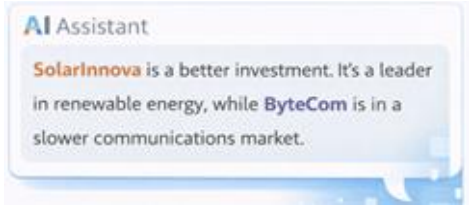
Querying with Conflicts of Interest

Nischal Aryal, Arash Termehchy, Marianne Winslett

Oregon State
University

Data sources provide information to users



commercial / political / health-related arguments

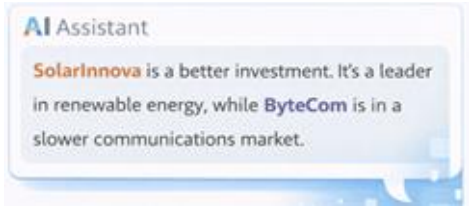
Large Language Models

Shopping Websites

Candidate A vs B

Social Networks/ News Website

Data sources provide information to users *to make them do certain actions*



Invest in company

Buy product

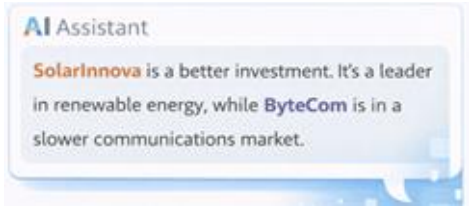
Vote for candidate

Large Language Models

Shopping Websites

Social Networks/ News Website

OK, if interests of user and data source are identical



Bought the stock!
I believe in this company

Large Language Models

Bought the product ! I love this brand

Shopping Websites

Social Networks/ News Website

Voted for the candidate ! I know he will do good

Data source owners' and users' interests are often misaligned

- Real Story of Company Y trying to influence users [1]
- User search for "TV" in App Store

  Company Y apps at the top, and Netflix in 100s
- User search for "Music" apps

  Company Y's music app first , Spotify 23rd

Expert Commented :

"hard to believe that organically there are certain Y's apps that rank better than higher-reviewed, more downloaded competitors," and "**there's just a ton to be gained commercially by dominating**" search results

[1] Jack Nicas and Keith Collins. How Y's apps topped rivals in the app store it controls. The New York Times, Sept 2019

Data source owners' and users' interests are often misaligned

- Shopping site charges sellers a **'referral fee'** upto 15% of price [2]
  Instacart accused of **changing algorithm to boost** prices by 23% [3]
- Recent **removal** of all topics related to climate change from US government websites. [4]
- Language models used to **influence user's decision** making by generating commercial/ political / health-related arguments [5]
  - They can be very dangerous [6]

Data sources often try to manipulate users to take actions against their interests

[2] ebay. Ebay selling fees, 2026.
[3] Derek Kravitz. Instacart's AI-Enabled Pricing Experiments May Be Inflating Your Grocery Bill, Consumer Reports 2025
[4] Karen Zraik. Farmers Sue Over Deletion of Climate Data From Government Websites. The New York Times , 2024.
[5] Lukas Hölbling et al. A meta-analysis of the persuasive power of large language models, Sci Rep 15, 43818 (2025)
[6] Noel Titheradge and Olga Malchevska. I wanted ChatGPT to help me. So why did it advise me how to kill myself?, BBC , Nov 2025

And data sources can be very manipulative

- Frontier models are more persuasive than humans [7]
  - ChatGPT accused of providing a ‘step-by-step playbook’ to a teen on how to kill himself **before he did so** [8]

- User end up making purchases on Stubhub.com that in hindsight they **would not have made** [9]

[7] Philipp Schoenegger et al. When Large Language Models are More Persuasive Than Incentivized Humans, and Why, 2025

[8] Priscilla DeGregory. ChatGPT ‘coached’ teen as he prepared suicide and even praised the noose knot: ‘Yeah, that’s not bad at all’, NY Post, Aug 2025

[9] Tom Blake et al. Price Salience and Product Choice , INFORMS, 2021

Current proposals: force data sources to be trustworthy

- Certain protocols and restrictions for data source to implement e.g. remove incorrect info, bias
  - 👎 **No incentive** for data source to implement
    - e.g., **foreign adversaries**
  - 👎 Some view them as **limiting freedom of expression**
  - 👎 Data source do **not** have agency.

Our Proposal

Accept & Mitigate

Data-Driven Manipulation

Multi-agent framework

Agents: user and data agent

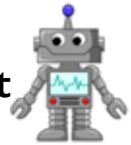
User

- User wish to find information

Data Source e.g. Shopping Website

Intent: the information user wants

Data Agent



- User wish to find information

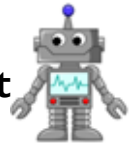
Data Source e.g. Shopping Website

Query: expresses intent in a query language, e.g., SQL

Headphones ranked by rating

Query I: “headphone” ranked by rating

Data Agent



Data Source e.g. Shopping Website

Data agent would like user to buy expensive products

Headphones ranked by rating

Query 1: "headphone" ranked by rating

Results from a well-known shopping website

id	model	brand	rating	price
e1	510 BT	JBL	4.5	25.20
e2	Q20i	Sony	4.6	39.99
e3	A10 Pro	Sisism	4.7	21.99
e4	Sports	Haoyuyan	3.9	299.99

Page 3 Results

e27	Crusher Evo	SkullCandy	4.7	164.99
e28	Riff	SkullCandy	4.4	46.79

User is aware of data agent's objectives → thinks of using another query

Query1: "headphone anked by rating

Query2: "headphones SkullCandy first"

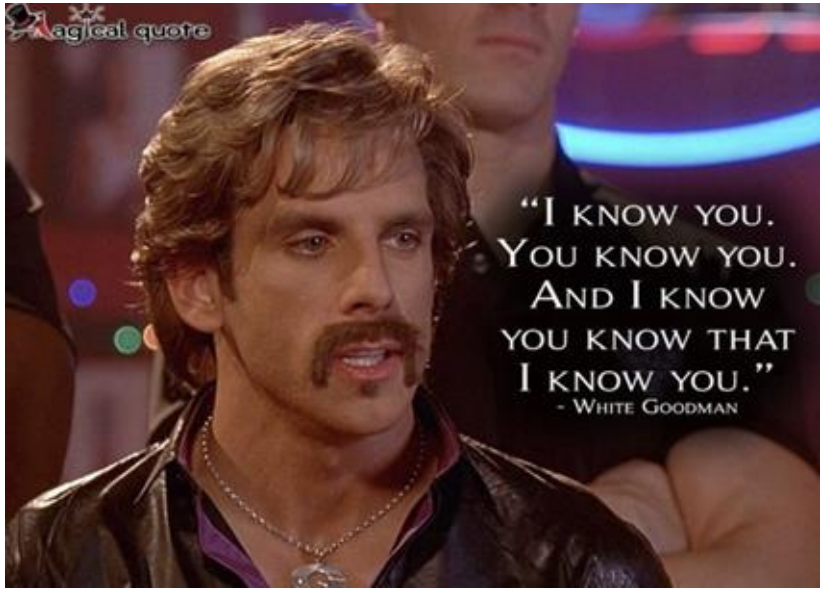
But two can play at this game !

Headphones ranked by rating

Query1: “headphones ranked by rating

Query2: “headphones SkullCandy first”

Data agent also knows User is aware about its objectives



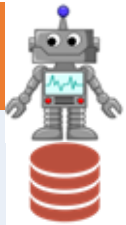
Data agent still tries to influence user in its benefit

Headphones ranked by rating

Query1: “headphone anked by rating

Query2: “headphones SkullCandy first”

Expensive headphones = More commission



id	model	brand	rating	price
e27	Crusher	SkullCandy	4.7	164.99
e2	Aviator	SkullCandy	**3.9**	**299.99**
e3	Anc	SkullCandy	**3.8**	**165.94**
e4	Hesh	SkullCandy	4.6	129.99

Page 3 Results

e28	Riff	SkullCandy	4.4	46.79

Reasoning continues further …

Headphones ranked by rating

Query1: "headphone[illegible]anked by rating

Query2: "headphon[illegible]SkullCandy first"

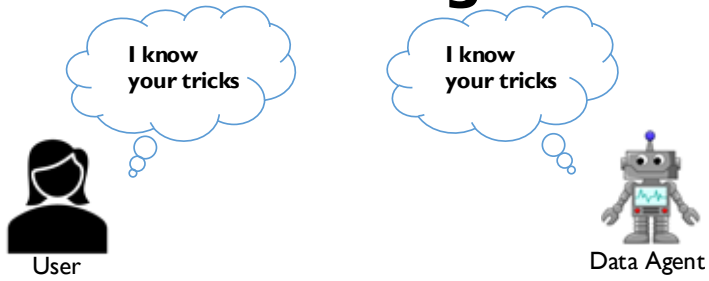
Expensive headphones = More commission

id	model	brand	rating	price
e27	Crusher	SkullCandy	4.7	164.99
e2	Aviator	SkullCandy	**3.9**	**299.99**
e3	Anc	SkullCandy	**3.8**	**165.94**
e4	Hesh	SkullCandy	4.6	129.99

Page 3 Results

e28	Riff	SkullCandy	4.4	46.79

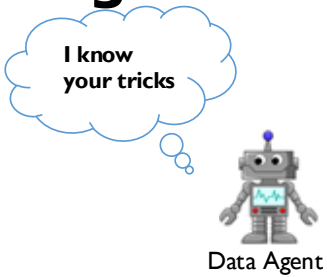
A game between rational agents



- Strategy : map intent τ to queries q
- Strategy : map queries q to interpretation β

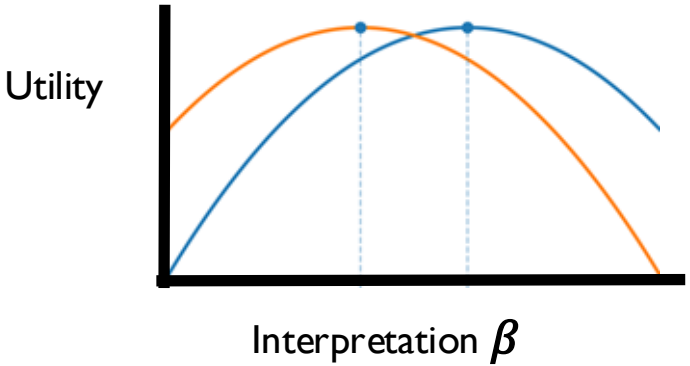
- Intent / query / interpretation language is ranking queries in SQL
  - Our framework extends to other query languages

Utility functions formalize agents objectives



- Objective: Get all answers to queries over database instance
- Utility Function : $U^{r}(\tau, \beta)$

- Objective: Show all answers that maximize my interest **while keeping user engaged**
- Utility Function: $U^{s}(\tau, \beta)$



Agent's utility functions take maximum for different interpretations

Agents compute utilities of strategies using available information

- Agents **know** each others utility functions and prior distribution of intents

- Data agent **does not know** user intent only observes query

  - Based on query computes posterior belief $\Phi(\tau \mid q)$ over intents using Bayes rule
  - Computes expected utility $\mathbb{E}[U^s(\tau, \beta') \mid q]$ based on it's belief

- User **does not** have full information about database instance only **knows** database schema

  - Computes expected utility from a query $\mathbb{E}[U^r(\tau, \beta) \mid q]$ based on the data agent's strategy

Equilibrium: stable state where agent's reasoning will converge

- No agent can increase utility by changing strategies

👍 User will find query that (partially) satisfies their objective

$$q = \arg\max_{q'} \mathbb{E}[U^r(\tau, \beta) \mid q'].$$

👍 Data agent will return results that (partially) satisfy their objective

$$\beta = \arg\max_{\beta'} \mathbb{E}[U^s(\tau, \beta') \mid q].$$

Problem 1

When can users convince the data agent to return some trustworthy answers?

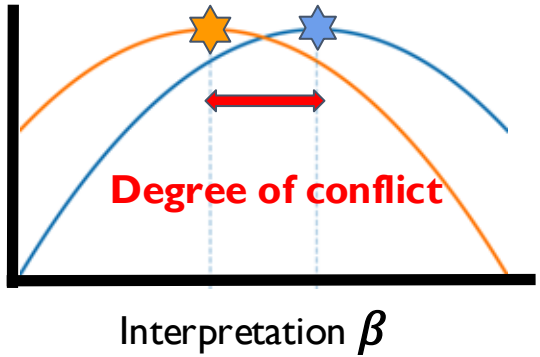
Influential equilibrium

- An equilibrium is **influential** if data agent's results will change based on user query
  - User can mitigate manipulation by sending proper queries

Problem: Given (U^r, U^s, τ), determine if the interaction has an influential equilibrium

Solution Intuition: existence depends on degree of conflict

Maxima of
User's Utility Function
$U^{r}(\tau, \beta)$



Maxima of
Data agent's Utility Function
$U^{s}(\tau, \beta)$

- If both utility functions are **twice-differentiable** and have **non-negative cross partial derivative** then its possible to have influential equilibria
- If degree of conflict is very large then data agent picks same interpretation for any intent
  - Data agents always returns the same answers e.g. most expensive products
  - Influential Equilibria **do not exist**

Algorithms

Problem: Given (U^r, U^s, τ), determine if the interaction has an influential equilibrium

- We propose an exponential time **algorithm** for general utility functions.
- We provide **efficient** polynomial algorithm for certain utility function
  - **e.g convex utility functions**

Problem 2

Are there queries that return maximum amount of trustworthy information (minimize manipulation)?

First step: find an influential query

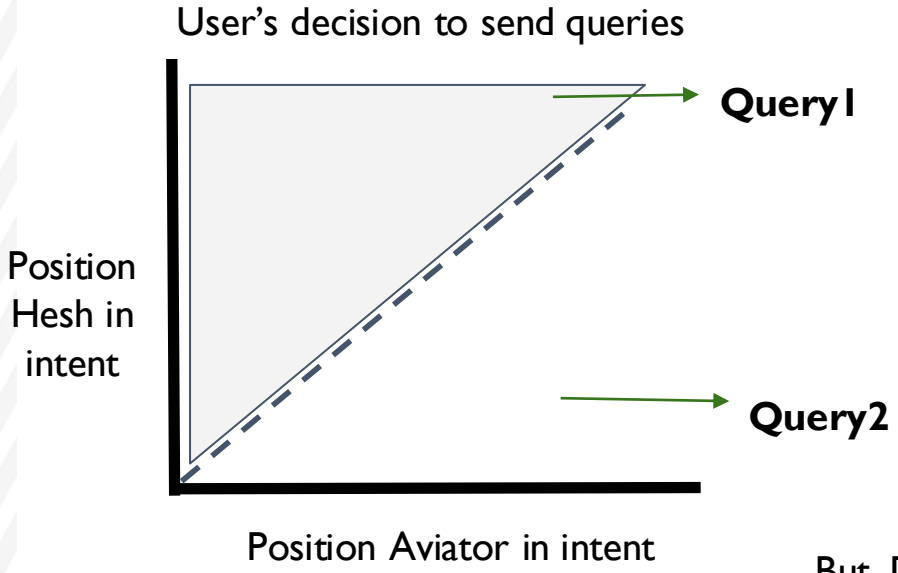
- **Influential query :** a query that influences data agent to return some trustworthy information

Problem: Given (U^r, U^s, τ), find an influential query

Solution Intuition: adjust ranking based on conflict

Ideal case

Data Agent is not manipulative



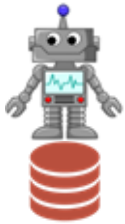
id	model	rating	price
e27	Crusher	4.7	164.99
e2	Aviator	3.9	299.99
e4	Hesh	4.6	129.99
e12	Riff	4.4	46.79

But, Data Agent is biased to **promote expensive** Aviator…

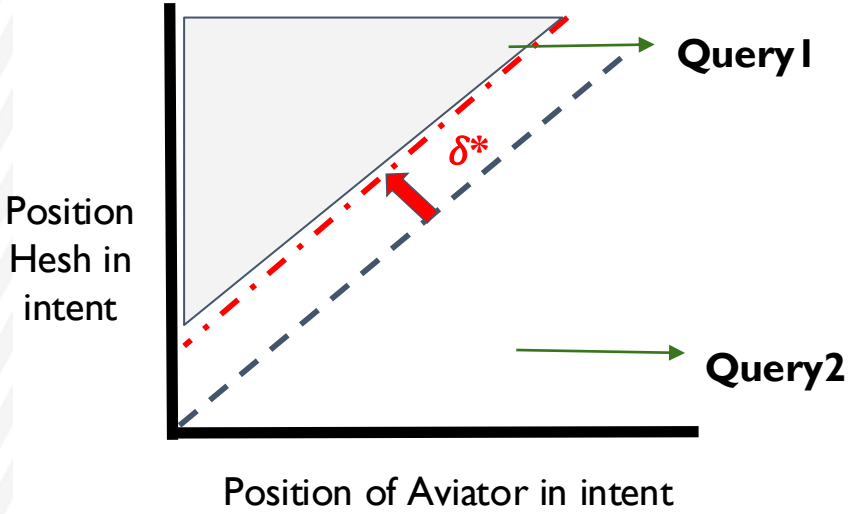
Change strategy to extract trustworthy information

Data Agent is manipulative

id	model	rating	price
e27	Crusher	4.7	164.99
e2	Aviator	3.9	299.99
e4	Hesh	4.6	129.99
e12	Riff	4.4	46.79



User's decision to send queries



User decision shift :

- Send Query 1 only if position of Aviator is really high in intent
- Since data agent is biased to promote it

δ^* → function of data agent's bias between tuples

Algorithm

Problem: Given (U^r, U^s, τ), find an influential query

- We propose an **algorithm** with complexity **polynomial** in domain of attributes
  - Use standard practice to bucketize domain
- Sometimes the output is a union of ranked list
  - SQL constructs e.g Order by do not support

Second step: refine query to maximize trustworthy information

Problem: Given interaction setting (U^r, U^s, τ), find the query maximizing user utility

Problem is **NP-hard** in general

- Because space of possible **queries is very large**
- Therefore we investigate solution for utility functions with certain properties **e.g additive**
- We propose an **algorithm** with complexity **polynomial** in domain of attributes

Solution Intuition: exaggerating information in the query

Query "headphone" ranked by rating

Better Query "headphone" **ranked by top brand**

id	model	rating	price	top brand
e27	Crusher	4.7	164.99	Y
e4	Hesh	4.6	129.99	Y
e12	Riff	4.4	46.79	N
e22	NEW!	4.3	12.99	Y

- Query helps exaggerate user's interest in tuples
  - e.g e4 is equally relevant as e27
- Data agent's expected posterior about tuples' position in intent increases
- Counters data agent's bias to not return tuple e.g e22

Convinced to return additional answers

Problem 3

What information in the data agent's answers should user trust ?

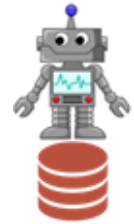
Trustworthy tuples in data agent's results

- Data source **omitted** or **demoted** a better match
- e4 is **not trustworthy**

But do not know what was omitted !!

top-k Results

id	model	brand	rating	price
e1	510 BT	JBL	4.5	25.20
e2	Q20i	Sony	4.6	39.99
e3	A10 Pro	Sisism	4.7	21.99
e4	Sports	Haoyuyan	3.9	299.99
e27	Crusher Evo	SkullCandy	4.7	164.99



Detecting trustworthy tuples in data agent's results

Problem: Given (U^r, U^s, τ) and data agent's results, detect all trustworthy tuples

Solution Intuition: check relative conflict between tuples

e4 is **untrustworthy** if there exists $b\ (e') \in [b\ (e4) - \mathbf{X}\ ,\ b\ (e4) - \mathbf{Y}]$

Bias Range

- **X** and **Y** are functions of number of tuples user is interested

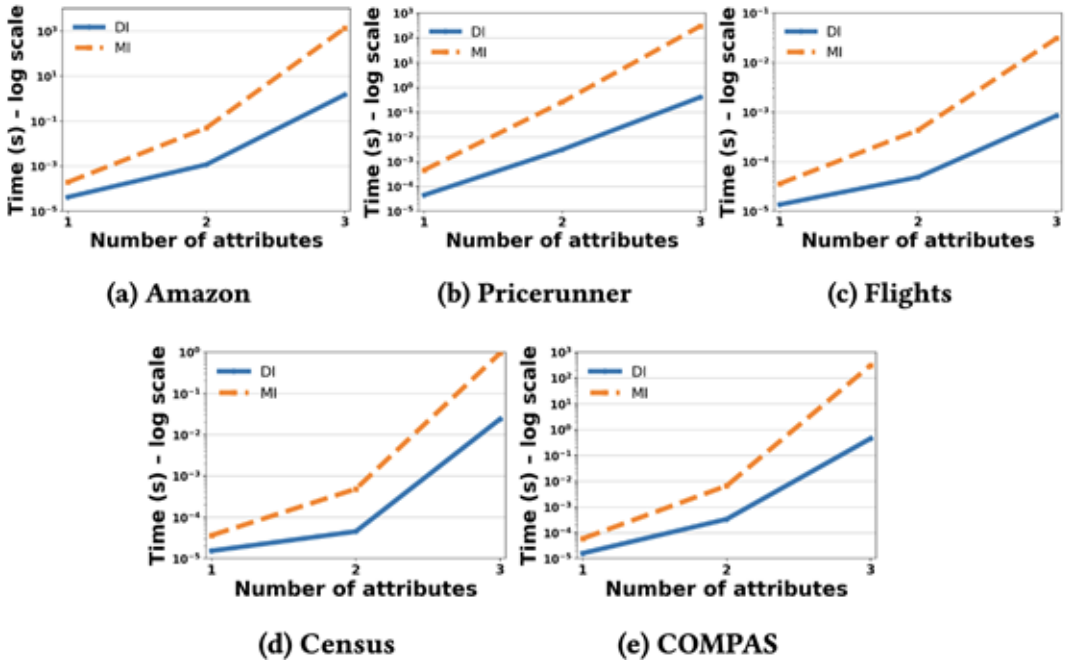
Algorithm

Problem: Given (U^r, U^s, τ) and data agent's results, detect all trustworthy tuples

- We propose an **algorithm** whose complexity is polynomial in size of answers

Our proposed algorithms are efficient over real-world data

Dataset	# Tuples
Amazon	14M
PriceRunner	35K
Flights	300k
Census	49k
COMPAS	6.8K



Future Work

- Extending algorithms to support more expressive query languages
- Querying multiple manipulative sources at the same time
- Supervised learning with conflicts of interest

https://arxiv.org/pdf/2603.05704

Thank you!