A Vision for Autonomous Data Agent Collaboration

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

Timo Eckmann & Carsten Binnig
Systems @ TU Darmstadt

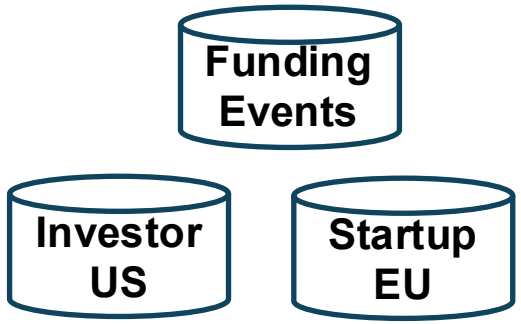
The Relational “Pipe Dream”

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

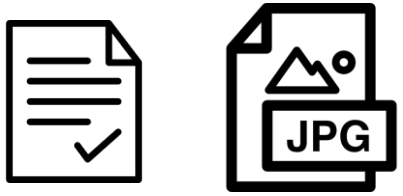
FundingDB			
Cmp-Name	**Cmp-Ctry**	**VC-Name**	**VC-Ctry**
TechBerlin	DE	Accel	USA
ParisAI	FR	Index	UK
MunichBio	DE	Sequoia	USA

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

Query-by-Integration

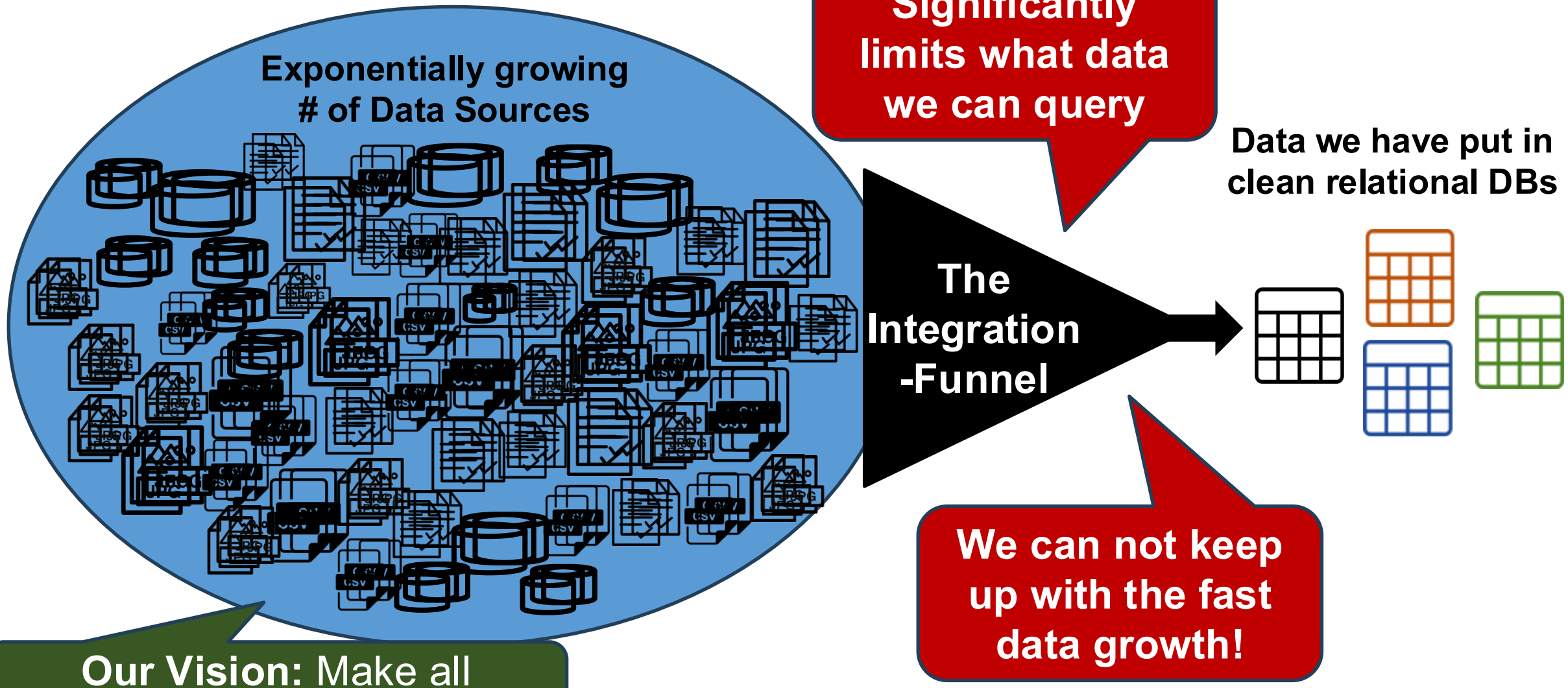
When did ParisAI get funding?

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

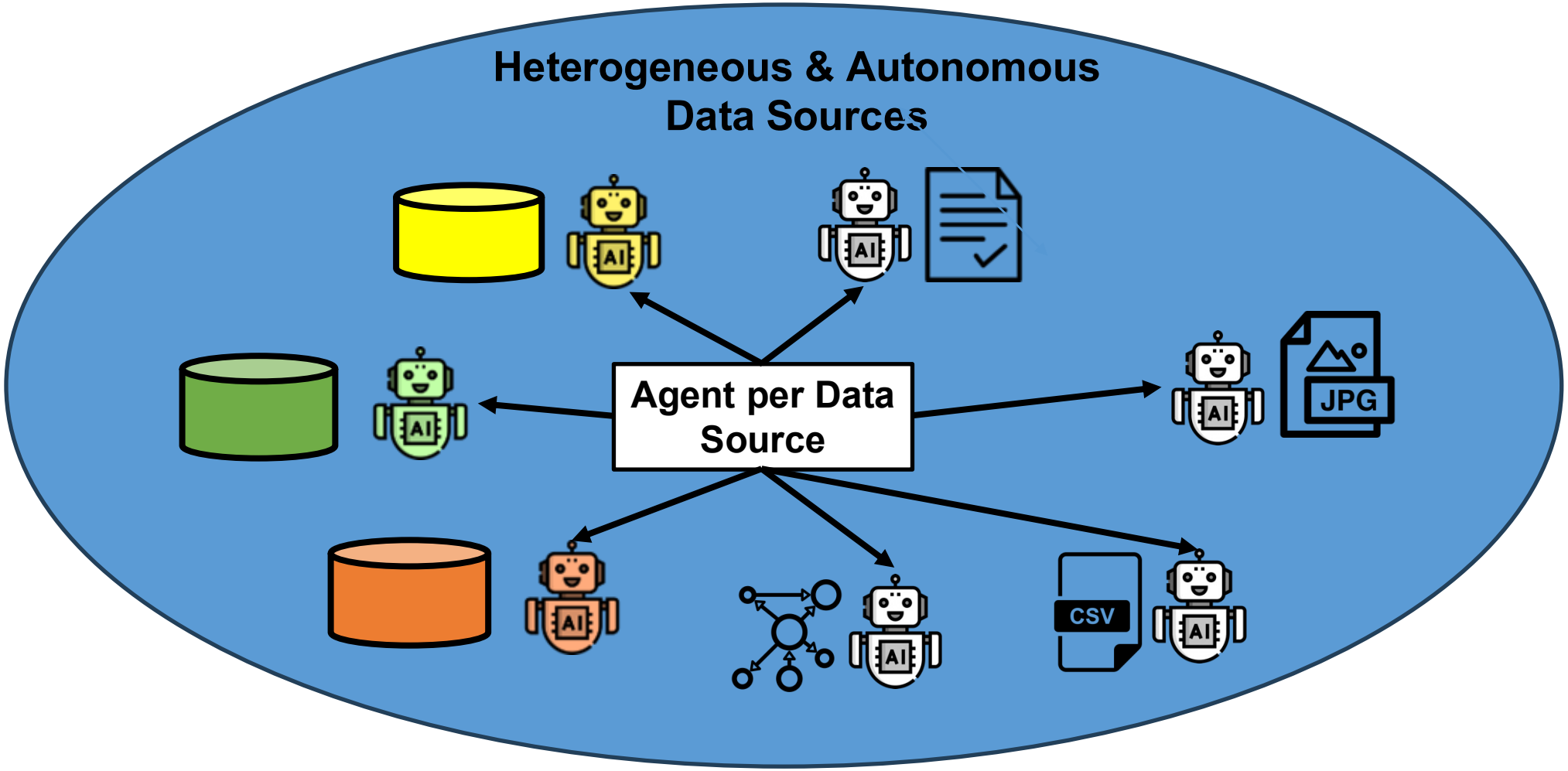
Find & integrate “Funding year”?

Limitation 2: Queries limited by integrated database

Why Query-by-Integration is doomed?

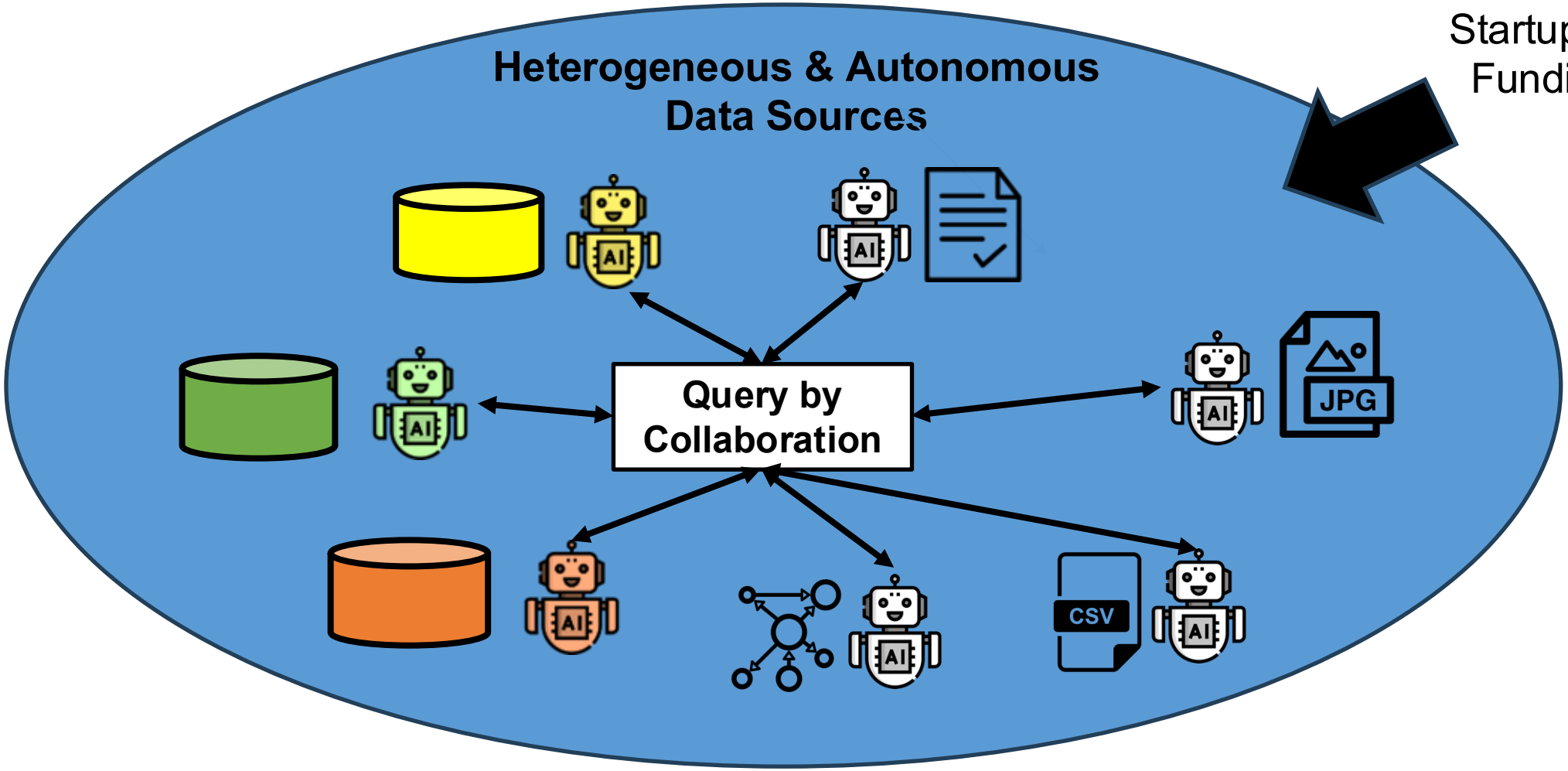


Our Vision: Query-by-Collaboration



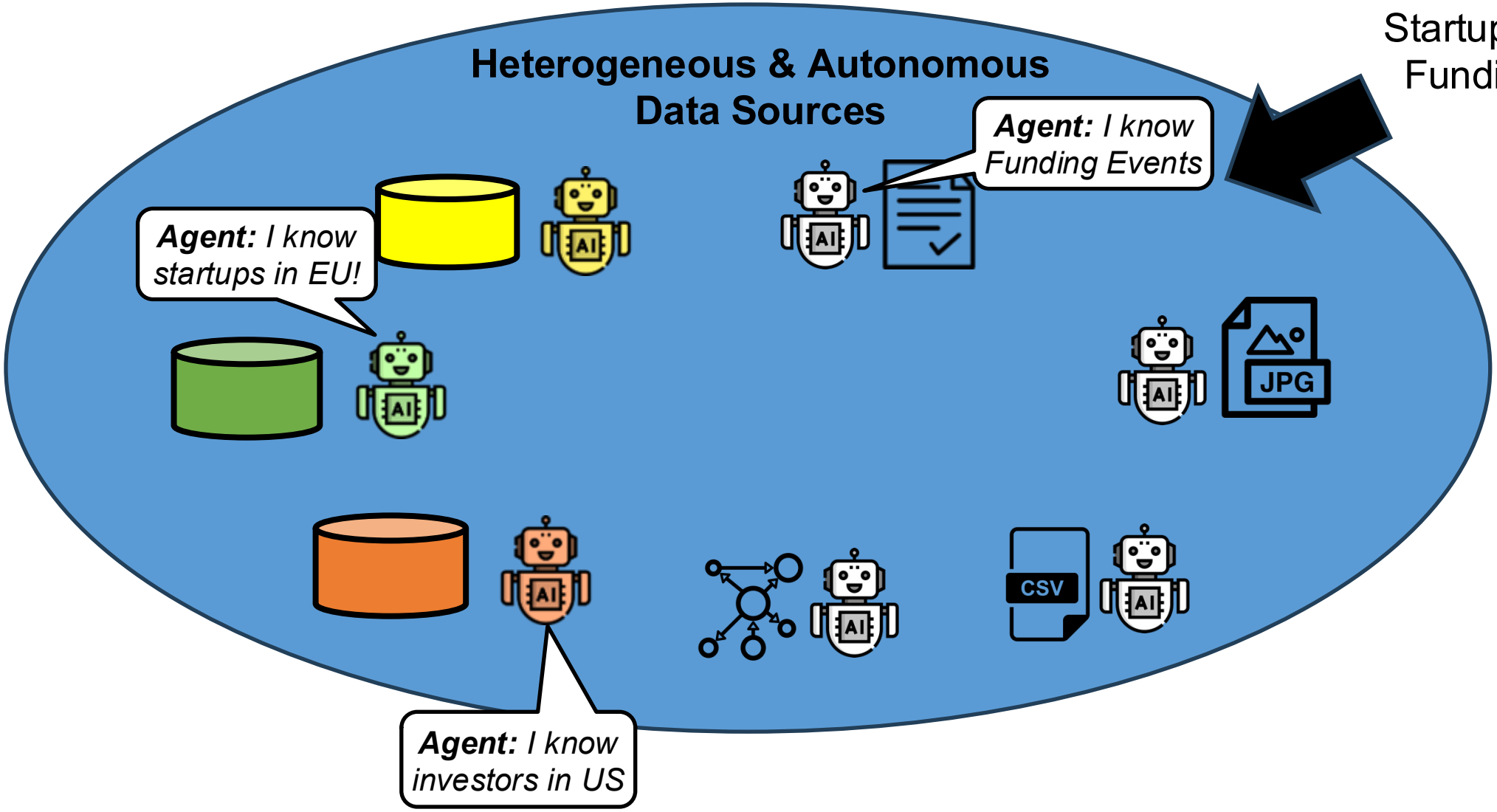
Our Vision: Query-by-Collaboration

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



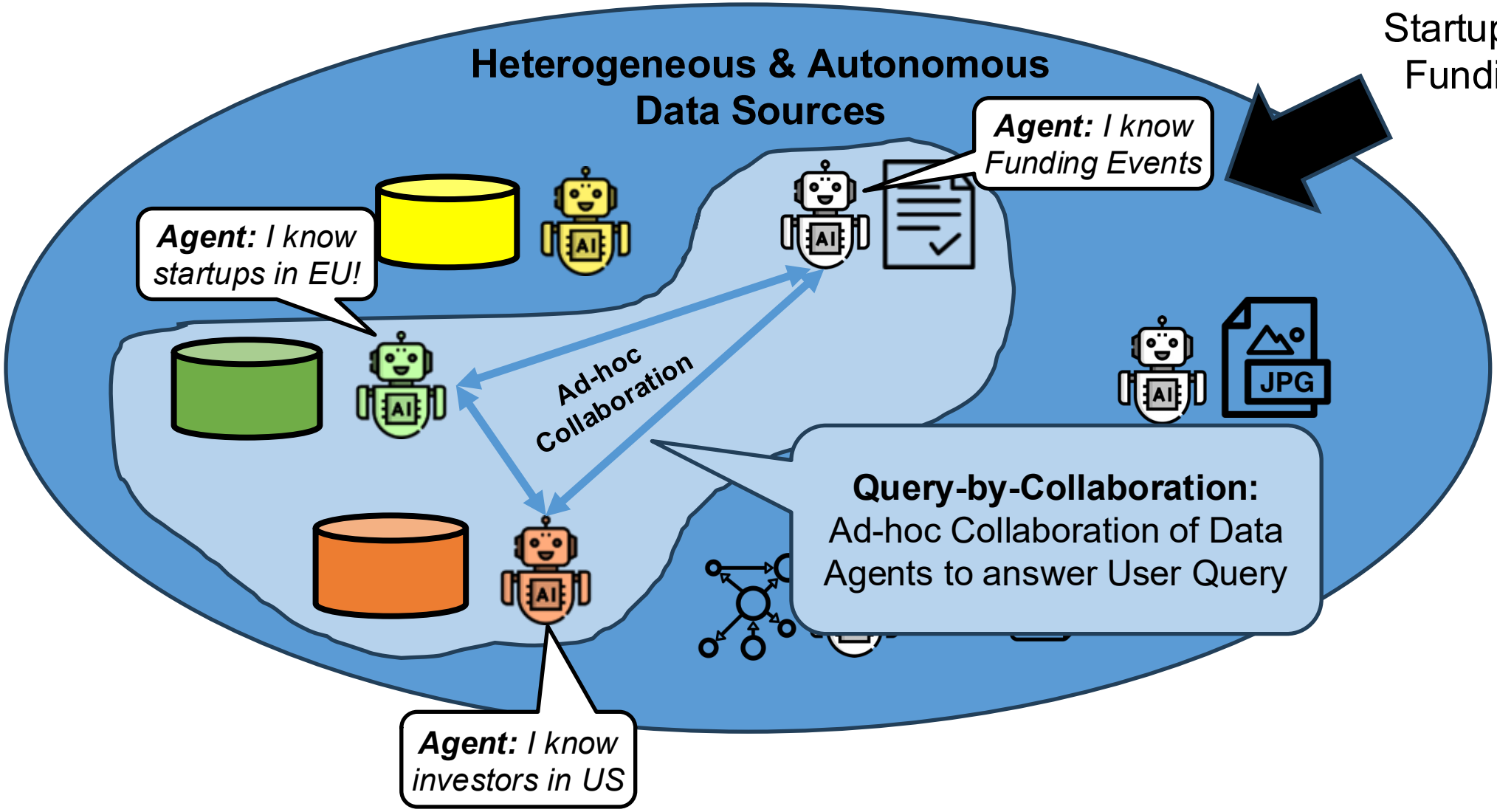
Our Vision: Query-by-Collaboration

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



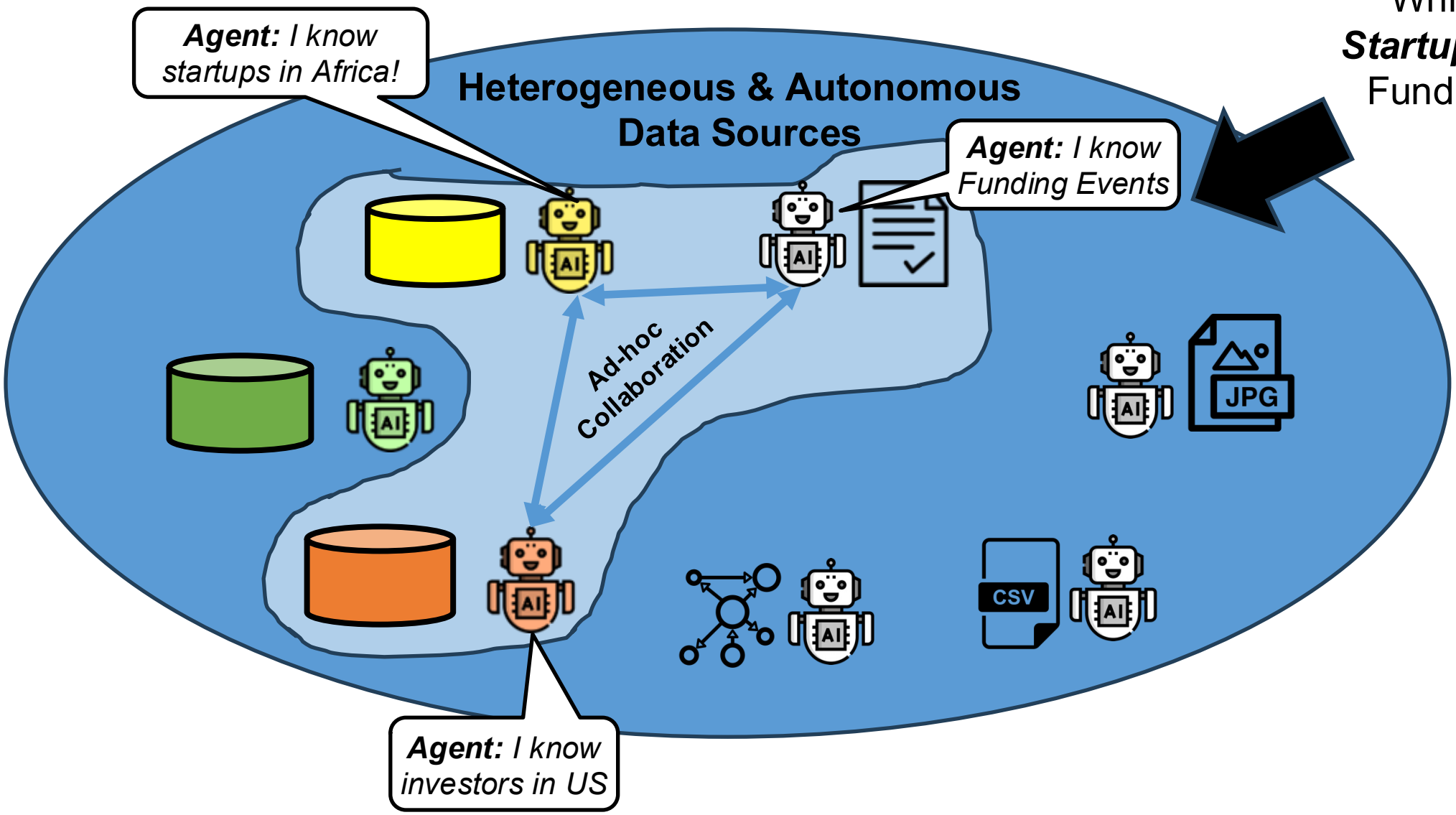
Our Vision: Query-by-Collaboration

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



Our Vision: Query-by-Collaboration

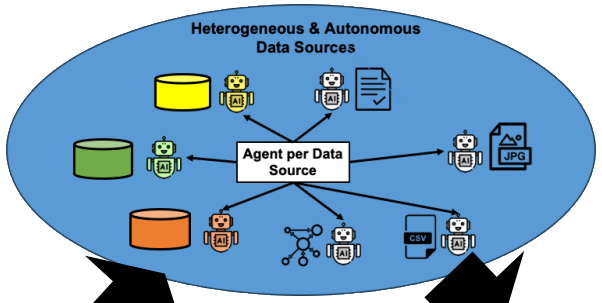
Another User Query:
Which ***African Startups*** Raised US Funding in 2024?



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)!



Which German Startups Raised US Funding in 2024?

NL Queries as users CAN not know all data

Query Coordinator

1. Select (initial) Agents & Sub-Queries

Which German Startups?

US VCs giving out funding in 2024

Partial Result

Partial Result

\+

BUT Missing Data: Funding Year!

Query Coordinator

2. Additional Agents & Sub-Queries

Funding years for German Startups?

Partial Result

Query Coordinator

Result Complete? Yes

Partial Results

\+ User Query

Result Agent

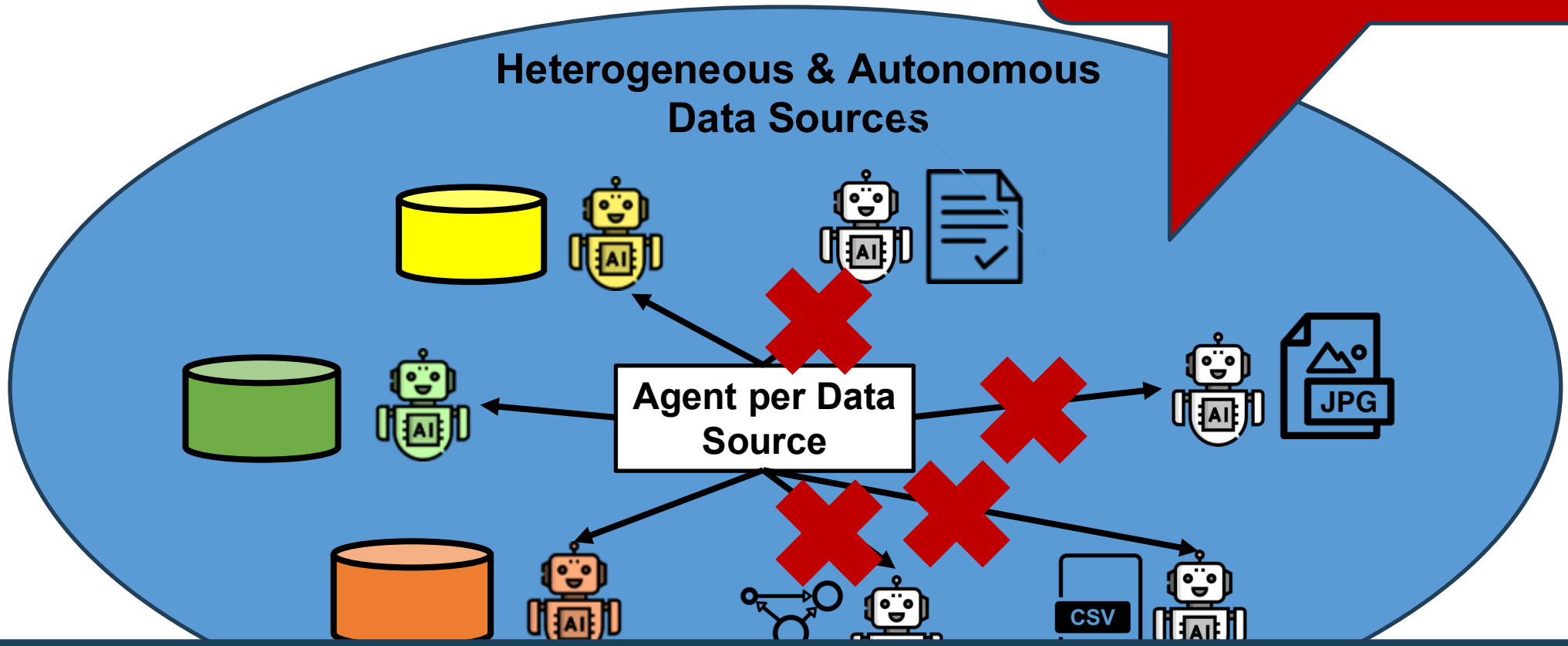
3. Result Synthesis

Merged & Aligned Result

Startup-Name
TechBerlin

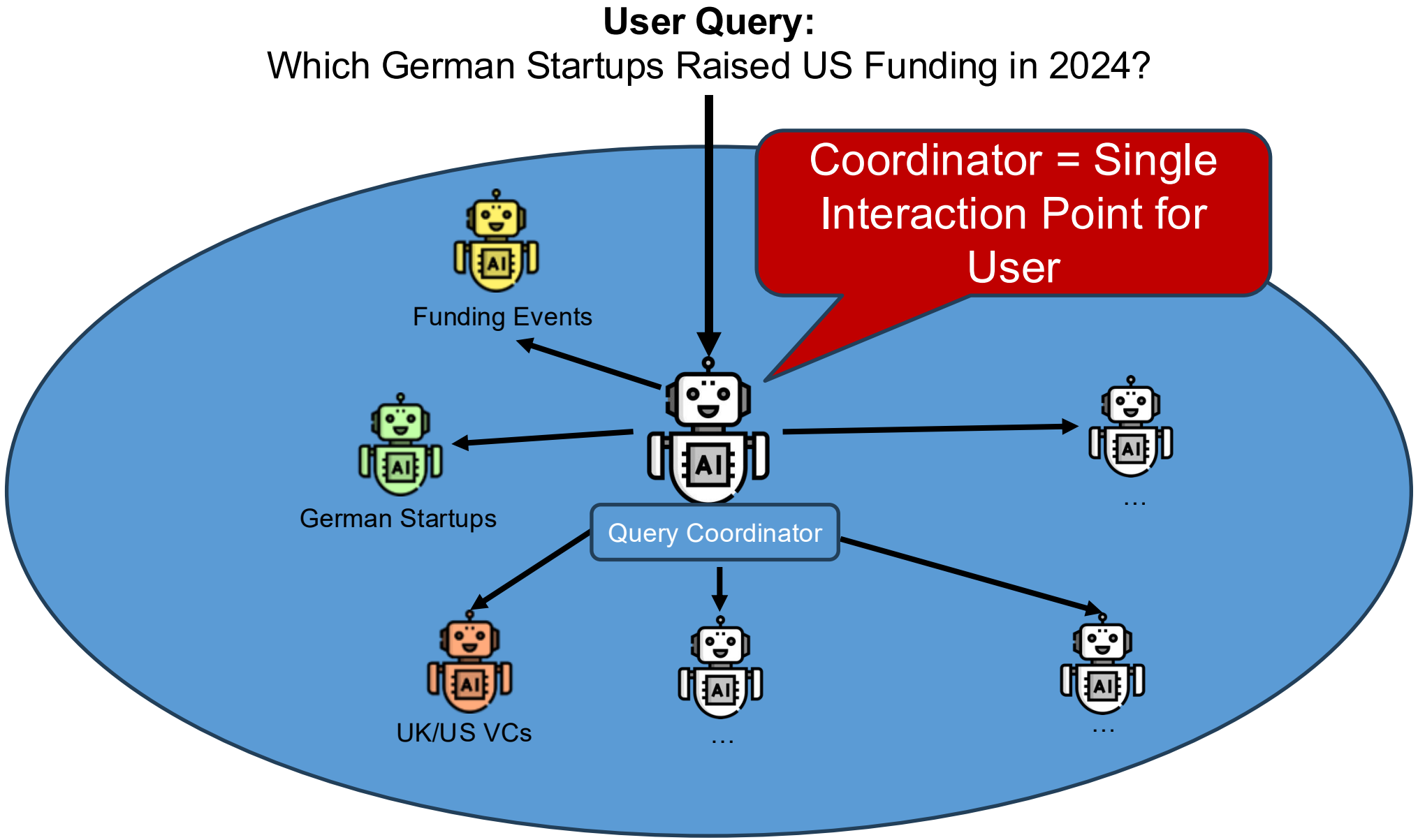
Our Prototype: QCP-DB

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

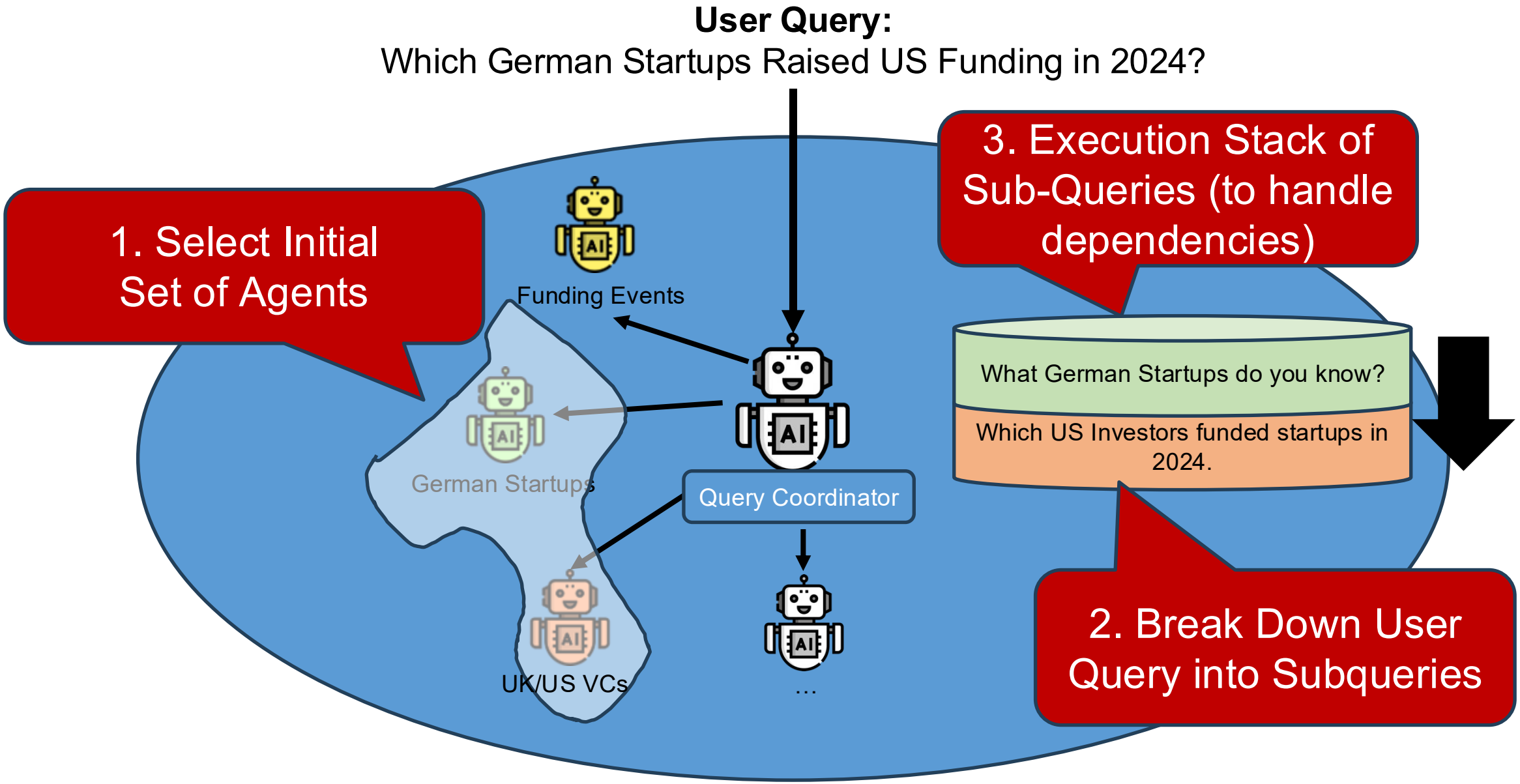


Goal of QCP-DB: Basic Understanding of how to design Query-by-Collaboration Systems

Our Prototype: QCP-DB

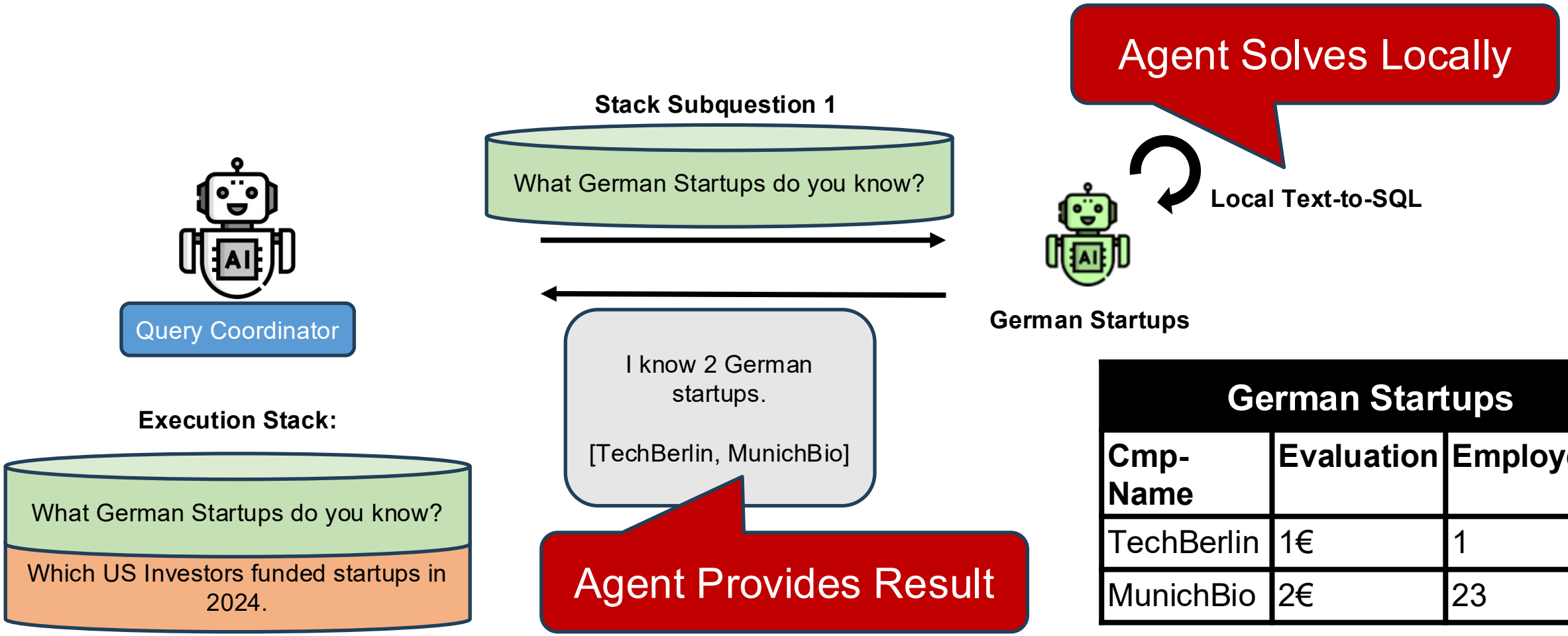


Our Prototype: QCP-DB



Our Prototype: QCP-DB Agent Level

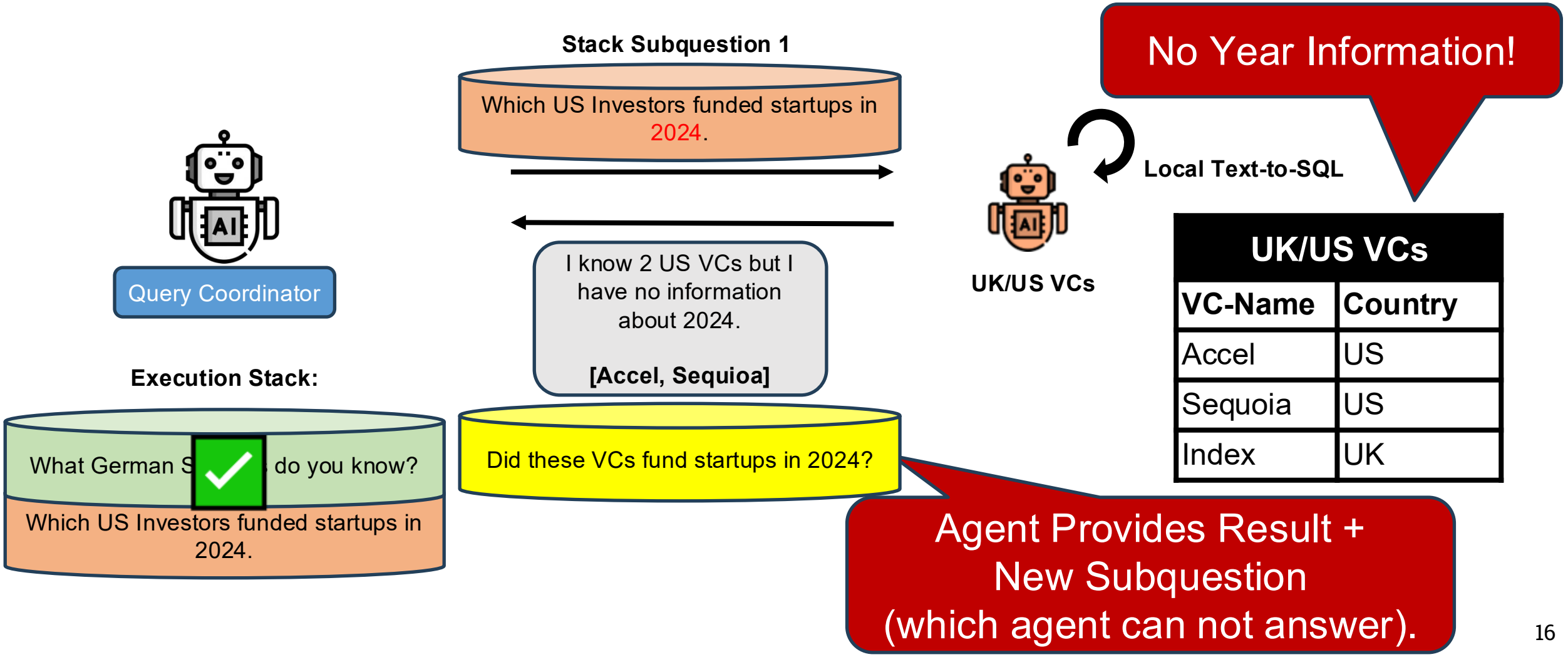
User Query:
Which German Startups Raised US Funding in 2024?



German Startups		
Cmp-Name	**Evaluation**	**Employees**
TechBerlin	1€	1
MunichBio	2€	23

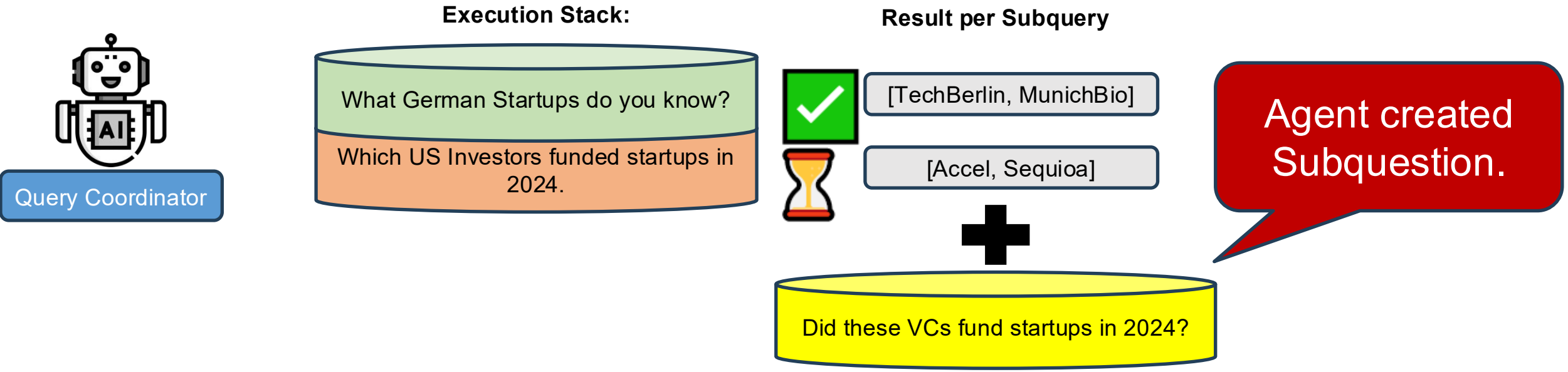
Our Prototype: QCP-DB Agent Level

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

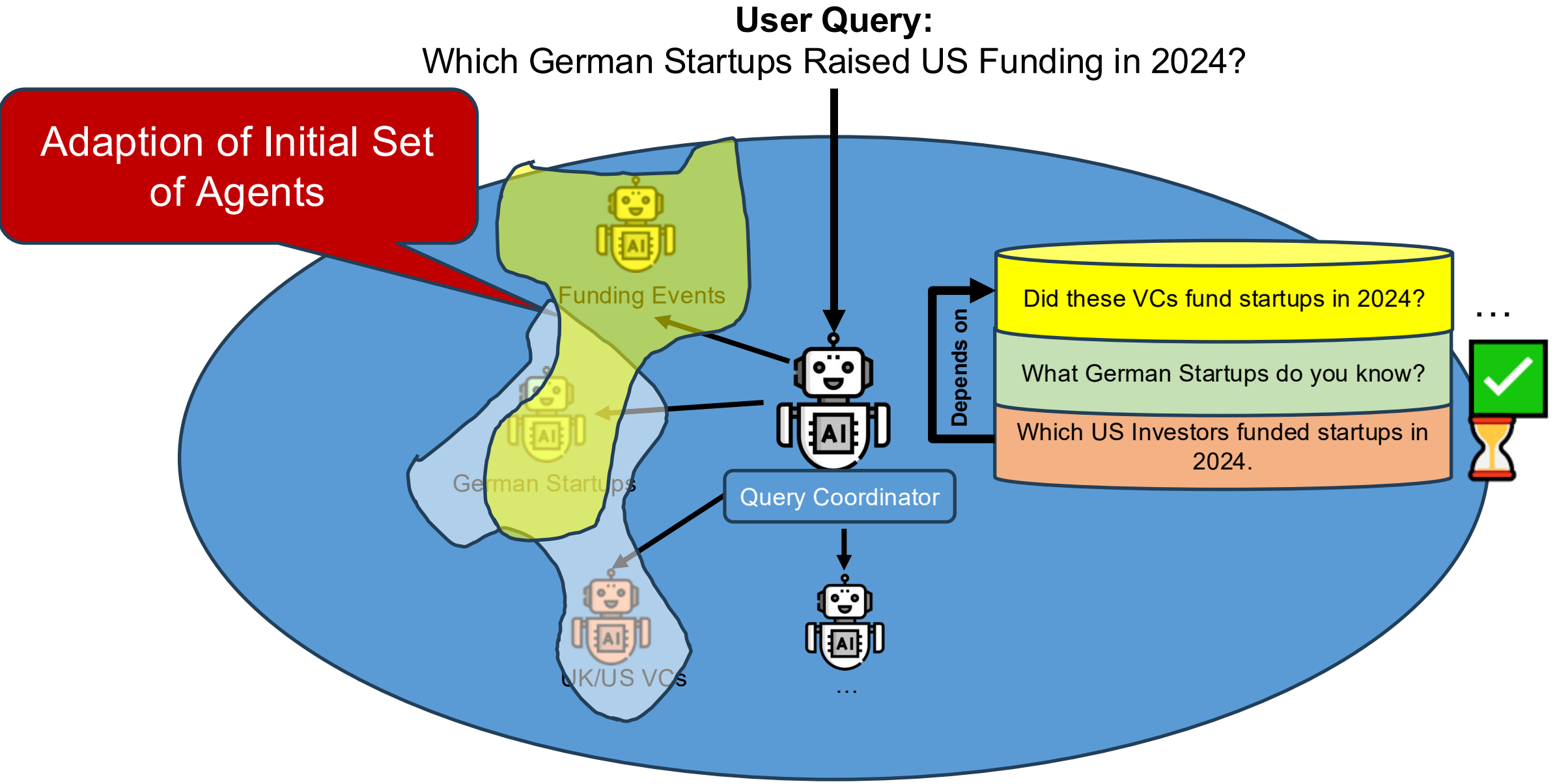


Our Prototype: QCP-DB Agent Level

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

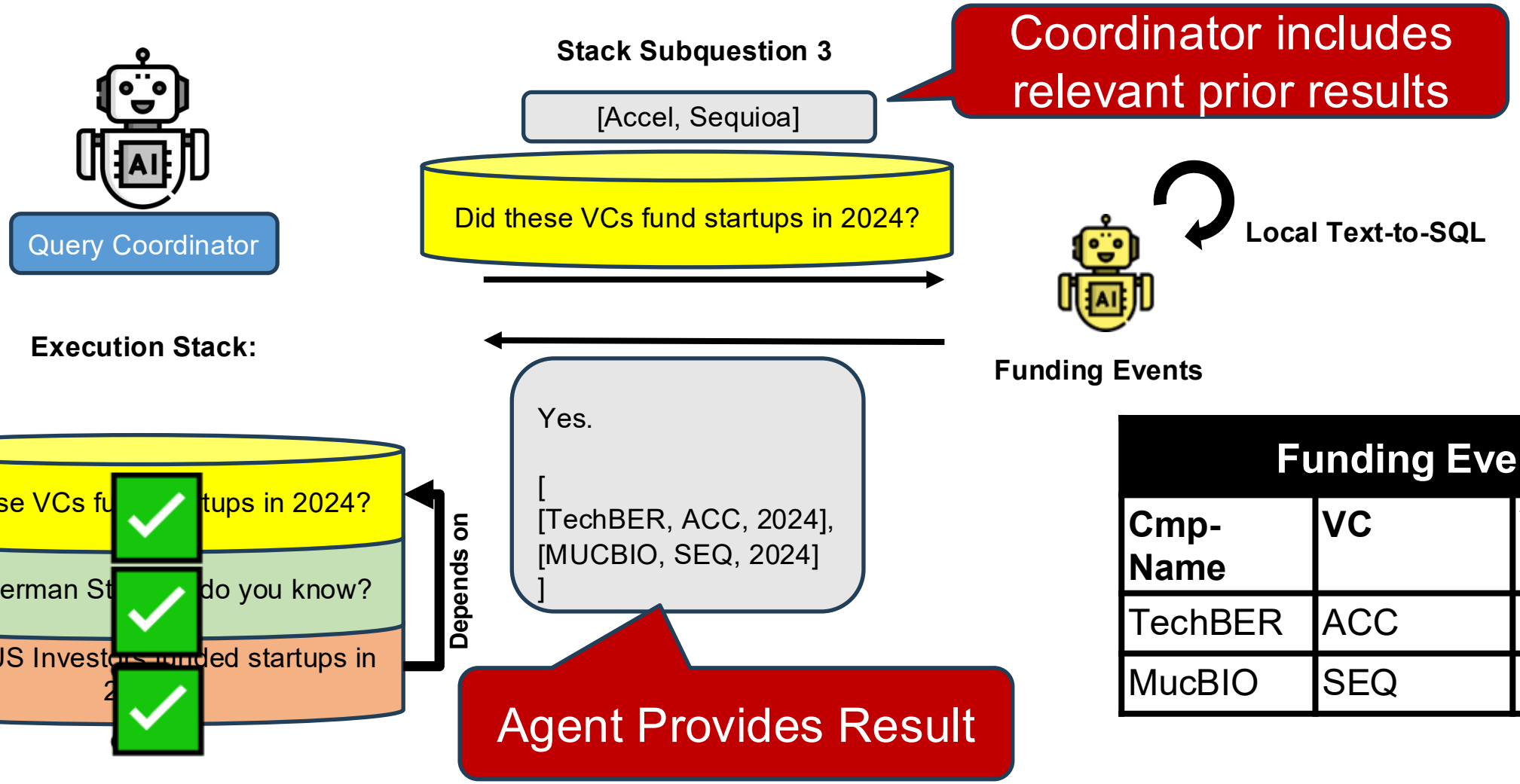


Our Prototype: QCP-DB Plan Adaption



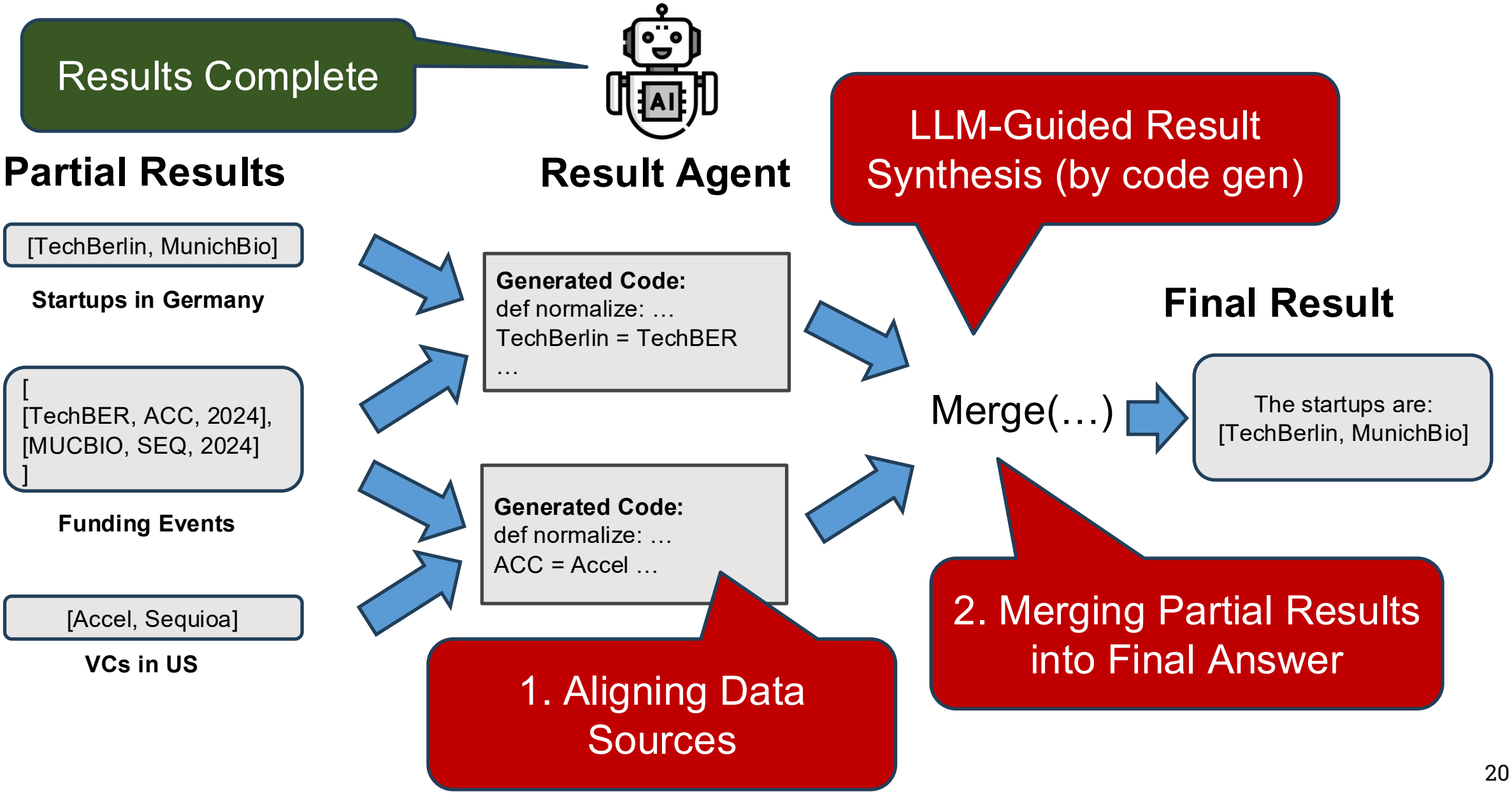
Our Prototype: QCP-DB Agent Level

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



Funding Events		
Cmp-Name	**VC**	**Year**
TechBER	ACC	2024
MucBIO	SEQ	2024

Our Prototype: QCP-DB Result Aggregation



QCP-DB: Evaluation

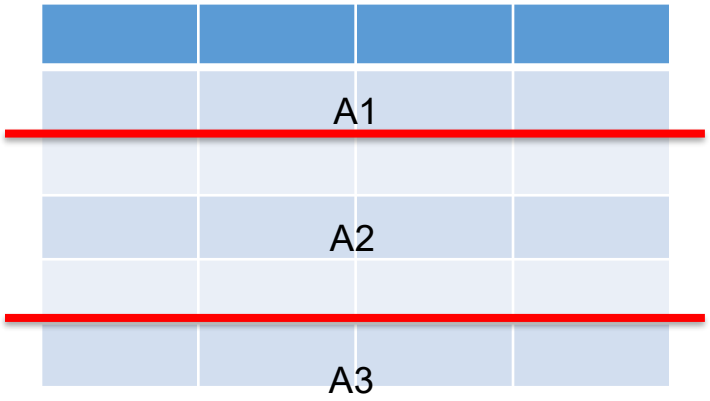
What do we need?
- Multiple independent data sources
- NL queries and ground truth results

Text-to-SQL benchmarks offers:
- NL questions and ground truth results
- BUT only in one clean DB and NOT in independent data sources

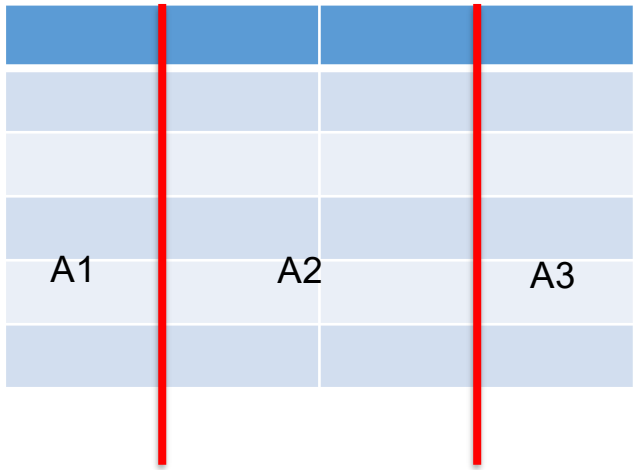
Our Fix: Split a Text-to-SQL data sets (BIRD) into multiple independent data sources + noise the splits to simulate heterogeneous schema/data

Evaluation Settings: Split Types

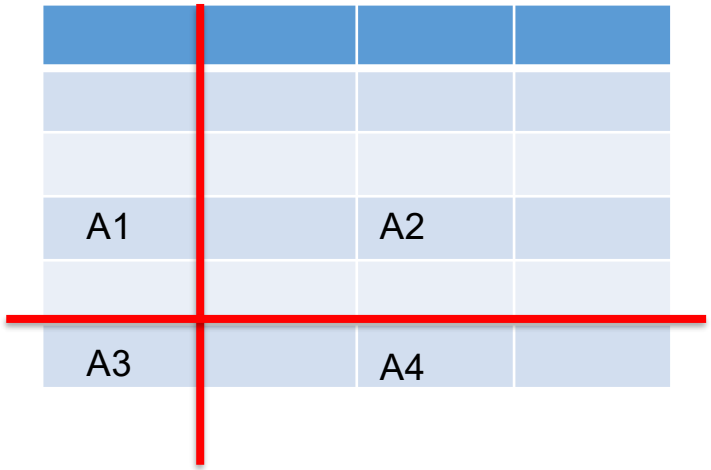
Horizontally (simple)



Vertically (simple)

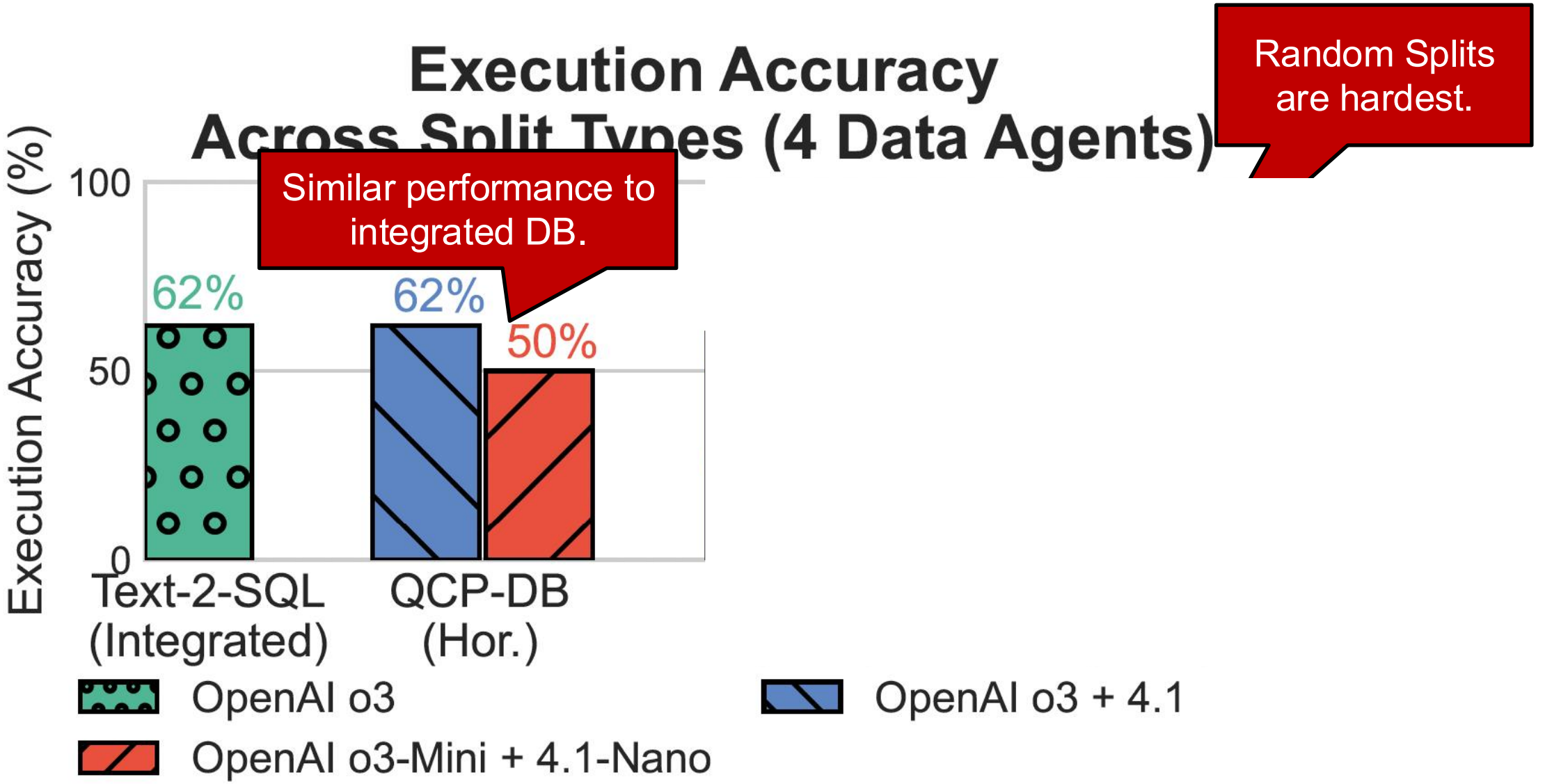


Randomly (difficult)

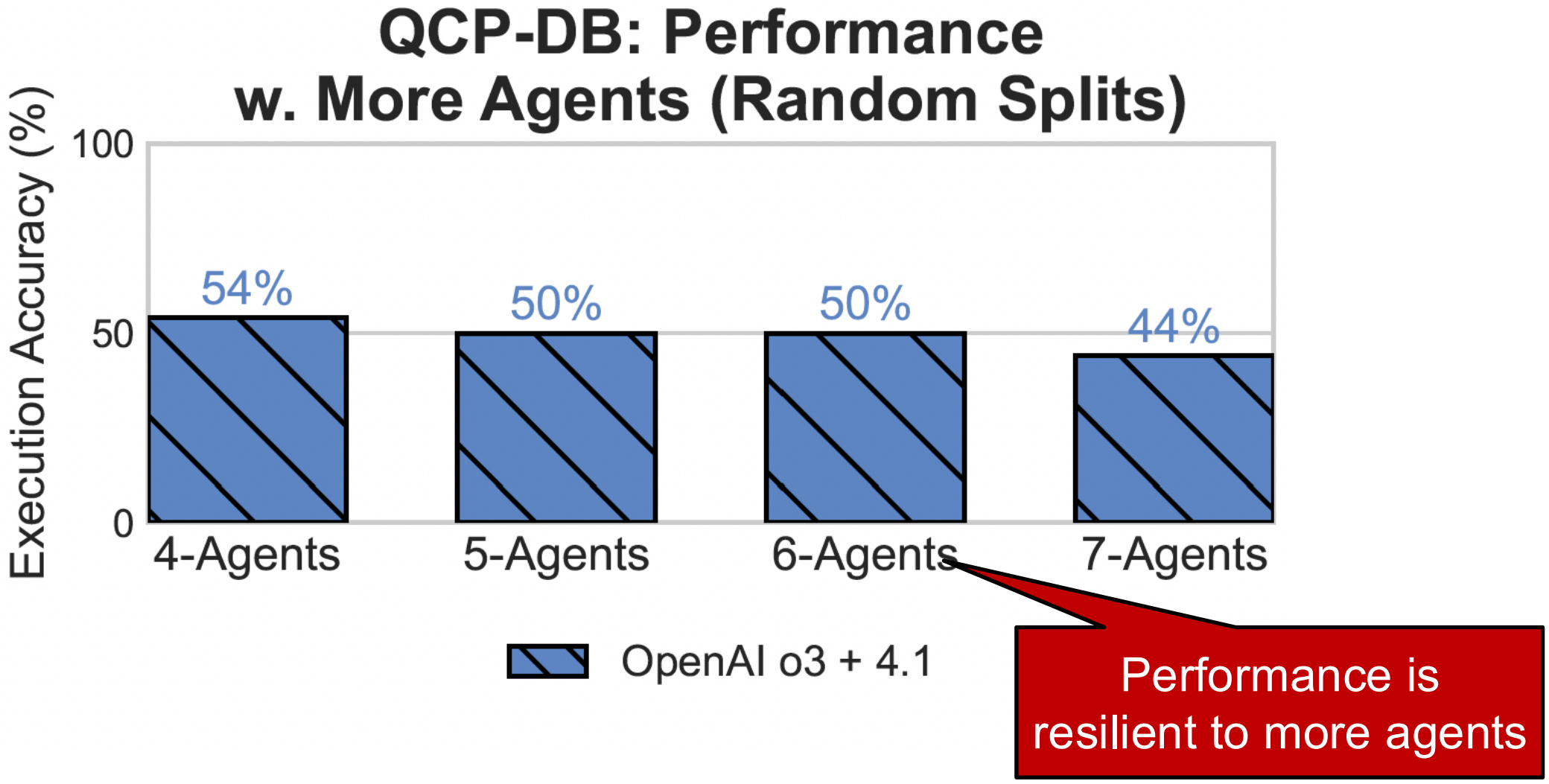


All splits are "noised" to have different representations in different heterogeneous data sources (e.g., United States vs US)

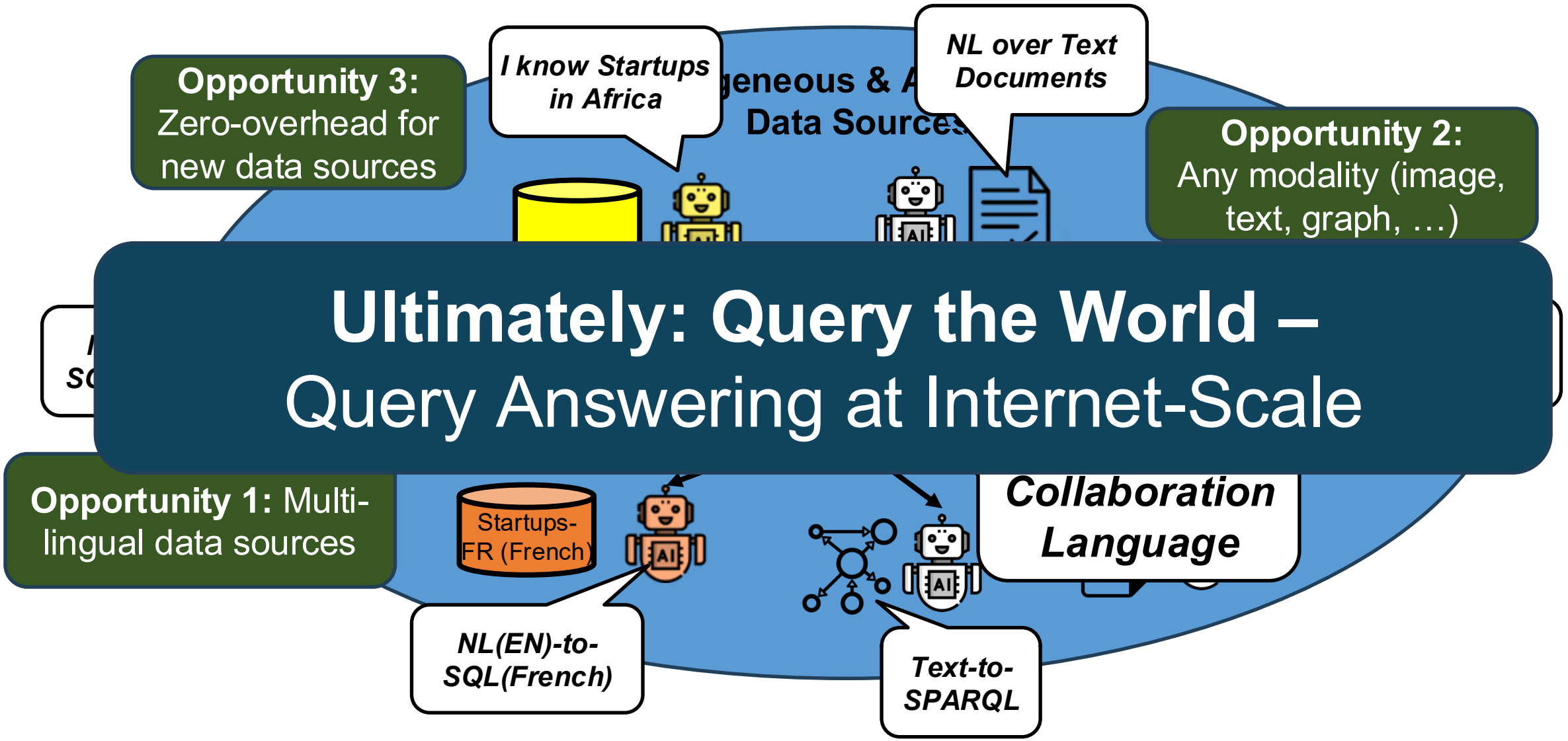
Evaluation: Initial Results



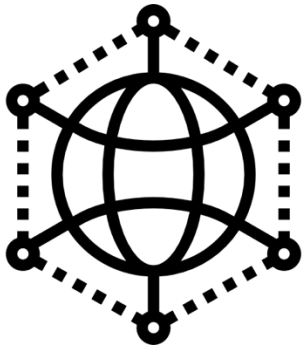
Evaluation: Impact of More Agents



Query-by-Collaboration: Opportunities



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