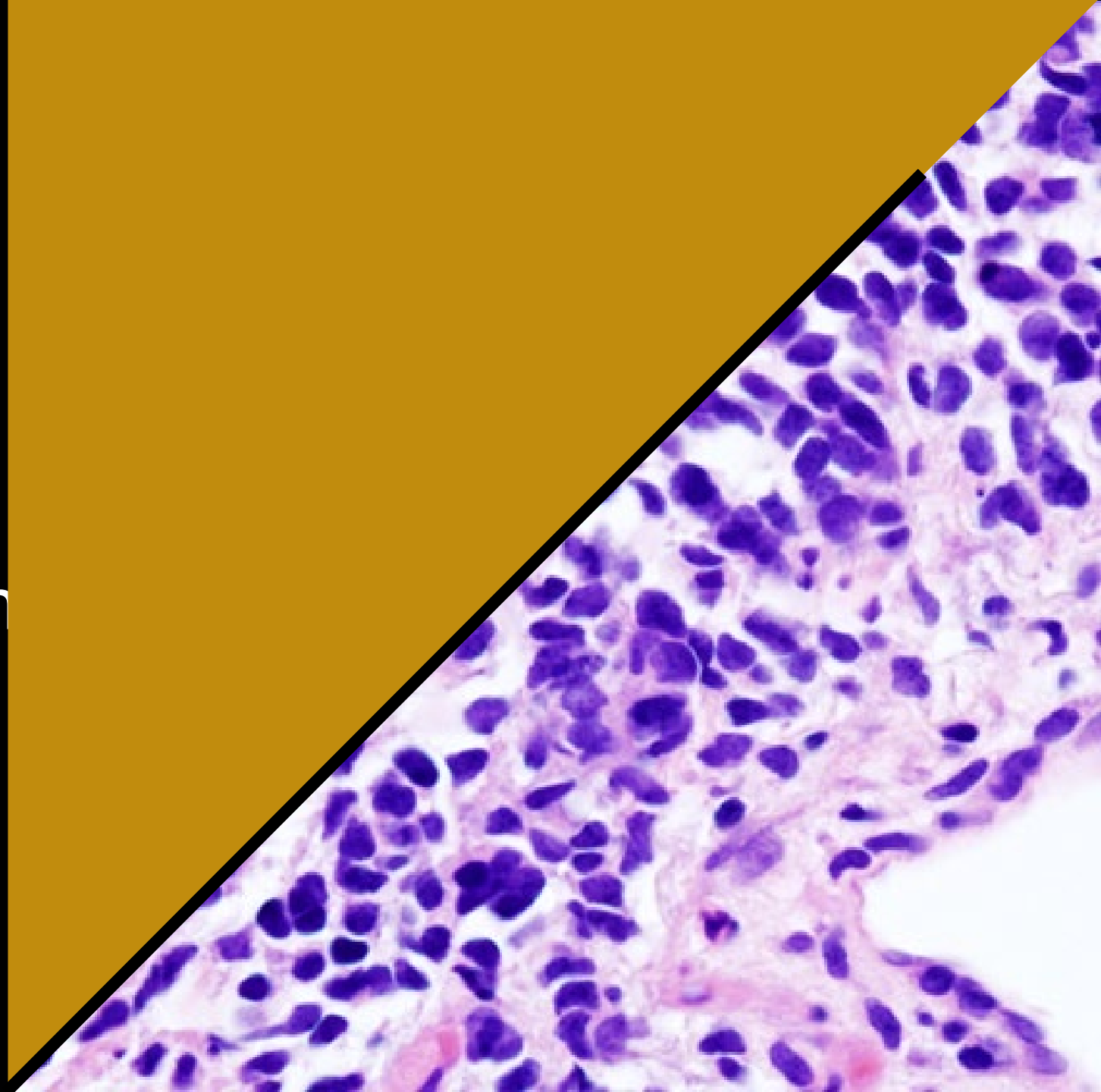
Agentic Data Systems and Multimodal Analytics

Gerardo Vitagliano

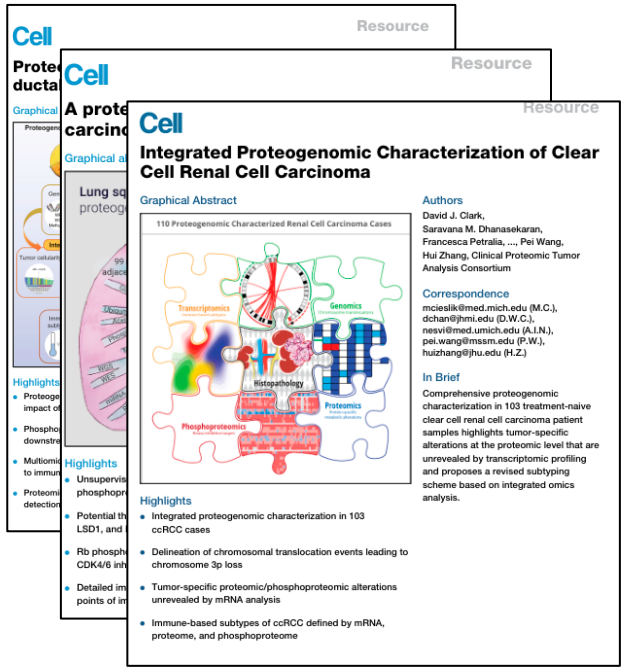
NHR4CES Workshop, May 11th

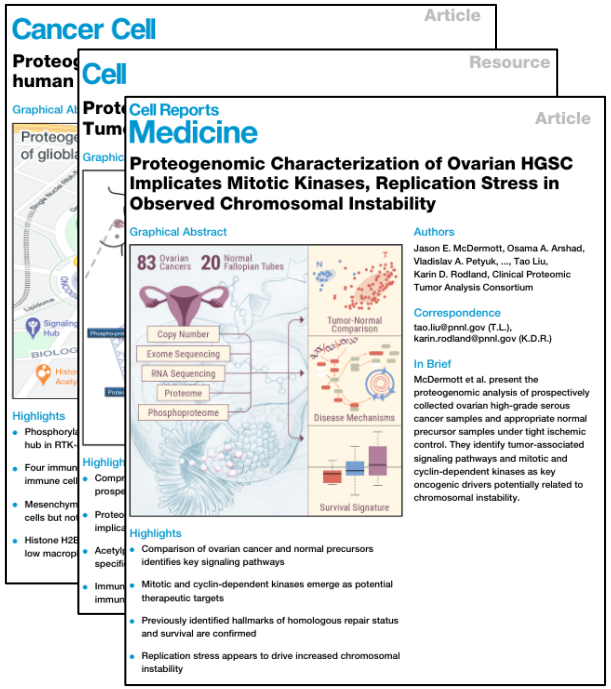
MIT CSAIL
Computer Science & Artificial Intelligence Laboratory

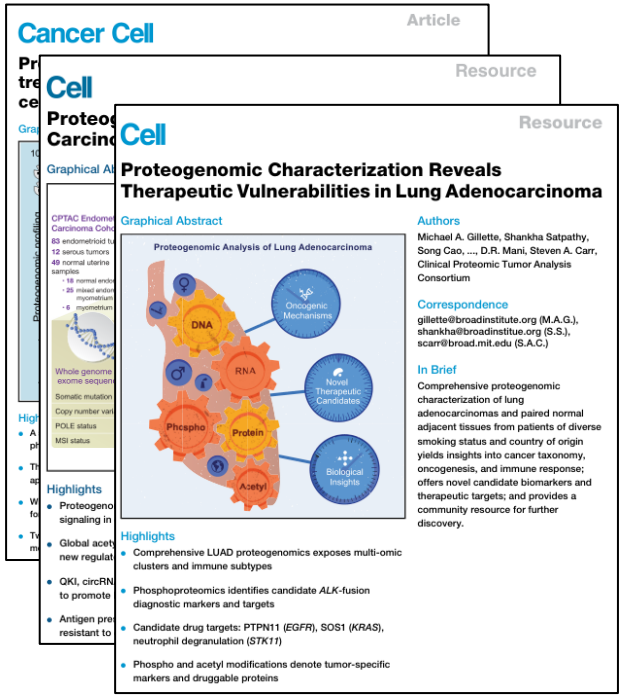


A complex analytical pipeline

- A researcher wants to run a medical study comparing patient data from several sources
  - The input contains PDF papers, images, genomics, and tabular data

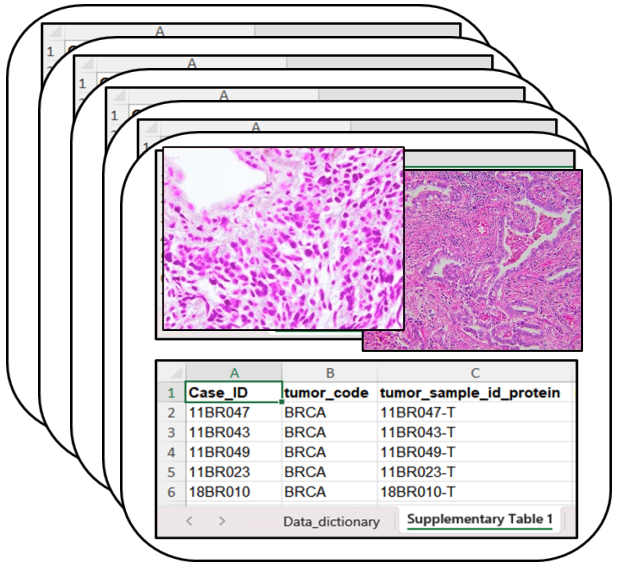


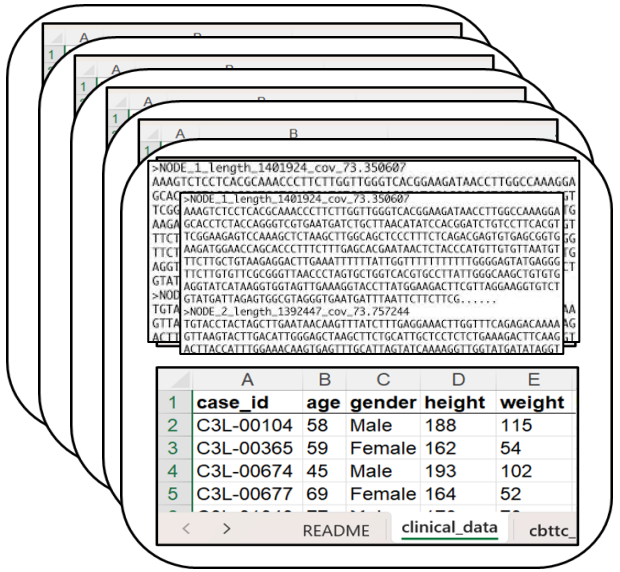


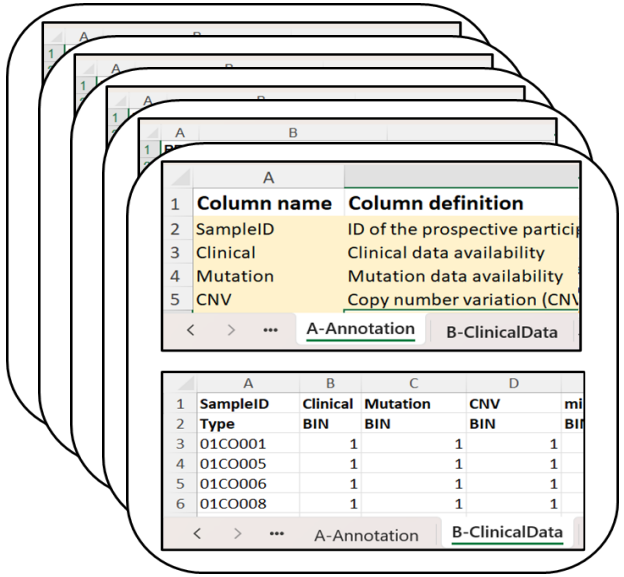


A complex analytical pipeline

- A researcher wants to run a medical study comparing patient data from several sources
  - The input contains PDF papers, images, genomics, and tabular data
  - …the linked datasets contain relevant as well as irrelevant tables

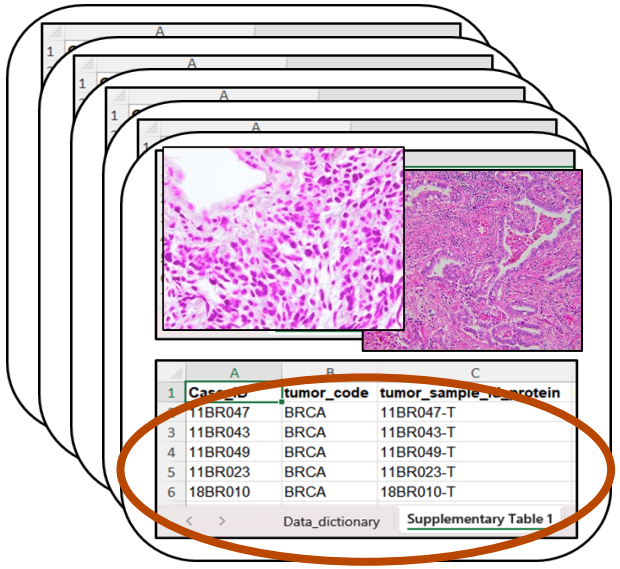


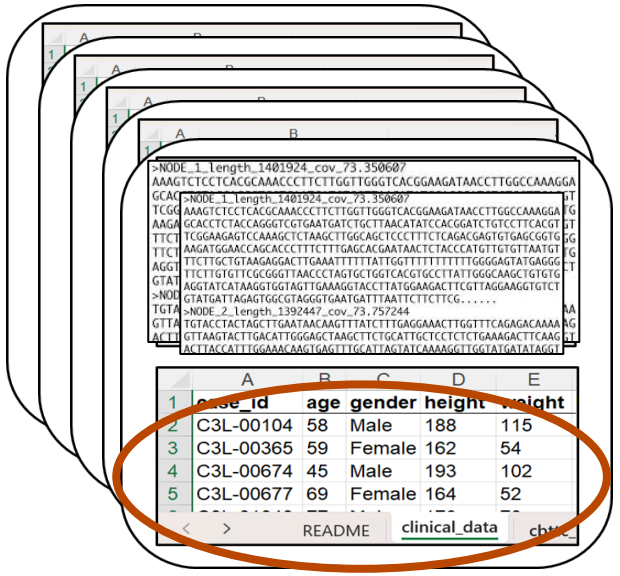


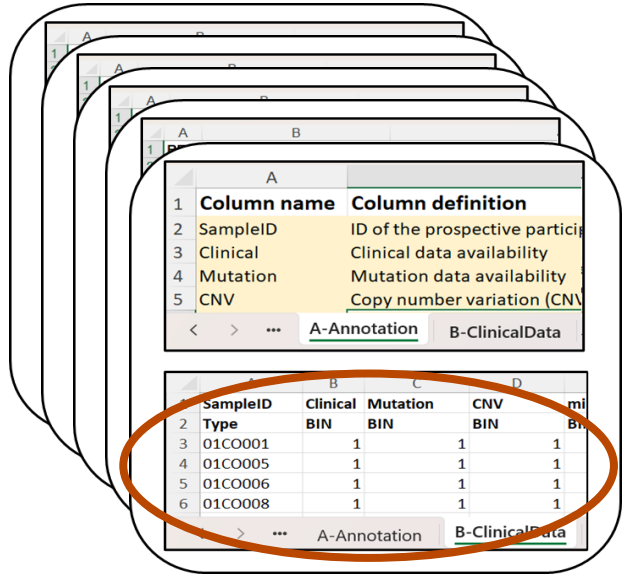


A complex analytical pipeline

- A researcher wants to run a medical study comparing patient data from several sources
  - The input contains PDF papers, images, genomics, and tabular data
  - …the linked datasets contain relevant as well as irrelevant tables
  - …the tables do not have the same schema



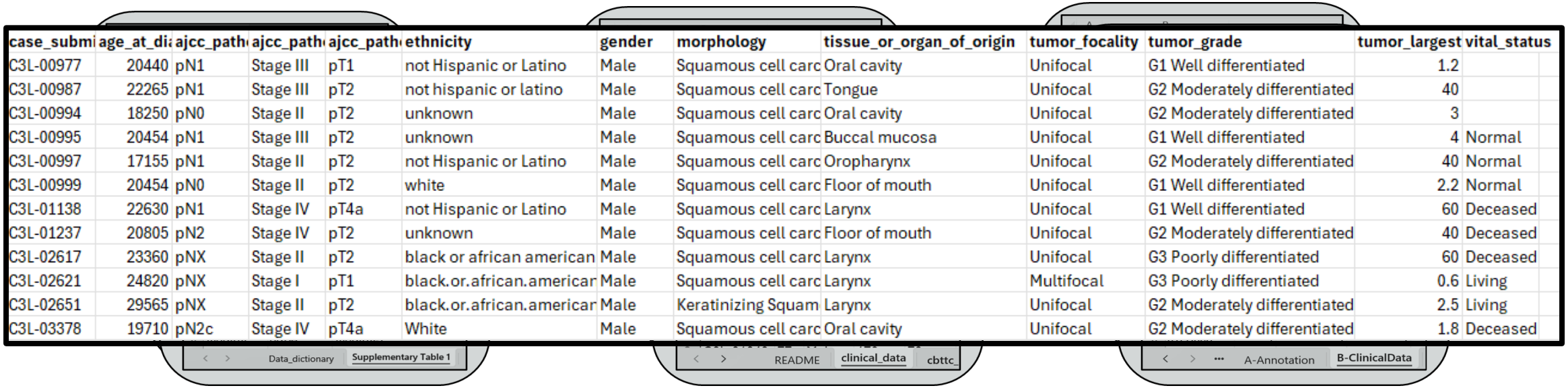




A complex analytical pipeline

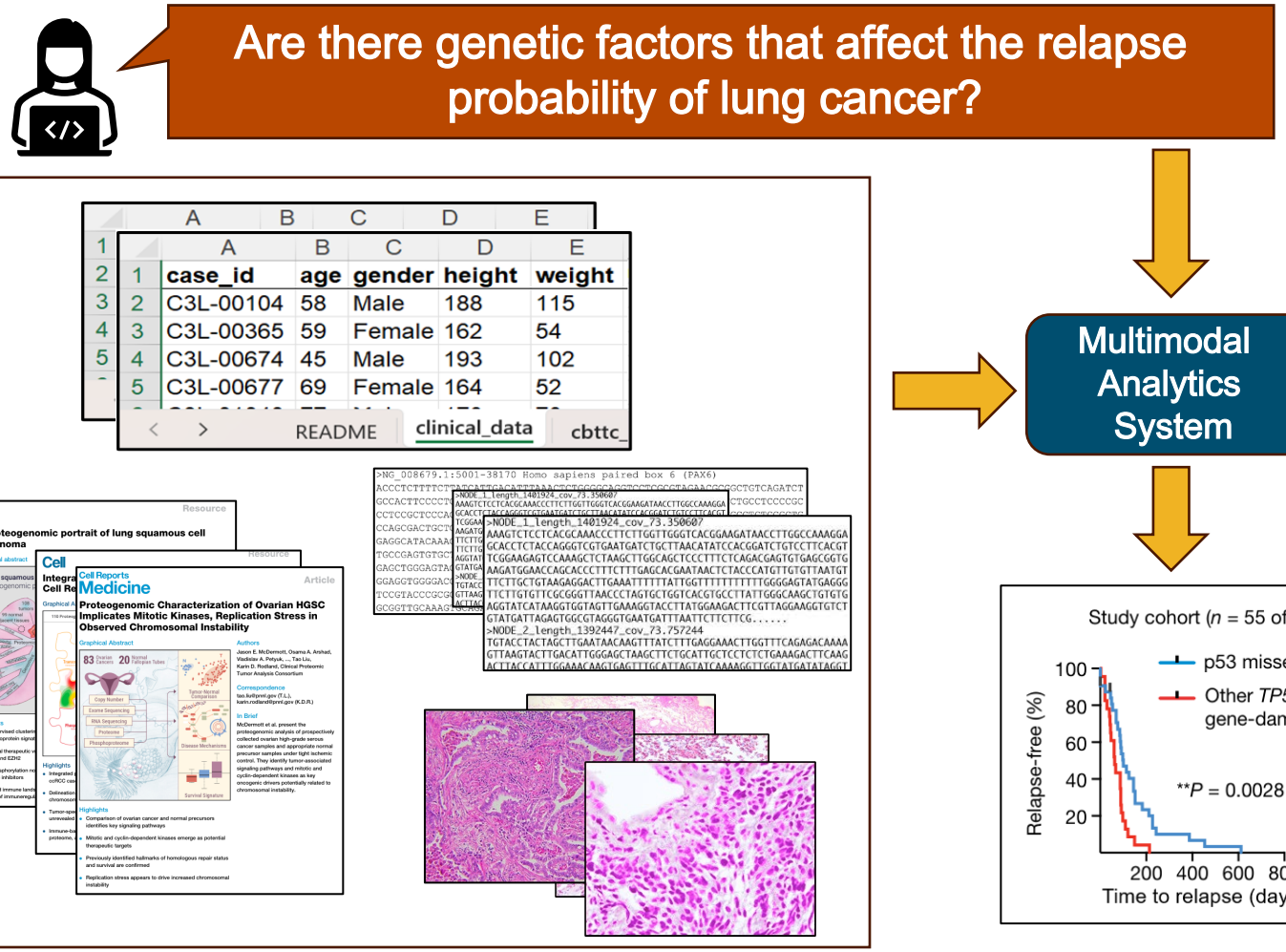
- A researcher w … several sources
  - The input …
  - …the linke…
  - …the table…

In the original study
data curation took 66 authors!



case_submi	age_at_di	ajcc_path	ajcc_path	ajcc_path	ethnicity	gender	morphology	tissue_or_organ_of_origin	tumor_focality	tumor_grade	tumor_largest	vital_status
C3L-00977	20440	pN1	Stage III	pT1	not Hispanic or Latino	Male	Squamous cell carc	Oral cavity	Unifocal	G1 Well differentiated	1.2	
C3L-00987	22265	pN1	Stage III	pT2	not hispanic or latino	Male	Squamous cell carc	Tongue	Unifocal	G2 Moderately differentiated	40	
C3L-00994	18250	pN0	Stage II	pT2	unknown	Male	Squamous cell carc	Oral cavity	Unifocal	G2 Moderately differentiated	3	
C3L-00995	20454	pN1	Stage III	pT2	unknown	Male	Squamous cell carc	Buccal mucosa	Unifocal	G1 Well differentiated	4	Normal
C3L-00997	17155	pN1	Stage II	pT2	not Hispanic or Latino	Male	Squamous cell carc	Oropharynx	Unifocal	G2 Moderately differentiated	40	Normal
C3L-00999	20454	pN0	Stage II	pT2	white	Male	Squamous cell carc	Floor of mouth	Unifocal	G1 Well differentiated	2.2	Normal
C3L-01138	22630	pN1	Stage IV	pT4a	not Hispanic or Latino	Male	Squamous cell carc	Larynx	Unifocal	G1 Well differentiated	60	Deceased
C3L-01237	20805	pN2	Stage IV	pT2	unknown	Male	Squamous cell carc	Floor of mouth	Unifocal	G2 Moderately differentiated	40	Deceased
C3L-02617	23360	pNX	Stage II	pT2	black or african american	Male	Squamous cell carc	Larynx	Unifocal	G3 Poorly differentiated	60	Deceased
C3L-02621	24820	pNX	Stage I	pT1	black.or.african.americar	Male	Squamous cell carc	Larynx	Multifocal	G3 Poorly differentiated	0.6	Living
C3L-02651	29565	pNX	Stage II	pT2	black.or.african.americar	Male	Keratinizing Squam	Larynx	Unifocal	G2 Moderately differentiated	2.5	Living
C3L-03378	19710	pN2c	Stage IV	pT4a	White	Male	Squamous cell carc	Oral cavity	Unifocal	G2 Moderately differentiated	1.8	Deceased

Our Vision



Are there genetic factors that affect the relapse probability of lung cancer?

case_id	age	gender	height	weight
C3L-00104	58	Male	188	115
C3L-00365	59	Female	162	54
C3L-00674	45	Male	193	102
C3L-00677	69	Female	164	52

README | clinical_data | cbttc_

Cell — A proteogenomic portrait of lung squamous cell carcinoma

Cell Reports Medicine — Proteogenomic Characterization of Ovarian HGSC Implicates Mitotic Kinases, Replication Stress in Observed Chromosomal Instability

>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)

>NODE_1_length_1401924_cov_73.350607

>NODE_2_length_1392447_cov_73.757244

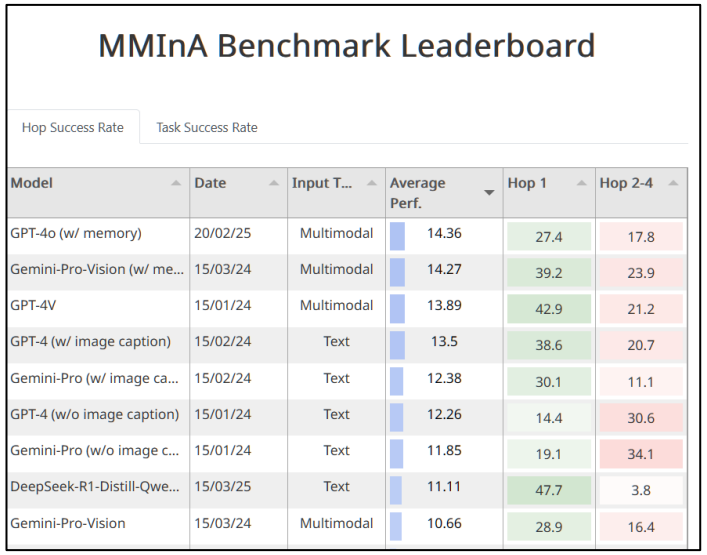
Multimodal Analytics System

Multimodal Analytics

- Analytical workloads
  - Large data inputs
  - Complex data relationships
  - Heavy computation loads

- OLAP systems for relational data
  - Require data engineering AND domain expertise
  - Rely on predefined schemas
  - Do not generalize across workloads

- Agentic-based systems
  - Not quite work yet for multimodal data [1,2]
  - Cost, time, and quality (!) uncertain



Model	Date	Input T...	Average Perf.	Hop 1	Hop 2-4
GPT-4o (w/ memory)	20/02/25	Multimodal	14.36	27.4	17.8
Gemini-Pro-Vision (w/ me...	15/03/24	Multimodal	14.27	39.2	23.9
GPT-4V	15/01/24	Multimodal	13.89	42.9	21.2
GPT-4 (w/ image caption)	15/02/24	Text	13.5	38.6	20.7
Gemini-Pro (w/ image ca...	15/02/24	Text	12.38	30.1	11.1
GPT-4 (w/o image caption)	15/01/24	Text	12.26	14.4	30.6
Gemini-Pro (w/o image c...	15/01/24	Text	11.85	19.1	34.1
DeepSeek-R1-Distill-Qwe...	15/03/25	Text	11.11	47.7	3.8
Gemini-Pro-Vision	15/03/24	Multimodal	10.66	28.9	16.4

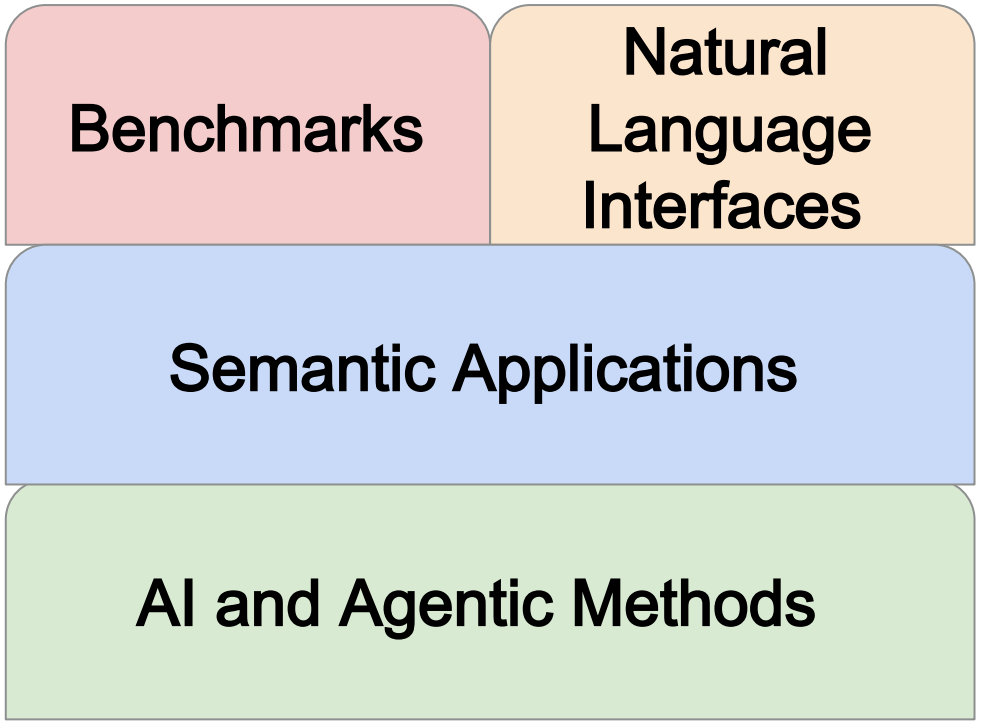
Model	Overall		
	OA(%)	SA(%)	AVG CS(%)
			Tool-Free
o4-mini-high	7.14	3.13	13.67
o4-mini	5.36	2.23	12.41
GPT-4.1	**7.59**	**5.36**	**14.68**
GPT-4o-2024-11-20	1.34	0.45	4.63
GPT-4o-mini	0.89	0.00	1.47
Gemini-2.5-Pro-Preview-05-06	6.31	4.50	11.56
Gemini-2.5-Flash-Preview-05-20	2.70	2.25	8.57

Table 1 Performance on MM-BrowseComp.

[1] Tian, Shulin, et al. "MMInA : Benchmarkingmultihop multimodal internet agents." *Findings of the ACL*2025.
[2] Li, Shilong, et al. "MM-browsecomp: A comprehensive benchmark for multimodal browsing agents." *arXiv preprint arXiv:2508.13186*2025

A Graphical Agenda

KramaBench
SemBench



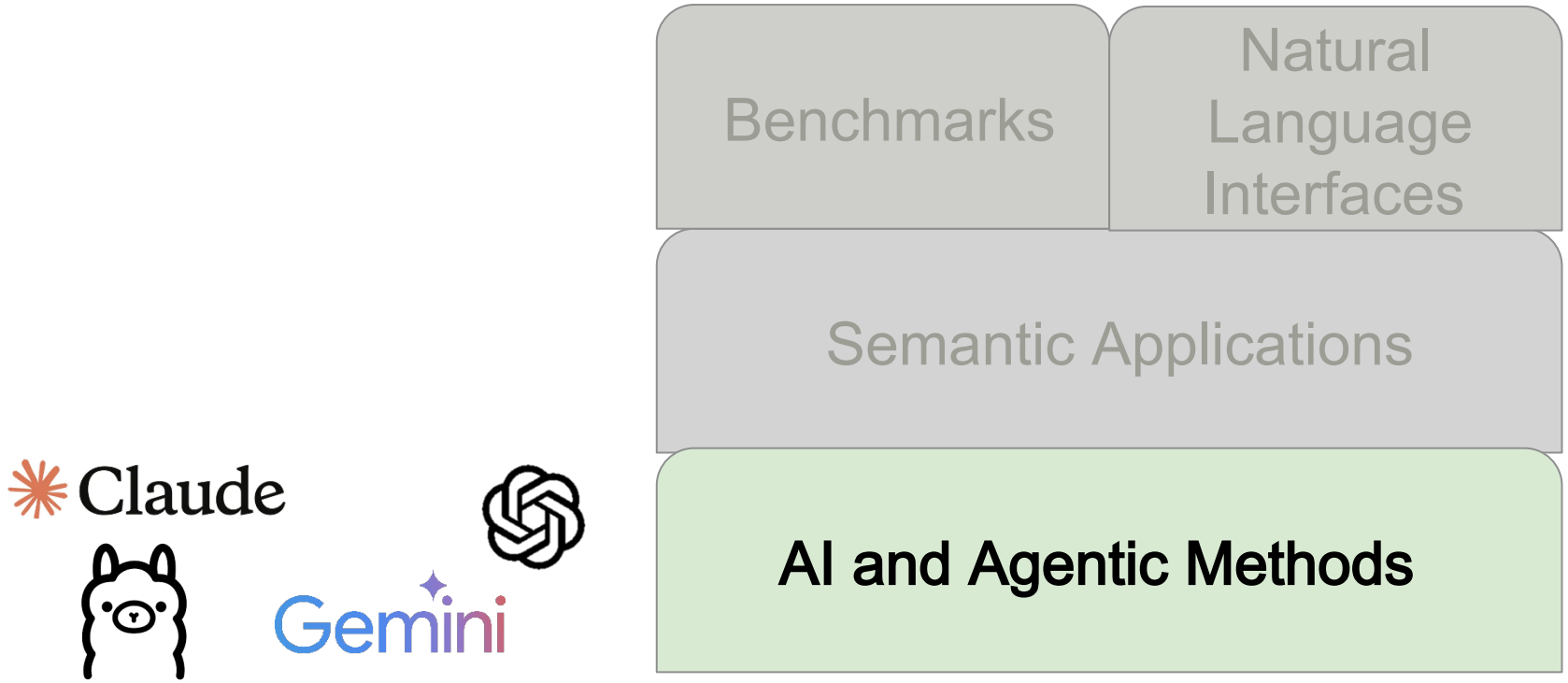
PalimpChat
Caravaggio

Palimpzest

Claude

Gemini

A Graphical Agenda



Infusing LLM with knowledge

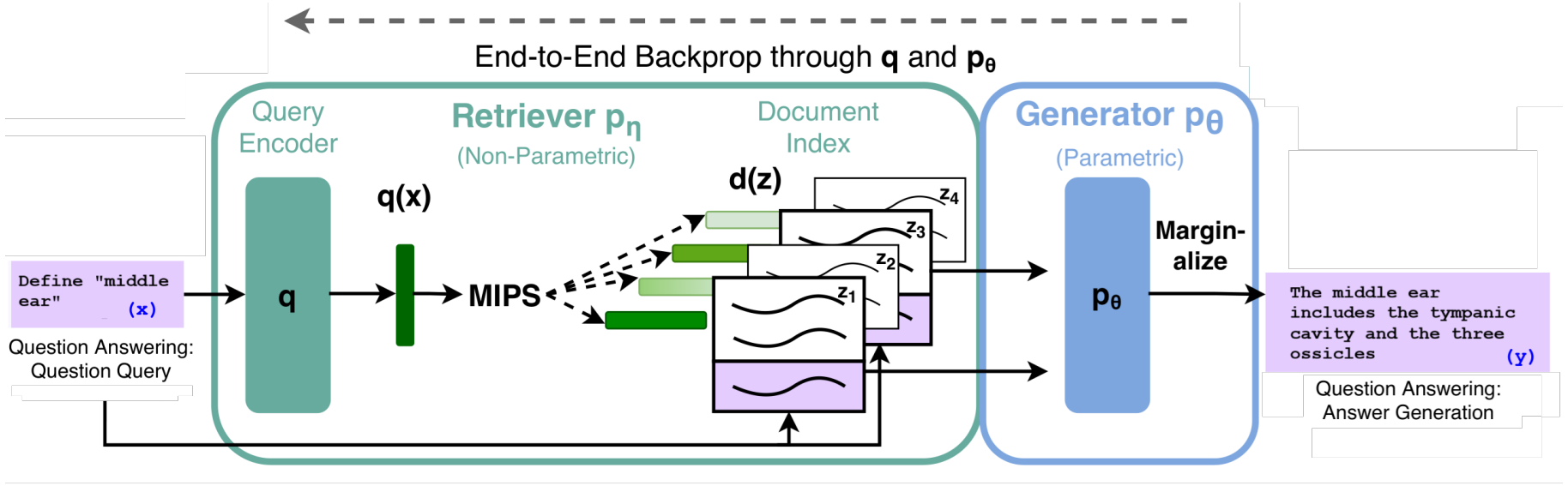
- The pre-training phase infuses LLMs with world knowledge
  - Called **parametric memory** since it's implicitly within the model neural parameters
  - This knowledge can be used to answer questions

- Why not use parametric knowledge for QA?
  - Answer requires data more recent than pre-training (expensive/impossible to re-train)
  - Answer requires private data
  - LLM might hallucinate or give unrealistic answer
  - Grounding answers on facts

- Idea: compound parametric knowledge with external, non-parametric knowledge

RAG–Retrieval Augmented Generation

- RAG–Retrieval Augmented Generation
  - Use a Dense Vector Representation to encode input query (x) and documents (z)
  - Inner vector product (e.g., cosine similarity) to rank documents
  - Populate context of generator model with relevant evidence



[3] P. Lewis*et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proceedings ofNeurIPS.* 2020

Agentic Retrieval

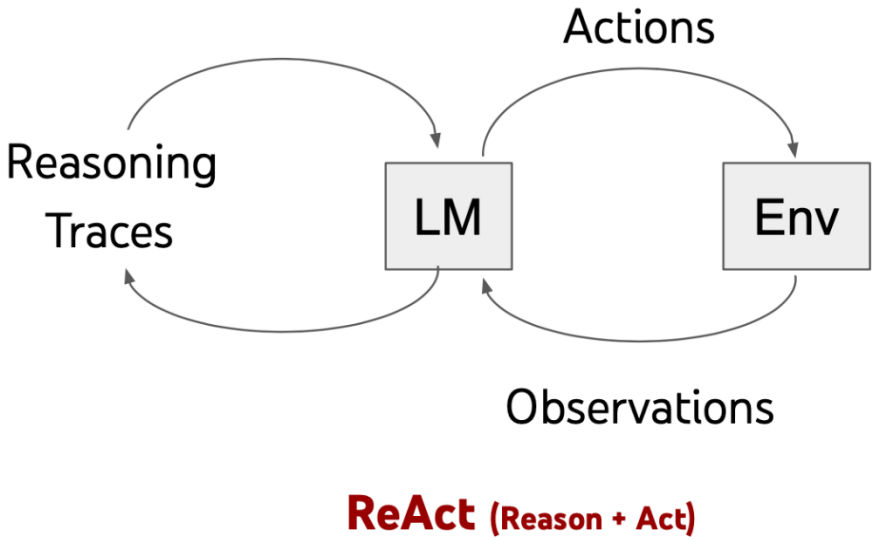
- Some questions require 'reasoning', e.g., mathematical problems
  - A breakthrough technique was the introduction of **Chain-of-Thought prompting** [4]
  - Main idea: leverage few-shot examples of **reasoning traces** to prompt models to "reason"

- Other questions require both reasoning AND information retrieval
  - Idea: interleave reasoning with retrieval, in a step-by-step pipeline: **ReAcT** [5], **FLARE** [6]
  - E.g., "How many rooms are in the hotel that is home to the Circus show Mystere"
    - Search query depends on intermediate results (first find which hotel, then room info)

[4] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of NeurIPS* 2022.
[5] S. Yao *et al.*, "ReAct: Synergizing Reasoning and Acting in Language Models," *Proceedings of ICLR* 2023.
[6] Z. Jiang *et al.*, "Active Retrieval Augmented Generation," in *Proceedings of EMNLP* 2023.

ReAct: Reason then Act

- ReAct: Reasoning + Acting with LLMs
  - Purely prompt-based technique (with in-context examples)
  - Three traces: 1. thought 2.action and 3. observation
  - Actions may be API calls, observations are the outputs
  - Reasoning stops after fixed number of steps

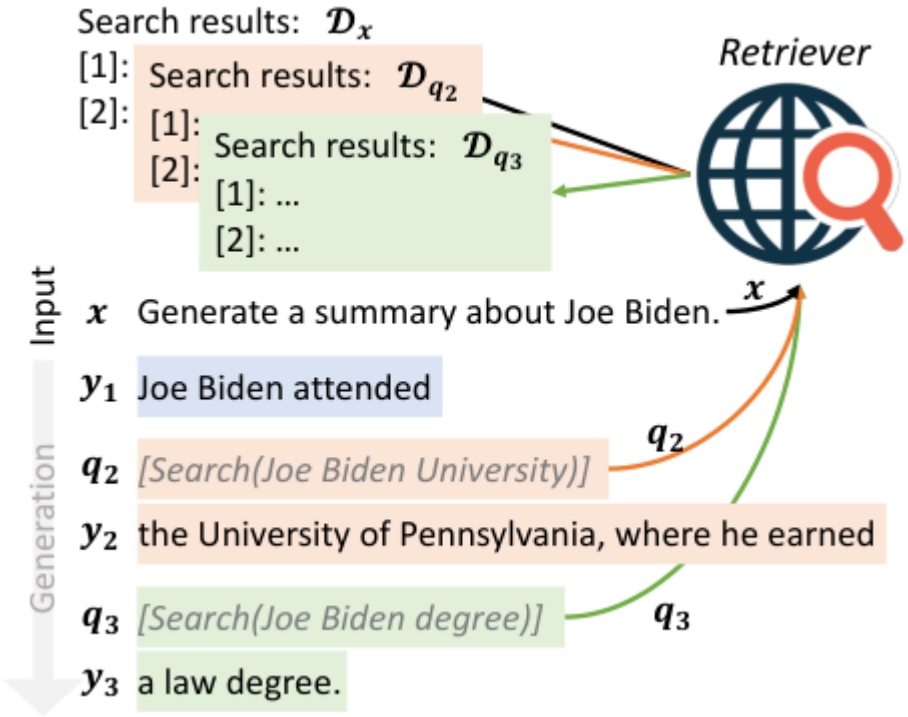


- Two approaches
  - Prompting: use 3-6 few-shot examples
  - Finetuning: use 3000 training samples

[5] S. Yao *et al.*, "ReAct: Synergizing Reasoning and Acting in Language Models," *Proceedings of ICLR* 2023.

FLARE: Active Retrieval

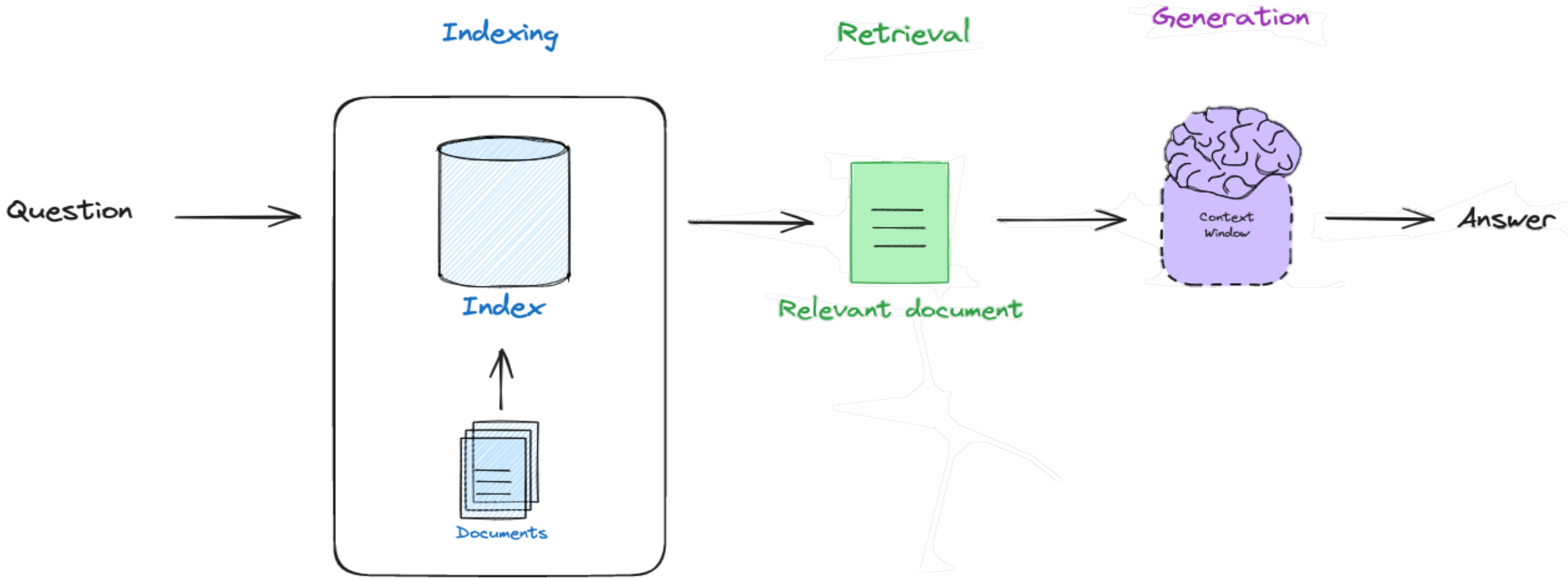
- FLARE: Forward Looking Active REtrieval
  - Generator decides "when" and "what" to retrieve
  - Both are dynamic based on intermediate outputs
  - When to retrieve:
    - If the answer lacks confidence (low logits)
  - What to retrieve:
    - Based on current outputs ("Forward") and not on previous generations



- Black box models: use prompting
- White box models: use logit activations

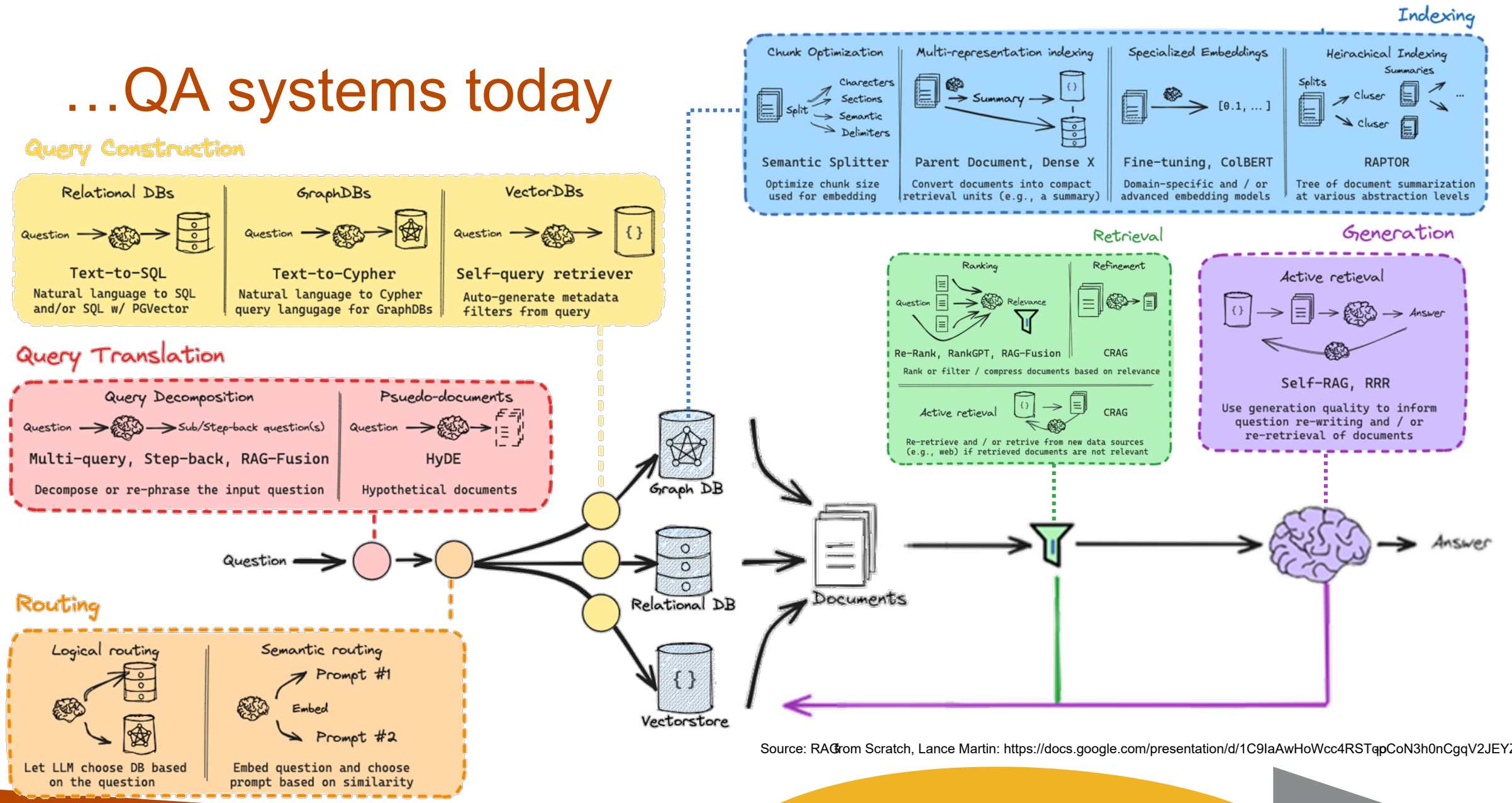
[6] Z. Jiang *et al.*, "Active Retrieval Augmented Generation," in *Proceedings of EMNLP* 2023.

What naïve RAG looked like….



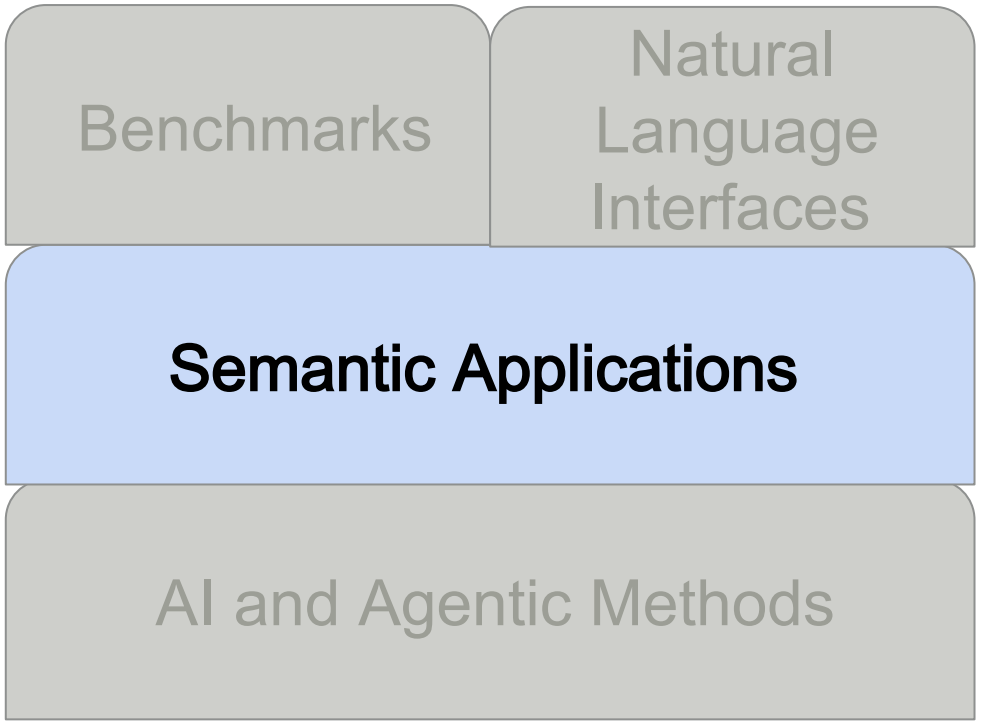
Source: RAG from Scratch, Lance Martin: https://docs.google.com/presentation/d/1C9IaAwHoWcc4RSTqpCoN3h0nCgqV2JEYZUJunv_9Q

…QA systems today



Source: RAG from Scratch, Lance Martin: https://docs.google.com/presentation/d/1C9IaAwHoWcc4RSTqpCoN3h0nCgqV2JEYZUJunv_9C

A Graphical Agenda



Palimpzest

Semantic Applications

```
import palimpzest as pz
papers = pz.Dataset("medical-pdf", schema=ScientificPaper)
table_urls = papers.map(pz.URL, "The URLs of the XLS tables from the page", "one-to-many")
patient_tables = table_urls.download()
patient_tables = patient_tables.filter("The table contains patient biometric data")
case_data = patient_tables.map(CaseData, desc="The patient data in the table", "one-to-many")
```

- Building data pipelines
  - Semantic: use natural language
  - Declarative: **what** not **how**
  - Automatic optimization

Introducing Palimpzest

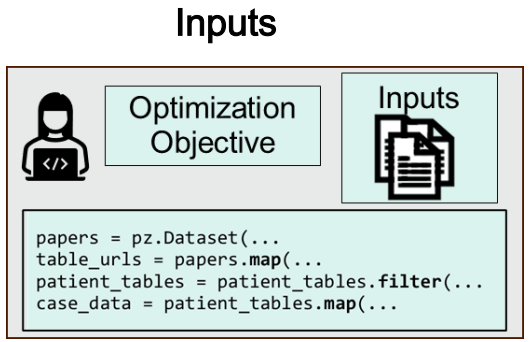
- A programming language that offers semantic operators as primitives
- Support for an (extensible) large number of under-the-hood operator implementations
- Optimize for the fastest, cheapest, or highest-quality implementation (or combination!)
- Declarative benefits: if models or prices or hardware change, it will choose differently

*A palimpsest is a physical document that has been erased and rewritten multiple times, just as our system rethinks and revises the best implementation

[7] C. Liu et al., "Palimpzest: Optimizing AI-Powered Analytics with Declarative Query Processing," in *Proceedings of CIDR*, 2025.

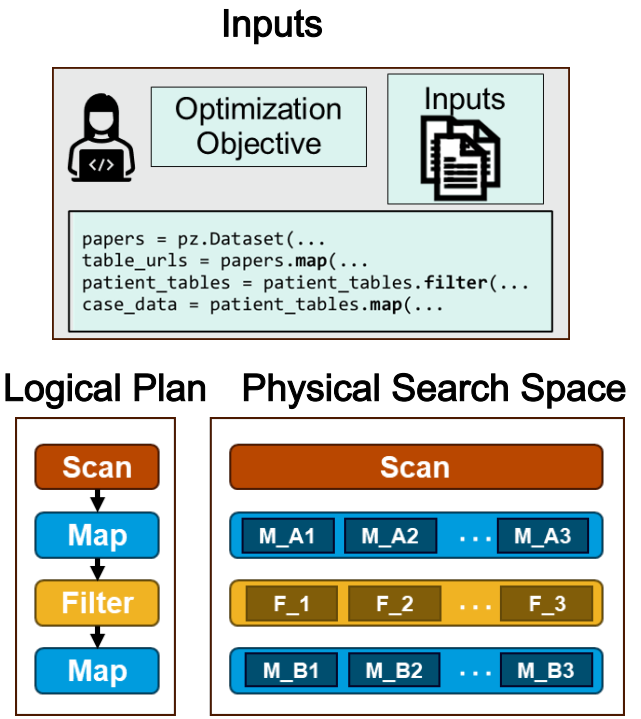
System Architecture

- Main workflow:
  - Users write declarative programs



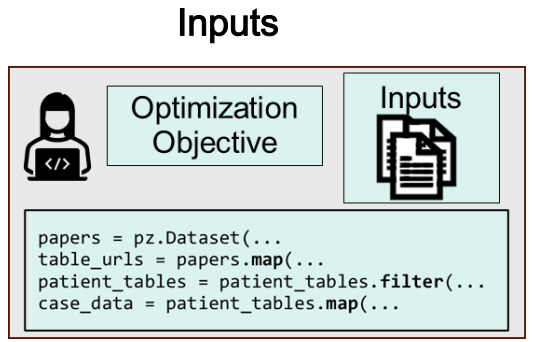
System Architecture

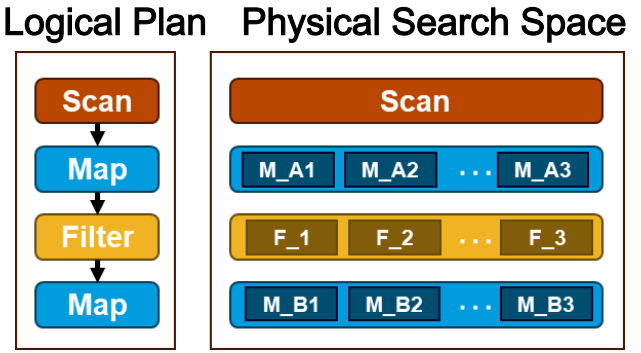
- Main workflow:
  - Users write declarative programs
  - Compile logical plans
  - Enumerate possible physical plans



System Architecture

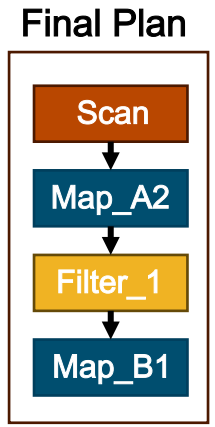
- Main workflow:
  - Users write declarative programs
  - Compile logical plans
  - Enumerate possible physical plans
  - Estimate the cost, quality, and latency
  - Choose the (pareto) optimal plan





Per-operator cost estimates

Operator	Quality	Cost	Latency
Filter_1	0.95	$0.001	0.0001s
Filter_2	0.75	$0.003	0.0003s
Map_A1	0.78	$0.540	4.5036s
Map_A2	0.31	$0.070	1.15210s
…	…	…	…

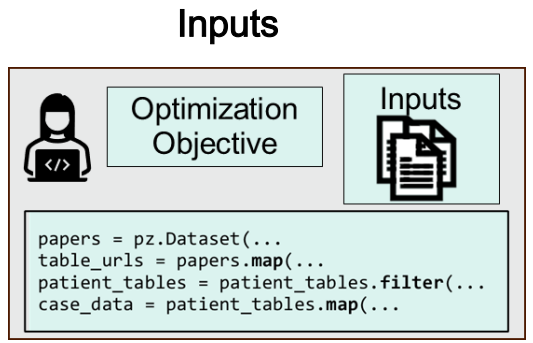


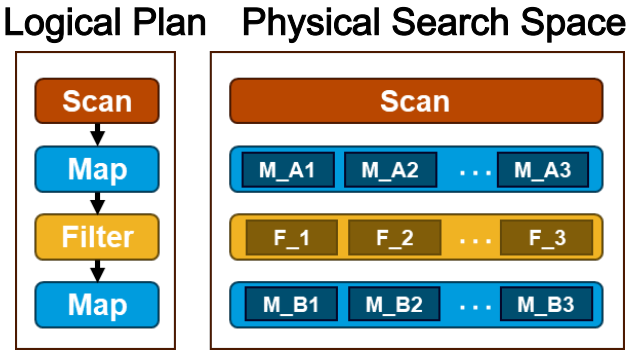
System Architecture

- Main workflow:
  - Users write declarative programs
  - Compile logical plans
  - Enumerate possible physical plans
  - Estimate the cost, quality, and latency
  - Choose the (pareto) optimal plan

- Example: physical implementations for **map**:
  - Single LLM prompting
  - RAG-based reduction of context
  - Mixture of Agents

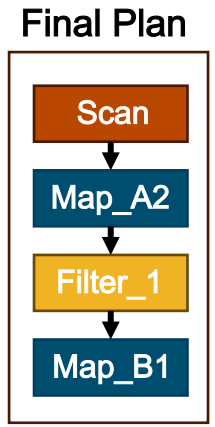
- Main challenges:
  - Cost, quality, runtime estimation
  - Search space is combinatorial!



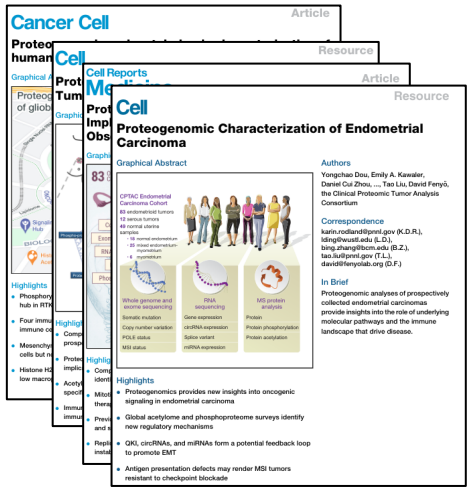


Per-operator cost estimates

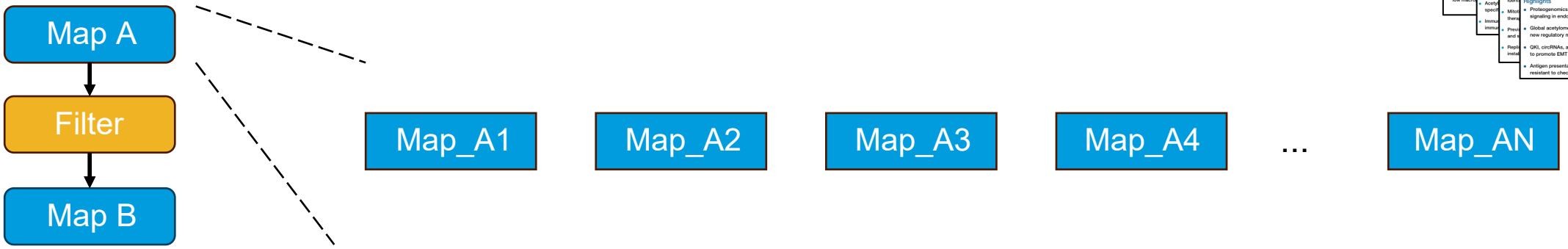
Operator	Quality	Cost	Latency
Filter_1	0.95	$0.001	0.0001s
Filter_2	0.75	$0.003	0.0003s
Map_A1	0.78	$0.540	4.5036s
Map_A2	0.31	$0.070	1.15210s
…	…	…	…



Insights: Abacus Cost Estimation



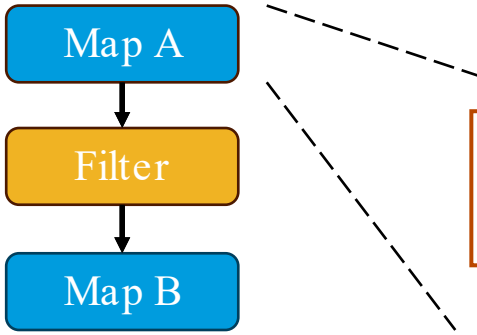
Logical Plan



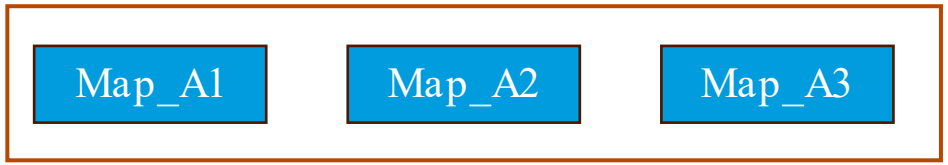
[8] M. Russo et al., "Abacus: A Cost-Based Optimizer for Semantic Operator Systems," *PVLDB*, 2026

Insights: Abacus Cost Estimation

Logical Plan

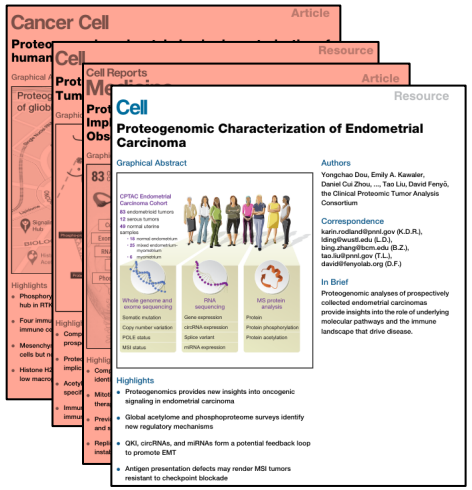


Sample k physical operators



Map_A4 ... Map_AN

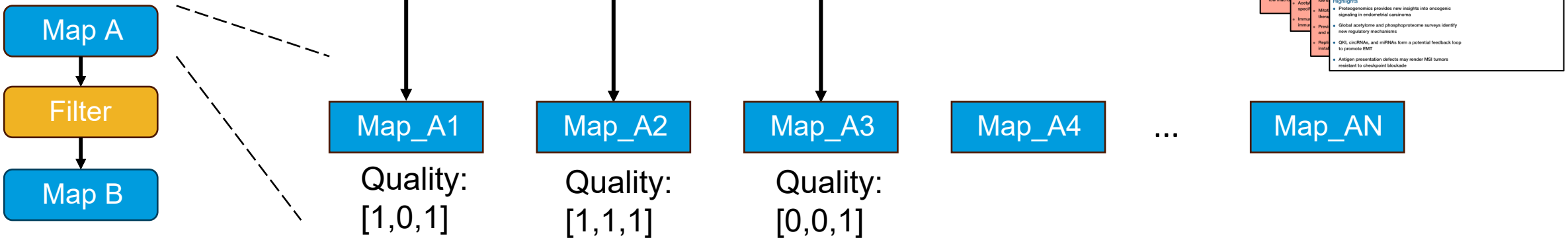
Sample j inputs



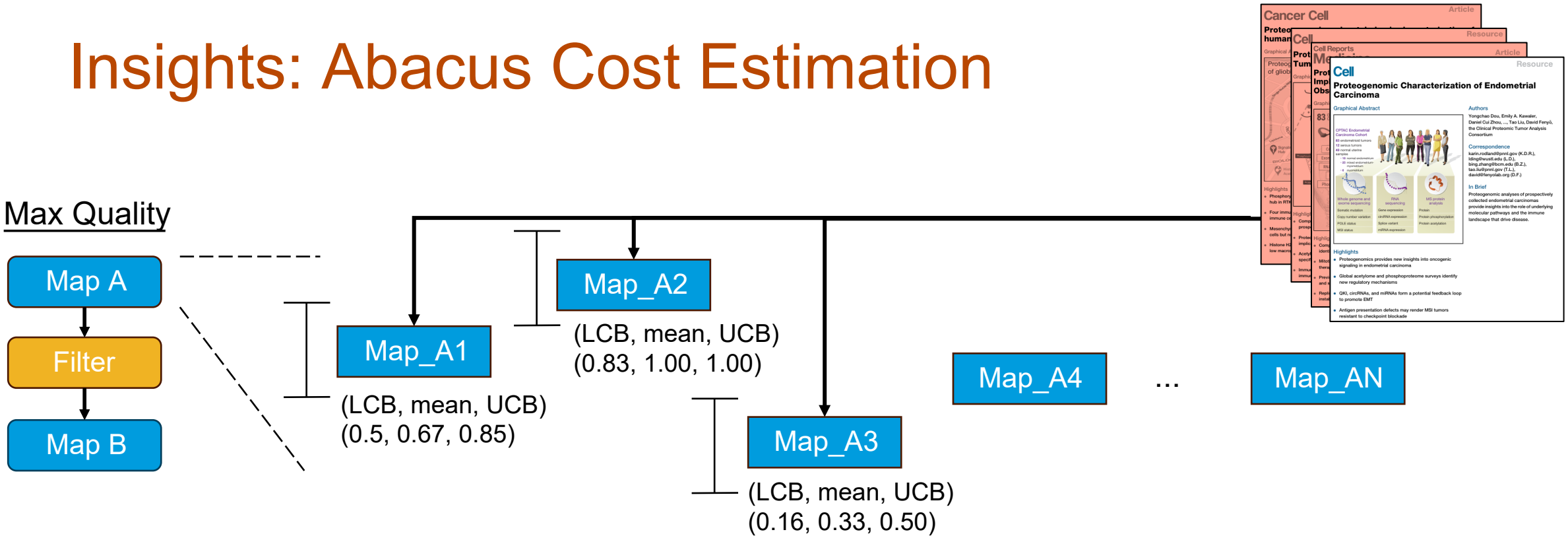
[8] M. Russo et al., "Abacus: A Cost-Based Optimizer for Semantic Operator Systems," *PVLDB*, 2026

Insights: Abacus Cost Estimation

Logical Plan

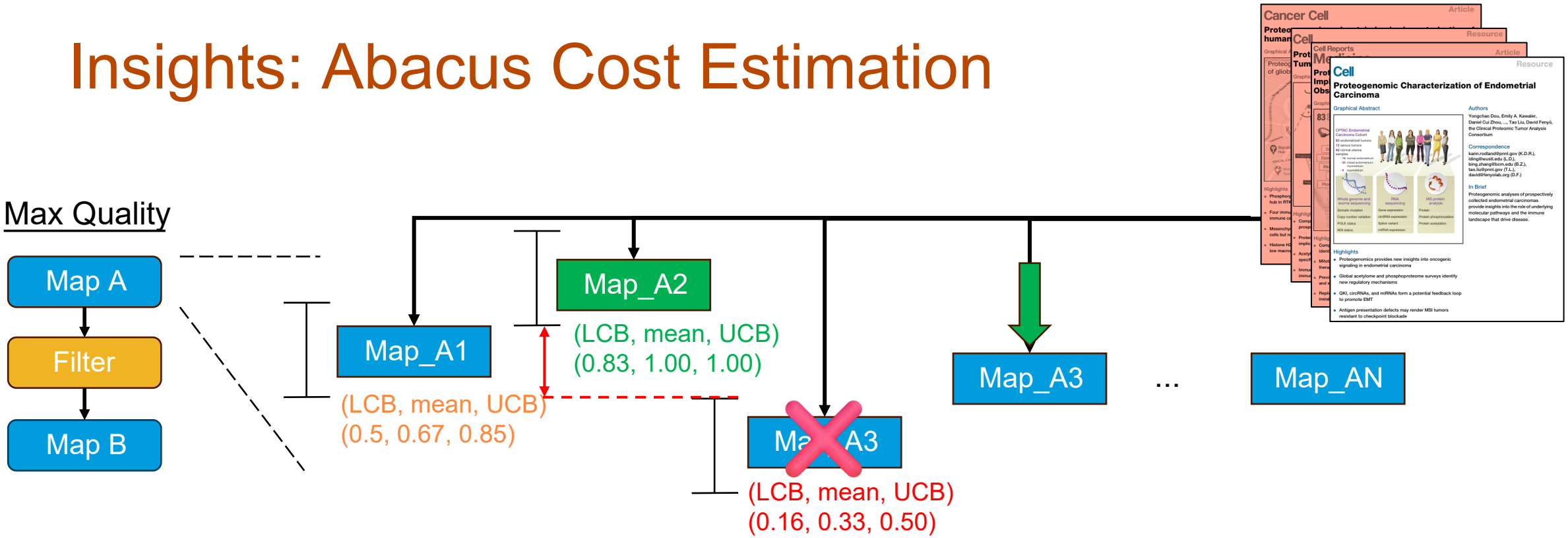


Insights: Abacus Cost Estimation



1. Estimate confidence bounds for each metric of interest & operator

Insights: Abacus Cost Estimation



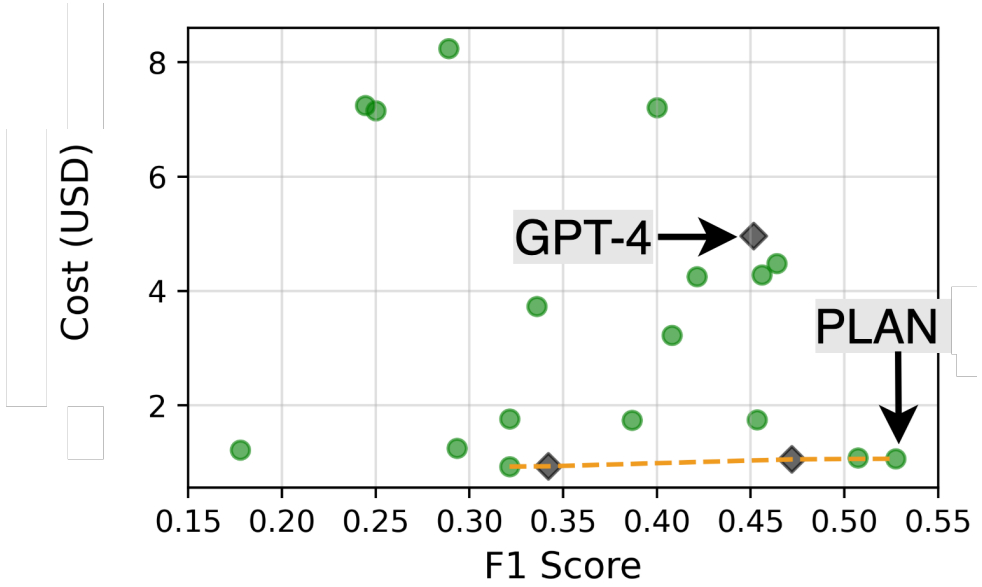
1. Estimate confidence bounds for each metric of interest & operator
2. Find pareto optimal operators based on mean
3. Remove operators outside of pareto boundaries
4. Select new operator to sample (until budget exhausted)

Experimental results

- Goal of our experiments:
  - Can we produce optimal plans?
  - Can we estimate and choose the correct plan?
  - Can we respect user preferences?

- Effect of constrained optimizations
  - Minimizing time: ~ 1.5x lower runtime
  - Minimizing costs: ~ 2.52x lower cost
  - Due to stochastic nature, not always wins

- Room for future work adding rules/implementations!



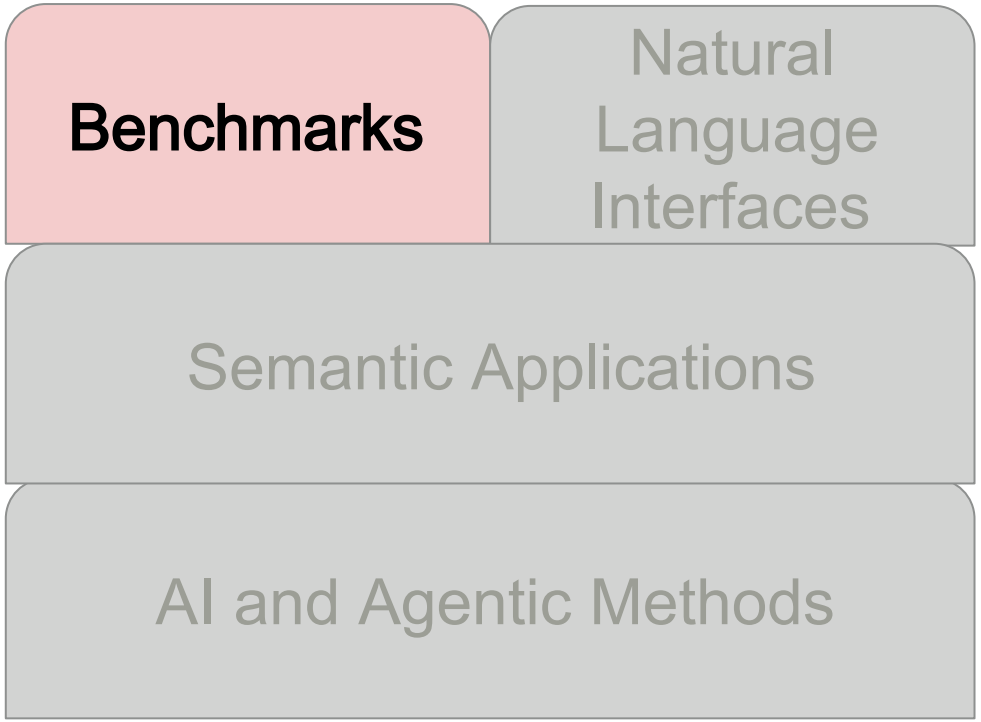
(Max Quality)	Quality	Total Cost ($)	Total Time (s)
BioDEX	0.261 ± 0.026	0.89 ± 0.11	450 ± 47
CUAD	0.662 ± 0.010	0.69 ± 0.05	450 ± 67
MMQA	0.304 ± 0.079	13.10 ± 10.6	1,149 ± 300

(MinCost)	Quality	Exec. Cost ($)	Reduction
BioDEX	0.21 ± 0.02	$0.28 ± 0.10	2.50x
CUAD	0.05 ± 0.02	$0.12 ± 0.01	4.25x
MMQA	0.31 ± 0.05	$16.0 ± 9.7	0.81x

(MinTime)	Quality	Exec. Time (s)	Reduction
BioDEX	0.21 ± 0.03	128 ± 50	1.15x
CUAD	0.10 ± 0.05	55 ± 18	2.4x
MMQA	0.28 ± 0.07	540 ± 382	1.02x

A Graphical Agenda

KramaBench
SemBench



HOTPOTQA5A8BC4695542995E66A47502

How many films she acted who was in lead role of Siva?

(More than 200)

- *Siva* (1989 Tamil film), a film starring Rajinikanth as the title character
- *Siva* (1989 Telugu film), an action film

Cast [edit]

- Rajinikanth as Siva (Tiger)
- Raghuvaran as John
- Shobana as Parvathy
- Sowcar Janaki as John's mother

In a career spanning over four decades, Shobana has starred in 230 films across several languages.[7][8] She has

BEERQA 0 18 8 7248 C0 C59 5E79 BDF235F5 1CEB8 6 D8 CD4 ADF7

For a role in what okay did contestants compete in the 2007 reality show based out of the united kingdom?

(Joseph)

Any Dream Will Do, is a 2007 talent show-themed television series produced by the BBC in the United Kingdom. It searched for a new, unknown lead to play Joseph in a West End revival of the Andrew Lloyd Webber musical *Joseph and the Amazing Technicolor Dreamcoat*.

MMQA: 26 71EDE0 DC29 6 6 D8 16 9 36 6 6 4 172BFA8 E

Was East Carolina University or the school where Richard Lindzen worked when he proposed the iris hypothesis the institution that was a conference chair of SIGDOC first?

(Massachuses Institute of Technology)

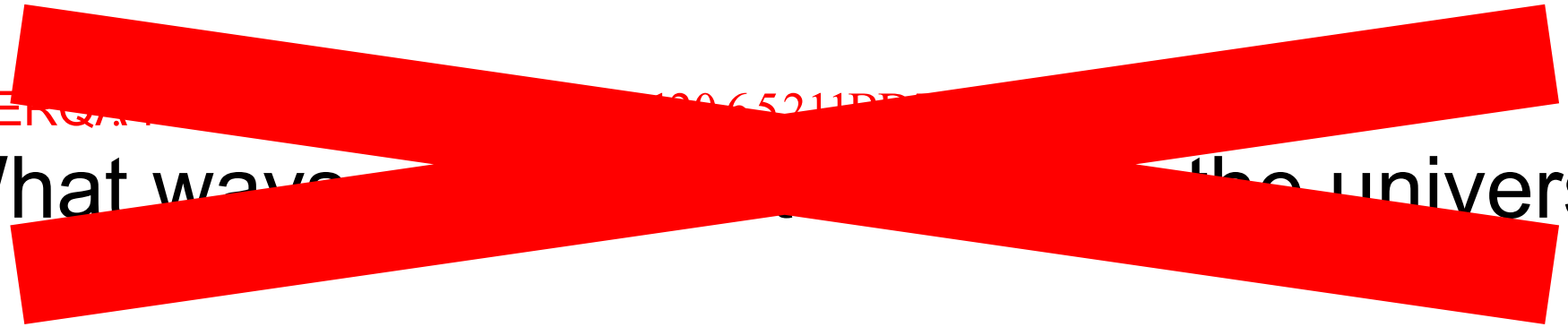
Conference Chairs

Year	Name	Institution	Location
2016	Michael Trice	Massachusetts Institute of Technology	Cambridge, Massachusetts, United States
2015-2014	Kathie Gossett	Iowa State University (ISU)	Ames, Iowa, United States
2013	Michael Albers	East Carolina University	Greenville, North Carolina, United States

Bonus!
IT'S WRONG!

BEERQA 4C5 19 370 6 4C2E6 776 38 120 6 5211BD750 BB45E9 E1

What ways can god interact with the universe?



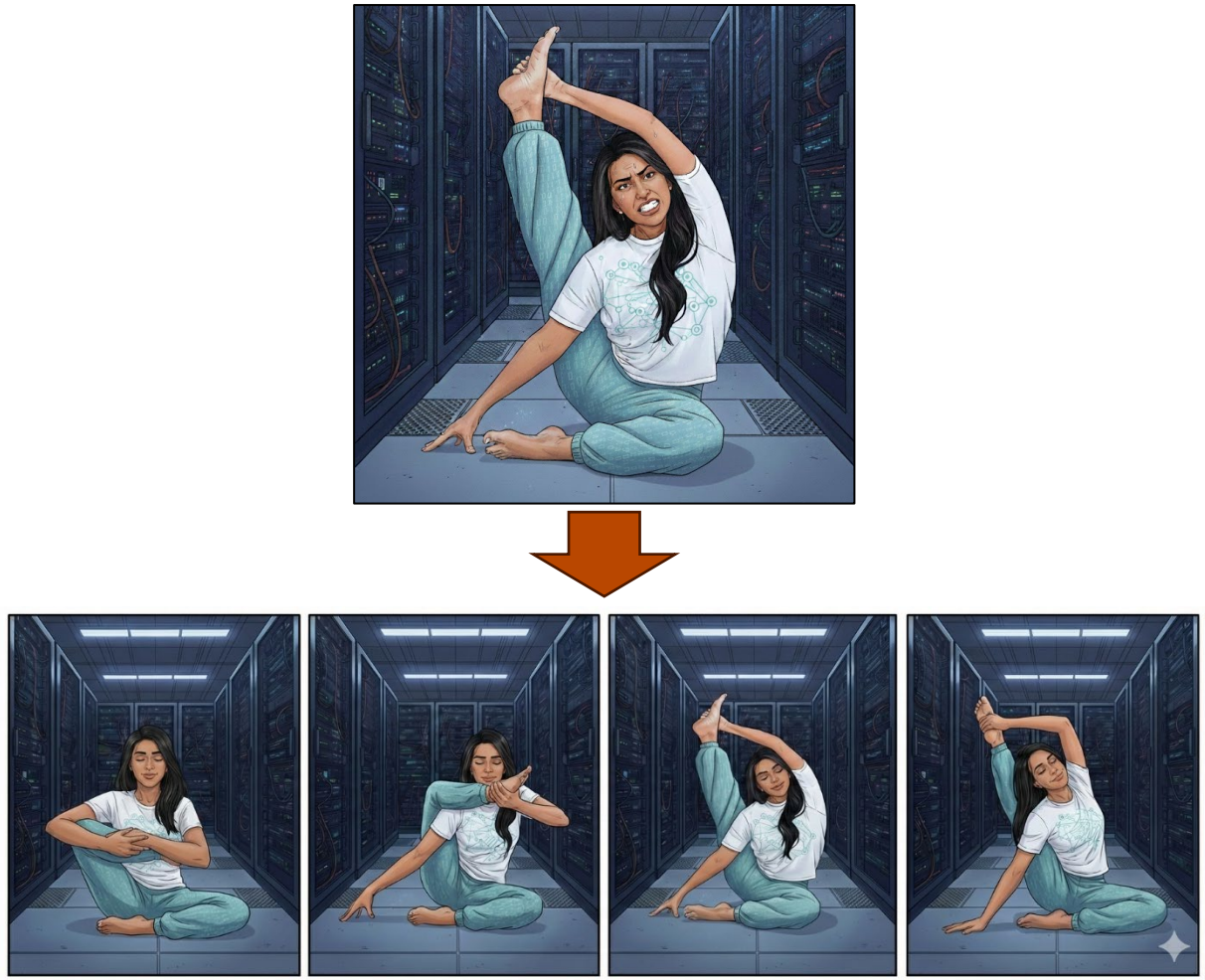
Are we really sure we want to ask the AI this question?

We need better benchmarks!

- Real-world data science is **complex:**
  - Data sources are multimodal, noisy, ambiguous
  - Domain knowledge might be required
  - It requires whole processing pipelines

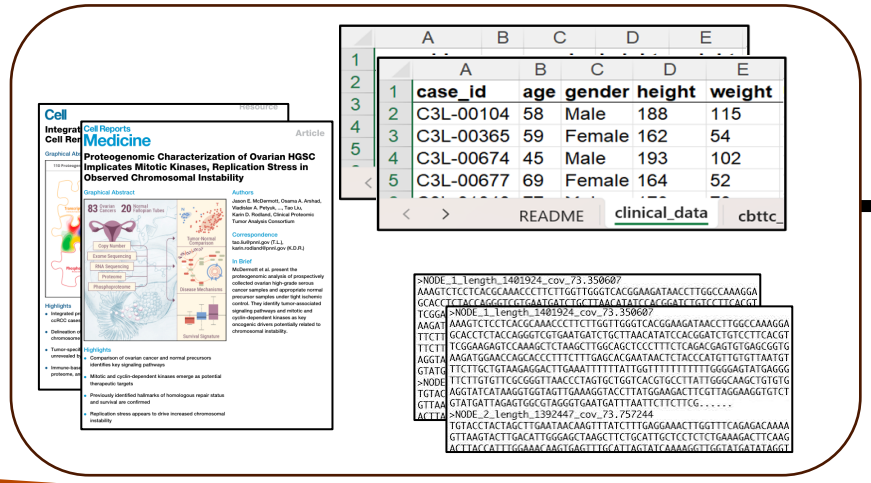
- But popular benchmarks evaluate ***isolated tasks***
  - *Ex:* Data Discovery, Coding, NL 2 SQL, …

- Our goal is to have encompassing benchmarks!

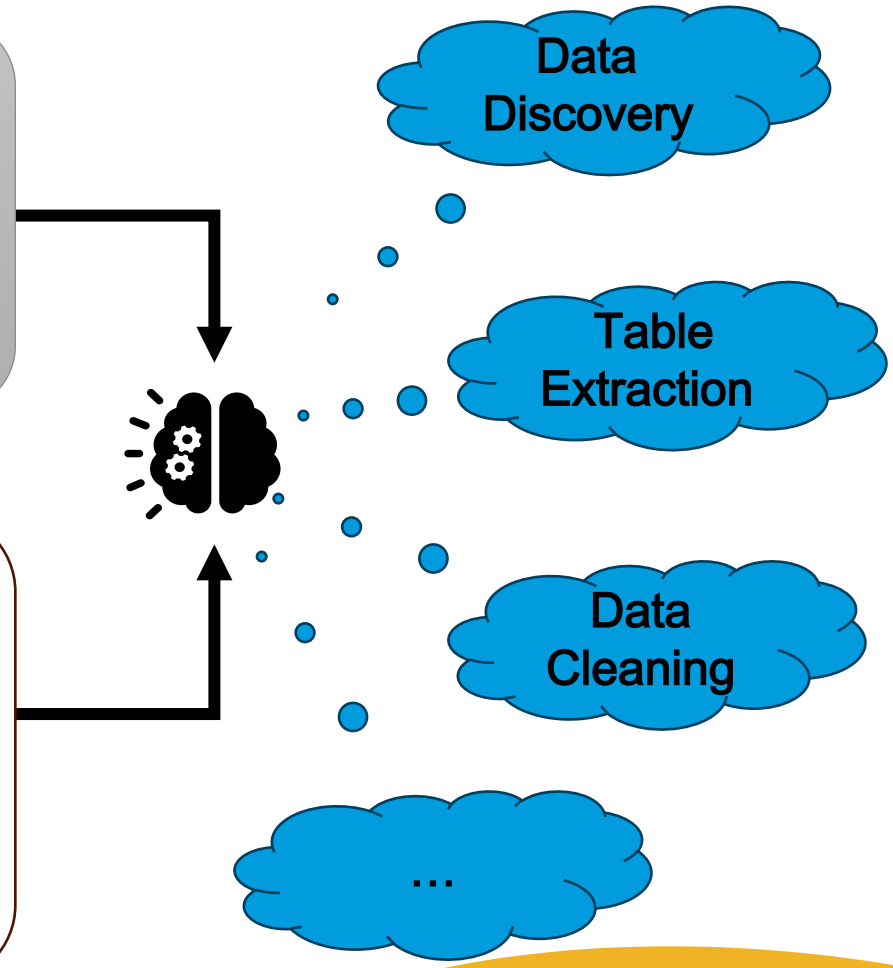


Real-world data science

Task What is the median number of variants perMbp for the serous tumor samples in the study? Round the result to 4 decimal places.



	case_id	age	gender	height	weight
2	C3L-00104	58	Male	188	115
3	C3L-00365	59	Female	162	54
4	C3L-00674	45	Male	193	102
5	C3L-00677	69	Female	164	52



2.6563

Introducing KramaBench

- Goal: evaluate **automation** of data science pipelines

- We collected data from 6 challenging domains
  - >100 pipelines reproducing published studies
  - Each pipeline is broken into sub-tasks
  - The tasks span ~1700 real world data files

- We provide reference solutions
  - Verified by external contributors
  - Annotated into key functionalities/sub-tasks
  - Enable evaluation of planning vs. execution

Domain	# tasks	# subtasks	# files	File Size
Archeology	12	71	5	7.5 MB
Astronomy	12	68	1556	486MB
Biomedical	9	38	7	175 MB
Environment	20	148	37	31 MB
Legal	30	188	136	1.3 MB
Wildfire	21	120	23	1 GB
Overall	**104**	**633**	**1764**	**1.7GB**

KRAMABENCH

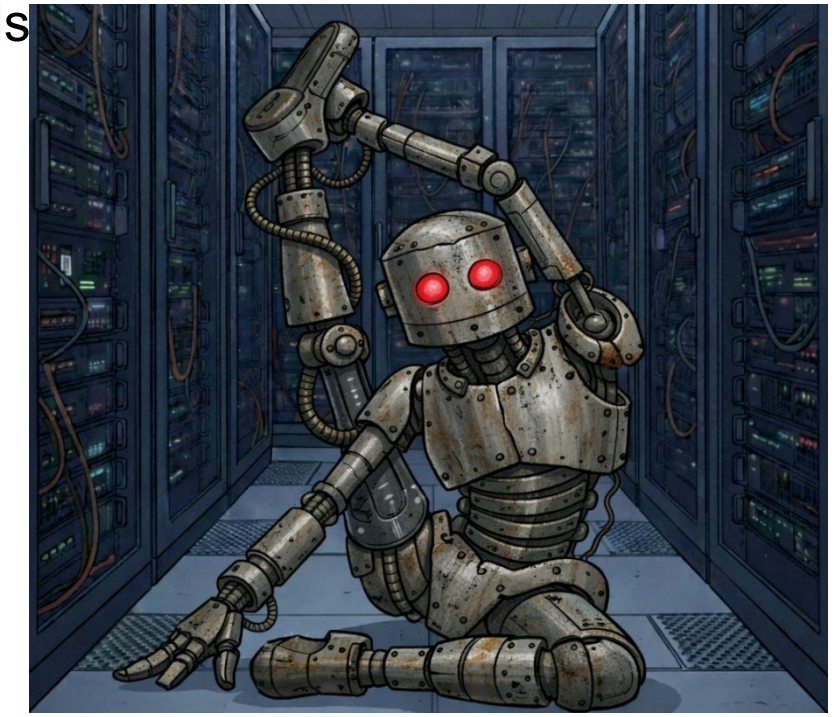
[10] Vitagliano et al., "KramaBench: A Benchmark for AI Systems on Data Intensive Tasks," in *Proceedings of ICLR*, 2026

Systems Under Test

- We tested three families of systems with multiple LLM backends
  - GPT family (o3, 4o), Claude 3.5, Llama 3.3, Deepseek R1

Systems Under Test

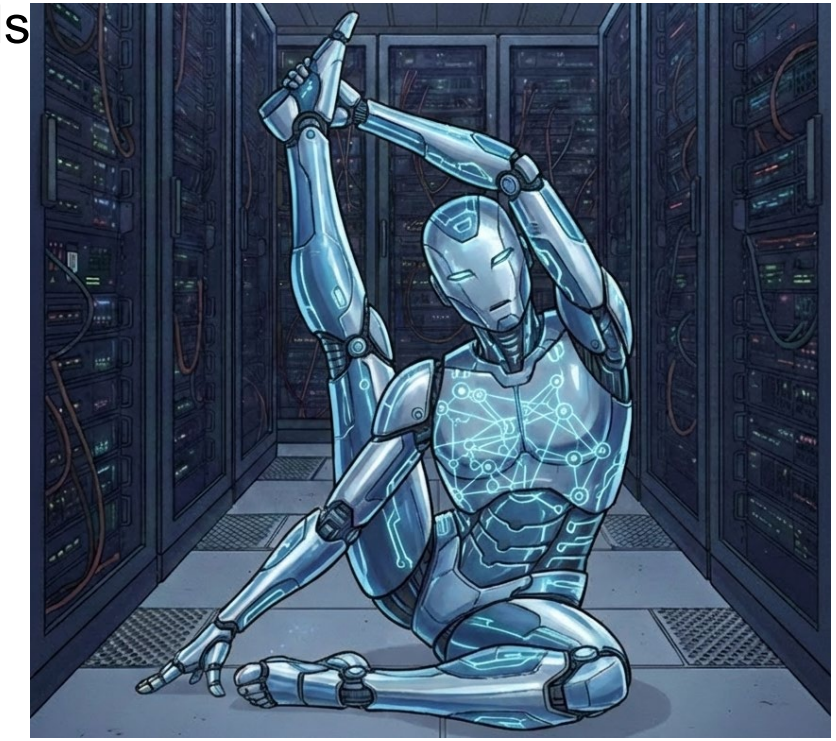
- We tested three families of systems with multiple LLM backends
  - GPT family (o3, 4o), Claude 3.5, Llama 3.3, Deepseek-R1
- **Naïve single-agent baselines**
  - One-shot pass
  - Sampled contents of the data lake



DS-GURU

Systems Under Test

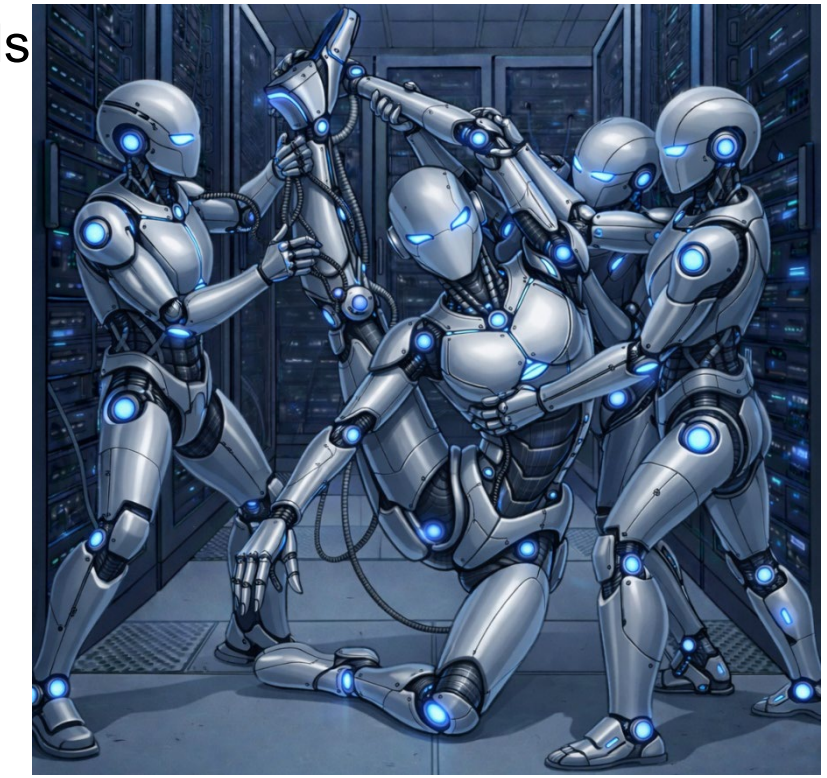
- We tested three families of systems with multiple LLM backends
  - GPT family (o3, 4o), Claude 3.5, Llama 3.3, Deepseek-R1
- Naïve single-agent baselines
  - One-shot pass
  - Sampled contents of the data lake
- **Robust single-agent baselines**
  - Few-shot debugging/iterating
  - Full contents of data lake



DS-GURU II

Systems Under Test

- We tested three families of systems with multiple LLM backends
  - GPT family (o3, 4o), Claude 3.5, Llama 3.3, Deepseek-R1
- Naïve single-agent baselines
  - One-shot pass
  - Sampled contents of the data lake
- Robust single-agent baselines
  - Few-shot debugging/iterating
  - Full contents of data lake
- **Deep Research and Multi-agent baselines**
  - Reflection and planning loops
  - Access to web search and tools



smolagents Reflexion

Pipeline design easier than implementation

- Best system **only obtains** 52.2% end-to-end accuracy

- Single-shot LLM fail on most tasks
- Few-shot show improvements
- Multi-agent reflection approaches SOTA

- Pipeline design is easier than implementation
  - Models identify required steps but can't solve them
  - Domain-specific knowledge, e.g., data cleaning

Baseline	End-to-End	Pipeline Design	Sub-task coding	Avg. Runtime / Task
DS-Guru	9.64%	40.60%	12.95%	0.54 min
DS-Guru II	22.08%	41.58%	19.75%	0.76 min
smolagents	50.64%	42.14%*	14.23%	6.10 min
OpenAI Deep Research	52.18%	—	—	10.35 min
Human Experts	71.07%	64.00%	—	4.36 min

SemBench

- We target semantic systems that optimize for processing large workloads
- Different from KramaBench, we assume the query implementation is provided

- Goal: evaluate **efficient** multimodal queries
  - Focus on solutions with **max quality** and **minimum cost**
  - Benchmark comprises 6 multimodal scenarios (text, image, tables, audio)
  - We formulate 61 "AI-powered" queries

```
SELECT Patient ID WHERE :
    AI('The patient is sick according to the symptoms.') ON Text description of symptoms
    AI('The patient is sick according to the audio.') ON Lung Audio
    AI('The patient is sick according to the X-ray image.') ON X Ray Image
    AI('The mole or skin patch is malignant/sick according to the image.') ON Skin mole Image
```

[11] J. Lao, et al. "SemBench: A Benchmark for Semantic Query Processing Engines," *Under revision*, 2026.

Systems Under Test

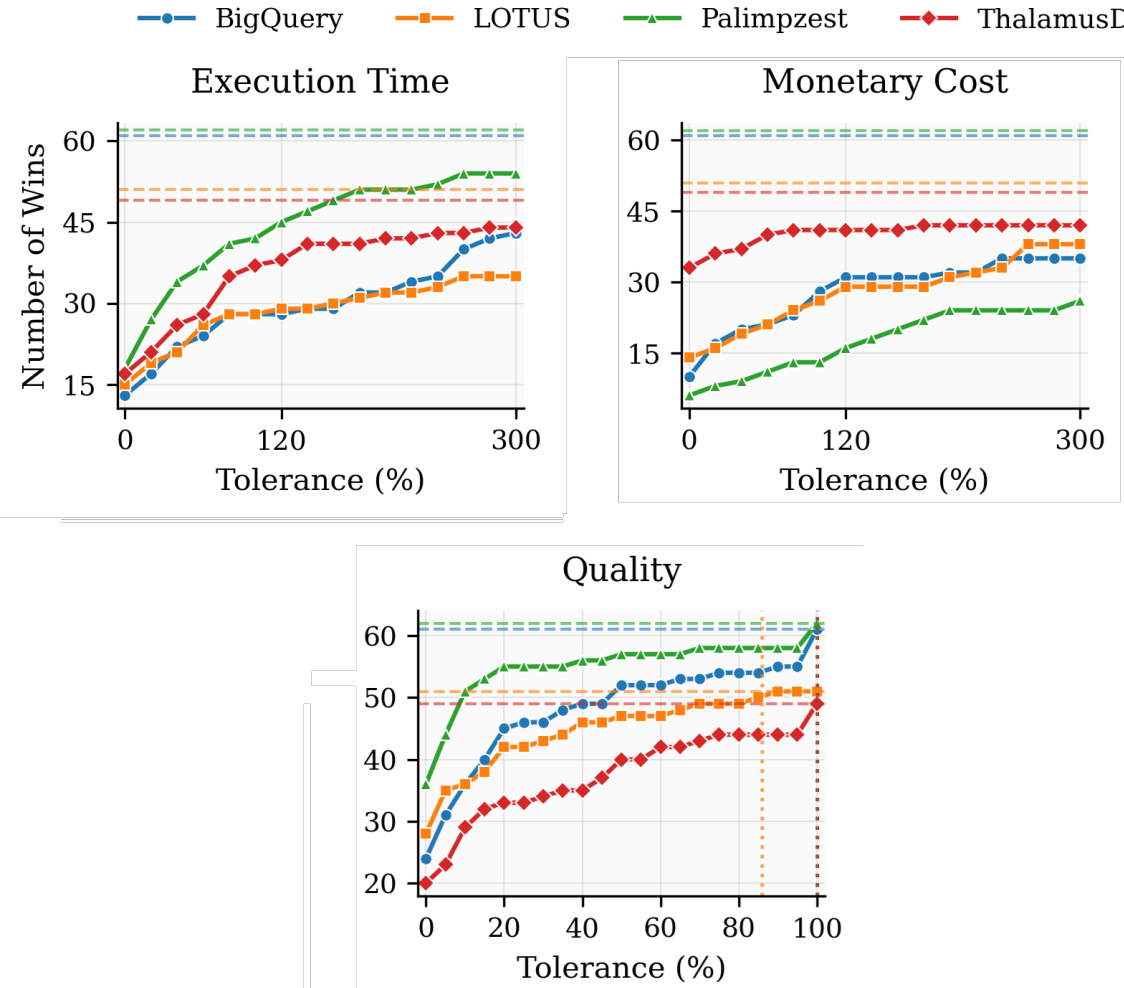
- We evaluated multiple semantic query processing engines
  - LOTUS, Palimpzest, ThalamusDB, Google BigQuery
  - Backend LLMs: Gemini, GPT-4o mini, GPT-5
  - Assume default configurations for systems

- Measurements
  - Relative error and F1 score
  - Execution time, monetary costs, memory consumption

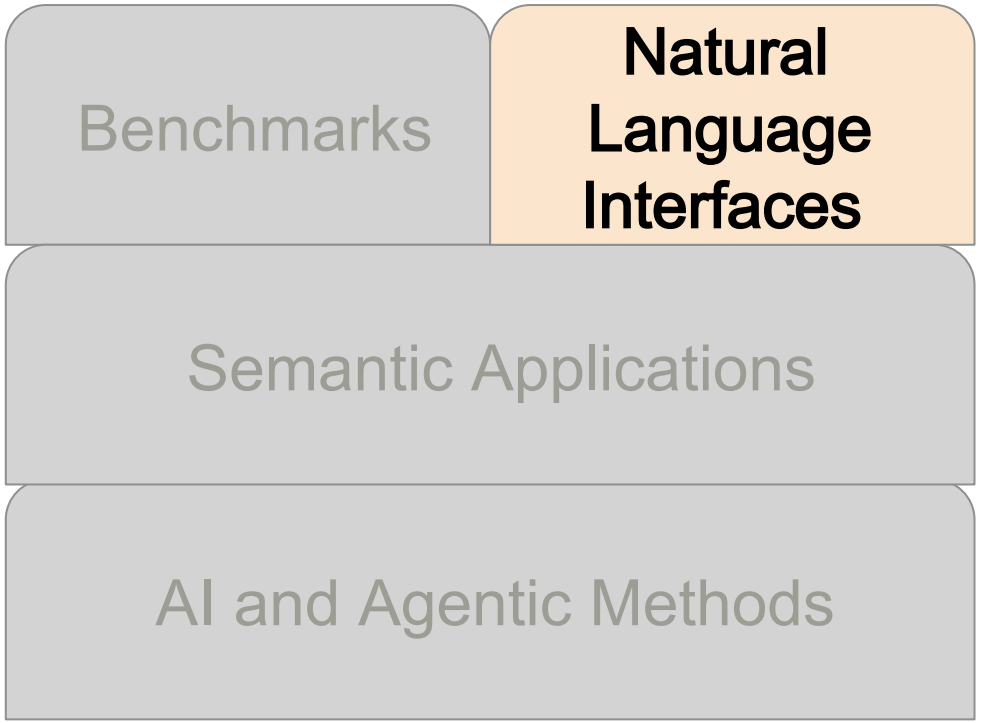
- Overall score: a system “wins” on a query if it's better than others
  - We allow a given tolerance

Semantic joins dominate costs

- General results:
  - No single system supports all queries
  - Large cost and quality variance
- System performances:
  - Multimodal join queries dominate cost and are the least supported
  - LIMIT queries are unoptimized (only post-processing)
  - Reasoning is costly but not effective
- Open challenges:
  - Adaptive LLM invocation and query plan optimization
  - Balancing accuracy and cost for multimodal joins



A Graphical Agenda



PalimpChat
Caravaggio

A data programming assistant

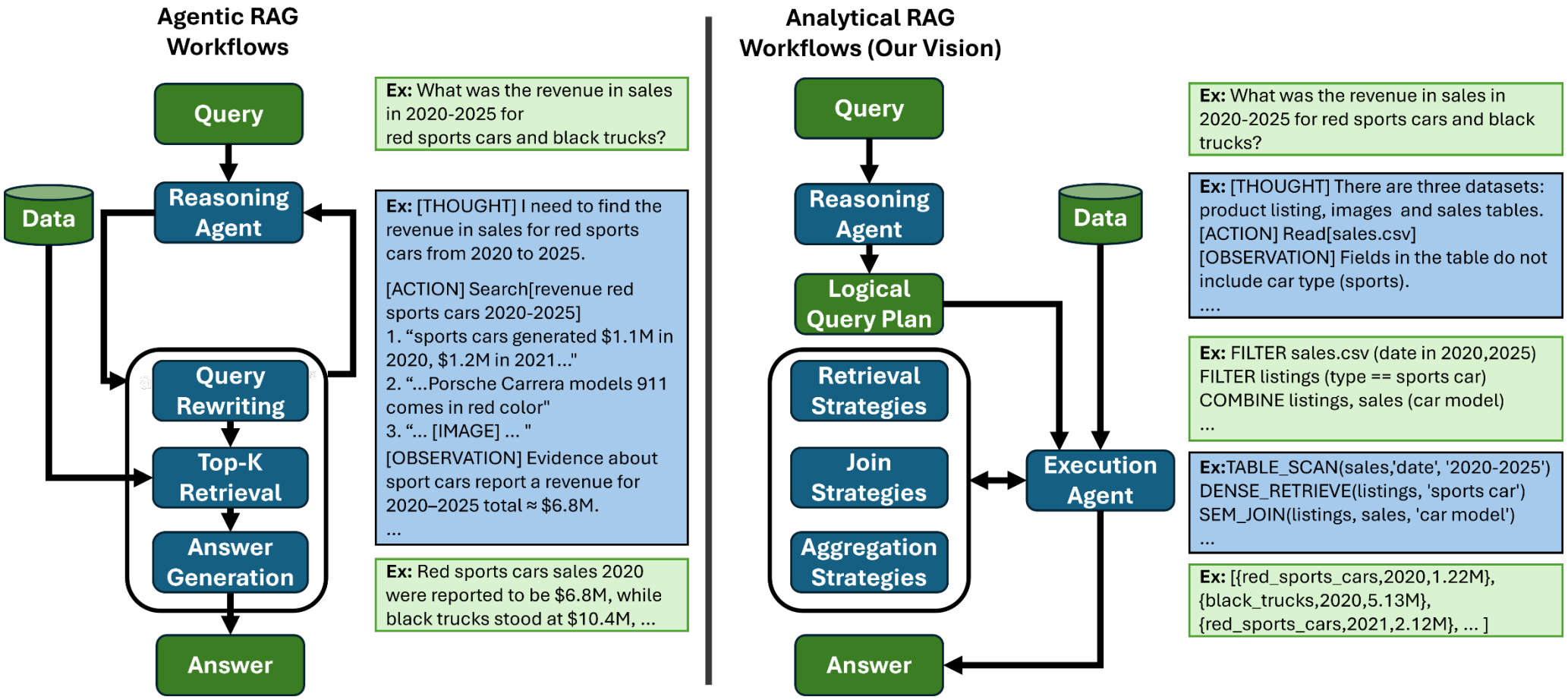
- Domain experts != coding experts
  - The less new tools, the better
  - Quick prototyping and debugging

- Designing a chat/notebook based copilot
  - A custom notebook frontend (based on Jupyter)
  - A coding assistant agent (based on ReAct)
  - A backend for data workloads (Palimpzest)

- The combination enables a new way to code data pipelines

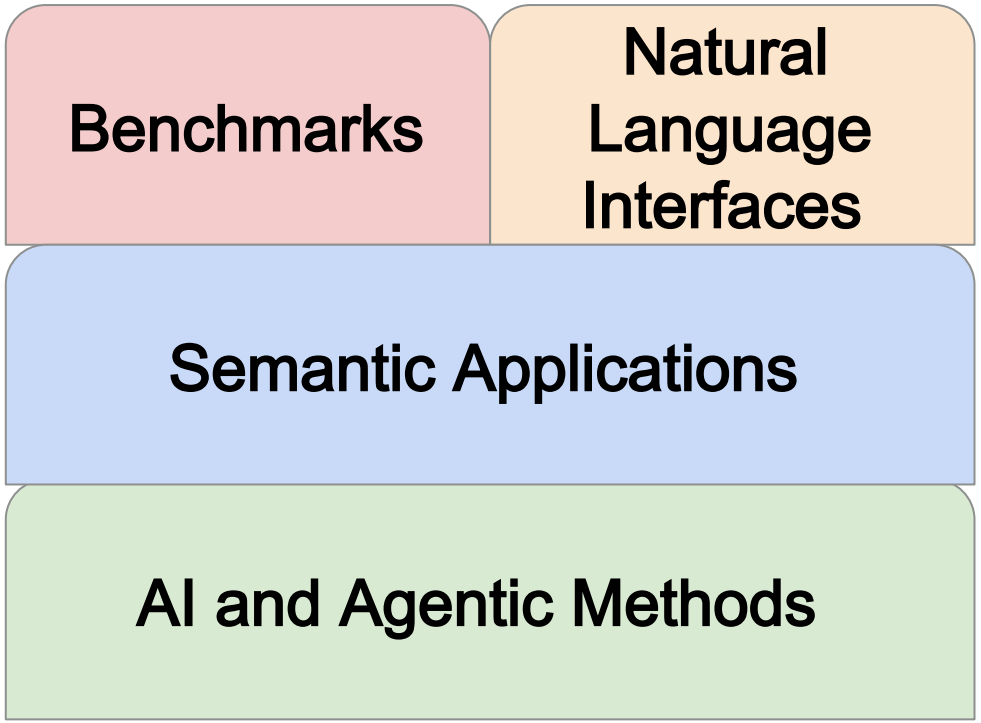
[9] C. Liu et al., "PalimpChat: Declarative and Interactive AI analytics". In *Companion of SIGMOD*, 2025

From ReActive to Proactive



A Graphical Agenda

KramaBench
SemBench



PalimpChat

Palimpzest

Claude

Gemini

Open Challenges

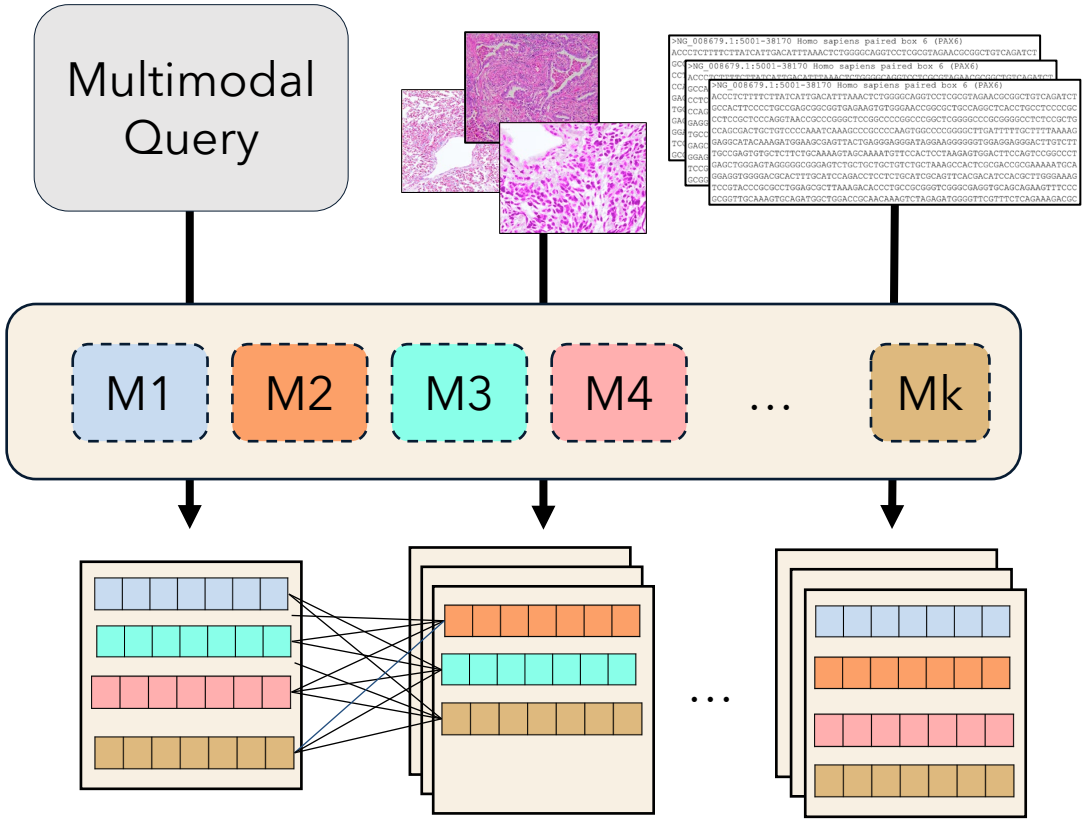
- **Logical Query Planning:**
  - How to derive a program from natural language sentence and unstructured data?
  - Challenges:
    - No fixed operators
    - No fixed schema, missing join keys
    - Ambiguous query intent

- **Query-driven indexing**
  - How to optimize the choice of chunking and indexing strategies for unstructured data based on the characteristics of a logical query plan?
  - Challenges:
    - Balance indexing costs
    - Infer attributes from context

Open Challenges

- **Native multimodal support:**
  - How to unify retrieval across heterogeneous modalities (text, image, audio, video)?
  - Challenges:
    - Alignment of modality-specific embeddings
    - Avoid mode collapse

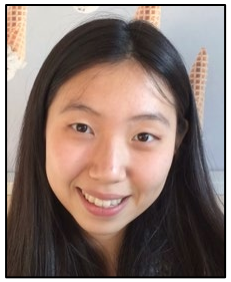
- **Semantic joins:**
  - How to design and optimize joins operations over unstructured data?
  - Challenges:
    - Noisy join attributes
    - Avoid combinatorial costs or cross-products



Acknowledgements


Matt Russo

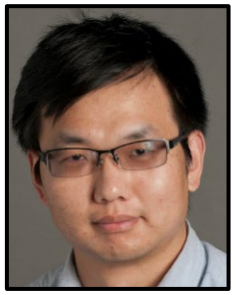

Sylvia Zhang


Eugenie Lai


Peter Baille Chen


Ferdi Kossmann


Zui Chen


Chunwei Liu


Mike Cafarella


Tim Kraska


Sam Madden

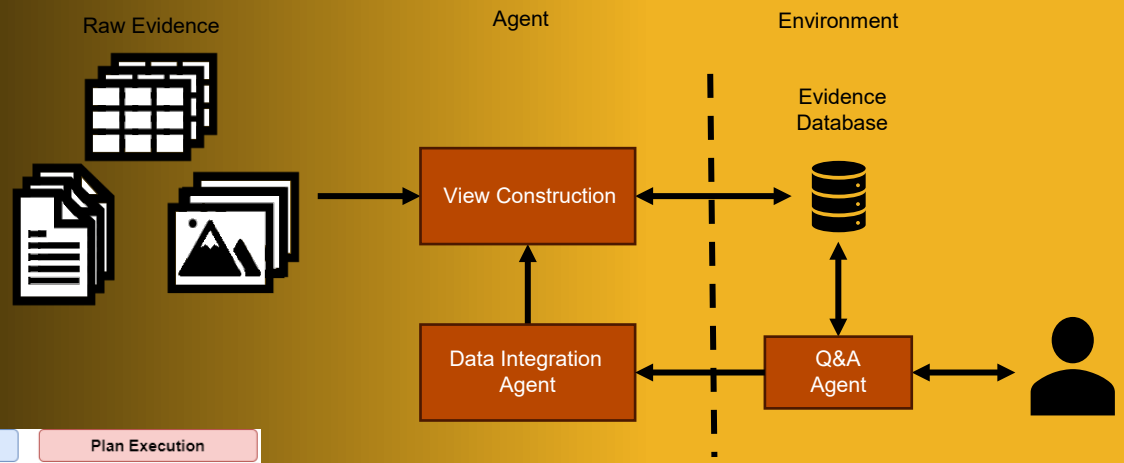

Michael Franklin

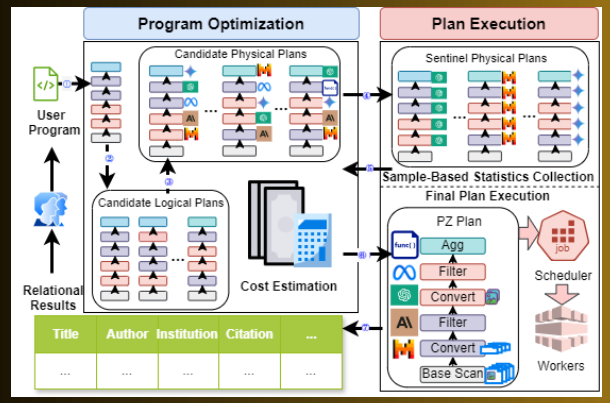

Lei Cao

Agentic Data Systems and Multimodal Analytics

Gerardo Vitagliano

MIT CSAIL Computer Science & Artificial Intelligence Laboratory





Benchmarks

Natural Language Interfaces

Semantic Applications

AI and Agentic Methods

References

[1] Tian, Shulin, et al. "MMInA: Benchmarking multihop multimodal internet agents." *Findings of the ACL* 2025.

[2] Li, Shilong, et al. "MM-browsecomp: A comprehensive benchmark for multimodal browsing agents." *arXiv preprint arXiv:2508.13186* 2025

[3] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proceedings of NeurIPS* 2020

[4] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of NeurIPS* 2022.

[5] S. Yao *et al.*, "ReAct: Synergizing Reasoning and Acting in Language Models," *Proceedings of ICLR* 2023.

[6] Z. Jiang *et al.*, "Active Retrieval Augmented Generation," in *Proceedings of EMNLP* 2023.

[7] C. Liu et al., "Palimpzest: Optimizing AI-Powered Analytics with Declarative Query Processing," in *Proceedings of CIDR*, 2025.

[8] M. Russo et al., "Abacus: A Cost-Based Optimizer for Semantic Operator Systems,", in *PVLDB*, 2026

[9] C. Liu et al., "PalimpChat: Declarative and Interactive AI analytics". In *Companion of SIGMOD*, 2025

[10] E. Lai et al., "KramaBench: A Benchmark for AI Systems on Data Intensive Tasks," in *Proceedings of ICLR*, 2026

[11] J. Lao, et al. "SemBench: A Benchmark for Semantic Query Processing Engines," *Under revision*, 2026.

References

- Repositories:
  - https://www.kramabench.org
  - https://github.com/mitdbg/palimpzest
  - https://github.com/jataware/bdf-pz
  - https://github.com/vitaglianog/caravaggio
  - https://github.com/mitdbg/kramabench
  - https://sembench.ngrok.io