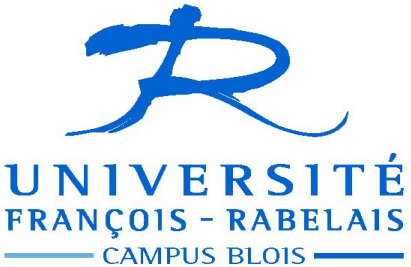


Query Recommendations for OLAP Discovery Driven Analysis

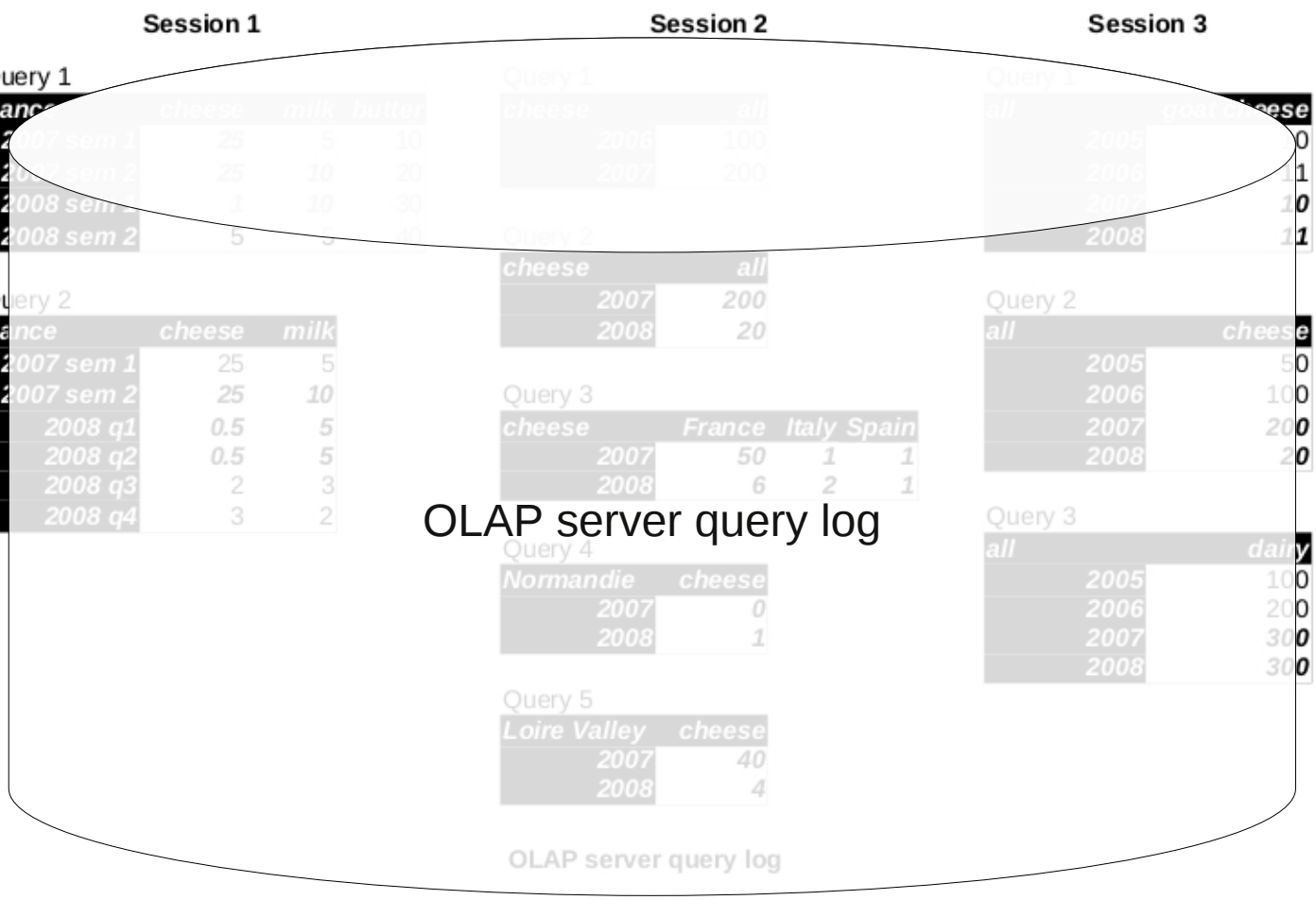
A. Giacometti, *P. Marcel*, E. Negre and A. Soulet

Université François Rabelais Tours, France
Laboratoire d'Informatique

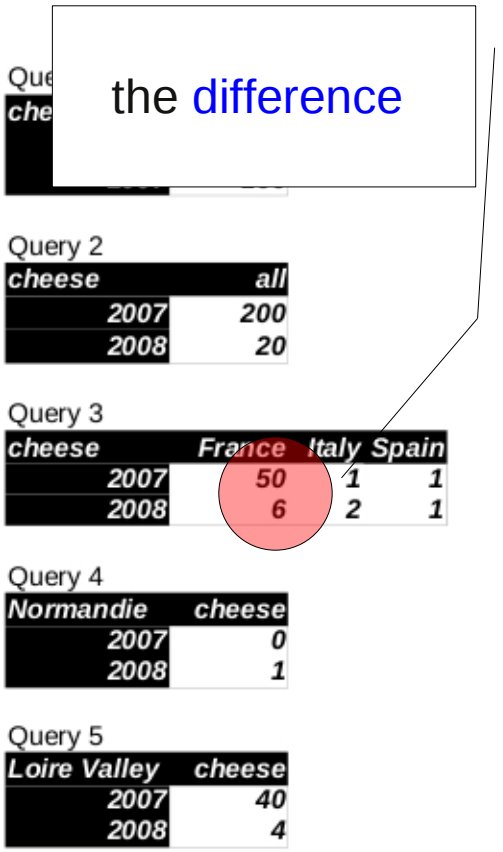
Outline

- Motivation
  - The big picture
  - Discovery driven analysis? Query recommendation?
- Query recommendation for discovery driven analysis?
  - difference pairs & difference queries
  - investigations
- Feasability?
- Conclusion

Motivation: the big picture



Discovery driven analysis?



Discovery driven analysis?

the difference

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

Loire Valley	cheese
2007	40
2008	4

Diff: drilldown to a pair that contributes a lot to the difference

Discovery driven analysis?

Relax: rollup to see if the difference is confirmed

the difference

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

Loire Valley	cheese
2007	40
2008	4

Diff: drilldown to a pair that contributes a lot to the difference

Discovery driven analysis?

Relax: rollup to see if the difference is confirmed

Except: drilldown to exceptions to the difference

Diff: drilldown to a pair that contributes a lot to the difference

the difference

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

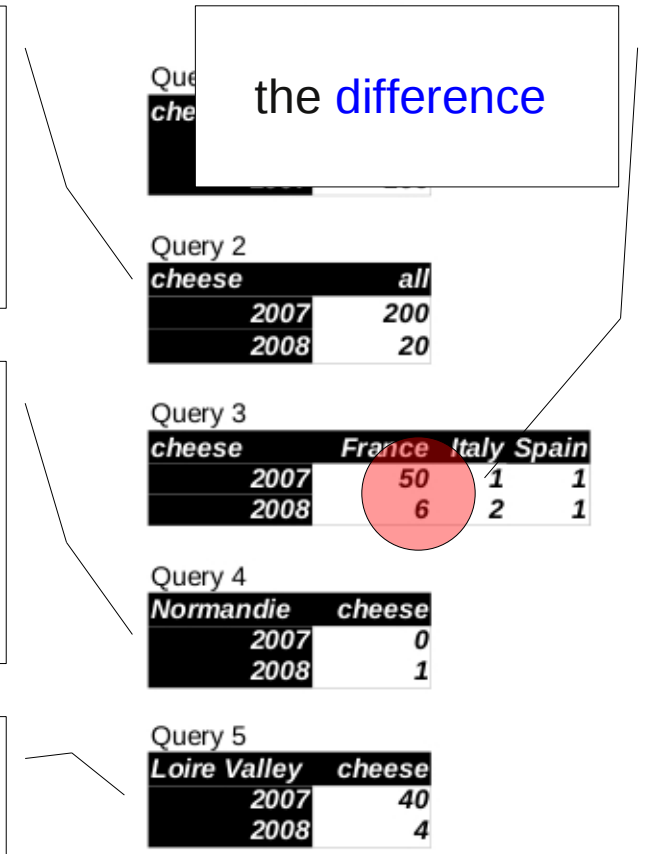
Loire Valley	cheese
2007	40
2008	4

Discovery driven analysis?

Relax: rollup to see if the difference is confirmed

Except: drilldown to exceptions to the difference

Diff: drilldown to a pair that contributes a lot to the difference



Motivation: to automate parts of the often tedious analysis

See Sarawagi's papers e.g., VLDB 1999 and 2001 but also Cariou & al. Dawak 2008

Query recommendation?

Session 1

Query 1

france	cheese	milk	butter
2007 sem 1	25	5	10
2007 sem 2	25	10	20
2008 sem 1	1	10	30
2008 sem 2	5	5	

Query 2

france	cheese
2007 sem 1	25
2007 sem 2	25
2008 q1	0.5
2008 q2	0.5
2008 q3	2
2008 q4	3

Session 2

Query 1

cheese	all
2006	100
2007	200

2008	6	2	1

Session 3

Query 1

all	goat chee
2005	10
2006	11
2007	10
2008	11

	cheese
2005	50
2006	100
2007	200
2008	20

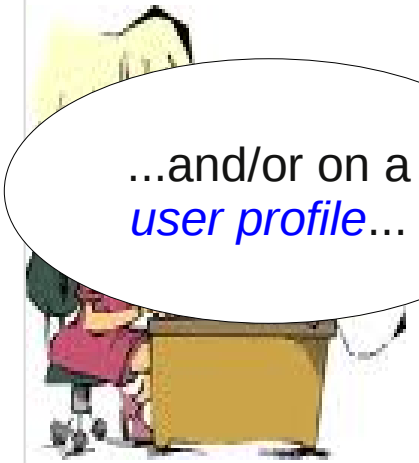
Query 3

	2008
USA	50

...and/or on the *current session*...

Based on the *query log*...

...and/or on a *user profile*...



Current session

...the *system* suggests to the *current user* this or that *particular query* for pursuing her *analysis*

See e.g., Jerbi & al. or Giacometti& al. Dawak 2009
But also Chatzopoulou & al. SSDBM 2009

Query recommandation for discovery driven analysis?

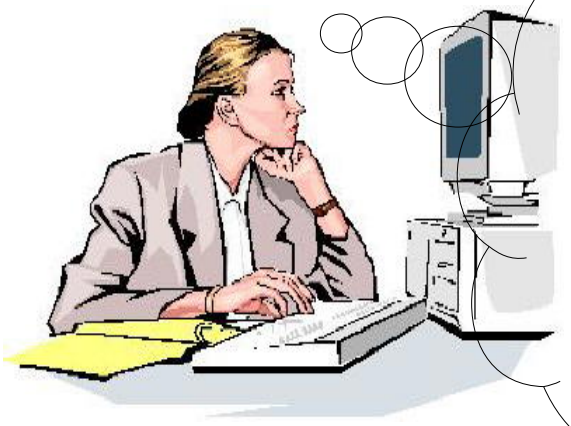
Session 1

Query 1

france	cheese	milk	butter
2007 sem 1	25	5	10
2007 sem 2	25	10	20
2008 sem 1	1	10	30
2008 sem 2	5	5	40

Query 2

france	cheese	milk
2007 sem 1	25	5
2007 sem 2	25	10
2008 q1	0.5	5
2008 q2	0.5	5
2008 q3	2	3
2008 q4	3	2



Session 2

Query 1

cheese	all
2006	100
2007	200

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

Loire Valley	cheese
2007	40
2008	4

interesting...

Session 3

Query 1

all	goat cheese
2005	10
2006	11
2007	10
2008	11

Query 2

all	cheese
2005	50
2006	100
2007	200
2008	20

Query 3

all	dairy
2005	100
2006	200
2007	300
2008	300

Hm this looks strange to me...

Current query

cheese	2007	2008
Europe	100	10
USA	50	5

Current session

Difference pairs and queries

difference pairs:
e.g., v/v' > ~10
for the same slice

Query 1

cheese	all
2006	100
2007	200

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

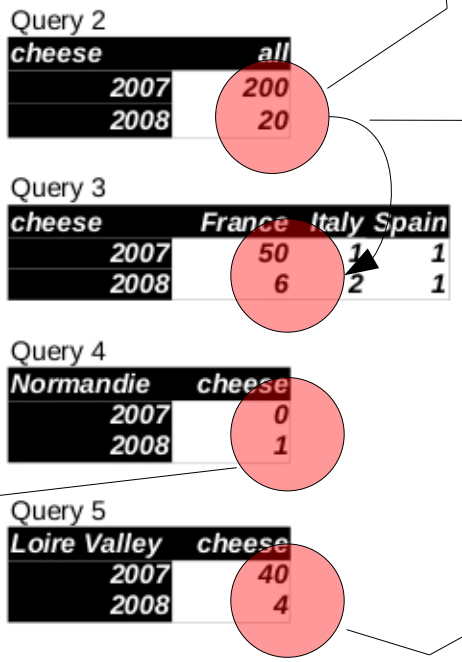
Loire Valley	cheese
2007	40
2008	4

Contains difference pairs so is a difference query

Difference pairs and queries

Most general difference pair of this log

This pair of Query 2 Generalizes that (i.e., is a rollup pair) of Query 3



Exception pair of the most general pair

Drilldown pair of the most general pair

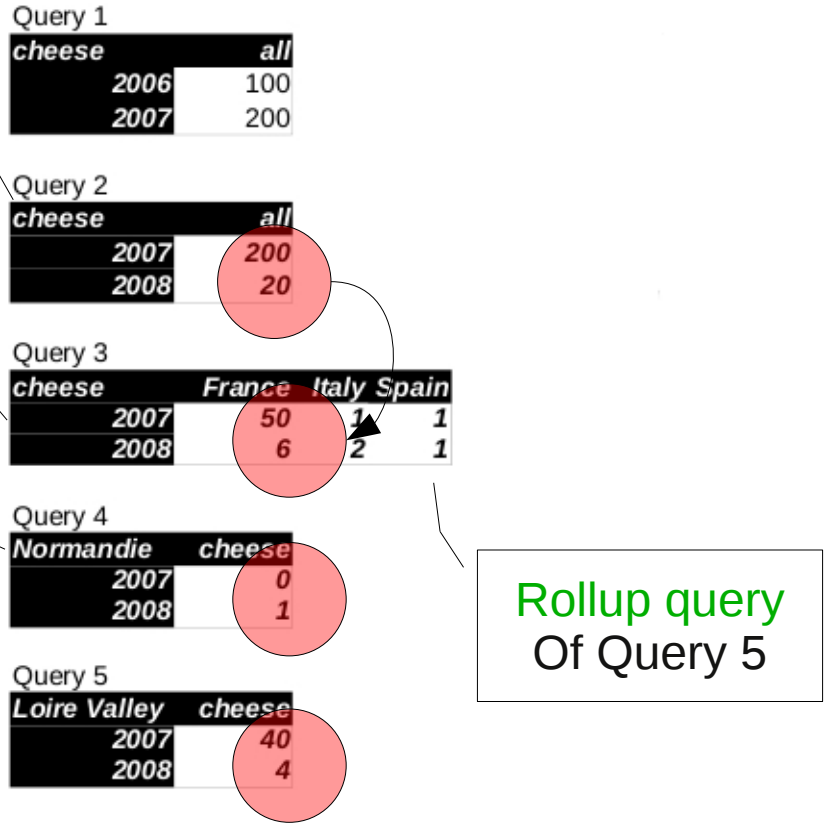
Difference pairs and queries

Most general query
of this log

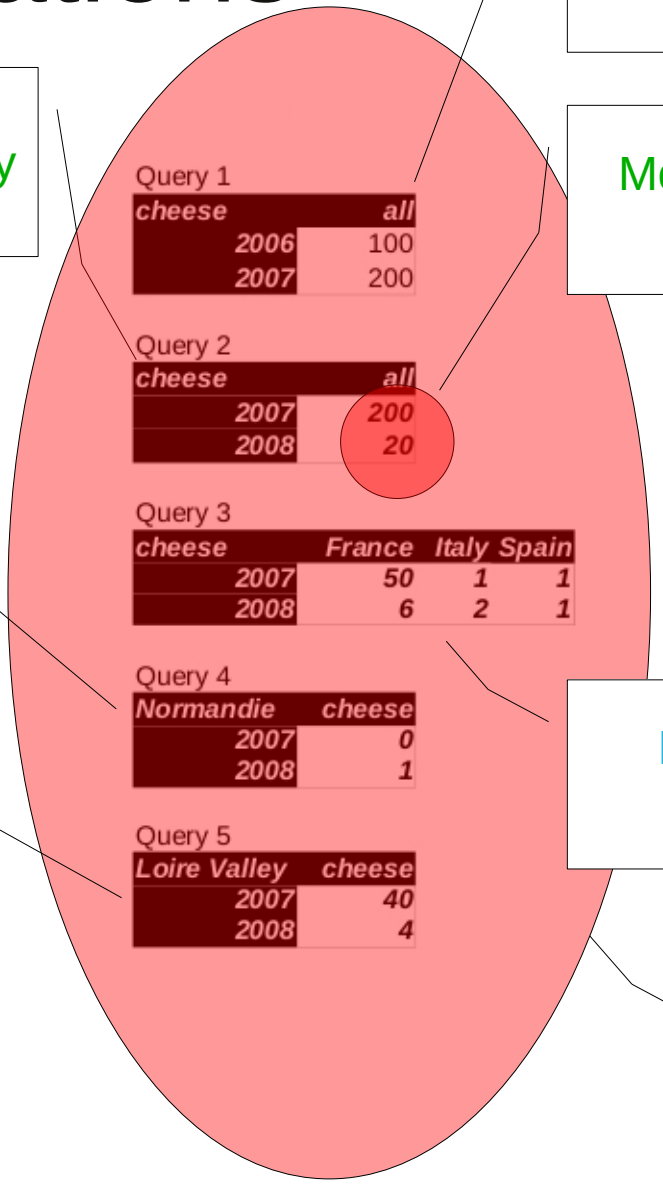
Drilldown query
of Query 2

Exception query
of Query 2

Rollup query
Of Query 5



Investigations



Most general query

Some other query

Most general pair

Exception query

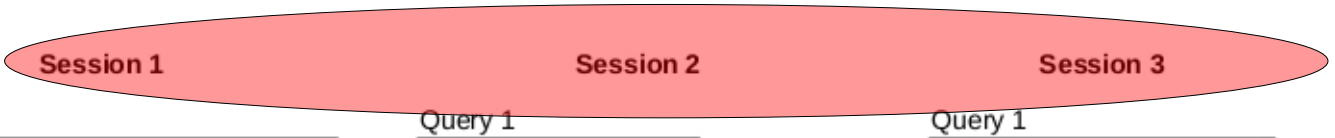
Drilldown query

Drilldown query

An investigation for session 2

Processing the log

1: Consider all sessions



Query 1

france	cheese	milk	butter
2007 sem 1	25	5	10
2007 sem 2	25	10	20
2008 sem 1	1	10	30
2008 sem 2	5	5	40

Query 2

france	cheese	milk
2007 sem 1	25	5
2007 sem 2	25	10
2008 q1	0.5	5
2008 q2	0.5	5
2008 q3	2	3
2008 q4	3	2

Query 1

cheese	all
2006	100
2007	200

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

Loire Valley	cheese
2007	40
2008	4

Query 1

all	goat cheese
2005	10
2006	11
2007	10
2008	11

Query 2

all	cheese
2005	50
2006	100
2007	200
2008	20

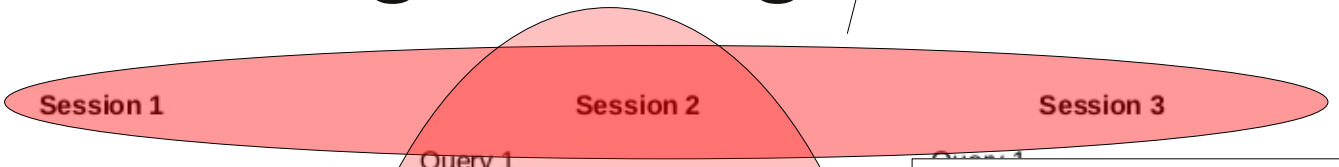
Query 3

all	dairy
2005	100
2006	200
2007	300
2008	300

OLAP server query log

Processing the log

1: Consider all sessions



2: consider all queries

Query 1

france	cheese	milk	butter
2007 sem 1	25	5	10
2007 sem 2	25	10	20
2008 sem 1	1	10	30
2008 sem 2	5	5	40

Query 2

france	cheese	milk
2007 sem 1	25	5
2007 sem 2	25	10
2008 q1	0.5	5
2008 q2	0.5	5
2008 q3	2	3
2008 q4	3	2

Query 1

cheese	all
2006	100
2007	200

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

Loire Valley	cheese
2007	40
2008	4

OLAP server query log

2008	11

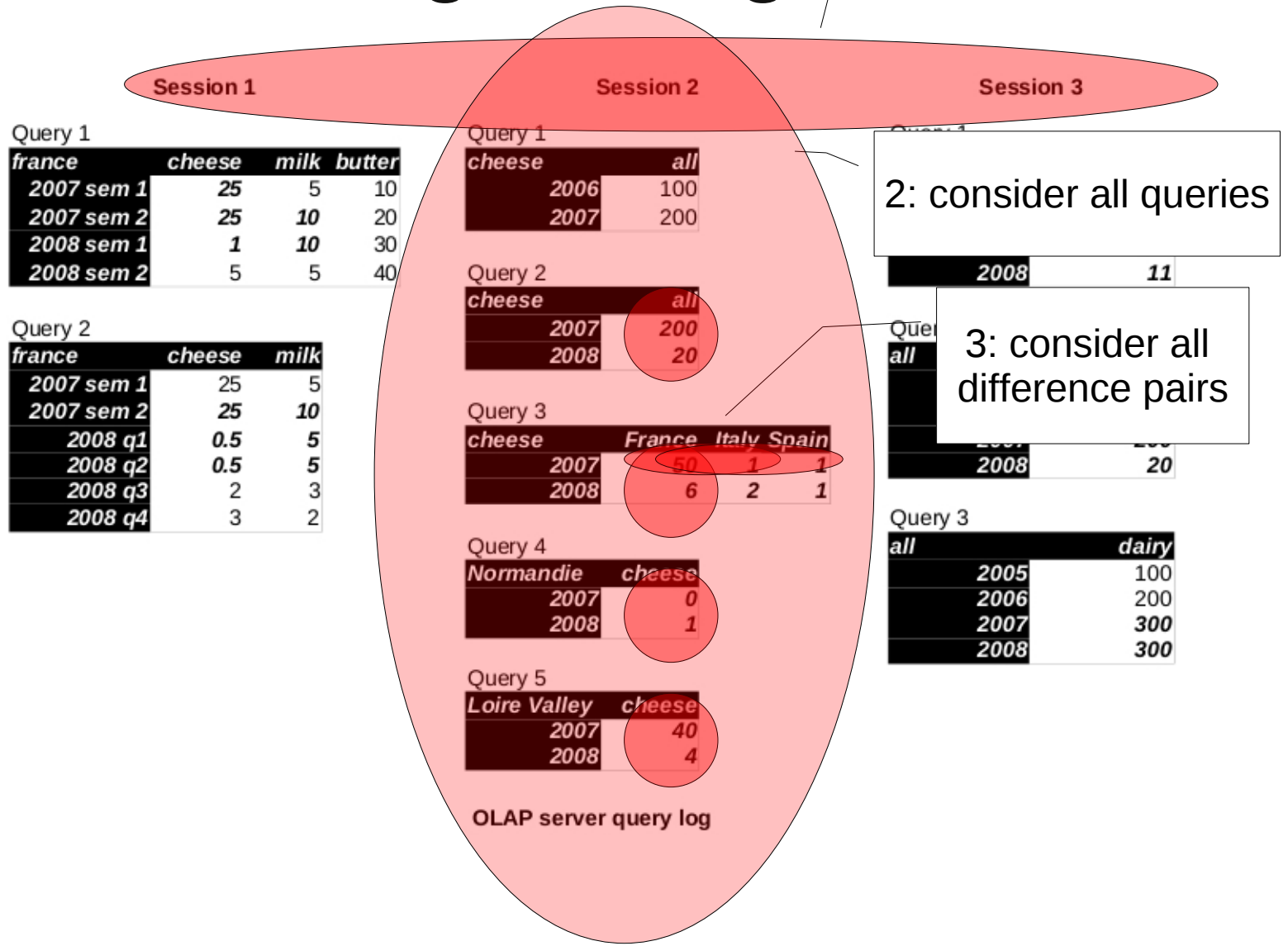
Query 2

all	cheese
2005	50
2006	100
2007	200
2008	20

Query 3

all	dairy
2005	100
2006	200
2007	300
2008	300

Processing the log



Query 1

france	cheese	milk	butter
2007 sem 1	25	5	10
2007 sem 2	25	10	20
2008 sem 1	1	10	30
2008 sem 2	5	5	40

Query 2

france	cheese	milk
2007 sem 1	25	5
2007 sem 2	25	10
2008 q1	0.5	5
2008 q2	0.5	5
2008 q3	2	3
2008 q4	3	2

Query 1

cheese	all
2006	100
2007	200

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

Loire Valley	cheese
2007	40
2008	4

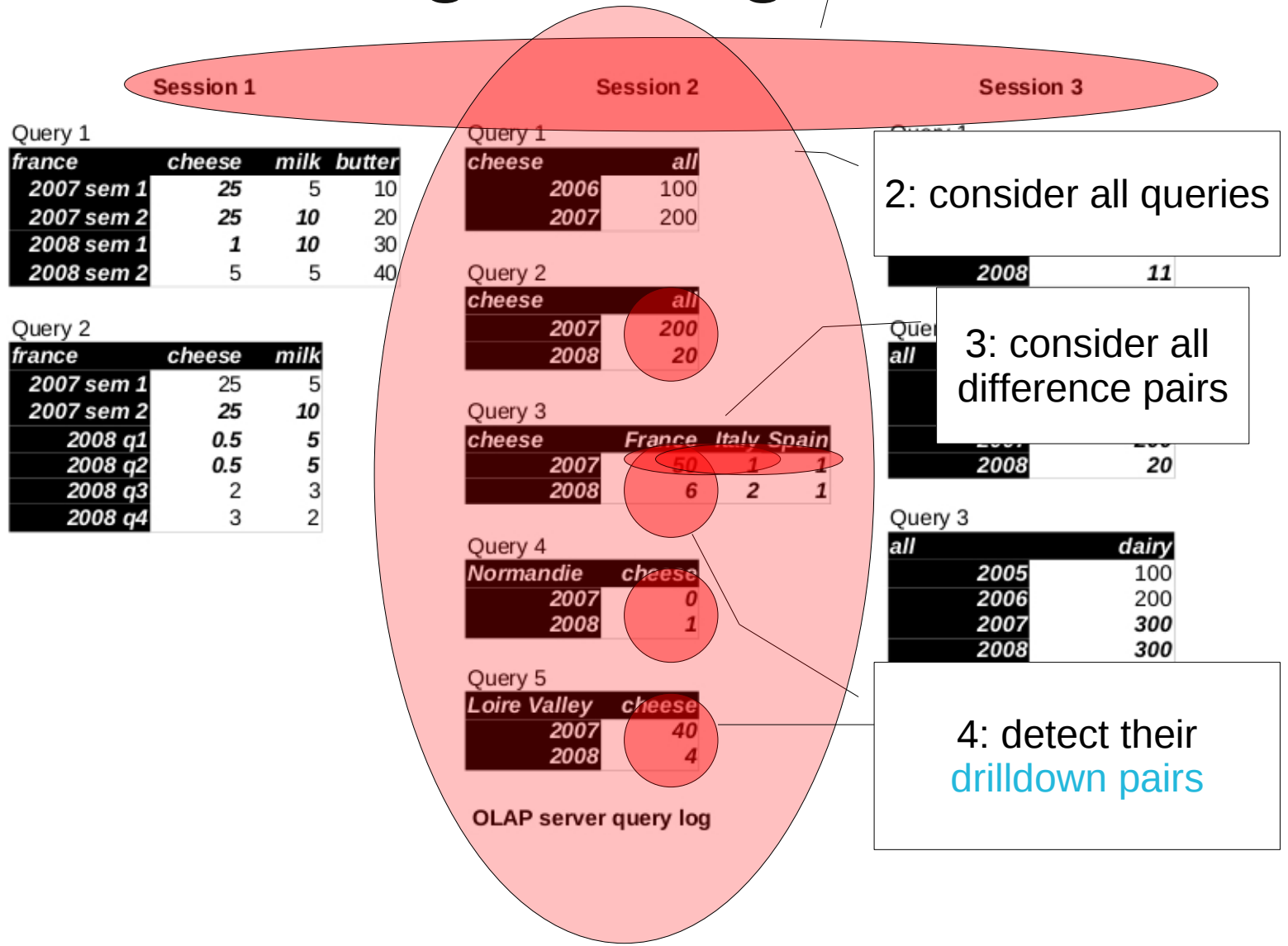
2008	11

all	
2008	20

Query 3

all	dairy
2005	100
2006	200
2007	300
2008	300

Processing the log



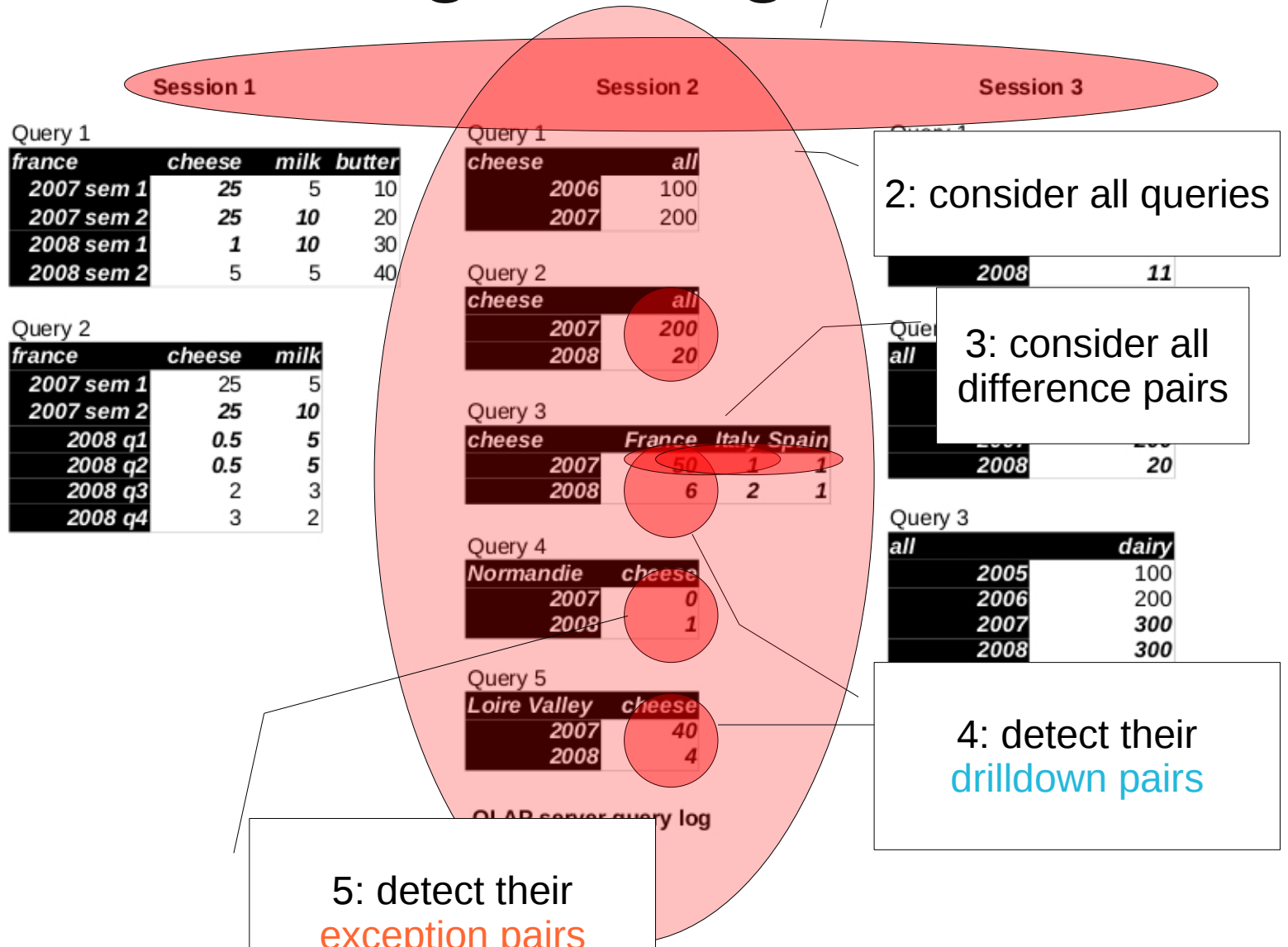
1: Consider all sessions

2: consider all queries

3: consider all difference pairs

4: detect their drilldown pairs

Processing the log



1: Consider all sessions

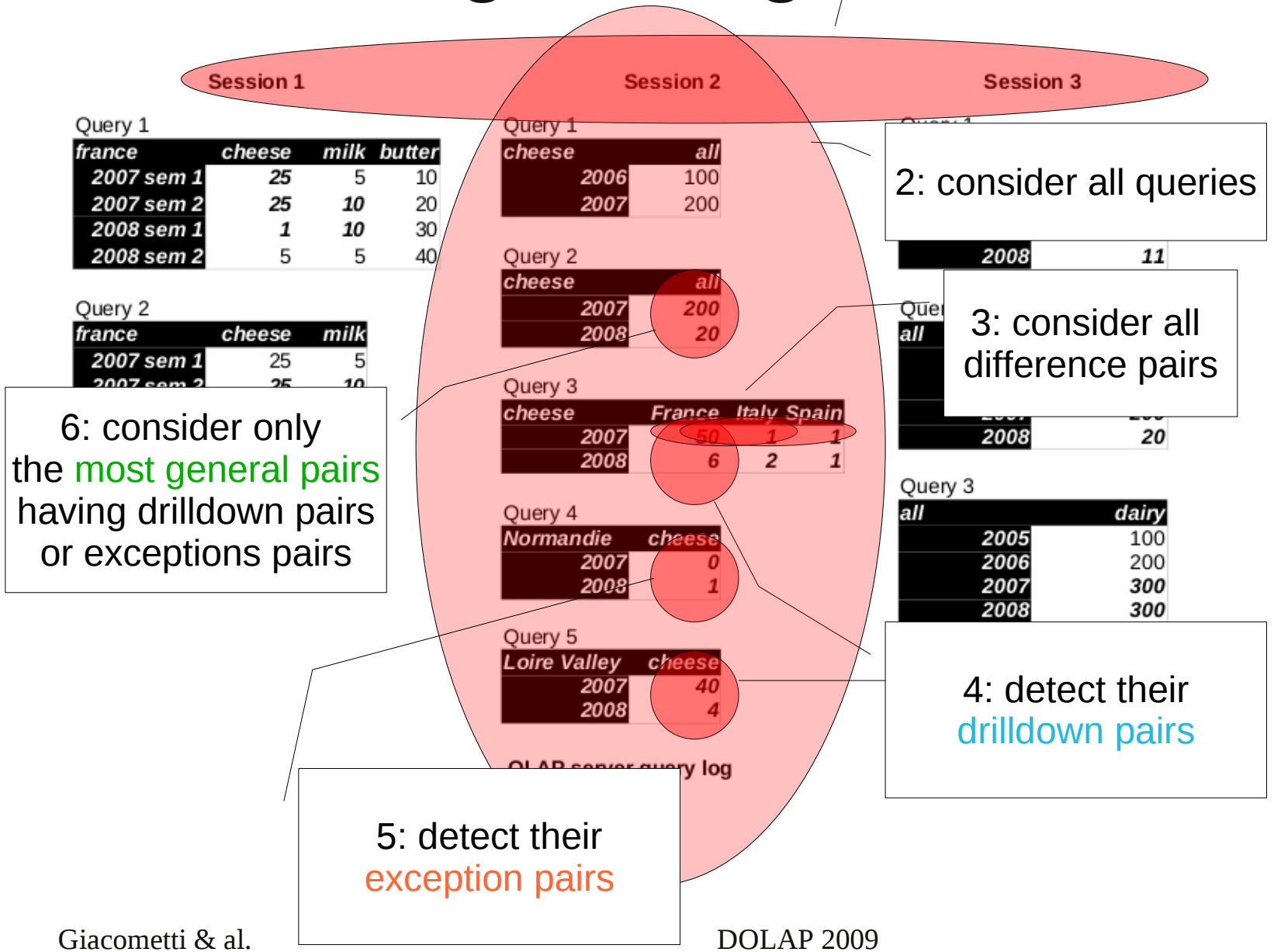
2: consider all queries

3: consider all difference pairs

4: detect their drilldown pairs

5: detect their exception pairs

Processing the log



1: Consider all sessions

2: consider all queries

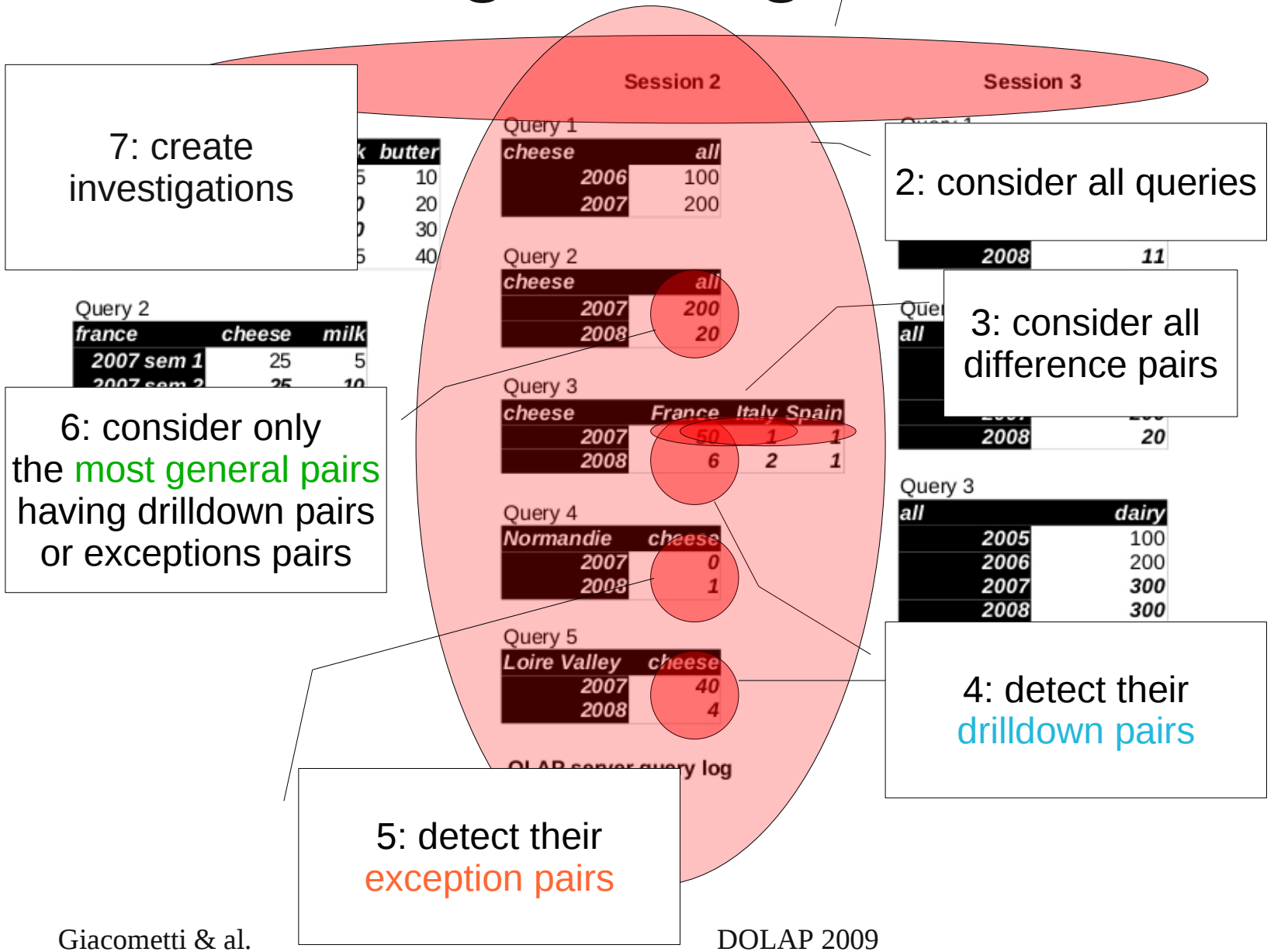
3: consider all difference pairs

4: detect their drilldown pairs

5: detect their exception pairs

6: consider only the most general pairs having drilldown pairs or exceptions pairs

Processing the log



Recommending

1: detect difference pairs

Session 1

Query 1

france	cheese	milk	butter
2007 sem 1	25	5	10
2007 sem 2	25	10	20
2008 sem 1	1	10	30
2008 sem 2	5	5	40

Query 2

france	cheese	milk
2007 sem 1	25	5
2007 sem 2	25	10
2008 q1	0.5	5
2008 q2	0.5	5
2008 q3	2	3
2008 q4	3	2

Session 2

Query 1

cheese	all
2006	100
2007	200

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

Loire Valley	cheese
2007	40
2008	4

OLAP server query log

Session 3

Query 1

all	goat cheese
2005	10
2006	11
2007	10
2008	11

Query 2

all	cheese
2005	50
2006	100
2007	200
2008	20

Query 3

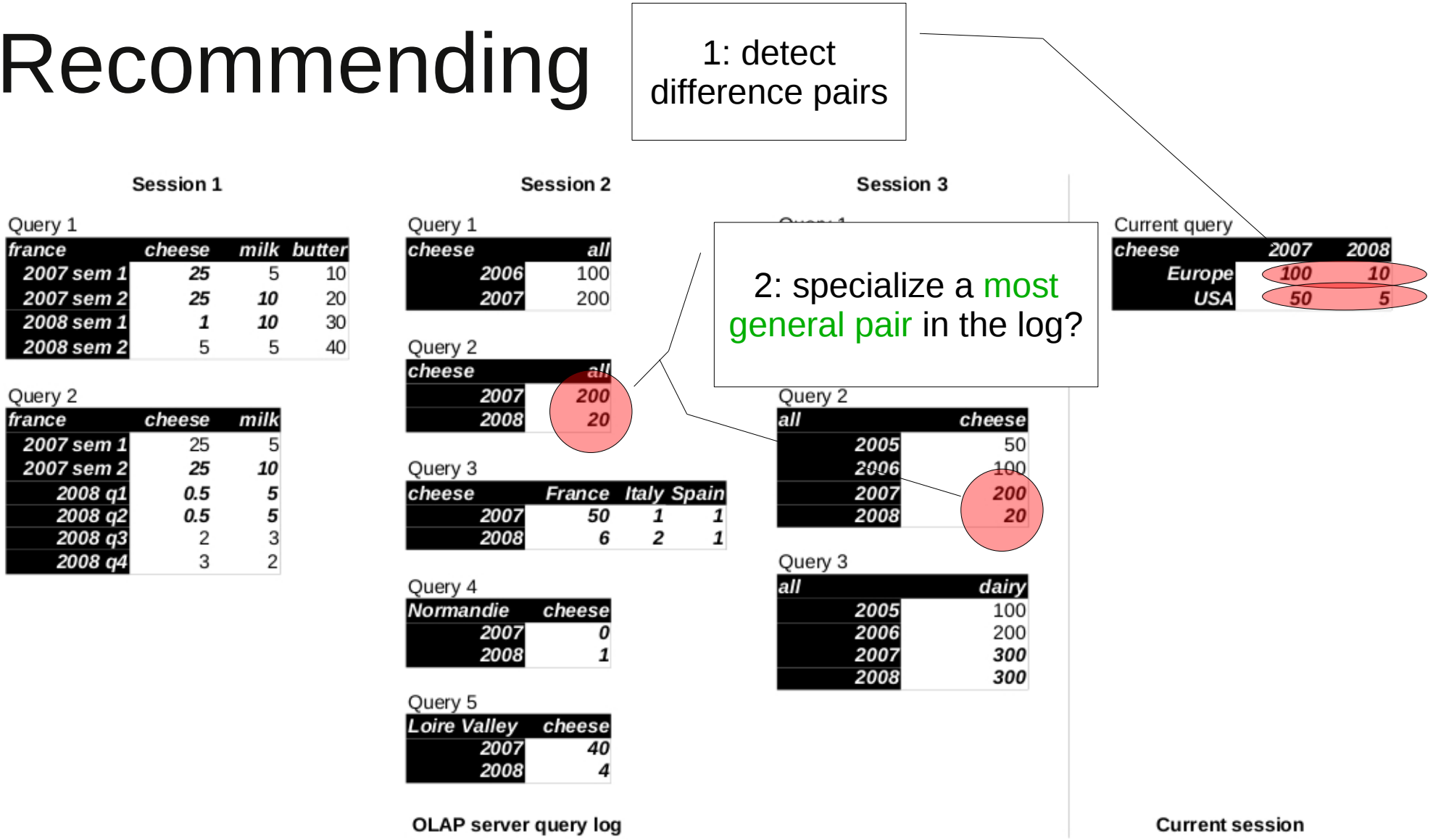
all	dairy
2005	100
2006	200
2007	300
2008	300

Current query

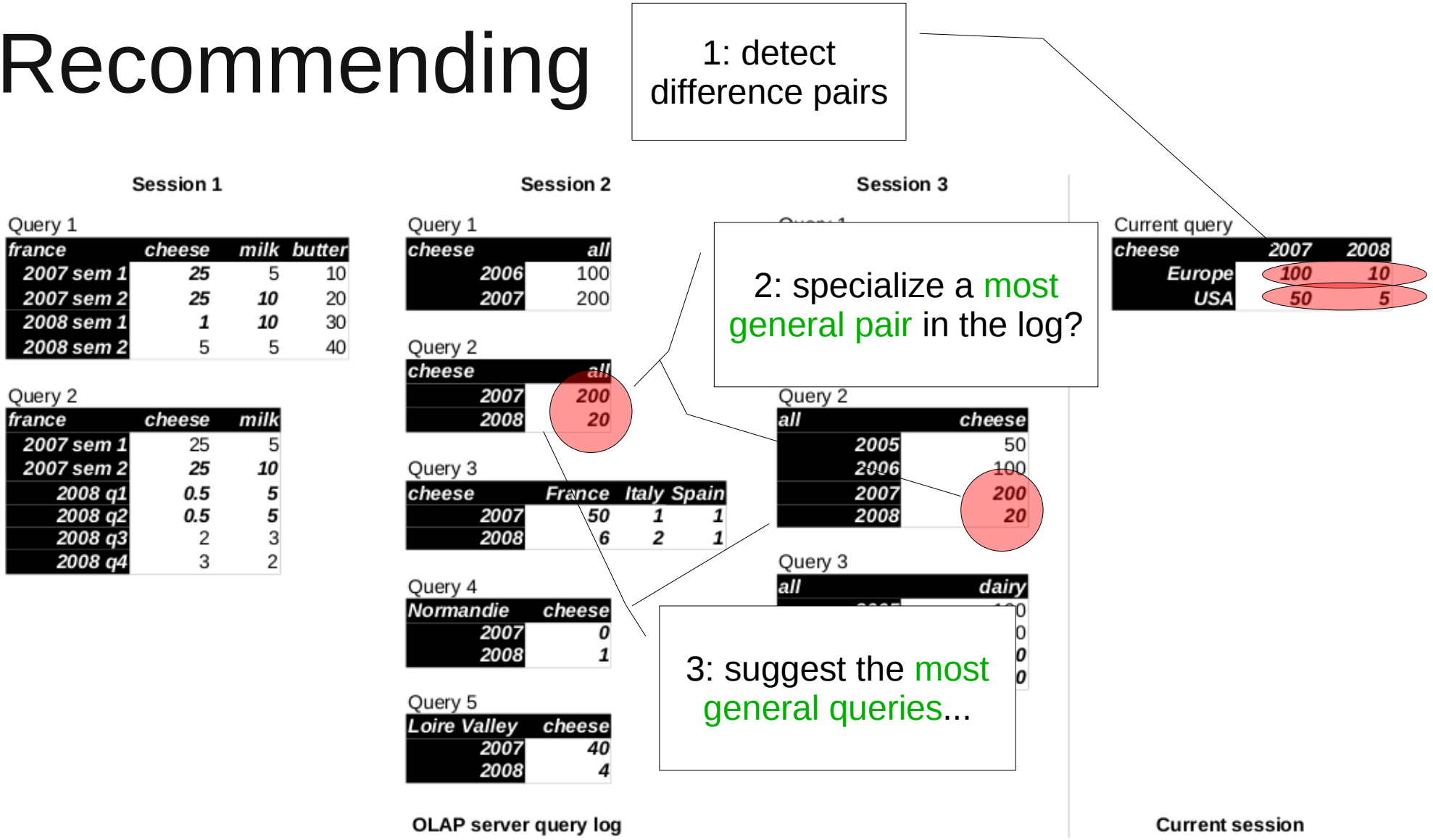
cheese	2007	2008
Europe	100	10
USA	50	5

Current session

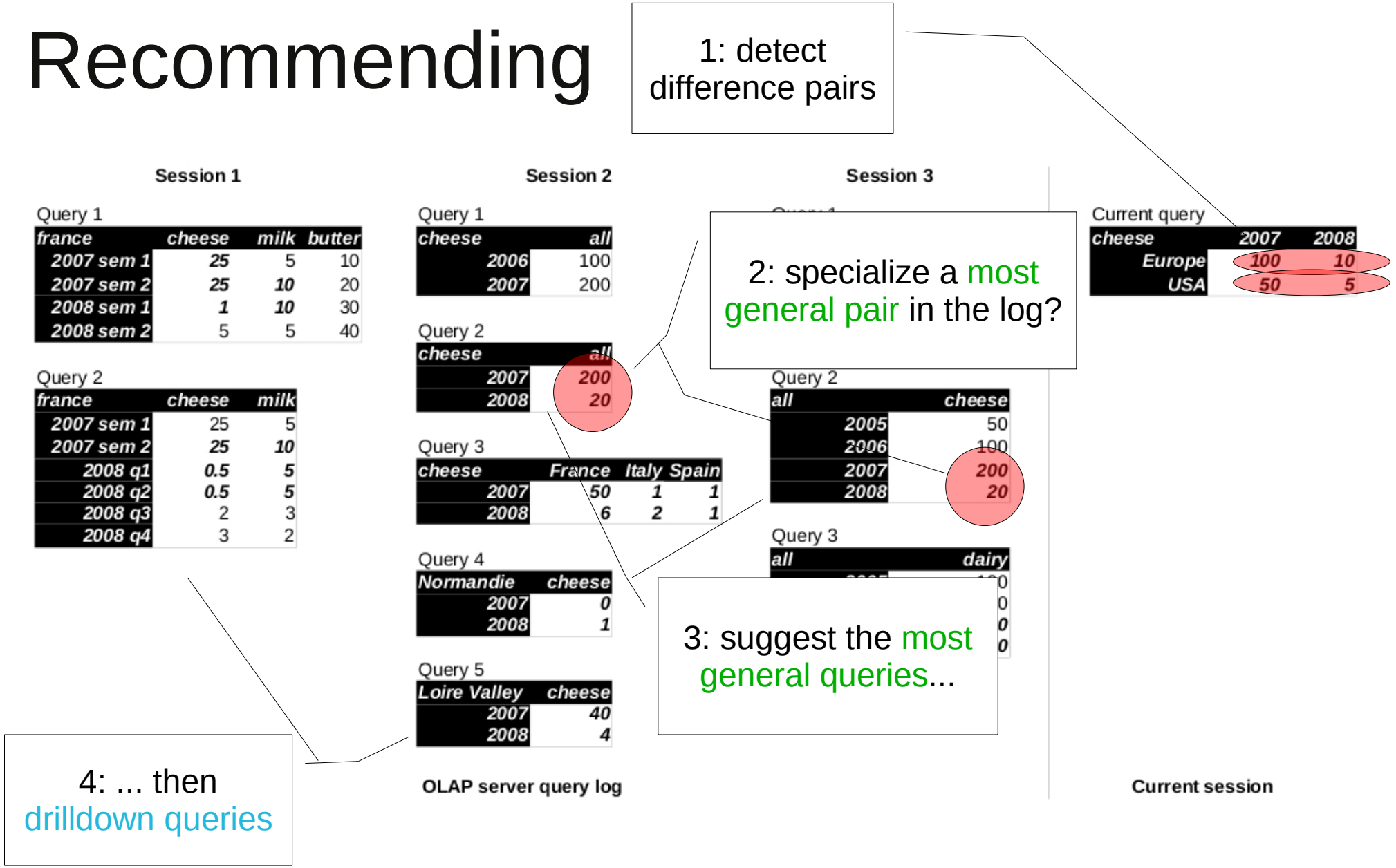
Recommending



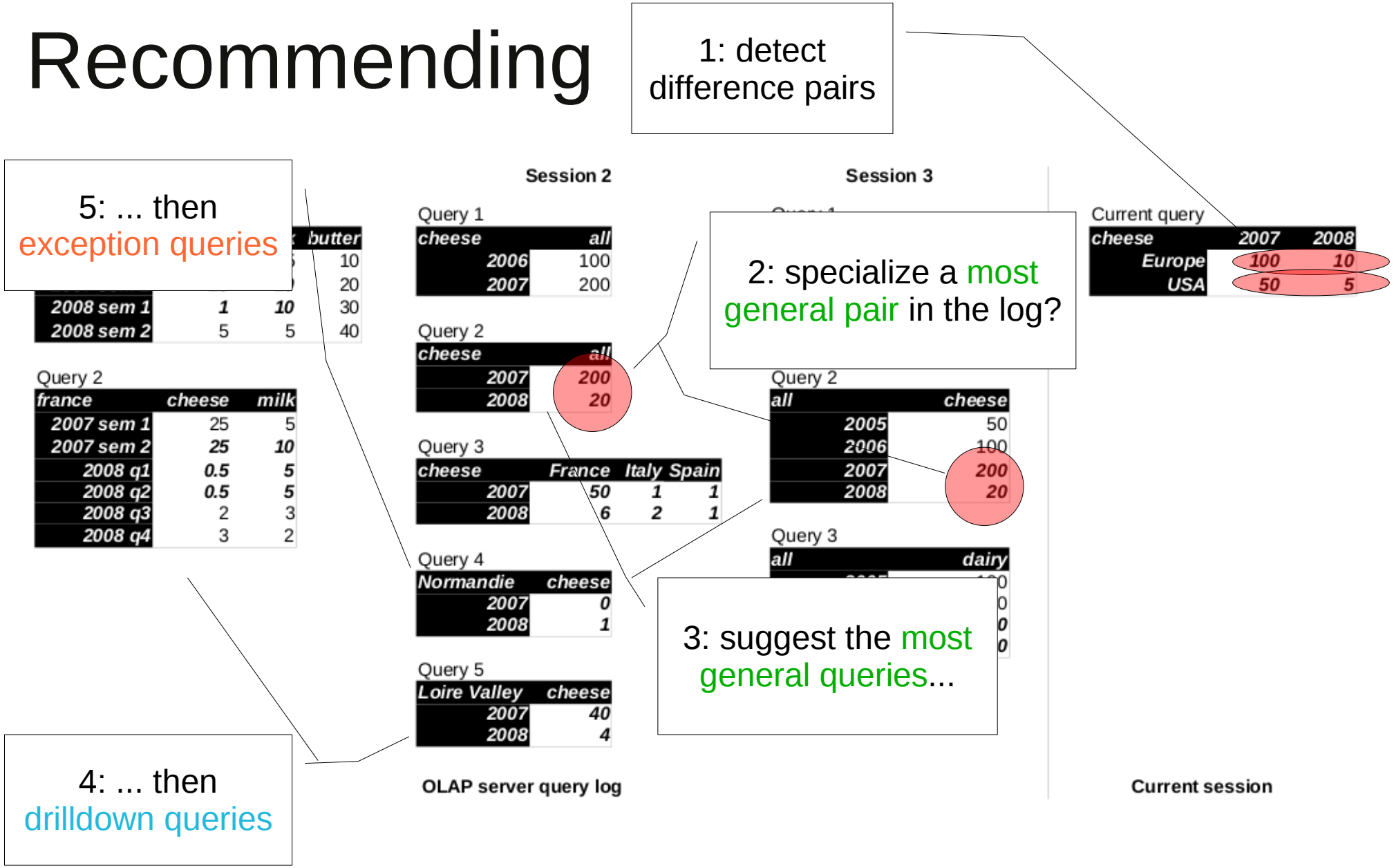
Recommending



Recommending



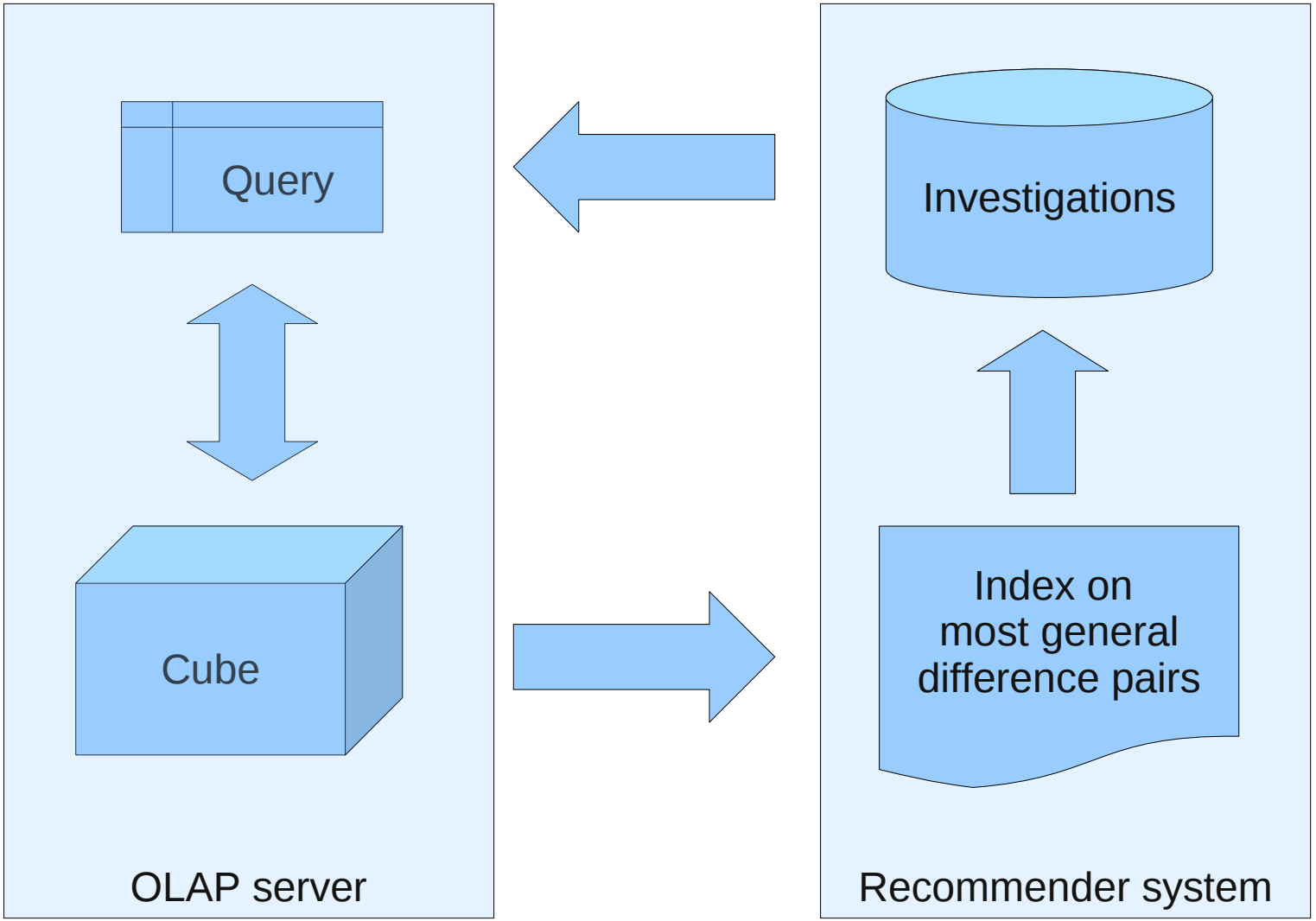
Recommending



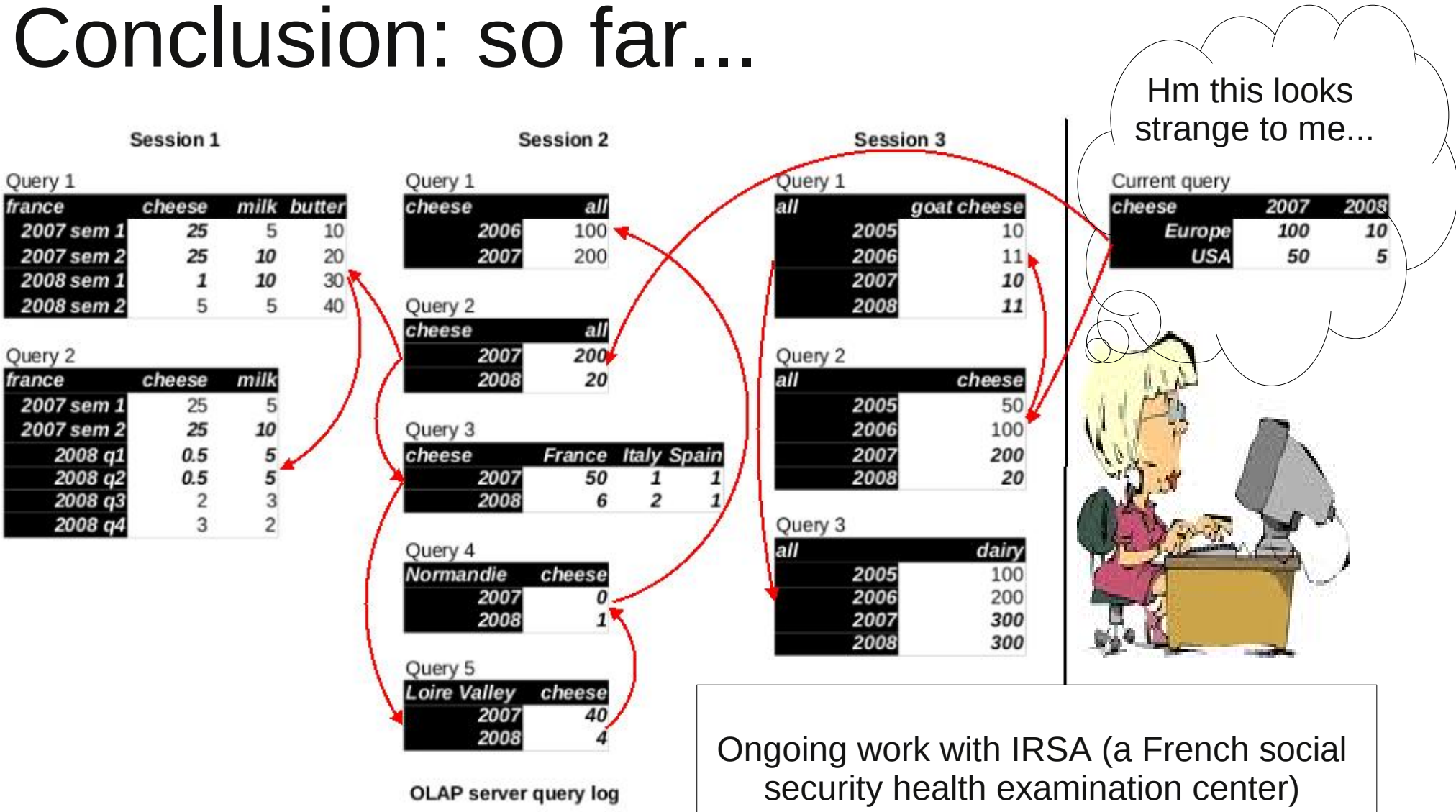
Feasability? So far...

- A naive implementation
  - Java, mondrian OLAP engine & Sarawagi's icube
  - The log is (of course) processed offline
  - But queries are resubmitted since results are needed
- Complexity of recommending
  - ndp(2(co+cp)+ni(co+cp))
    - ndp: number of difference pairs in the current query
    - ni: number of inverstigations
    - co: cost of Sarawagi's icube operators
    - cp: cost of presenting the result

Perspective: A possible architecture



Conclusion: so far...



Ongoing work with IRSA (a French social security health examination center) to analyze over 500.000 health care examination questionnaires

Conclusion: ... what next?

Session 1

Query 1

france	cheese	milk	butter
2007 sem 1	25	5	10
2007 sem 2	25	10	20
2008 sem 1	1	10	30
2008 sem 2	5	5	40

Query 2

france	cheese	milk
2007 sem 1	25	5
2007 sem 2	25	10
2008 q1	0.5	5
2008 q2	0.5	5
2008 q3	2	3
2008 q4	3	2

Session 2

Query 1

cheese	all
2006	100
2007	200

Query 2

cheese	all
2007	200
2008	20

Query 3

cheese	France	Italy	Spain
2007	50	1	1
2008	6	2	1

Query 4

Normandie	cheese
2007	0
2008	1

Query 5

Loire Valley	cheese
2007	40
2008	4

OLAP server query log

Session 3

Query 1

all	goat cheese
2005	10
2006	11
2007	10
2008	11

Query 2

all	cheese
2005	50
2006	100
2007	200
2008	20

Query 3

all	dairy
2005	100
2006	200
2007	300
2008	300

Conclusion: ... what next?

Session 1

Query 1

france	cheese	milk	butter
2007 sem 1	25	5	10
2007 sem 2	25	10	20
2008 sem 1	1	10	30
2008 sem 2			

Query 2

france	chee
2007 sem 1	
2007 sem 2	
2008 q1	
2008 q2	
2008 q3	
2008 q4	

Session 2

Query 1

cheese	all
2006	100
2007	200

Query 5

Loire Valley	cheese
2007	40
2008	4

OLAP server query log

Session 3

Query 1

all	goat cheese
2005	10
2006	11
2007	10
2008	11

	cheese
2005	50
2006	100
2007	200
2008	20

	dairy
2005	100
2006	200
2007	300
2008	300

Can the system combine queries?
Are there queries of better quality?
...

Hm this looks strange to me...

Current query

cheese	2007	2008
Europe	100	10
USA	50	5

I prefer to see exceptions better than to rollup
And I also prefer to see Salespersons better than Locations...

Conclusion: ... what next?

Session 1

Query 1

france	cheese	milk	butter
2007 sem 1	25	5	10
2007 sem 2	25	10	20
2008 sem 1	1	10	30
2008 sem 2			

Query 2

france	chee
2007 sem 1	
2007 sem 2	
2008 q1	
2008 q2	
2008 q3	
2008 q4	

Session 2

Query 1

cheese	all
2006	100
2007	200

Session 3

Query 1

all	goat cheese
2005	10
2006	11
2007	10
2008	11

	cheese
2005	50
2006	100
2007	200
2008	20

	dairy
2005	100
2006	200
2007	300
2008	300

Can the system combine queries?
Are there queries of better quality?
...

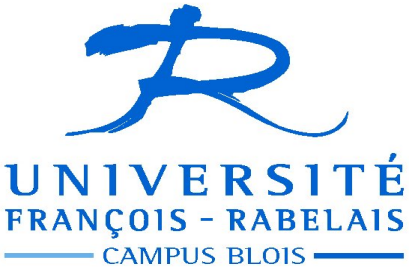
The bigger picture: see Khoussainova & al. CIDR 2009
'A case for a collaborative query management system'

Hm this looks strange to me...

Current query

cheese	2007	2008
Europe	100	10
USA	50	5

I prefer to see exceptions better than to rollup
And I also prefer to see Salespersons better than Locations...



Query Recommendations for OLAP Discovery Driven Analysis

Thanks for your attention,
Any questions?

Advantage of the approach

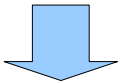
- Compared to simply using Sarawagi's operators
  - This is collaborative filtering
    - what other users found is relevant
    - Filters out what is not in the log
  - Answer is a query, not a set of cells
  - One step further in automating the analysis

Complexity

- Of processing the log:
  - nq*ndp (co + nq log nq +nq*ndp(2co+1))
    - ndp: number of difference pairs in the log
    - nq: number of queries in the log
    - co: cost of Sarawagi's icube operators

Item recommendation vs query recommendation

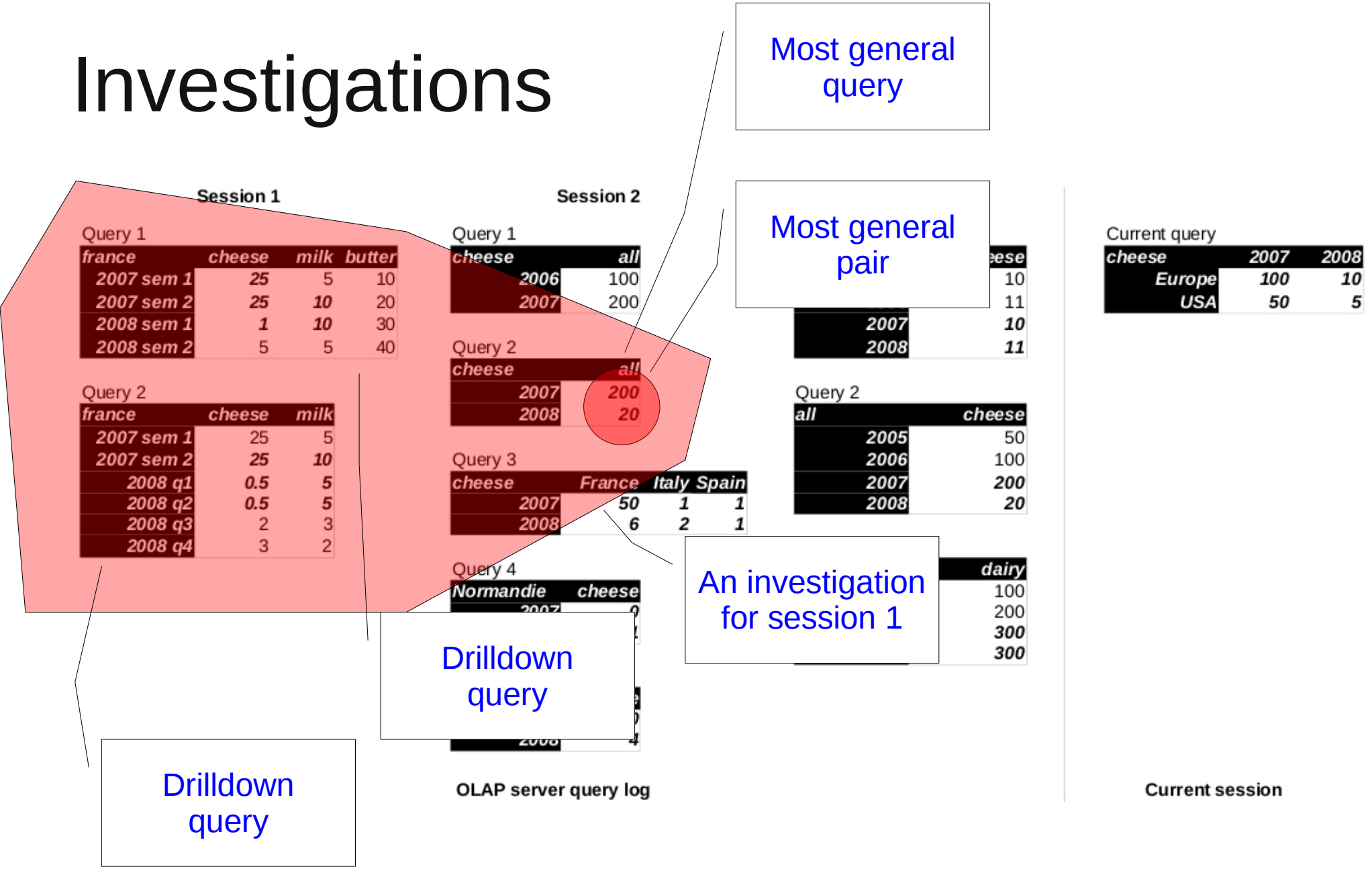
e-commerce	Item 1	...	Item m
User 1	rating_1_1		??
...		...	
User n	??		rating_n_m



OLAP	query 1	...	query m
session 1	rating_1_1		??
...		...	
session n	??		rating_n_m

- Very large
- very sparse
- Compute ??
- Recommend highest

Investigations



Difference pairs and queries