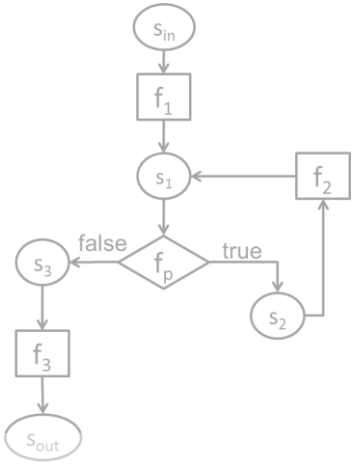
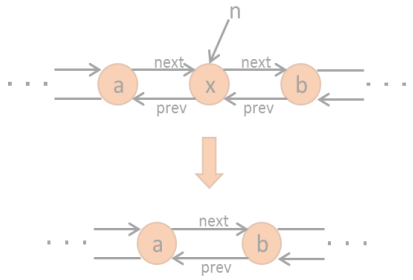
$\exists c \forall in\ Q(c, in)$

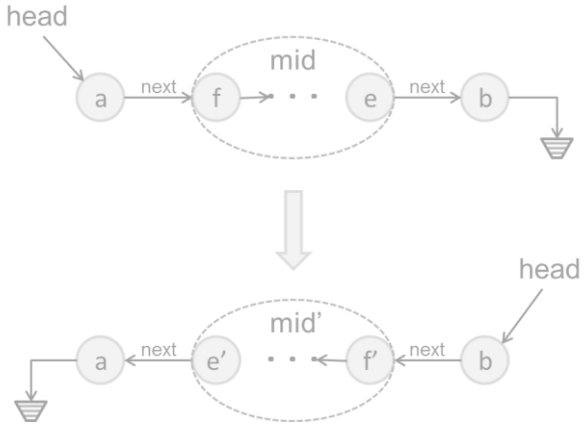
```
/* Average of x and y without using x+y (avoid overflow)*/
int avg(int x, int y){
  int t = expr({x/2, y/2, x%2, y%2, 2 }, {PLUS, DIV});
  assert t == (x+y)/2;
  return t;
}
```

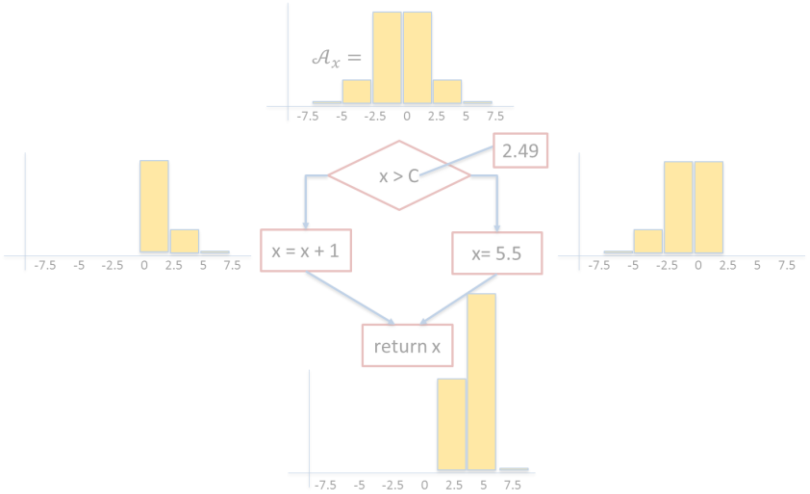





```
{
  s = n.succ;
  p = n.pred;
  p.succ = s;
  s.pred = p;
}
```

Module III: Applications of Synthesis





$\varphi(p)$

$Sk[c](in)$

Logistics

Project presentations

- Tuesday Mar 17, 3-6pm; CSE 2154
- 20 min per team (15 min presentation + questions)
- Structure: motivation, **demo**, technique, evaluation
- Talk to me if you can't make it

Project reports

- Due on Mar 20 (start working on them now!)
- Format: see course organization page (3-5 pages, SIGPLAN format)

Lecture 15
Overview of Applications

Nadia Polikarpova

Applications of synthesis

We have seen:

- End-user spreadsheet programming [FlashFill, BlinkFill]
- Superoptimization [Stoke]

Today:

- Custom data structures
- Data extraction and data wrangling
- Databases

Thursday: synthesis as AI

Next week: synthesis for programmers

Applications of synthesis

→ Custom data structures

- Loncaric, Torlak, Ernst: Fast synthesis of fast collections. PLDI'16

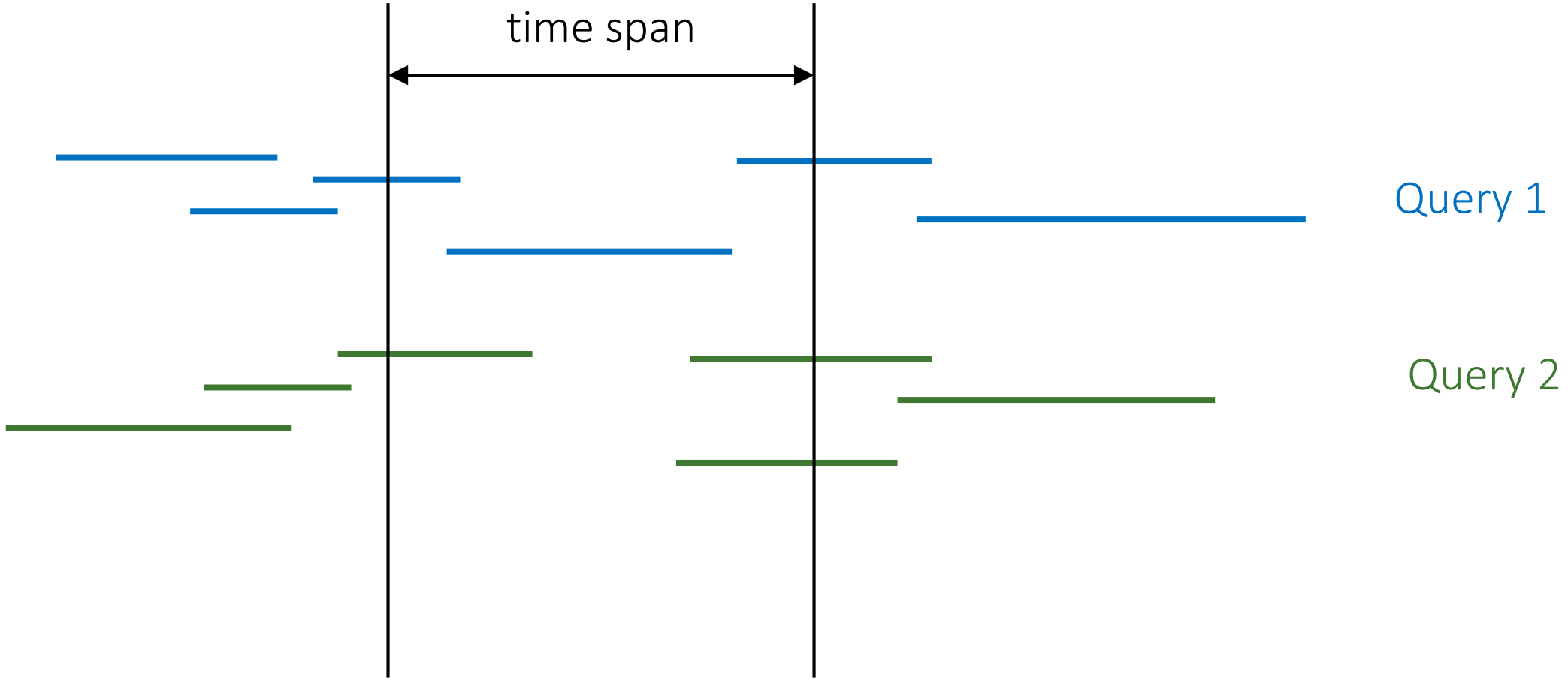
Data extraction and data wrangling

Databases

Fast Synthesis of Fast Collections

[Loncaric et al. PLDI'16]

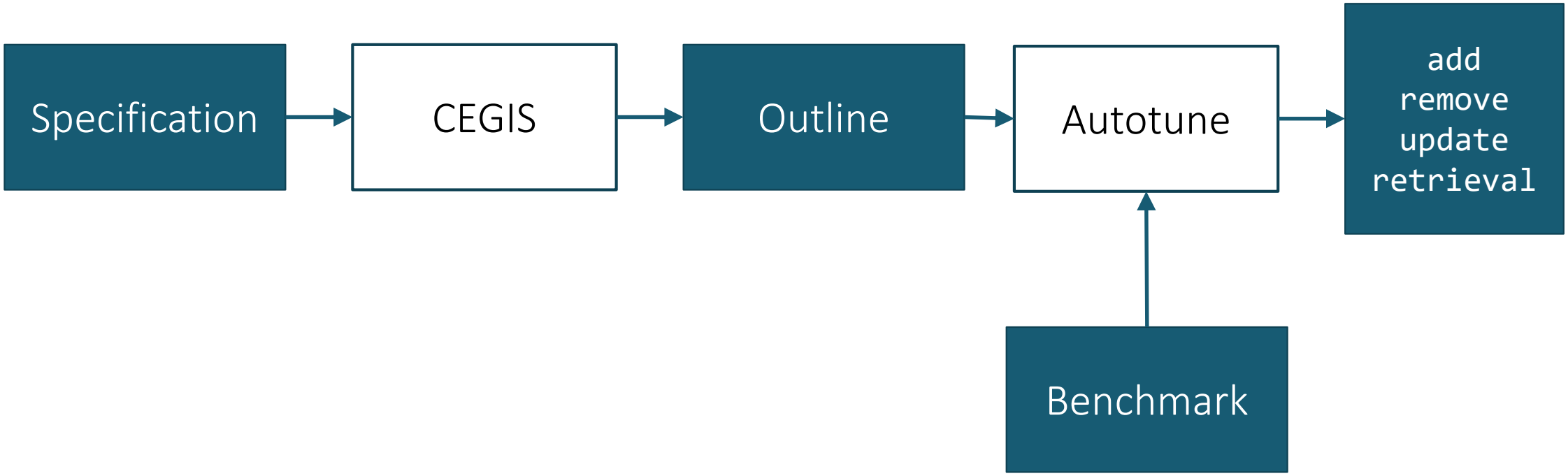
Myria distributed database: needs to retrieve all queries in a given timespan



Specification

```
fields
  queryId:long, subqueryId:long,
  fragmentId:int, opId:int,
  startTime:long, endTime:long,

  assume startTime <= endTime

query getAnalyticsInTimespan(
    v_queryId:long, v_subqueryId:long,
    v_fragmentId:int,
    v_start:long, v_end:long)
  assume v_start <= v_end

  queryId == v_queryId and
  subqueryId == v_subqueryId and
  fragmentId == v_fragmentId and
  startTime < v_end and
  endTime >= v_start

costmodel myria-cost.java
```

The Cozy workflow



Results

Match performance of hand-written code on four real-world applications:

- Myria (a distributed data-base)
- Bullet (a physics simulation library)
- ZTopo (a topographic map viewer)
- Sat4J (a SAT solver)

Applications of synthesis

Custom data structures

→ Data extraction and data wrangling

- Vu Le, Sumit Gulwani: FlashExtract: a framework for data extraction by examples. PLDI’14
- Inala, Singh: WebRelate: integrating web data with spreadsheets using examples. POPL’17
- Feng, Martins, Van Geffen, Dillig, Chaudhuri: Component-based synthesis of table consolidation and transformation tasks from examples. PLDI'17

Databases

FlashExtract

[Le, Gulwani. PLDI'14]

Problem: extract data from semi-structured sources (e.g. log file) into a list of records

User input:

- output schema
- highlights examples of fields

Search strategy: VSA

```
DLZ - Summary Report
"Sample ID:,""5007-01"""
"Sample Date/Time:,""Wednesday, May 30, 2006 00:43:51"""
Intensities
"I/S,""Analyte"",""Mass"",""Conc. Mean"",""Unit"",""Conc. SD"",""RSD"",""Mean"""
"|-,""Be"",9,0.070073,""ug/L"",0.009,12.542,121.334"
"|>,""Sc"",45,,""ug/L"",,,404615.043"
"| ,""Ti"",48,10.653153,""ug/L"",0.847,7.949,181379.200"
"|-,""Se"",82,1.009204,""ug/L"",0.026,2.613,457.487"
"|-,""Sr"",88,20.163079,""ug/L"",2.005,9.943,718014.023"
"|>,""Rh"",103,,""ug/L"",,,438976.176"
DLZ - Summary Report
"Sample ID:,""5007-02"""
"Sample Date/Time:,""Wednesday, May 30, 2006 01:02:38"""
Intensities
"I/S,""Analyte"",""Mass"",""Conc. Mean"",""Unit"",""Conc. SD"",""RSD"",""Mean"""
"| ,""Mn"",55,71.705740,""ug/L"",0.350,0.489,2428667.736"
"| ,""Co"",59,0.131132,""ug/L"",0.004,3.315,3606.816"
"|-,""Ba"",138,129.339264,""ug/L"",3.088,2.387,4648771.382"
"|-,""Hf"",178,,""ug/L"",,,338359.496"
"| ,""Tl"",205,2.876992,""ug/L"",0.730,25.380,129217.588"
"| ,""Pb"",208,3.671043,""ug/L"",0.026,0.702,228830.402"
```

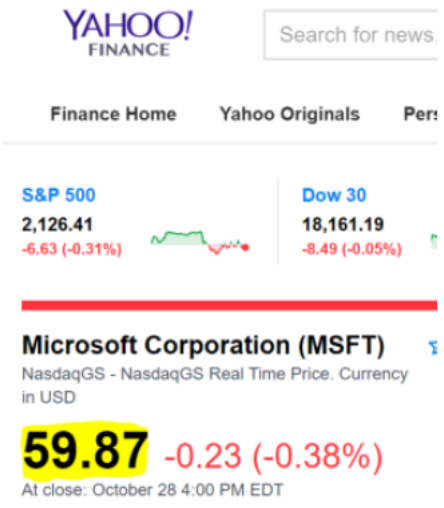
WebRelate

[Inala, Singh. POPL'17]

Problem: extract data from web pages into spreadsheets

User input: navigate to a webpage and select content

	Company	URL	Stock price
1	MSFT	https://finance.yahoo.com/q?s=msft	59.87
2	AMZN	**https://finance.yahoo.com/q?s=amzn**	**775.88**
3	AAPL	**https://finance.yahoo.com/q?s=aapl**	**113.69**
4	TWTR	**https://finance.yahoo.com/q?s=twtr**	**17.66**
5	T	**https://finance.yahoo.com/q?s=t**	**36.51**
6	S	**https://finance.yahoo.com/q?s=s**	**6.31**

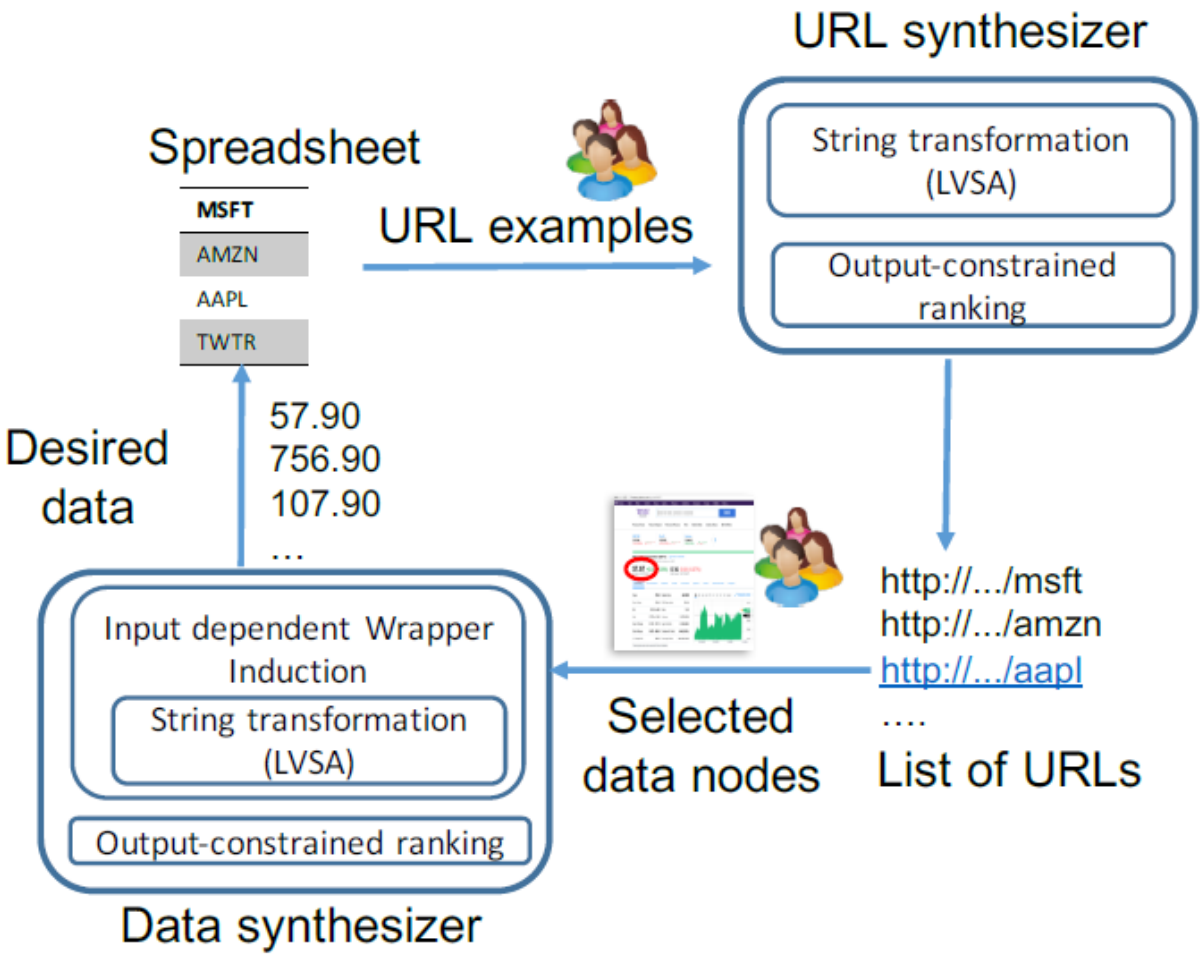


WebRelate

Search strategy: VSA

Optimizations:

- Layered VSA (URLs are too long for FlashFill-style VSAs)
- Output-constrained synthesis: we know the space of possible outputs



Morpheus

[Feng et al. PLDI'17]

Problem: table data wrangling

User input: input-output examples (small tables)

Search strategy: enumerative search with deduction

id	year	A	B
1	2007	5	10
2	2009	3	50
1	2007	5	17
2	2009	6	17

→

id	*A_2007*	*B_2007*	*A_2009*	*B_2009*
1	*5*	*10*	*5*	*17*
2	*3*	*50*	*6*	*17*

Morpheus: TDP with deduction

[Feng et al'17]

Input-output examples

ID	Name	Age
1	Alice	8
2	Bob	18
3	Tom	12

→

ID	Name	Age
2	Bob	18
3	Tom	12

Synthesizer

straight-line program

Components

with partial specifications!

```
select : Table → [Col] → Table
```

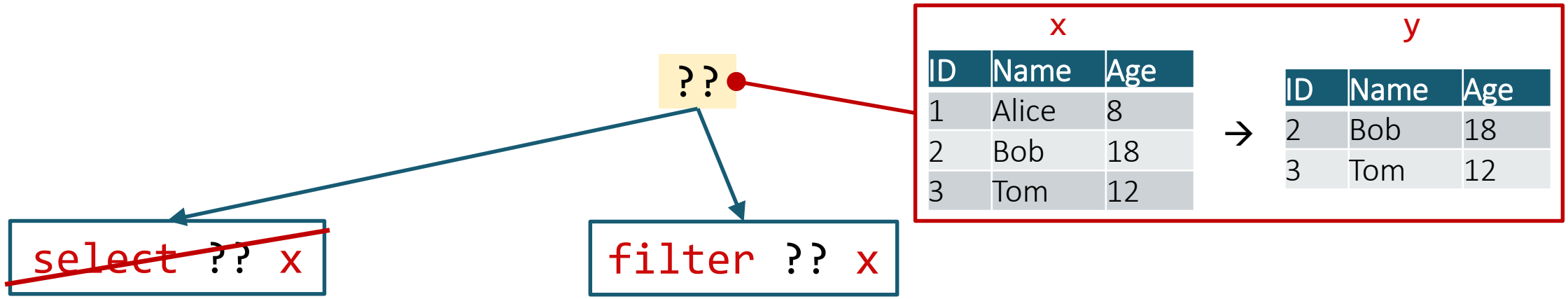
```
out.rows = in.rows
&& out.cols < in.cols
```

```
filter : Table → (Row → Bool) → Table
```

```
our.rows < in.rows
&& out.cols = in.cols
```

Morpheus: TDP with deduction

[Feng et al'17]



$$\exists x\ y\colon y.\mathrm{rows} = x.\mathrm{rows} \wedge y.\mathrm{cols} < x.\mathrm{cols} \wedge x.\mathrm{rows} = 3 \wedge x.\mathrm{cols} = 3 \wedge y.\mathrm{rows} = 2 \wedge y.\mathrm{cols} = 3$$

SMT → UNSAT

$$\exists x\ y\colon y.\mathrm{rows} < x.\mathrm{rows} \wedge y.\mathrm{cols} = x.\mathrm{cols} \wedge x.\mathrm{rows} = 3 \wedge x.\mathrm{cols} = 3 \wedge y.\mathrm{rows} = 2 \wedge y.\mathrm{cols} = 3$$

SMT → SAT

```
select : Table → [Col] → Table
filter : Table → (Row → Bool) → Table
```

```
out.rows = in.rows && out.cols < in.cols
our.rows < in.rows && out.cols = in.cols
```

Applications of synthesis

Custom data structures

Data extraction and data wrangling

→ Databases

- Wang, Cheung, Bodík: Synthesizing highly expressive SQL queries from input-output examples. PLDI'17
- Yaghmazadeh, Wang, Dillig, Dillig: SQLizer: Query Synthesis from Natural Language. OOPSLA'17
- Singh, Meduri, Elmagarmid, Madden, Papotti, Quiané-Ruiz, Solar-Lezama, Tang: Synthesizing Entity Matching Rules by Examples. VLDB'17

Scythe

[Wang et al. PLDI'17]

Problem: SQL query synthesis

User input: input-output examples (small tables)

T_1

id	date	uid
1	12/25	1
2	11/21	1
4	12/24	2

T_2

oid	val
1	30
1	10
1	10
2	50
2	10

T_{out}

c_0	c_1	c_2	c_3	c_4
1	12/25	1	1	30
4	12/24	2	2	10

Output:

```
Select *
From   (Select *
        From   T1
        Where  T1.date = 12/24
               Or T1.date = 12/25) T3
Join   (Select   oid, Max(val)
        From     (Select *
                  From   T2
                  Where  T2.val < 50) T4
        Group By oid) T5
On     T3.uid = T5.oid
```

Scythe: technique

[Wang et al. PLDI'17]

Sketch generation via bottom-up enumeration with OE

- sketches are SQL queries with holes for predicates
- similar idea to Cosi

Predicate synthesis via bottom-up enumeration with optimizations

SQLizer

[Yaghmazadeh et al. 2017]

Problem: SQL query synthesis

User input: natural language + DB schema

"Find the number of papers in OOPSLA 2010"

Output:

```
SELECT count(Publication.pid)
FROM Publication JOIN Conference ON Publication.cid = Conference.cid
WHERE Conference.name = "OOPSLA" AND Publication.year = 2010
```

SQLizer: techniuqes

[Yaghmazadeh et al. OOPSLA'17]

Sketch generation via semantic parsing

- similar idea to Cosi and Scythe

Quantitative type inhabitation

- deductive synthesis that also deduces weights

Sketch refinement

- most similar to program repair

Entity matching

[Singh et al. VLDB'17]

Entity matching: which rows correspond to the same person?

Goal: more interpretable results than existing approaches (e.g. decision trees)

Search strategy: Sketch + techniques for handling noise

(a) D_1: **an instance of schema** R

	name	address	email	nation	gender
r_1	Catherine Zeta-Jones	9601 Wilshire Blvd., Beverly Hills, CA 90210-5213	c.jones@gmail.com	Wales	F
r_2	C. Zeta-Jones	3rd Floor, Beverly Hills, CA 90210	c.jones@gmail.com	US	F
r_3	Michael Jordan	676 North Michigan Avenue, Suite 293, Chicago		US	M
r_4	Bob Dylan	1230 Avenue of the Americas, NY 10020		US	M

(b) D_2: **An instance of the schema** S

	name	apt	email	country	sex
s_1	Catherine Zeta-Jones	9601 Wilshire, 3rd Floor, Beverly Hills, CA 90210	c.jones@gmail.com	Wales	F
s_2	B. Dylan	1230 Avenue of the Americas, NY 10020	bob.dylan@gmail.com	US	M
s_3	Micheal Jordan	427 Evans Hall #3860, Berkeley, CA 94720	jordan@cs.berkeley.edu	US	M

Entity matching

[Singh et al. VLDB'17]