Adaptive Schema Databases

William Spoth[b], Bahareh Sadat Arab[i], Eric S. Chan[o], Dieter Gawlick[o],
Adel Ghoneimy[o], Boris Glavic[i], Beda Hammerschmidt[o], Oliver Kennedy[b],
Seokki Lee[i], Zhen Hua Liu[o], Xing Niu[i], **Ying Yang[b]**

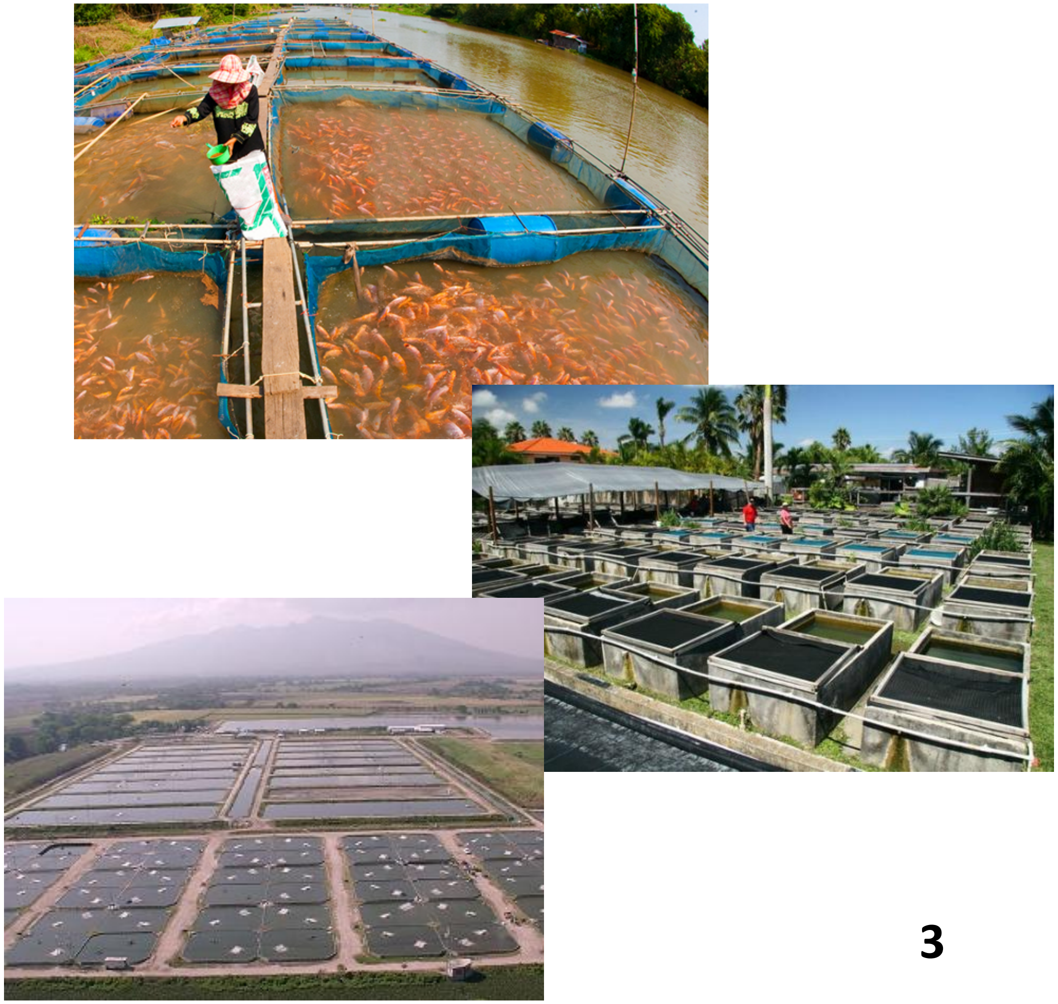
b: University at Buffalo   i: Illinois Inst. Tech.   o: Oracle

Adaptive Schema Databases

Classic relational database

- Navigational and organizational purpose

retain discovery, good performance and space, reusable.



Classic relational database

- But... High upfront cost and inflexible

BigData/NOSQL

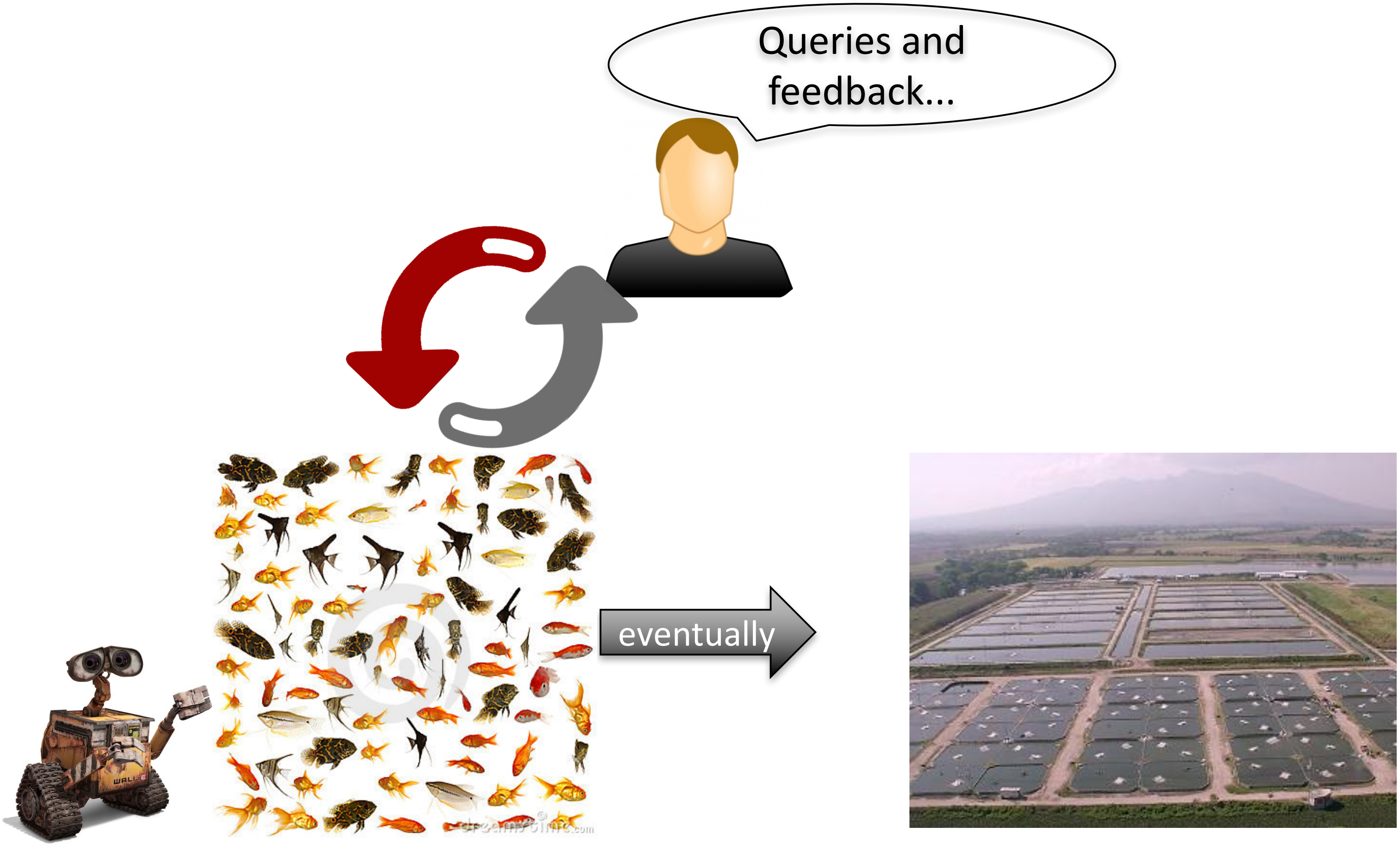
- Data can be used immediately.

BigData/NOSQL

- But... Sacrifice navigational and Performance benefit and may end up with duplicate of work

Adaptive Schema Databases

- Bridge the gap between relational database and NoSQl.

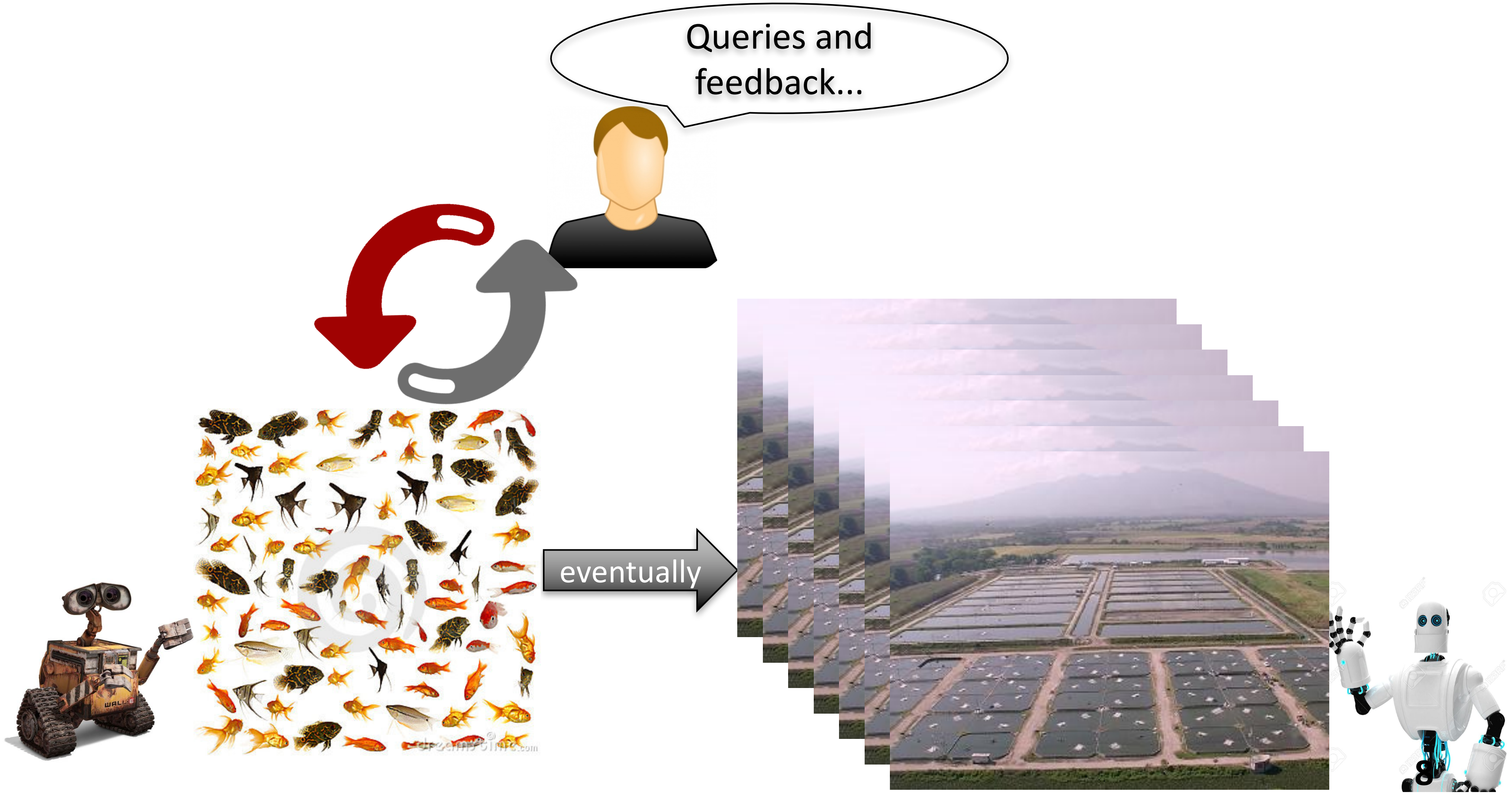


Queries and feedback...

eventually

Adaptive Schema Databases

- Bridge the gap between relational database and NoSQl.

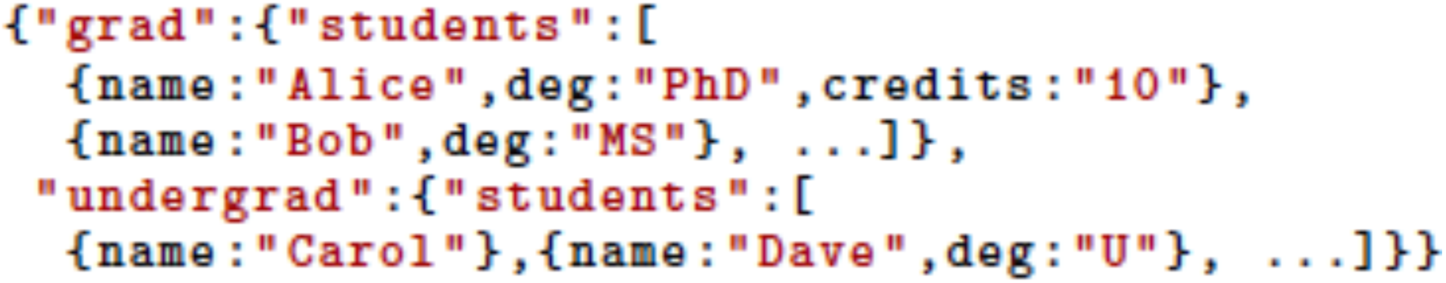


Queries and feedback...

eventually

Adaptive Schema Databases

Input:



```
{"grad":{"students":[
  {name:"Alice",deg:"PhD",credits:"10"},
  {name:"Bob",deg:"MS"}, ...]},
 "undergrad":{"students":[
  {name:"Carol"},{name:"Dave",deg:"U"}, ...]}}
```

Queries:

SELECT name FROM Undergrad UNION
SELECT name FROM Grad

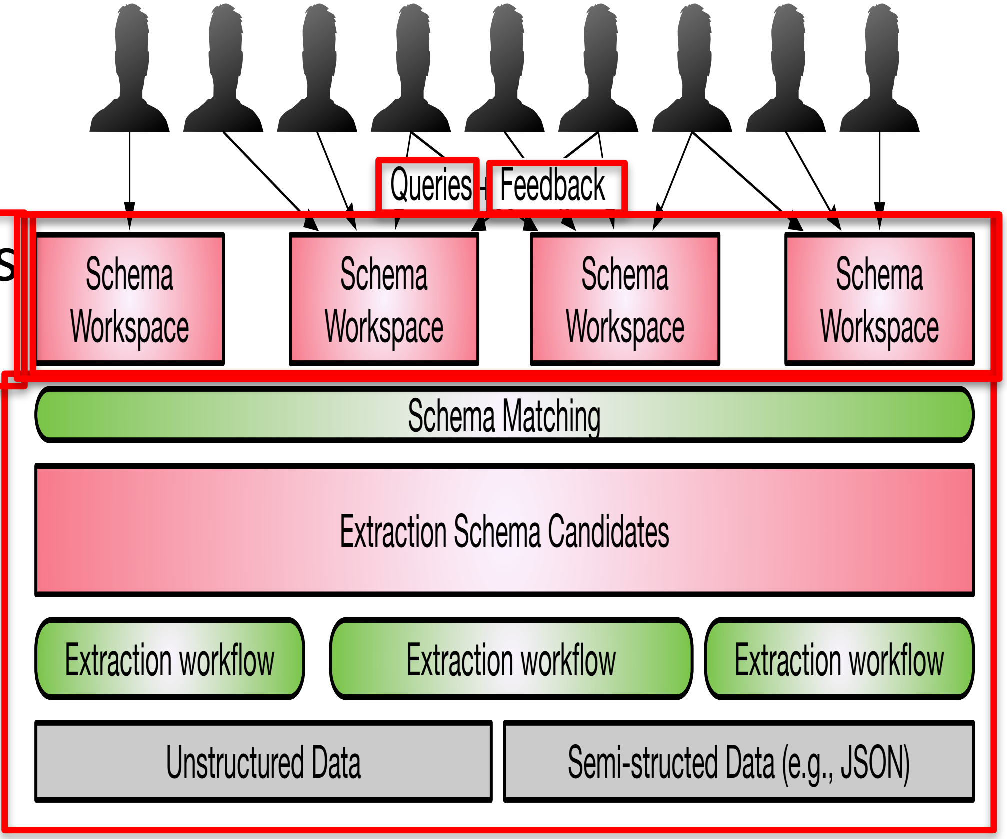
SELECT deg FROM Grad

SELECT name FROM Student

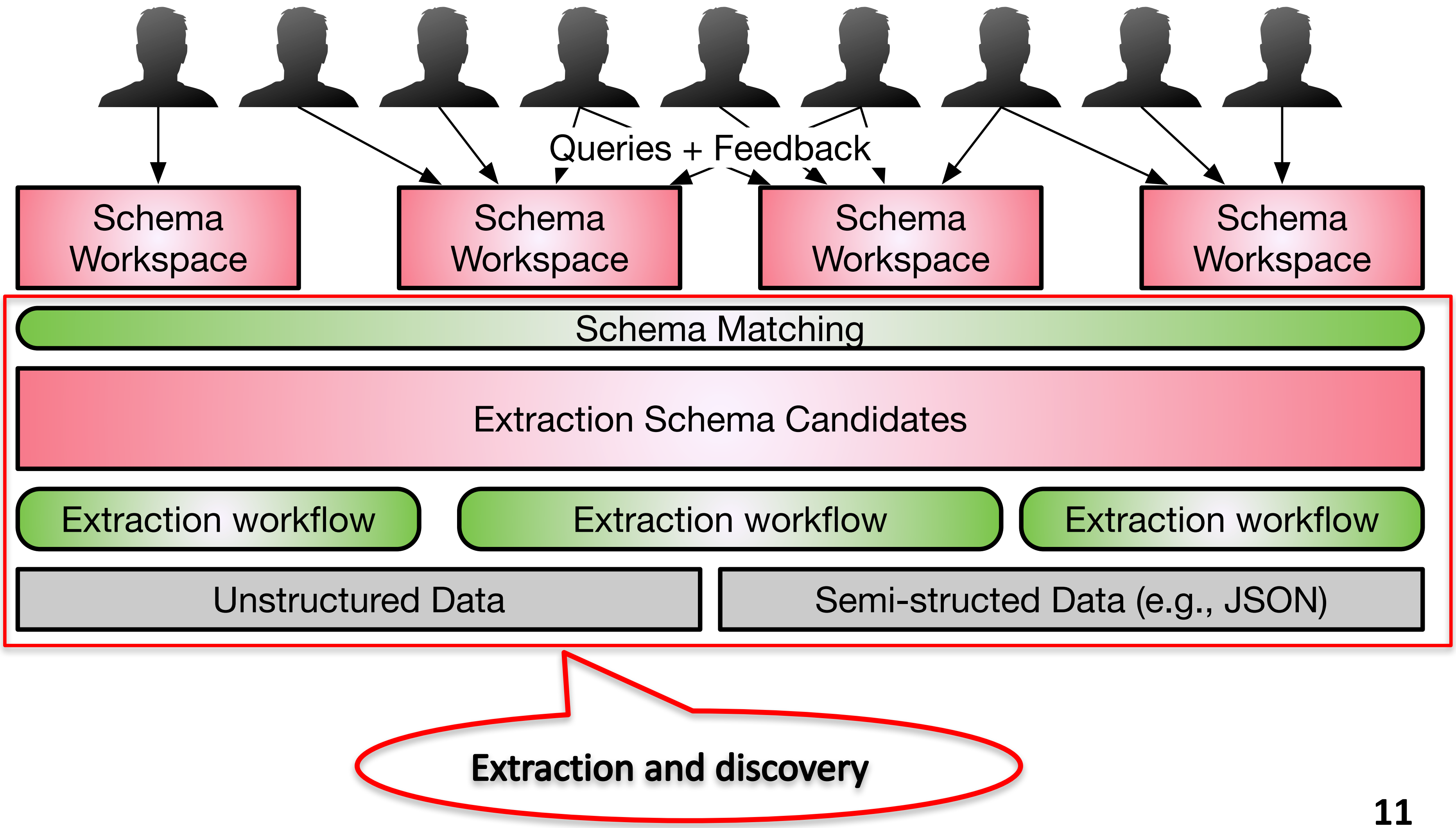
…

Outline

- Extraction and discovery
- Adaptive, personalized schemas from queries
- Explanations and feedback
- Adaptive organization
- Conclusions and future work



Extraction



Queries + Feedback

Schema Workspace

Schema Workspace

Schema Workspace

Schema Workspace

Schema Matching

Extraction Schema Candidates

Extraction workflow

Extraction workflow

Extraction workflow

Unstructured Data

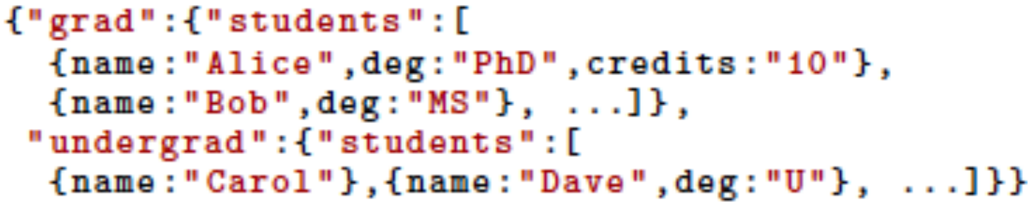
Semi-structed Data (e.g., JSON)

Extraction and discovery

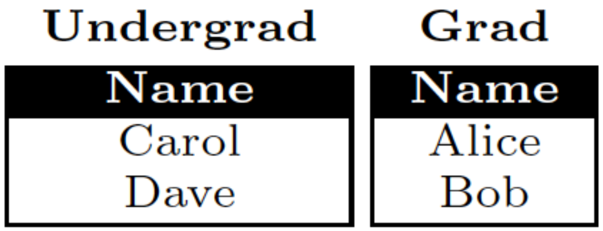
Extraction

- ASD extracts schema candidate set

Given input:



```
{"grad":{"students":[
  {name:"Alice",deg:"PhD",credits:"10"},
  {name:"Bob",deg:"MS"}, ...]},
 "undergrad":{"students":[
  {name:"Carol"},{name:"Dave",deg:"U"}, ...]}}
```



Undergrad

Name
Carol
Dave

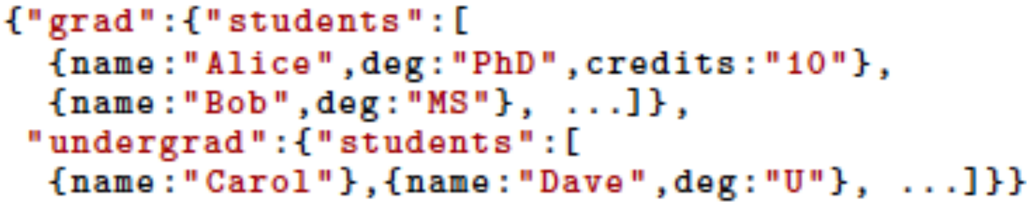
Grad

Name
Alice
Bob

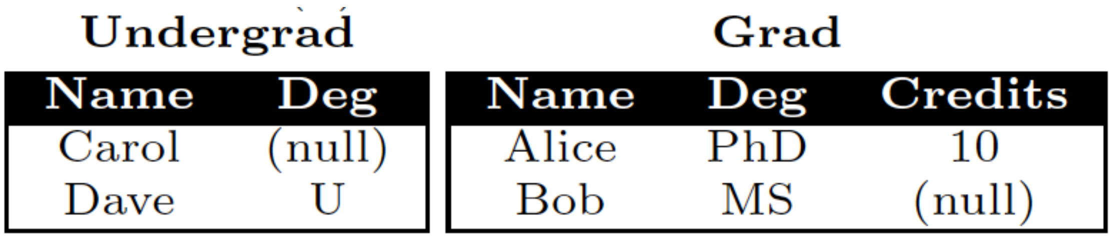
Extraction

- ASD extracts schema candidate set

Given input:



```
{"grad":{"students":[
  {name:"Alice",deg:"PhD",credits:"10"},
  {name:"Bob",deg:"MS"}, ...]},
 "undergrad":{"students":[
  {name:"Carol"},{name:"Dave",deg:"U"}, ...]}}
```



Undergrad

Name	Deg
Carol	(null)
Dave	U

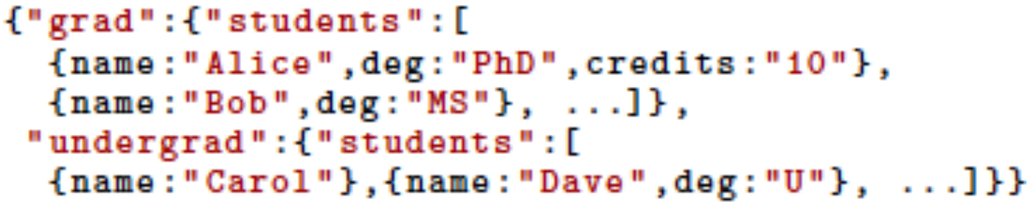
Grad

Name	Deg	Credits
Alice	PhD	10
Bob	MS	(null)

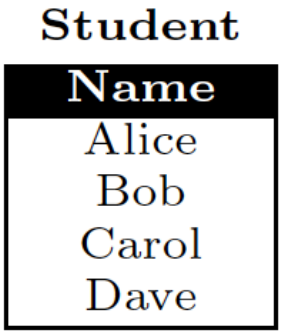
Extraction

- ASD extracts schema candidate set

Given input:



```
{"grad":{"students":[
  {name:"Alice",deg:"PhD",credits:"10"},
  {name:"Bob",deg:"MS"}, ...]},
 "undergrad":{"students":[
  {name:"Carol"},{name:"Dave",deg:"U"}, ...]}}
```



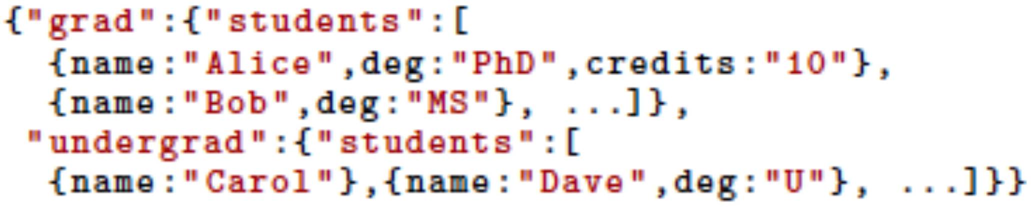
Student

Name
Alice
Bob
Carol
Dave

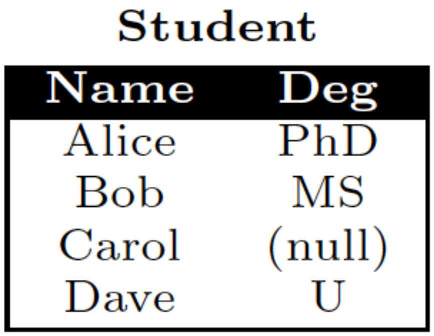
Extraction

- ASD extracts schema candidate set

Given input:



```
{"grad":{"students":[
  {name:"Alice",deg:"PhD",credits:"10"},
  {name:"Bob",deg:"MS"}, ...]},
 "undergrad":{"students":[
  {name:"Carol"},{name:"Dave",deg:"U"}, ...]}}
```



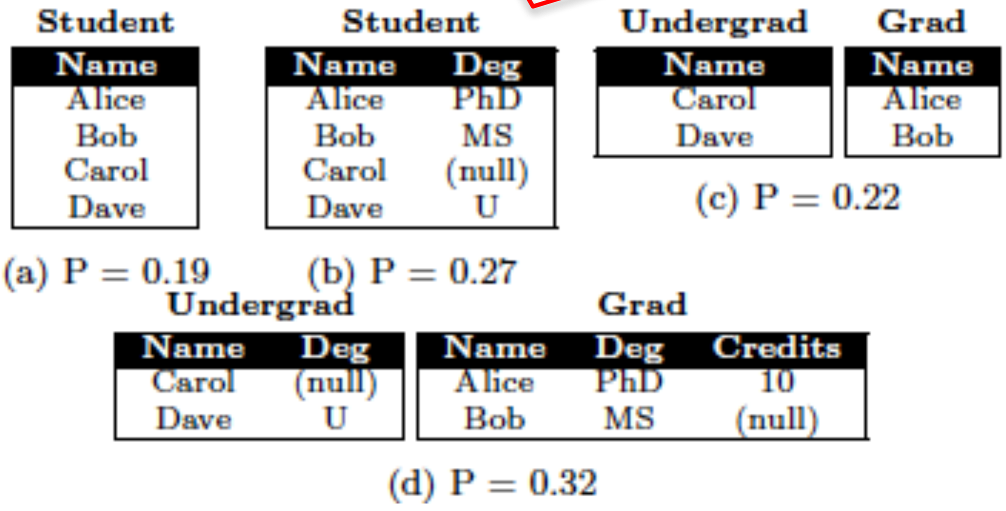
Student

Name	Deg
Alice	PhD
Bob	MS
Carol	(null)
Dave	U

Discovery

- ASD extracts schema candidate set

schema candidate set $C_{ext=}\{S_{ext}, P_{ext}\}$,
where S_{ext} is a set of candidate schemas,
P_{ext} is a probability distribution over these schemas.



Student

Name
Alice
Bob
Carol
Dave

(a) P = 0.19

Student

Name	Deg
Alice	PhD
Bob	MS
Carol	(null)
Dave	U

(b) P = 0.27

Undergrad

Name
Carol
Dave

Grad

Name
Alice
Bob

(c) P = 0.22

Undergrad

Name	Deg
Carol	(null)
Dave	U

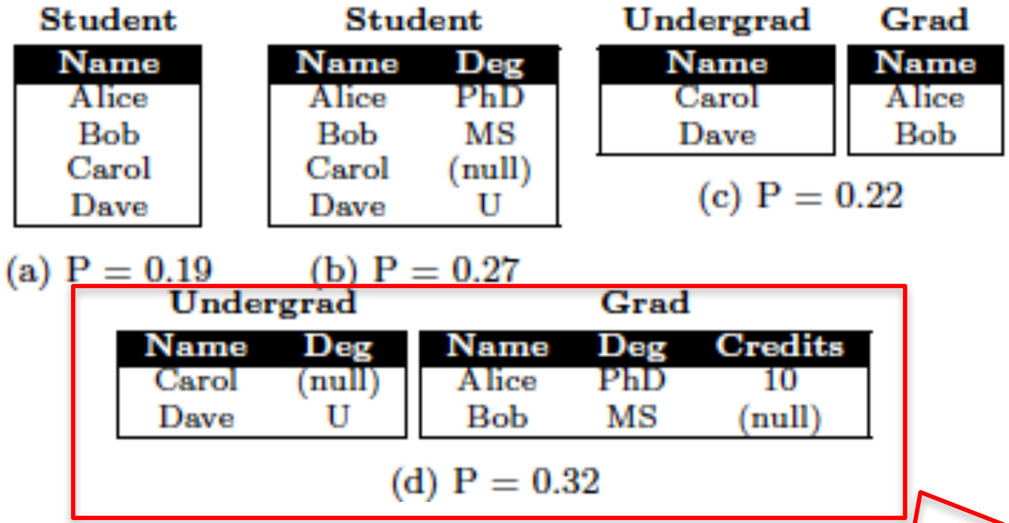
Grad

Name	Deg	Credits
Alice	PhD	10
Bob	MS	(null)

(d) P = 0.32

Discovery

- ASD extracts schema candidate set



Student

Name
Alice
Bob
Carol
Dave

(a) P = 0.19

Student

Name	Deg
Alice	PhD
Bob	MS
Carol	(null)
Dave	U

(b) P = 0.27

Undergrad

Name
Carol
Dave

Grad

Name
Alice
Bob

(c) P = 0.22

Undergrad

Name	Deg
Carol	(null)
Dave	U

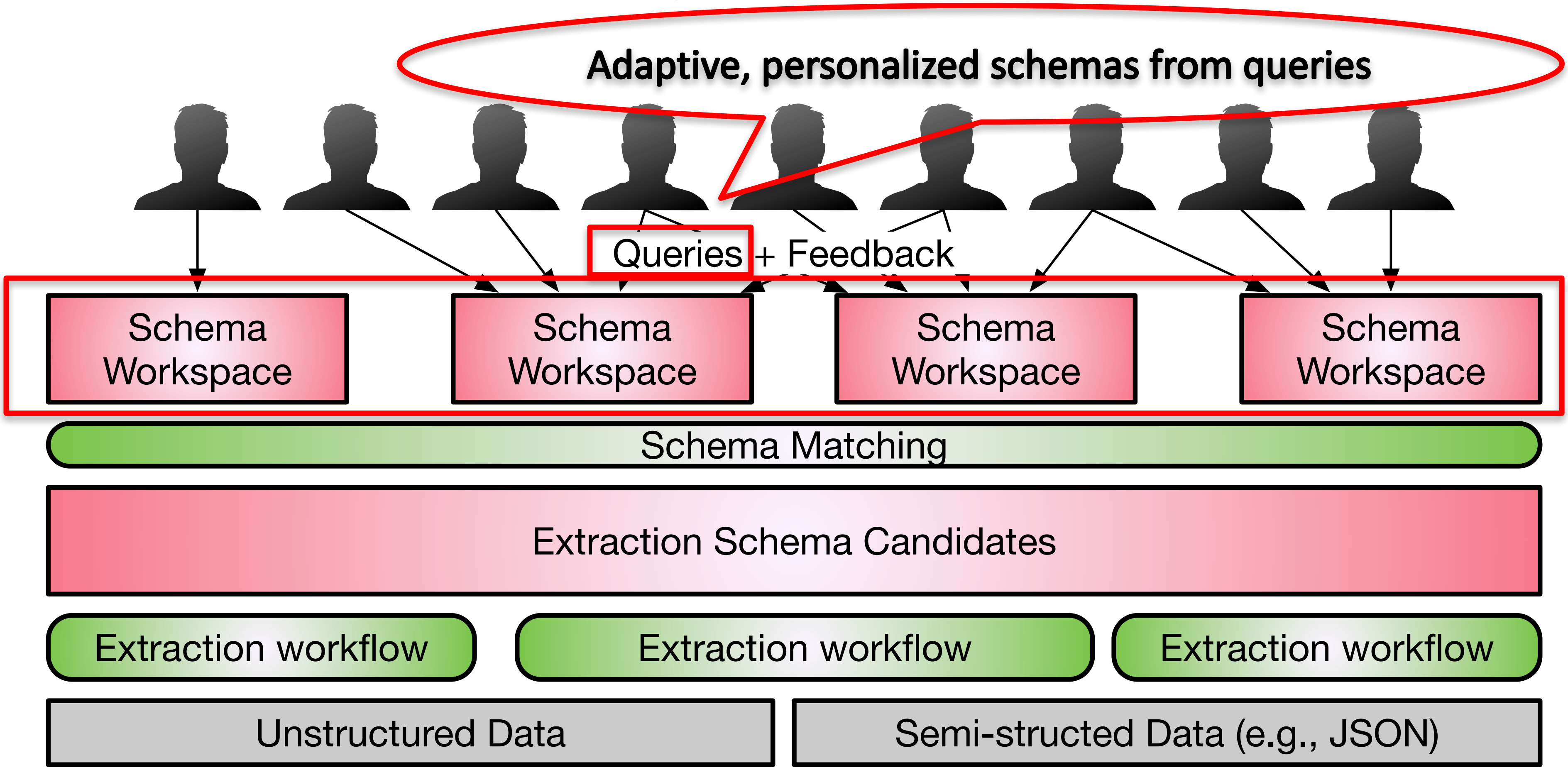
Grad

Name	Deg	Credits
Alice	PhD	10
Bob	MS	(null)

(d) P = 0.32

Smax:
the best guess schema

Adaptive, personalized schemas from queries



Adaptive, personalized schemas

- ASD maintains a set of schema workspaces W={$W_{1,\ldots,}$$W_n$}.

Initially, W={}

Finding Schemas from Queries

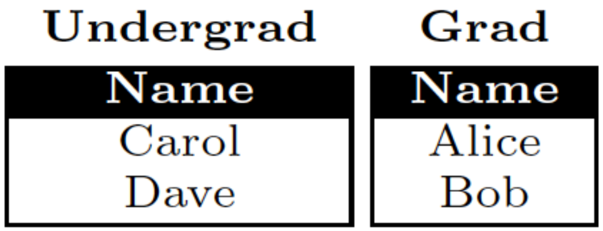
- ASD maintains a set of schema workspaces $W=\{W_{1,\ldots,}W_n\}$.

```
Query 1:   SELECT name FROM Undergrad UNION
                                  SELECT name FROM Grad
```

Finding Schemas from Queries

- ASD maintains a set of schema workspaces W={$W_{1,…,}W_n$}.

```
Query 1:   SELECT name FROM Undergrad UNION
                                       SELECT name FROM Grad
```



Undergrad

Name
Carol
Dave

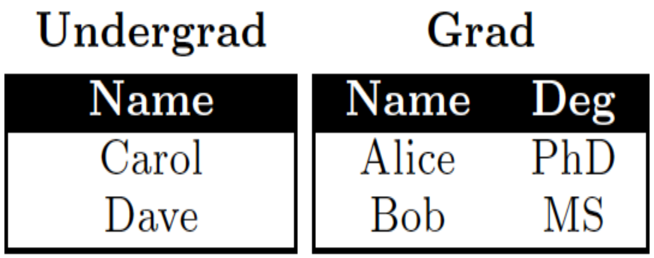
Grad

Name
Alice
Bob

Finding Schemas from Queries

- ASD maintains a set of schema workspaces $W=\{W_{1,\ldots,}W_n\}$.

Query 2: SELECT deg FROM Grad



Undergrad

Name
Carol
Dave

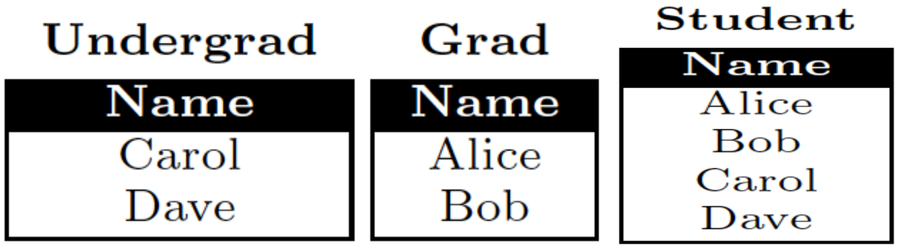
Grad

Name	Deg
Alice	PhD
Bob	MS

Synthesizing Tables

- ASD maintains a set of schema workspaces $W=\{W_{1,\ldots,}W_n\}$.

Query 3: SELECT name FROM Student



Undergrad

Name
Carol
Dave

Grad

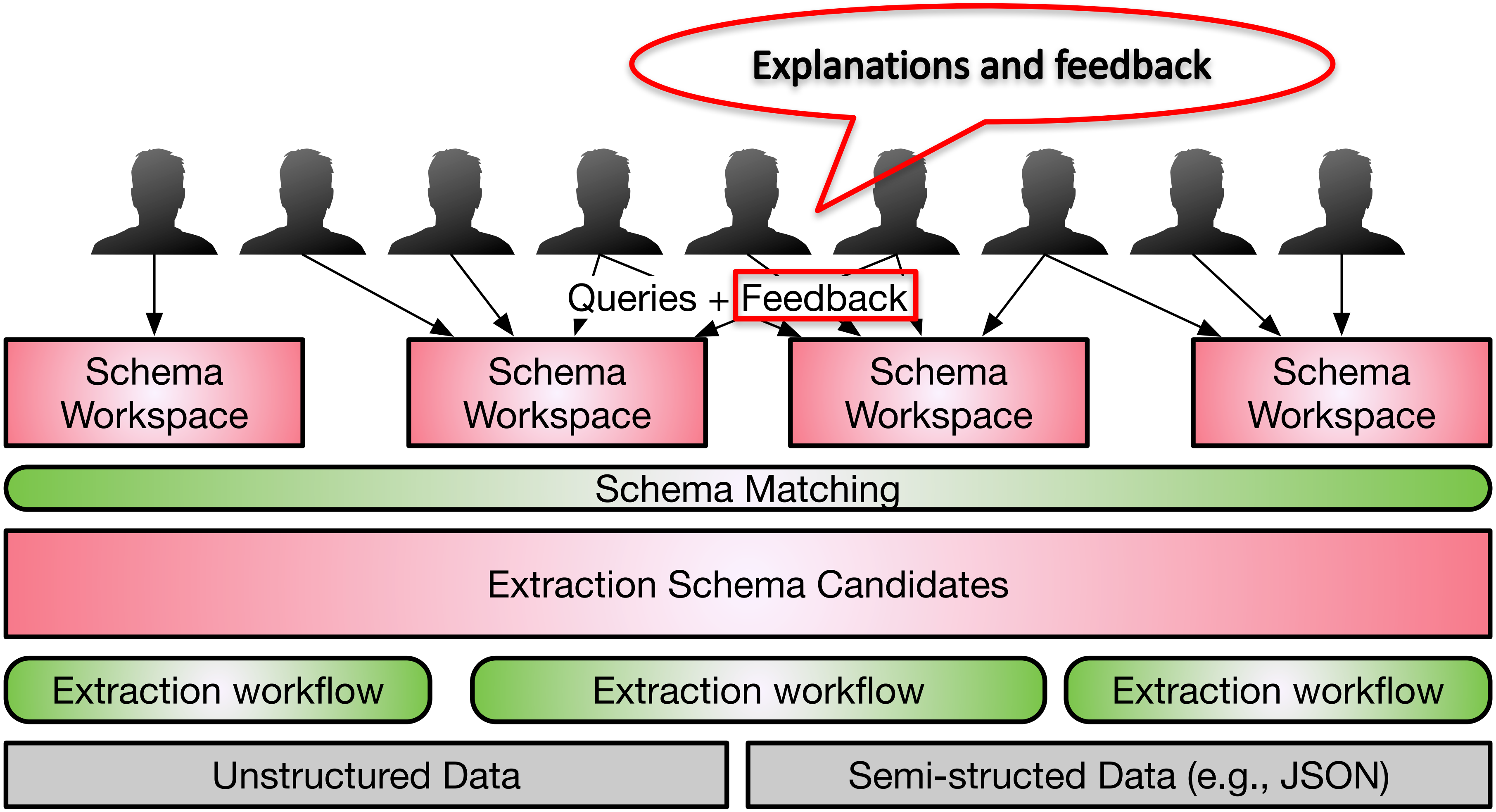
Name
Alice
Bob

Student

Name
Alice
Bob
Carol
Dave

W_1 = (S_1={Undergrad(name)},P_1=0.27),
(S_1={Grad(name)},P_1=0.23),
(S_1={Undergrad(name), Grad(name)}, P_1=0.5)

Explanations and feedback



What might go wrong

Extraction errors appear in three forms:

(1) A query incompatible with S_{max}

(2) An update with data that violates S_{max}

(3) An extraction error presented to user

We provide: (1) explanation of results

(2) provenance

(3) Warn the analyst with ambiguity

(4) Explain the ambiguity

(5) Evaluate the magnitude of ambiguity

(6) Assist the analyst to resolve the ambiguity

Types of errors

ASD interacts with the outside world: Schema, Data, and Update.

Schema interactions: When a query incompatible with S_{max} and the workspace

Data interactions: provenance for attribute and row level ambiguity.

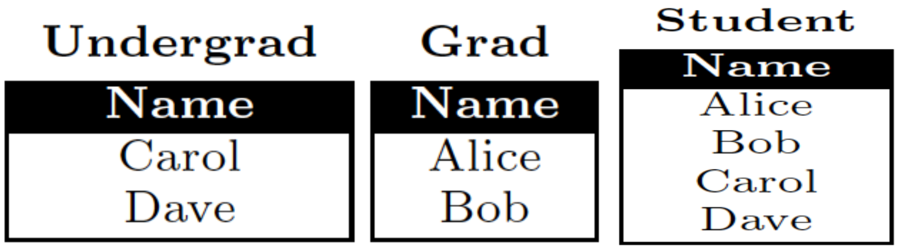
Update interactions:

- represent schema mismatches as missing values.
- resolve data errors with a probabilistic repair.
- upgrade her schema to match the changes.
- checkpoint her workspace and ignore new updates.

Explanations and feedback

Condition 2: Query from unknown schema elements:

SELECT name FROM Student



Undergrad

Name
Carol
Dave

Grad

Name
Alice
Bob

Student

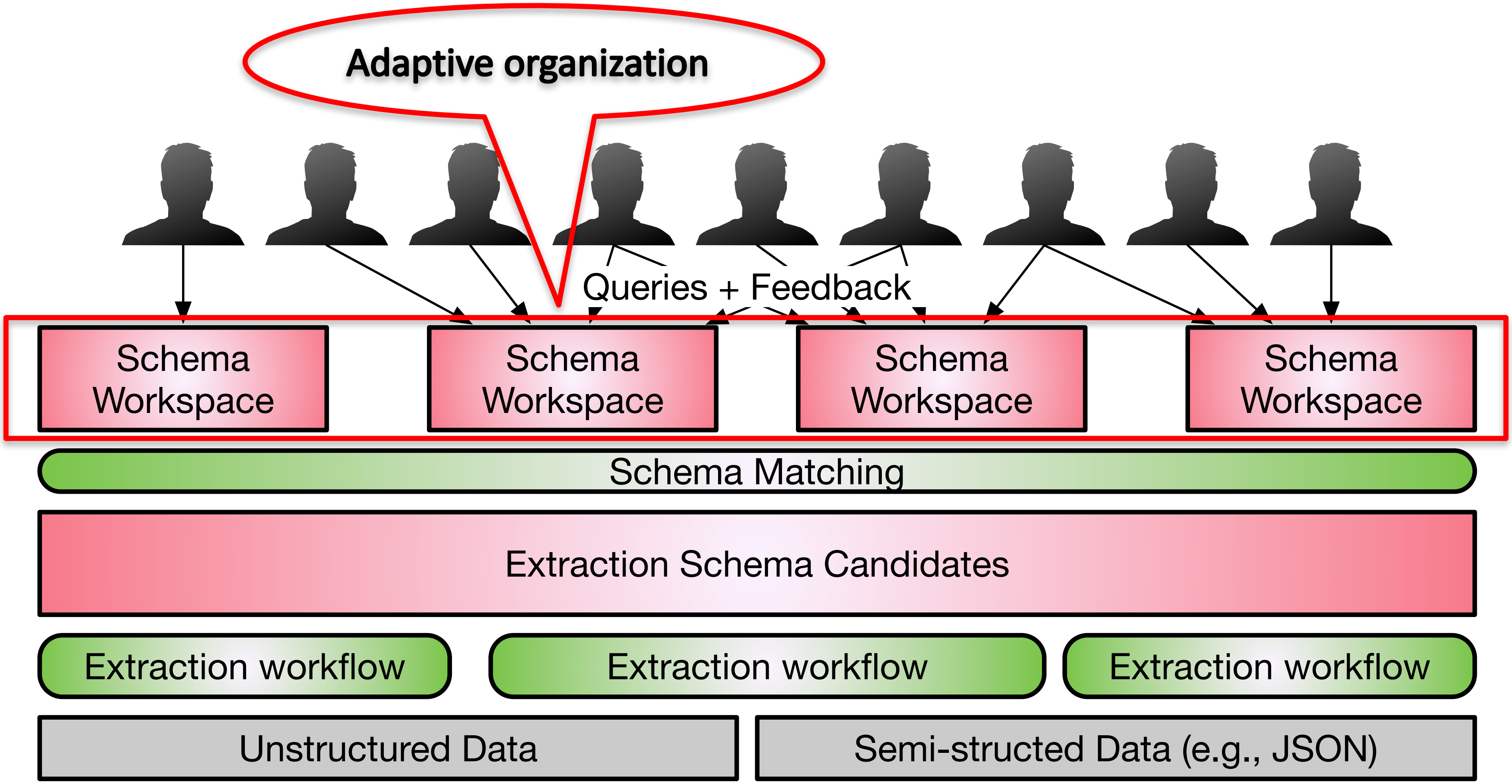
Name
Alice
Bob
Carol
Dave

W_1 = (S_1={Undergrad(name)},P_1=0.27),
(S_1={Grad(name)},P_1=0.23),
(S_1={Undergrad(name), Grad(name)}, P_1=0.5)

→

Explanations:
We match Student with both Grad and Undergrad

Adaptive organization



Adaptive organization

Trade-off between storing data in its native format and based on a specific schema.

What is the challenge? Many workspaces, add table to the schema, ….

Challenges and Possible Solutions:

- We want multiple personalized schemas

  1. Relational workspace schema is essentially a *view* over raw data. Materializing view can be used.

  2. Use existing *adaptive physical design* and *caching* techniques.

- Shared materializations

  1. Incremental materialized view maintenance. Leverage techniques from revision control systems.

  2. View selection problem.

Conclusions and future work

ASD bridges the gap between relational databases and NoSQL.

- ***Discovery***: Help user explore and understand new data by providing an outline of the available information. ***Done***
- ***Materialization***: Adopt work on adaptive data structures. ***Partially done***
- ***Data Synthesis***: Synthesis new tables and attributes from existing data. ***Done***
- ***Conflict Response***:
  - Versioning or branching the schema.
  - Log analysis to help users assess the impact of schema revisions.