

Fancy types for provenance

James Cheney
University of Edinburgh

The provenance crisis

The provenance crisis



UAL Shares Fall as Old Story Surfaces Online - WSJ.com - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://online.wsj.com/article/SB12 Google

SEPTEMBER 9, 2008

UAL Shares Fall as Old Story Surfaces Online

The mysterious appearance on the Internet of a nearly six-year-old news story about UAL Corp.'s 2002 bankruptcy-court filing caused investors to dump the stock Monday.

After trading near \$12.50 a share early Monday, stock in United Airlines' parent quickly fell to \$3 on the Nasdaq Stock Market on heavy volume before trading was halted and the company issued a statement saying that reports of a new Chapter 11 filing were "completely untrue."

Once trading resumed 90 minutes later, UAL shares rebounded, but they still closed off 11% for the day at \$10.92. Nasdaq, a unit of Nasdaq OMX Group Inc., ...

The provenance crisis



The provenance crisis

UAI Shares Fall as Old Story Surfaces Online - WSJ.com - Mozilla Firefox

SCIENTIFIC PUBLISHING

A Scientist's Nightmare: Software Problem Leads to Five Retractions

Until recently, Geoffrey Chang's career was on a trajectory most young scientists only dream about. In 1999, at the age of 28, the protein crystallographer landed a faculty position at the prestigious Scripps Research Institute in San Diego, California. The next year, in a ceremony at the White House, Chang received a

2001 *Science* paper, which described the structure of a protein called MsbA, isolated from the bacterium *Escherichia coli*. MsbA belongs to a huge and ancient family of molecules that use energy from adenosine triphosphate to transport molecules across cell membranes. These so-called ABC transporters perform many

The provenance crisis



SCIENTIFIC PUBLISHING

A Scientist's Nightmare: Software Problem Leads to Five Retractions

Until recently, Geoffrey Chang's career was on a trajectory most young scientists only dream about. In 1999, at the age of 28, the protein crystallographer landed a faculty position at the prestigious Scripps Research Institute in San Diego, California. The next year, in a ceremony at the White House, Chang received a

2001 *Science* paper, which described the structure of a protein called MsbA, isolated from the bacterium *Escherichia coli*. MsbA belongs to a huge and ancient family of molecules that use energy from adenosine triphosphate to transport molecules across cell membranes. These so-called ABC transporters perform many

The provenance crisis

The screenshot shows the BBC News website interface. At the top, there's a navigation bar with 'BBC Mobile' and links for 'News', 'Sport', 'Weather', 'iPlayer', 'TV', 'Radio', and 'More'. A search box is on the right. Below the navigation bar is a red banner with 'NEWS' and 'LIVE BBC NEWS CHANNEL'. The main content area features a headline: "'Show Your Working': What 'ClimateGate' means". Below the headline is a 'VIEWPOINT' section with a photo of Mike Hulme and Jerome Ravetz. The article text discusses the 'ClimateGate' affair and the need for better communication. On the left, there's a sidebar with navigation links like 'News Front Page', 'World', 'UK', 'England', etc. On the right, there's a 'THE GREEN ROOM' section with a sub-article 'Pinch of salt'.

UAL Shares Fall as Old Story Surfaces Online - WSJ.com - Mozilla Firefox

File Edit View History Bookmarks Tools Help

BBC Mobile News Sport Weather iPlayer TV Radio More Search the BBC

NEWS LIVE BBC NEWS CHANNEL

Page last updated at 14:56 GMT, Tuesday, 1 December 2009

E-mail this to a friend Printable version

'Show Your Working': What 'ClimateGate' means

VIEWPOINT
Mike Hulme and Jerome Ravetz

The "ClimateGate" affair - the publication of e-mails and documents hacked or leaked from one of the world's leading climate research institutions - is being intensely debated on the web. But what does it imply for climate science? Here, Mike Hulme and Jerome Ravetz say it shows that we need a more concerted effort to explain and engage the public in understanding the processes and practices of science and scientists.

THE GREEN ROOM
A weekly series of thought-provoking pieces on environmental topics

Pinch of salt
Idea that the world production must do wrong"

Your comments

RECENT ARTICLES

SCIE
A S
Pro
Until re
a trajec
about.
crystal
the pre
San Di
emony

The provenance crisis



SCIENTIFIC PUBLISHING

A Scientist's Nightmare: Software Problem Leads to Five Retractions

A screenshot of the BBC News website. The page features the BBC logo, navigation links for News, Sport, Weather, iPlayer, TV, Radio, and More. The main headline is "'Show Your Working': What 'ClimateGate' means". Below the headline is a 'VIEWPOINT' section by Mike Hulme and Jerome Ravetz, discussing the 'ClimateGate' affair. There is also a 'THE GREEN ROOM' section with a sub-headline 'Pinch of salt' and a 'RECENT ARTICLES' section.

The provenance crisis



SCIENTIFIC PUBLISHING

A Scientist's Nightmare: Software Problem Leads to Five Retractions

BBC News
NEWS
Page last updated at 14:56 GMT, Tuesday, 1 December 2009
'Show Your Working': What 'ClimateGate' means
VIEWPOINT
Mike Hulme and Jerome Ravetz
The "ClimateGate" affair - the publication of e-mails and documents hacked or leaked from one of the world's leading climate research institutions - is being intensely debated on the web. But what does it imply for climate science? Here, Mike Hulme and Jerome Ravetz say it shows that we need a more concerted effort to explain and engage the public in understanding the processes and practices of science and scientists.
THE GREEN ROOM
A weekly series of thought-provoking pieces on environmental topics
Pinch of salt
Idea that the world production must do wrong
RECENT ARTICLES

nature
www.nature.com/nature Vol 442 | Issue no. 7098 | 6 July 2006
Illuminating the black box
Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.
This journal aims to publish papers that are not only interesting and thought-provoking, but reproducible and useful. In order to do this, novel materials and reagents need to be carefully described and readily available to interested scientists.
That might seem obvious. But despite the efforts of our editors and referees, papers in the biological sciences are still being submitted — and often the reagents used in the practice, we can see from researchers' comments. Some of these technologies, for example involving the production of antibodies, established didn't want the author to reveal the sequences, as this would jeopardize its *raison d'être*. This kind of stalemate matters, because it prevents the replication of experiments and inhibits the selection of appropriate controls in subsequent work.
Some authors claim replication is possible without full sequence information or the details of novel compounds. They say that the

Science POLICYFORUM
COMPUTER SCIENCE
Accessible Reproducible Research
Jill P. Mesirov
As use of computation in research grows, new tools are needed to expand recording, reporting, and reproduction of methods and data.

The New York Times Science
NYTimes: Home - Site Index - Archive - Help
Nobel Laureate Retracts Two Papers Unrelated to Her Prize
By KENNETH CHANG
Published: September 23, 2010
Linda B. Buck, who shared a 2004 Nobel Prize in Physiology or Medicine, apologized for

Where to start?

Where to start?

Definition 4.1 (Tuple Derivation for an Operator). Let Op be any relational operator over tables T_1, \dots, T_m , and let $T = Op(T_1, \dots, T_m)$ be the table that results from applying Op to T_1, \dots, T_m . Given a tuple $t \in T$, we define t 's derivation in T_1, \dots, T_m according to Op to be $Op_{\langle T_1, \dots, T_m \rangle}^{-1}(t) = \langle T_1^*, \dots, T_m^* \rangle$, where T_1^*, \dots, T_m^* are maximal subsets of T_1, \dots, T_m such that

(a) $Op(T_1^*, \dots, T_m^*) = \{t\}$.

(b) $\forall T_i^* : \forall t^* \in T_i^* : Op(T_1^*, \dots, \{t^*\}, \dots, T_m^*) \neq \emptyset$.

We also say that $Op_{T_i}^{-1}(t) = T_i^*$ is t 's derivation in T_i , and each tuple t^* in T_i^* contributes to t , for $i = 1..m$.

Where to start?

Definition 6. (Witness Basis) Consider a normal form query Q . The *witness basis* for a singular value t with respect to Q and D , denoted as $W_{Q,D}(t)$, is:

- (1) If Q is of the form $Q_1 \sqcup \dots \sqcup Q_n$ then $W_{Q,D}(t) = W_{Q_1,D}(t) \cup \dots \cup W_{Q_n,D}(t)$.
- (2) If Q is of the form $\{e \mid p_0 \in e_0, \dots, p_n \in e_n, \text{condition}\}$, let Ψ be the set of all valuations on the variables of Q such that “where” clause of Q holds under each valuation in Ψ . Then, $W_{Q,D}(t) = \{\llbracket p_0 \rrbracket_\psi \sqcup \dots \sqcup \llbracket p_n \rrbracket_\psi \mid \psi \in \Psi, t = \llbracket e \rrbracket_\psi\}$. Note that e_i ($0 \leq i \leq n$) is a database constant since Q is in normal form.
- (3) Otherwise, $W_{Q,D}(t) = \{\}$.

More generally, for any well-formed query Q , we can define the witness basis by extending (2) as follows. We partition the set of $p_i \in e_i$ in the “where” clause of Q into two parts: $S_1 = \{p_i \mid e_i \text{ is the database constant } D\}$ and $S_2 = \{(p_i, e_i) \mid p_i \text{ is a pattern matched against a query } e_i\}$. We use p_0^1, \dots, p_k^1 to denote the members of S_1 and $(p_0^2, e_0^2), \dots, (p_m^2, e_m^2)$ to denote the members of S_2 . Let Ψ be the set of all valuations on the variables of Q such that for each valuation in Ψ , “where” clause of Q holds. Then $W_{Q,D}(t) = \{P_1 \sqcup P_2 \mid \psi \in \Psi, t \sqsubseteq \llbracket e \rrbracket_\psi, P_1 = \llbracket p_0^1 \rrbracket_\psi \sqcup \dots \sqcup \llbracket p_k^1 \rrbracket_\psi, P_2 = w_1 \sqcup \dots \sqcup w_m \text{ where } w_i \in W_{\psi(e_i^2), D}(\llbracket p_i^2 \rrbracket_\psi)\}$. For a compound value t , the witness basis is the product of individual witness basis of singular values making up t . That is, consider $t = t_1 \sqcup \dots \sqcup t_m$ where each t_i is singular. Then $W_{Q,D}(t) = \{w_1 \sqcup \dots \sqcup w_m \mid w_i \in W_{Q,D}(t_i)\}$. \square

Def
tiona
the
 $t \in$
 $Op_{\langle T \rangle}$
 $T_1,$

(a)

(b)

We
 T_i^*

Where to start?

Definition 6. (Witness Basis) Consider a normal form query Q . The *witness basis* for $l:v$ where v is an atomic value, denoted as $\Gamma_{Q,D}(l:v)$ with respect to Q and D , is defined as below:

- (1) If $Q = Q_1 \sqcup \dots \sqcup Q_n$ then $\Gamma_{Q,D}(l:v) = \Gamma_{Q_1,D}(l:v) \cup \dots \cup \Gamma_{Q_n,D}(l:v)$.
- (2) If Q has the form $\{e \mid p_0 \in e_0, \dots, p_n \in e_n, \text{condition}\}$, let Ψ be the set of valuations on the variables of Q such that the “where” clause of Q holds under each valuation and $\psi(e)$ contains $l:v$. For each $\psi \in \Psi$, let p_{x_ψ} denote the path in e that points to a variable x_ψ such that there exists p' and p'' so that $l = p'.p''$ and $\psi(p_{x_\psi}) = p'$ and $\psi(x_\psi)(p'') = v$. Then, $\Gamma_{Q,D}(l:v) = \{([\![p_0]\!]_\psi \sqcup \dots \sqcup [\![p_n]\!]_\psi, S) \mid \psi \in \Psi, S = \{\psi(p'_i).p'' \mid p'_i \text{ is the path that points to variable } x_\psi \text{ in pattern } p_i, 0 \leq i \leq n\}\}$.
- (3) Otherwise, $\Gamma_{Q,D}(l:v) = \{\}$.

More generally, the derivation basis of $l:v$ where v is a compound value is defined to be the derivation basis of all possible (path,value) pairs $p':v'$ such that $p':v'$ points to a value in v . The derivation basis for multiple (path,value) pairs is defined to be the product of the derivation basis of individual (path,value) pairs. That is, $\Gamma_{Q,D}(p_1:v_1, p_2:v_2) = \Gamma_{Q,D}(p_1:v_1) * \Gamma_{Q,D}(p_2:v_2) = \{(w_1 \sqcup w_2, P_1 \cup P_2) \mid (w_1, P_1) \in \Gamma_{Q,D}(p_1:v_1), (w_2, P_2) \in \Gamma_{Q,D}(p_2:v_2)\}$. \square

Def
tiona
the
 $t \in$
 $Op_{\langle T \rangle}$
 $T_1,$
(a) Mo
(b) by
clau
 S_2
We
 T_i^*
 $S_2.$
valu
 $[e]_\psi$
For
basi
each

Where to start?

Definition 6. (Witness Basis) Consider a normal form query Q . The *witness basis* for $l:v$ where v is an atomic value is defined as follows:

Definition 8. (Derivation Basis) Consider a normal form query Q and a derivation D . The *derivation basis* for $l:v$ where v is an atomic value is defined as below:

- (1) If $Q = Q_1 \sqcup \dots \sqcup Q_n$ then $\Gamma_{Q,D}(l:v) = \bigcup_{i=1}^n \Gamma_{Q_i,D}(l:v)$.
- (2) If Q has the form $\{e \mid p_0 \in e_0, \dots, p_n \in e_n\}$ then $\Gamma_{Q,D}(l:v) = \{([p_0]_\psi \sqcup \dots \sqcup [p_n]_\psi, S) \mid \psi \in \Psi, S = \{\psi(p'_i).p'' \mid p'_i \text{ is the path that points to variable } x_\psi \text{ in pattern } p_i, 0 \leq i \leq n\}\}$.
- (3) Otherwise, $\Gamma_{Q,D}(l:v) = \{\}$.

Not compositional

More generally, the derivation basis of $l:v$ where v is a compound value is defined to be the derivation basis of all possible (path,value) pairs $p':v'$ such that $p':v'$ points to a value in v . The derivation basis for multiple (path,value) pairs is defined to be the product of the derivation basis of individual (path,value) pairs. That is, $\Gamma_{Q,D}(p_1:v_1, p_2:v_2) = \Gamma_{Q,D}(p_1:v_1) * \Gamma_{Q,D}(p_2:v_2) = \{(w_1 \sqcup w_2, P_1 \cup P_2) \mid (w_1, P_1) \in \Gamma_{Q,D}(p_1:v_1), (w_2, P_2) \in \Gamma_{Q,D}(p_2:v_2)\}$. \square

Def
tiona
the
 $t \in$
 $Op_{\langle T \rangle}$
 $T_1,$
(a) Mo
by
(b) cla
 S_2
We
 T_i^*
 $S_2.$
valu
 $[e]_\psi$
For
basi
each

Where to start?

Definition 6. (Witness Basis) Consider a normal form query Q . The *witness basis* for $l:v$ where v is an atomic value is defined as follows:

(1) **Definition 8. (Derivation Basis)** Consider a normal form query Q and D , is defined as below:

(2) *Derivation basis* for $l:v$ where v is an atomic value to Q and D , is defined as below:

(1) If $Q = Q_1 \sqcup \dots \sqcup Q_n$ then $\Gamma_{Q,D}(l:v)$ is the set of the form $\{e \mid p_0 \in e_0, \dots, p_n \in e_n\}$ where e_i is a path that points to a variable x_i of Q such that the "where" clause of Q holds and $\psi(e)$ contains $l:v$. For each $\psi \in \Psi$, let p_{x_ψ} denote the path that points to a variable x_ψ such that there exists p' and p'' such that $\psi(p_{x_\psi}) = p'$ and $\psi(x_\psi)(p'') = v$. Then, $\Gamma_{Q,D}(l:v) = \{\psi(p_{x_\psi}, S) \mid \psi \in \Psi, S = \{\psi(p'_i) \cdot p'' \mid p'_i \text{ is the path that points to pattern } p_i, 0 \leq i \leq n\}\}$.

More generally, the derivation basis of $l:v$ where v is a compound value is defined to be the derivation basis of all possible (path,value) pairs $p':v'$ such that $p':v'$ points to a value in v . The derivation basis for multiple (path,value) pairs is defined to be the product of the derivation basis of individual (path,value) pairs. That is, $\Gamma_{Q,D}(p_1:v_1, p_2:v_2) = \Gamma_{Q,D}(p_1:v_1) * \Gamma_{Q,D}(p_2:v_2) = \{(w_1 \sqcup w_2, P_1 \cup P_2) \mid (w_1, P_1) \in \Gamma_{Q,D}(p_1:v_1), (w_2, P_2) \in \Gamma_{Q,D}(p_2:v_2)\}$. \square

Hard to separate "policy" from "mechanism"

Not compositional

Where to start?

Definition 6. (Witness Basis) Consider a normal form query Q . The *witness basis* for $l:v$ is defined as follows:

(1) **Definition 8. (Derivation Basis)** Consider a normal form query Q and a value v . A *derivation basis* for $l:v$ where v is an atomic value is a set of paths P such that $P \vdash Q$ and D , is defined as below:

(1) If $Q = Q_1 \sqcup \dots \sqcup Q_n$ then $\Gamma_{Q,D}(l:v)$ is defined as follows:

(1) If $Q = Q_1 \sqcup \dots \sqcup Q_n$ then $\Gamma_{Q,D}(l:v)$ is the set of paths P of the form $\{e \mid p_0 \in e_0, \dots, p_n \in e_n\}$ such that p_i is a path to a variable x_i of Q such that x_i is a variable of Q such that the "where" clause of Q holds and $\psi(e)$ contains l . For each $\psi \in \Psi$, let p_{x_ψ} denote the path to a variable x_ψ such that there exists p' and p'' such that $\psi(p_{x_\psi}) = p'$ and $\psi(x_\psi) = p''$.

Hard to separate "policy" from "mechanism"

Not compositional

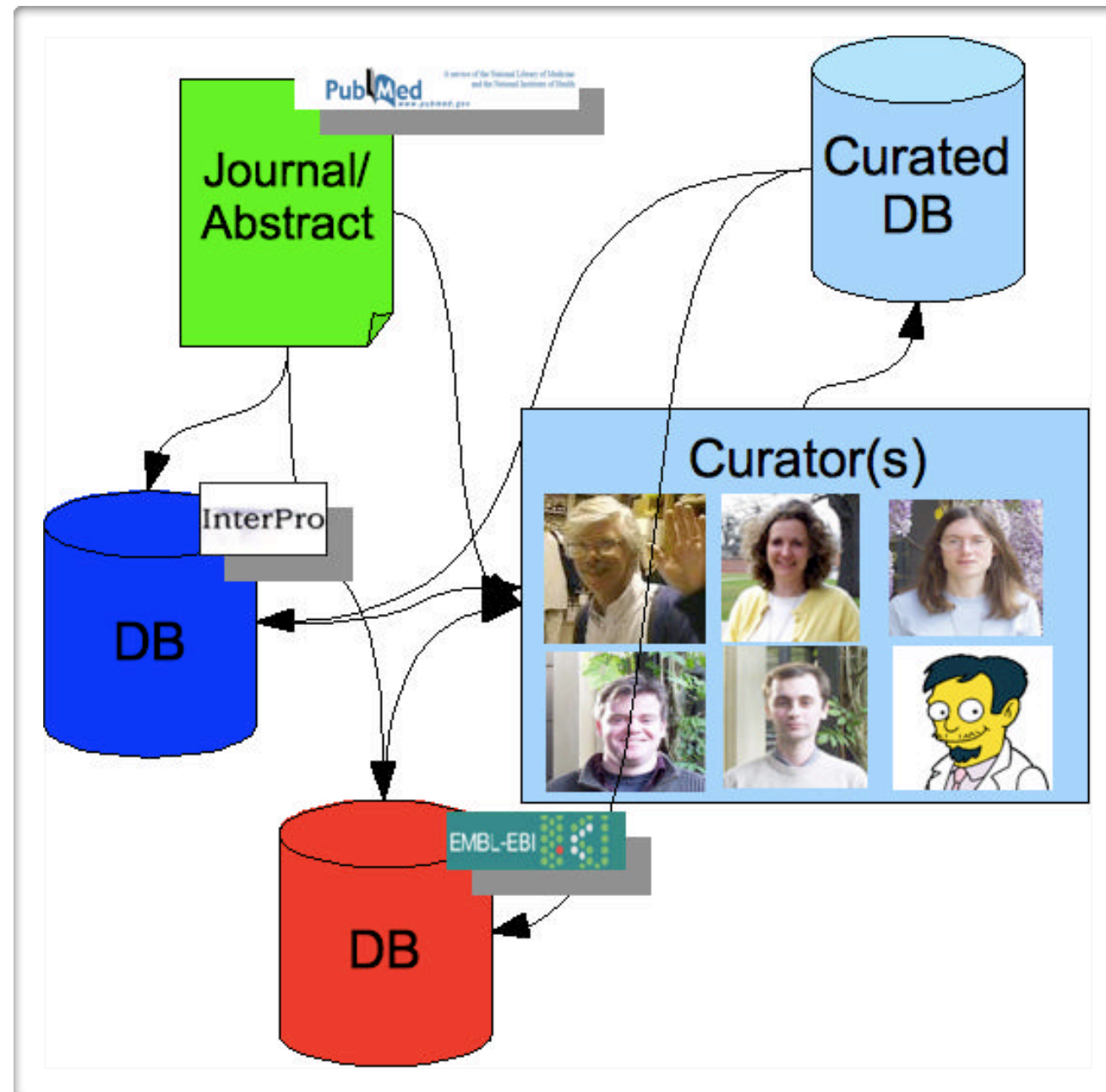
Hard to implement

More generally, the derivation basis of $l:v$ where v is a compound value is defined to be the derivation basis of all possible $(\text{path}, \text{value})$ pairs $p':v'$ such that $p':v'$ points to a value in v . The derivation basis for multiple $(\text{path}, \text{value})$ pairs is defined to be the product of the derivation basis of individual $(\text{path}, \text{value})$ pairs. That is, $\Gamma_{Q,D}(p_1:v_1, p_2:v_2) = \Gamma_{Q,D}(p_1:v_1) * \Gamma_{Q,D}(p_2:v_2) = \{(w_1 \sqcup w_2, P_1 \cup P_2) \mid (w_1, P_1) \in \Gamma_{Q,D}(p_1:v_1), (w_2, P_2) \in \Gamma_{Q,D}(p_2:v_2)\}$. \square

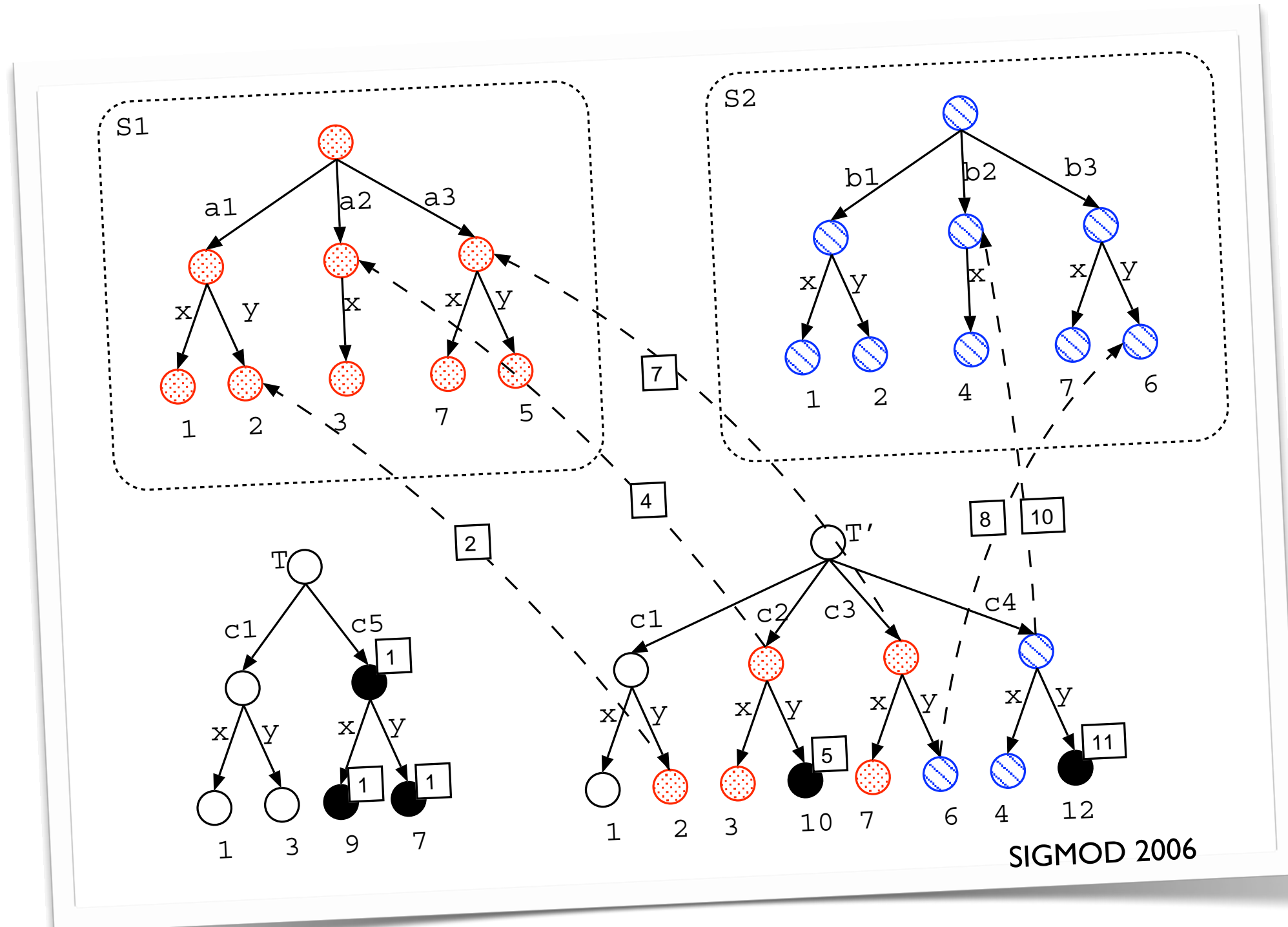
Databases and programming languages

- Database query languages **are** purely functional
 - optimization by equational rewriting basis of £10⁹ DB industry
- Programming languages ideas can...
 - Help in analyzing, optimizing database queries (types, compilation, equational rewriting)
 - Integrate database or Web capabilities into higher-level languages (LINQ, Links)
- Database ideas can...
 - Lead to new programming idioms (Datalog, atomicity, STM)
 - Open up new problem spaces (high-level updates, **provenance**)

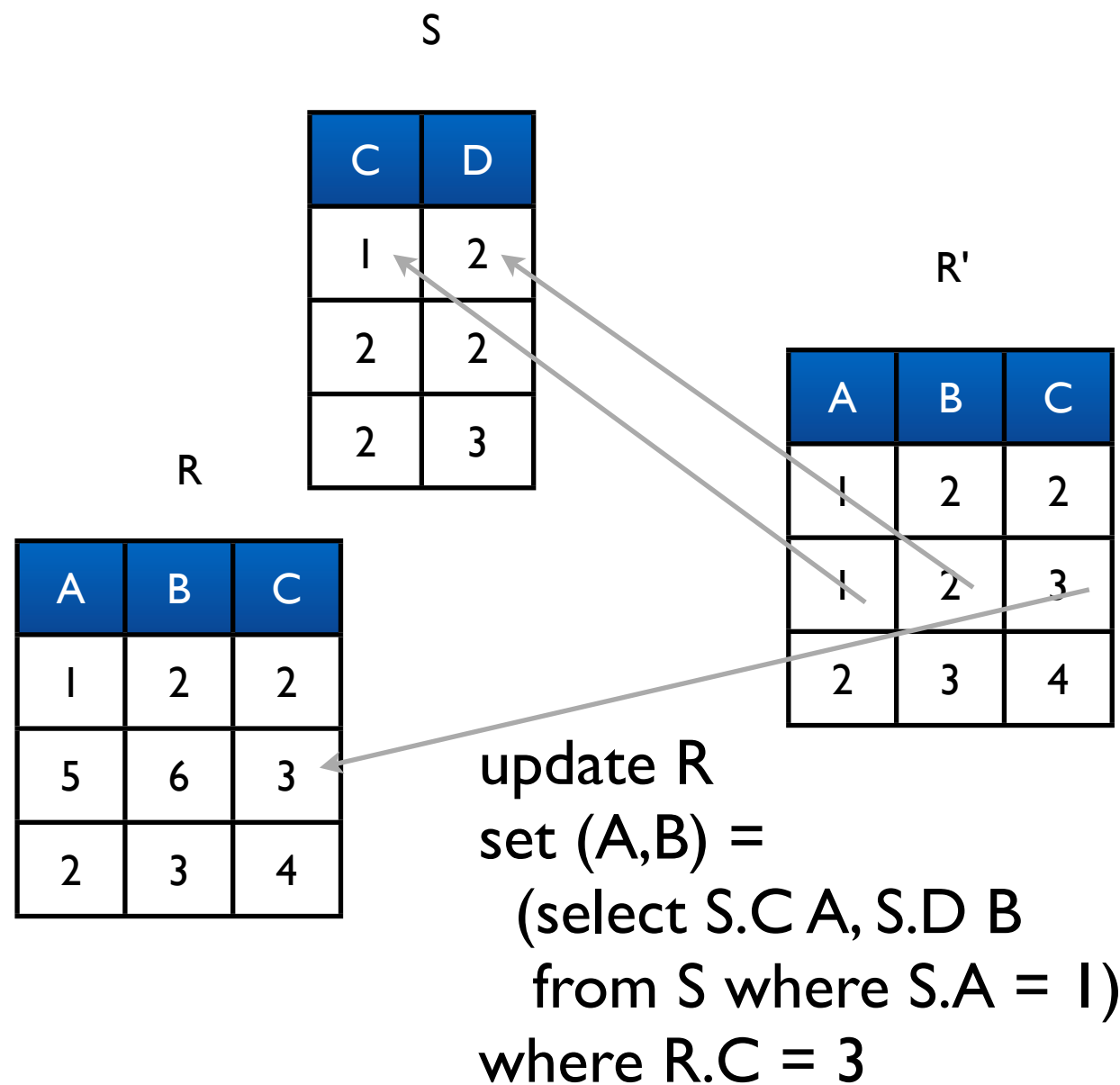
Provenance in curated databases



Provenance in curated databases



Provenance in curated databases



ICDT 2007/TODS 2008

Formalization

- Consider types:

$T ::= \text{int} \mid \dots \mid T * T \mid T \text{ set}$

- And expressions:

- $e ::= x \mid \text{let } x = e_1 \text{ in } e_2 \mid i \mid \dots$

- $\mid (e_1, e_2) \mid \pi_i(e)$

- $\mid \emptyset \mid \{e\} \mid e_1 \cup e_2 \mid \cup\{e_2 \mid x \leftarrow e_1\}$

Type translation

- Translate:

$$P[\text{int}] = \text{int} * 'a \text{ option}$$

$$P[T_1 * T_2] = P[T_1] * P[T_2] * 'a \text{ option}$$

$$P[T \text{ set}] = P[T] \text{ set} * 'a \text{ option}$$

Annotations 'a represent "pointers" to optional sources

Term Translation

Given $x_1:T_1, \dots, x_n:T_n \vdash e : T$

Want $P[e]$ such that

$x_1: P[T_1] \dots x_n: P[T_n] \vdash P[e] : P[T]$

s.t. each SOME-pointer points to the "source"

Simple cases:

$$P[x] = x$$

$$P[i] = (i, \text{NONE})$$

$$P[\text{let } x = e_1 \text{ in } e_2] = \text{let } x = P[e_1] \text{ in } P[e_2]$$

Pairs

$$P[(e_1, e_2)] = (P[e_1], P[e_2], \text{NONE})$$

$$P[\pi_i(e)] = \pi_i(\pi_1(P[e]))$$

Sets

$$P[\emptyset] = (\emptyset, \text{NONE})$$

$$P[e_1 \cup e_2] = \text{let } (v_1, _)$$

$$(v_2, _) = P[e_2]$$

$$\text{in } (v_1 \cup v_2, \text{NONE})$$

$$P[\{e\}] = (\{P[e]\}, \text{NONE})$$

$$P[\cup\{e_2 \mid x \leftarrow e_1\}] = (\cup\{P[e_2] \mid x \leftarrow \pi_1(P[e_1])\}, \text{NONE})$$

Dependency provenance

[CAA DBPL07, MSCS11]

Teams							Players					Result	
name	team	gp	win	loss	tie	otl	name	team	gp	g	a	name	g
string	string	int	int	int	int	int	string	string	int	int	int	string	int
name	team	gp	win	loss	tie	otl	name	team	gp	g	a	name	g
Ottawa Senators	OTT	65	36	29	0	6	Malkin	PIT	65	36	48	Iginla	38
Dallas Stars	DAL	68	41	27	0	5	Ovechkin	WAS	64	48	34		
Detriot Red Wings	DET	65	42	23	0	6	Lecavalier	TAM	63	32	46		
Philadelphia Flyers	PHI	64	32	32	0	7	Iginla	CGY	64	38	37		
Pittsburgh Penguins	PIT	65	36	29	0	7	Spezza	OTT	59	25	50		
Calgary Flames	CGY	64	33	31	0	9	Alfredsson	OTT	~42	35	40		
Washington Capitals	WAS	64	29	35	0	8	Datsyuk	DET	65	23	51		
Tampa Bay Lightning	TAM	63	25	38	0	7	Ribeiro	DAL	65	26	48		
Colorado Avalanche	COL	65	33	32	0	6	Zetterberg	DET	58	35	38		
							St Louis	TAM	63	21	50		
							Richards	PHI	62	23	43		
							Stastny	COL	50	19	33		

cf. Dependency Core Calculus [Abadi et al. 1999]

Dependency provenance

[CAA DBPL07.MSCS11]

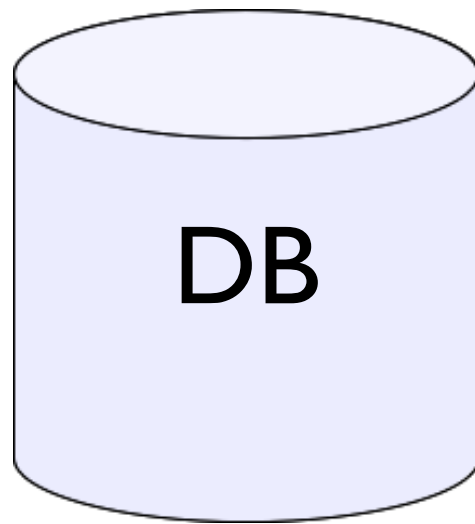
Teams							Result			
name	team	gp	win	loss	tie	of	name	g	g	
string	string	int	int	int	int	int	string	int	int	
Ottawa Senators	OTT	65	36	29	0	6	Ovechkin	WAS	64	48
Dallas Stars	DAL	68	41	27	0	5	Lecavalier	TAM	63	32
Detriot Red Wings	DET	65	42	23	0	6	Iginla	CGY	64	38
Philadelphia Flyers	PHI	64	32	32	0	7	Spezza	OTT	59	25
Pittsburgh Penguins	PIT	65	36	29	0	7	Alfredsson	OTT	~42	35
Calgary Flames	CGY	64	33	31	0	9	Datsyuk	DET	65	23
Washington Capitals	WAS	64	29	35	0	8	Ribeiro	DAL	65	26
Tampa Bay Lightning	TAM	63	25	38	0	7	Zetterberg	DET	58	35
Colorado Avalanche	COL	65	33	32	0	6	St Louis	TAM	63	21
							Richards	PHI	62	23
							Stastny	COL	50	19

name	g
Iginla	38

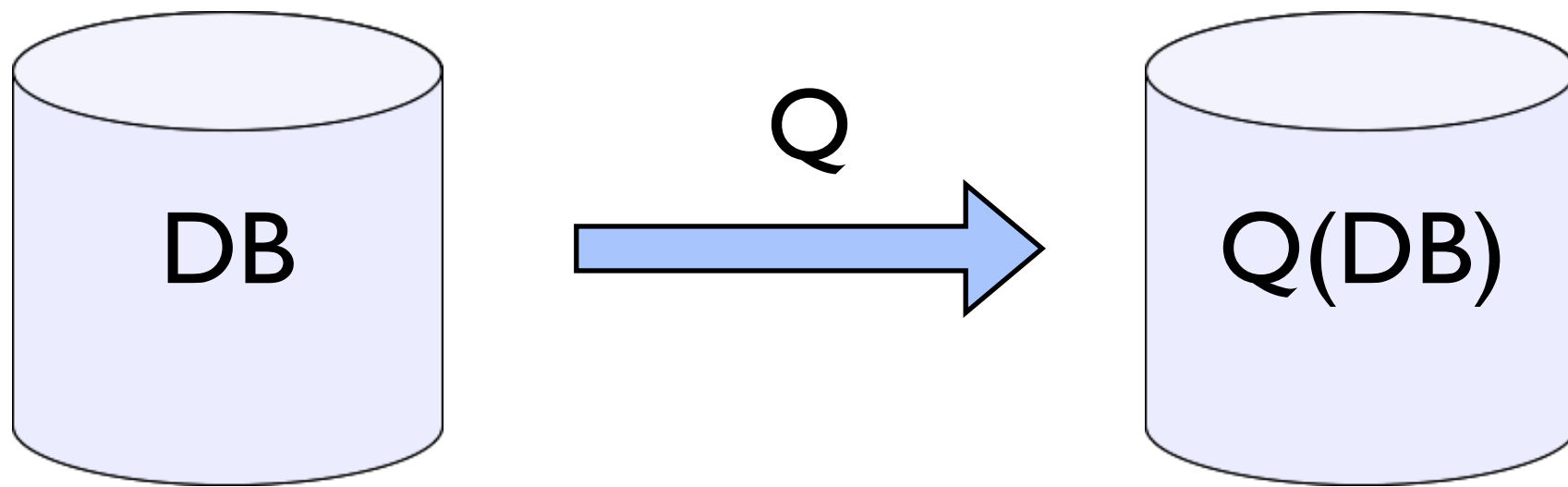
"Good" players on winning teams

cf. Dependency Core Calculus [Abadi et al. 1999]

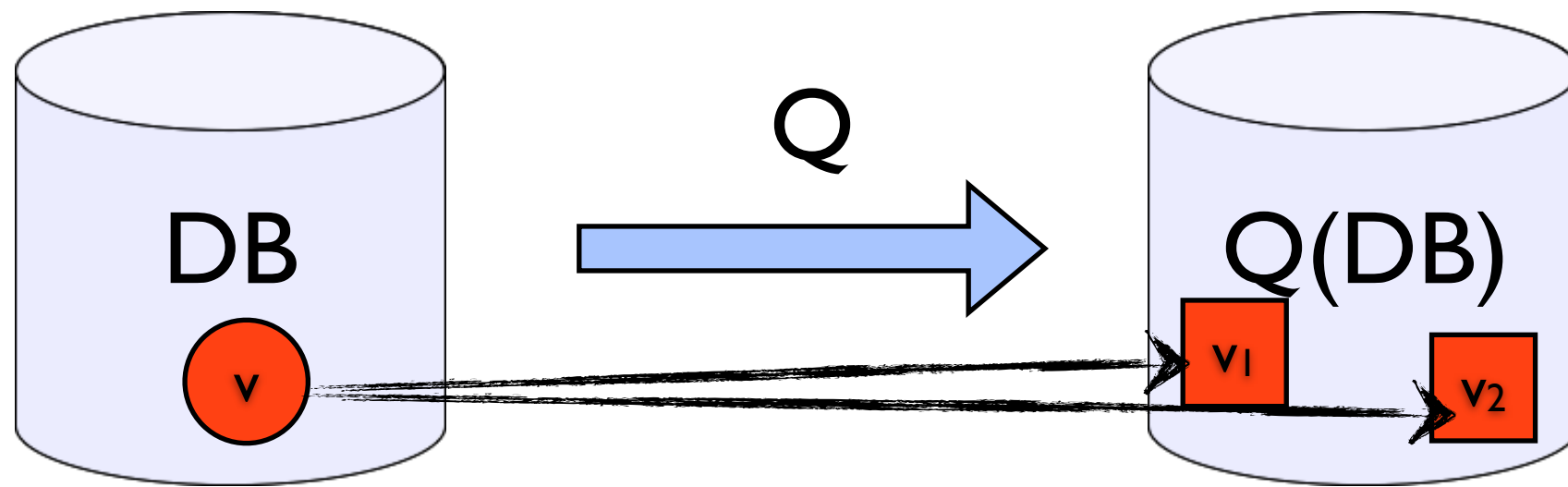
Basic idea



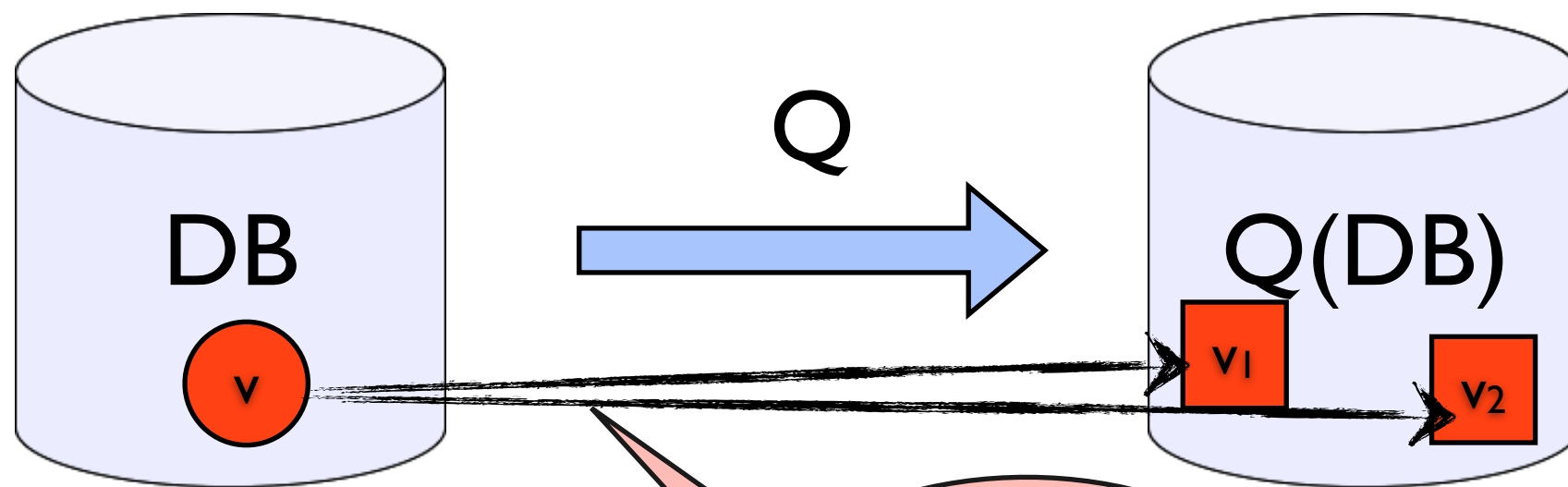
Basic idea



Basic idea

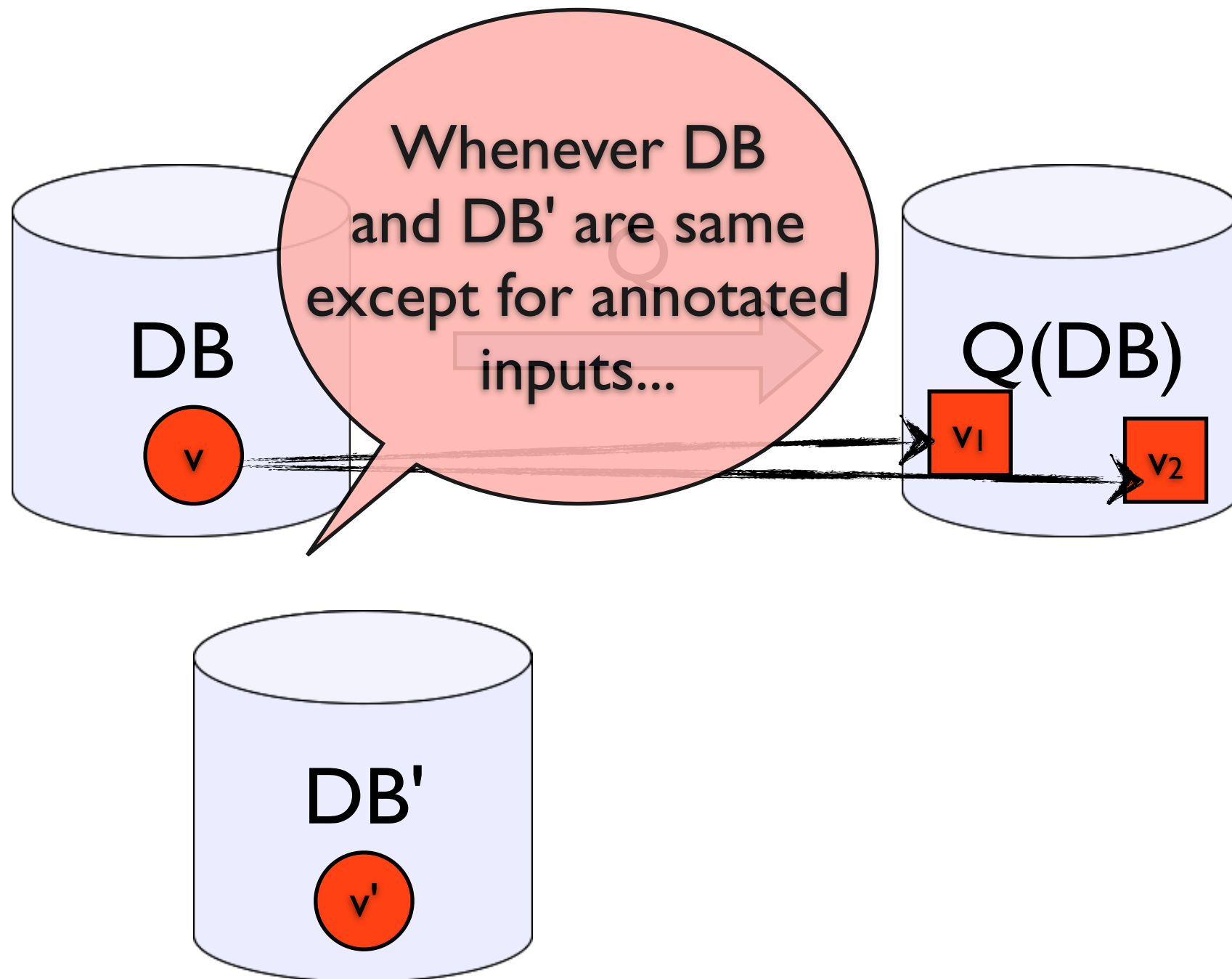


Basic idea

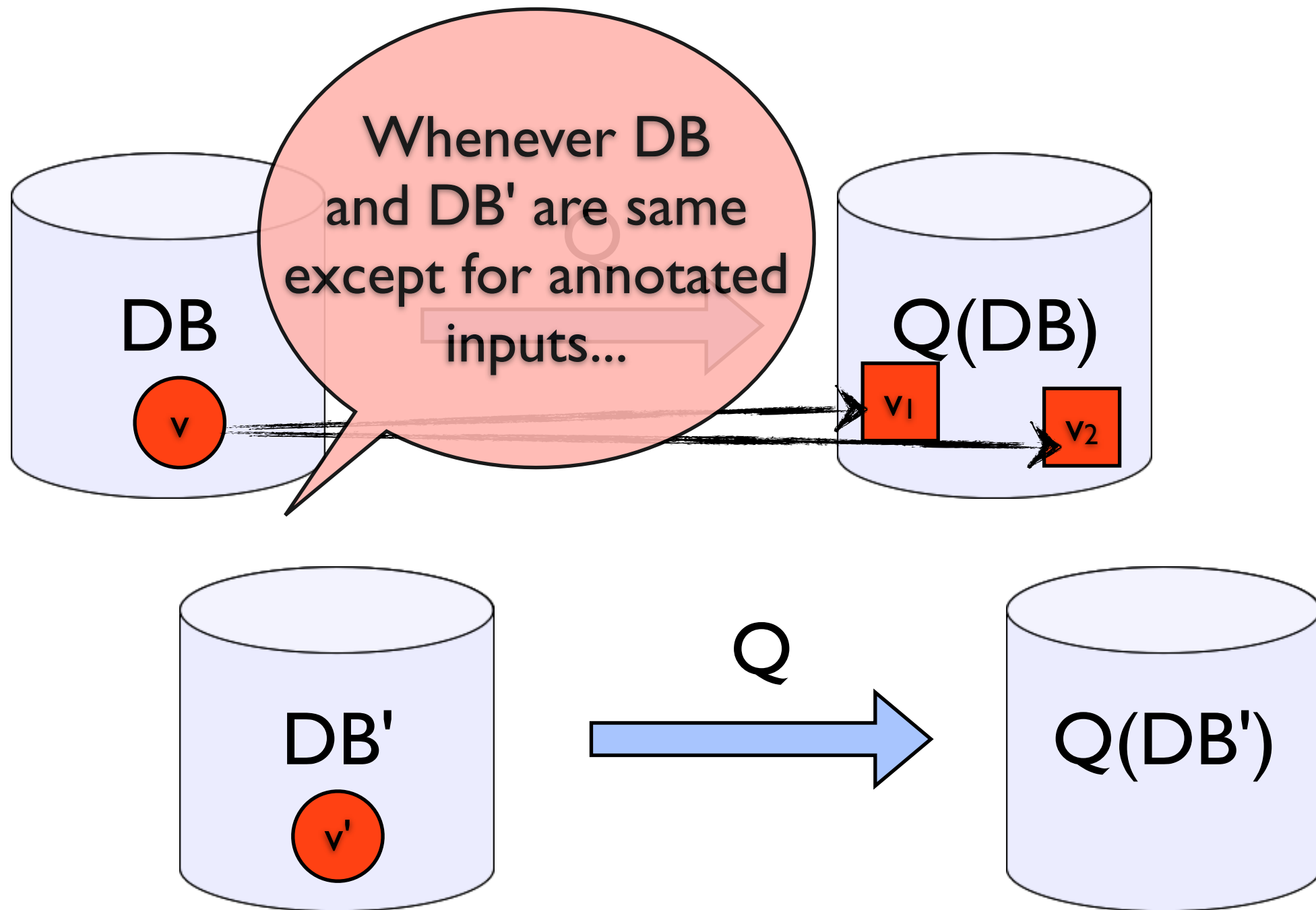


Link output parts
to sets of input parts
such that...

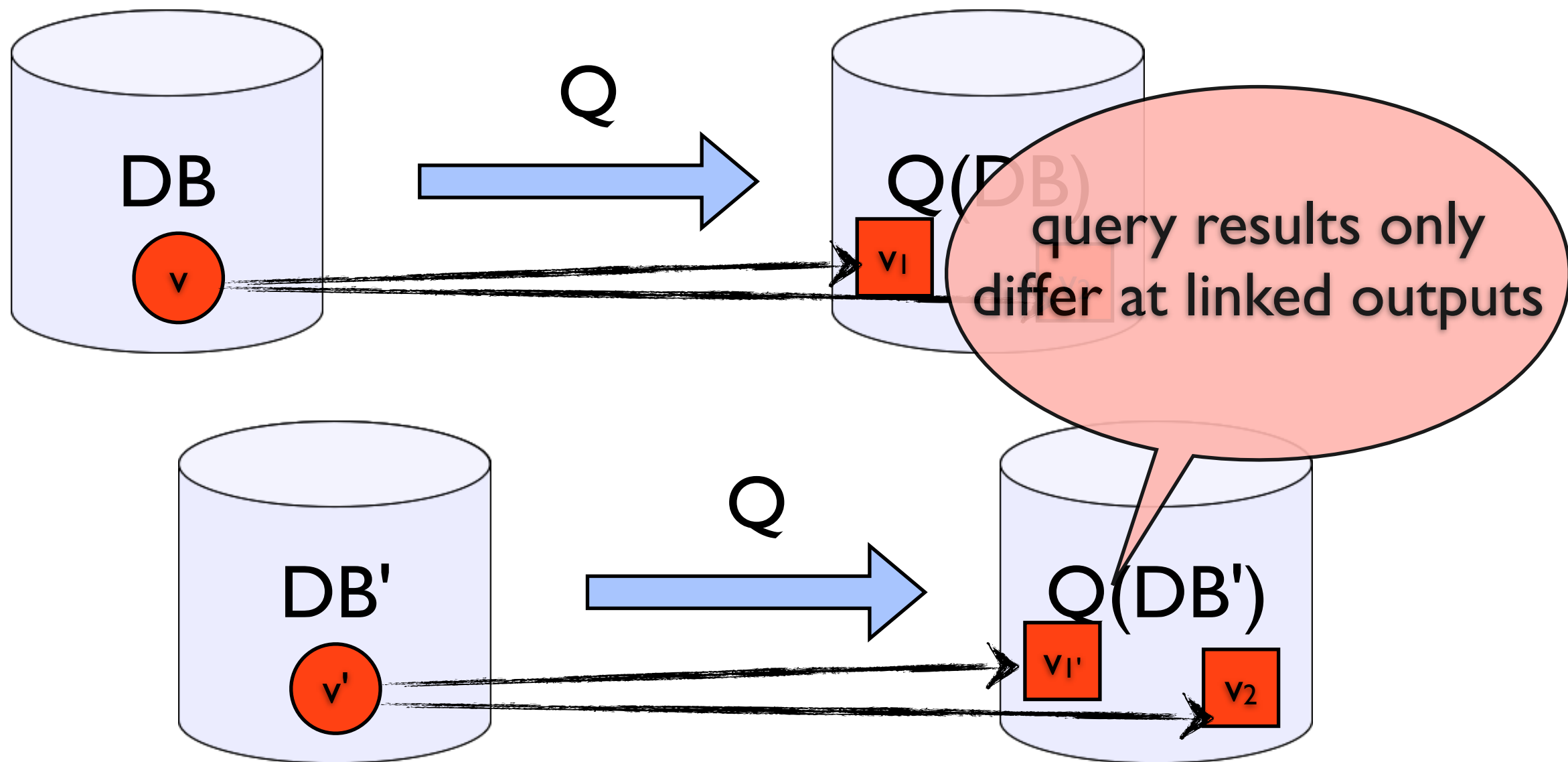
Basic idea



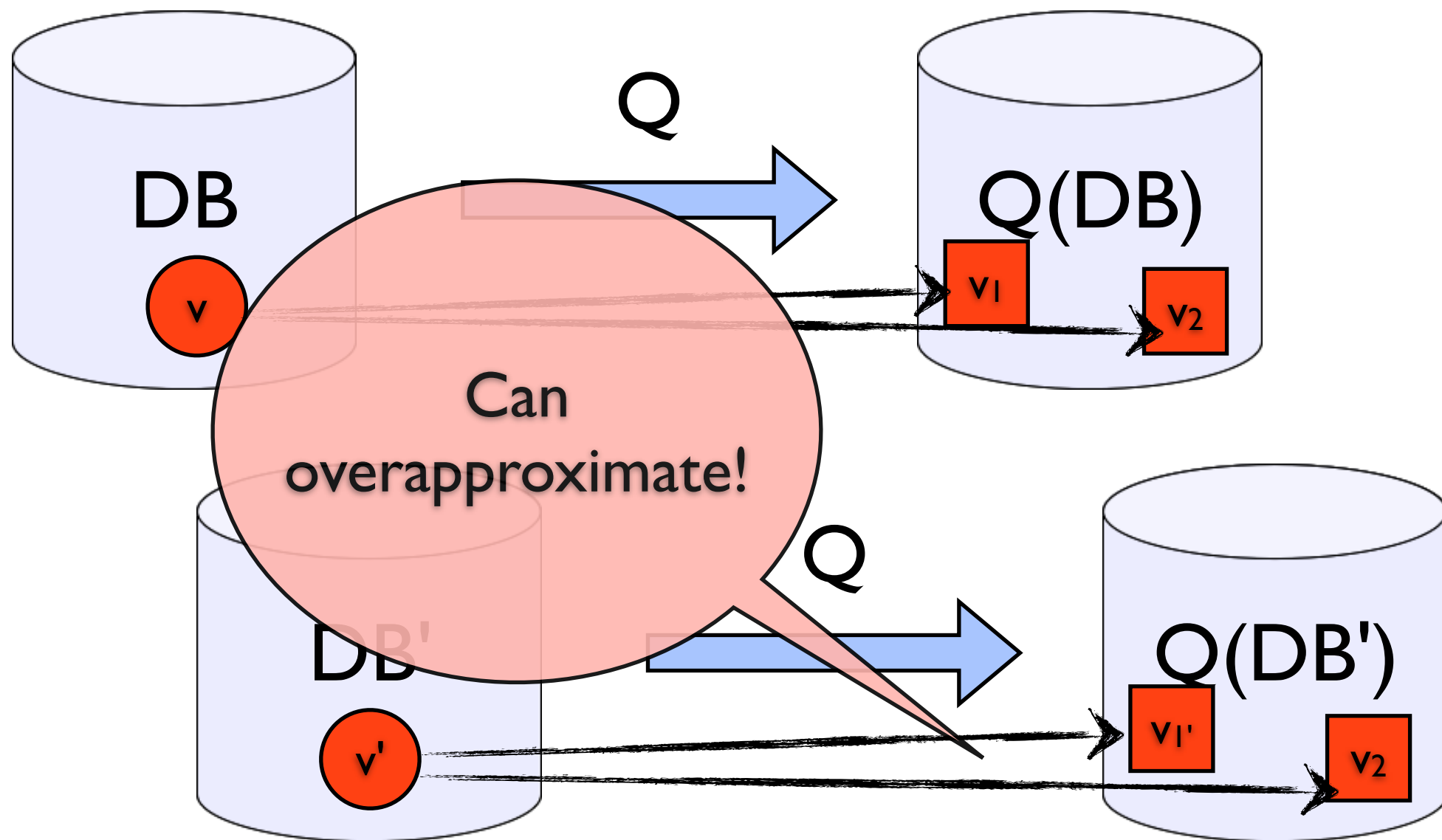
Basic idea



Basic idea



Basic idea



Formalization

- Consider types:

$T ::= \text{int} \mid \dots \mid T * T \mid T \text{ set}$

- Translate:

$P[\text{int}] = \text{int} * 'a \text{ set}$

$P[T_1 * T_2] = P[T_1] * P[T_2] * 'a \text{ set}$

$P[T \text{ set}] = P[T] \text{ set} * 'a \text{ set}$

Annotations 'a represent "pointers" to sets of sources

Term Translation

Given $x_1:T_1, \dots, x_n:T_n \vdash e : T$

Want $P[e]$ such that

$x_1: P[T_1] \dots x_n: P[T_n] \vdash P[e] : P[T]$

s.t. all "dependencies" are captured.

Simple cases:

$$P[x] = x$$

$$P[i] = (i, \emptyset)$$

$$P[\text{let } x = e_1 \text{ in } e_2] = \text{let } x = P[e_1] \text{ in } P[e_2]$$

Pairs

$$P[(e_1, e_2)] = (P[e_1], P[e_2], \emptyset)$$

$$P[\pi_i(e)] = \text{let } (v, a) = P[e] \\ \quad (v_i', b) = \pi_i(v) \\ \text{in } (v_i, a \cup b)$$

Sets

$$P[\emptyset] = (\emptyset, \emptyset)$$

$$P[e_1 \cup e_2] = \text{let } (v_1, a_1) = P[e_1]$$

$$(v_2, a_2) = P[e_2]$$

$$\text{in } (v_1 \cup v_2, a_1 \cup a_2)$$

$$P[\{e\}] = (\{P[e]\}, \emptyset)$$

$$P[\cup\{e_2 \mid x \leftarrow e_1\}] = \text{let } (v, a) = P[e_1]$$

$$\text{in } (\cup\{P[e_2] \mid x \leftarrow v\}, a)$$

Question

- The translations seem to have a lot in common...
- Can we implement them "once and for all"
 - generic/dynamic typing?
 - dependent types?
- Can we implement them in a way that runs efficiently against database?

Links

- Currently supports superset of NRC core-language
 - Higher-order, impure features
 - Effect typing allows safe combination, query extraction [Cooper 2009]
- Ferry [Grust et al. 2010]: extending to support nested data
 - number of queries bounded by types, not data

Generic/Dependent Links?

- Ur/WEB also supports some generic web programming
- Would like to write something like this:

```
type family P a Int = (Int, a)
```

```
type family P a (b, c) =  
    (P a b, P a c, a)
```

```
type family P a [b] = ([P a b], a)
```

Dependent/Generic Links?

- Would like to write something like this:

`whereprov :: Exp e t' →`

`Exp (P e (Maybe a))`

`(P t' (Maybe a))`

`whereprov (Const c) =`

`Pair (Const c) Nothing`

`whereprov (Var x) = Var x`

`...`

Other ways forward?

- Haskell: GADTs + type families/type-level computation + HaskellDB?
- Agda: dependent types ✓, but not DB?
- Idris: dependent types ✓; can we implement query normalization & DB communication as a EDSL?
- Ur/Web: Maybe already has enough GP, but still learning
- (ideally: compile Links to another language that has a mature compiler :)

Database Wiki

- Idea: Wiki-like Web interface to (semi)structured data

- Joint work with Buneman, Mueller, Lindley
- Prototype showcases prior research on provenance, archiving, annotation, security
 - to present at workshop on "Biological Wikis" [NETTAB 2010]

CIA World Factbook

Facts about Belgium

There is an interesting note about land use in Belgium:

Land use		
Name	Land use	
Note	includes Luxembourg (2005)	
Name	Text	Rank
arable land	27.42%	
permanent crops	0.69%	
other	71.89%	

se Wiki

ce to (semi)structured data

Mueller, Lindley

research on provenance,
rity

biological Wikis" [NETTAB 2010]

Database Wiki

- Idea: Wiki-like Web interface to (semi)structured data



- Joint work with Buneman, Mueller, Lindley
- Prototype showcases prior research on provenance, archiving, annotation, security
- to present at workshop on "Biological Wikis" [NETTAB 2010]

CIA World Factbook

- Idea:

lata



Page

Title

There is an interesting note about land use in Belgium:

```
?wpath://COUNTRY[NAME='Belgium']/  
CATEGORY[NAME='Geography']/  
PROPERTY[NAME='Land use']?
```

- Joint
- Proto
- archiv
- to p

,

2010]

Save

Database Wiki

- Idea: Wiki-like Web interface to (semi)structured data



- Joint work with Buneman, Mueller, Lindley
- Prototype showcases prior research on provenance, archiving, annotation, security
- to present at workshop on "Biological Wikis" [NETTAB 2010]

Database



- Idea: Wiki-like Web interface to



- Joint work with Buneman, Mue
- Prototype showcases prior rese
archiving, annotation, security
- to present at workshop on "Biolo

CIA World Factbook

Belgium > Geography > Land use

Edit

Name	Land use
Note	includes Luxembourg (2005)
Sub-Property	
Name	arable land
Text	27.42%
Sub-Property	

Database Wiki

- Idea: Wiki-like Web interface to (semi)structured data



Database Wiki
iDEA The University of Guelph

Edit View Settings

CIA World Factbook

Facts about Belgium

There is an interesting note about land use in Belgium:

Land use

Name	Land use
Name	includes Luxembourg (2005)

Name	Text	Rank
arable land	27.42%	
permanent crops	0.69%	
other	71.89%	



Database Wiki
iDEA The University of Guelph

CIA World Factbook

Page

Title Facts about Belgium

There is an interesting note about land use in Belgium:

```
?paths://COUNTRY[NAME="Belgium"]/  
CATEGORY[NAME="Geography"]/  
PROPERTY[NAME="Land use"]?
```

Save



Database Wiki
iDEA The University of Guelph

CIA World Factbook

Belgium - Geography - Land use

Edit

Name Land use

Note includes Luxembourg (2005)

Sub-Property

Name arable land

Text 27.42%

Sub-Property

- Joint work with Buneman, Mueller, Lindley
- Prototype showcases prior research on provenance, archiving, annotation, security
- to present at workshop on "Biological Wikis" [NETTAB 2010]

Conclusions

- Provenance techniques can be defined as "type-dependent types/functions"
- Complex provenance transformations challenging to implement against real DBs
- Combining Links, Ur/WEB or LINQ with generic or dependent typing might be a good way to proceed