ProDIS

Provenance for Data-Intensive Systems

ProDIS: Provenance for Data-Intensive Systems

- Databases, Data Mining, Data Science…
- Highly complex logic, Big Data

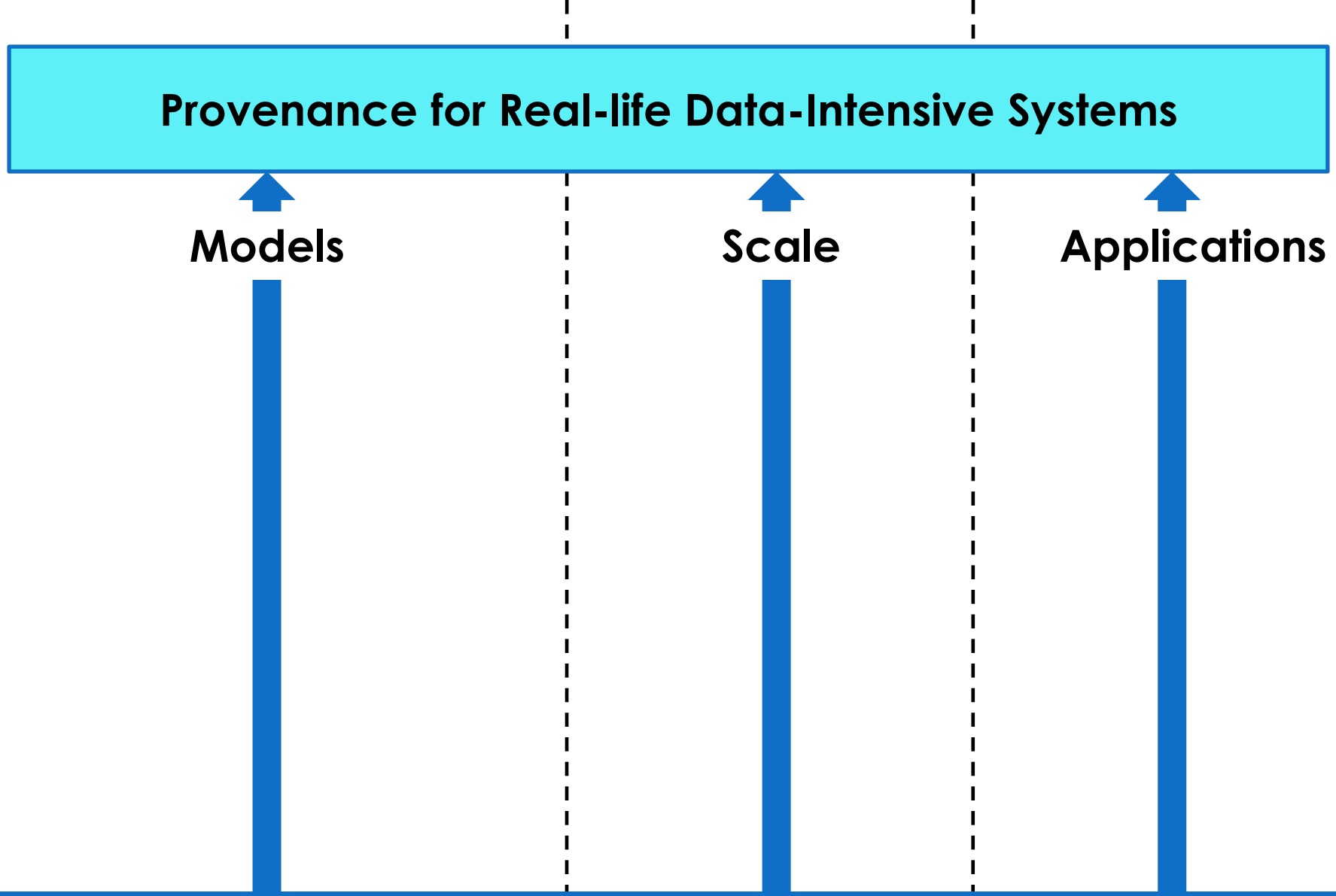
Provenance of output is typically unknown

- Why, what if, what data was used, can we trust?...
- Without answers to these questions, results may be useless/harmful
  - Medical recommendations, loan request rejections..

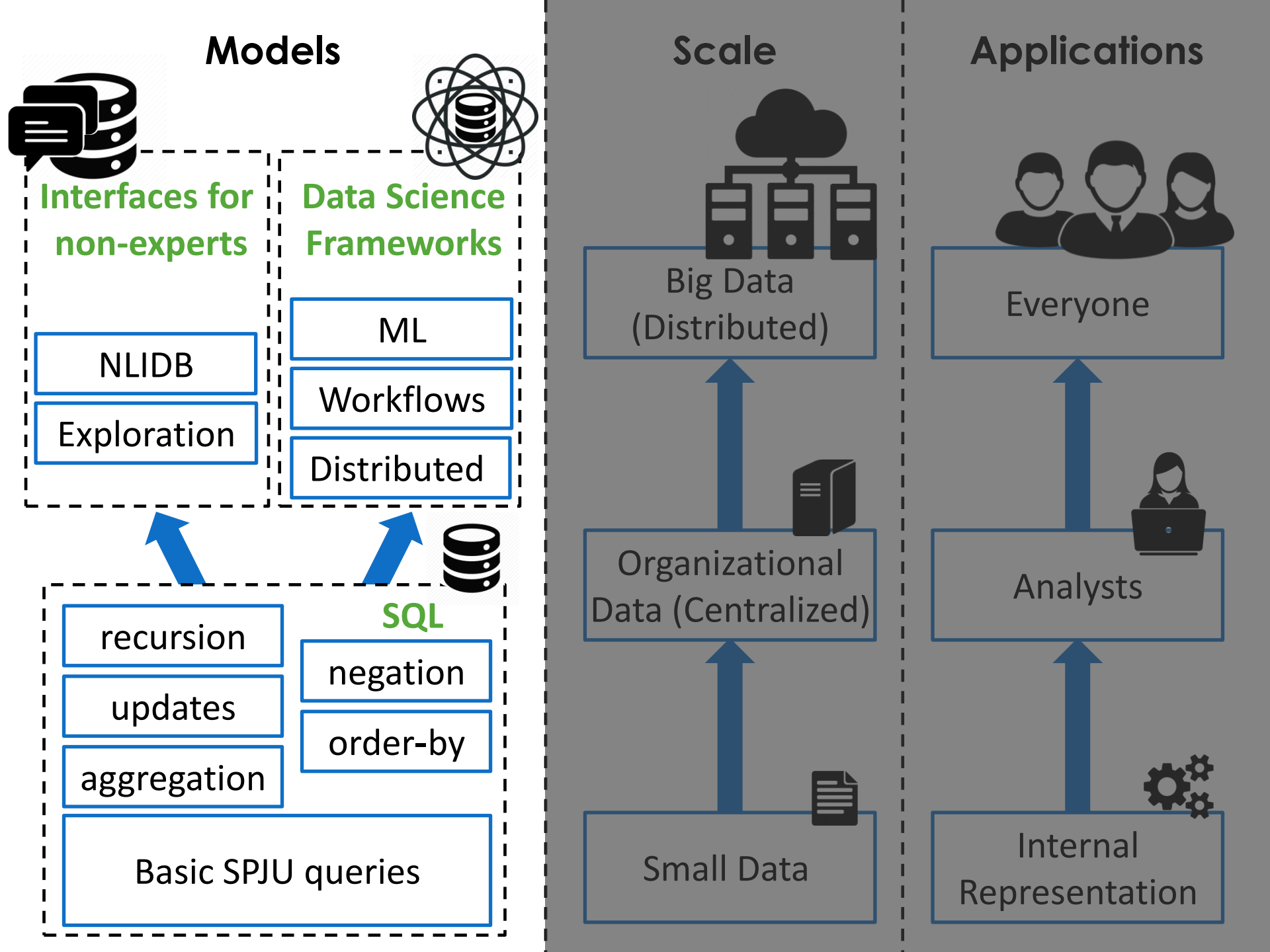
Imagine a world where computation results are
accounted for and explained

Systems would be transparent and controllable,
and the results credible and reusable

Provenance for Real-life Data-Intensive Systems



Data Provenance: theory and algorithms


Models
Scale
Applications
Interfaces for non-experts
Data Science Frameworks
NLIDB
Exploration
ML
Workflows
Distributed
SQL
recursion
updates
aggregation
negation
order-by
Basic SPJU queries
Big Data (Distributed)
Organizational Data (Centralized)
Small Data
Everyone
Analysts
Internal Representation

Models

"Return customers with overdraft events after 2006"

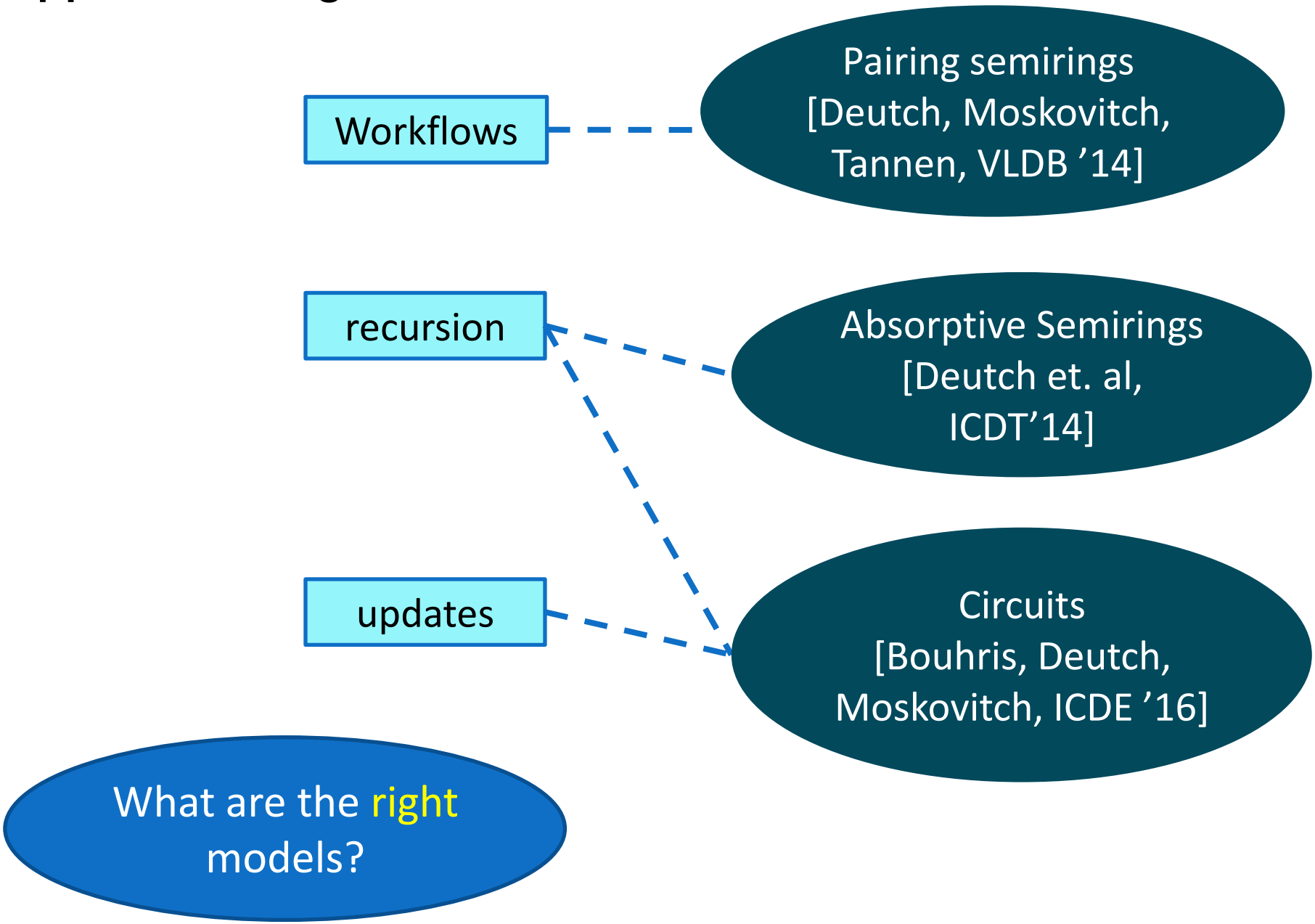
eventid	sum	type	due	Prov
1	50000	overdraft	2012	p1
2	400000	mortgage	2014	p2
3	2000000	overdraft	2010	p3

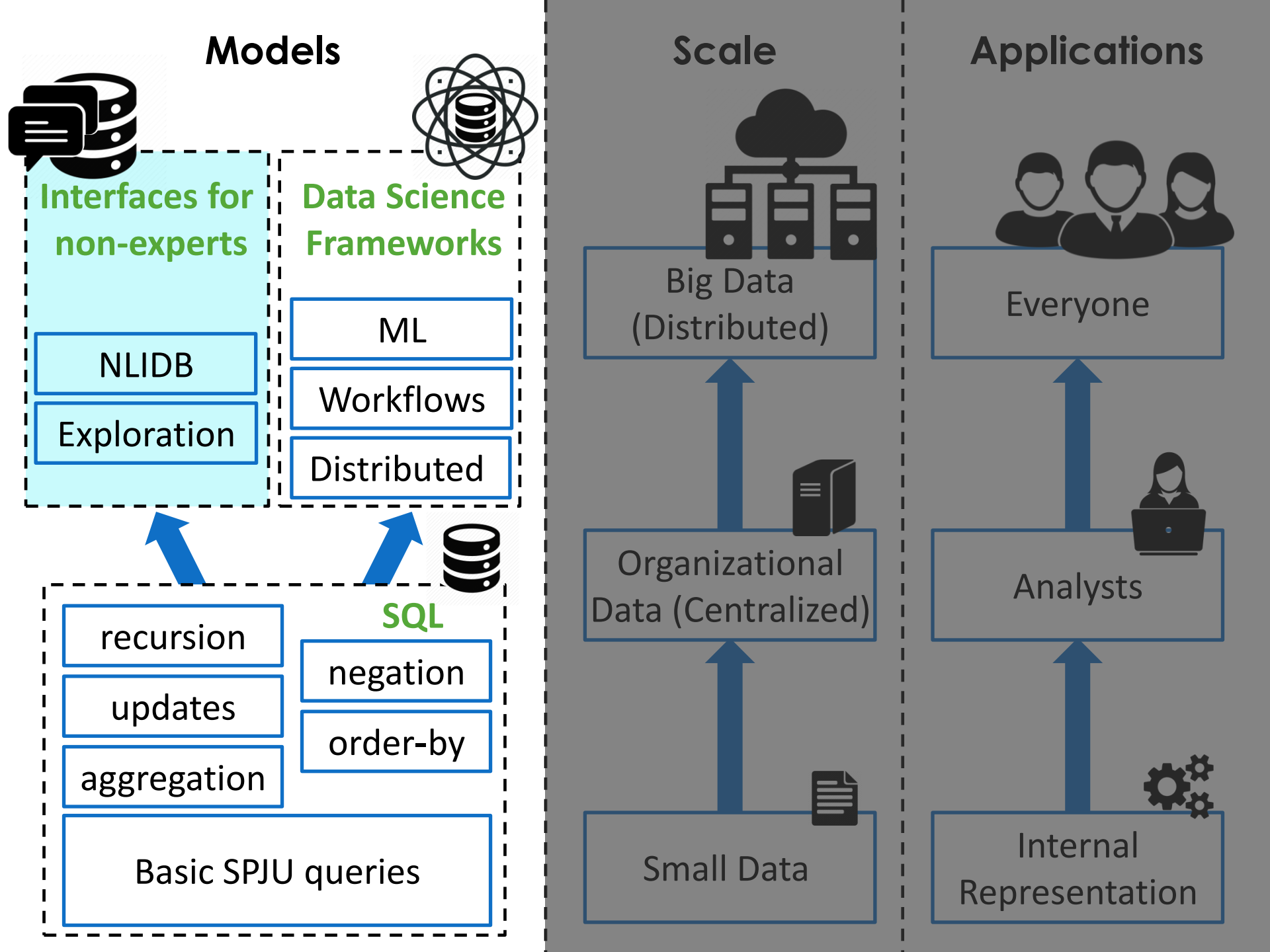
custname	eventid	prov
Smith	1	c1
Smith	3	c2
Roth	2	c3

custname	prov
Smith	p1·c1+p3·c2

[Green et. al, PODS 2007]

Models
Approach I: Algebraic Provenance




Models
Scale
Applications
Interfaces for non-experts
NLIDB
Exploration
Data Science Frameworks
ML
Workflows
Distributed
SQL
recursion
updates
aggregation
negation
order-by
Basic SPJU queries
Big Data (Distributed)
Organizational Data (Centralized)
Small Data
Everyone
Analysts
Internal Representation

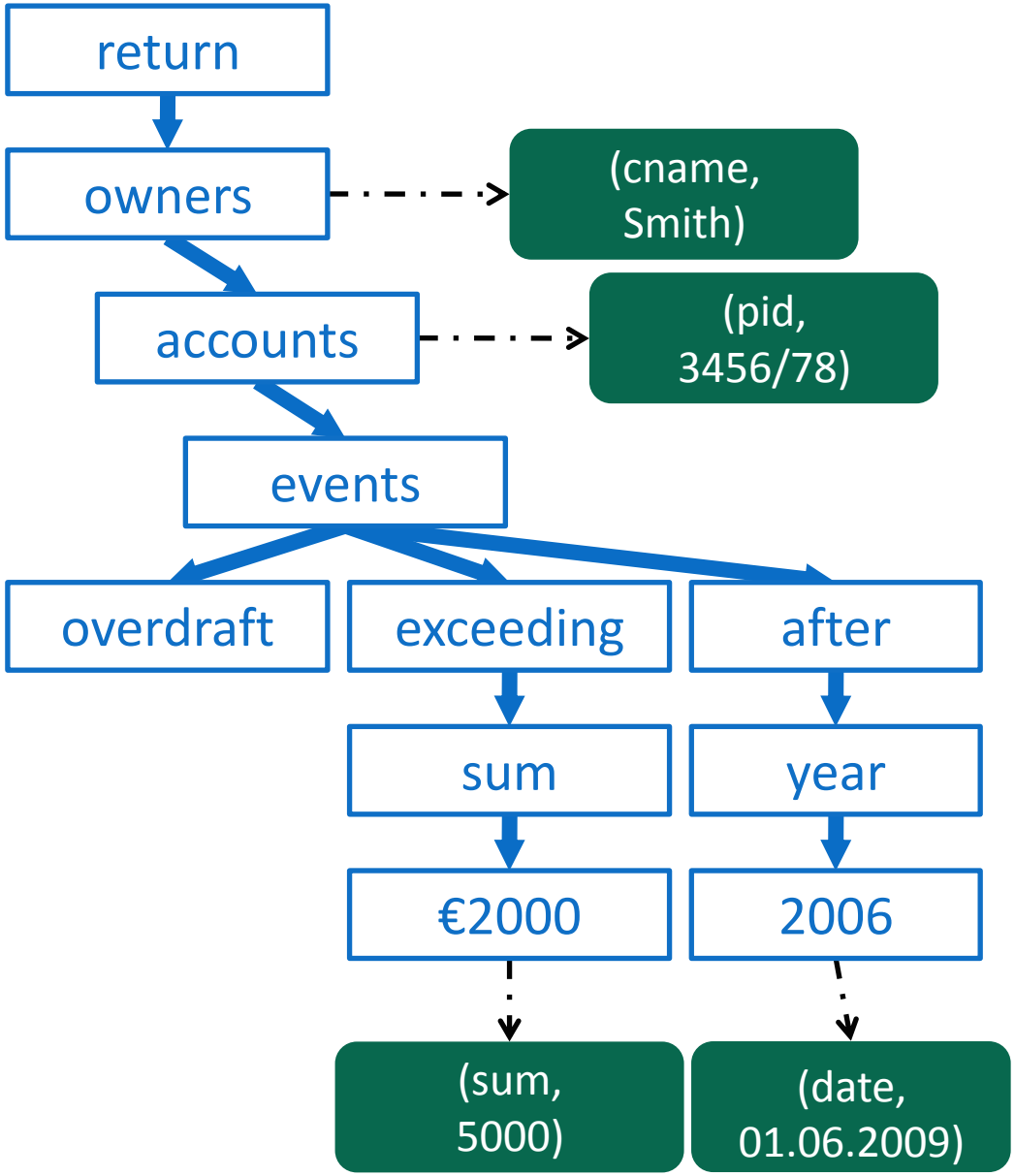
Models
Approach II: Interaction-Based Provenance

NLIDB

"Return owners of accounts with overdraft events exceeding a sum of €2000 after the year 2006"

...

"Why is 'Smith' an answer?"



[Deutch, Frost, Gilad, VLDB '17 best paper]

Models
Approach II: Interaction-Based Provenance

Data Science Frameworks

ML

Workflows

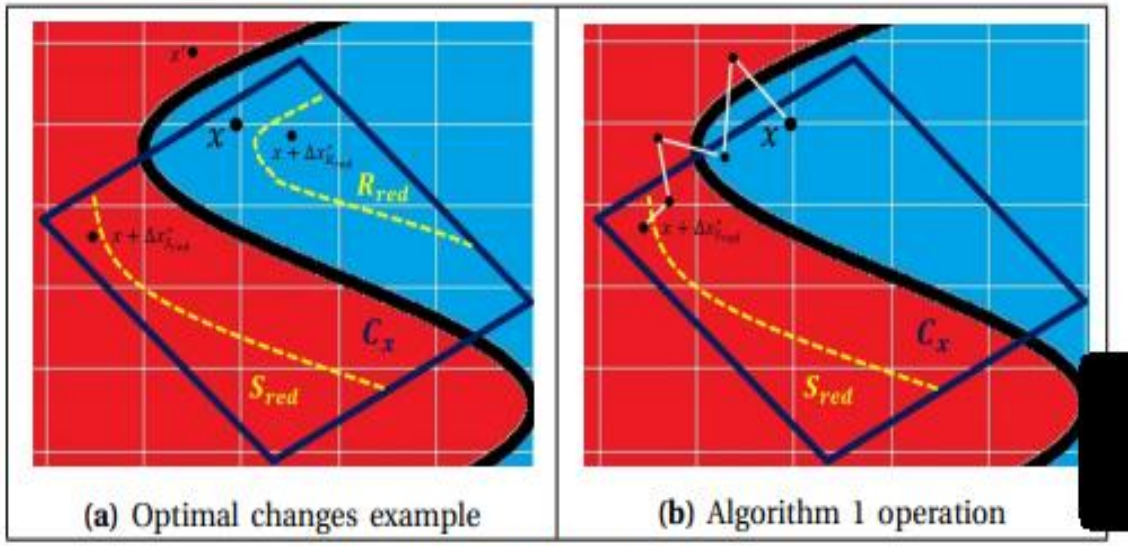
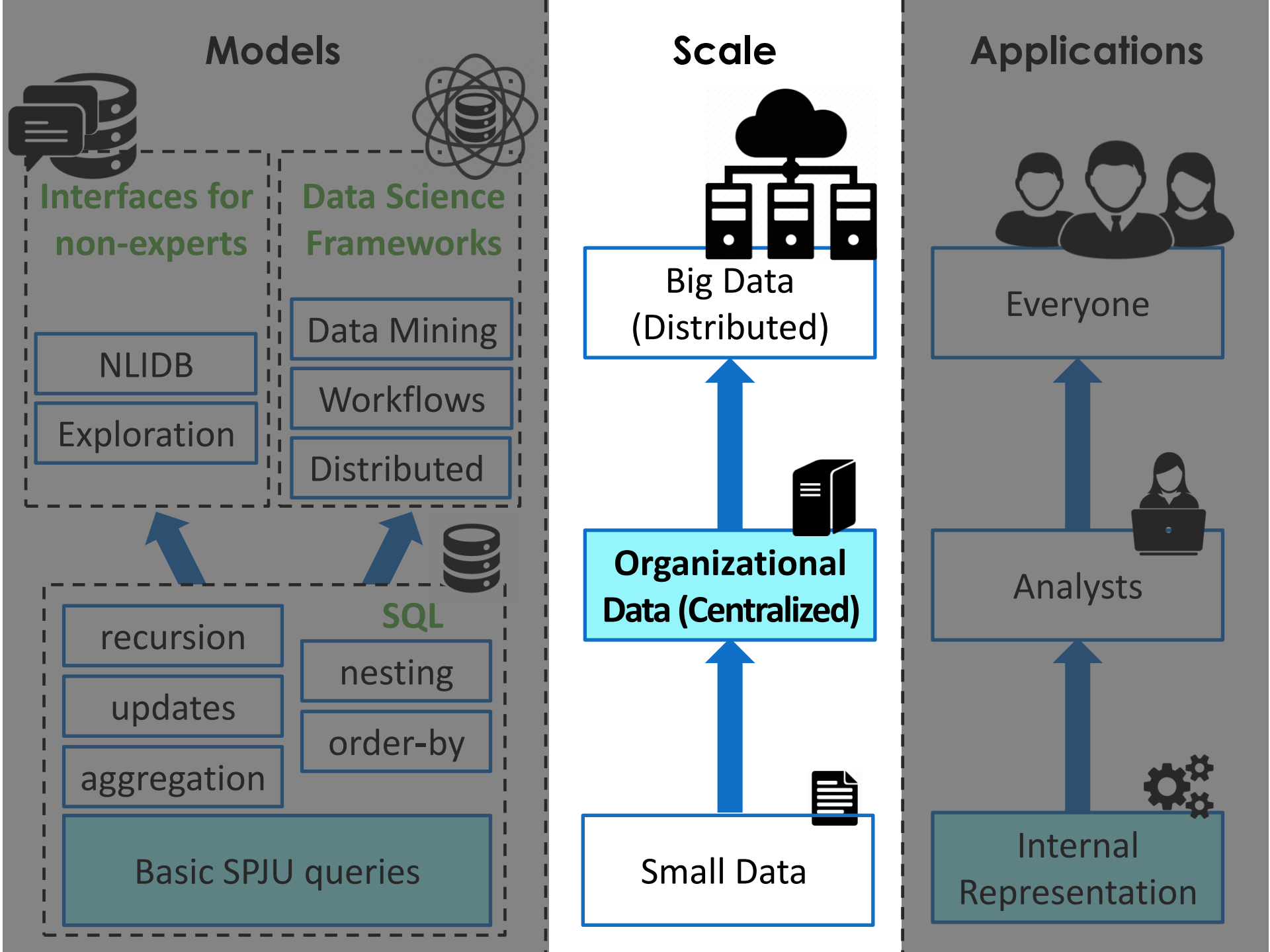


Figure 2: Optimal changes

Integrity Constraint	Formula Example	Project Constraint
Domain	$\forall t \in R, t.v_i \in [a, b]$	$v_i \geq a \wedge v_i \leq b$
Check	$\forall t \in R, t.v_i \leq t.v_j$	$v_i - v_j \leq 0$
CFD	$\forall t \in R, t.v_i = V_1 \rightarrow t.v_j = V_2$	$v_i \neq V_1 \vee v_j = V_2$
Unary Denial	$\forall t \in R, \neg(\bigwedge_{i=1}^{k} t.v_i = V_i)$	$\bigvee_{i=1}^{k} v_i \neq V_i$

D.Deutch and N.Frost, **Constraints-based Explanations of Classifications**, ICDE 2019


Models
Interfaces for non-experts
NLIDB
Exploration
Data Science Frameworks
Data Mining
Workflows
Distributed
SQL
recursion
updates
aggregation
nesting
order-by
Basic SPJU queries
Scale
Big Data (Distributed)
Organizational Data (Centralized)
Small Data
Applications
Everyone
Analysts
Internal Representation

```
SELECT Customer.cname
FROM Customer, Ownership, Product,
Assoc, Event, DebtEvent, Currency
WHERE Customer.cid = Ownership.cid
AND Ownership.pid = Product.pid
AND Product.type LIKE '%account'
AND Product.pid=Assoc.pid AND
Assoc.eid=Event.eid AND
Event.date > '01.01.2007' AND
Event.eid = DebtEvent.eid
DebtEvent.sum > 2000
```

"Return owners of accounts with overdraft events exceeding a sum of €2000 after the year 2006"

Organizational Data (Centralized)

Basic SPJU queries

Internal Representation

cname	prov
Smith	C0123·01325·P85335·A821 ·E23874·DE23874·CU2+ C0123·01325·P85335·A43 ·E9873·DE9873·CU2+ …
Jones	C8432·012387·P1248 ·E2384·DE2384·CU …

"Return owners of accounts with overdraft events exceeding a sum of €2000 after the year 2006"

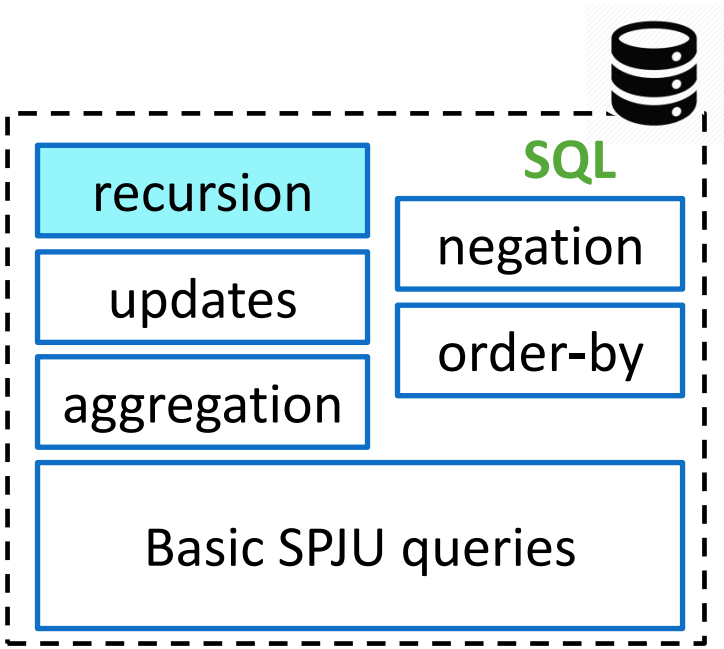
PTIME but practically inefficient

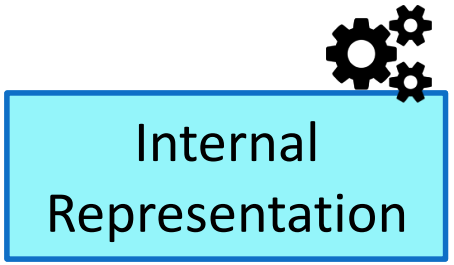
Organizational Data (Centralized)

Basic SPJU queries

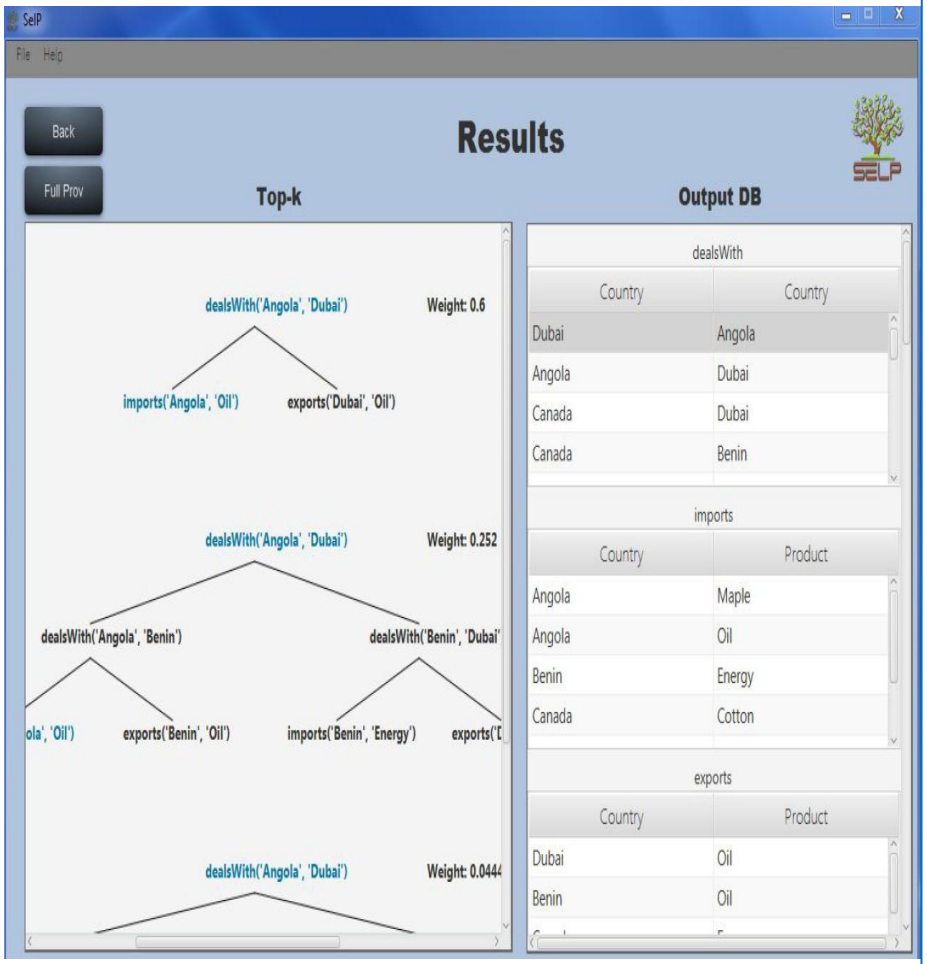
Internal Representation

Super-polynomial lower bound for datalog [Deutch et. al, ICDT ’14]





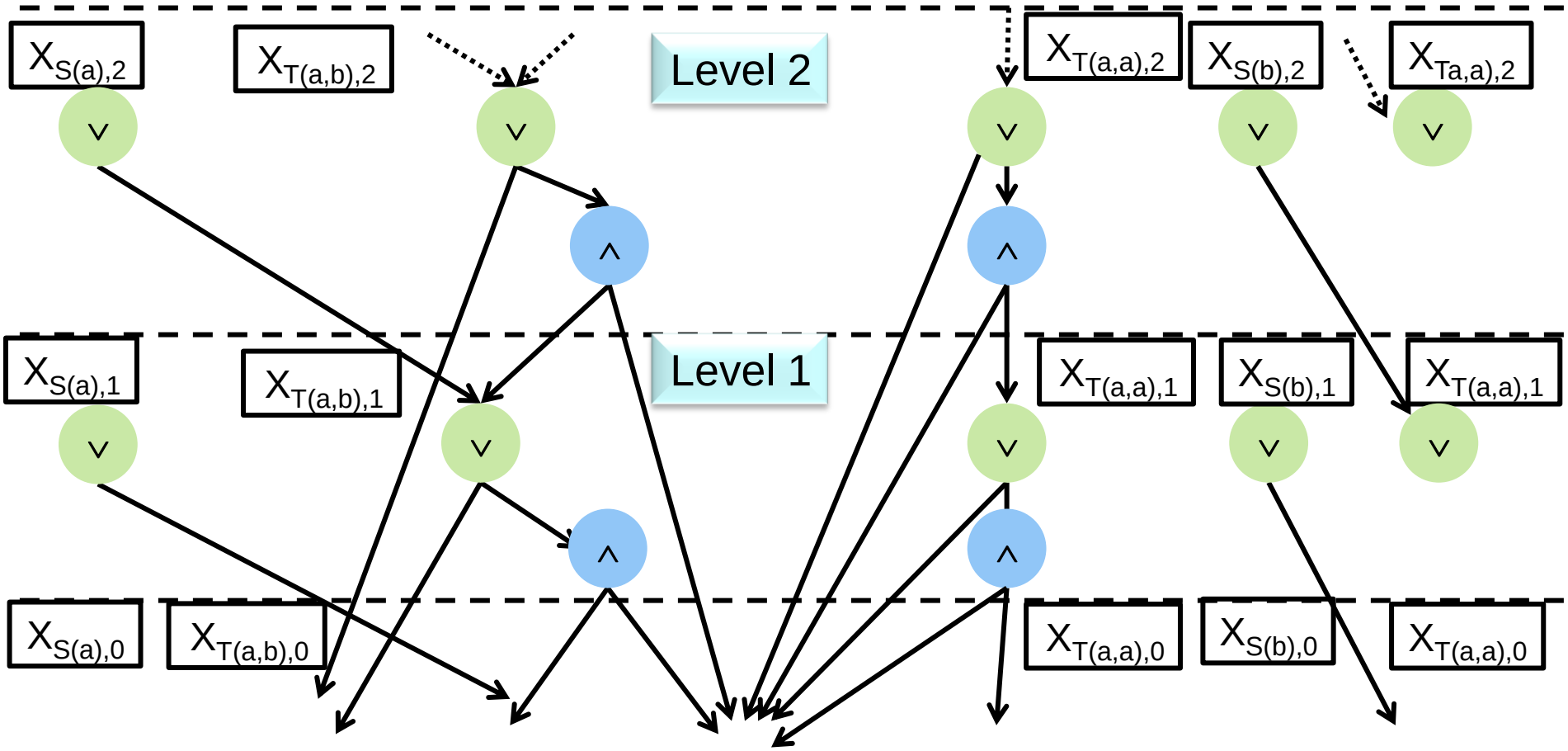
Scalable Provenance Solutions
Approach I: Selective Provenance Tracking



[Deutch, Gilad, Moskovitch, VLDB ’15, VLDB Journal ‘18]

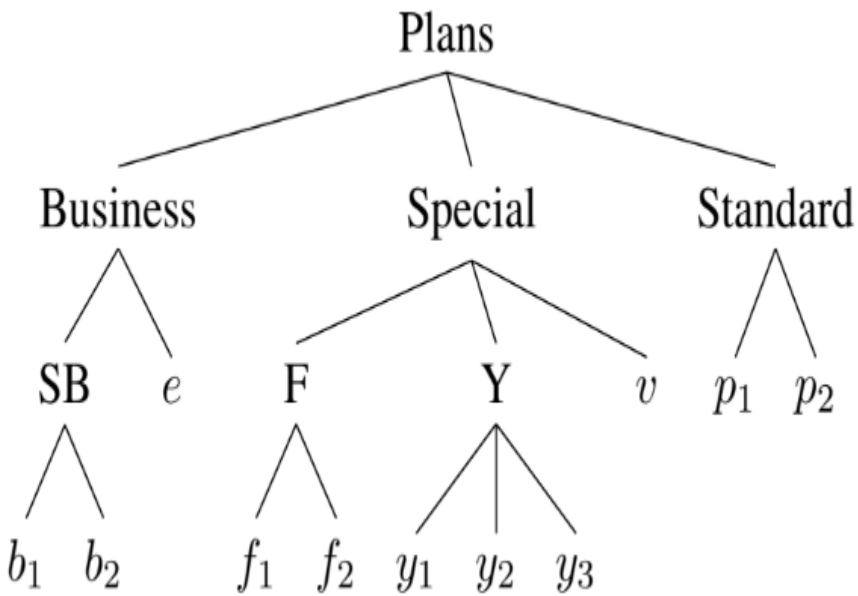
[Bouhris, Deutch, Moskovitch, ICDE ‘16]

Scalable Provenance Solutions
Approach II: Summarization



[Deutch et. al, **Provenance for Datalog Circuits**, ICDT '14]

Scalable Provenance Solutions
Approach III: Abstraction



[Deutch,Moskovitch, Rinetzky,
Hypothetical Reasoning Via Provenance Abstraction, SIGMOD '19]

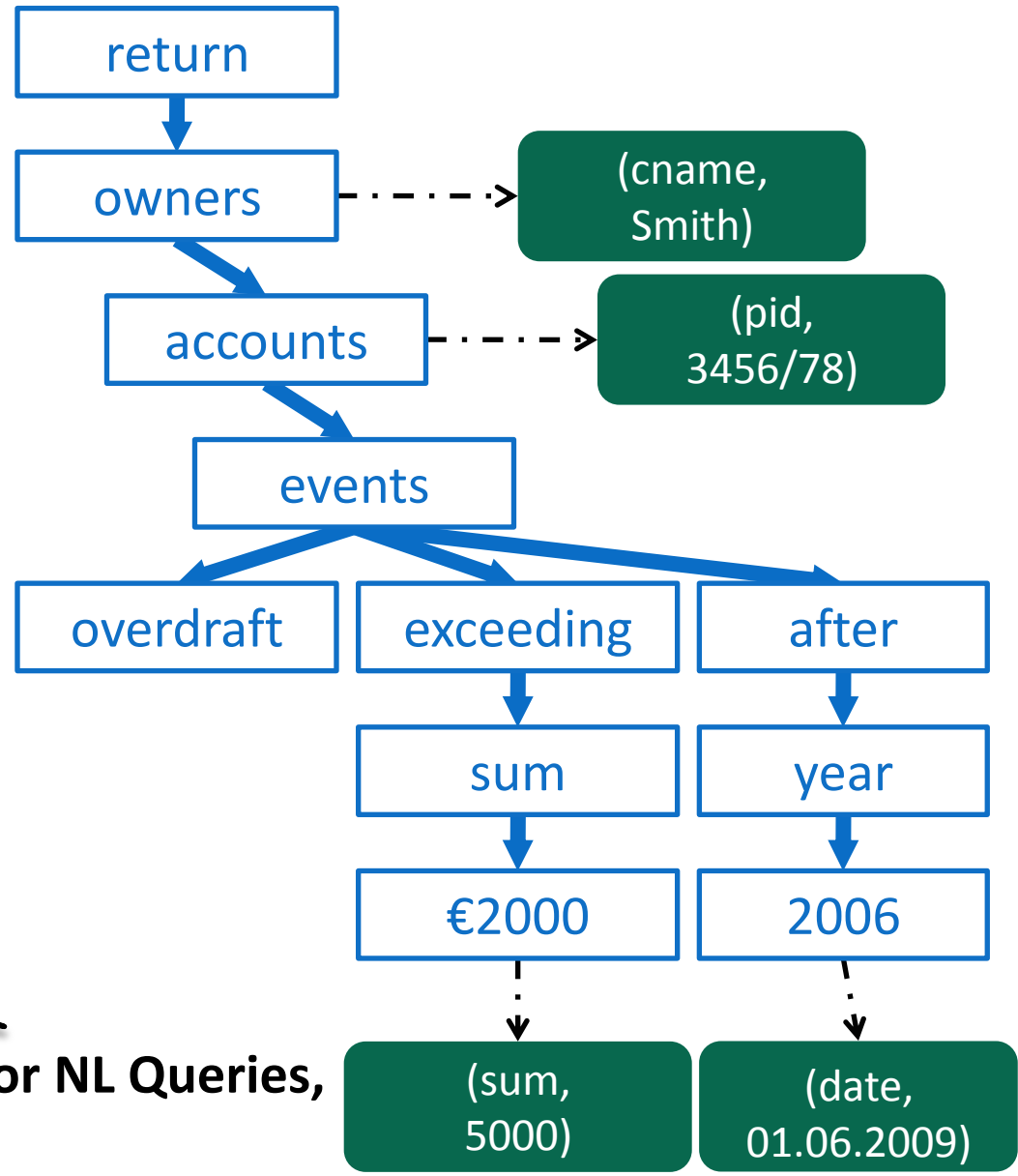
Scalable Provenance Solutions
Approach IV: Interaction Based

"Return owners of accounts with overdraft events exceeding a sum of €2000 after the year 2006"

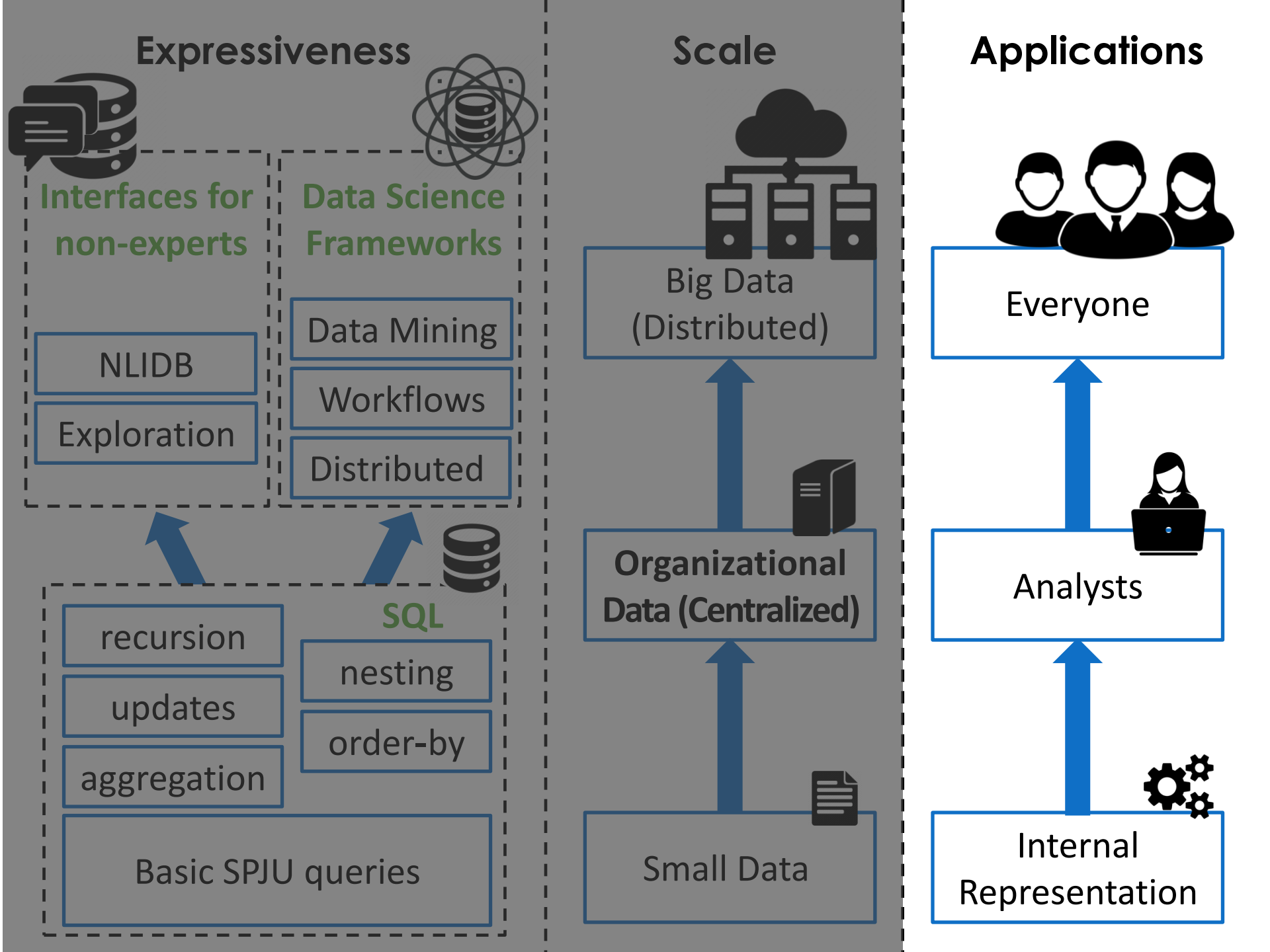
...

"Why is 'Smith' an answer?"

"Smith is the owner of 3 accounts with 13 overdraft events of a total sum of €30000 in 01.02.2009-01.05.2010"



[Deutch, Frost, Gilad, **Provenance for NL Queries,** VLDB '17 best paper]


Expressiveness
Scale
Applications
Interfaces for non-experts
Data Science Frameworks
NLIDB
Exploration
Data Mining
Workflows
Distributed
SQL
recursion
updates
aggregation
nesting
order-by
Basic SPJU queries
Big Data (Distributed)
Organizational Data (Centralized)
Small Data
Everyone
Analysts
Internal Representation

Provenance Applications

"Why is 'Smith' an answer?"

"On what sources is the 'Smith' answer based on?"

"Smith is the owner of 3 accounts with 13 overdraft events of a total sum of €30000 in 01.02.2009-01.05.2010"

"Citibank combined with American Express and independently BNP Paribas combined with Visa "

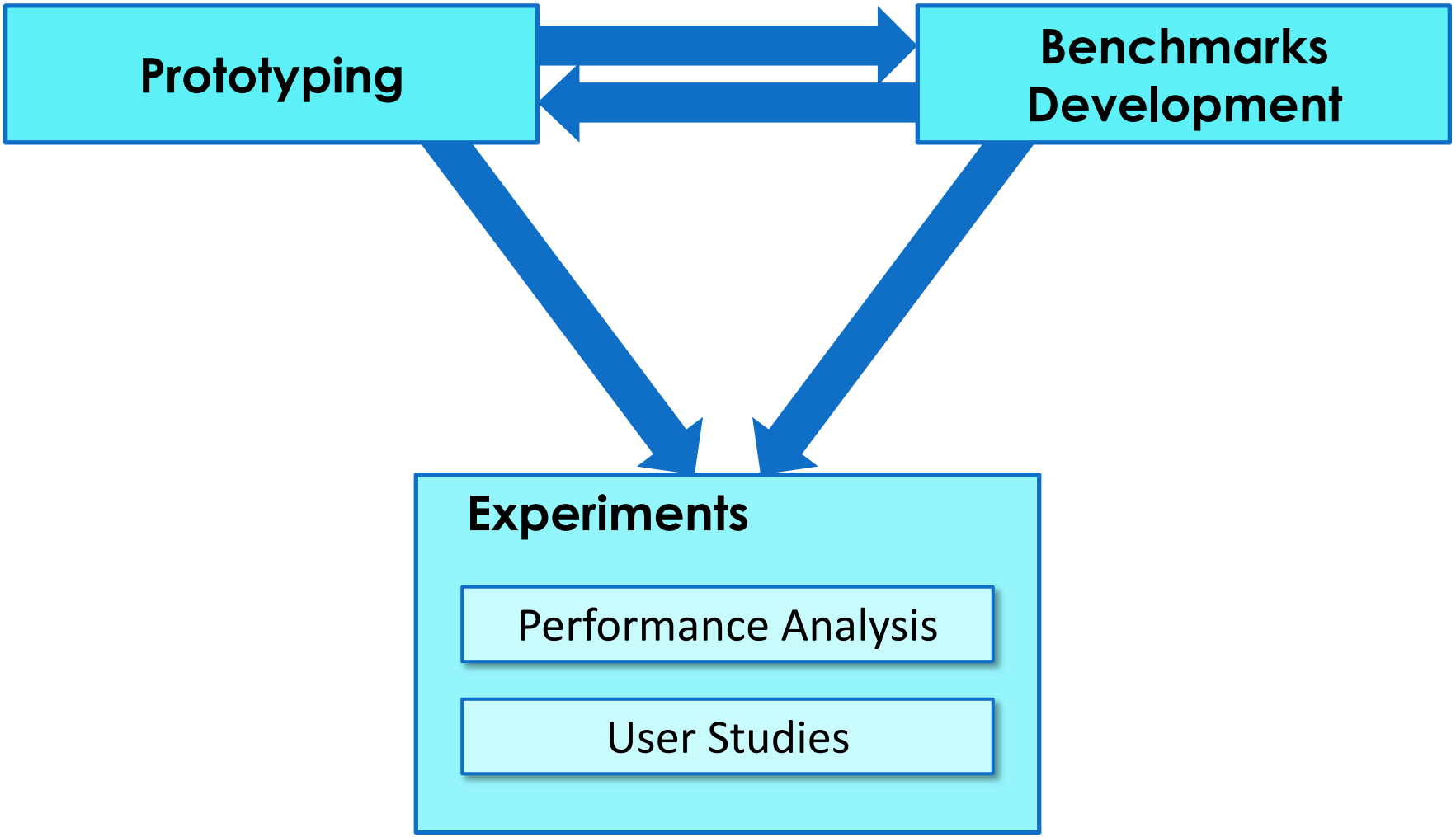
"How could 'Smith' Become a non-answer?"

"What if a particular overdraft event of 'Smith' is excused?"

"Remove overdraft event of date 01.04.2009 of sum €10000 "

"'Smith' would still be an answer"

Implementation and Evaluation



Vision: a world where computation results are **accounted for and explained**

Essence of computation
• why • what if • trust • …

Provenance for Real-life Data-Intensive Systems

Models | **Scale** | **Applications**

Data Provenance: theory and algorithms