25th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data
DOLAP 2023

Cube query answering via the results of previous cube queries

Panos Vassiliadis

https://www.cs.uoi.gr/~pvassil/projects/olap_III/index.html

Department of Computer Science and Engineering
University of Ioannina, Hellas

The author has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call Research - Create - Innovate (prj code:T2EDK-02848).

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο
Περιφερειακής Ανάπτυξης

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΟΙΚΟΝΟΜΙΑΣ & ΑΝΑΠΤΥΞΗΣ
ΕΙΔΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΤΠΑ & ΤΣ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΑνΕΚ

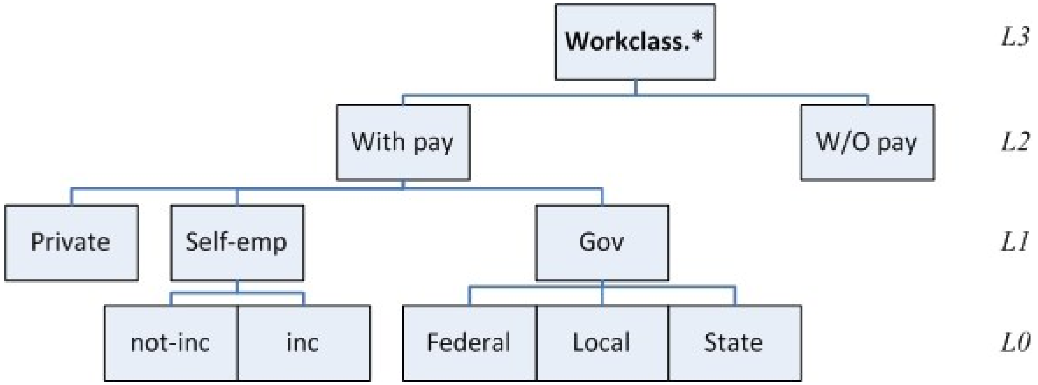
ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ
ΚΑΙΝΟΤΟΜΙΑ

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

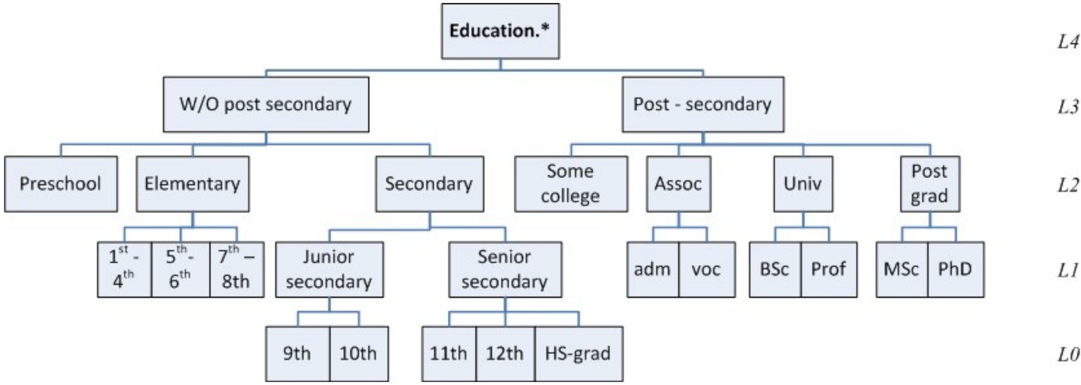
Hierarchical multidimensional data spaces are future, not past

- Multidimensional data are…
- … intuitive, for people to work with
- … powerful, because of the inherent hierarchies



Workclass.*
With pay
W/O pay
Private
Self-emp
Gov
not-inc
inc
Federal
Local
State
L3
L2
L1
L0

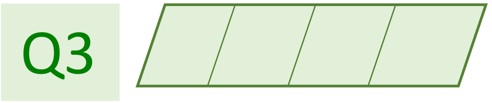
Simplicity is the fundamental virtue of cubes and cube queries



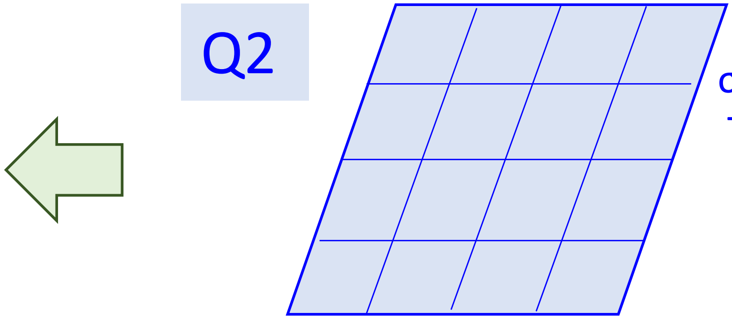
Education.*
W/O post secondary
Post - secondary
Preschool
Elementary
Secondary
Some college
Assoc
Univ
Post grad
1st - 4th
5th - 6th
7th – 8th
Junior secondary
Senior secondary
adm
voc
BSc
Prof
MSc
PhD
9th
10th
11th
12th
HS-grad
L4
L3
L2
L1
L0

Problem statement

- How can we compute the contents of a new cube query, by reusing the existing contents of the result of a previous cube query?



σ:
Time.Year ∈ {2019},
WC.L2∈{WithPay},
Edu.L3∈{Post-Secondary}



σ:
Time.Year ∈ {2018,2019},
WC.ALL∈{All},
Edu.L3∈{Post-Secondary}

Schema:
[Time.Year, WC.L2, Edu.L2],[sum(TaxPaid)]

Schema:
[Time.Month, WC.L1, Edu.L2], [sum(TaxPaid)]

But … hasn’t this been solved years ago? …

- **Query containment**: determine whether the set of tuples computed by a query Q is always a subset of the set of tuples produced by a query Q’ independently of the database contents
- **View usability**: determine whether a query Q defined over a set of relations SQ can be answered via a view V (and possibly a set of auxiliary relations SV ⊆ SQ) such that the resulting set of tuples is identical, independently of the contents of the underlying database
- **Query rewriting**: how the query Q must be rewritten in order to be answered via V
- Long list of papers for the relational case; only a couple of papers for query expressions involving different levels in the hierarchy with (a) simple expressions or (b) a reasoner-based approach
- We need a simple and straightforward way to …
  a. … check whether a cube is usable for the computation of another cube, via formal tests over the combination of selection conditions and groupers, for a fairly broad range of expressions…
  b. .. …perform the computation

How can we compute the contents of a new cube query, by reusing the existing contents of the result of a previous cube query?

- In this paper, we address the *usability problem* of computing a new cube query c^n from the cells of a previous one, c^b, defined at a different level of abstraction; we introduce the respective test as well as a rewriting algorithm
- This is part of a broader effort to provide a model and an algebra on how two cubes can be related

LONG VERSION

Panos Vassiliadis. A Cube Algebra with Comparative Operations: Containment, Overlap, Distance and Usability

https://arxiv.org/abs/2203.09390

A Cube Algebra with Comparative Operations:
Containment, Overlap, Distance and Usability

Cube query usability and rewriting: the recipe

- How can we compute the result of a query on the basis of the result of a previous query, in the context of a multidimensional hierarchical space?
- The recipe:
  - Provide a formal **model** for data and queries; pay attention to define query **semantics** with respect to the most detailed data of the multidimensional space
  - Introduce **equivalent expressions** at different levels of detail
  - Isolate the cases where the **grouper- and filter- parts of a query can collaborate** to produce meaningful results
  - Introduce a usability **test** and a rewriting **algorithm**

Model for data and queries

- We assume **dimensions**, with **hierarchies** of **levels** as context for the facts
- The Cartesian product of dimensions creates a **multidimensional space**, within which, **cubes** carrying measures exist.
- A **cube query** specifies
  a. the detailed data set over which it is imposed,
  b. the selection condition that isolates the records that qualify for further processing,
  c. the aggregator levels, that determine the level of coarseness for the result, and
  d. an aggregation over the measures of the underlying cube that accompanies the aggregator levels in the final result.

Cube Queries

q = DS^0, ϕ, $[L_1, \ldots, L_n, M_1, \ldots, M_m]$, $[agg1\,(M^0{}_1), \ldots, aggm(M^0{}_m)]$

SELECT $L_1, \ldots, L_n$, $agg_1(M_1^0)$ AS $M_1, \ldots, agg_1(M_m^0)$ AS M_m
FROM DS^0 NATURAL JOIN D$_1$ . . . NATURAL JOIN D$_n$
WHERE ϕ^0
GROUP BY $L_1, \ldots, L_n$

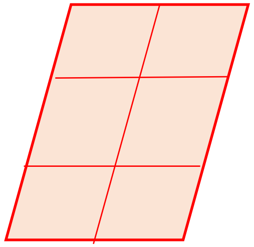
Selection condition ϕ : conjunctive selection condition with the following constraints:

(a) it involves a single composite conjunctive expression,

(b) there is exactly one atom per dimension of the schema of DS

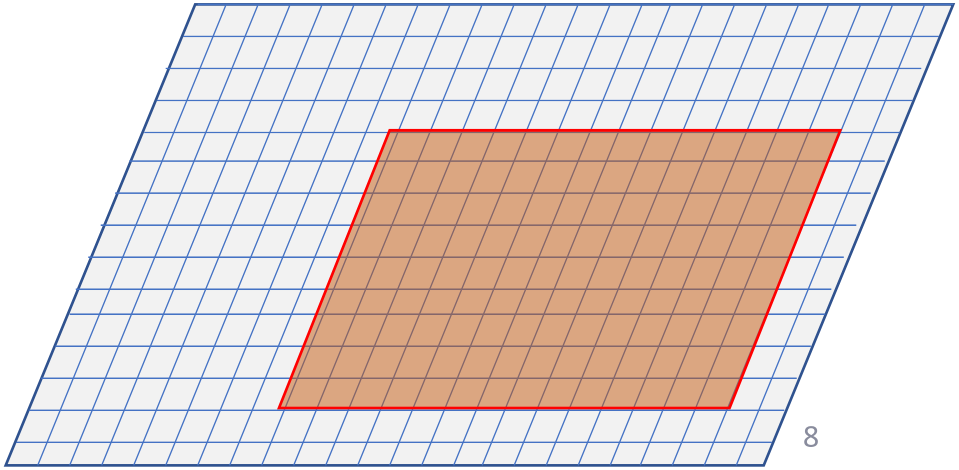
(c) the atoms of the condition are all of the form $L \in \{v_1, \ldots, v_k\}$, $v_i \in dom(L)$

Q1



Schema:
[Time.Month,
WC.L1, Edu.L2],
[sum(TaxPaid)]

σ:
Time.Year ∈ {2019,2020},
WC.L2∈{WithPay},
Edu.L2∈{Univ,Post-Grad}

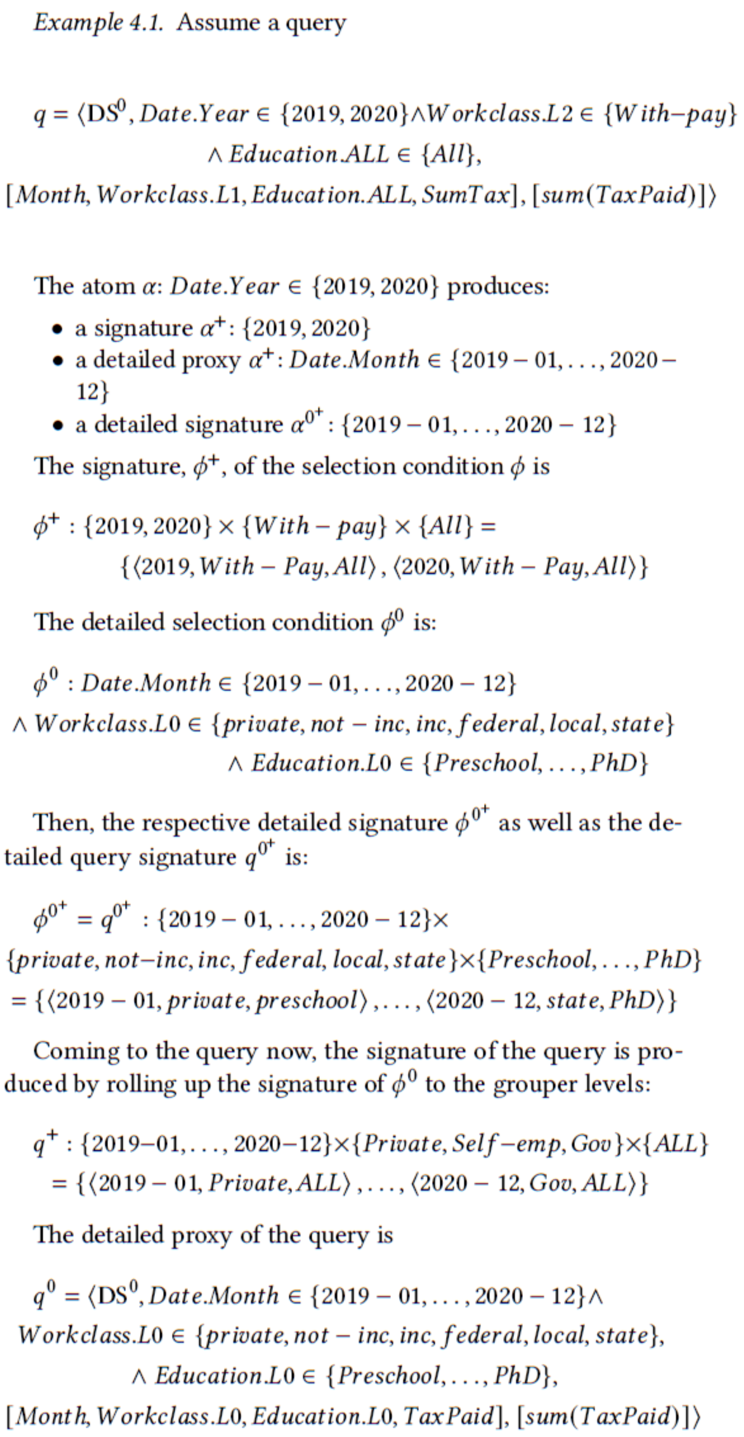


Equivalent detailed selection condition ϕ^0 : each atom of ϕ is mapped to

$$L_0 \in U, \text{ with } U = \bigcup_{i=1}^{k} desc_L^{L_0}(v_i)$$

Proxies: what does equivalent mean?

- A **proxy** is an equivalent expression at a different level of detail that by construction covers exactly the same subset of the multidimensional space, albeit at different level of coarseness.
- The **signature** of a construct is a set of coordinates that characterize the subset of the multidimensional space "framed“ by the construct.
- **Areas** are sets of cells within the bounds of a signature.



Example 4.1. Assume a query

$$q = \langle \mathrm{DS}^0, Date.Year \in \{2019, 2020\} \wedge Workclass.L2 \in \{With{-}pay\}$$
$$\wedge\, Education.ALL \in \{All\},$$
$$[Month, Workclass.L1, Education.ALL, SumTax], [sum(TaxPaid)]\rangle$$

The atom α: $Date.Year \in \{2019, 2020\}$ produces:

- a signature α^+: $\{2019, 2020\}$
- a detailed proxy α^+: $Date.Month \in \{2019-01, \ldots, 2020-12\}$
- a detailed signature α^{0^+}: $\{2019-01, \ldots, 2020-12\}$

The signature, ϕ^+, of the selection condition ϕ is

$$\phi^+ : \{2019, 2020\} \times \{With - pay\} \times \{All\} =$$
$$\{\langle 2019, With - Pay, All\rangle, \langle 2020, With - Pay, All\rangle\}$$

The detailed selection condition ϕ^0 is:

$$\phi^0 : Date.Month \in \{2019-01, \ldots, 2020-12\}$$
$$\wedge\, Workclass.L0 \in \{private, not - inc, inc, federal, local, state\}$$
$$\wedge\, Education.L0 \in \{Preschool, \ldots, PhD\}$$

Then, the respective detailed signature ϕ^{0^+} as well as the detailed query signature q^{0^+} is:

$$\phi^{0^+} = q^{0^+} : \{2019-01, \ldots, 2020-12\} \times$$
$$\{private, not{-}inc, inc, federal, local, state\} \times \{Preschool, \ldots, PhD\}$$
$$= \{\langle 2019-01, private, preschool\rangle, \ldots, \langle 2020-12, state, PhD\rangle\}$$

Coming to the query now, the signature of the query is produced by rolling up the signature of ϕ^0 to the grouper levels:

$$q^+ : \{2019{-}01, \ldots, 2020{-}12\} \times \{Private, Self{-}emp, Gov\} \times \{ALL\}$$
$$= \{\langle 2019-01, Private, ALL\rangle, \ldots, \langle 2020-12, Gov, ALL\rangle\}$$

The detailed proxy of the query is

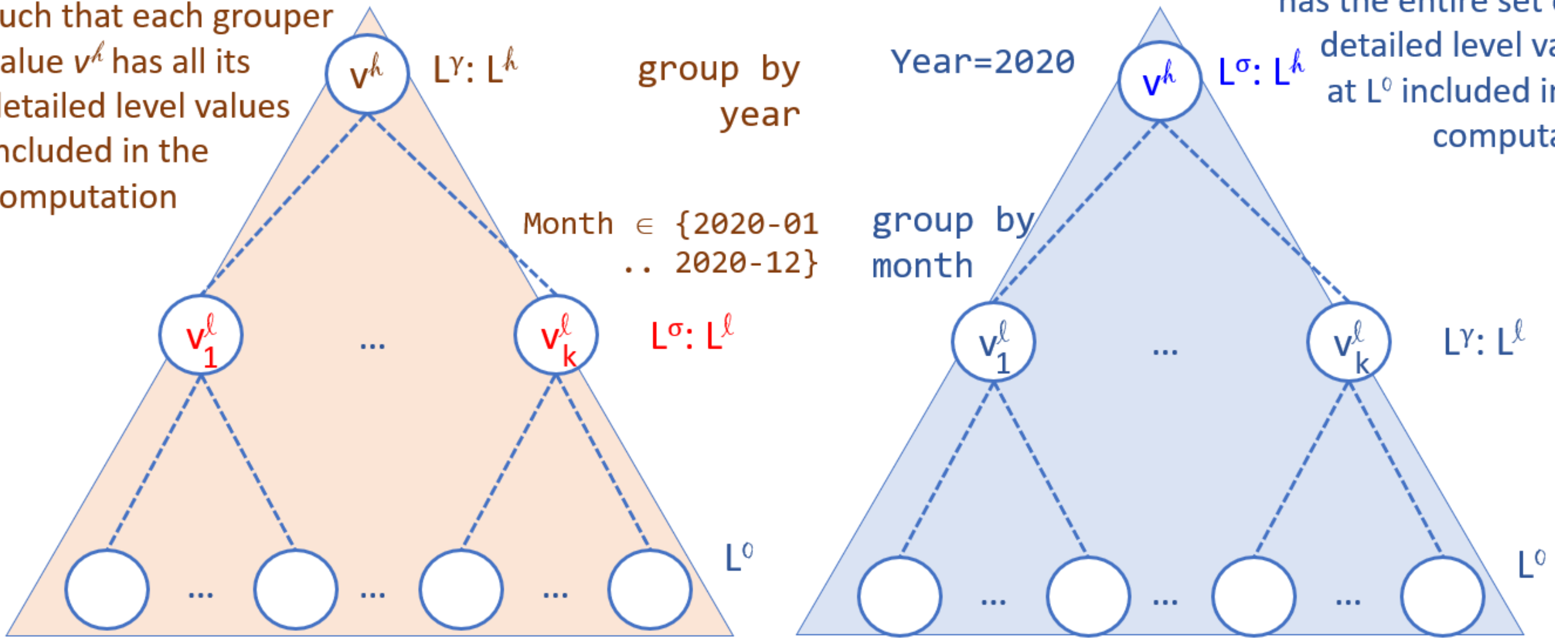
$$q^0 = \langle \mathrm{DS}^0, Date.Month \in \{2019-01, \ldots, 2020-12\} \wedge$$
$$Workclass.L0 \in \{private, not - inc, inc, federal, local, state\},$$
$$\wedge\, Education.L0 \in \{Preschool, \ldots, PhD\},$$
$$[Month, Workclass.L0, Education.L0, TaxPaid], [sum(TaxPaid)]\rangle$$

Model
Equiv. expr.
Rollability
Test & Algo

Perfect Rollability

If $L^{h} = L^{\gamma}$ and $L^{\ell} = L^{\sigma}$, i.e., $\mathbf{L^{\sigma} < L^{\gamma}}$, then *the entire set of desc(v^{h}) at L^{ℓ} must be included in the selection condition,* such that each grouper value v^{h} has all its detailed level values included in the computation

If $L^{h} = L^{\sigma}$ and $L^{\ell} = L^{\gamma}$, i.e., $\mathbf{L^{\gamma} \leq L^{\sigma}}$, then by definition, each grouper value v^{ℓ}_{i} has the entire set of its detailed level values at L^{0} included in the computation



v^{h} L^{γ}: L^{h}

group by year

Month ∈ {2020-01 .. 2020-12}

v^{ℓ}_{1} … v^{ℓ}_{k} L^{σ}: L^{ℓ}

… … … L^{0}

Year=2020

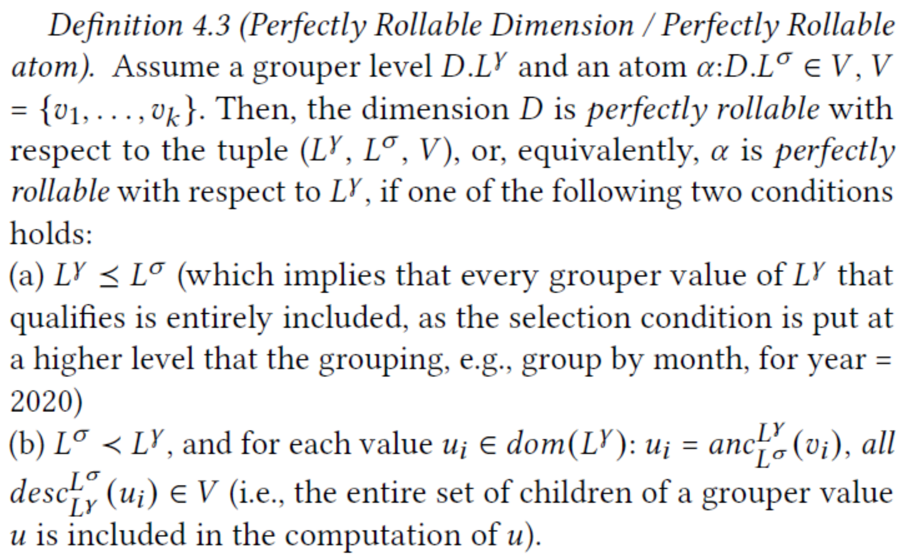
v^{h} L^{σ}: L^{h}

group by month

v^{ℓ}_{1} … v^{ℓ}_{k} L^{γ}: L^{ℓ}

… … … L^{0}

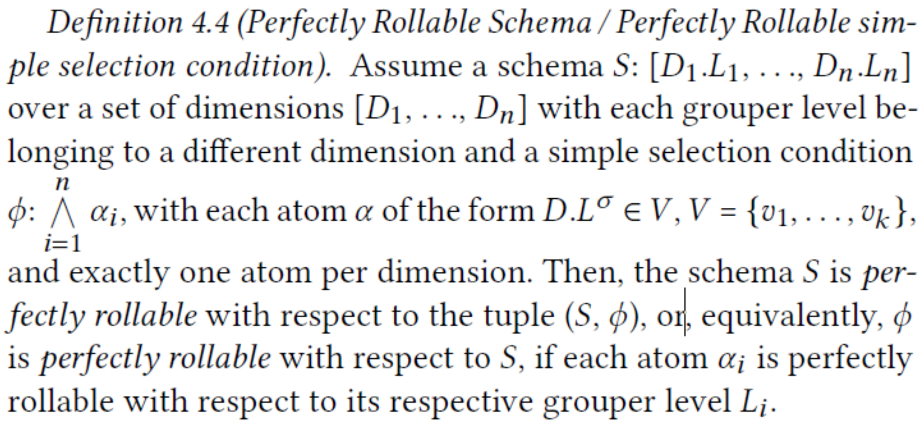
Perfect Rollability



Definition 4.3 (Perfectly Rollable Dimension / Perfectly Rollable atom). Assume a grouper level $D.L^Y$ and an atom α:$D.L^\sigma \in V$, $V = \{v_1, \ldots, v_k\}$. Then, the dimension D is *perfectly rollable* with respect to the tuple (L^Y, L^σ, V), or, equivalently, α is *perfectly rollable* with respect to L^Y, if one of the following two conditions holds:

(a) $L^Y \preceq L^\sigma$ (which implies that every grouper value of L^Y that qualifies is entirely included, as the selection condition is put at a higher level that the grouping, e.g., group by month, for year = 2020)

(b) $L^\sigma \prec L^Y$, and for each value $u_i \in dom(L^Y)$: $u_i = anc_{L^\sigma}^{L^Y}(v_i)$, *all* $desc_{L^Y}^{L^\sigma}(u_i) \in V$ (i.e., the entire set of children of a grouper value u is included in the computation of u).

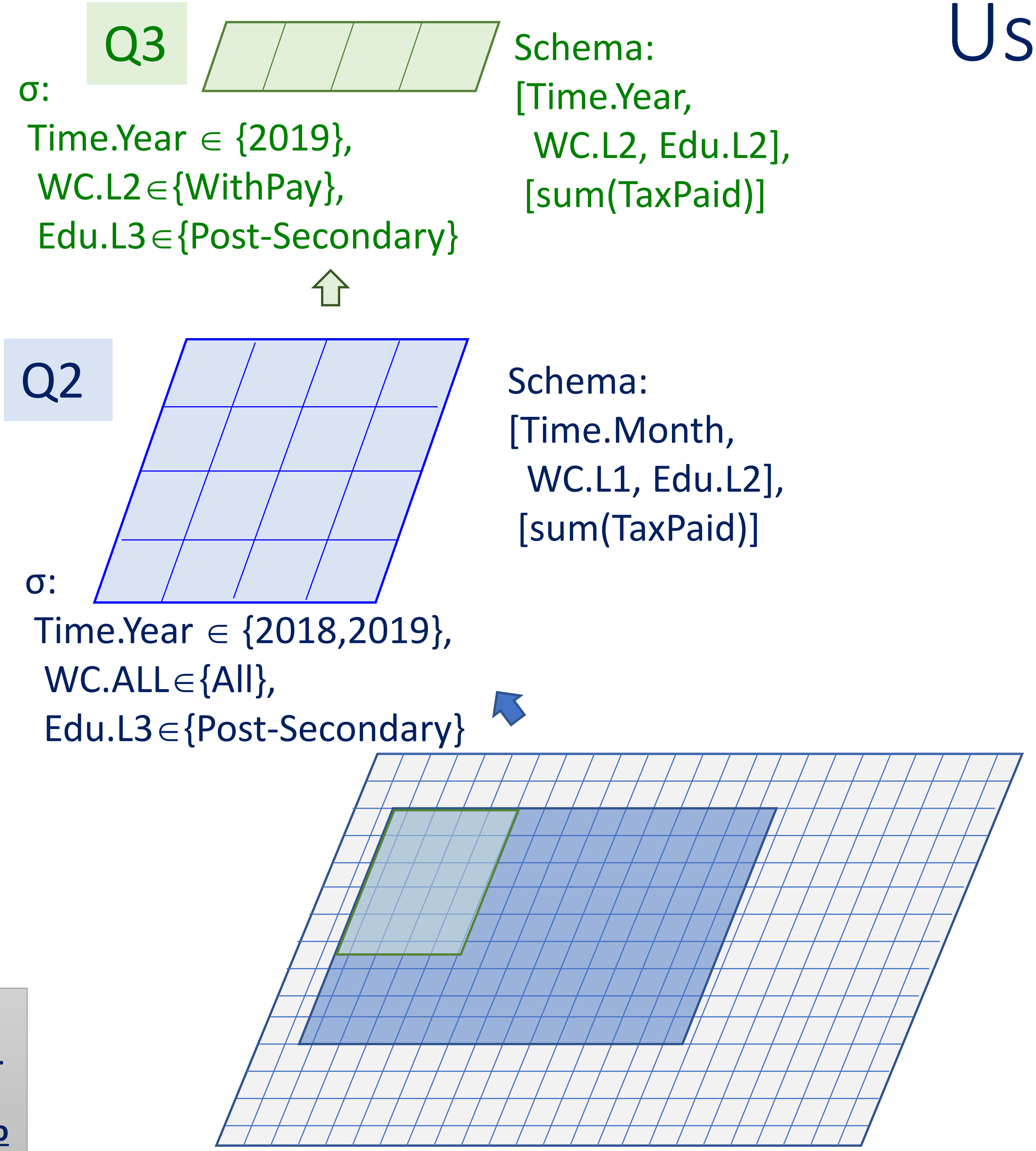


Definition 4.4 (Perfectly Rollable Schema / Perfectly Rollable simple selection condition). Assume a schema S: $[D_1.L_1, \ldots, D_n.L_n]$ over a set of dimensions $[D_1, \ldots, D_n]$ with each grouper level belonging to a different dimension and a simple selection condition ϕ: $\bigwedge_{i=1}^{n} \alpha_i$, with each atom α of the form $D.L^\sigma \in V$, $V = \{v_1, \ldots, v_k\}$, and exactly one atom per dimension. Then, the schema S is *perfectly rollable* with respect to the tuple (S, ϕ), or, equivalently, ϕ is *perfectly rollable* with respect to S, if each atom α_i is perfectly rollable with respect to its respective grouper level L_i.

- The perfectly rollable condition is a "clean" characterization stating that if we group by a level L on any possible data set, then, the resulting grouper values of L will be produced by the entire population of their descendants at lower levels (in fact, as far as the semantics are concerned: the most detailed one)
- Perfect rollability guarantees that, given a simple selection condition on a dimension and a grouper level, there are no grouper cells in the result of a cube that could be computed on the basis of only a subset of their detailed descendants, but rather, the entire range of descendant values are taken into consideration for their computation.

Usability and Rewriting

- Can we compute a query result based on the query result of a previous query?
- If yes, how?



Q3

σ:
Time.Year ∈ {2019},
WC.L2∈{WithPay},
Edu.L3∈{Post-Secondary}

Schema:
[Time.Year,
WC.L2, Edu.L2],
[sum(TaxPaid)]

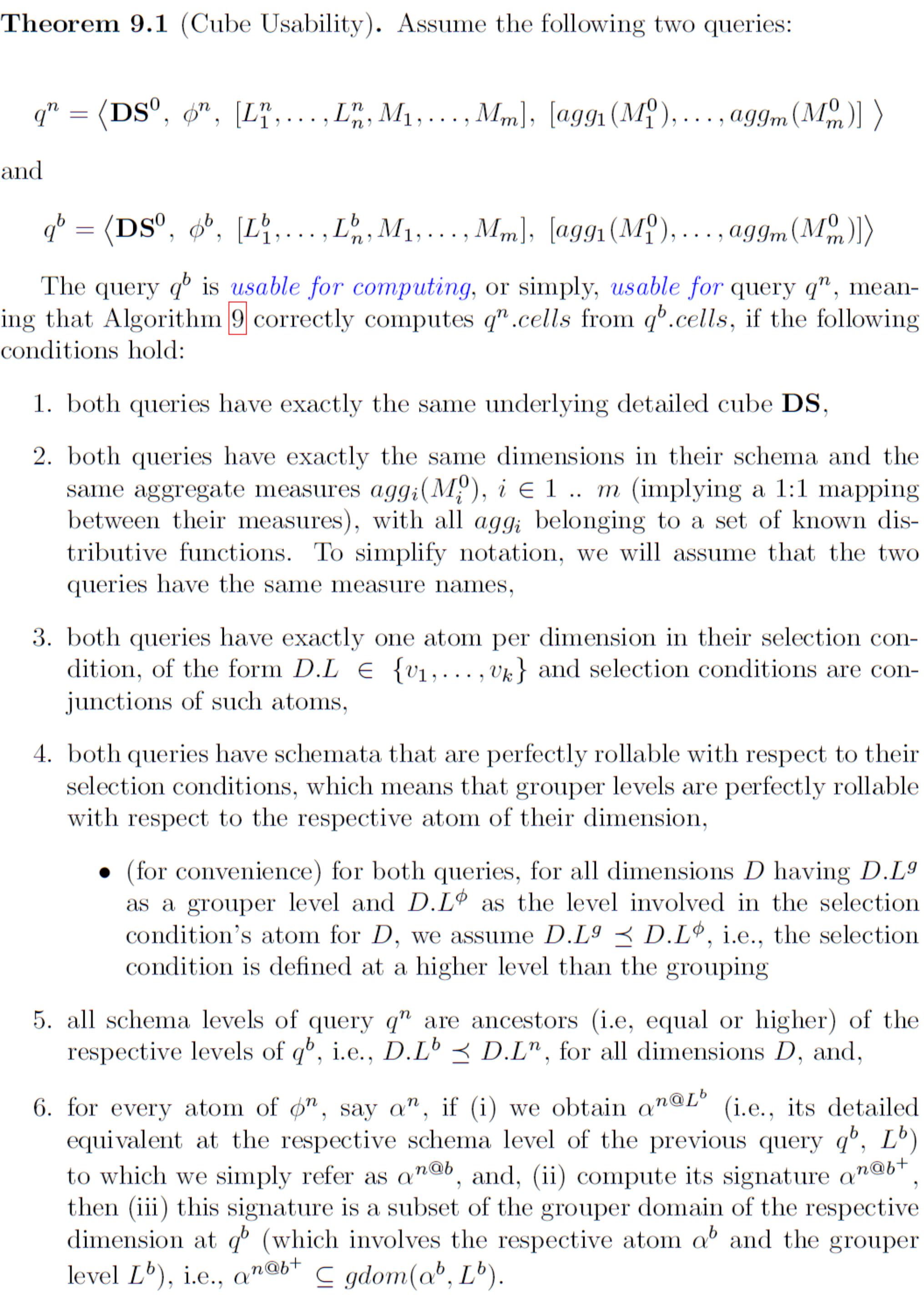
Q2

Schema:
[Time.Month,
WC.L1, Edu.L2],
[sum(TaxPaid)]

σ:
Time.Year ∈ {2018,2019},
WC.ALL∈{All},
Edu.L3∈{Post-Secondary}

Model
Equiv. expr.
Rollability
Test & Algo

Check and algorithm



Theorem 9.1 (Cube Usability). Assume the following two queries:

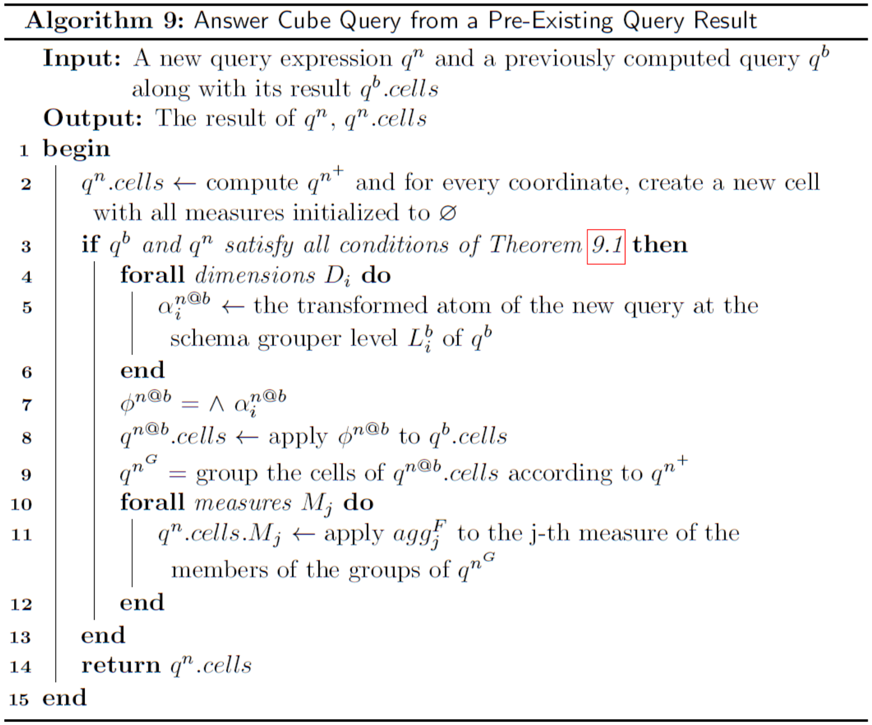
$$q^n = \langle \mathbf{DS}^0,\ \phi^n,\ [L_1^n, \dots, L_n^n, M_1, \dots, M_m],\ [agg_1(M_1^0), \dots, agg_m(M_m^0)]\ \rangle$$

and

$$q^b = \langle \mathbf{DS}^0,\ \phi^b,\ [L_1^b, \dots, L_n^b, M_1, \dots, M_m],\ [agg_1(M_1^0), \dots, agg_m(M_m^0)]\rangle$$

The query q^b is *usable for computing*, or simply, *usable for* query q^n, meaning that Algorithm 9 correctly computes $q^n.cells$ from $q^b.cells$, if the following conditions hold:

1. both queries have exactly the same underlying detailed cube **DS**,
2. both queries have exactly the same dimensions in their schema and the same aggregate measures $agg_i(M_i^0)$, $i \in 1\ ..\ m$ (implying a 1:1 mapping between their measures), with all agg_i belonging to a set of known distributive functions. To simplify notation, we will assume that the two queries have the same measure names,
3. both queries have exactly one atom per dimension in their selection condition, of the form $D.L \in \{v_1, \dots, v_k\}$ and selection conditions are conjunctions of such atoms,
4. both queries have schemata that are perfectly rollable with respect to their selection conditions, which means that grouper levels are perfectly rollable with respect to the respective atom of their dimension,
   - (for convenience) for both queries, for all dimensions D having $D.L^g$ as a grouper level and $D.L^\phi$ as the level involved in the selection condition's atom for D, we assume $D.L^g \preceq D.L^\phi$, i.e., the selection condition is defined at a higher level than the grouping
5. all schema levels of query q^n are ancestors (i.e, equal or higher) of the respective levels of q^b, i.e., $D.L^b \preceq D.L^n$, for all dimensions D, and,
6. for every atom of ϕ^n, say α^n, if (i) we obtain $\alpha^{n@L^b}$ (i.e., its detailed equivalent at the respective schema level of the previous query q^b, L^b) to which we simply refer as $\alpha^{n@b}$, and, (ii) compute its signature $\alpha^{n@b^+}$, then (iii) this signature is a subset of the grouper domain of the respective dimension at q^b (which involves the respective atom α^b and the grouper level L^b), i.e., $\alpha^{n@b^+} \subseteq gdom(\alpha^b, L^b)$.



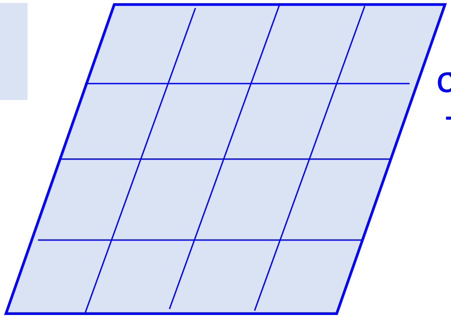
Algorithm 9: Answer Cube Query from a Pre-Existing Query Result

```
Input: A new query expression $q^n$ and a previously computed query $q^b$
       along with its result $q^b.cells$
Output: The result of $q^n$, $q^n.cells$
begin
    $q^n.cells \leftarrow$ compute $q^{n^+}$ and for every coordinate, create a new cell
      with all measures initialized to $\varnothing$
    if $q^b$ and $q^n$ satisfy all conditions of Theorem 9.1 then
        forall dimensions $D_i$ do
            $\alpha_i^{n@b} \leftarrow$ the transformed atom of the new query at the
              schema grouper level $L_i^b$ of $q^b$
        end
        $\phi^{n@b} = \wedge\ \alpha_i^{n@b}$
        $q^{n@b}.cells \leftarrow$ apply $\phi^{n@b}$ to $q^b.cells$
        $q^{n^G}$ = group the cells of $q^{n@b}.cells$ according to $q^{n^+}$
        forall measures $M_j$ do
            $q^n.cells.M_j \leftarrow$ apply $agg_j^F$ to the j-th measure of the
              members of the groups of $q^{n^G}$
        end
    end
    return $q^n.cells$
end
```

Q3

σ:
Time.Year ∈ {2019},
WC.L2∈{WithPay},
Edu.L3∈{Post-Secondary}

Q2



σ:
Time.Year ∈ {2018,2019},
WC.ALL∈{All},
Edu.L3∈{Post-Secondary}

Schema:
[Time.Year, WC.L2, Edu.L2],[sum(TaxPaid)]

Schema:
[Time.Month, WC.L1, Edu.L2], [sum(TaxPaid)]

	Schema L^g			Atoms L^ϕ		
	Time	WC	Edu	Time	WC	Edu
Q3	Year	L2	L2	Year ∈ {2019}	WC.L2 ∈ {WithPay}	Edu.L3 ∈ {Post-Sec}
Q2	Month	L1	L2	Year ∈ {2018, 2019}	WC.ALL∈{All}	Edu.L3 ∈ {Post-Sec}

	Time	WC	Edu
Q^{3+}	{2019}	X {WithPay}	X {SomeColl., Assoc., Univ., PostGrad}
$\phi^{3@2+}$ $Q^{3@2+}$	{2019-01, … 2019-12}	X {Private, SelfEmp, Gov}	X {SomeColl., Assoc., Univ., PostGrad}
ϕ^{2+}	{2018-01, … 2018-12, 2019-01, … 2019-12}	X {Private, SelfEmp, Gov, W/OPay}	X {SomeColl., Assoc., Univ., PostGrad}

Thank you!

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Ταμείο
Περιφερειακής Ανάπτυξης

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΟΙΚΟΝΟΜΙΑΣ & ΑΝΑΠΤΥΞΗΣ
ΕΙΔΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΤΠΑ & ΤΣ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΑνΕΚ

ΕΠΑνΕΚ 2014-2020
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΑΝΤΑΓΩΝΙΣΤΙΚΟΤΗΤΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΤΗΤΑ
ΚΑΙΝΟΤΟΜΙΑ

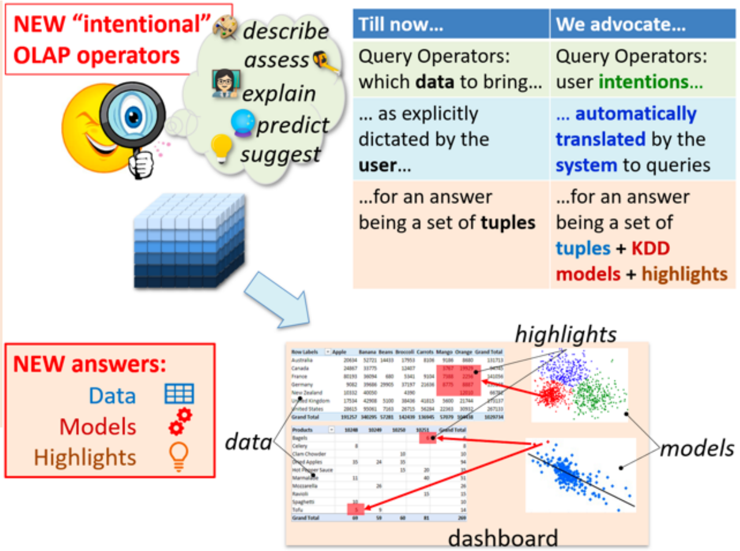
ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

- We have provided a method for computing a new cube from a previous one, defined at a different level of abstraction.
- The basis of the method is perfect rollability, a property characterizing the combination of selection conditions and groupers that guarantees the correct computation of aggregate measures.
- Future work can also target operators comparing cubes for intrinsic properties of their cells (e.g., hidden correlations, predictions, classifications) that have to be decided via the application of knowledge extraction operators to the results, or the detailed areas, of the contrasted cubes.

This work is part of the OLAP III effort:
http://www.cs.uoi.gr/~pvassil/projects/olap_III/

We pursue a revolution to both what a query and what a query answer is and envision OLAP to come with 3 properties: Intentional querying, Intelligent results, Interesting highlights



Till now...	We advocate...
Query Operators: which **data** to bring...	Query Operators: user **intentions**...
... as explicitly dictated by the **user**...	... **automatically translated** by the **system** to queries
...for an answer being a set of **tuples**	...for an answer being a set of **tuples** + **KDD models** + **highlights**