Mikaël

Provenance cycluits

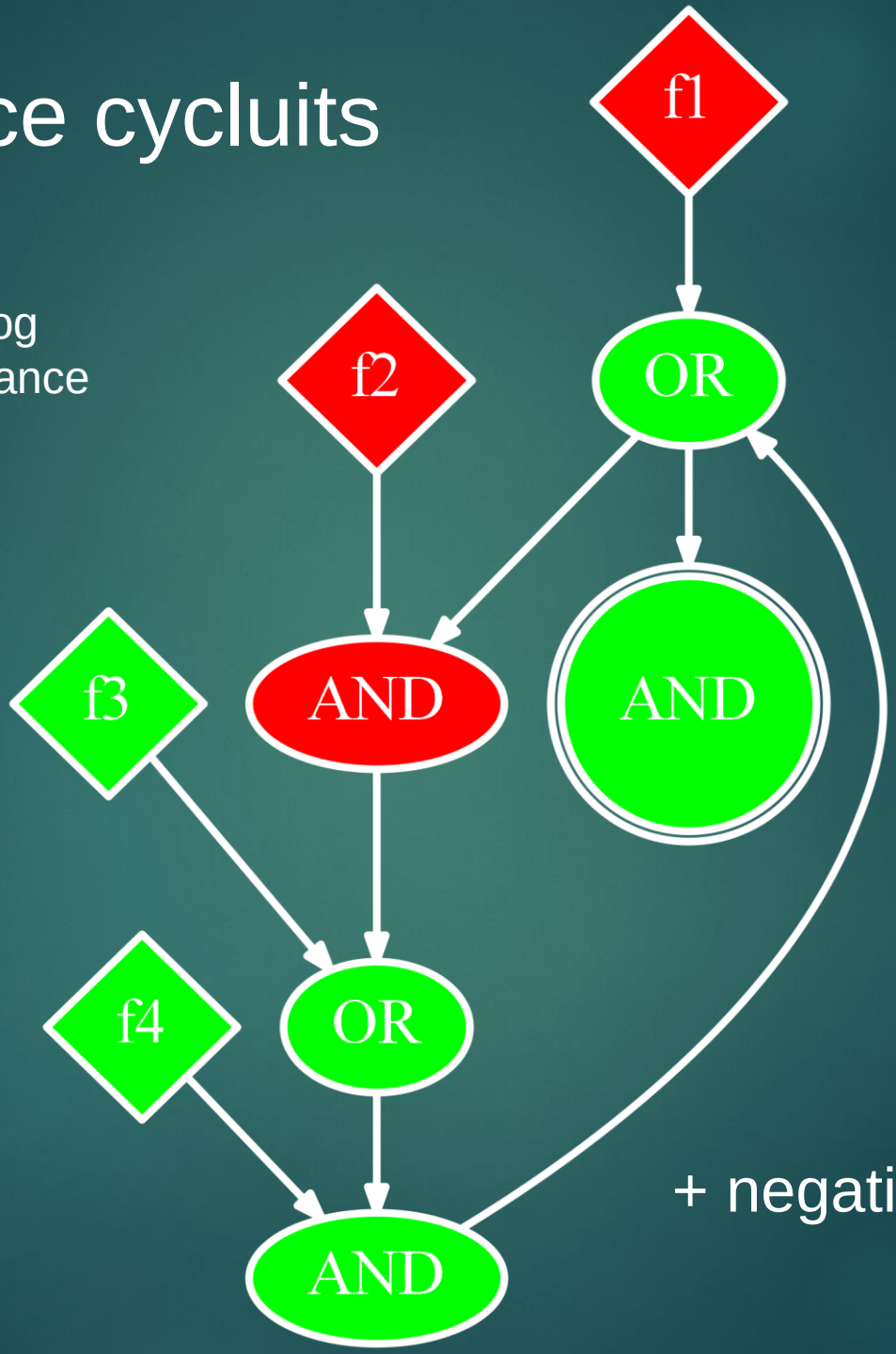
Some ICG-Datalog program, some instance I with facts {f1, f2, f3, f4}



Amarilli Antoine, Maniu Silviu, Monet Mikaël

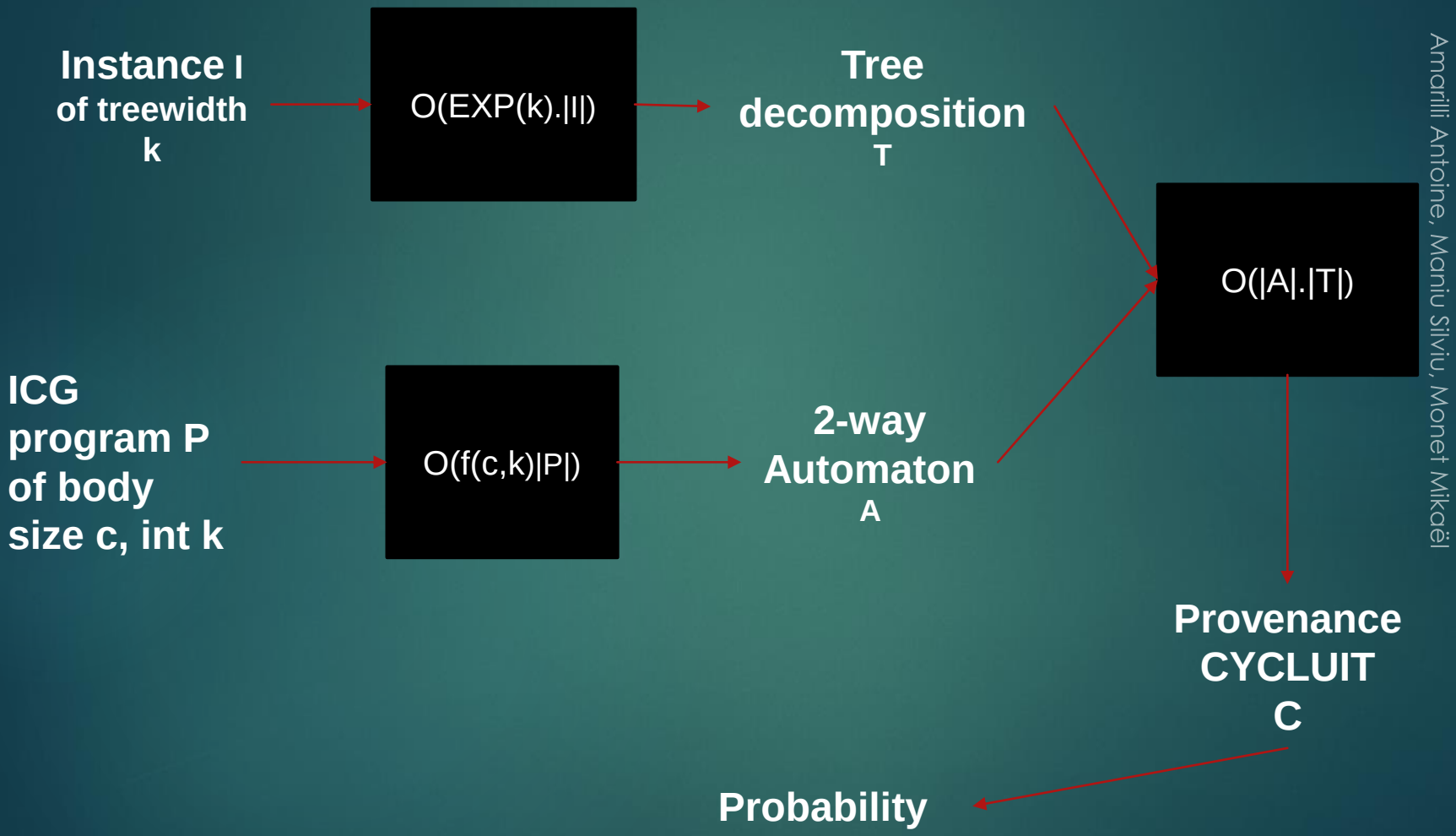
Provenance cycluits

Some ICG-Datalog program, some instance I with facts {f1, f2, f3, f4}

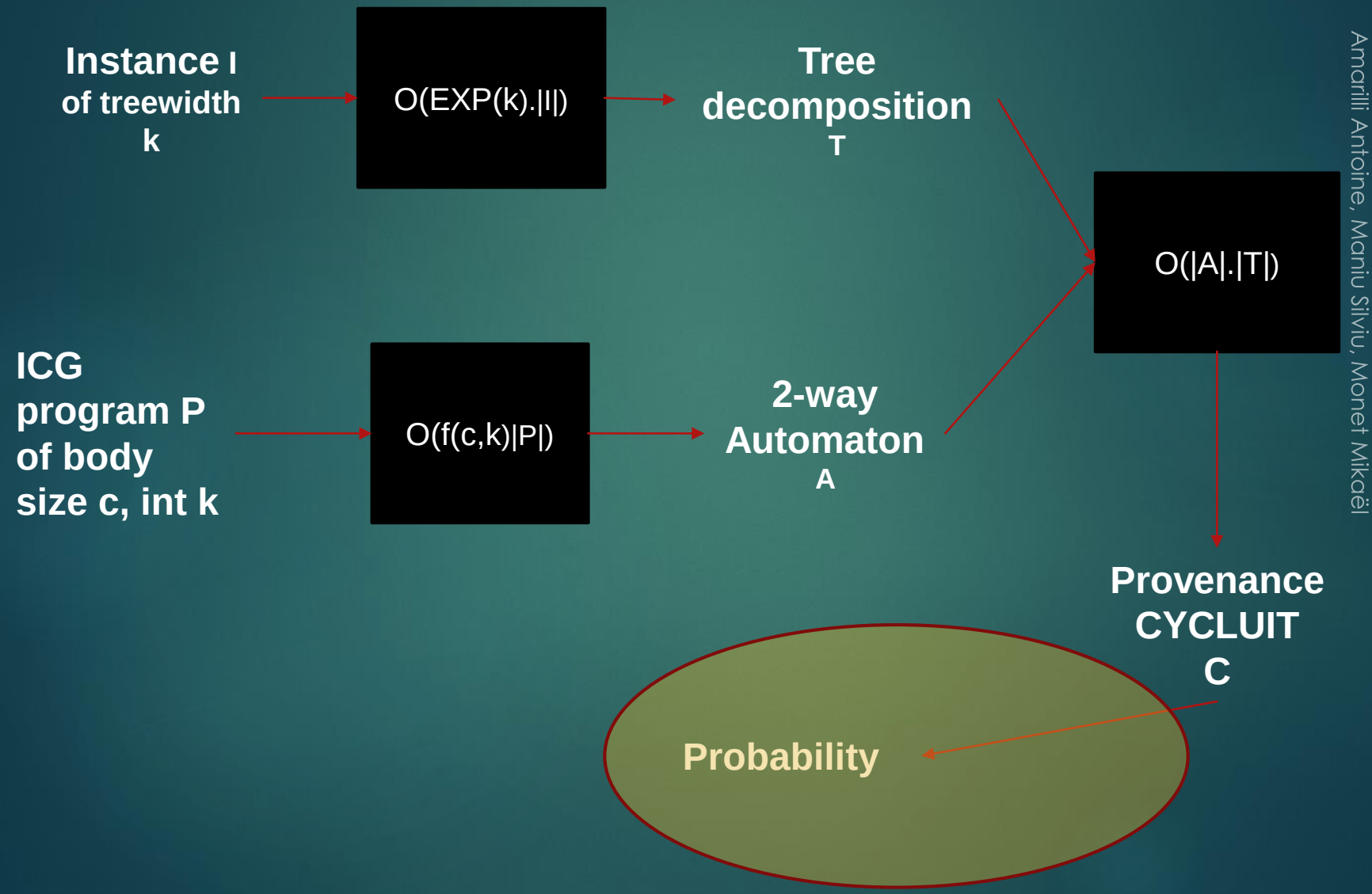


+ negations (stratified)

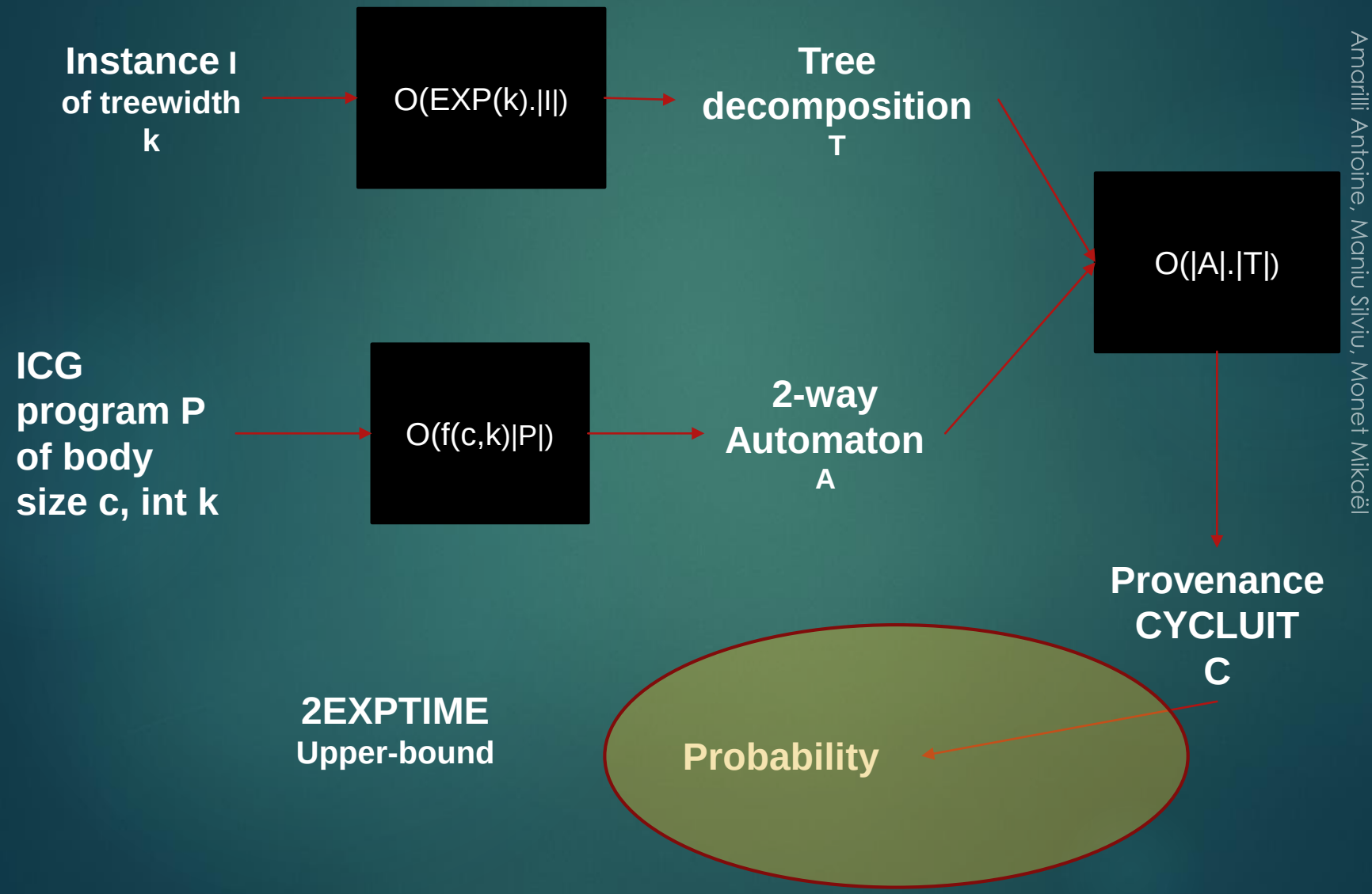
Amarilli Antoine, Maniu Silviu, Monet Mikaël



Amarilli Antoine, Maniu Silviu, Monet Mikaël



Amarilli Antoine, Maniu Silviu, Monet Mikaël



Amarilli Antoine, Maniu Silviu, Monet Mikaël

Bad news…

- We proved that:

Path queries on tree instances (treewidth = 1) is already #P-hard. (reduction from #MONOTONE-2-SAT)

- Still, we obtain a 2EXPTIME combined complexity upperbound

Treelike datasets

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Treelike datasets

- Transportation networks

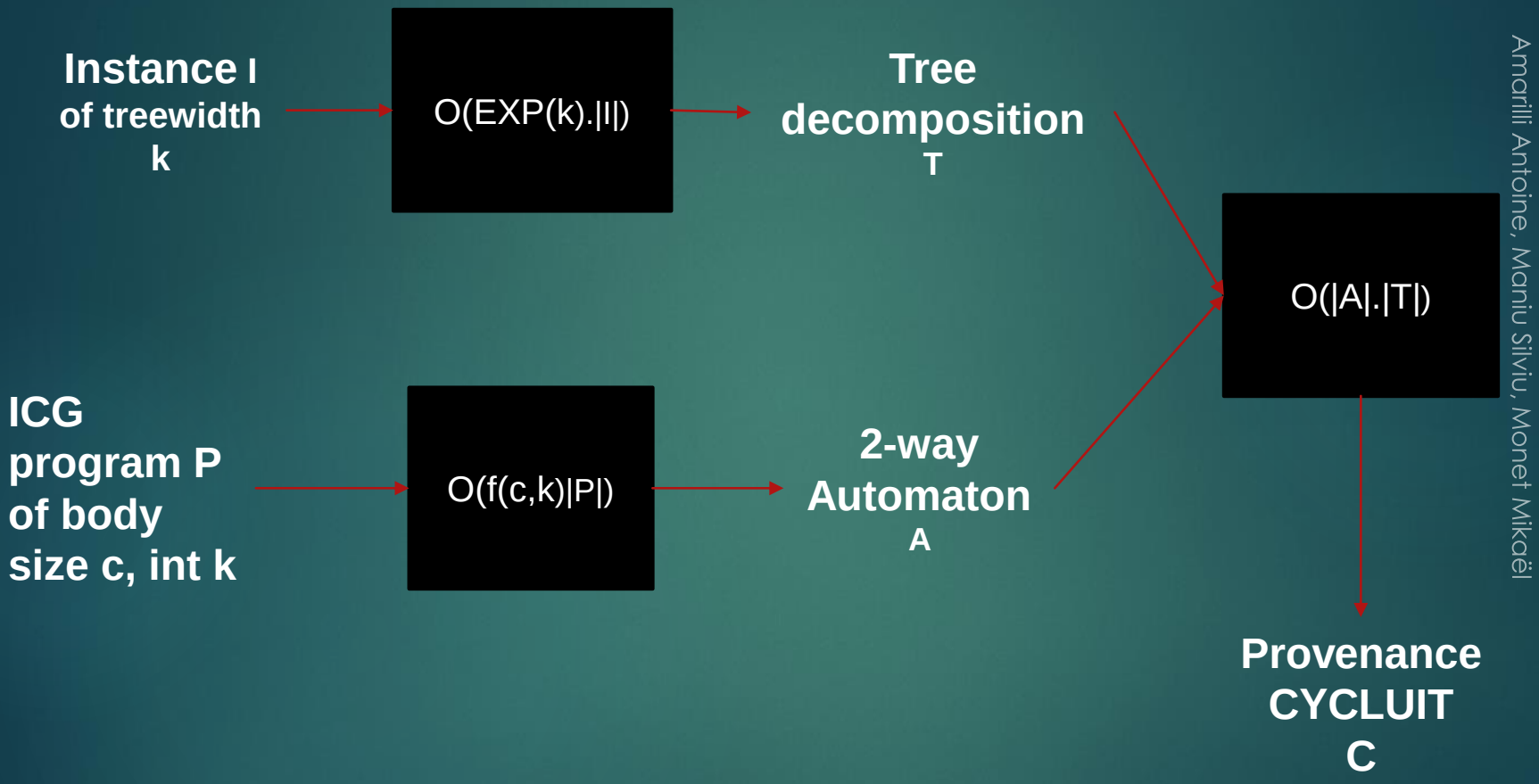
Treelike datasets

- Transportation networks
- Partial decompositions

Amarilli Antoine, Maniu Silviu, Monet Mikaël

Treelike datasets

- Transportation networks
- Partial decompositions
- Query-specific decompositions



Amarilli Antoine, Maniu Silviu, Monet Mikaël