

A Vision for Autonomous Data Agent Collaboration

From Query-by-Integration to Query-by-Collaboration

Timo Eckmann & Carsten Binnig
Systems @ TU Darmstadt



SYSTEMS



TECHNISCHE
UNIVERSITÄT
DARMSTADT

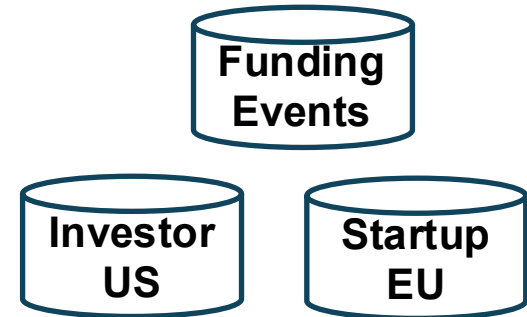
The Relational “Pipe Dream”

```
SELECT Cmp-Name FROM FundingDB  
WHERE Cmp-Cntry='DE' AND VC-Cntry='US'
```

FundingDB			
Cmp-Name	Cmp-Ctry	VC-Name	VC-Ctry
TechBerlin	DE	Accel	USA
ParisAI	FR	Index	UK
MunichBio	DE	Sequoia	USA

The Relational Pipe Dream: (1) Users need to put data “only” in a clean relational DB, (2) then you can query it!

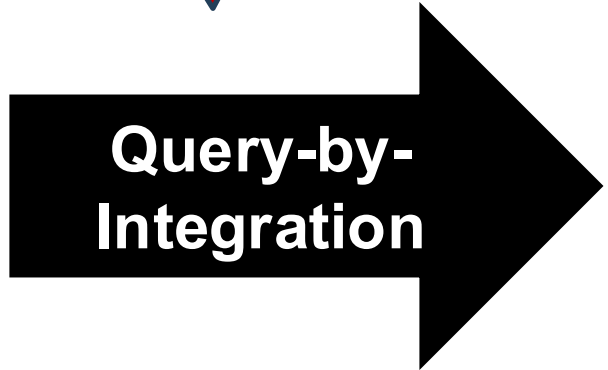
The “Dirty” Secret: Query-by-Integration



Data Sources



Limitation 1:
High overheads for data integration



When did ParisAI get funding?

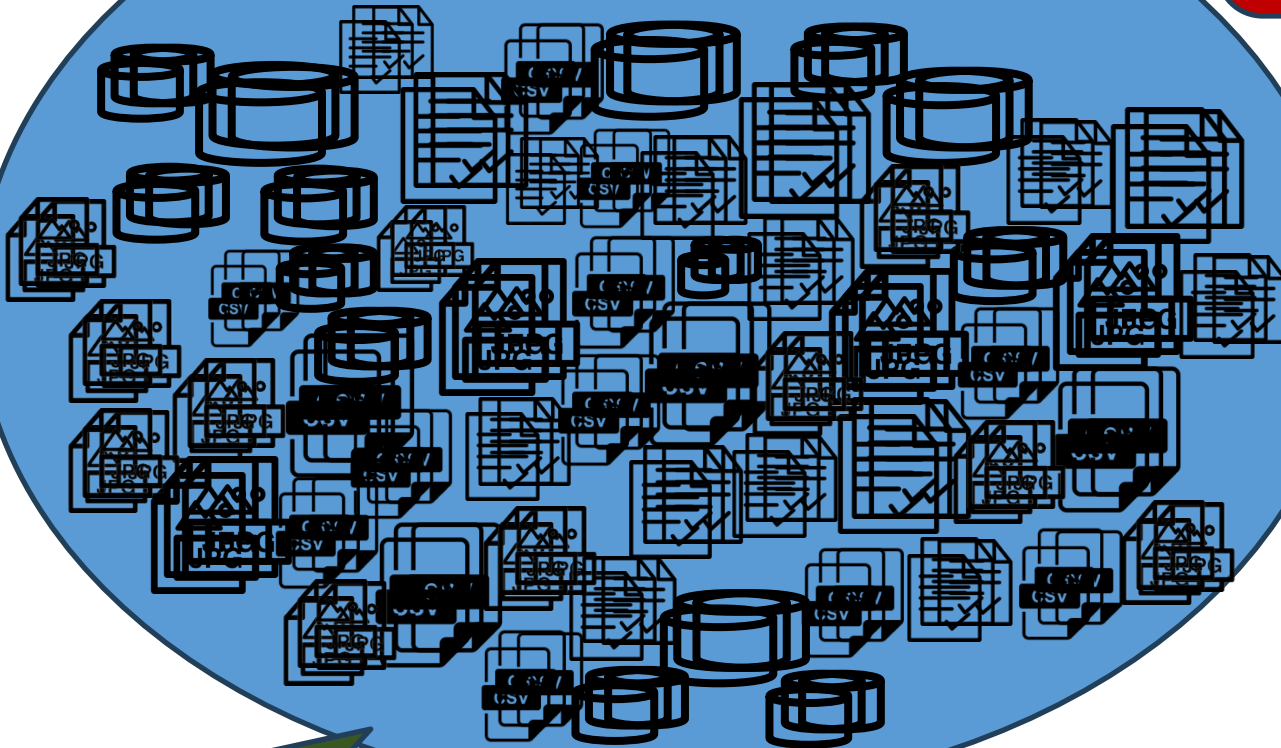
FundingDB			
Cmp-Name	Cmp-Ctry	VC-Name	VC-Ctry
TechBerlin	DE	Accel	USA
ParisAI	FR	Index	UK
MunichBio	DE	Sequoia	USA

Find & integrate “Funding year”?

Limitation 2:
Queries limited by integrated database

Why Query-by-Integration is doomed?

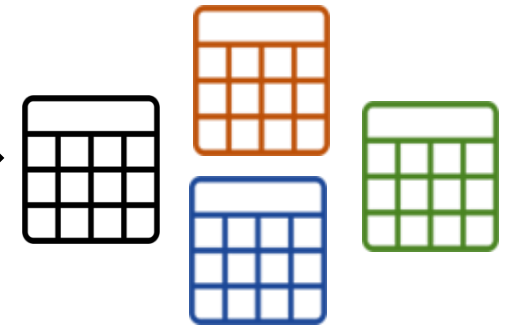
Exponentially growing
of Data Sources



Significantly
limits what data
we can query

The
Integration
-Funnel

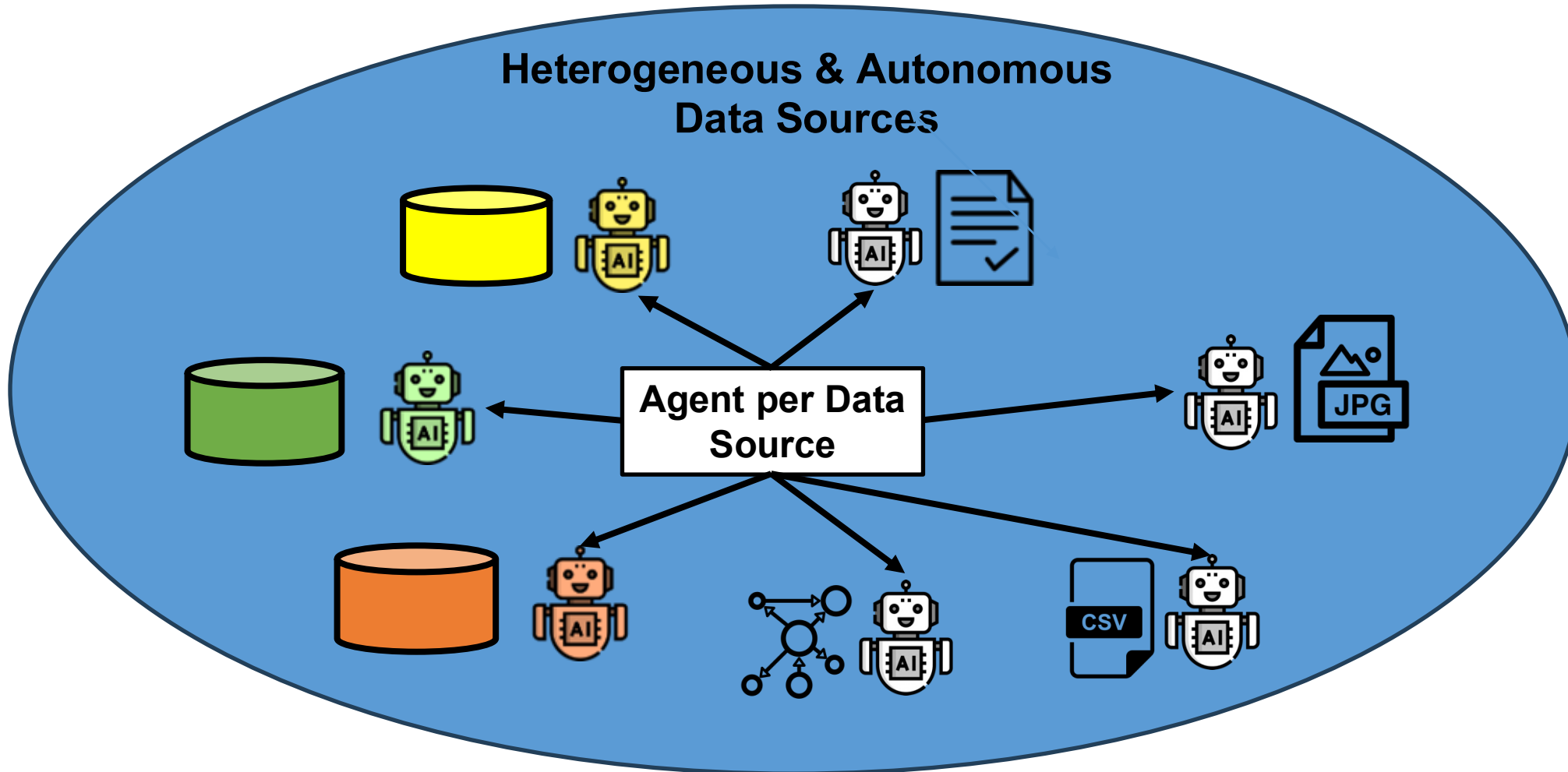
Data we have put in
clean relational DBs



Our Vision: Make all
data of the world queryable!

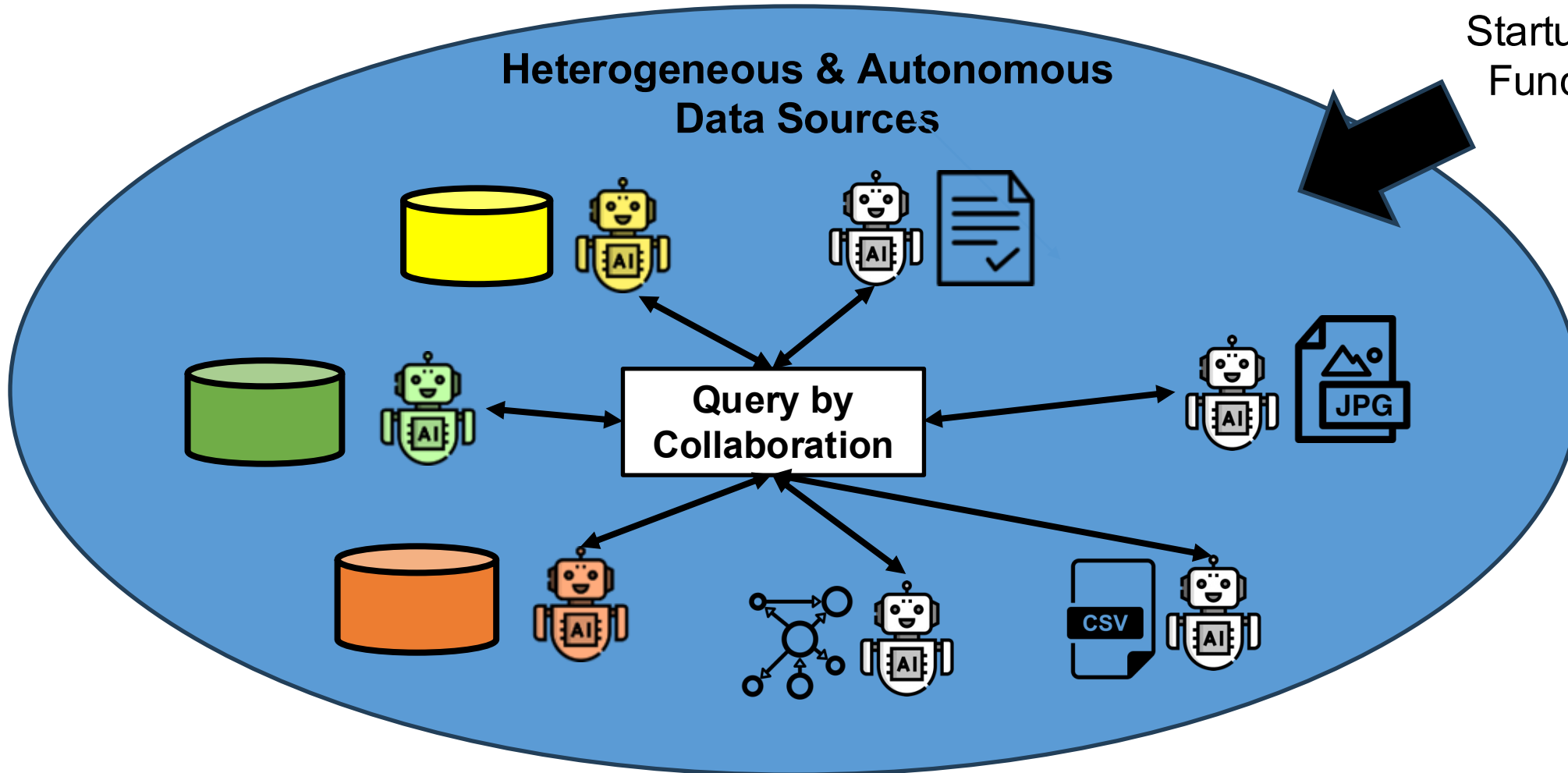
We can not keep
up with the fast
data growth!

Our Vision: Query-by-Collaboration



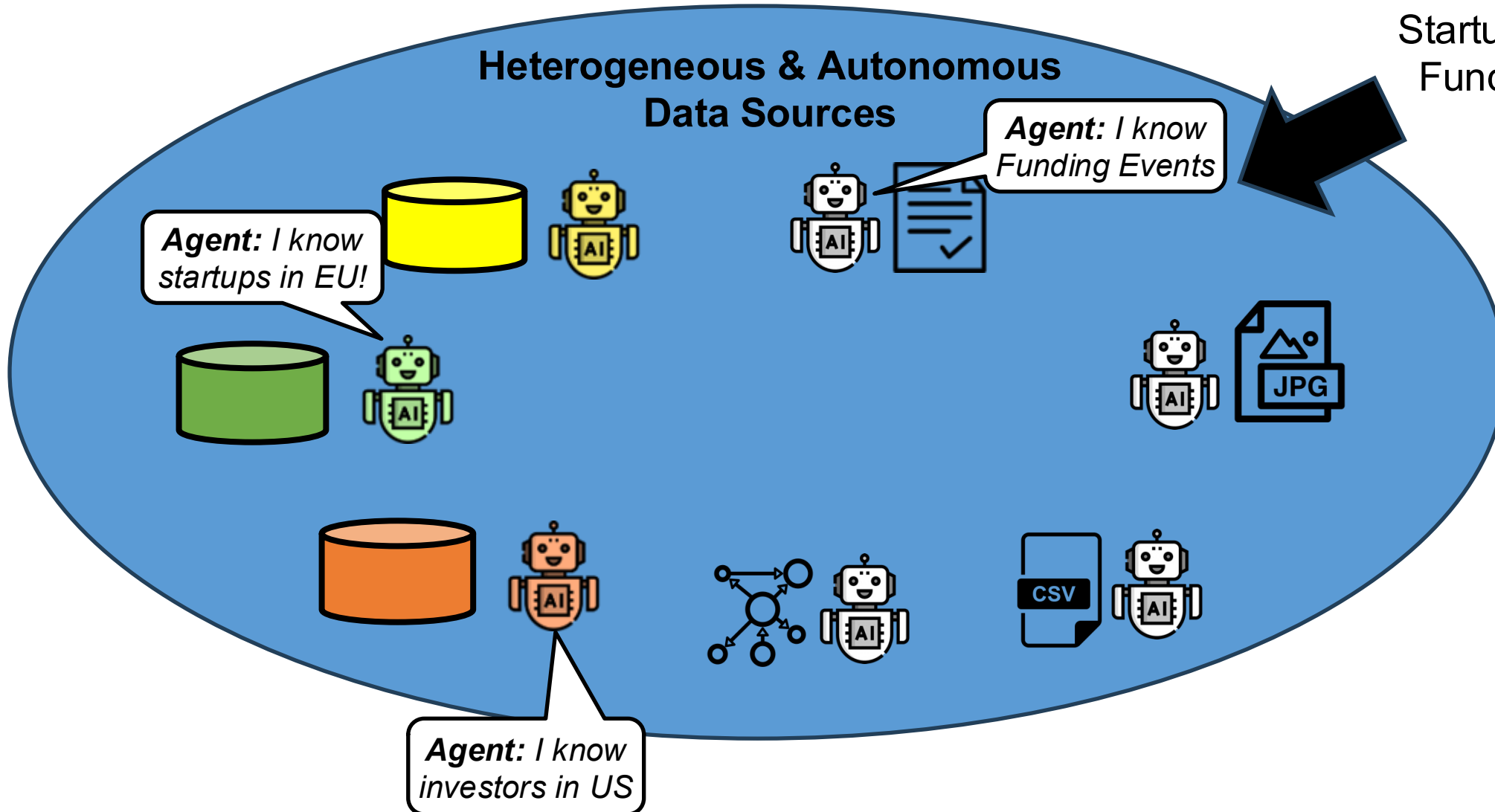
Our Vision: Query-by-Collaboration

User Query:
Which German
Startups Raised US
Funding in 2024?



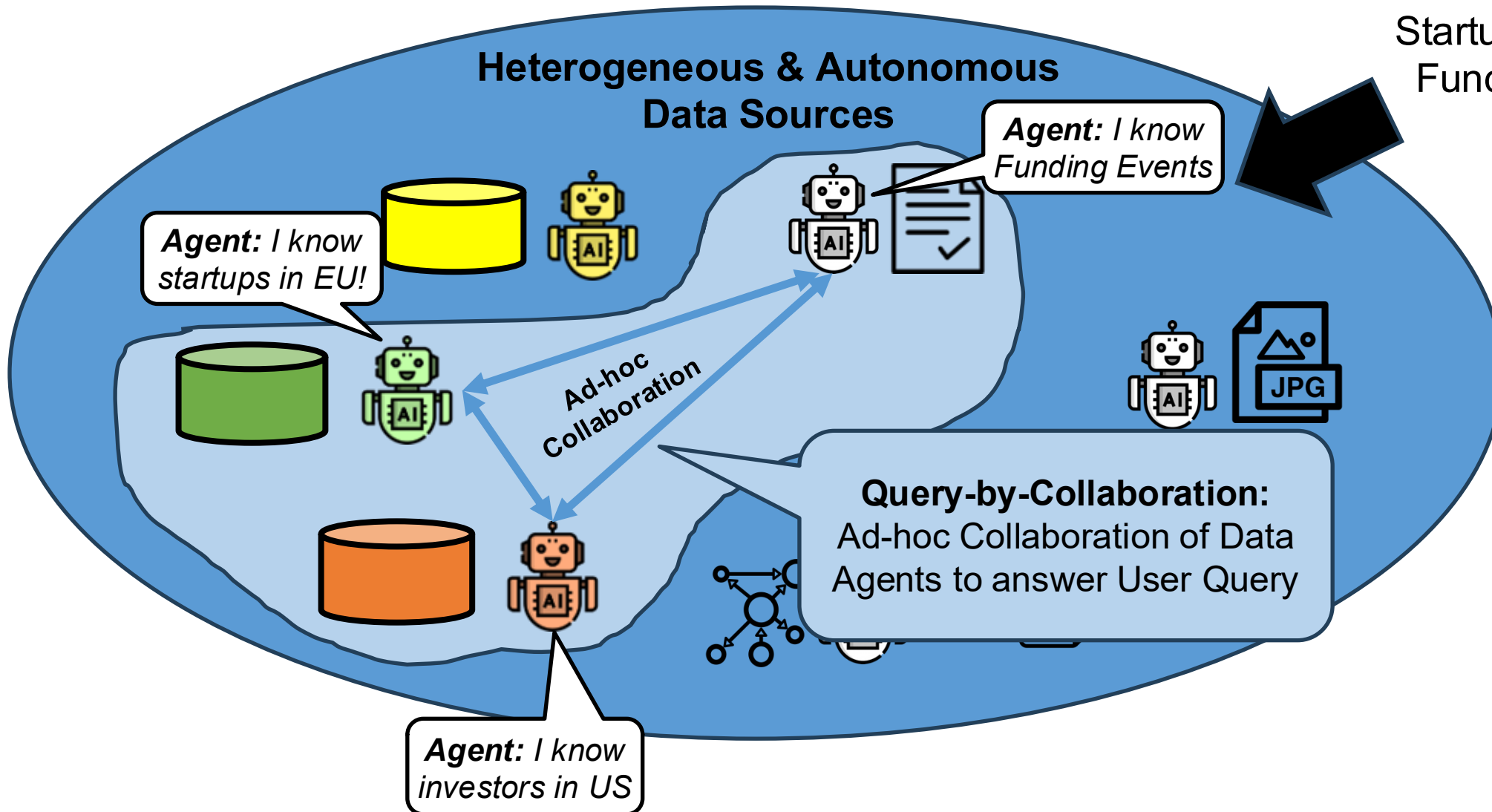
Our Vision: Query-by-Collaboration

User Query:
Which German
Startups Raised US
Funding in 2024?



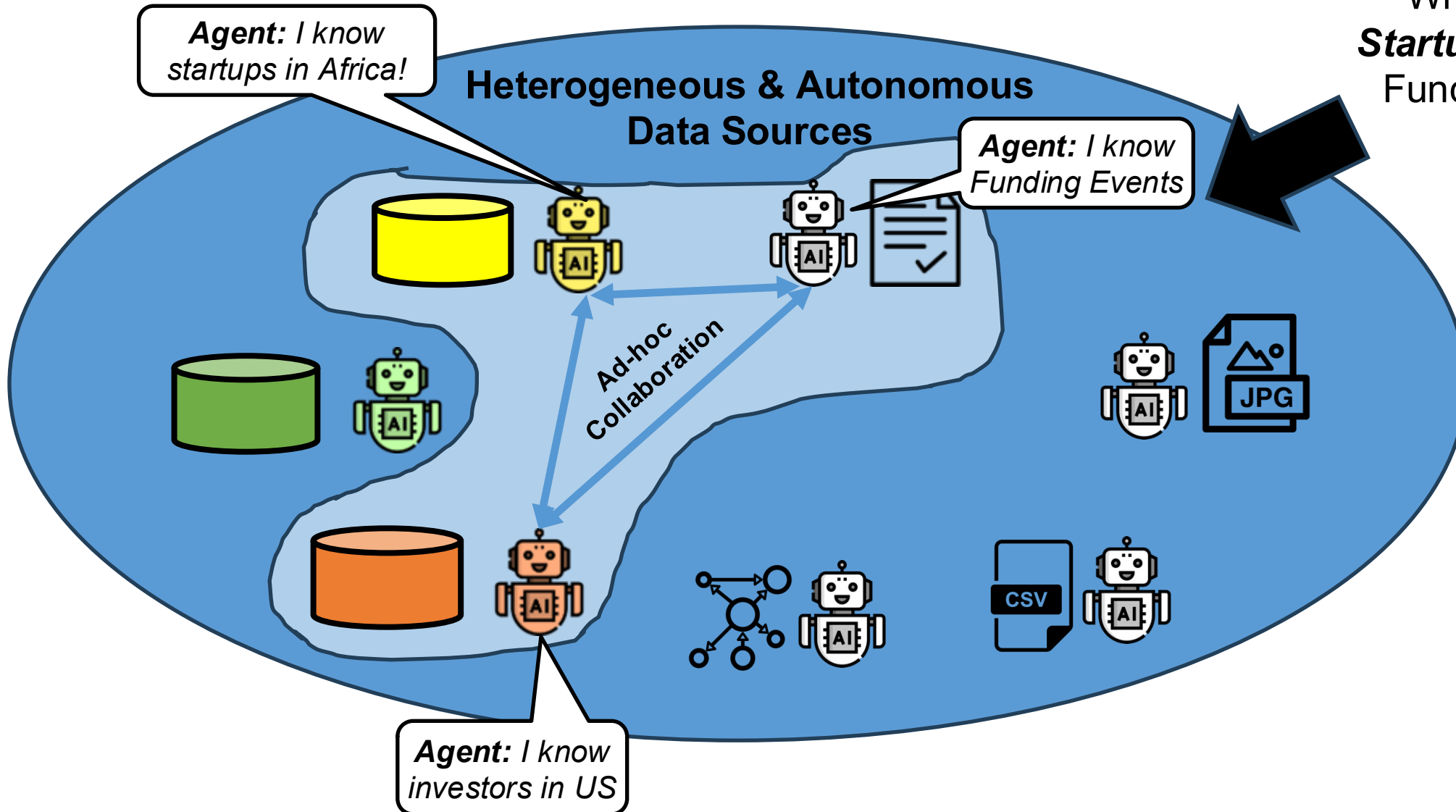
Our Vision: Query-by-Collaboration

User Query:
Which German
Startups Raised US
Funding in 2024?



Our Vision: Query-by-Collaboration

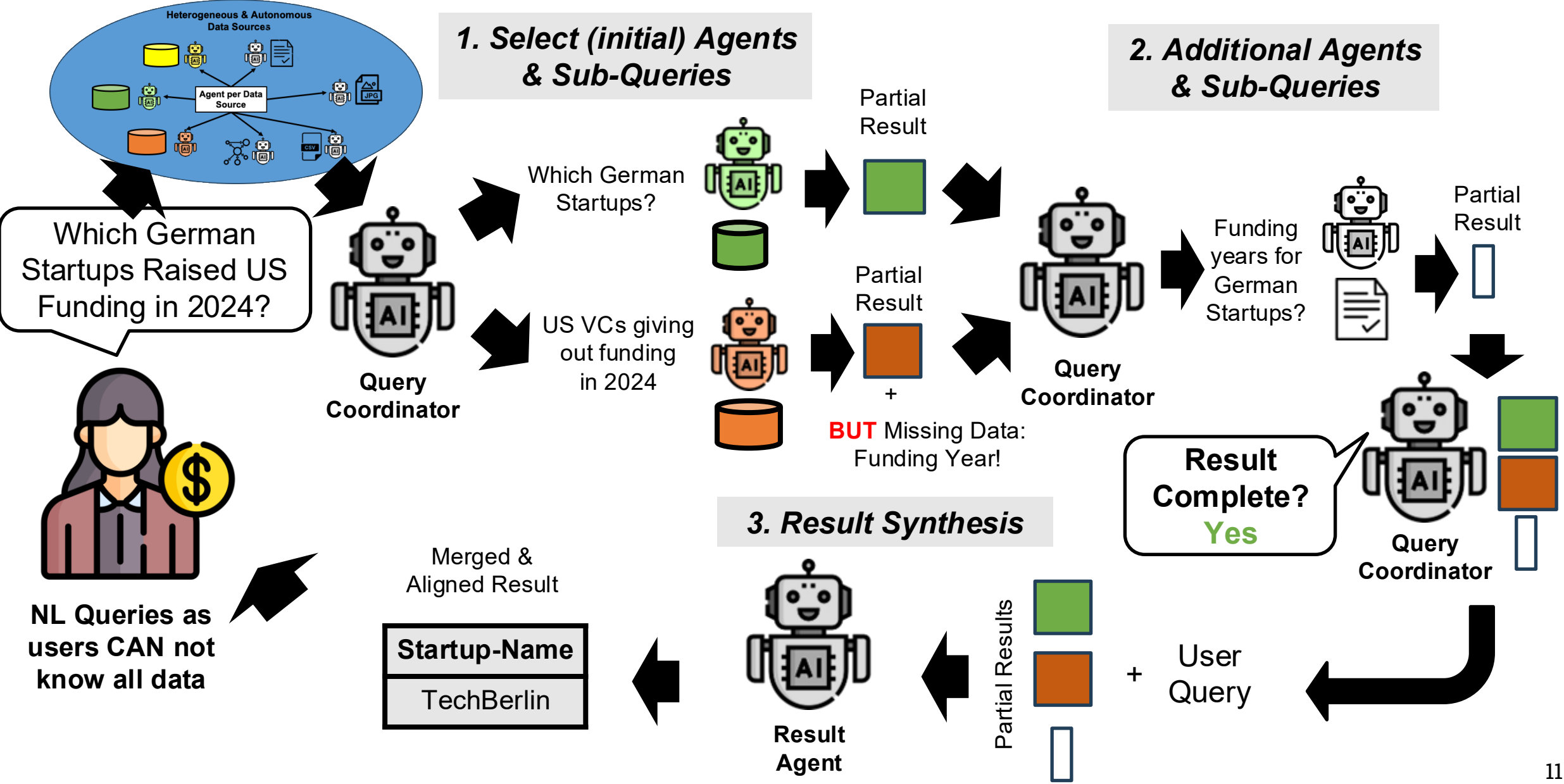
Another User Query:
Which *African*
Startups Raised US
Funding in 2024?



How? The Query Collaboration Protocol (QCP)!

QCP = A protocol on how data agents can collaborate for query answering by executing partial queries and sharing intermediate results without a global schema

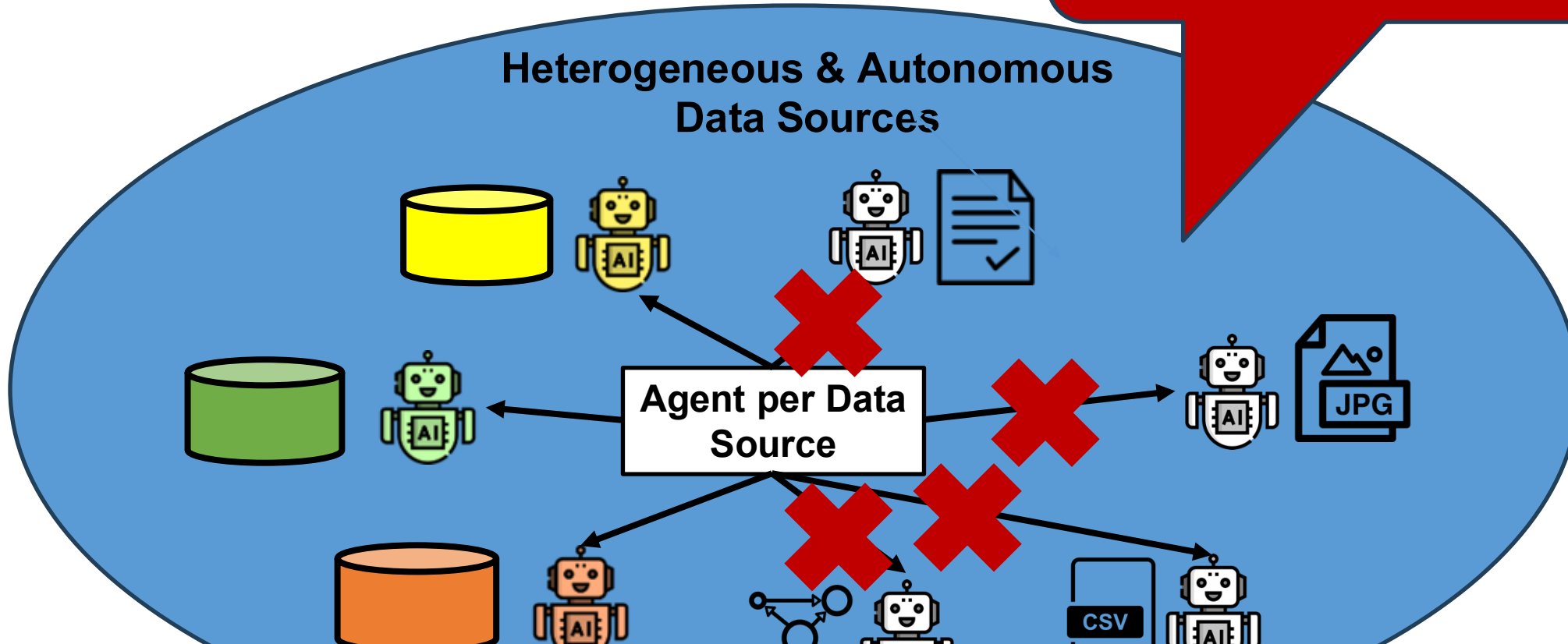
How? The Query Collaboration Protocol (QCP)!



Our Prototype: QCP-DB

Our Limitations:

- Only Relational Data (so far)
- Rather small scale in # of agents

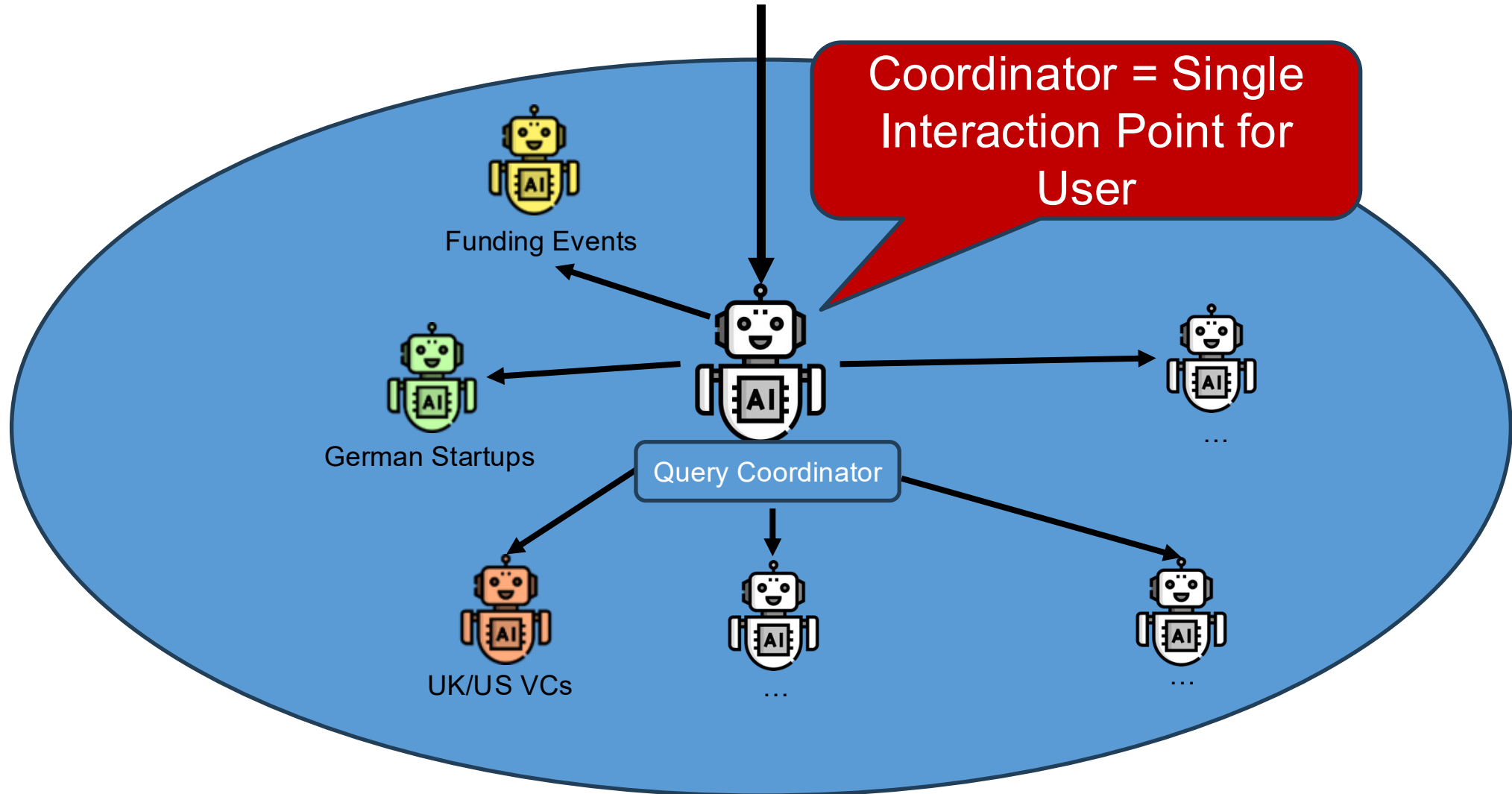


Goal of QCP-DB: Basic Understanding of how to design Query-by-Collaboration Systems

Our Prototype: QCP-DB

User Query:

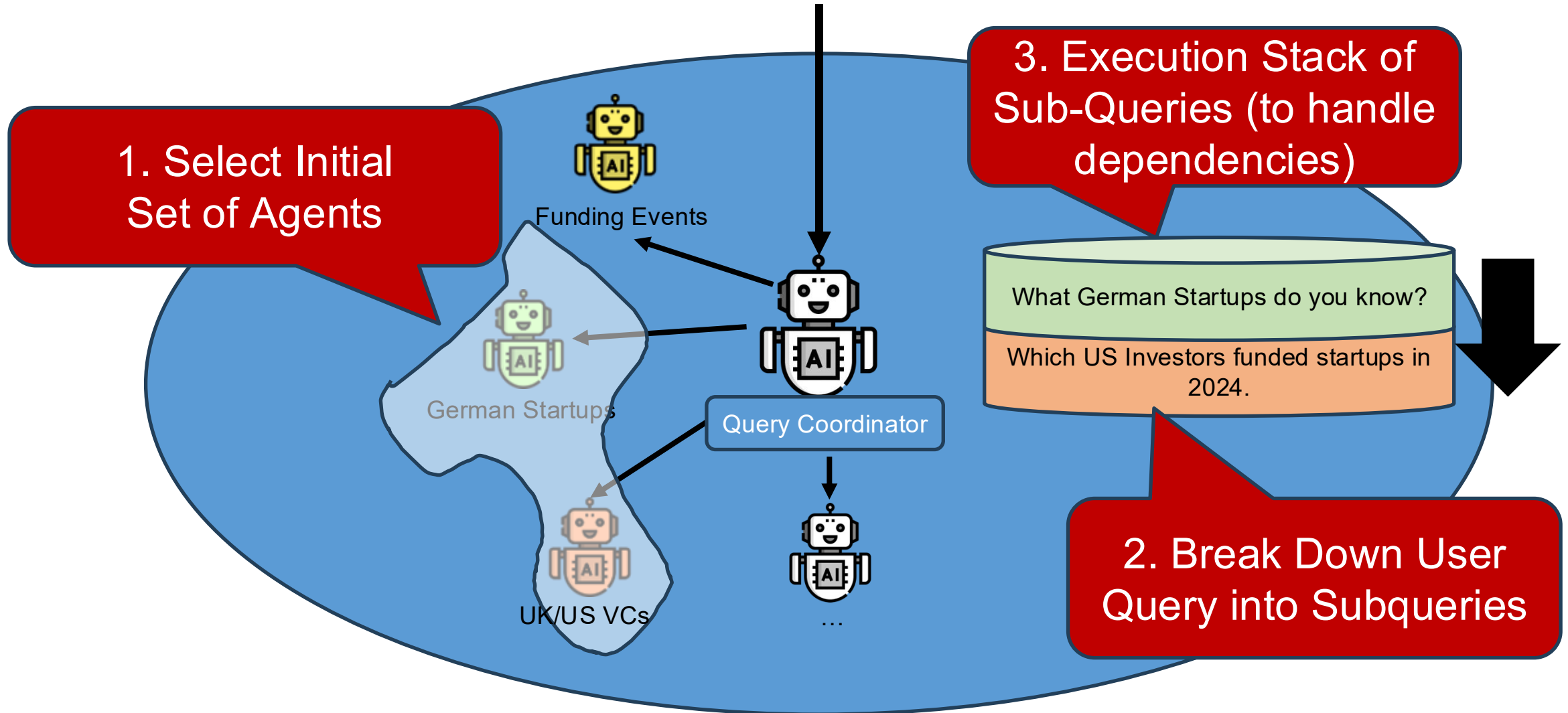
Which German Startups Raised US Funding in 2024?



Our Prototype: QCP-DB

User Query:

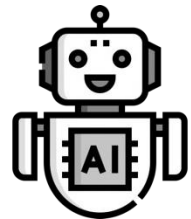
Which German Startups Raised US Funding in 2024?



Our Prototype: QCP-DB Agent Level

User Query:

Which German Startups Raised US Funding in 2024?



Query Coordinator

Execution Stack:

What German Startups do you know?
Which US Investors funded startups in 2024.

Stack Subquestion 1

What German Startups do you know?

I know 2 German startups.

[TechBerlin, MunichBio]

Agent Provides Result

Agent Solves Locally



Local Text-to-SQL

German Startups

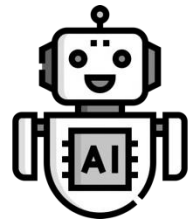
German Startups

Cmp-Name	Evaluation	Employees
TechBerlin	1€	1
MunichBio	2€	23

Our Prototype: QCP-DB Agent Level

User Query:

Which German Startups Raised US Funding in 2024?



Query Coordinator

Execution Stack:

What German Startups do you know? 

Which US Investors funded startups in 2024.

Stack Subquestion 1

Which US Investors funded startups in 2024.

I know 2 US VCs but I have no information about 2024.

[Accel, Sequoia]

Did these VCs fund startups in 2024?



UK/US VCs

No Year Information!

Local Text-to-SQL

UK/US VCs

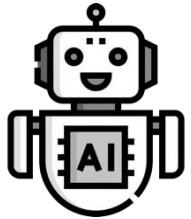
VC-Name	Country
Accel	US
Sequoia	US
Index	UK

Agent Provides Result + New Subquestion (which agent can not answer).

Our Prototype: QCP-DB Agent Level

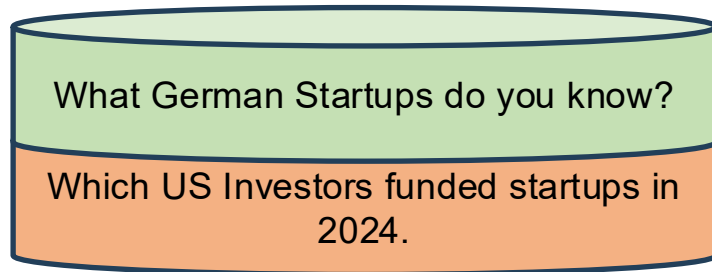
User Query:

Which German Startups Raised US Funding in 2024?

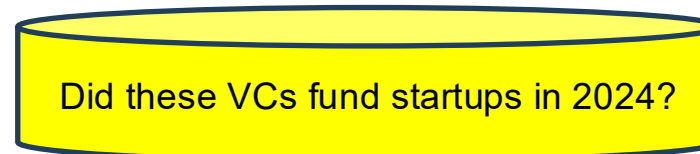
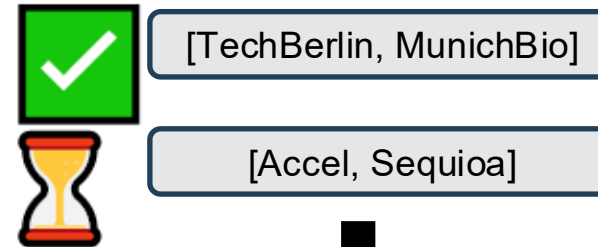


Query Coordinator

Execution Stack:



Result per Subquery



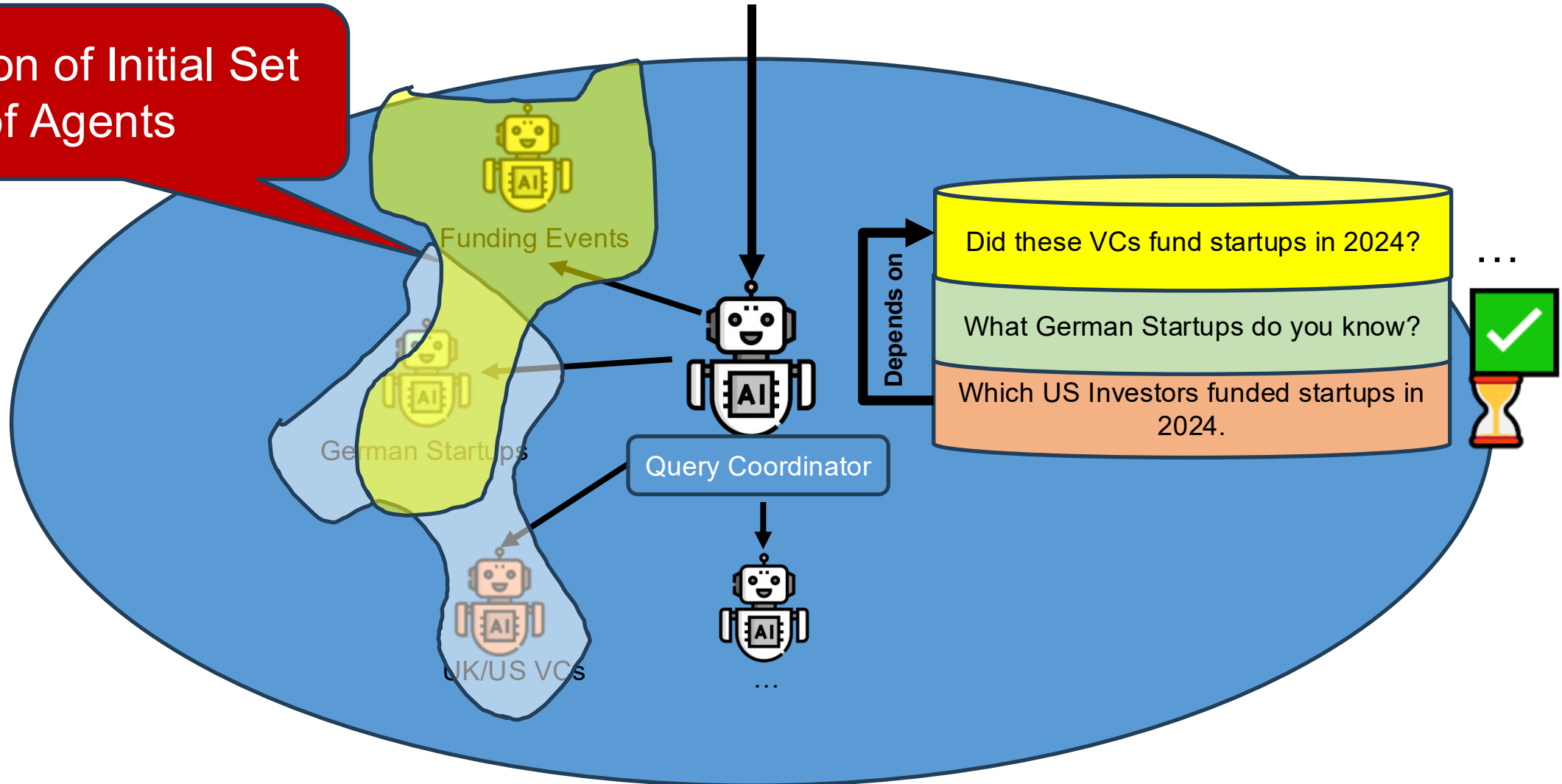
Agent created Subquestion.

Our Prototype: QCP-DB Plan Adaption

User Query:

Which German Startups Raised US Funding in 2024?

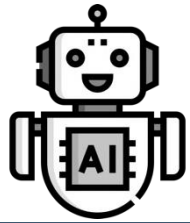
Adaption of Initial Set of Agents



Our Prototype: QCP-DB Agent Level

User Query:

Which German Startups Raised US Funding in 2024?



Query Coordinator

Stack Subquestion 3

[Accel, Sequoia]

Coordinator includes relevant prior results

Did these VCs fund startups in 2024?



Local Text-to-SQL

Execution Stack:

Funding Events

Yes.

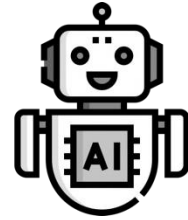
[
[TechBER, ACC, 2024],
[MUCBIO, SEQ, 2024]
]

Agent Provides Result

Funding Events

Cmp-Name	VC	Year
TechBER	ACC	2024
MucBIO	SEQ	2024

Our Prototype: QCP-DB Result Aggregation



Results Complete

Partial Results

[TechBerlin, MunichBio]

Startups in Germany

[
[TechBER, ACC, 2024],
[MUCBIO, SEQ, 2024]
]

Funding Events

[Accel, Sequoia]

VCs in US

Result Agent

```
Generated Code:  
def normalize: ...  
TechBerlin = TechBER  
...
```

```
Generated Code:  
def normalize: ...  
ACC = Accel ...
```

LLM-Guided Result
Synthesis (by code gen)

Final Result

Merge(...)

The startups are:
[TechBerlin, MunichBio]

1. Aligning Data
Sources

2. Merging Partial Results
into Final Answer

QCP-DB: Evaluation

What do we need?

- Multiple independent data sources
- NL queries and ground truth results

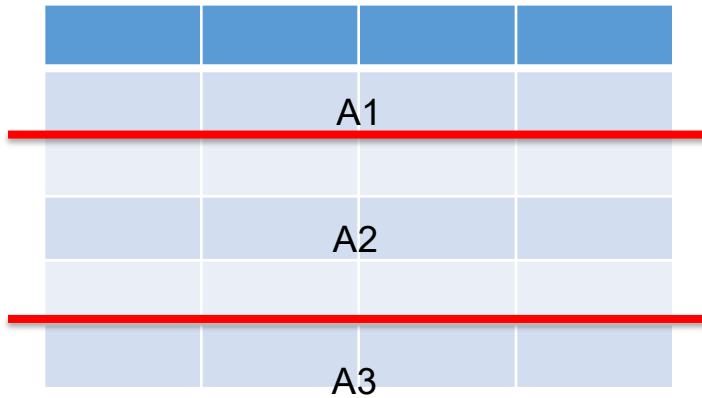
Text-to-SQL benchmarks offers:

- NL questions and ground truth results
- BUT only in one clean DB and NOT in independent data sources

Our Fix: Split a Text-to-SQL data sets (BIRD) into multiple independent data sources + noise the splits to simulate heterogeneous schema/data

Evaluation Settings: Split Types

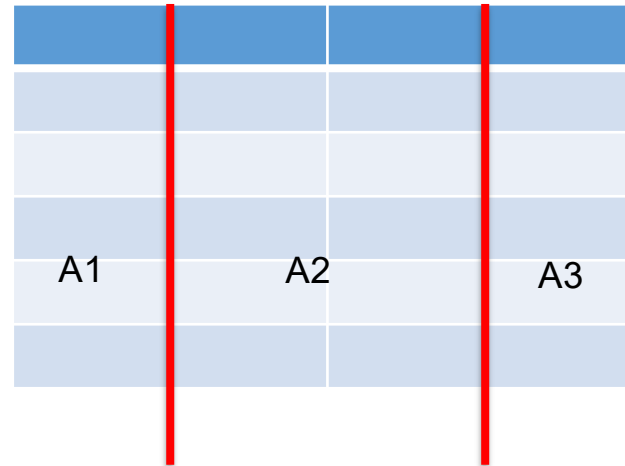
Horizontally
(simple)



A 4x4 grid with a blue header row. Two horizontal red lines divide the grid into three rows. The middle row is labeled 'A1', the second row from the bottom is labeled 'A2', and the bottom row is labeled 'A3'.

		A1	
		A2	
		A3	

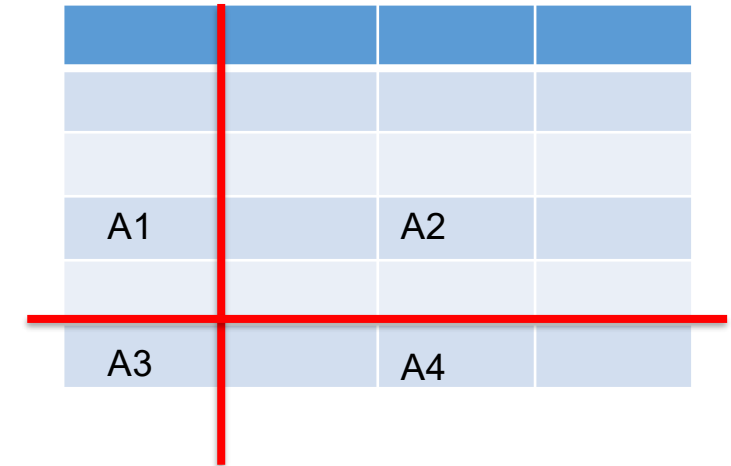
Vertically
(simple)



A 4x4 grid with a blue header row. Two vertical red lines divide the grid into three columns. The leftmost cell is labeled 'A1', the middle cell is labeled 'A2', and the rightmost cell is labeled 'A3'.

A1		A2	
			A3

Randomly
(difficult)

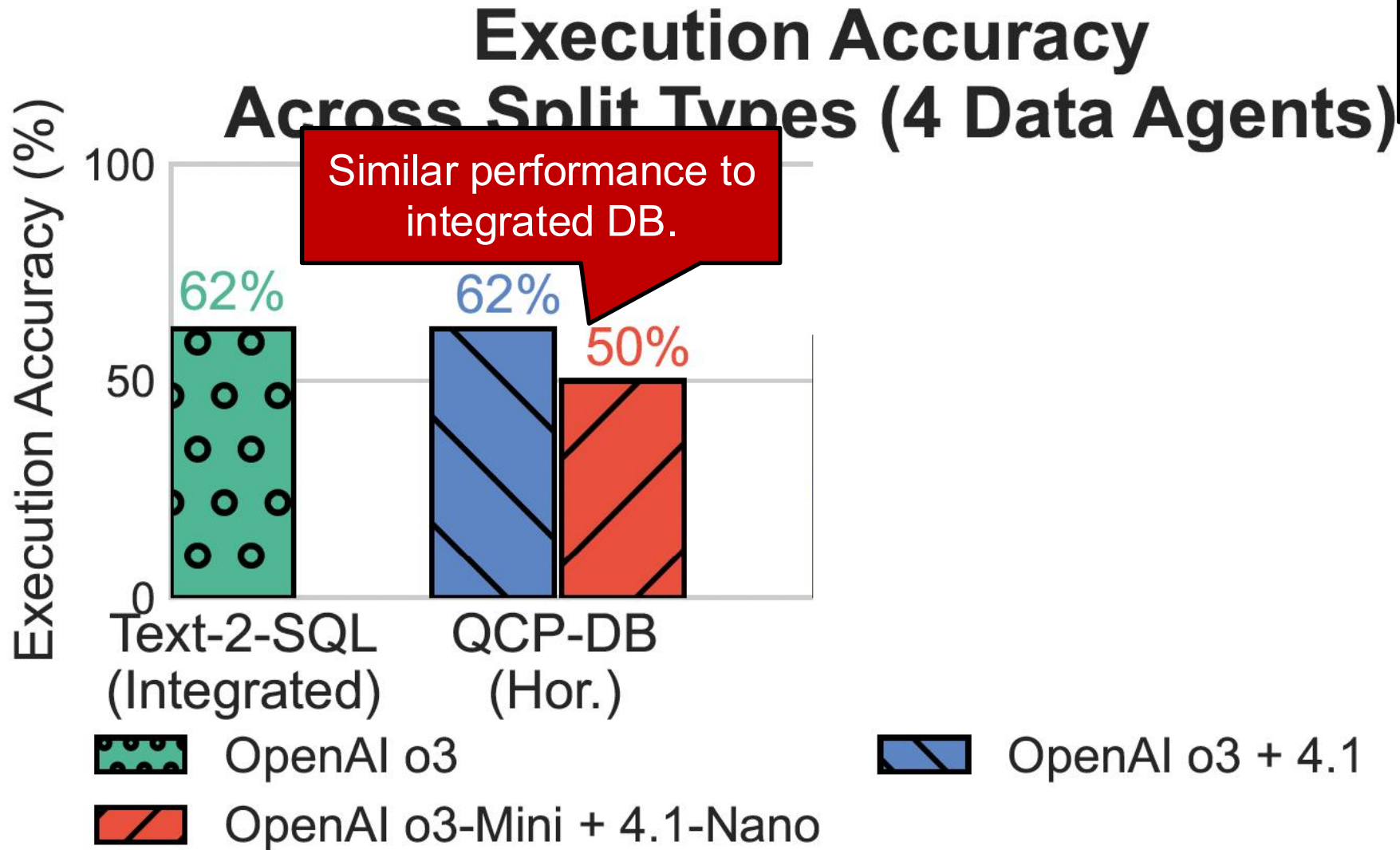


A 4x4 grid with a blue header row. A vertical red line is between the first and second columns, and a horizontal red line is between the second and third rows. The cell at row 2, column 1 is labeled 'A1', the cell at row 2, column 2 is labeled 'A2', the cell at row 3, column 1 is labeled 'A3', and the cell at row 3, column 2 is labeled 'A4'.

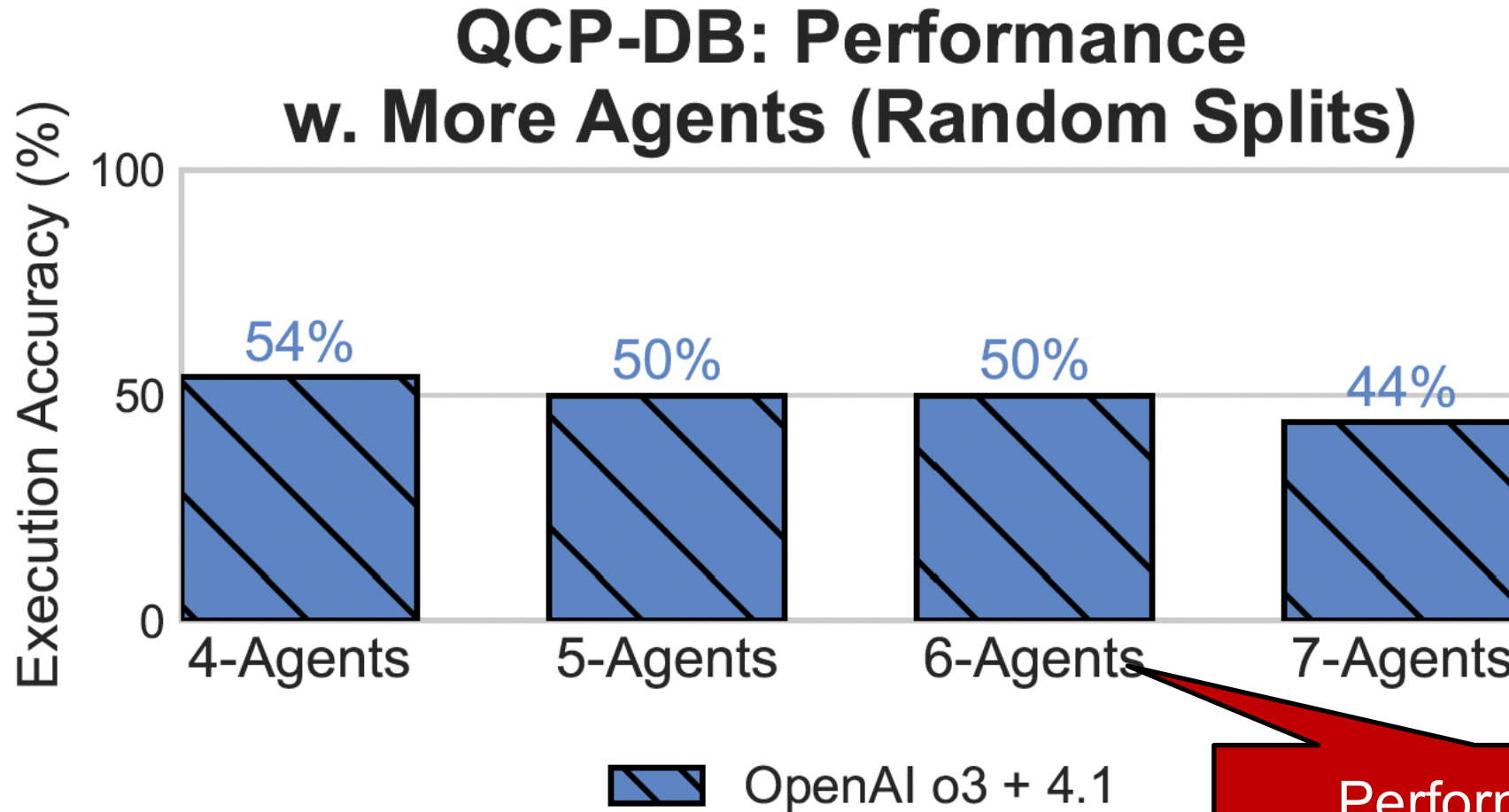
A1		A2	
A3		A4	

All splits are “noised” to have different representations in different heterogeneous data sources (e.g., United States vs US)

Evaluation: Initial Results



Evaluation: Impact of More Agents



Performance is resilient to more agents

Query-by-Collaboration: Opportunities

Opportunity 3:
Zero-overhead for
new data sources

*I know Startups
in Africa*

Heterogeneous & A
Data Sources

*NL over Text
Documents*

Opportunity 2:
Any modality (image,
text, graph, ...)

**Ultimately: Query the World –
Query Answering at Internet-Scale**

Opportunity 1: Multi-
lingual data sources

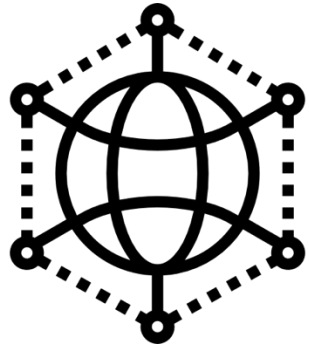
Startups-
FR (French)

*NL(EN)-to-
SQL(French)*

**Collaboration
Language**

*Text-to-
SPARQL*

Query-by-Collaboration: Future Directions



Scale from Few Data Agents to Internet-Scale

Query execution at increasingly larger scale?



Query Execution & Optimization is different

Different from Classical DBs: “plan” is not known upfront, trustworthiness of sources



Other Open Challenges

Completeness of results, complex forms of overlap, missing data, ...

Thanks for your Attention

