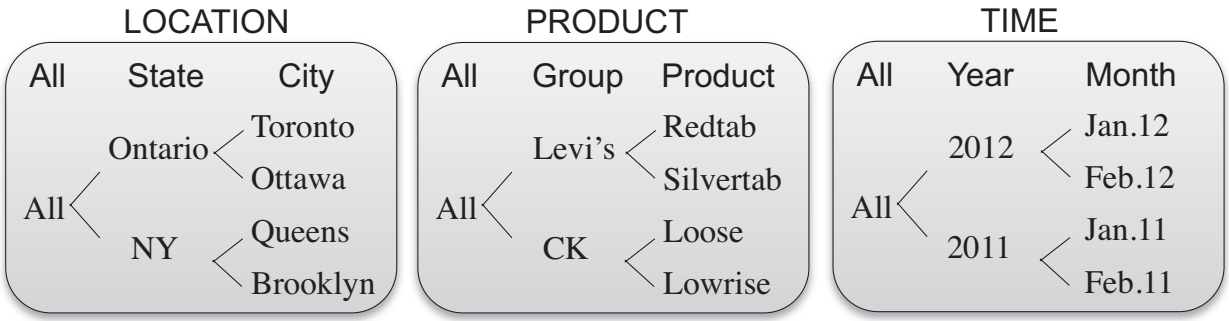
Consider the following scenario…

The cube (portion)



	Redtab		Silvertab	
	Jan.11	Feb.11	Jan.11	Feb.11
Queens	50	40	30	40
Brooklyn	10	20	10	0
Toronto	0	10	0	10
Ottawa	0	10	0	10

The first query:

- Total of sales for all products, all years, all locations?
- System answers: 640

The second query:

- Drill-down and slice: 2011 monthly sales per product per state?
- System answers: (Ontario, Levi's, Feb.11) and (NY, CK, 2011) as expected, but:

		Redtab	Silvertab	Loose	Lowrise
Ontario	Jan.11	0	0	10	10
	Feb.11			10	10
NY	Jan.11	60	40		
	Feb.11	60	40		

The third query:

- Drill down to cities
- System answers: (Ontario, All, 2011) and (NY, CK, 2011) as expected, but:

		Redtab	Silvertab
Jan.11	Queens	50	30
	Brooklyn	10	10
Feb.11	Queens	40	40
	Brooklyn	20	0

UNIVERSITÉ
FRANÇOIS - RABELAIS
TOURS

Towards Intensional Answers to OLAP Queries for Analytical Sessions

Patrick Marcel, Rokia Missaoui, Stefano Rizzi

DOLAP 2012

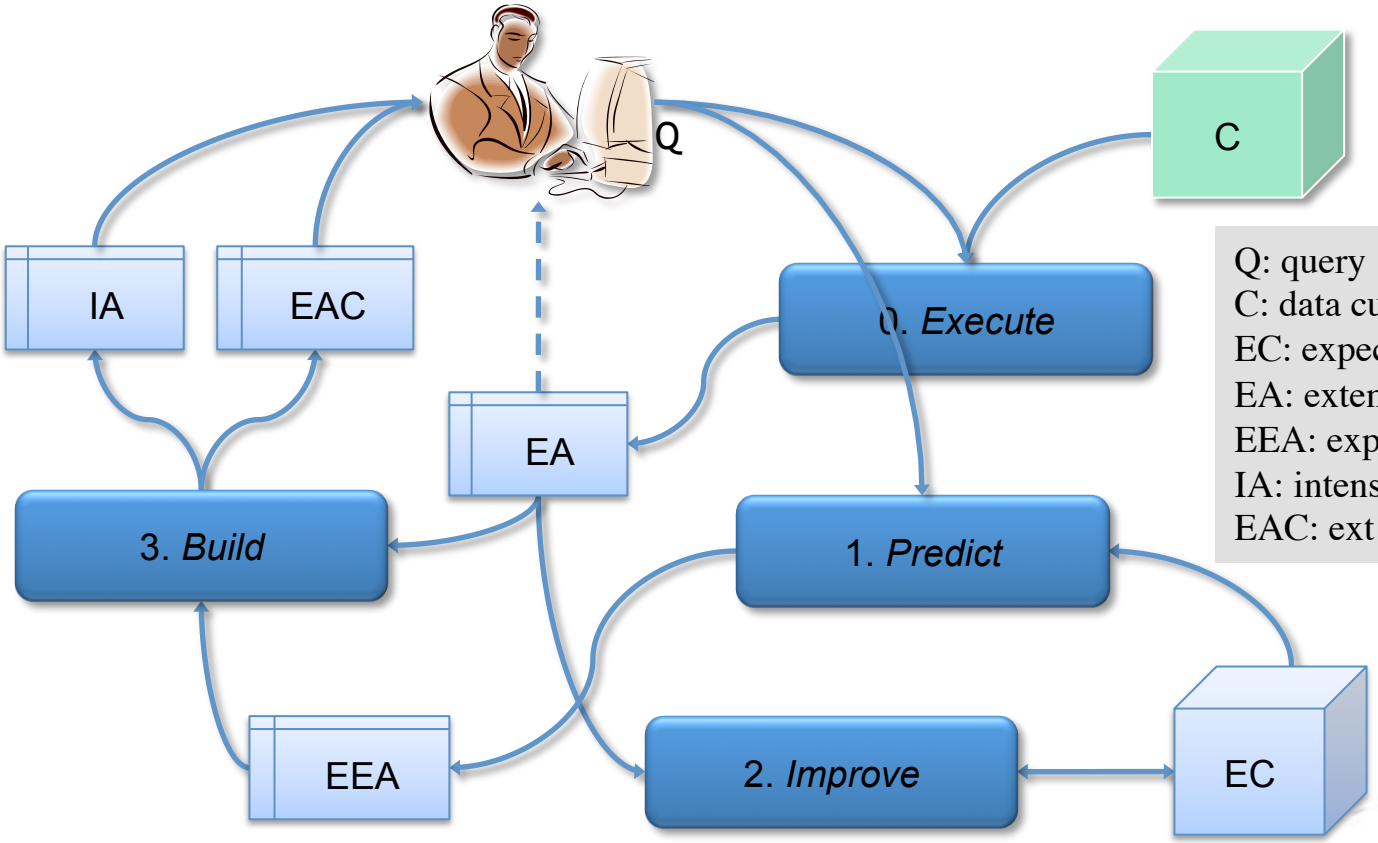
Outline

- Motivation
- The approach
- An instance of the approach
- Future directions

Motivation

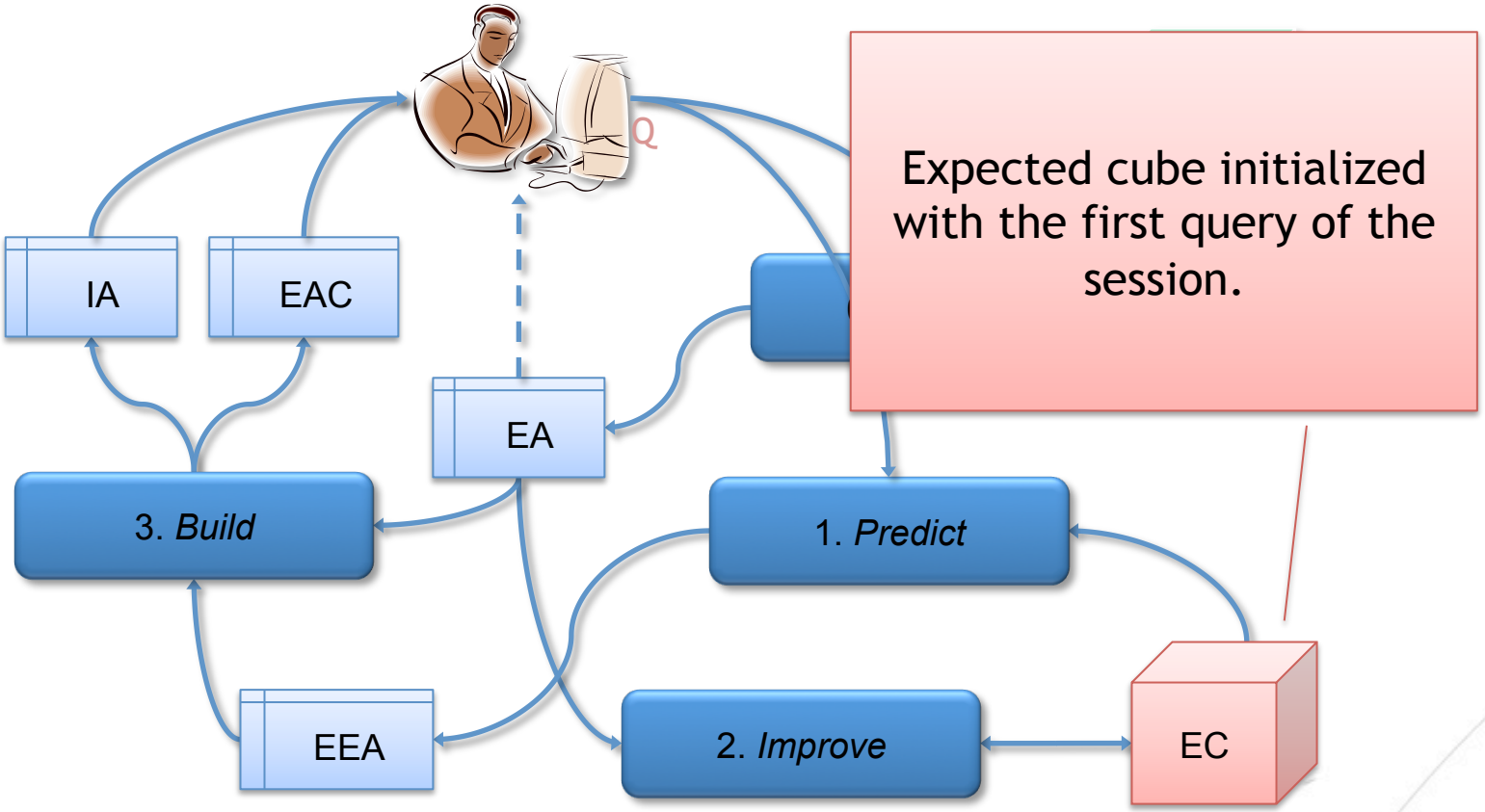
- Intensional Answers (IA)?
  - Concise description of the answer
- OLAP Queries (OQ)?
  - Known ☺
- Analytical Sessions (AS)?
  - Sequence of OLAP queries
- IA2OQ4AS
  - Leveraging past queries to reduce the size of the answer

Approach overview

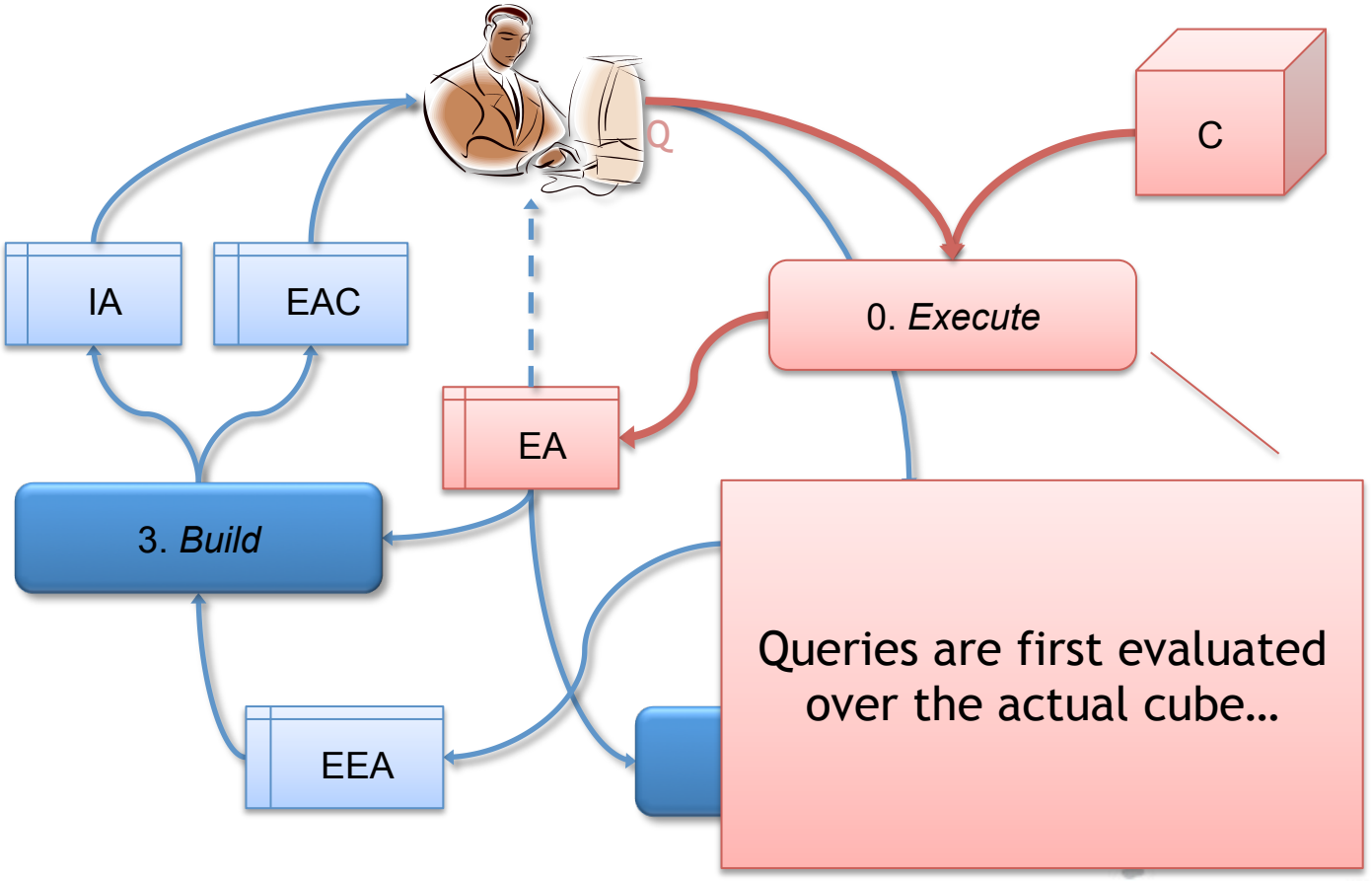


Q: query
C: data cube
EC: expected cube
EA: extensional answer
EEA: expected ext. answer
IA: intensional answer
EAC: ext. answer complement

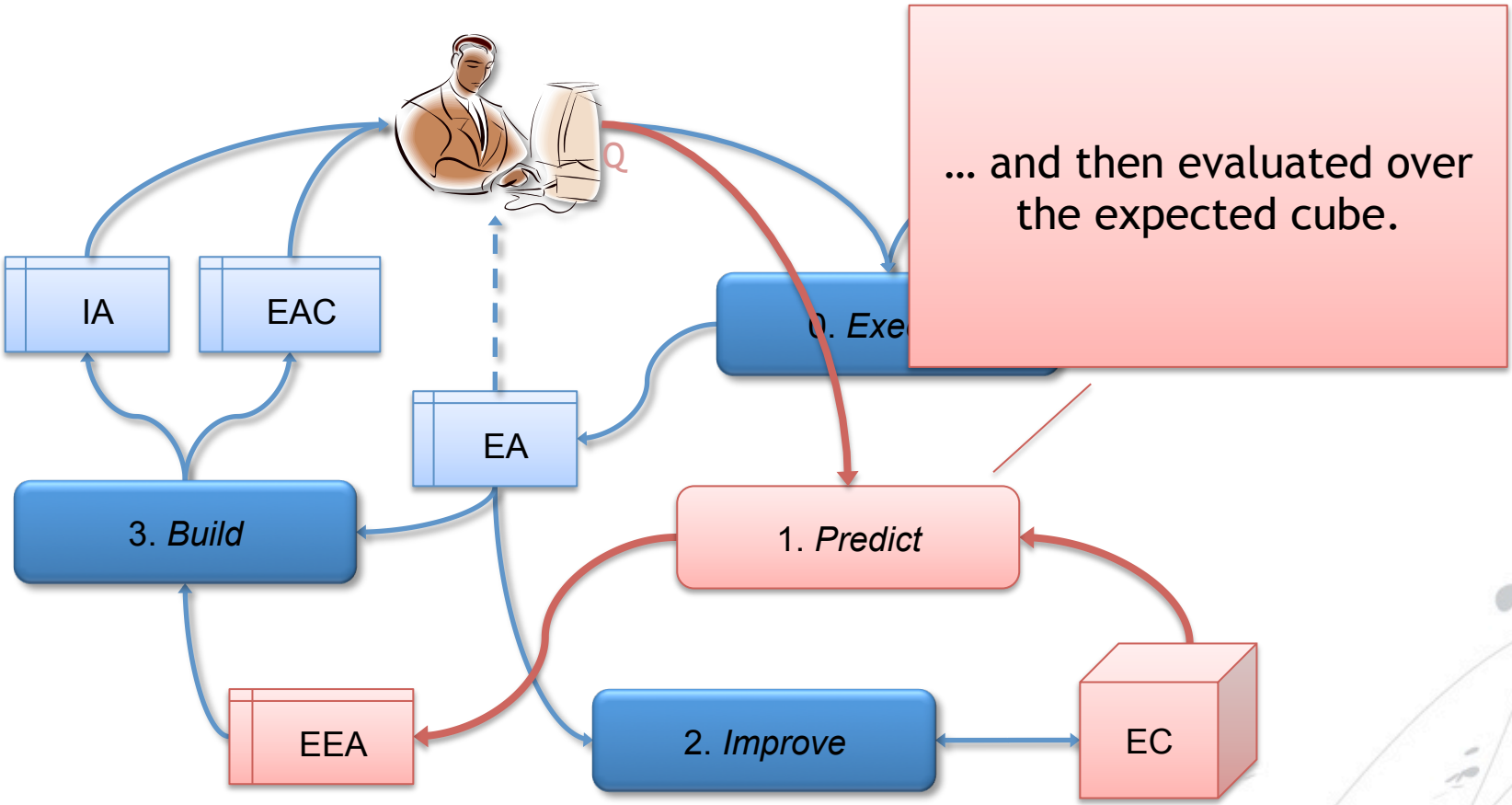
Startup



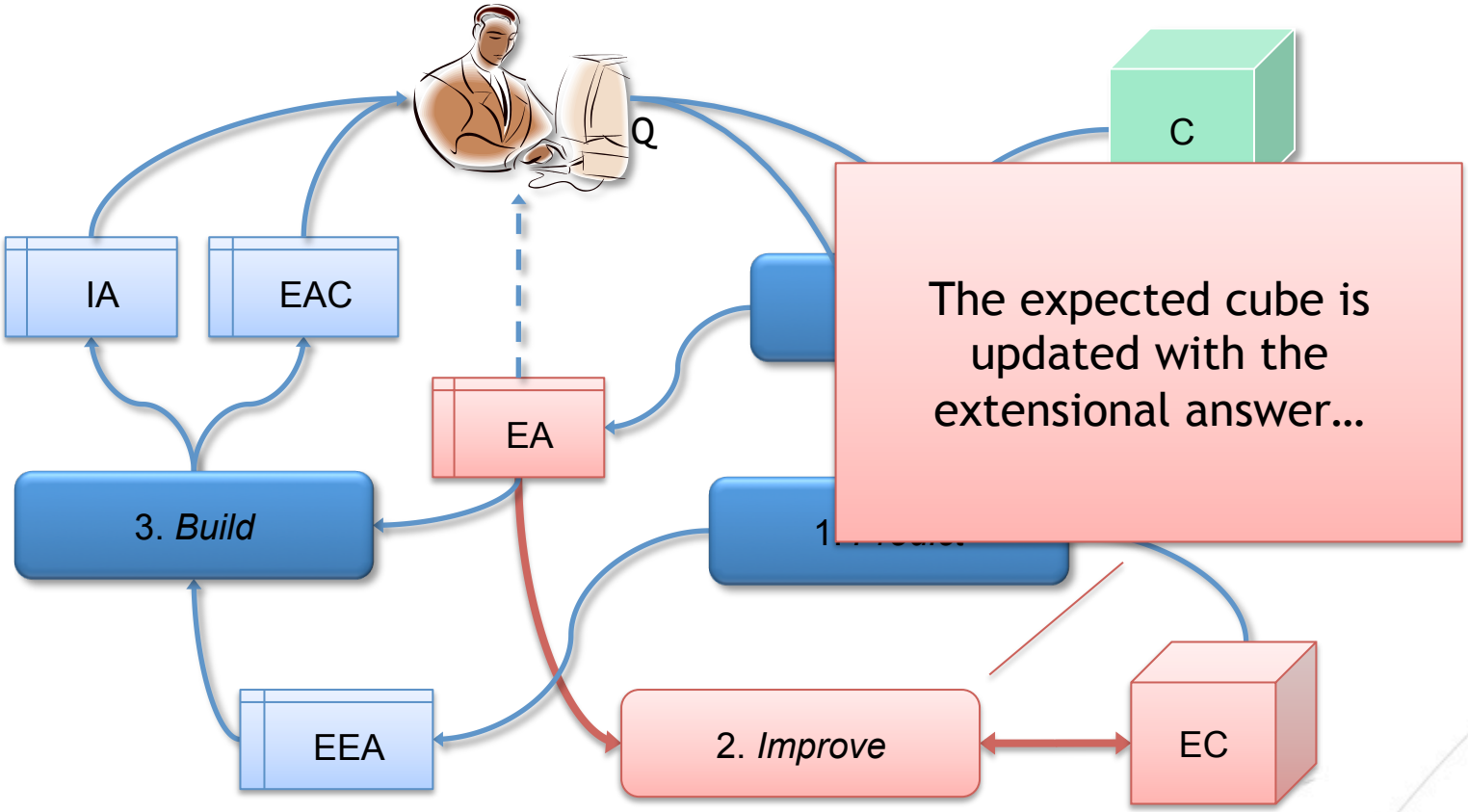
Execute



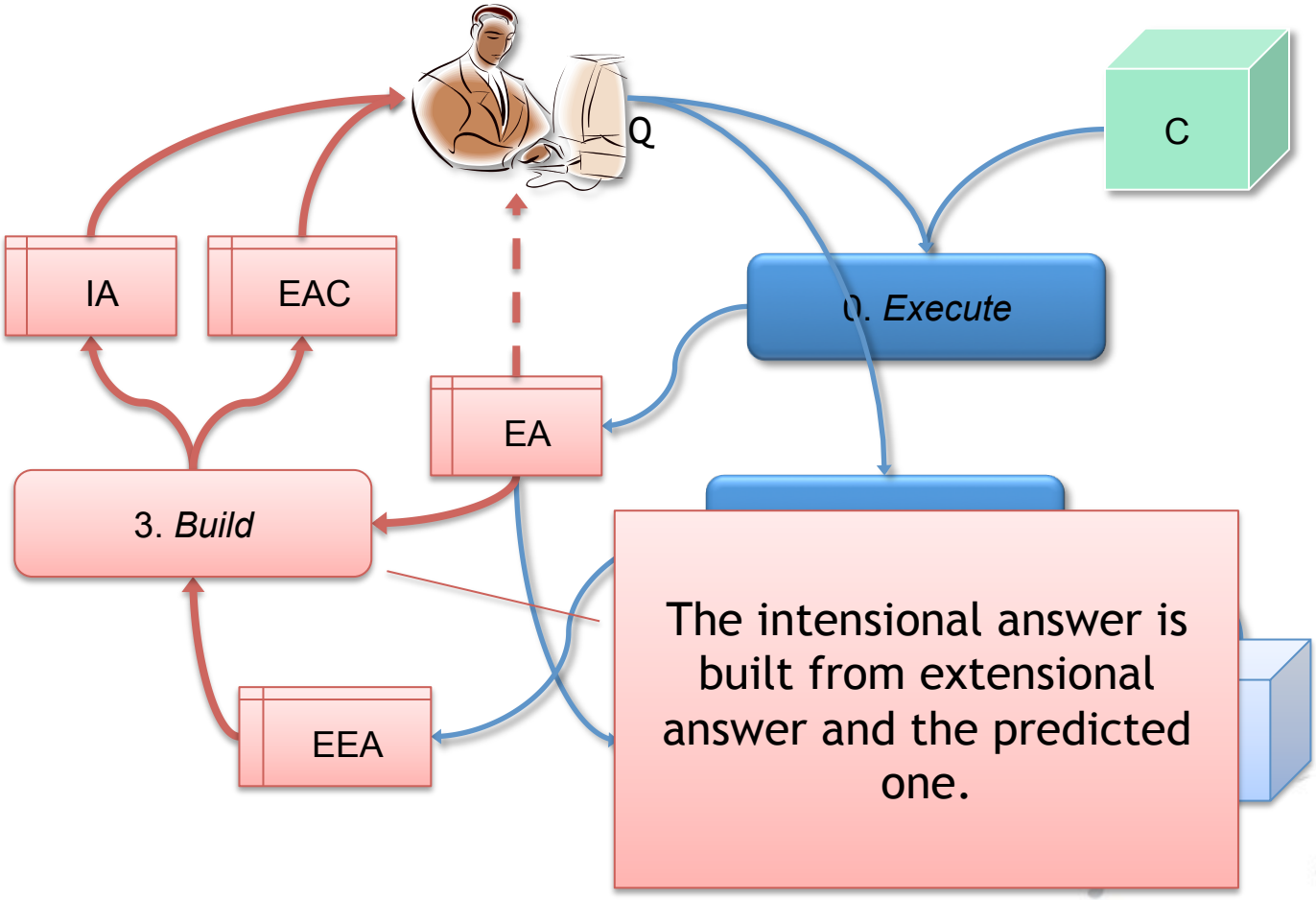
Predict



Improve



Build



An instance of the framework

- Relying on past contributions
  - Cube modeling
    - Using maximum entropy principle
      - like in [Sarawagi, VLDB'00], [Palpanas & al., TKDE05]
  - Intensional answers:
    - Information theoretic characterization
      - like in [Chum & Muntz, VLDB'88]
    - Using hierarchies to build the IA
      - like in [Park & Yoon, HICSS'99]

Improve the expected cube

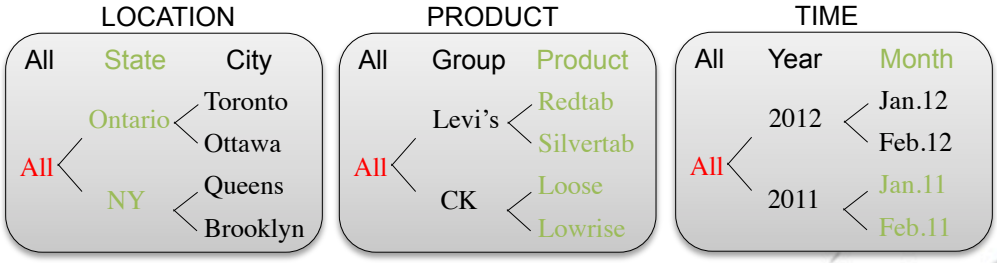
- The estimated values are those that:
  - maximize the uniformity of data values
  - maintain the value of known aggregates
- Estimates are scored:
  - A confidence score is computed from a distance in the group by lattice
  - the more precise the aggregate used for estimation, the more accurate the estimate, and the higher the score

Example

- System answers: grand total is 640
- Expected 2011 monthly sales per product per state is:

		Redtab	Silvertab	Loose	Lowrise
Ontario	Jan.11	20	20	20	20
	Feb.11	20	20	20	20
NY	Jan.11	20	20	20	20
	Feb.11	20	20	20	20

- Confidence of estimates is: avg(0.5, 0, 0)=0.17



Predict the extensional answer

- Run the query over the expected cube:
  - Not only the slice requested by the query
  - But also slices with the same schema that have the best confidence score

Example

- If the facts and their scores are:

f	$Agg(f)$	$conf(f)$
$\langle\langle$Ontario,CK,2012$\rangle, 80\rangle$	$\langle\langle$All,All,2012$\rangle, 320\rangle$	*0.4*
$\langle\langle$NY,CK,2012$\rangle, 80\rangle$	$\langle\langle$All,All,2012$\rangle, 320\rangle$	*0.4*
$\langle\langle$Ontario,CK,2011$\rangle, 80\rangle$	$\langle\langle$All,CK,2011$\rangle, 120\rangle$	*0.9*
$\langle\langle$NY,CK,2011$\rangle, 80\rangle$	$\langle\langle$NY,All,2011$\rangle, 280\rangle$	*0.9*

- If the query asks for the 2012 sales of CK by state, then slice (All,CK,2011) will be used to adjust the prediction

Example

- If the facts and their scores are:

f	$Agg(f)$	$conf(f)$
$\langle\langle$Ontario,CK,2012$\rangle$, 80$\rangle$	$\langle\langle$All,All,2012$\rangle$, 320$\rangle$	*0.4*
$\langle\langle$NY,CK,2012$\rangle$, 80$\rangle$	$\langle\langle$All,All,2012$\rangle$, 320$\rangle$	*0.4*
$\langle\langle$Ontario,CK,2011$\rangle$, 80$\rangle$	$\langle\langle$All,CK,2011$\rangle$, 120$\rangle$	*0.9*
$\langle\langle$NY,CK,2011$\rangle$, 80$\rangle$	$\langle\langle$NY,All,2011$\rangle$, 280$\rangle$	*0.9*

- If the query asks for the 2012 sales of CK by state, then slice (All,CK,2011) will be used to adjust the prediction
- Expected value for (Ontario,CK,2012) is:

$$\frac{320 \times \frac{80}{120} \times 0.9 + 320 \times \frac{80}{320} \times 0.4}{0.9 + 0.4} = 172.3$$

Build the intensional answer

- EA and EEA are compared
- Regions are labeled *"as expected"* if
  - Deviations in the region are low
  - Déviations in the region are homogeneous
- Regions are delimited using granularity levels coarser than the ones of the EA

Example

- Expected answer:

		Redtab	Silvertab	Loose	Lowrise
Ontario	Jan.11	20	20	20	20
	Feb.11	20	20	20	20
NY	Jan.11	20	20	20	20
	Feb.11	20	20	20	20

- Extensional answer:

		Redtab	Silvertab	Loose	Lowrise
Ontario	Jan.11	0	0	10	10
	Feb.11	20	20	10	10
NY	Jan.11	60	40	20	20
	Feb.11	60	40	20	20

- (Ontario, Levi's, Feb.11) and (NY, CK, 2011) as expected

Conclusion

- A generic and flexible framework for computing intensional answers to OLAP queries
  - Leverages the previous queries of the session
  - Helps reduce the answer's size
- An instance of the framework
  - Assumes uniformly distributed values
  - Specific details and procedures

Future directions

- Different instantiations of the 3 steps
  - Alternatives to uniform assumption
  - A different kind of intensional answer
  - User's profile to generate the IA
- Coupling with a recommendation engine
  - Among various possible queries, recommend the one whose answer diverges the most from the user's expectation

Thanks for your attention