OVERVIEW OF DATA EXPLORATION TECHNIQUES

Stratos Idreos, Olga Papaemmanouil, Surajit Chaudhuri

SIGMOD 2015, Melbourne

USER INTERACTION

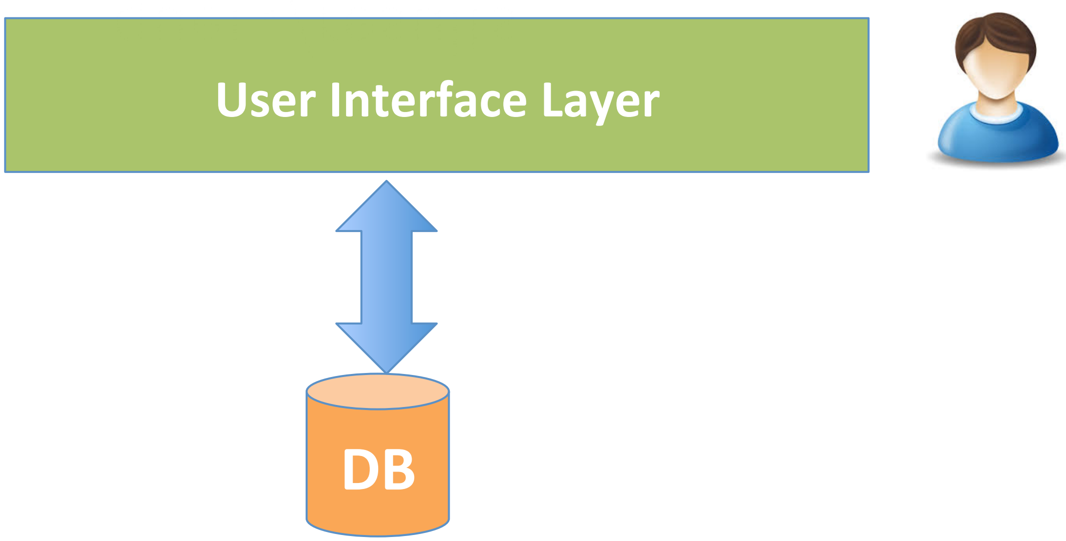
express interests

collaborate

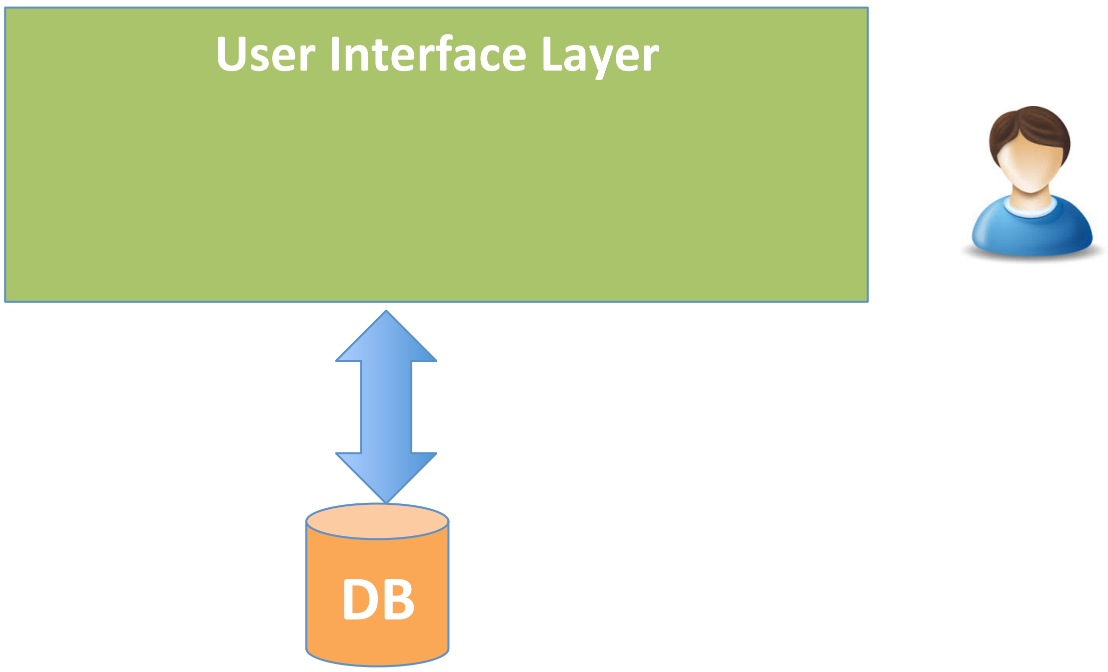
query/results recommendations

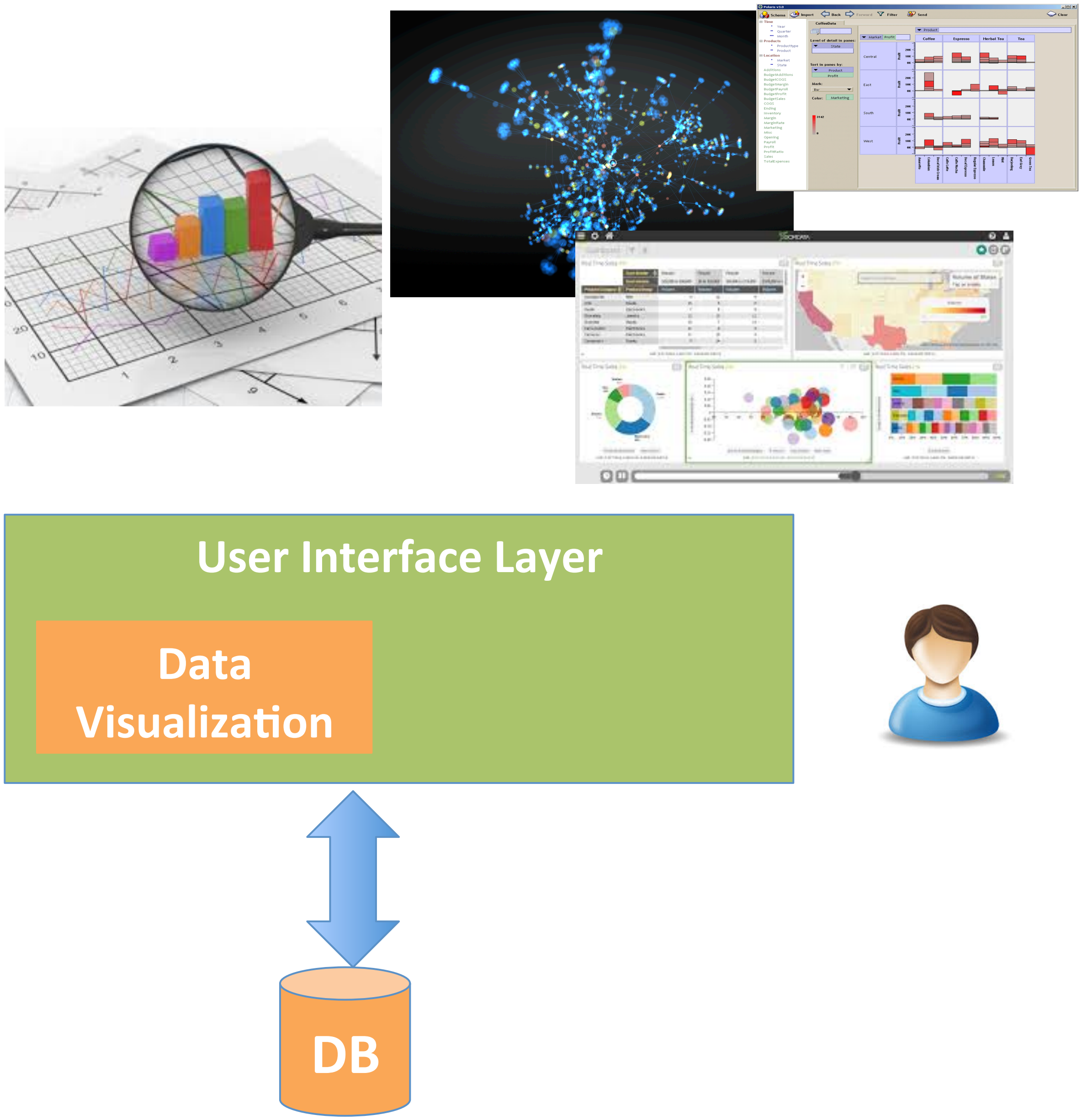
annotate

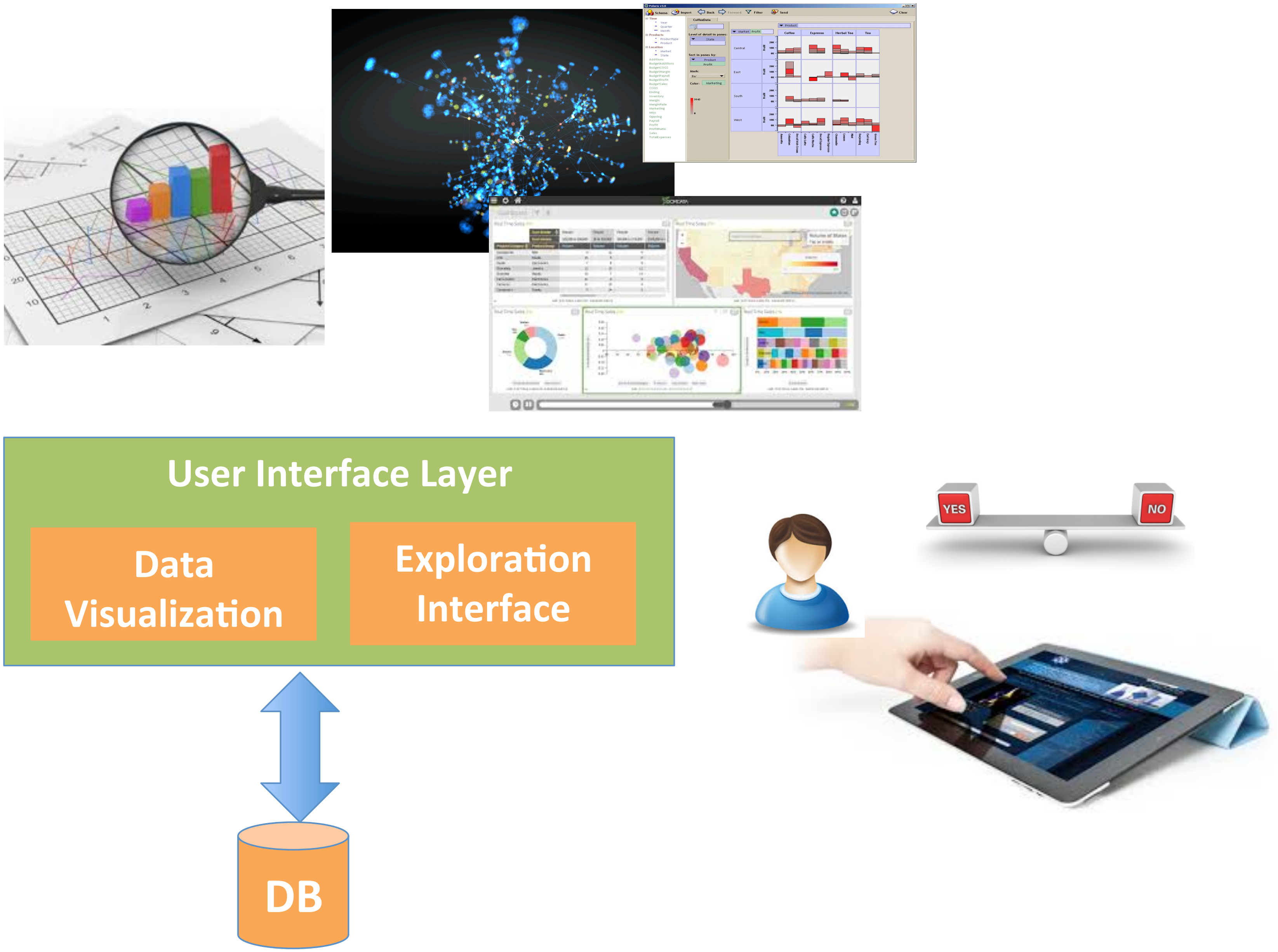
visualize results



assisted query formulation

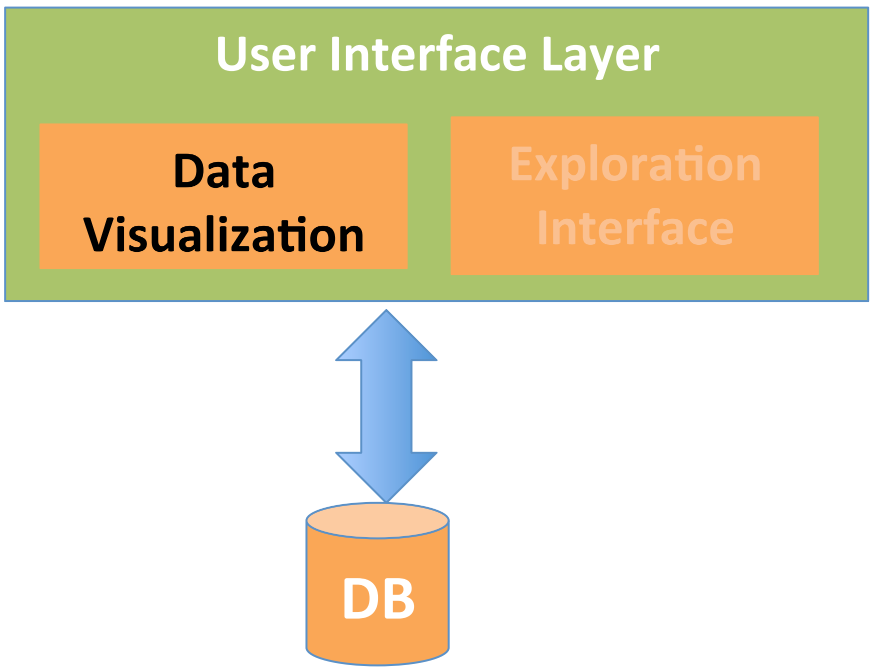

User Interface Layer
DB


User Interface Layer
Data Visualization
DB


User Interface Layer
Data Visualization
Exploration Interface
DB
YES
NO

data visualization

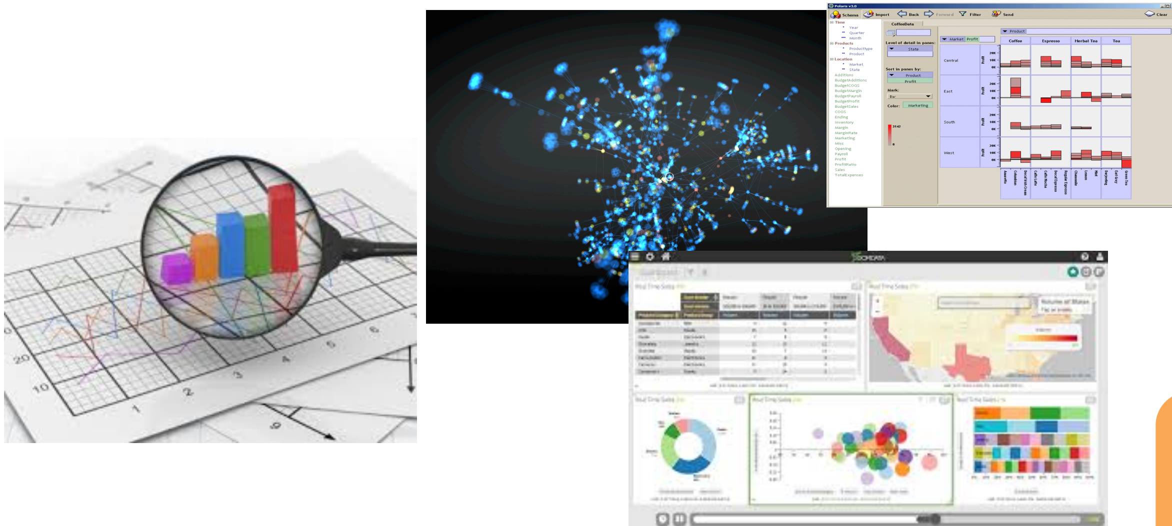
visualization tools



visual optimizations

automatic visualization

data visualization



User Interface Layer

Data Visualization

Exploration Interface

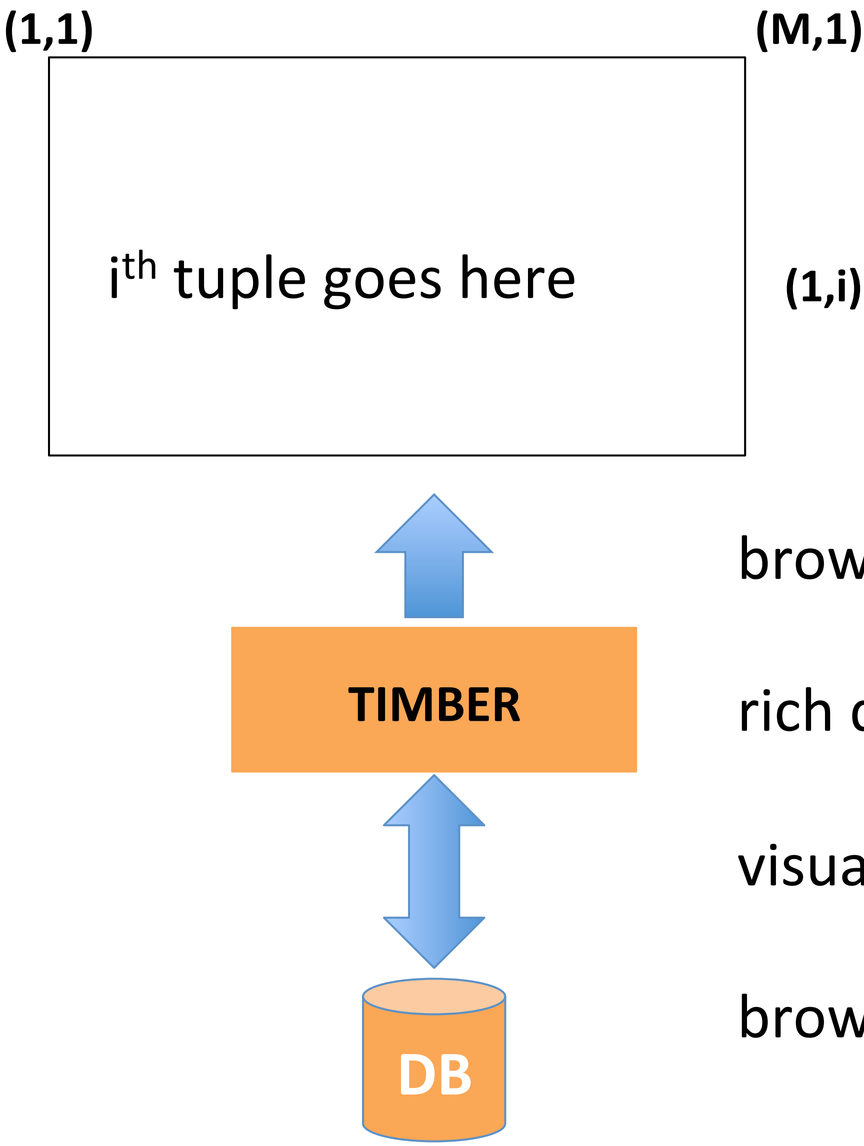
DB

visualization tools

visual optimizations

automatic visualization

Back in 1982...



window-based "sophisticated" browser for relational DBs

browser for multiple relations/tuples

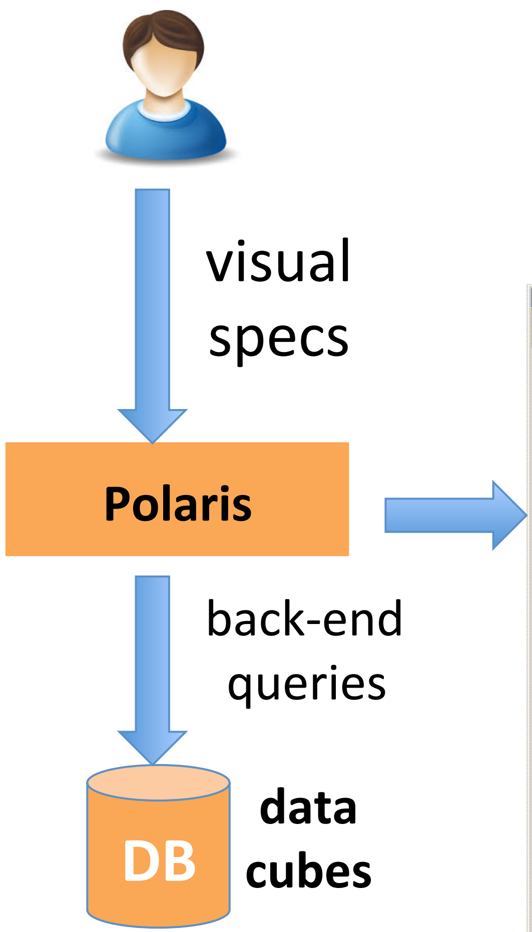
rich query language for icon-oriented DBs

visual editor of text objects

browser for geographical data

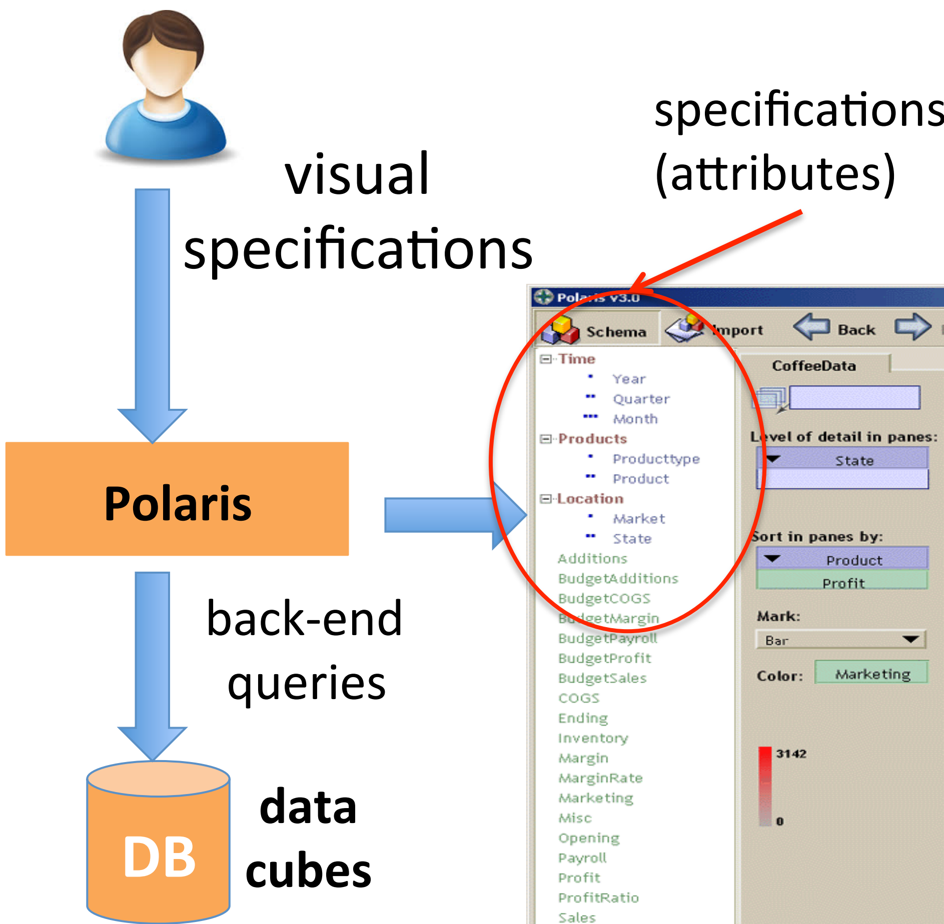
TIMBER, VLDB'82

user-driven visualizations



Polaris, INFOVIS ‘02

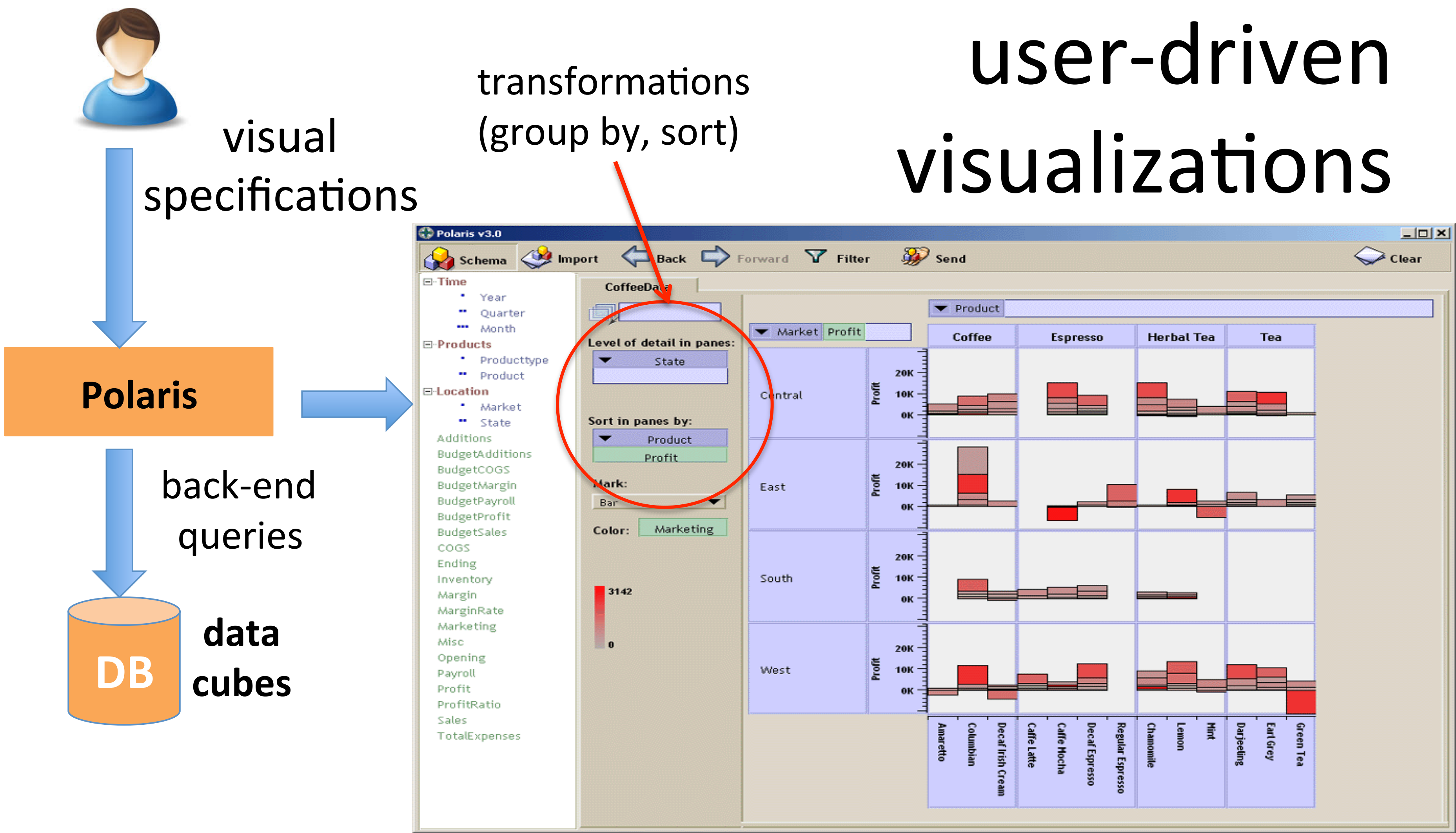
user-driven visualizations



back-end queries: data selection, partition into panes

Polaris, INFOVIS 2002

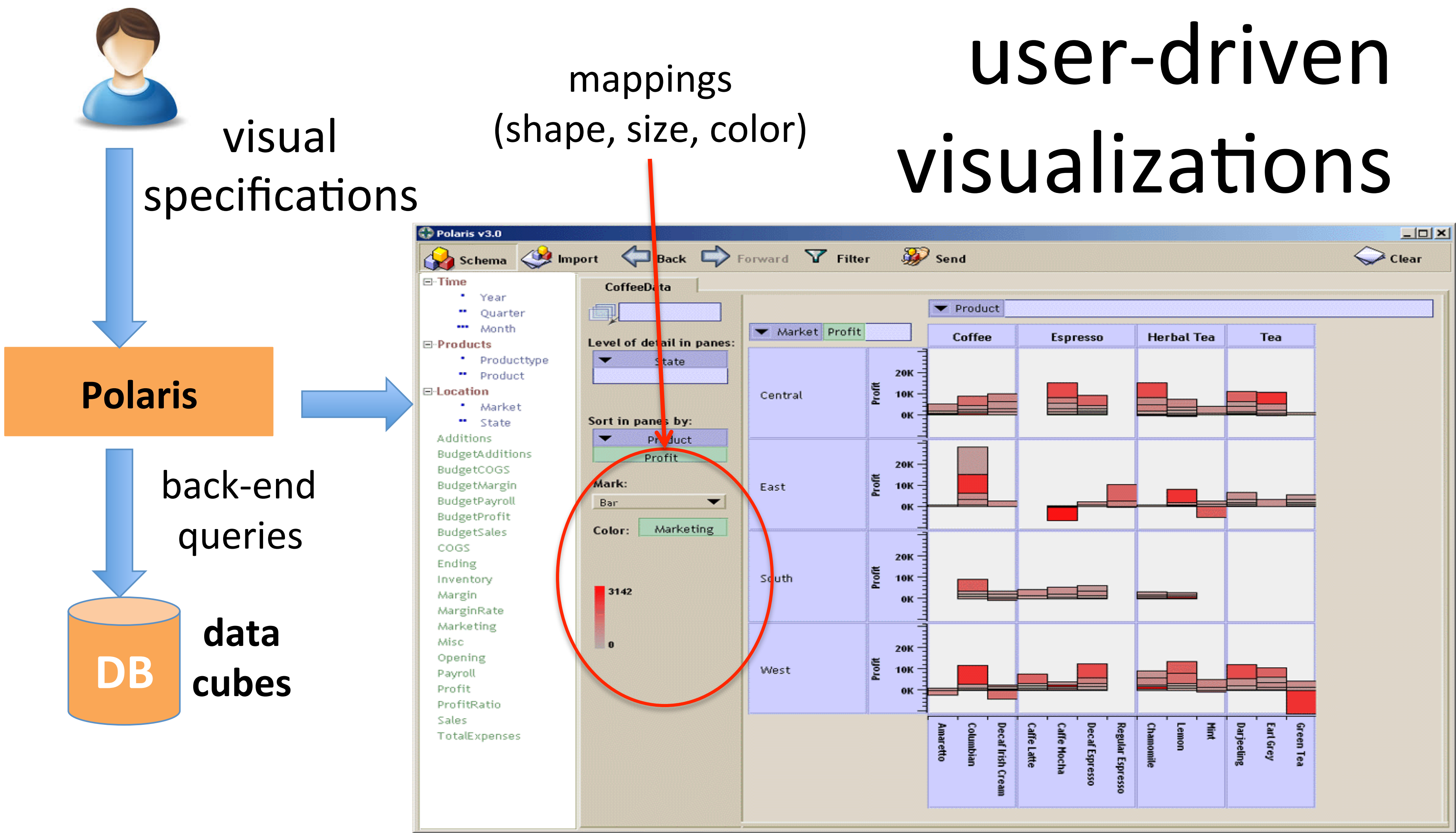
user-driven visualizations



back-end queries: data transformations
(group, sort, aggregate within each pane)

Polaris, INFOVIS 2002

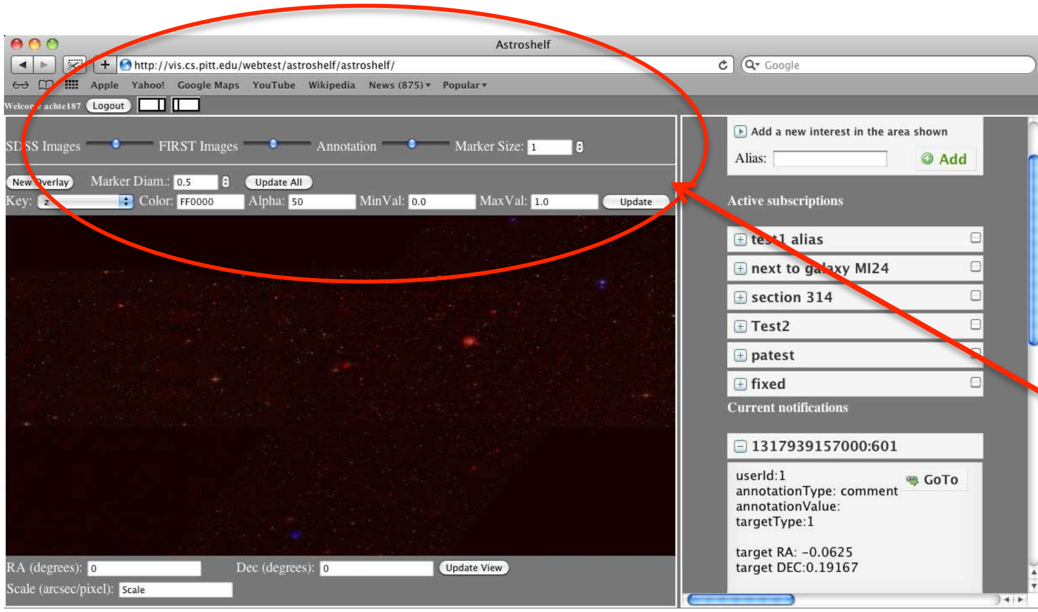
user-driven visualizations



back-end queries: graphical transformations (renter and visualize)

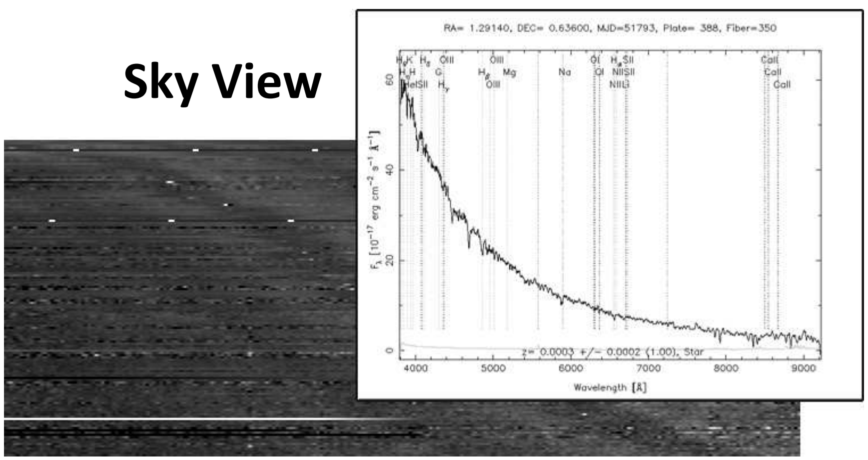
Polaris, INFOVIS 2002

collaborative exploration



live annotations

Sky View

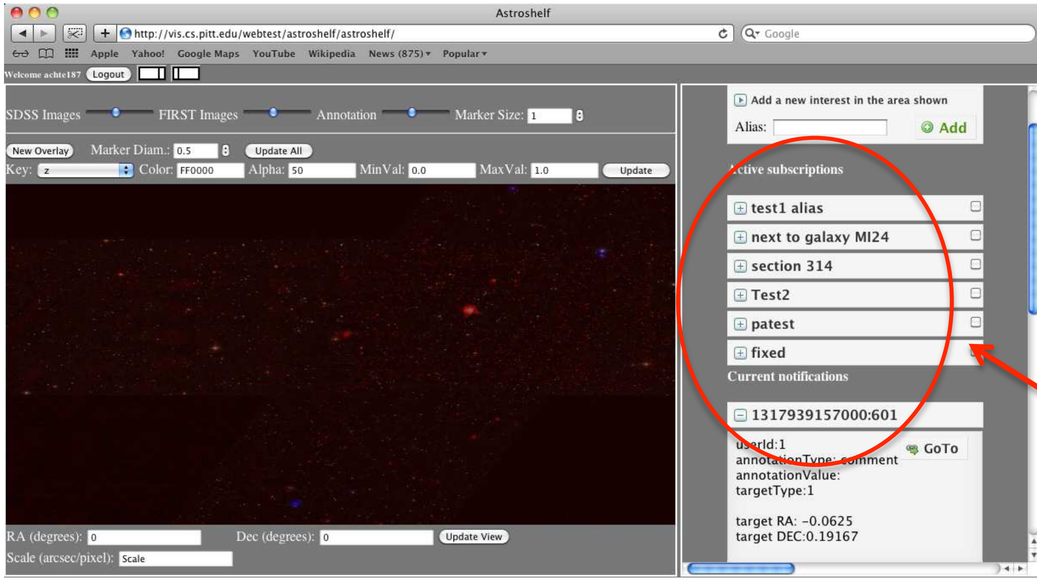


exploration for
sky objects/patterns

AstroShelf, SIGMOD '12

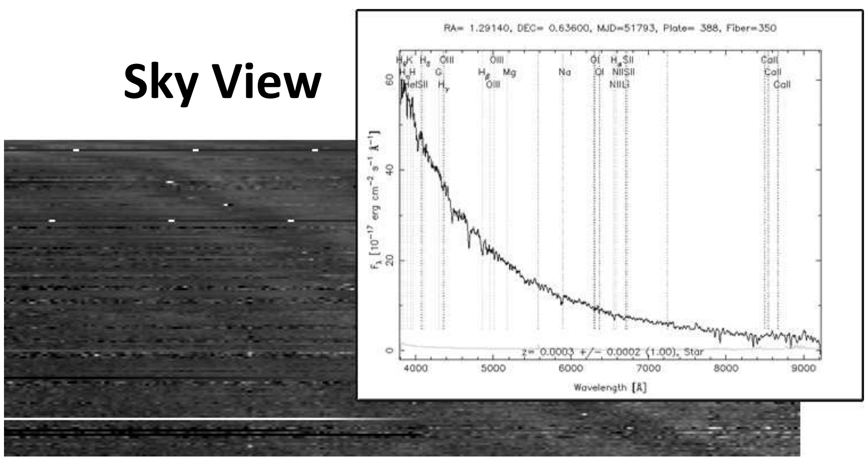
collaborative exploration

Live Annotations



subscriptions to interesting objects

Sky View

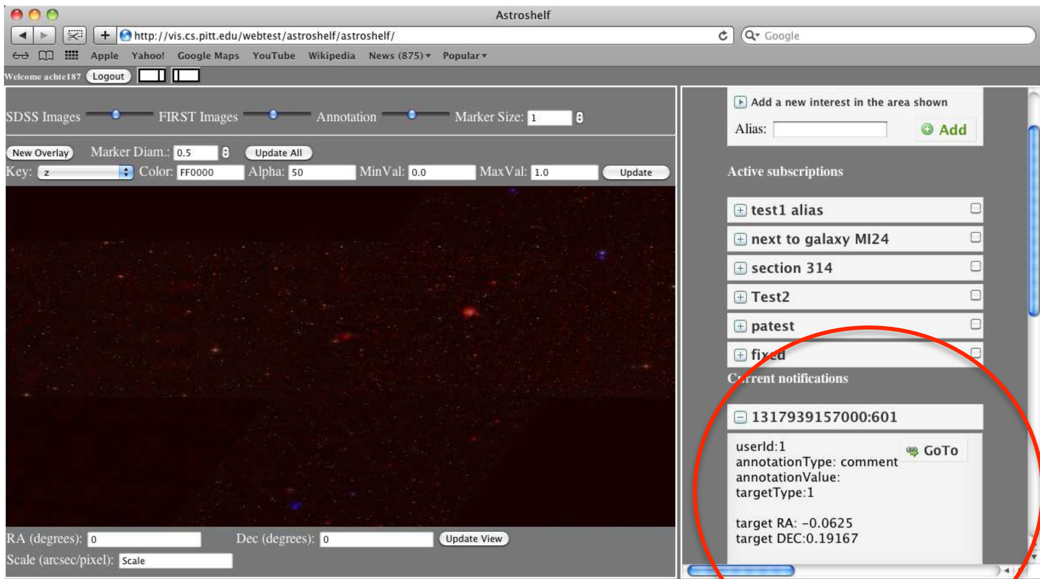


exploration for sky objects/patterns

AstroShelf, SIGMOD '12

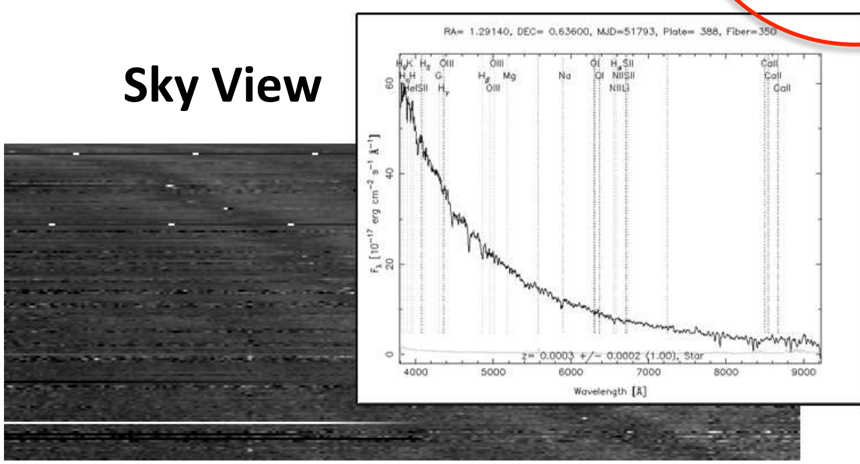
collaborative exploration

Live Annotations



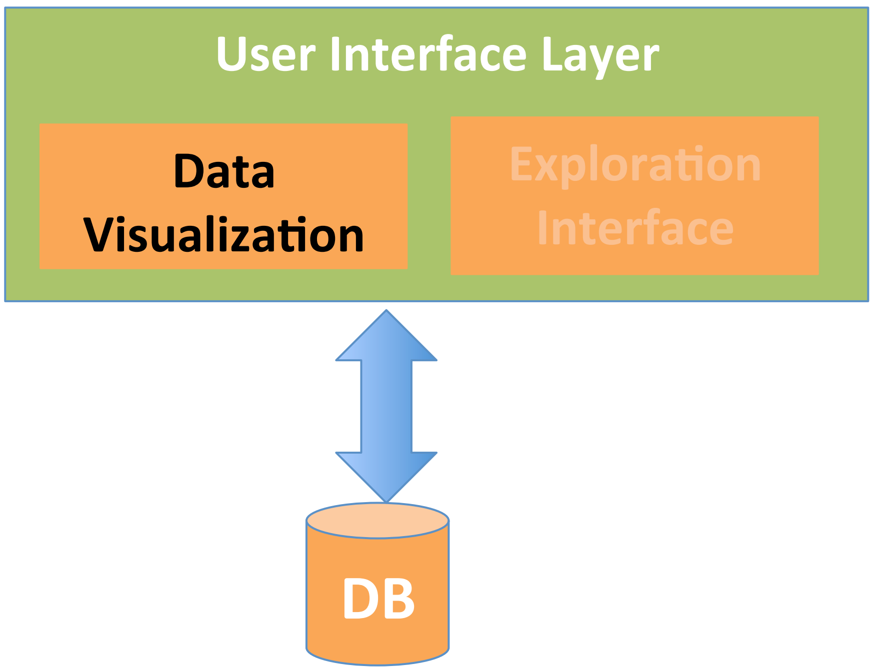
stream based notifications

Sky View



exploration for sky objects/patterns

data visualization

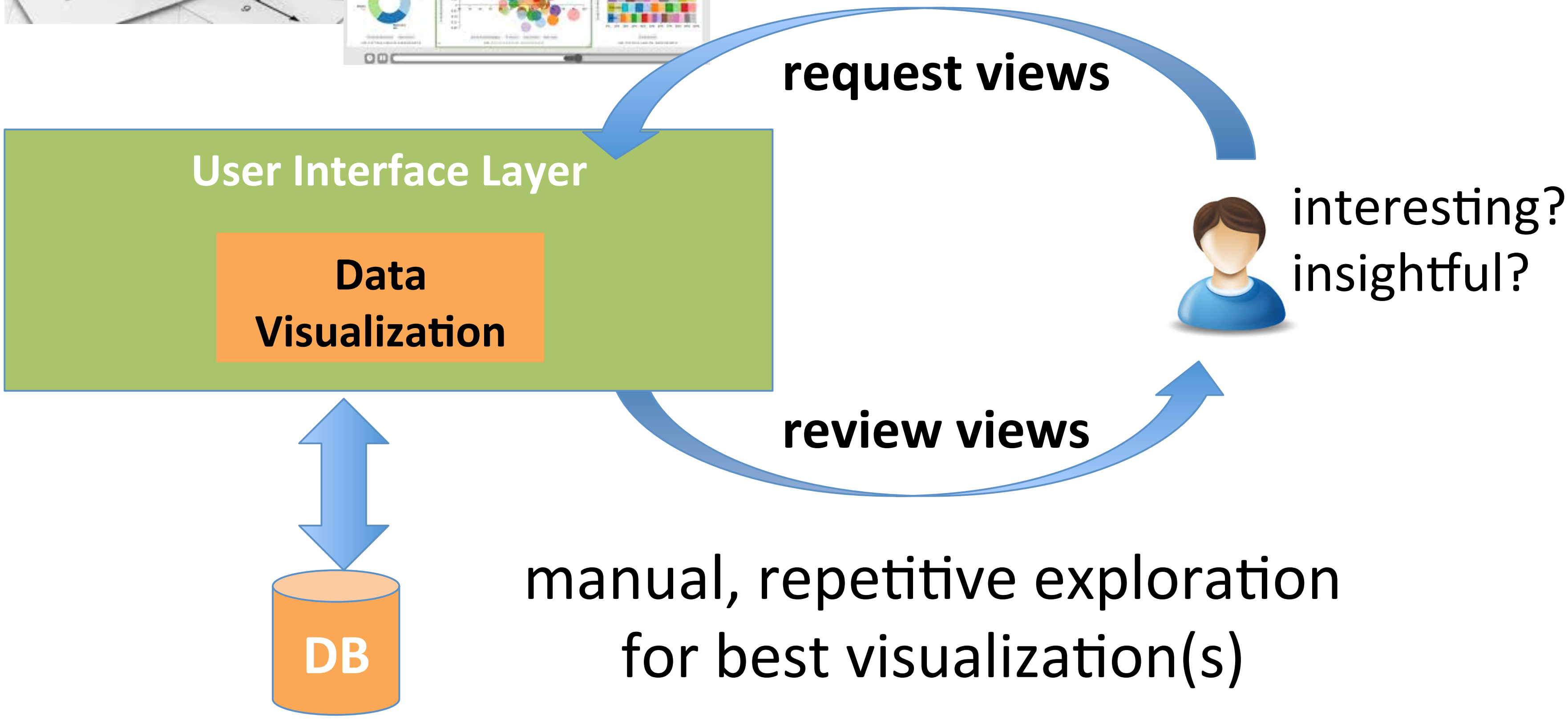


visualization tools

visual optimizations

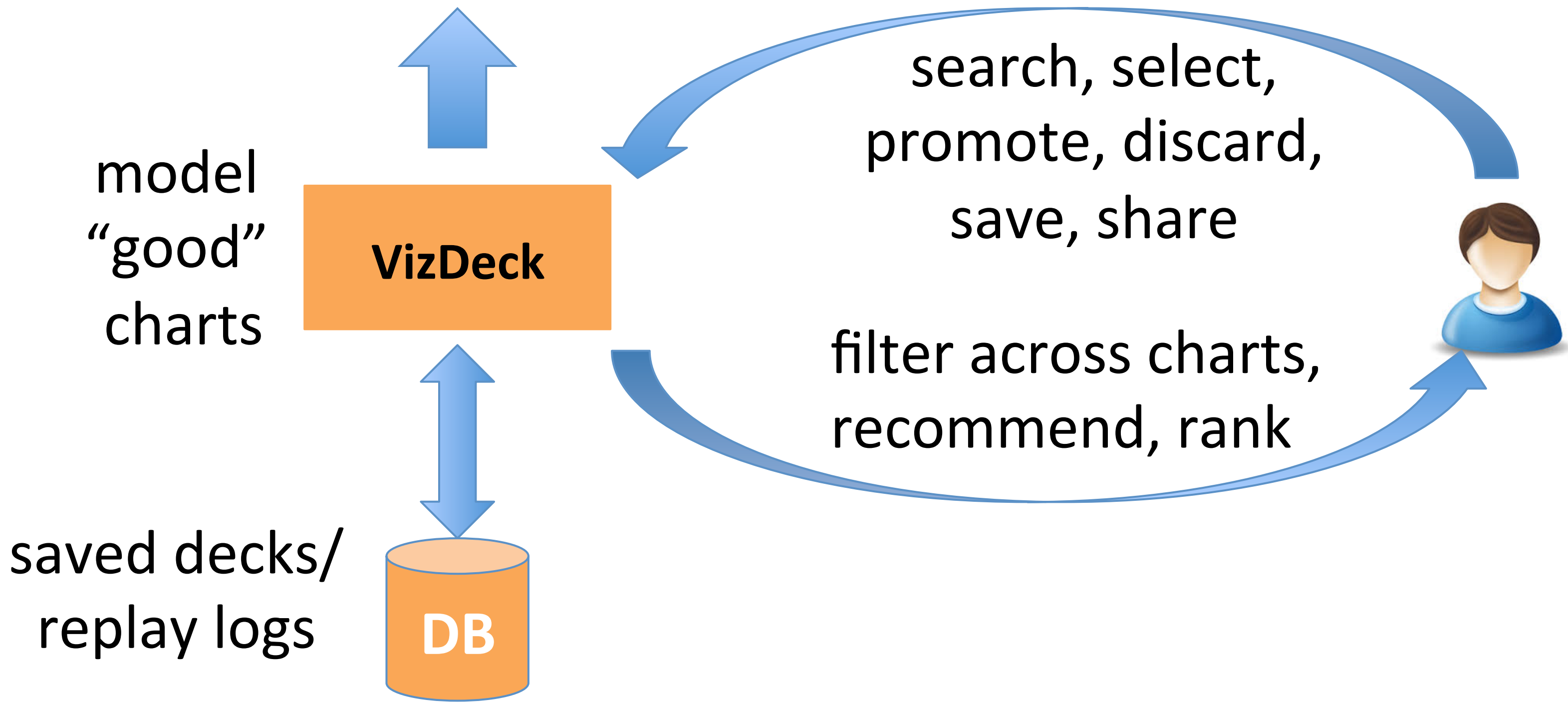
automatic visualization

automatic visualization



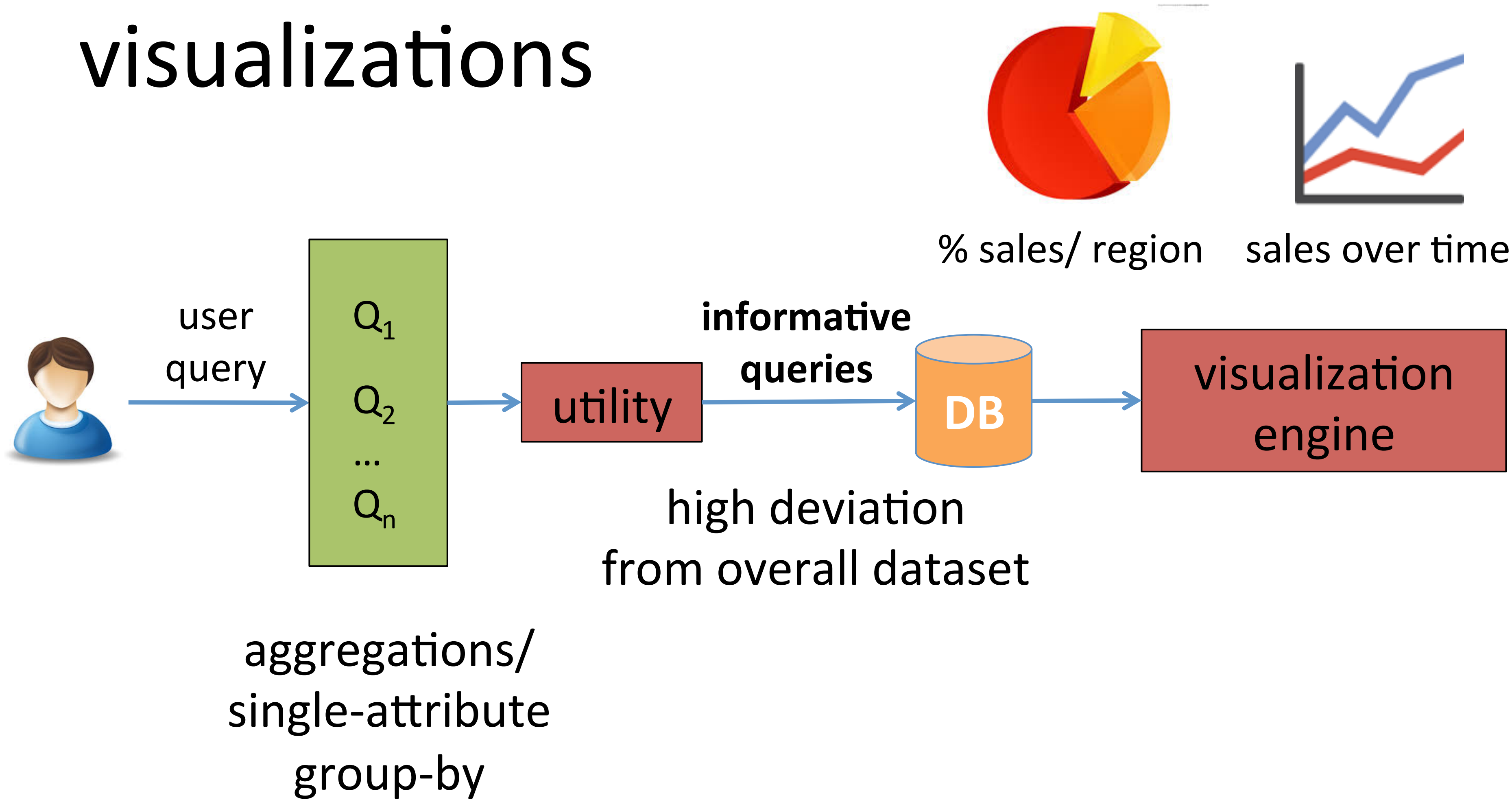
manual, repetitive exploration
for best visualization(s)

auto-ranked visualizations



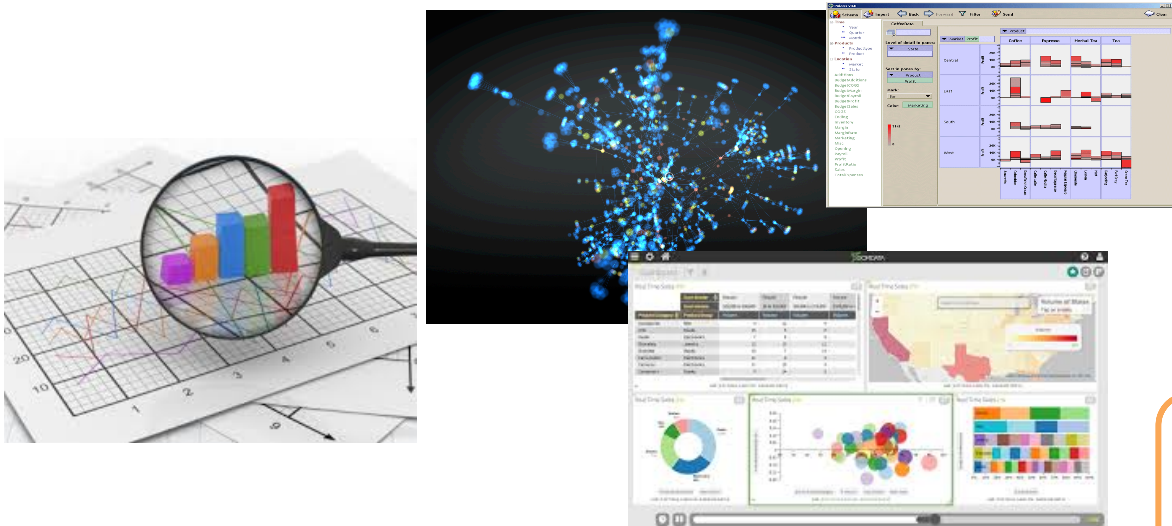
VizDeck, SIGMOD ‘12

automatic visualizations

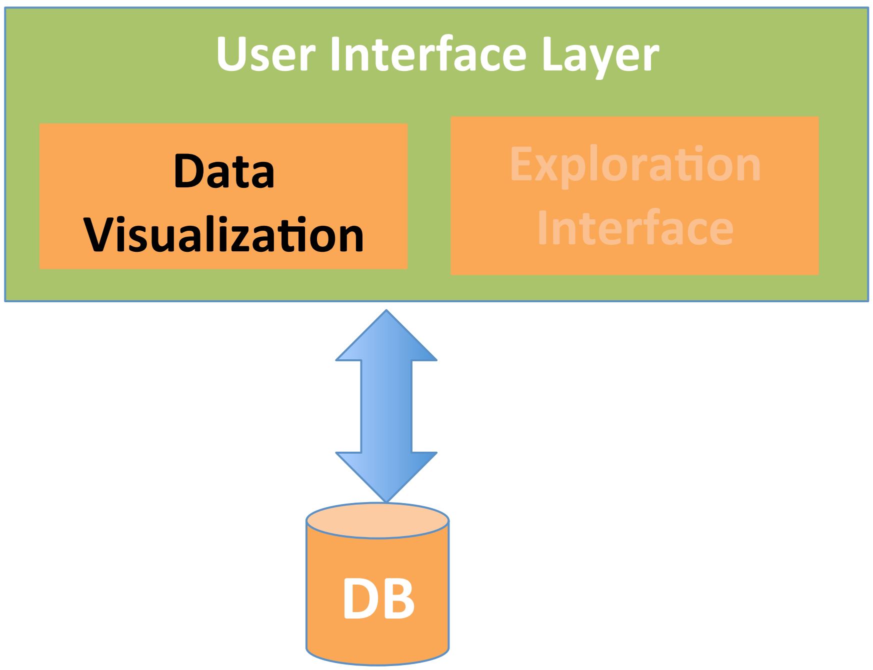


SeeDB, PVLDB'13

data visualization



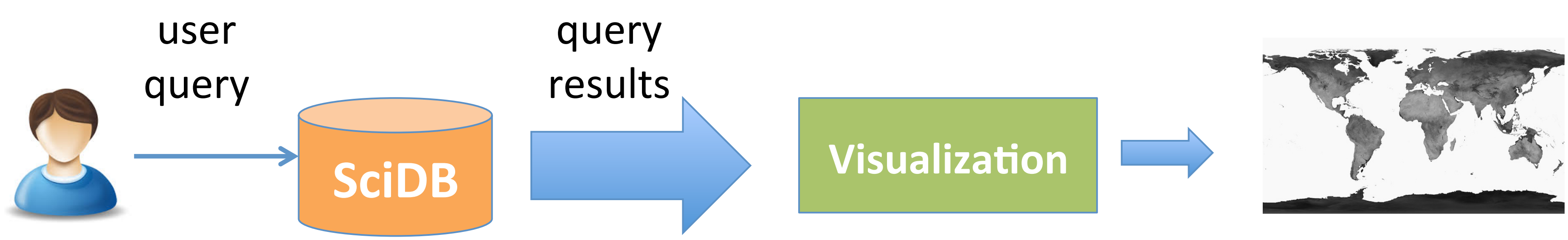
visualization tools



visual optimizations

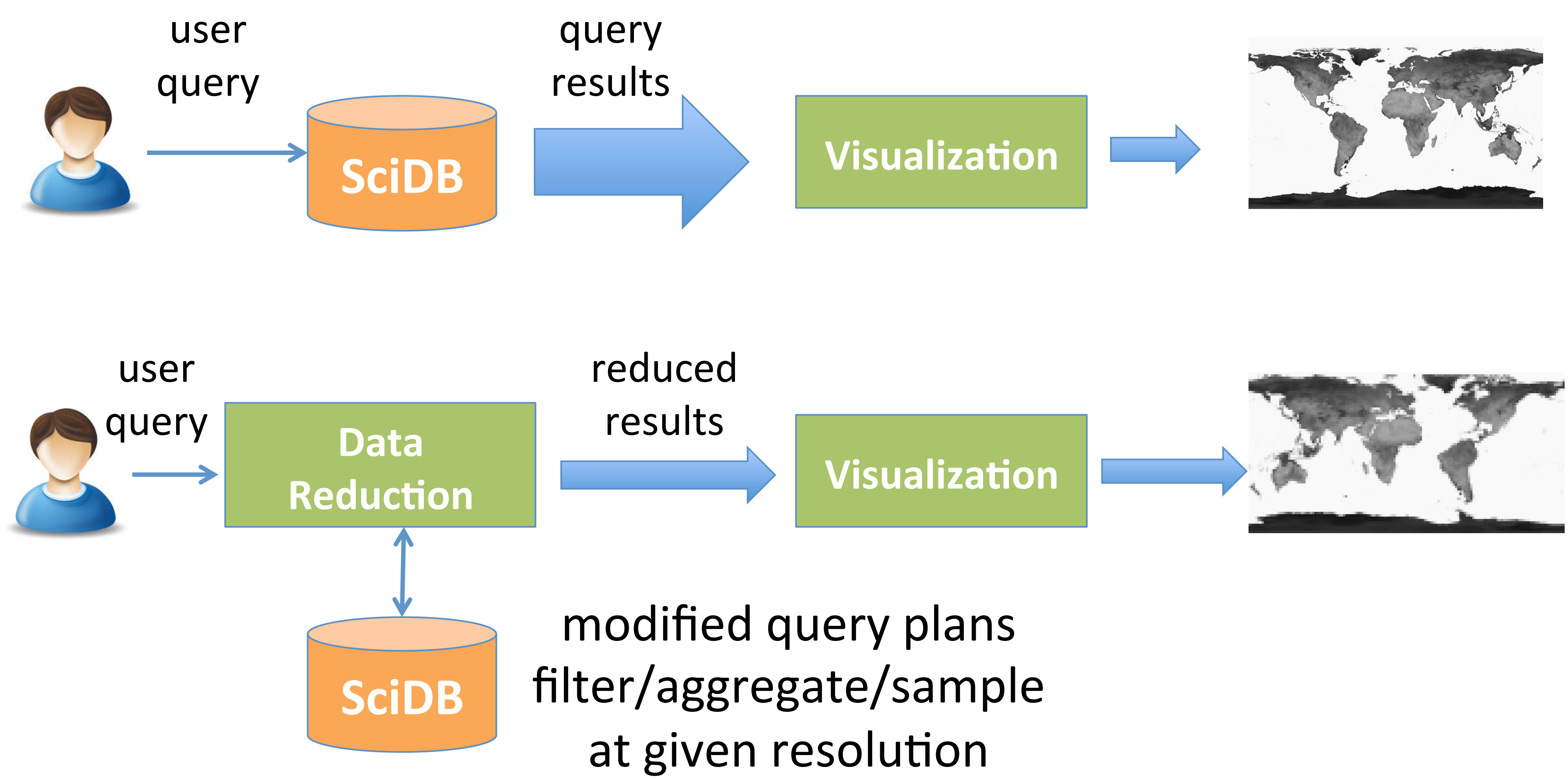
automatic visualization

resolution reduction



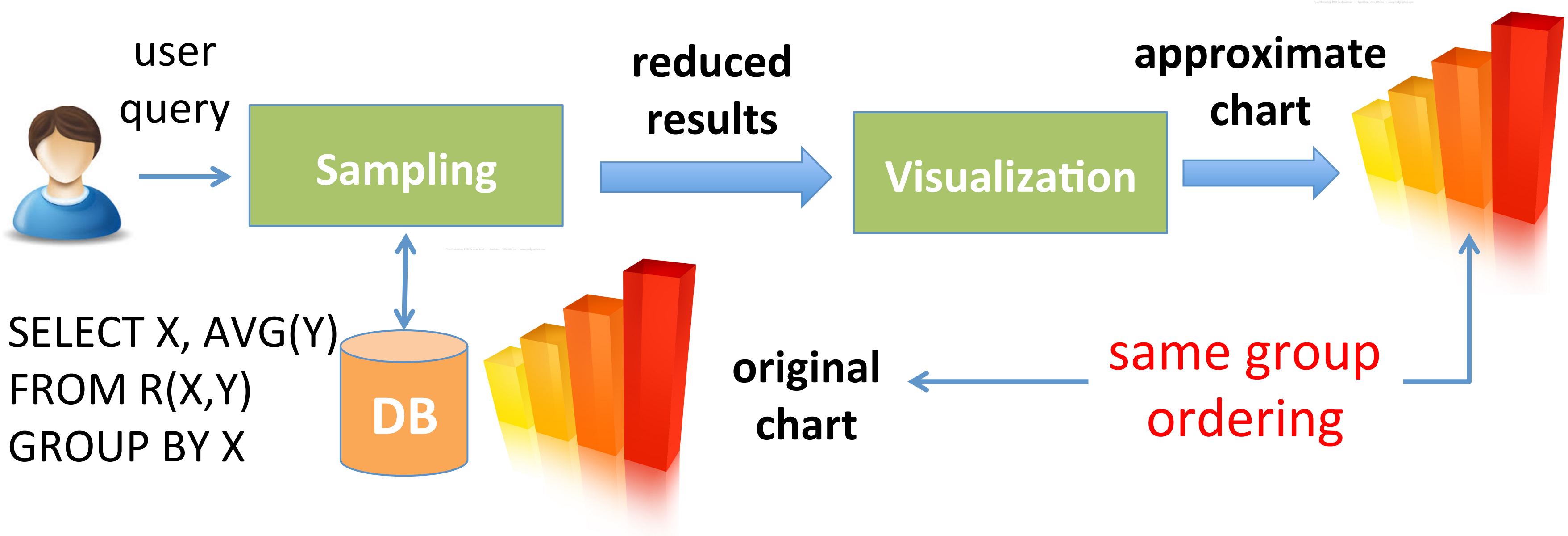
expensive, ineffective on big data sets

resolution reduction



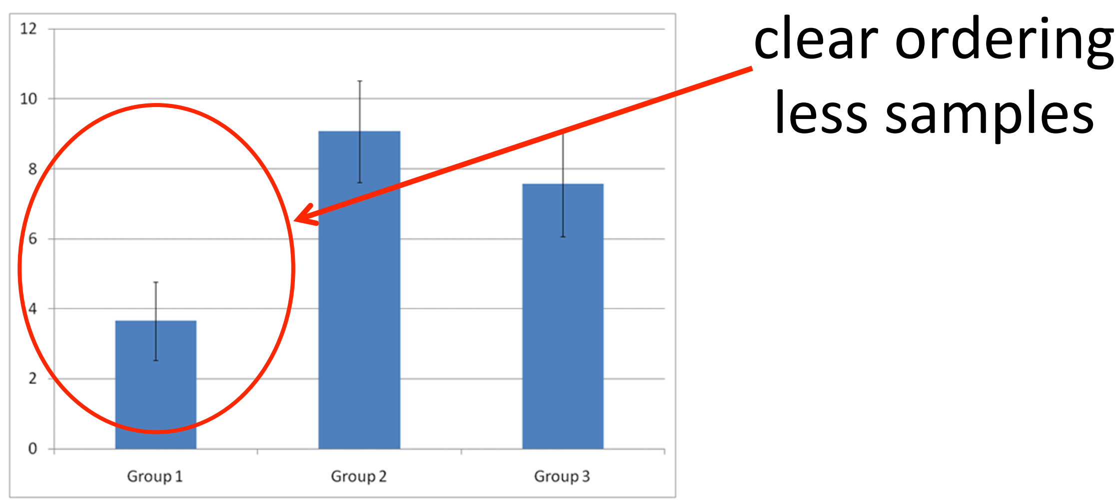
Scalar, Big Data Vis '13

approximate visualizations



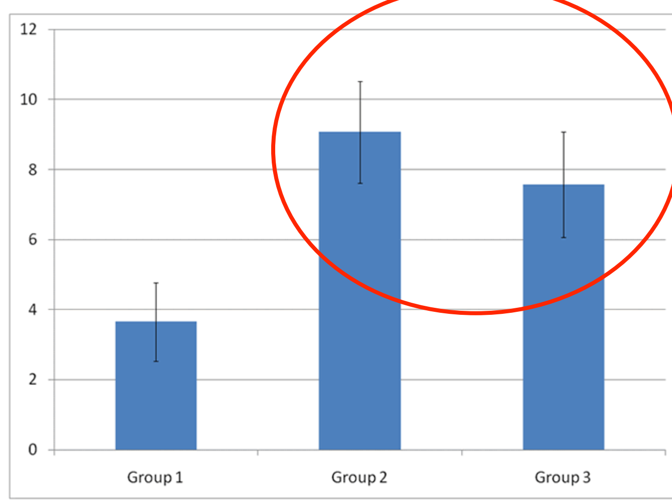
Blais et al, PVLDB ‘15

approximate visualizations



Blais et al, PVLDB ‘15

approximate visualizations

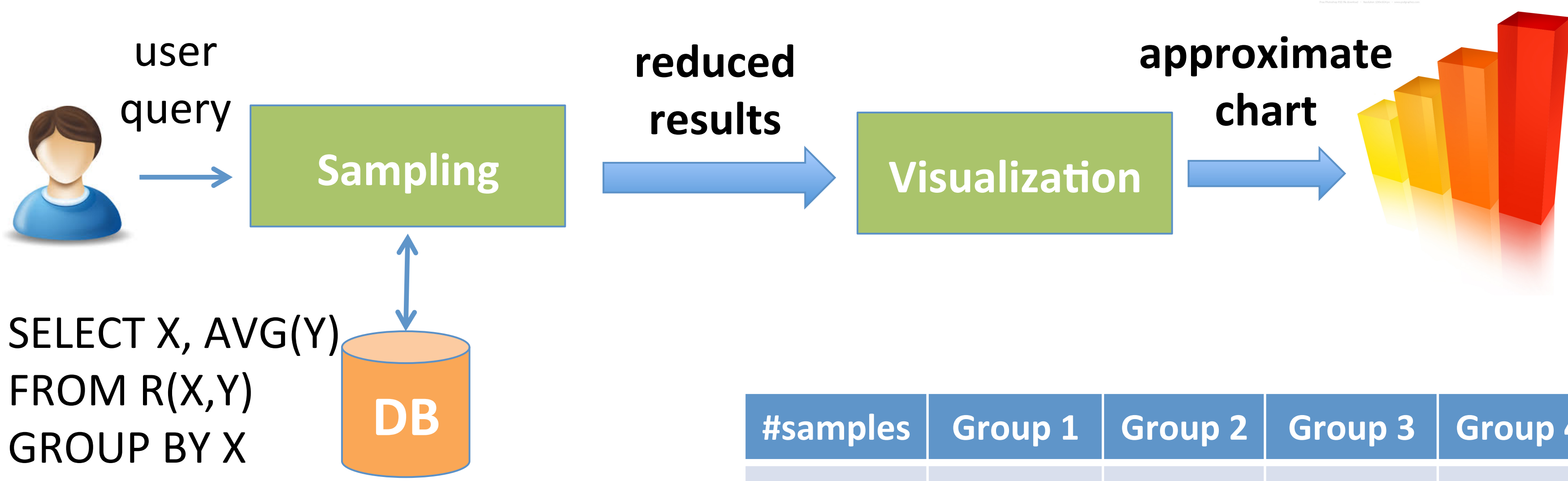


correct order?
sample more

min # samples for correct order?

Blais et al, PVLDB '15

approximate visualizations

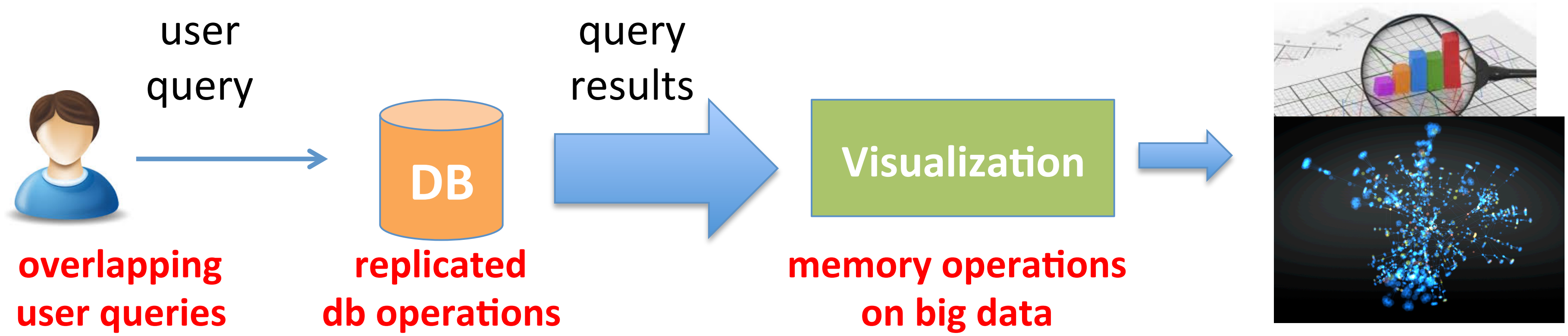


#samples	Group 1	Group 2	Group 3	Group 4
1	**[60,90]**	[20,50]	[10,40]	[40,70]
20	**[64,84]**	[30,48]	[15,35]	[45,65]
21	[66,84], I	[30,48]	[17,35]	[46,64]
70	[66,84], I	[40,47]	[17,32], I	[46,53]

sampling phases/ confidence intervals

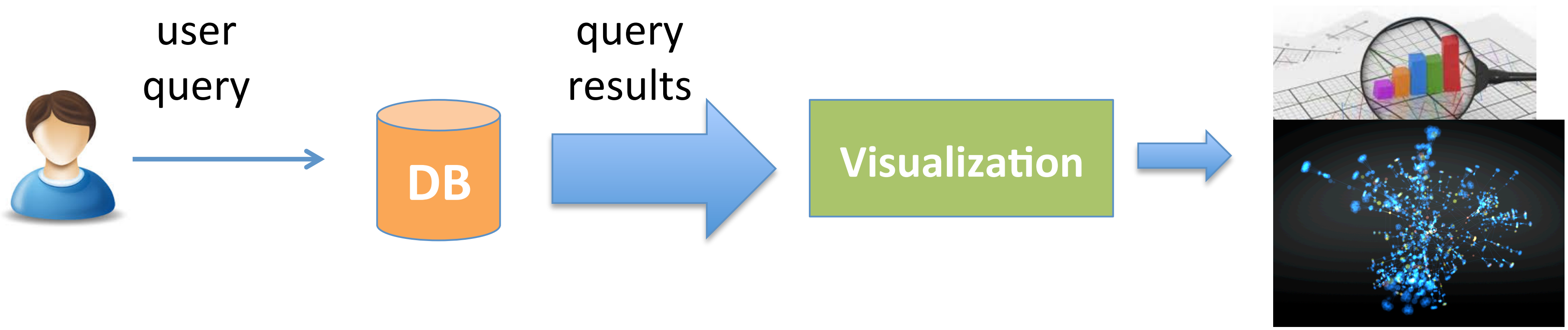
Blais et al, PVLDB '15

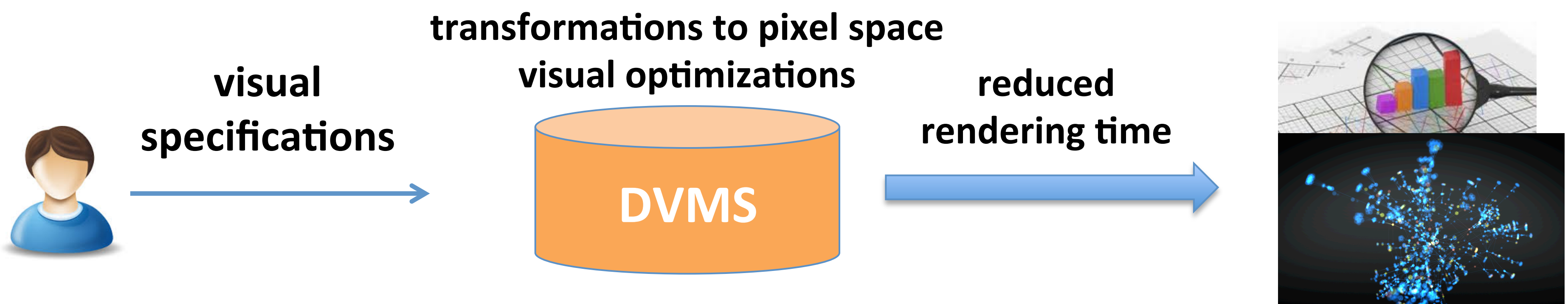
visualization management



Ermac, PVLDB '14

visualization management

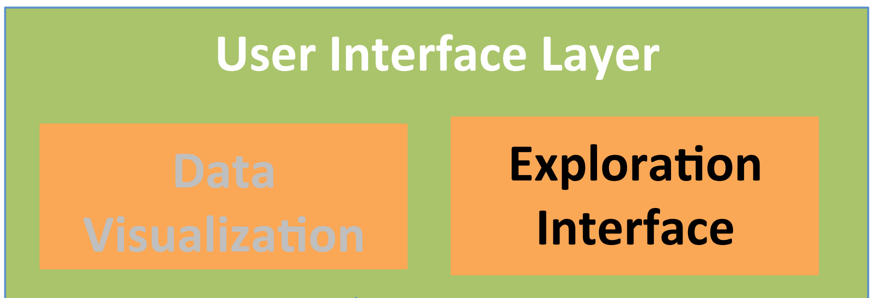




logical visual plans ➔ physical query plans

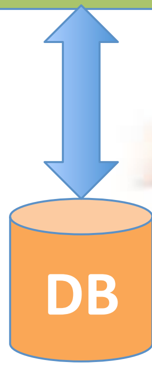
Ermac, PVLDB ‘14

exploration interfaces



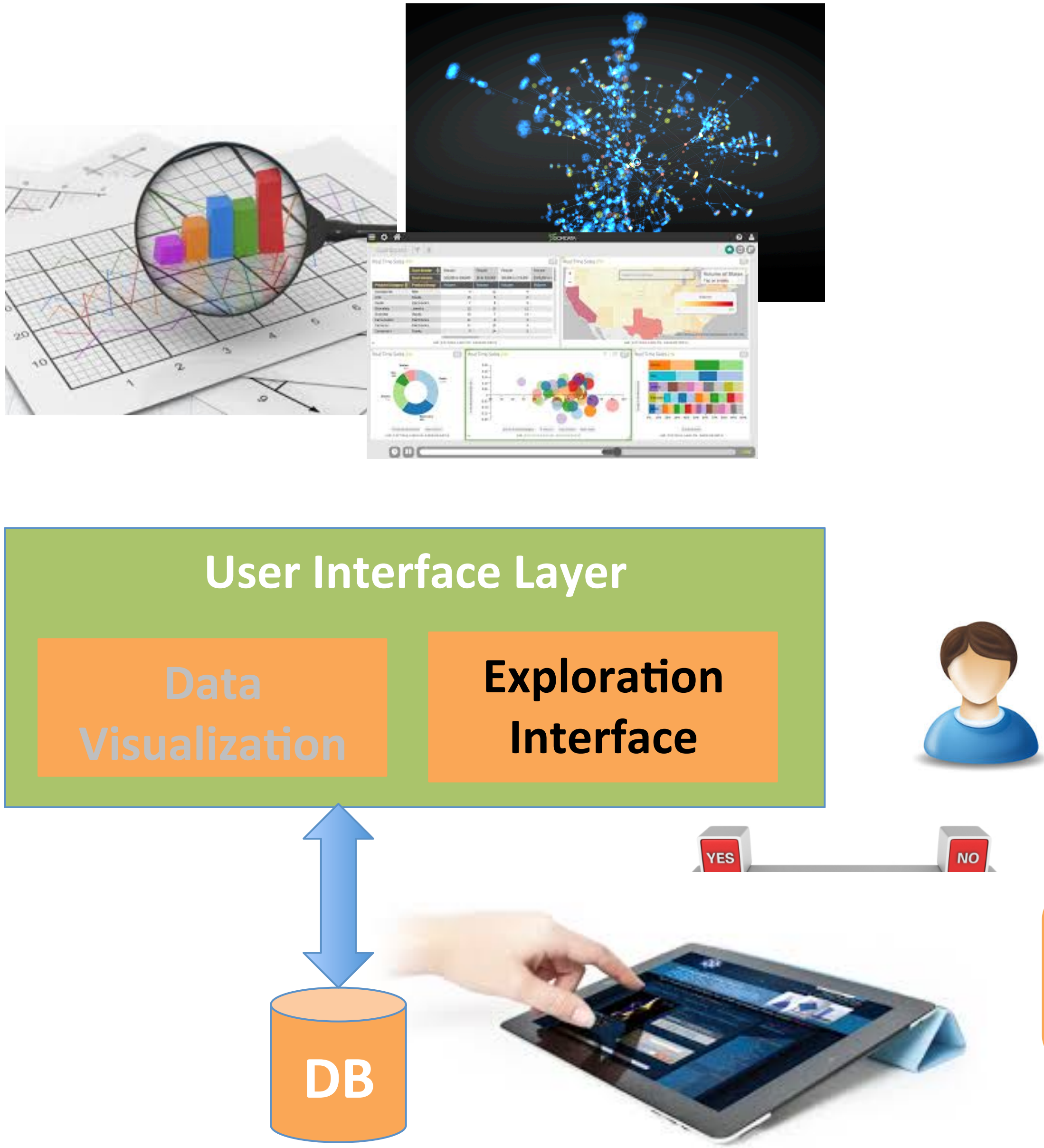
automatic exploration

assisted query formulation



novel query interfaces

exploration interfaces

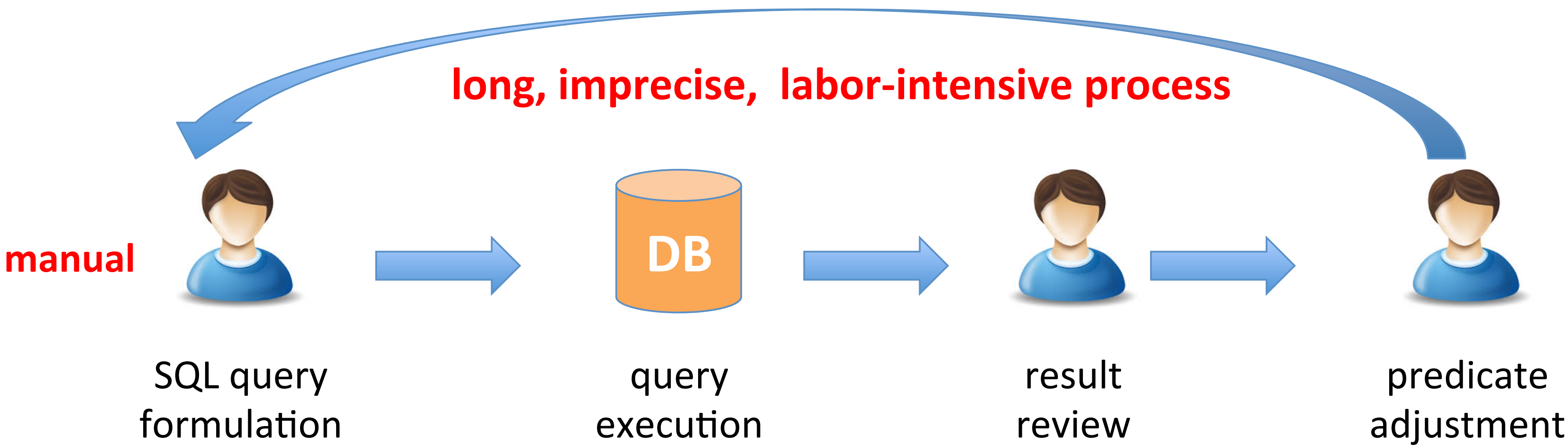


automatic exploration

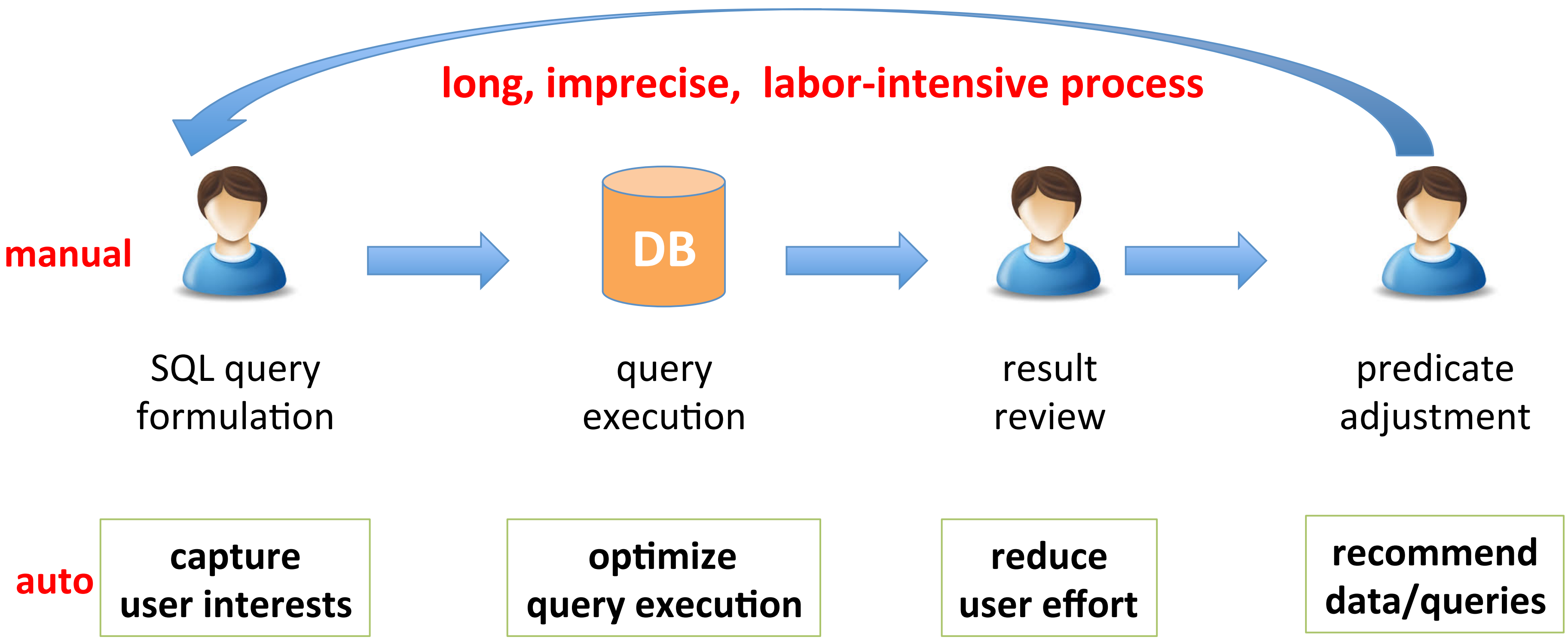
assisted query formulation

novel query interfaces

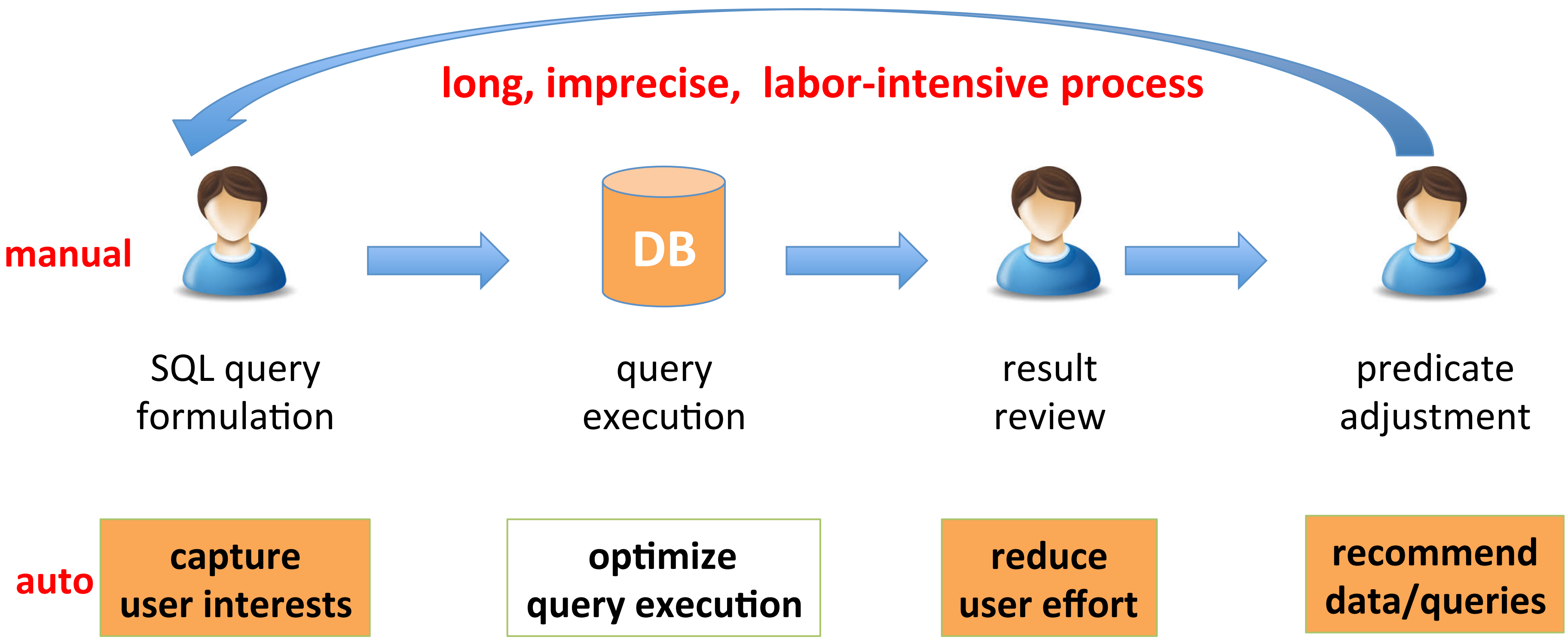
manual vs automatic data exploration



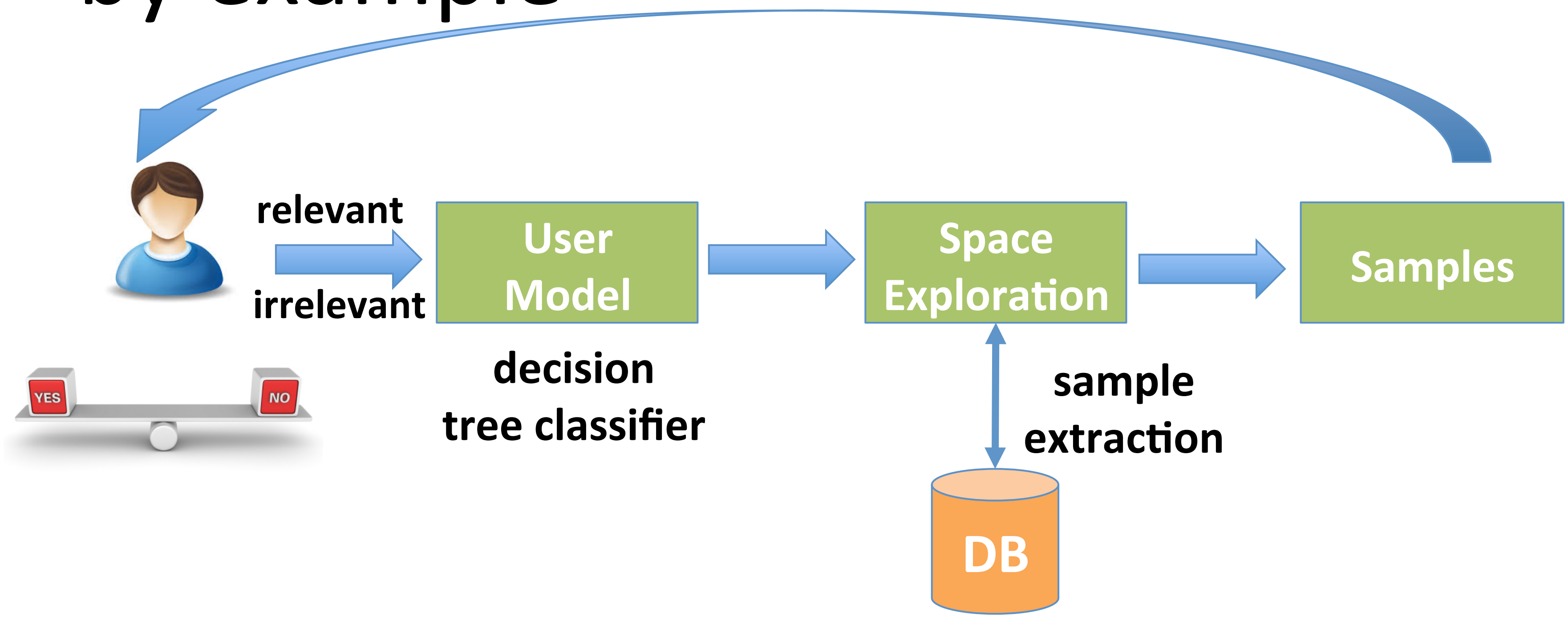
manual vs automatic data exploration



manual vs automatic data exploration



explore by example

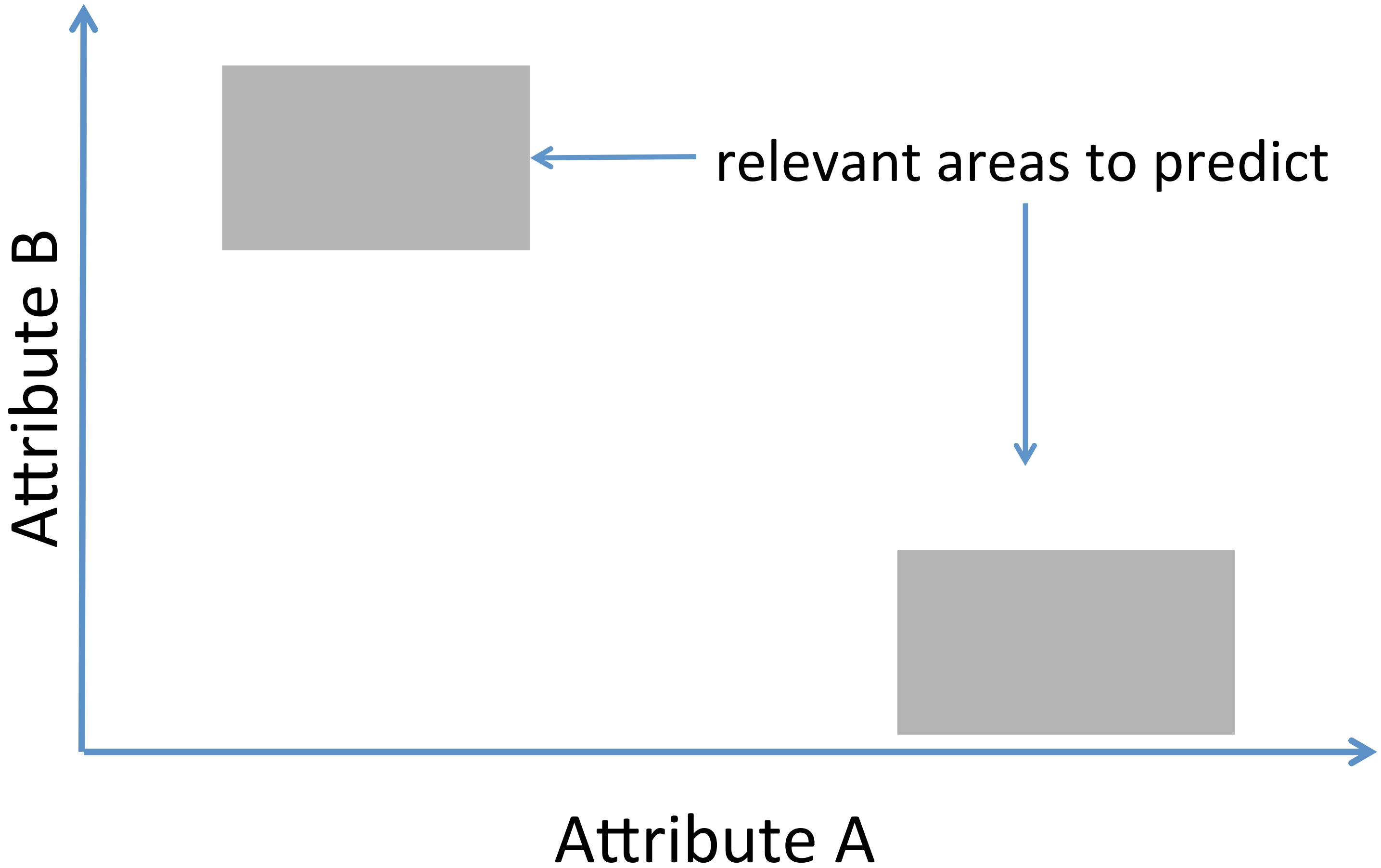


effectiveness vs efficiency

sampling areas? sampling size?

AIDE, SIGMOD'14/ VLDB'15

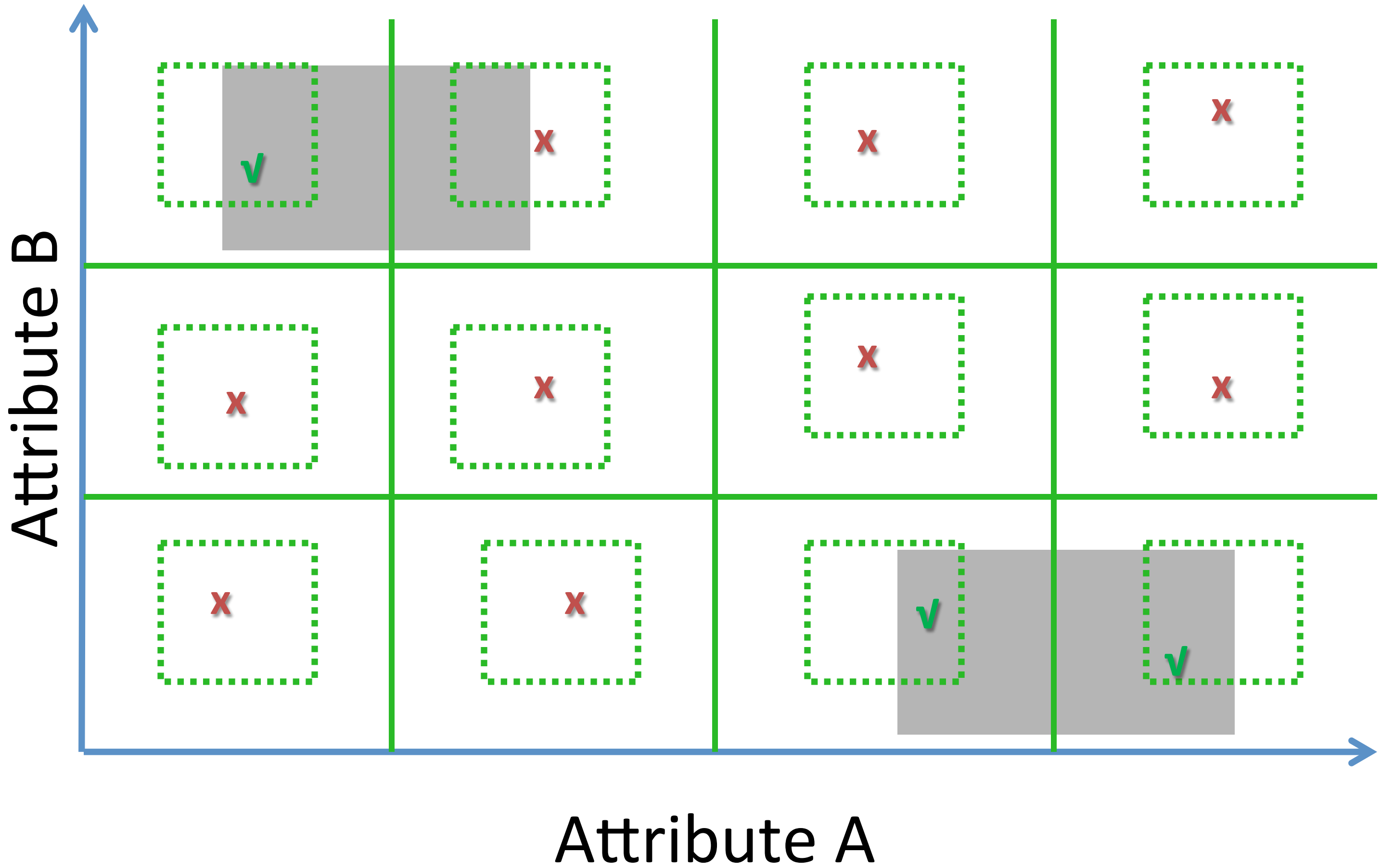
explore by example



AIDE, SIGMOD'14/ VLDB'15

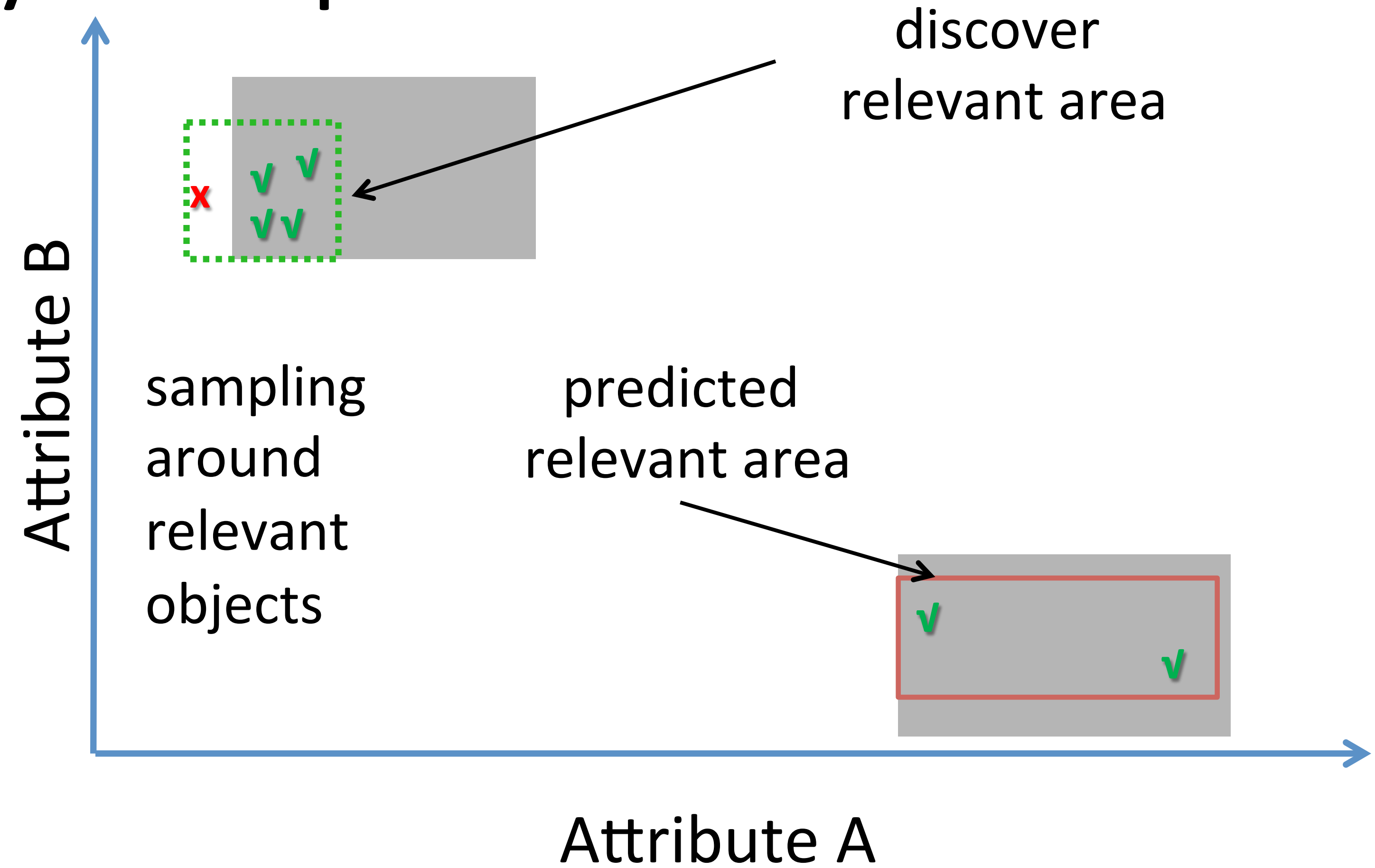
explore by example

uniform sampling across domain



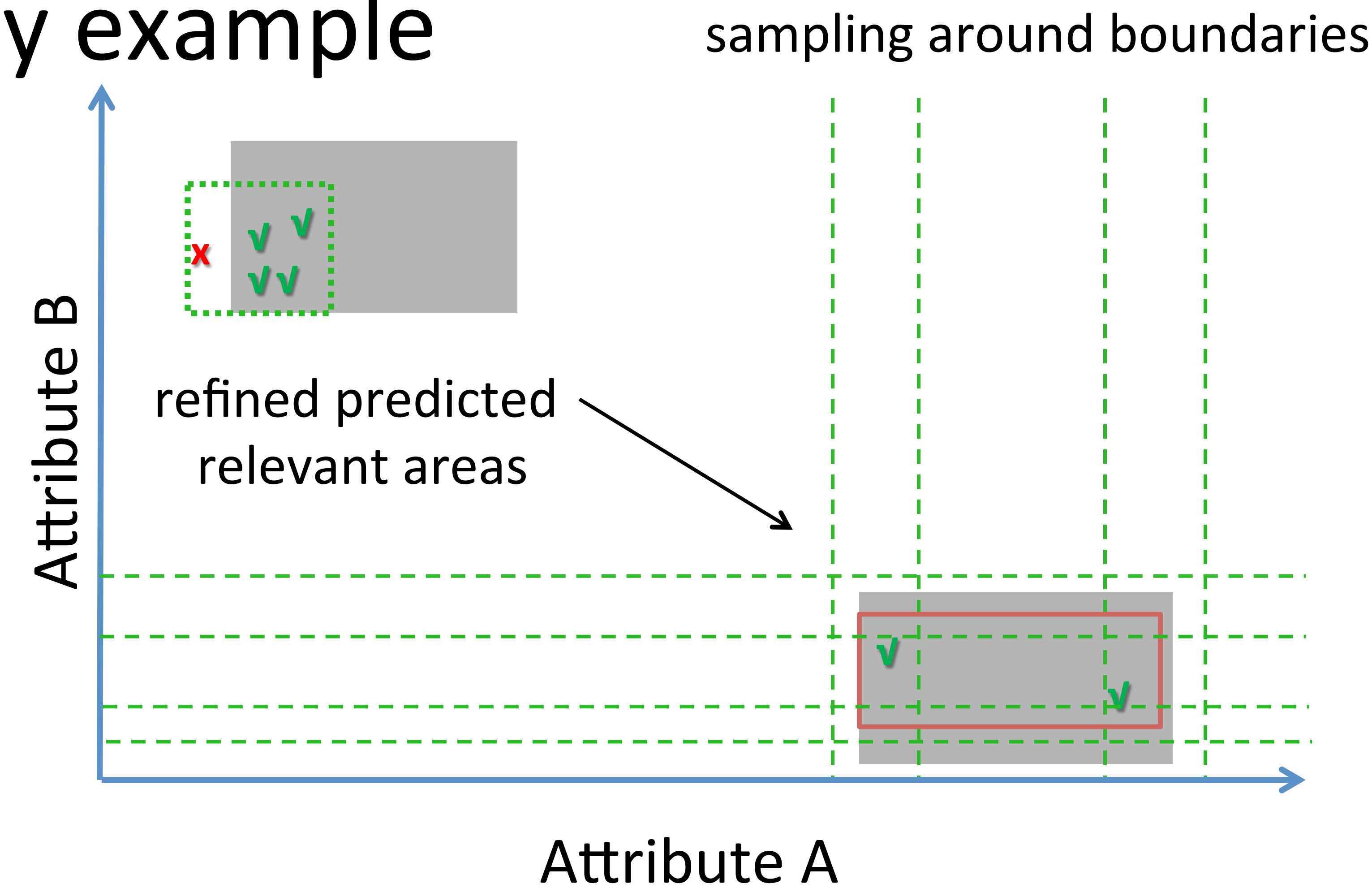
AIDE, SIGMOD'14/ VLDB'15

explore by example



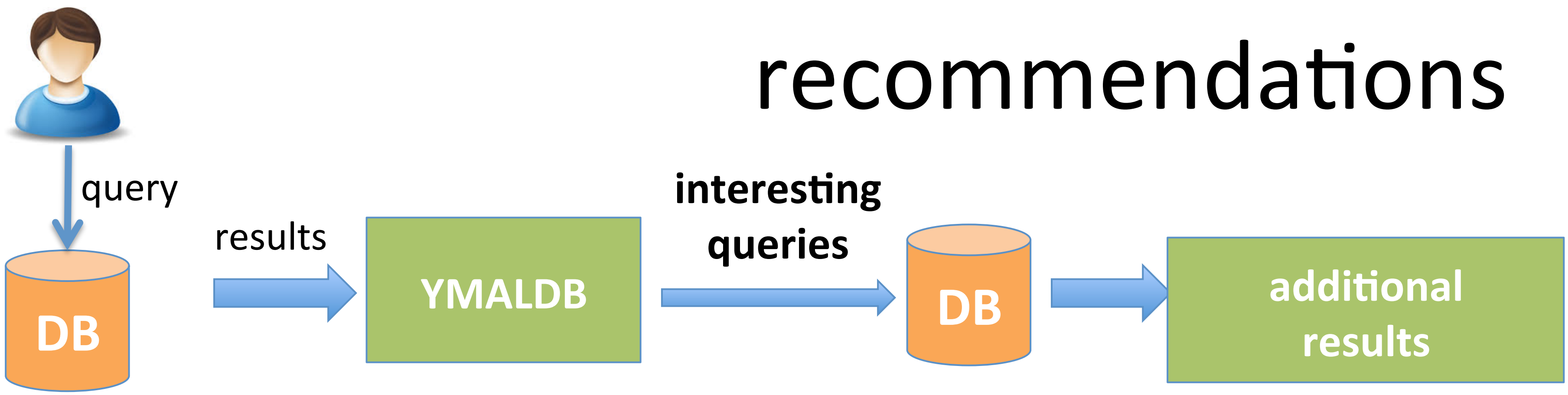
AIDE, SIGMOD'14/ VLDB'15

explore by example



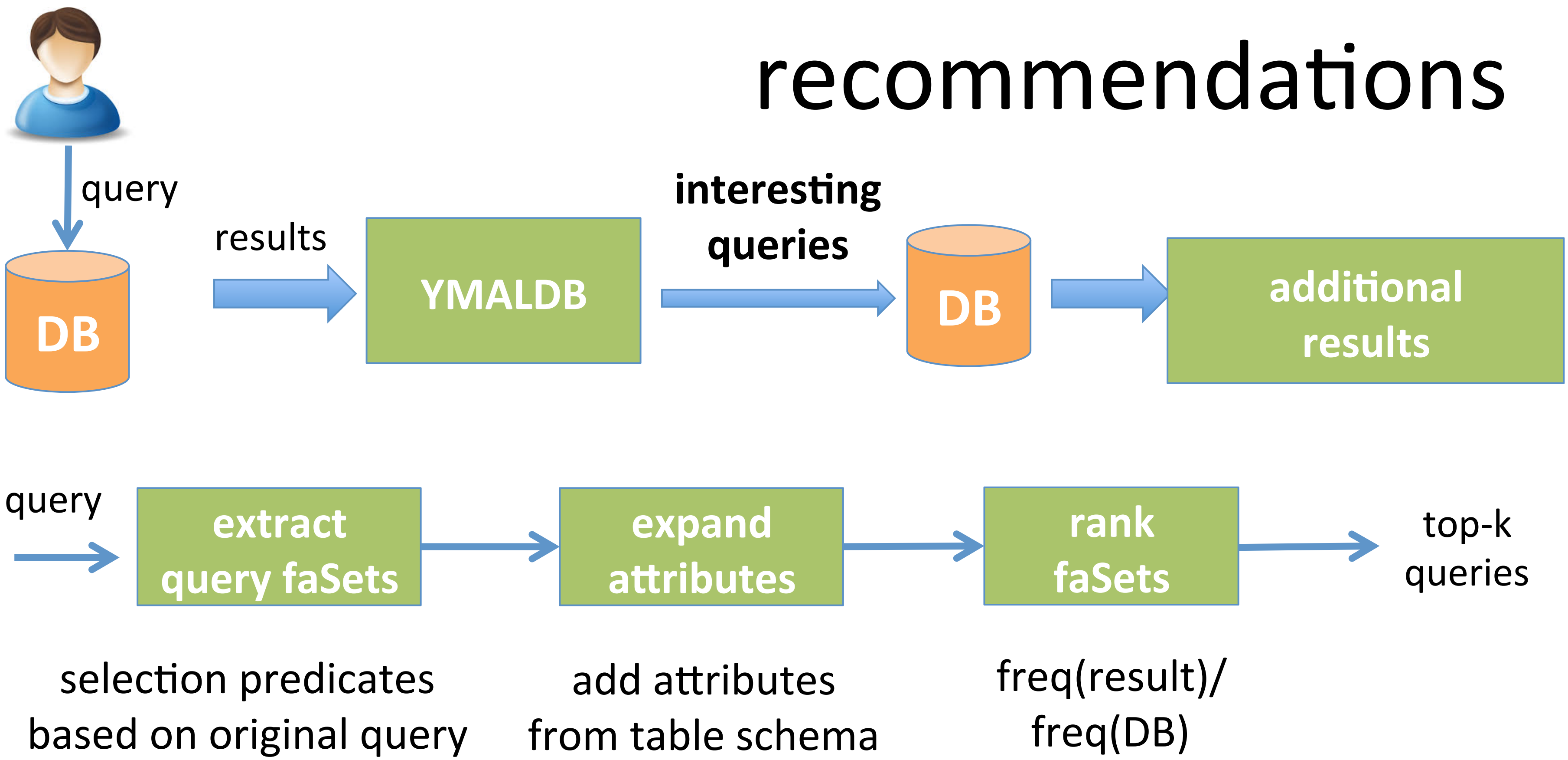
AIDE, SIGMOD'14/ VLDB'15

result recommendations



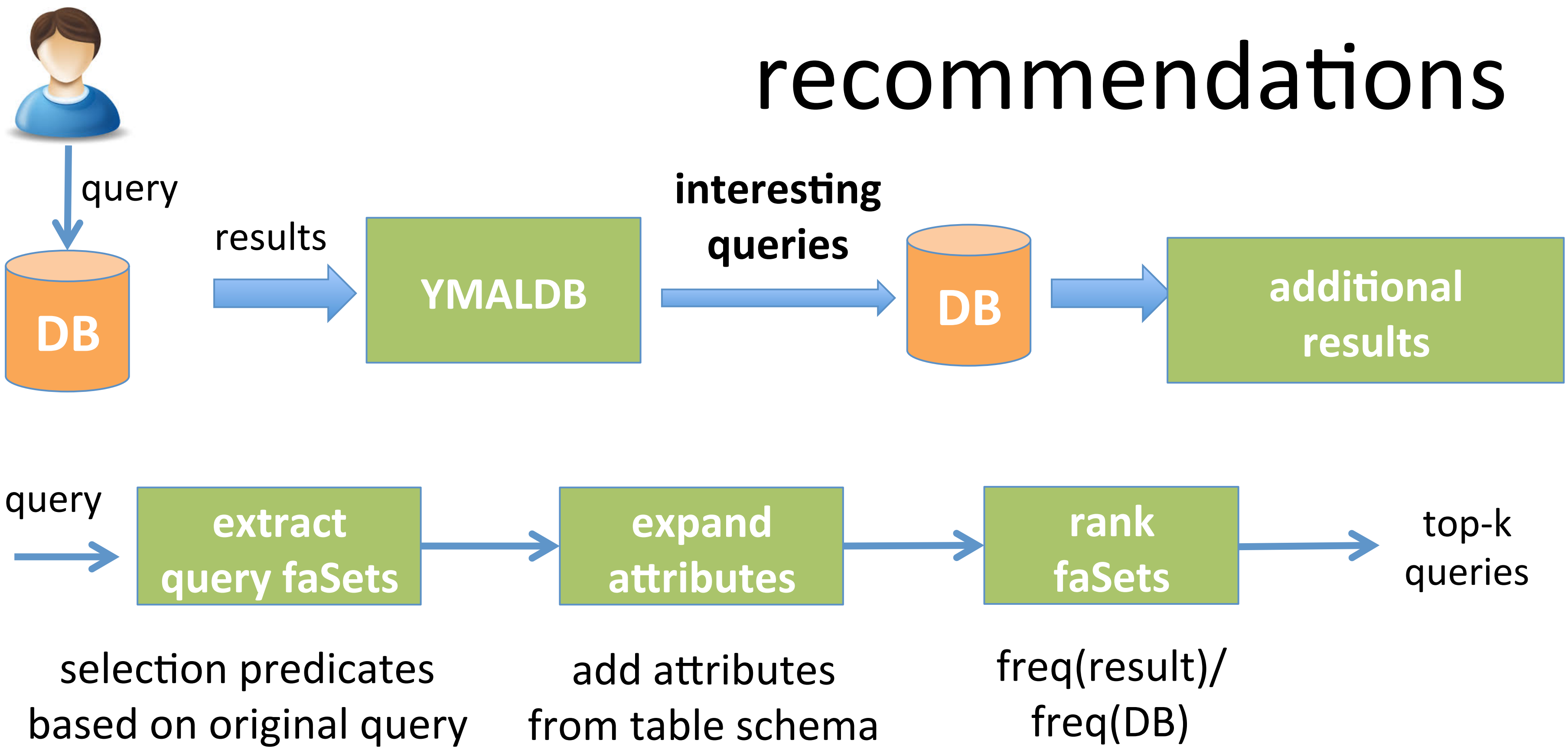
YMALDB, VLDBJ '13

result recommendations



YMALDB, VLDBJ ’13

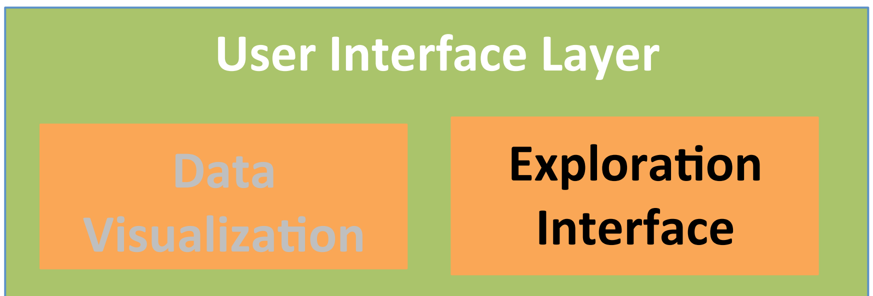
result recommendations

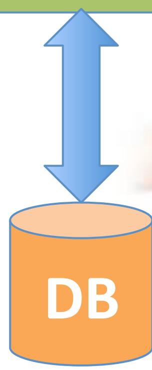


title, year, genre of Scorsese movies + *title, year, genre, **country** of Scorsese movies* = *many Scorsese movies are related to Italy*

YMALDB, VLDBJ '13

exploration interfaces



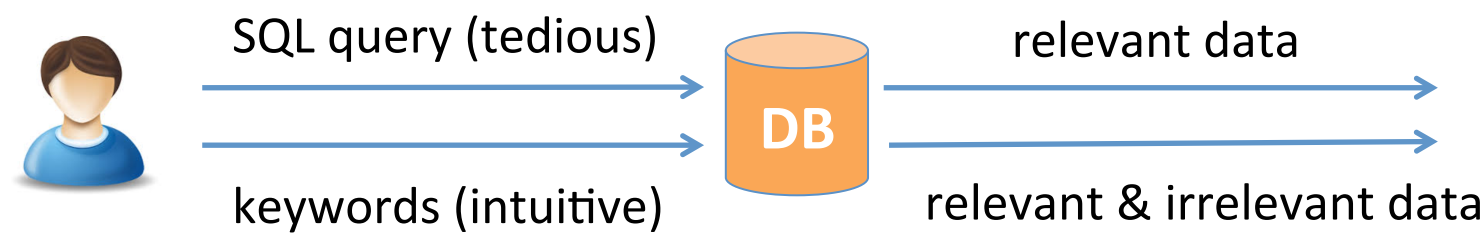


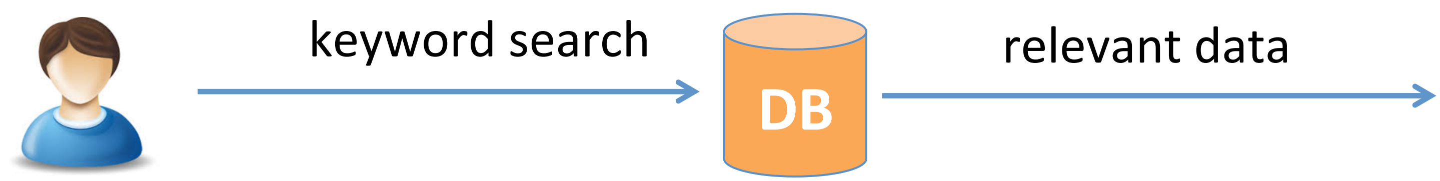
automatic exploration

assisted query formulation

novel query interfaces

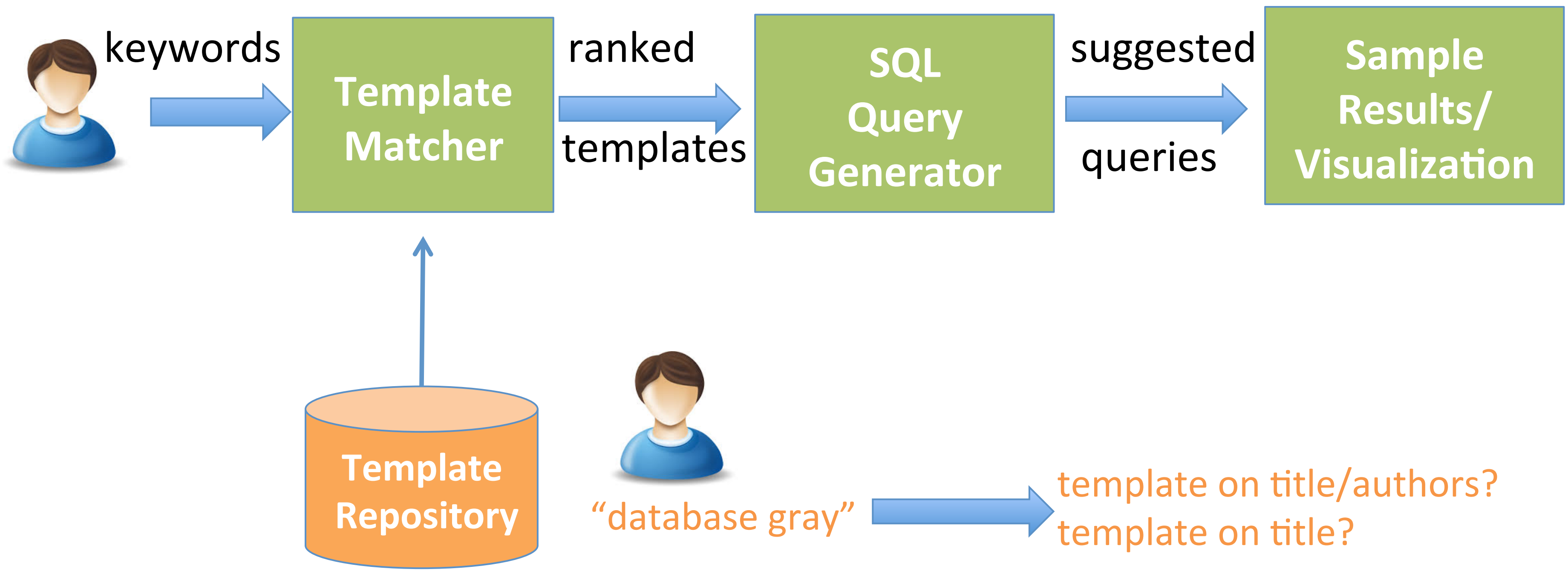
keyword-based query suggestions





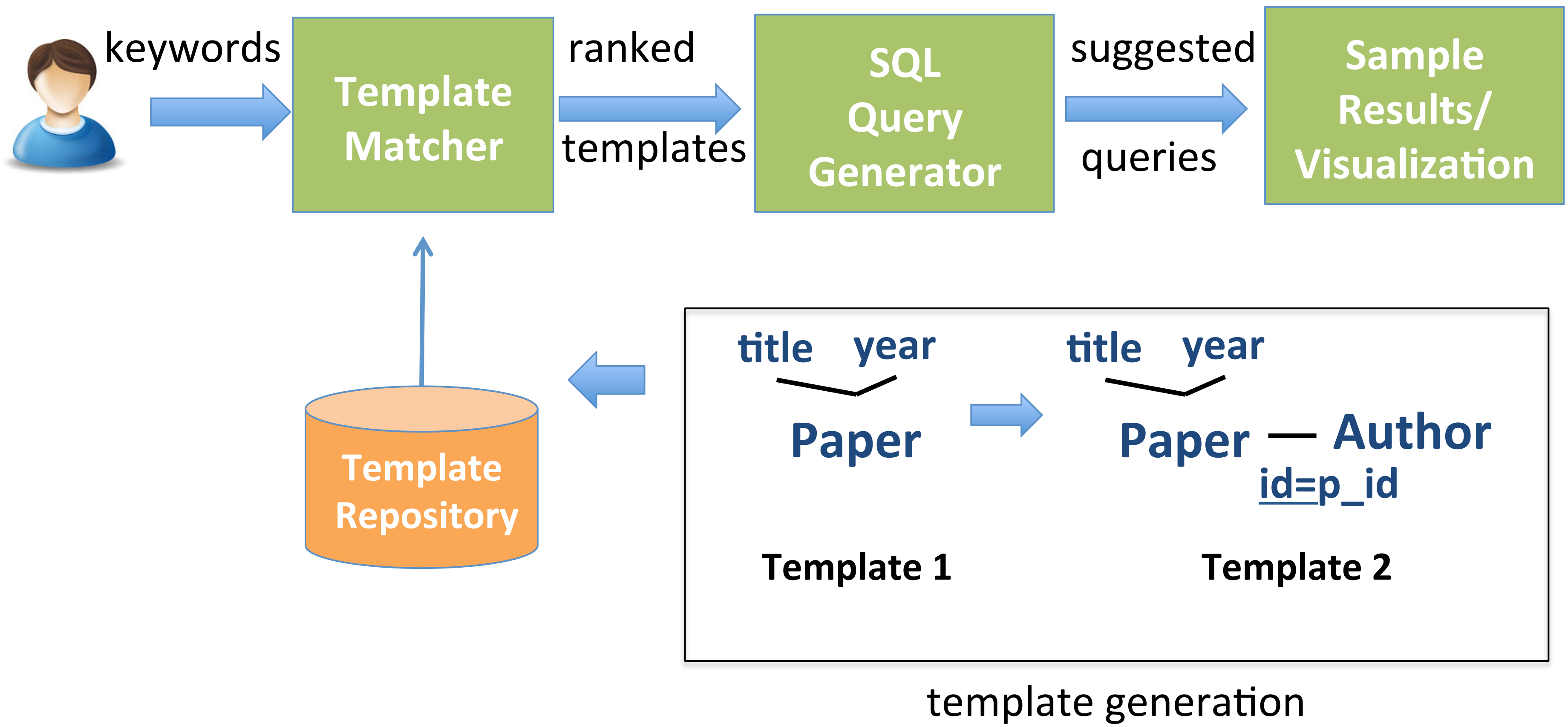
how we can discover relevant queries?

keyword-based query suggestions



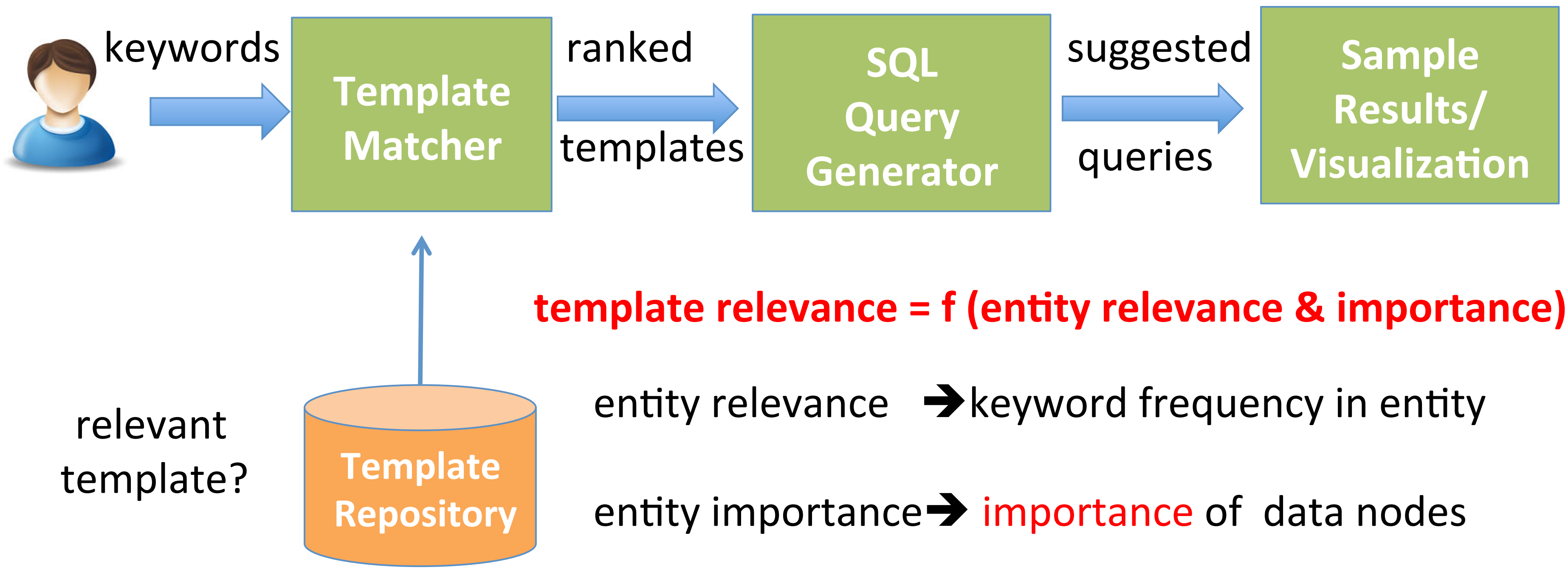
SQLSUGG, ICDE’11

keyword-based query suggestions



SQLSUGG, ICDE'11

keyword-based query suggestions



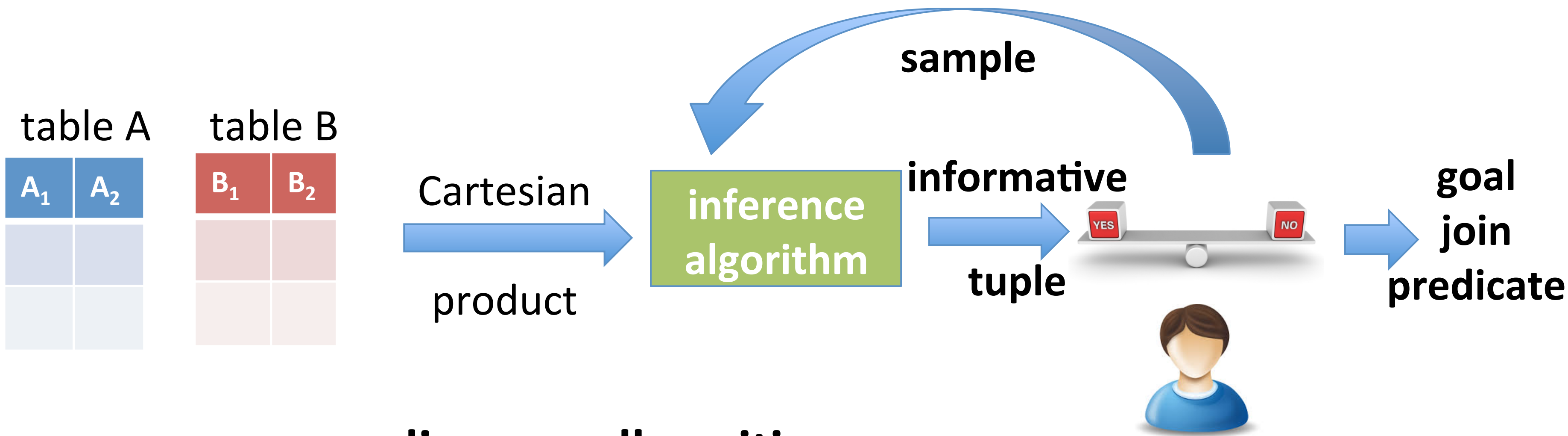
template relevance = f (entity relevance & importance)

entity relevance → keyword frequency in entity

entity importance → importance of data nodes

SQLSUGG, ICDE'11

equi-join inference

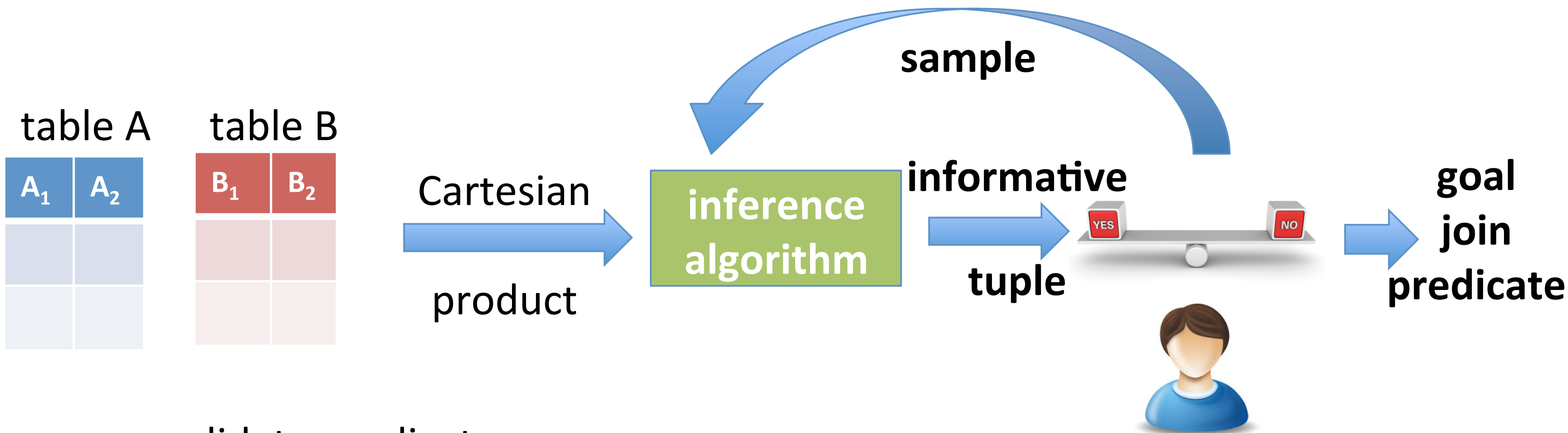


goal predicate: **discover all positives**
eliminate all negatives

minimize user effort

Bonifati et al, EDBT'14

equi-join inference



candidate predicates

(A1, B1) (A1, B2)	**(A1, B1)** **(A2, B1)**	(A1, B1) (A2, B2)
(A1, B1)	**(A1, B2)**	(A2, B1)

prune predicates with uninformative tuples

label tuple that prunes as many predicates as possible

Bonifati et al, EDBT'14

graphical query specification

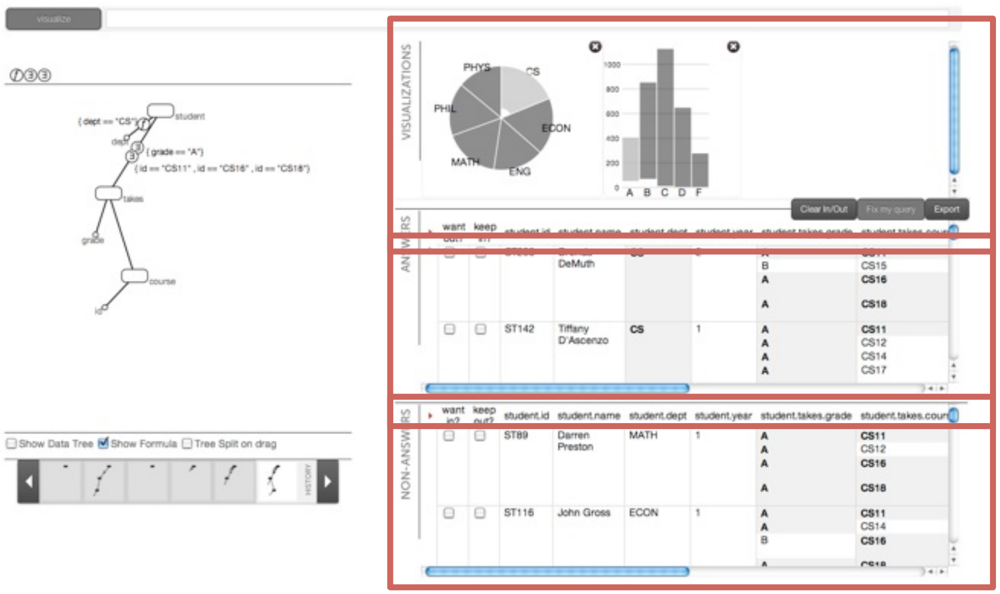
result visualization

answers

non-answers

DataPlay, PVLDB ’13

graphical query specification



result visualization

answers

non-answers

pivot relation → add, remove query constraints ⇄ query /visualization recommendations

semantic query tuning by local syntactic modifications

DataPlay, PVLDB '13

graphical query specification

result visualization

answers

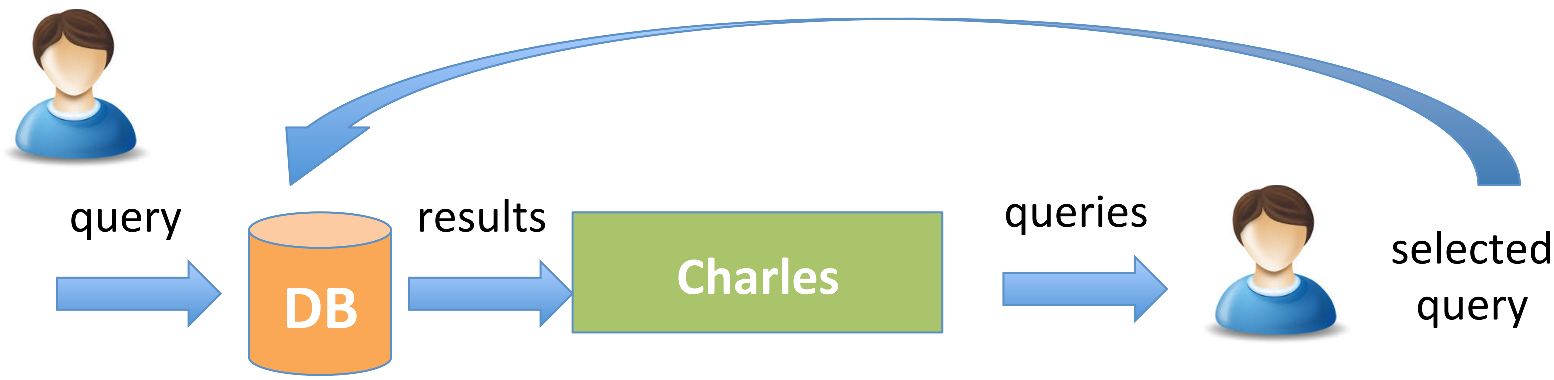
non-answers

pivot relation → add, remove results ⇄ query corrections

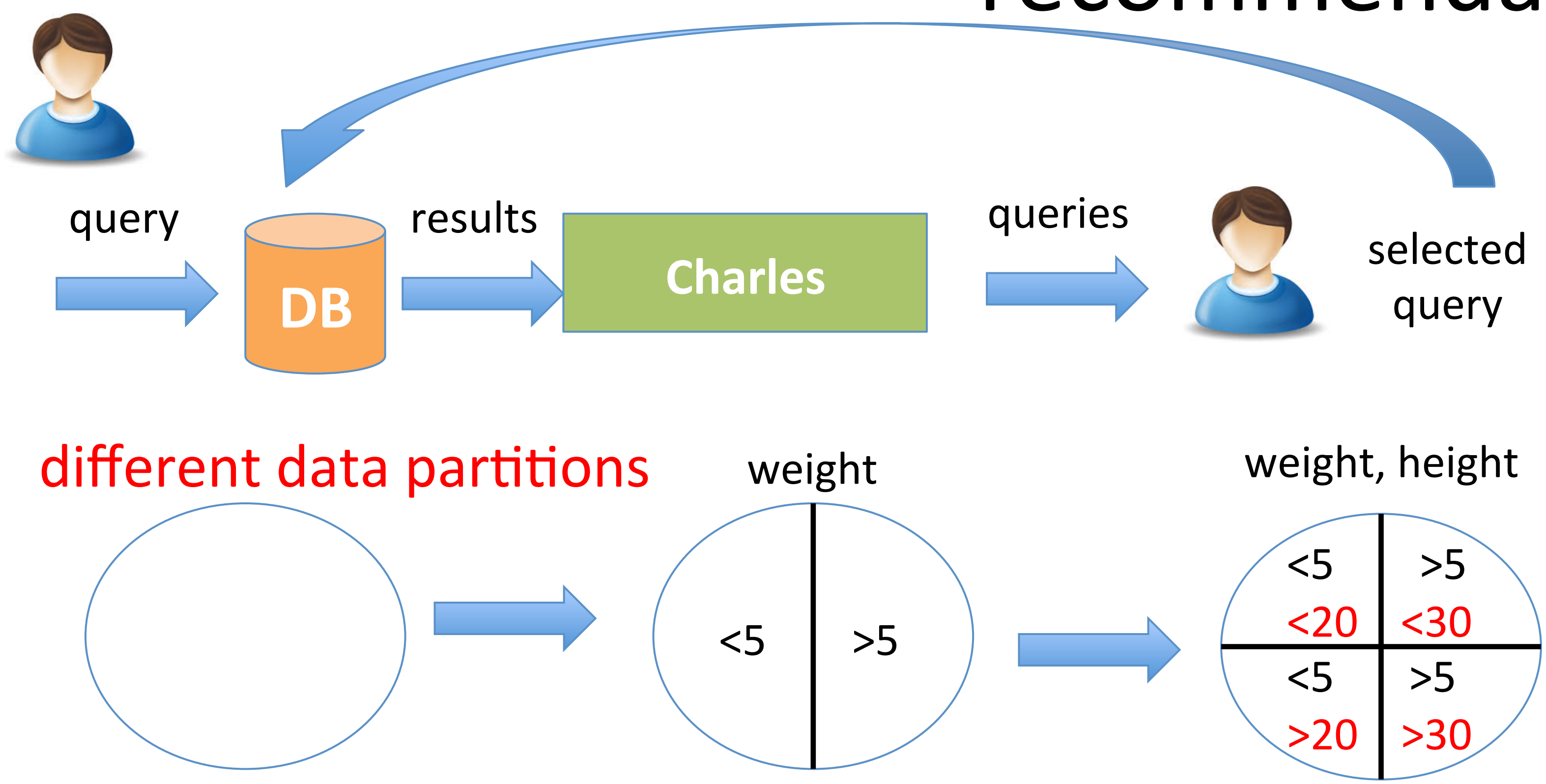
search limited to local modifications

DataPlay, PVLDB '13

query recommendations



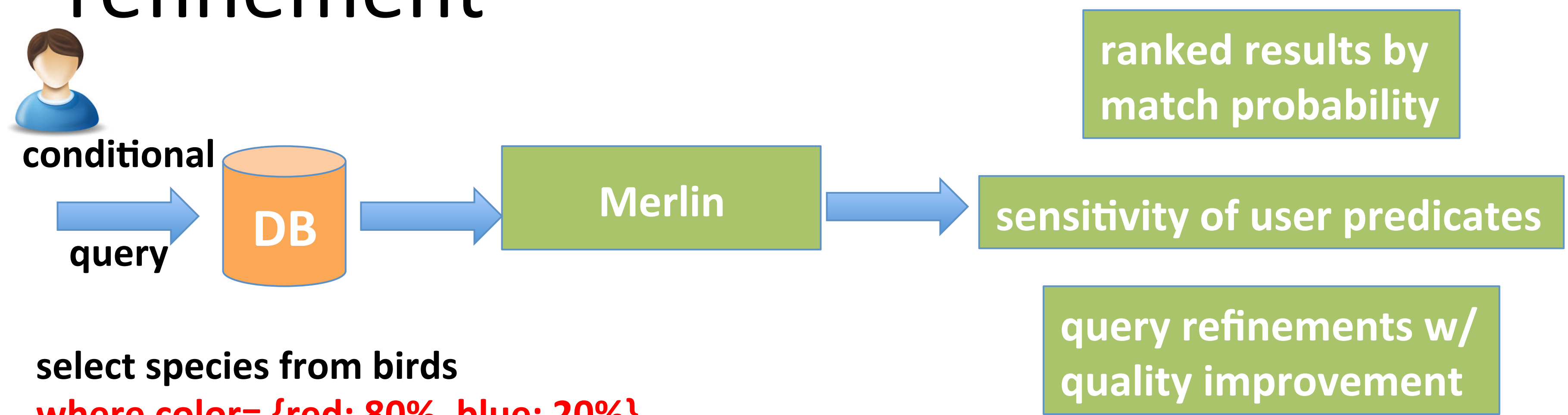
query recommendations



quality: simplicity, breadth, balance

Charles, CIDR '13

query refinement



select species from birds
where color= {red: 80%, blue: 20%}

rank	species
1	Bluebird
2	Blue Jay

attr	sensitivity
color	18.6

impact on ranking

attr	quality score
size	83.3
legcolor	57.1

remaining attributes

result quality if added in the query

Merlin, ICDE '14

exploration interfaces

User Interface Layer

Data Visualization

Exploration Interface

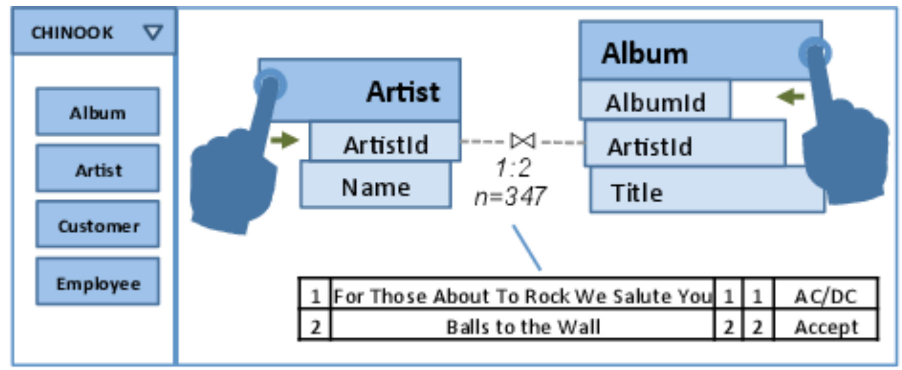
DB

automatic exploration

assisted query formulation

novel query interfaces

no-keyboard interfaces



query context + gesture recognition = query intend

query space + search pattern = query template

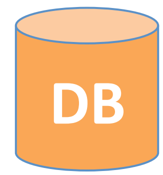
GestureDB, CIDR '13

no-keyboard interfaces

novel database kernel

touch input

quick response



touch recognition → gesture recognition → map touch to operators

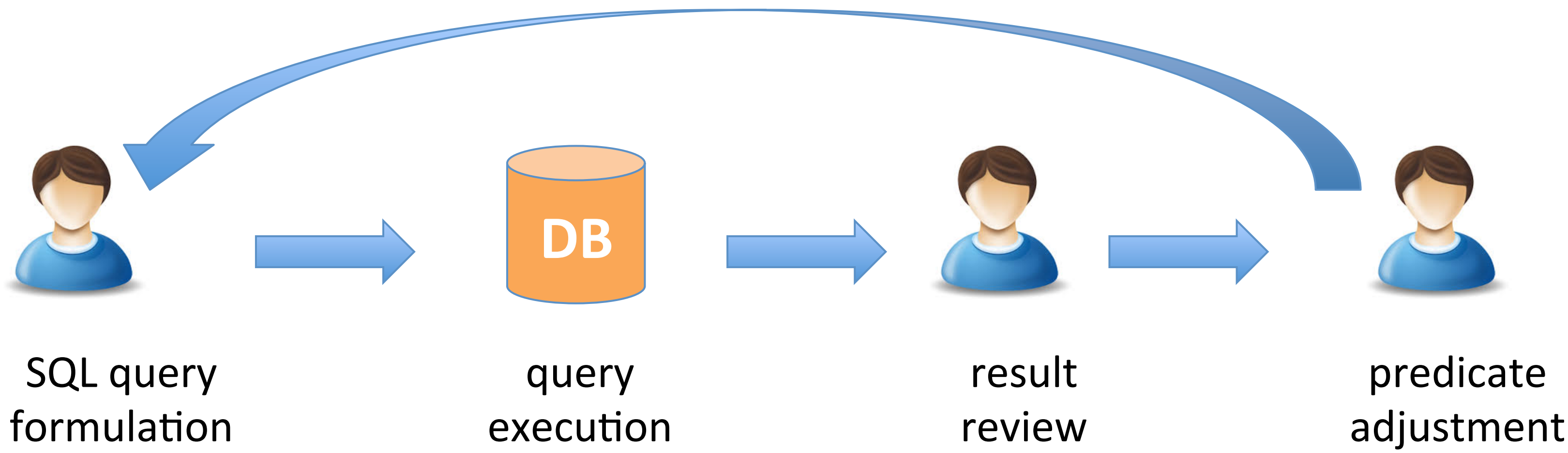
dbTouch, CIDR ’13

Interactive Exploration through

Data Prefetching & Query Approximation

MIDDLEWARE TECHNIQUES

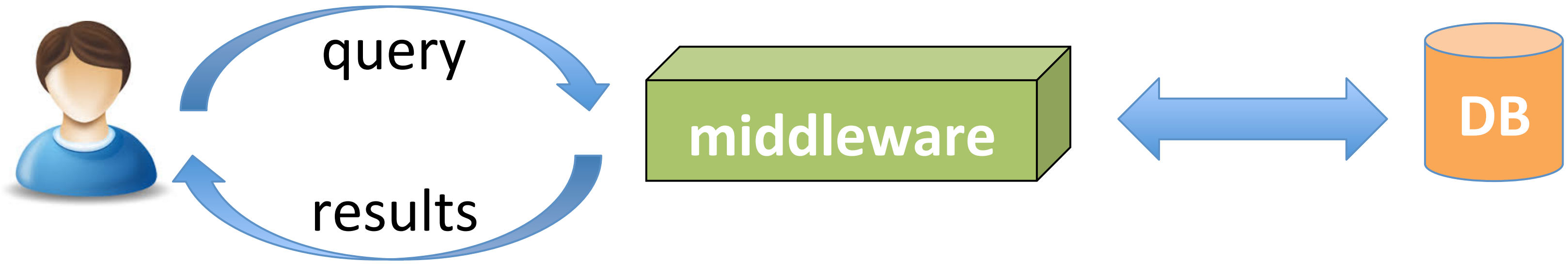
interactive data exploration



ad-hoc, non-optimized, labor-intensive process

interactive: small latency bounds on user wait time

middleware optimizations



prefetching

speculative query execution

result reuse

structure-aware prefetching

query approximation

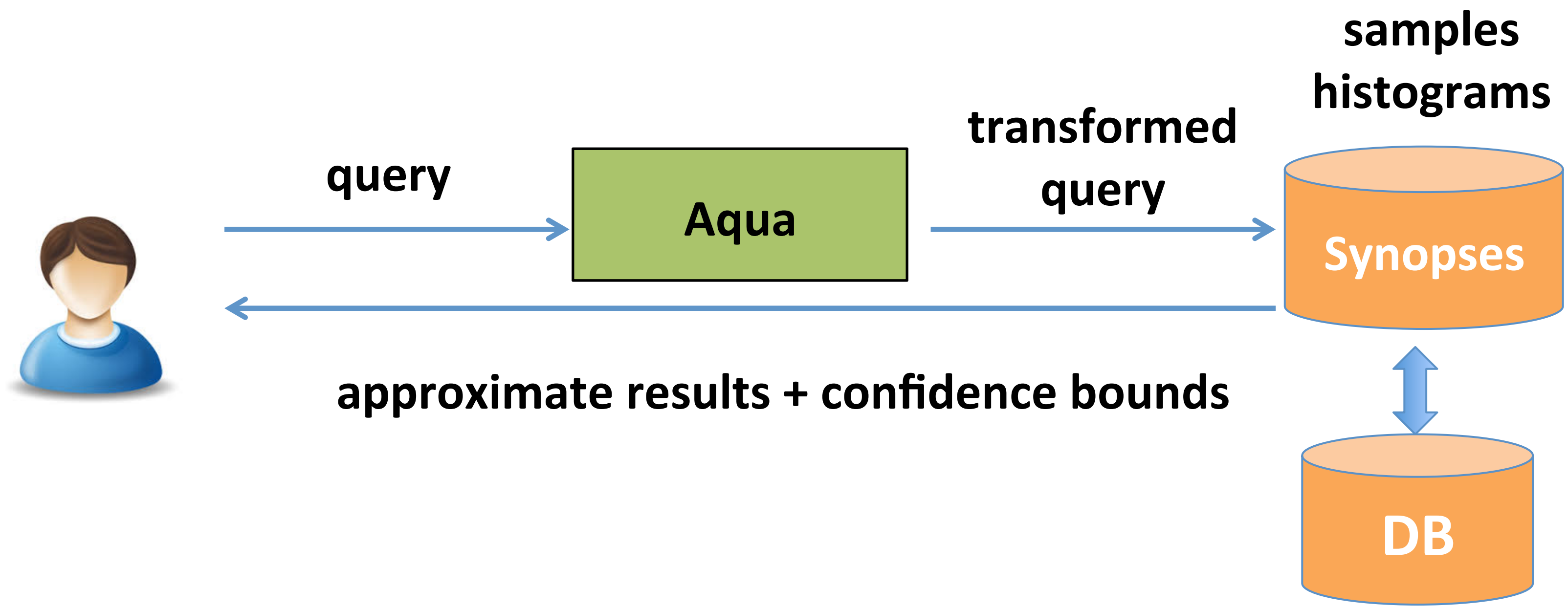
online processing

sample-based processing

sample-based processing

- accuracy vs response times
- sample construction & selection
- error approximation

off-line data synopses



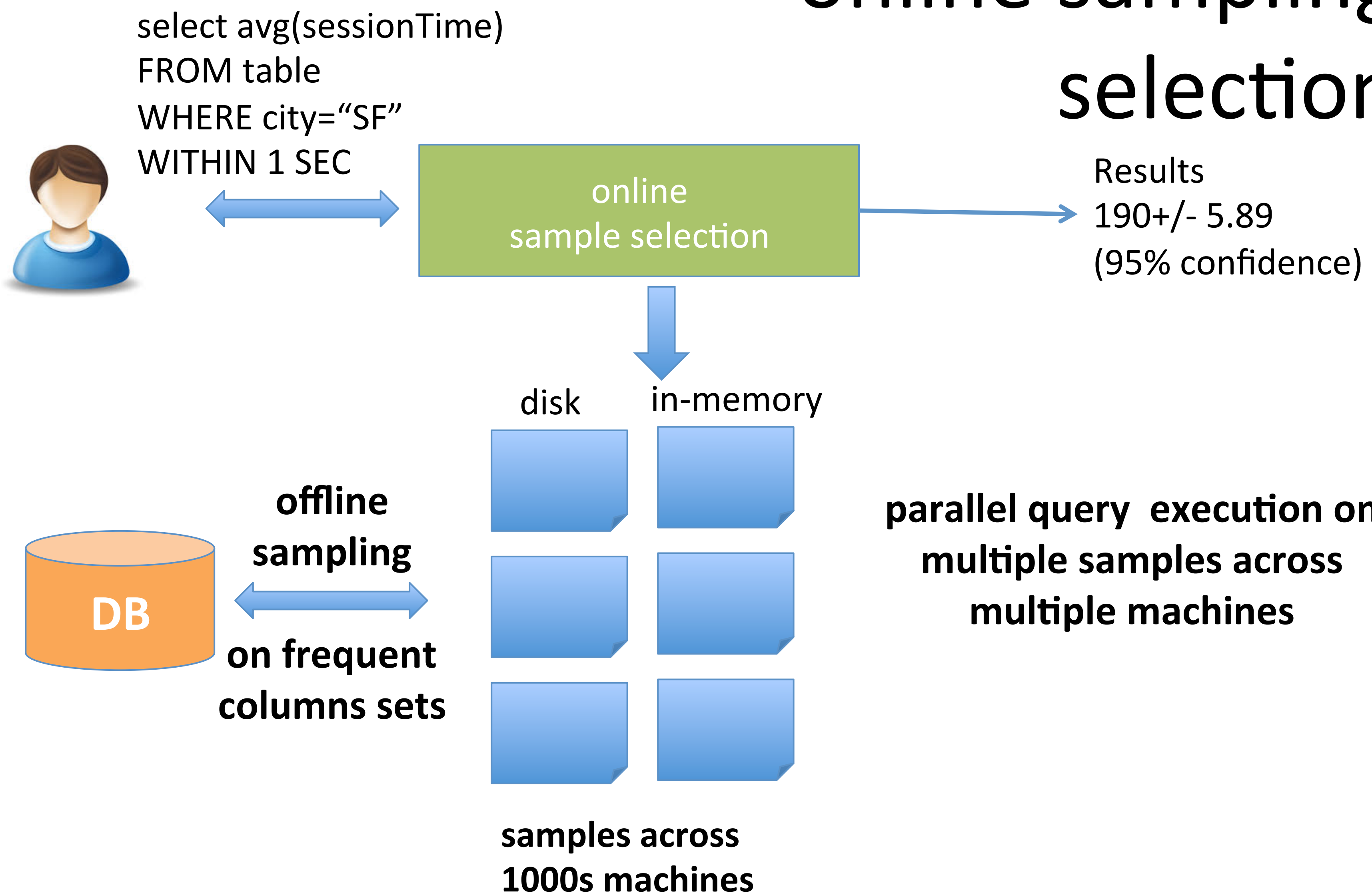
join synopses: sample distinguished joins

congressional samples: biased sampling for group-by queries

incremental maintenance: equi-depth & compressed histograms

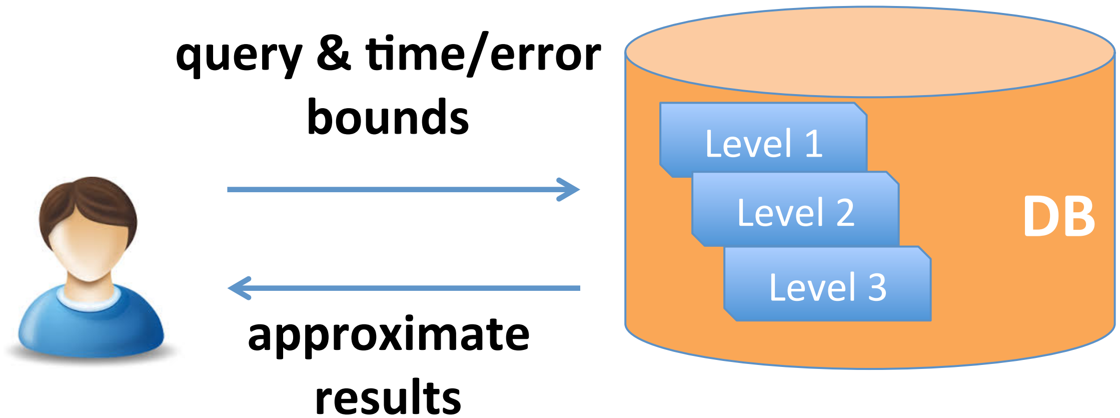
Aqua, SIGMOD ’99

online sampling selection



BlinkDB, EuroSys '13

data impressions

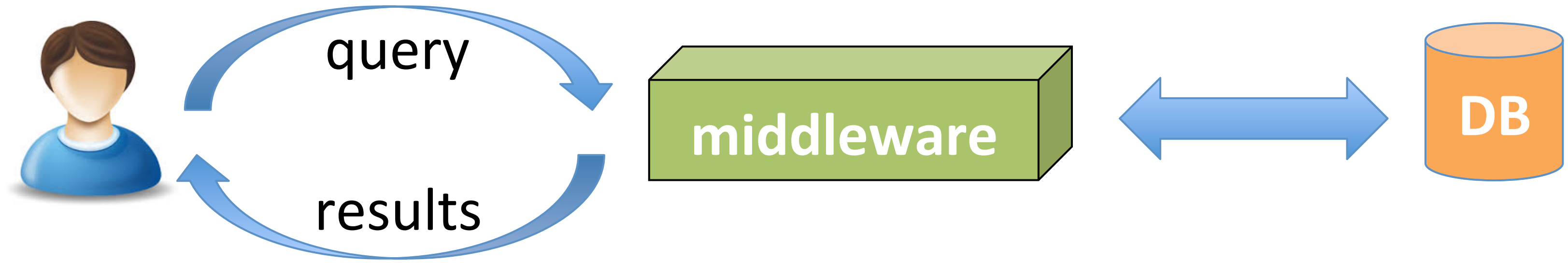


impressions during data loading

adaptive sampling to exploration focus + multi layer sampling and processing to meet user bounds

SciBORG, CIDR ’11

middleware optimizations



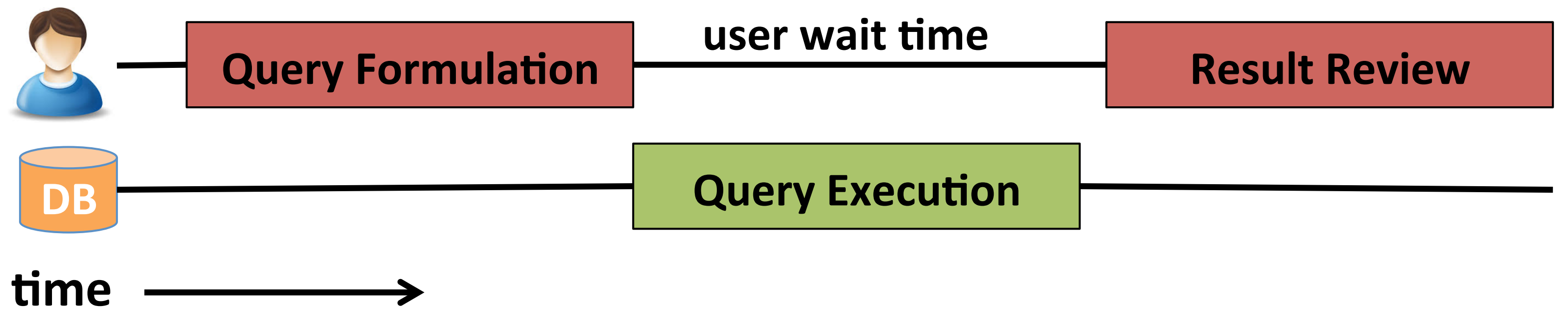
query approximation

- **online processing**
- **sample-based processing**

prefetching

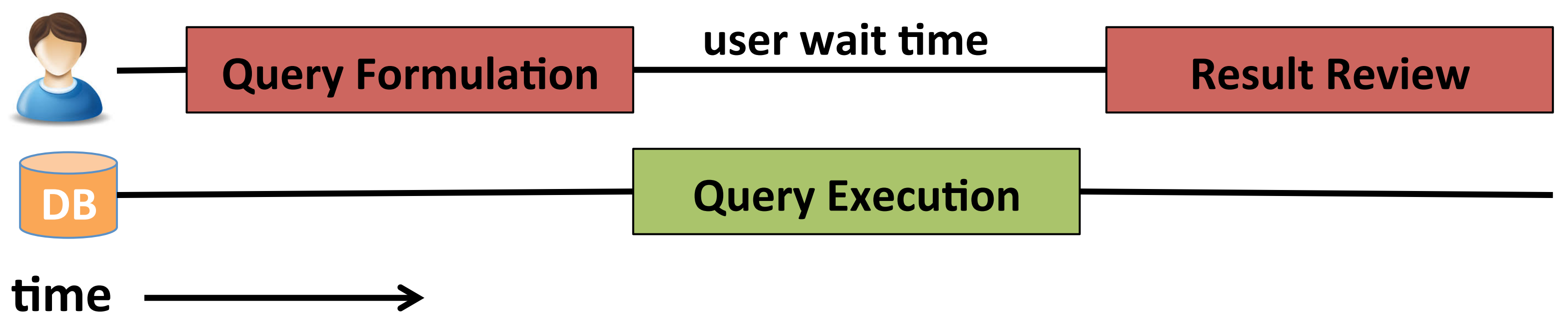
- **speculative query execution**
- **result reuse**
- **structure-aware prefetching**

speculative query execution



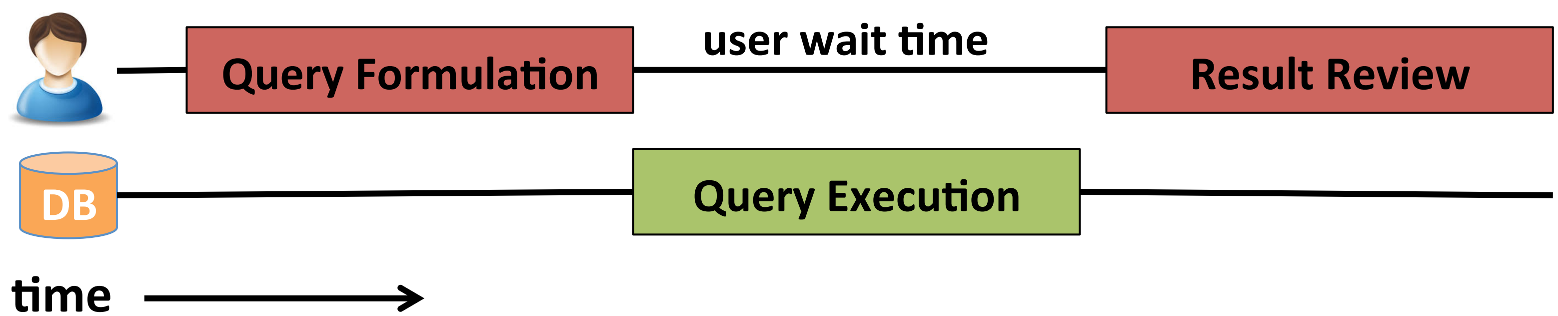
1. predict follow-up queries
2. execute queries
3. cache results

speculative query execution

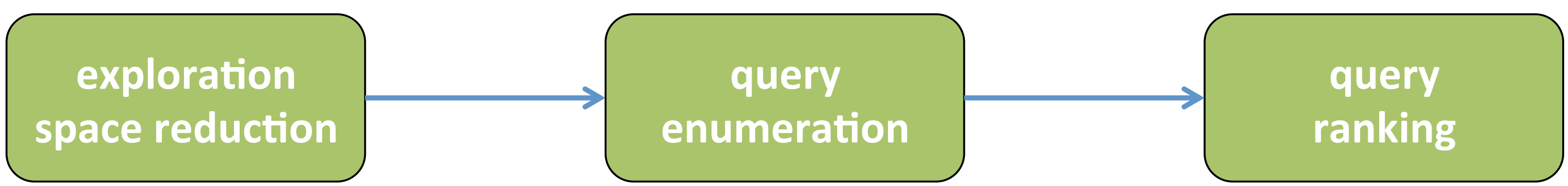


1. predict follow-up queries
2. execute queries
3. cache results

speculative query execution



1. predict follow-up queries
2. execute queries
3. cache results

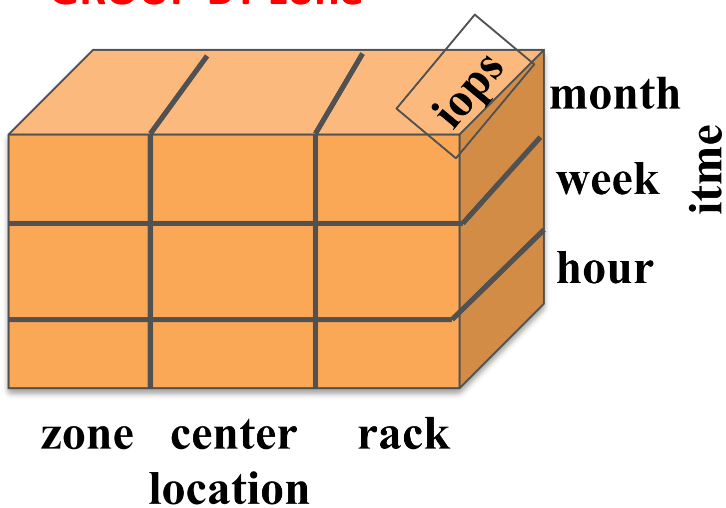


cube exploration

exploration space reduction

user query

SELECT AVG (iops) FROM events
WHERE month="m1" AND week="w1"
GROUP BY zone

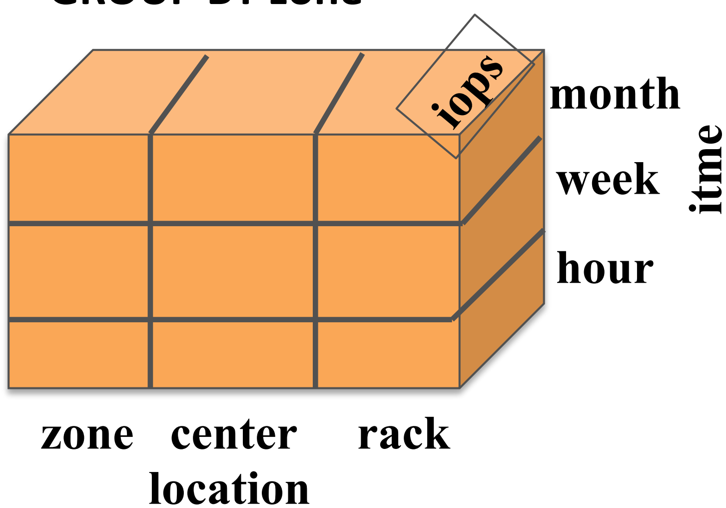


cube exploration

exploration
space reduction

user query

SELECT AVG (iops) FROM events
WHERE month="m1" AND week="w1"
GROUP BY zone



cube exploration operators

WHERE month="m1" — **parent**

WHERE month="m1"
AND week="w1
AND hour="h1" — **child**

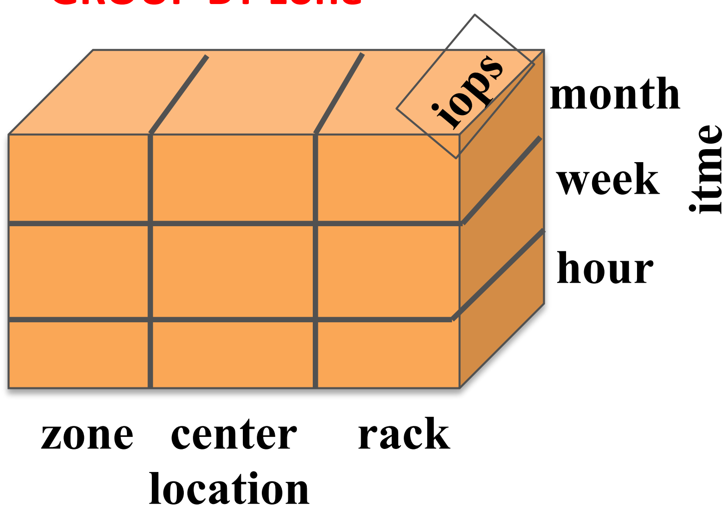
WHERE month="m1"
AND week="w2" — **sibling**

DICE, ICDE '14

cube exploration

user query

SELECT AVG (iops) FROM events
WHERE month="m1" AND week="w1"
GROUP BY zone



speculative queries

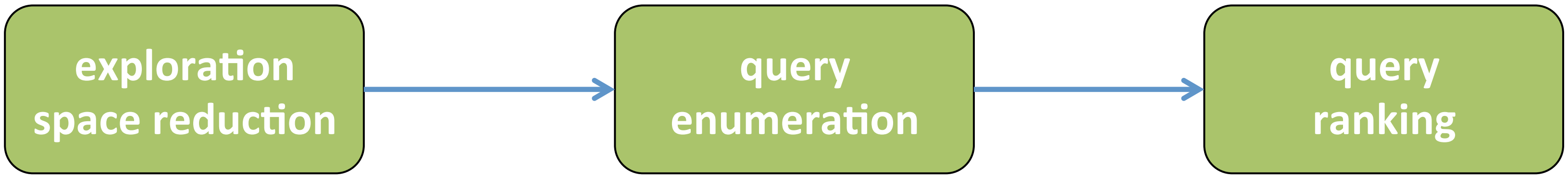
Q(month="m1")
…
Q(month = "m12")

Q(hour ="h1")
…
Q(hour =" h24")

Q(week="w2")
…
Q(week="w3")

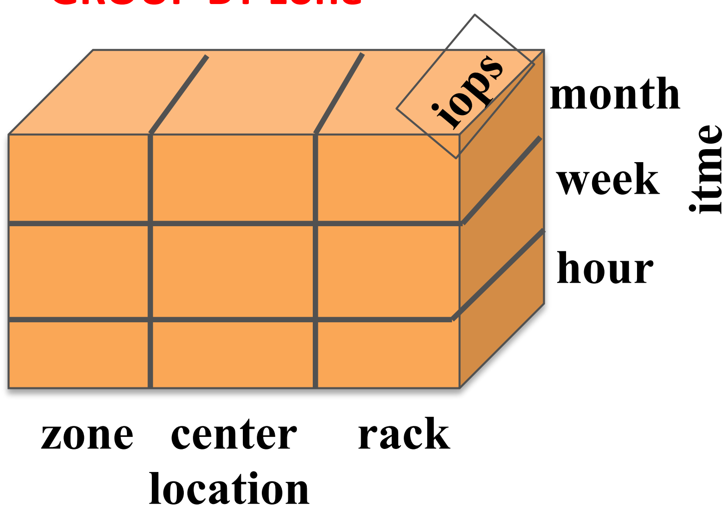
DICE, ICDE '14

cube exploration



user query

SELECT AVG (iops) FROM events
WHERE month="m1" AND week="w1"
GROUP BY zone



speculative queries

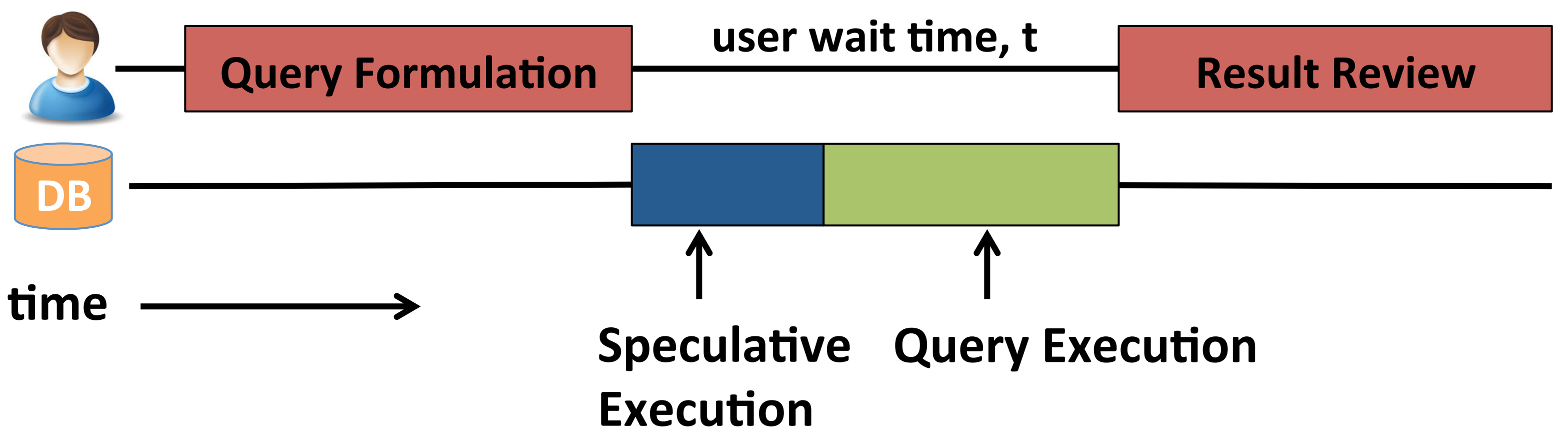
Q(month="m1")
…
Q(month = "m12")

Q(hour ="h1")
…
Q(hour =" h24")

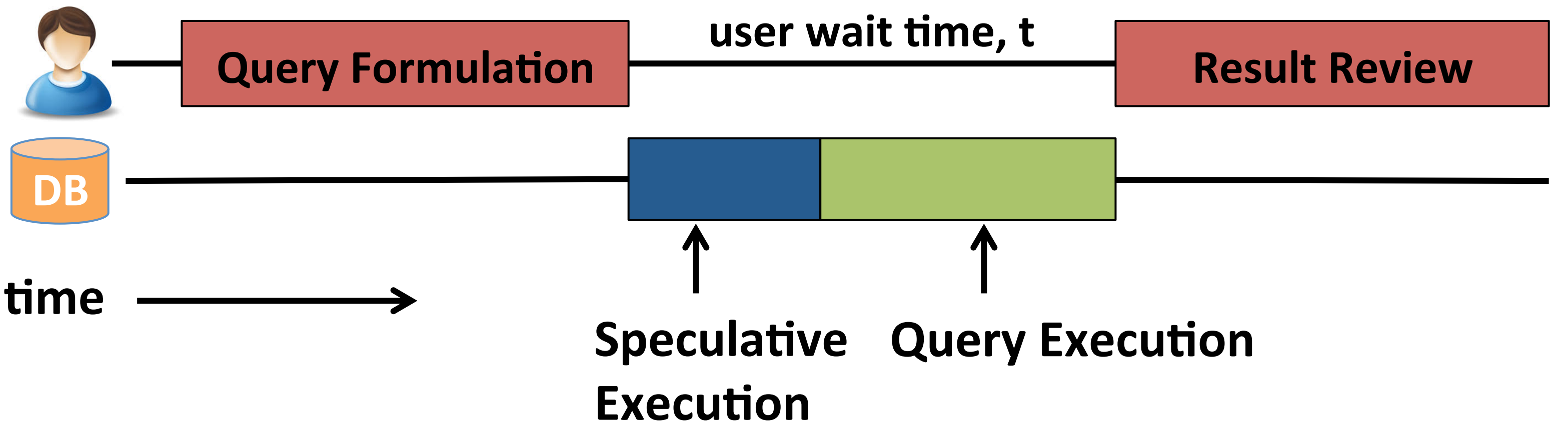
Q(week="w2")
…
Q(week="w3")

DICE, ICDE '14

cube exploration



cube exploration



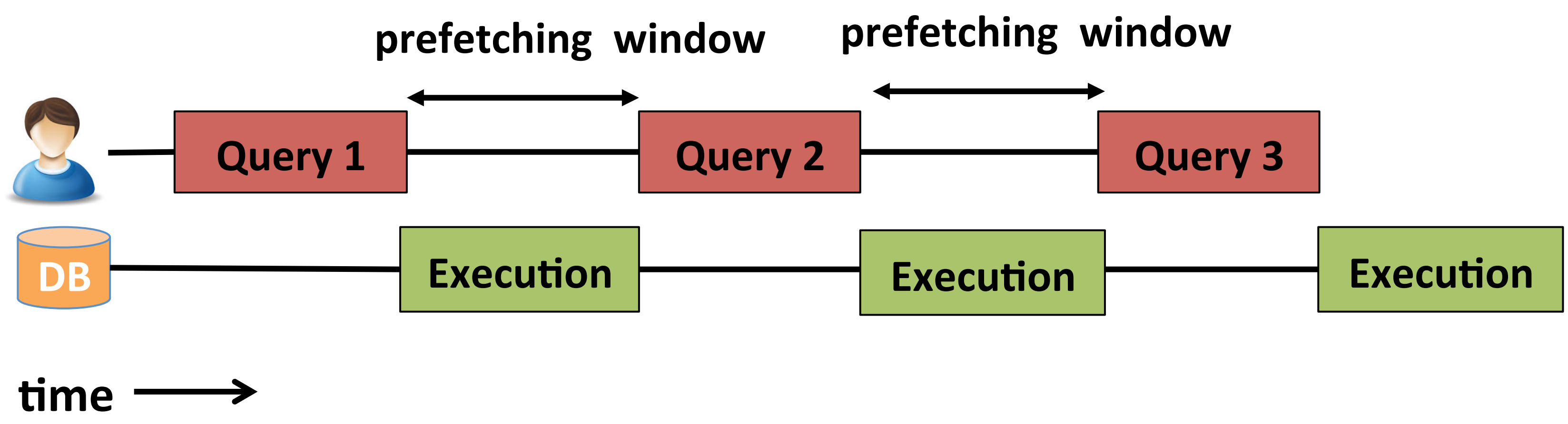
QUERY	Probability	Exec Time
Q_1	0.3	22
Q_2	0.25	20
Q_3	0.25	35
Q_4	0.15	70
Q_5	0.05	35

maximize query probability

total speculation time < t

DICE, ICDE '14

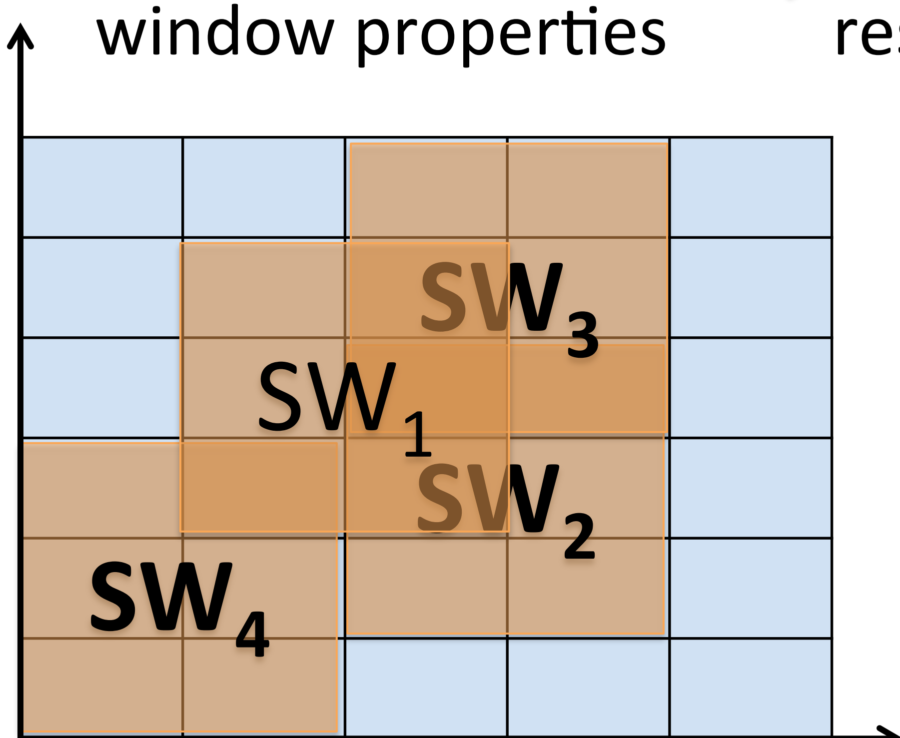
result reuse



- identify (likely) overlapping results
- cache them
- reduce query execution time (user wait time)

semantic windows

user-defined window properties → overlapping results/windows



2D exploration space

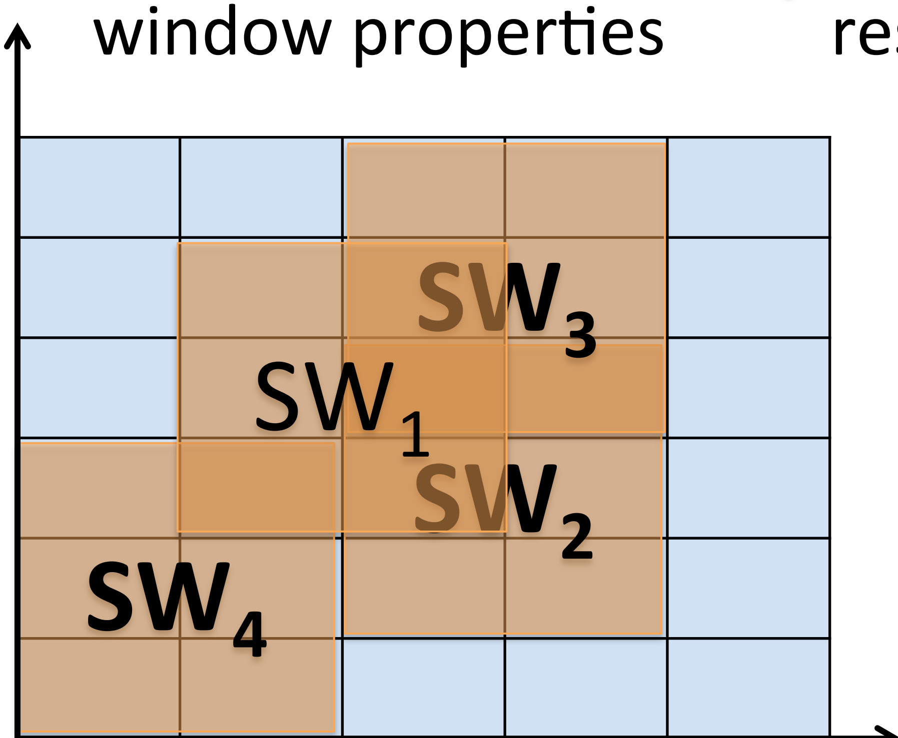
window prefetching

which order?

Kalinin et al, SIGMOD '14

semantic windows

user-defined window properties → overlapping results/windows

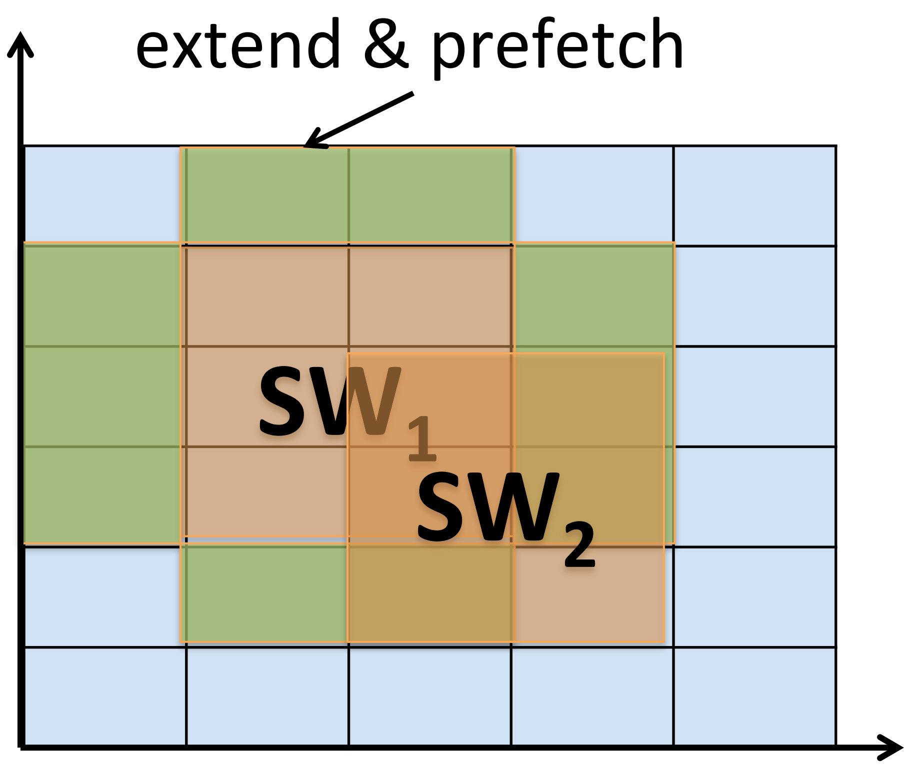


2D exploration space

utility-based result ranking
&
result prefetching

Kalinin et al, SIGMOD ’14

semantic windows

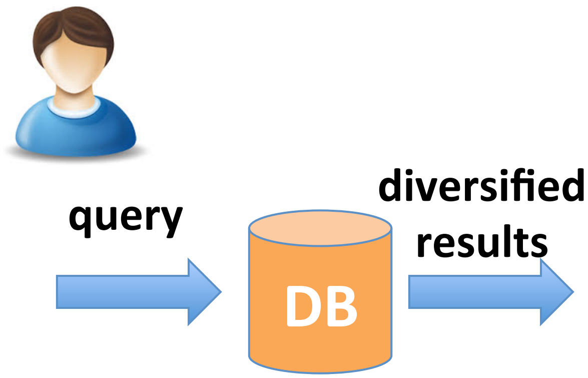


online performance
vs
query completion time

adjust prefetching size
to output progress

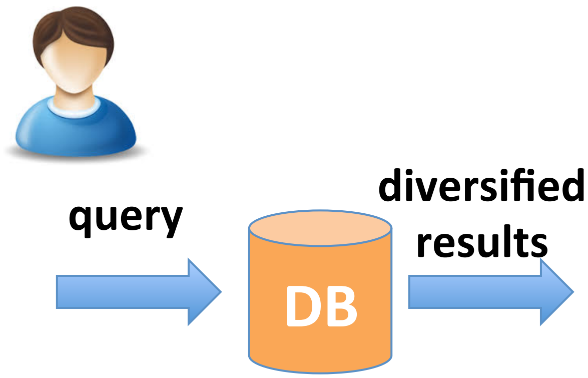
Kalinin et al, SIGMOD '14

data diversification



k representative tuples with max total pairwise distance

data diversification



k representative tuples with max total pairwise distance

Max Diversified Set Search

Query Output → **Diversified Output k= 3**

$\mathbf{T_1}$
$\mathbf{T_2}$
$\mathbf{T_3}$
$\mathbf{T_4}$
$\mathbf{T_5}$

$d(T_1, T_3)$

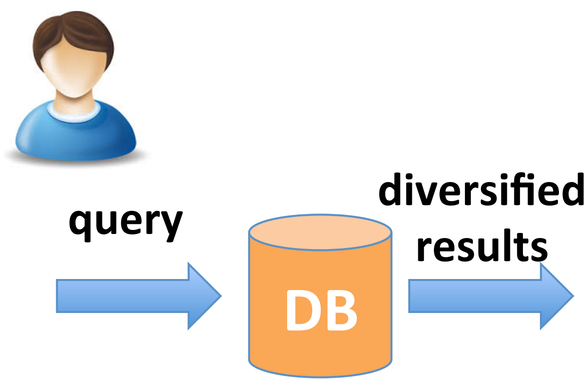
$d\ (T_2, T_3)$

← random tuple

$d(T_4, T_3)$

$d(T_5, T_3)$

data diversification

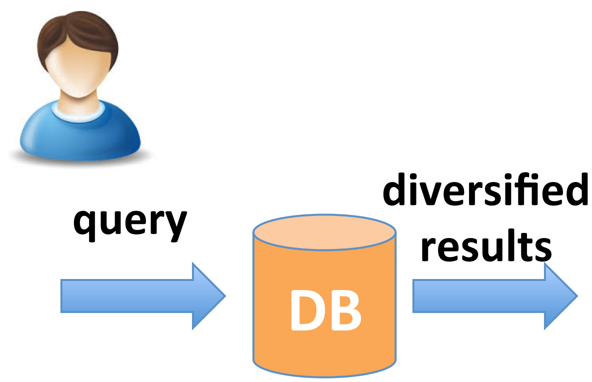


k representative tuples with max total pairwise distance

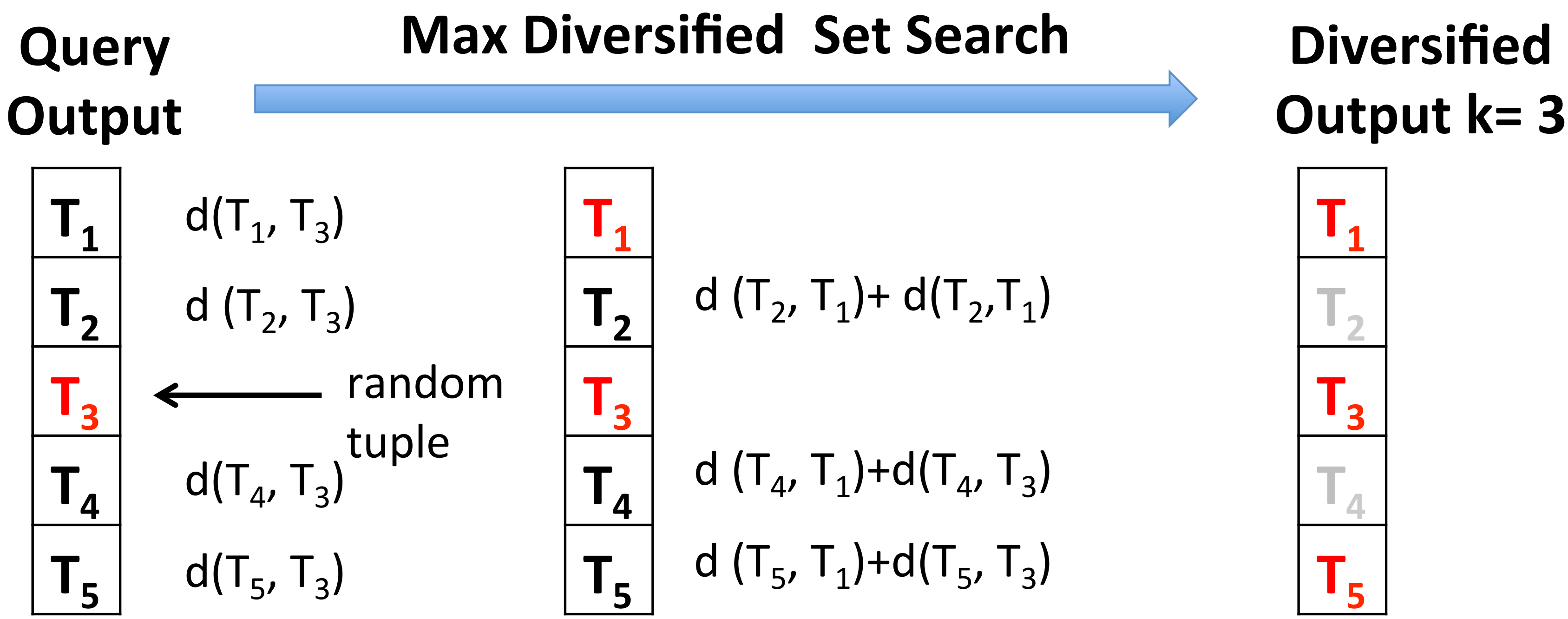
Query Output → **Max Diversified Set Search** → **Diversified Output k= 3**

$\mathbf{T_1}$	$d(T_1, T_3)$		$\mathbf{T_1}$	
$\mathbf{T_2}$	$d(T_2, T_3)$		$\mathbf{T_2}$	$d(T_2, T_1) + d(T_2, T_3)$
$\mathbf{T_3}$	←	random tuple	$\mathbf{T_3}$	
$\mathbf{T_4}$	$d(T_4, T_3)$		$\mathbf{T_4}$	$d(T_4, T_1) + d(T_4, T_3)$
$\mathbf{T_5}$	$d(T_5, T_3)$		$\mathbf{T_5}$	$d(T_5, T_1) + d(T_5, T_3)$

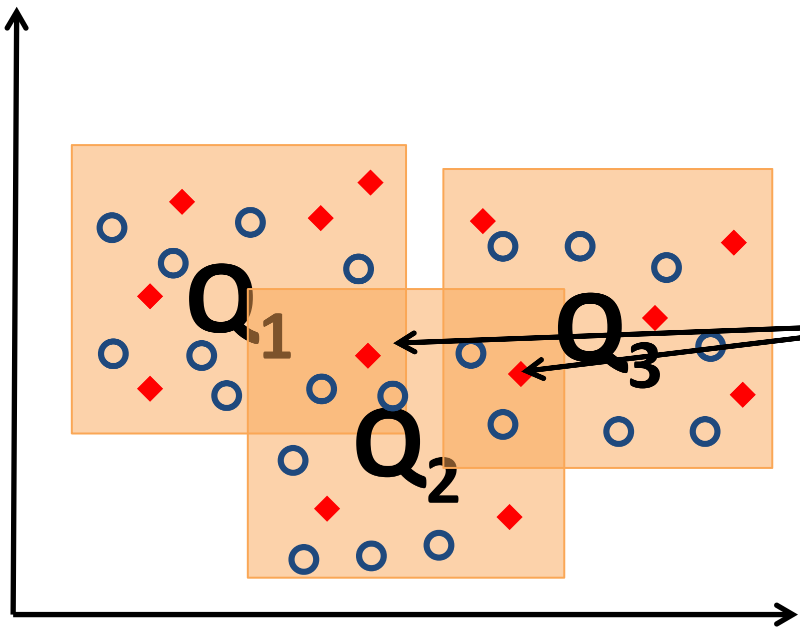
data diversification



k representative tuples with max total pairwise distance



interactive data diversification



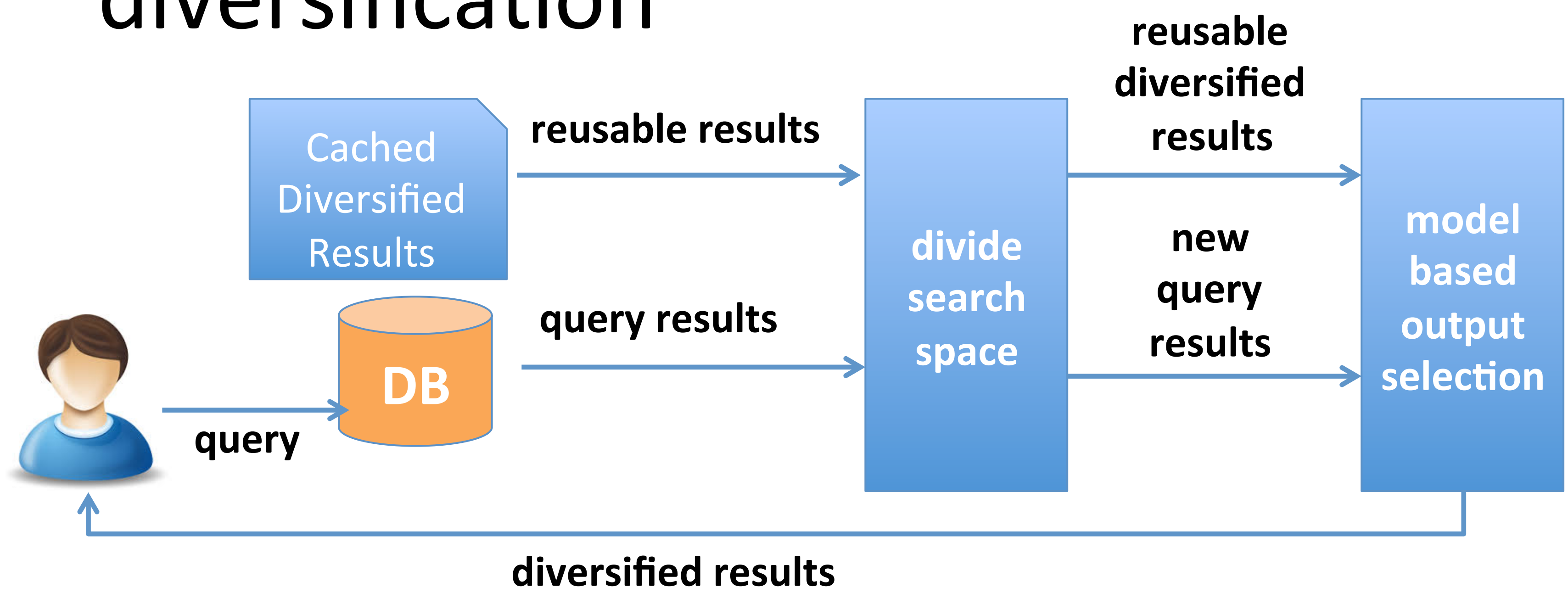
overlapping diversified results

long Time-To-Insight

- cache diversified results and use most promising
- regression model predicts max diversification of a set

DivIDE, SSDBM'14

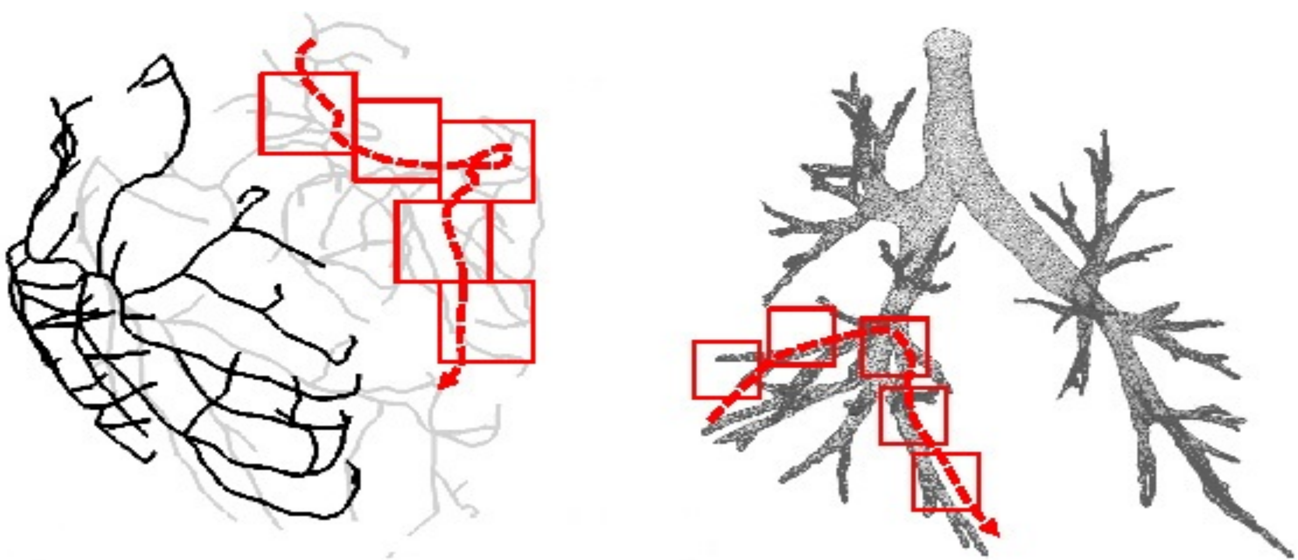
interactive data diversification



- search space pruning through regression model
- best/first fit search for max total diversification among cached and new results

structure-aware prefetching

- prefetching for interactive spatial query sequences
- model structures of past spatial queries in graph
- identify guiding structure in past two queries : iterative pruning
- cache the predicted next location



SCOUT, VLDB'12