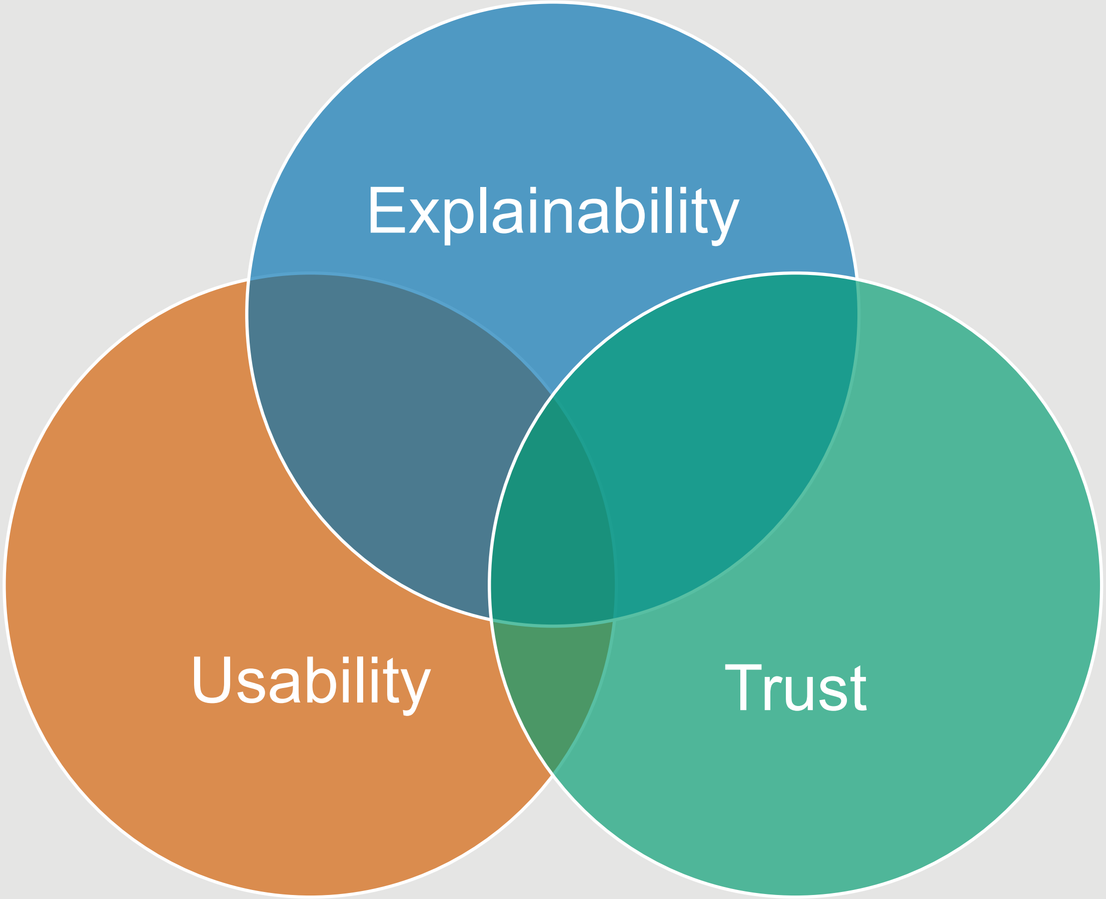
UMass Amherst | College of Information & Computer Sciences

Enhancing Usability and Explainability of Data Systems

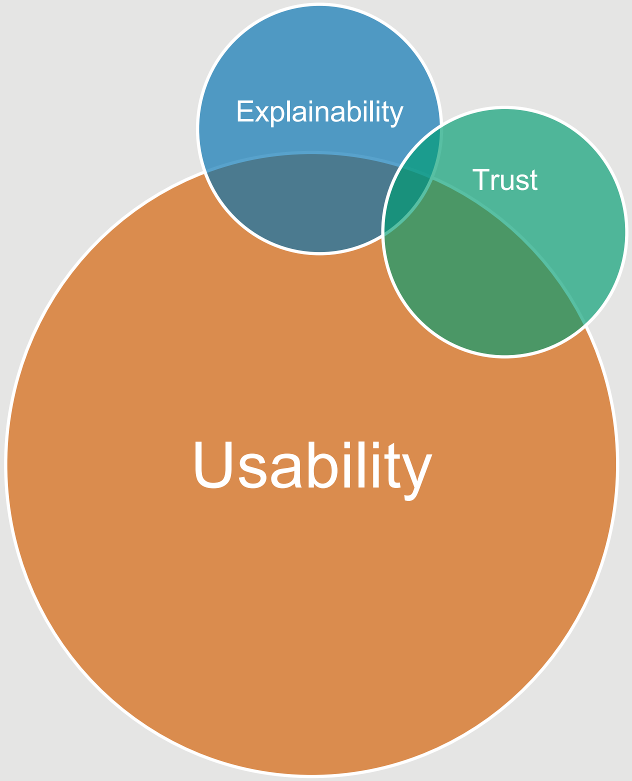
Anna Fariha

Advisor: Alexandra Meliou

Democratization of data systems



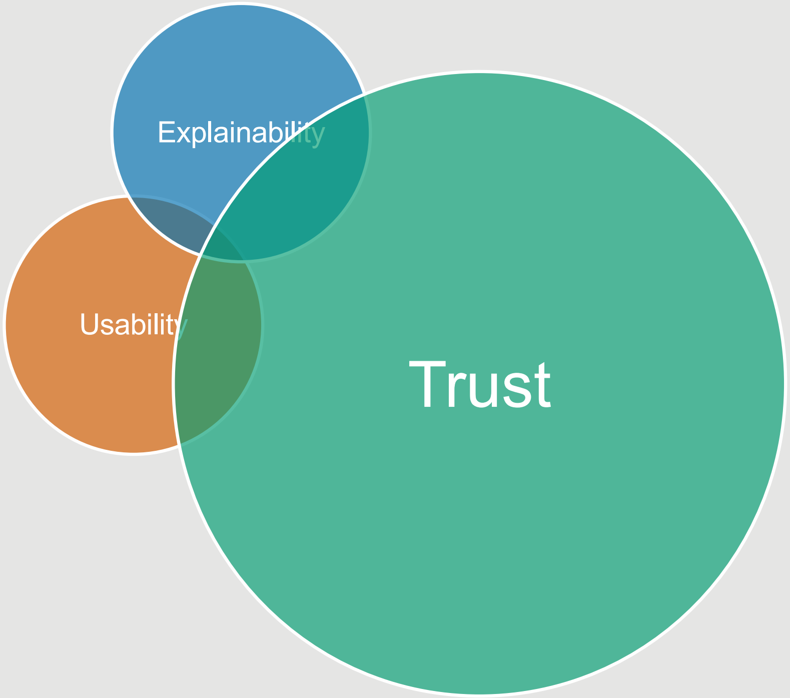
Usability



Makes data systems accessible to non-expert users.

- Applications
  - Data access
    - Querying relational databases
  - Data integration
  - Data transformation
  - Data visualization
  - Data summarization
    - Text document summarization

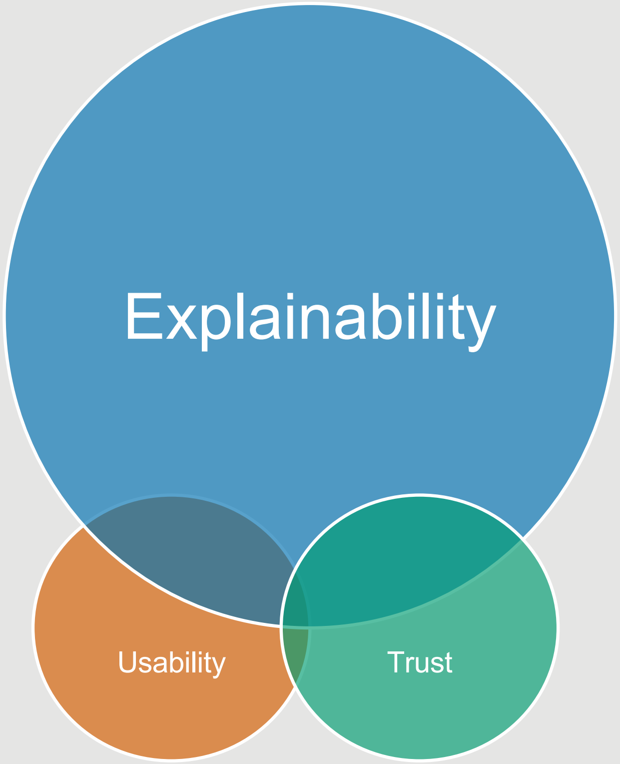
Trust



Enhances people's confidence towards data systems.

- Applications
  - Artificial intelligence and machine learning
    - Model predictions
  - Novel interaction mechanisms
    - Programming by example

Explainability



Increases transparency of data systems.

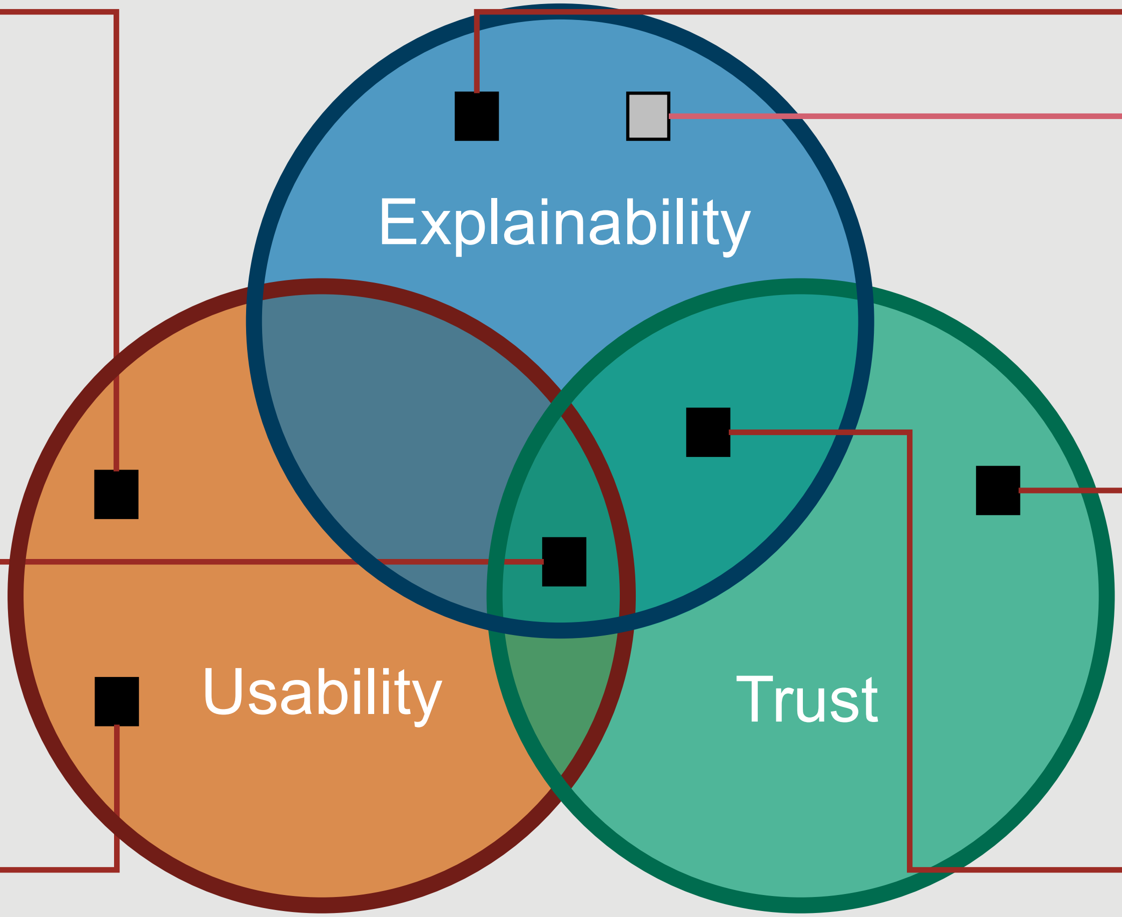
- **Applications**
  - Machine learning
    - Model predictions
  - Distributed systems
    - Concurrent applications
  - Data evolution
    - Why/how two databases differ?
  - Fairness in algorithms/software

Dissertation outline

Query by Example (QBE)
[VLDB 2019]
[SIGMOD 2018] (demo)

Comparative User Study: QBE vs SQL
[CHI 2020]*

Data Summarization by Example
[VLDB 2020] (demo)



Adaptive Interventional Debugging
[SIGMOD 2020]

Data Change Explanation
WORK IN PROGRESS

Conformance Constraints: Trusted ML
[SIGMOD 2021]*

Explaining Tuple Non-conformance
[SIGMOD 2019] (demo)

* under submission/revision

Part 1: Usability of Data Systems

Usability

Are data systems accessible to non-experts?

Microsoft SQL Server

SQLite

MySQL

ORACLE DB Management

Who are our most valuable customers?

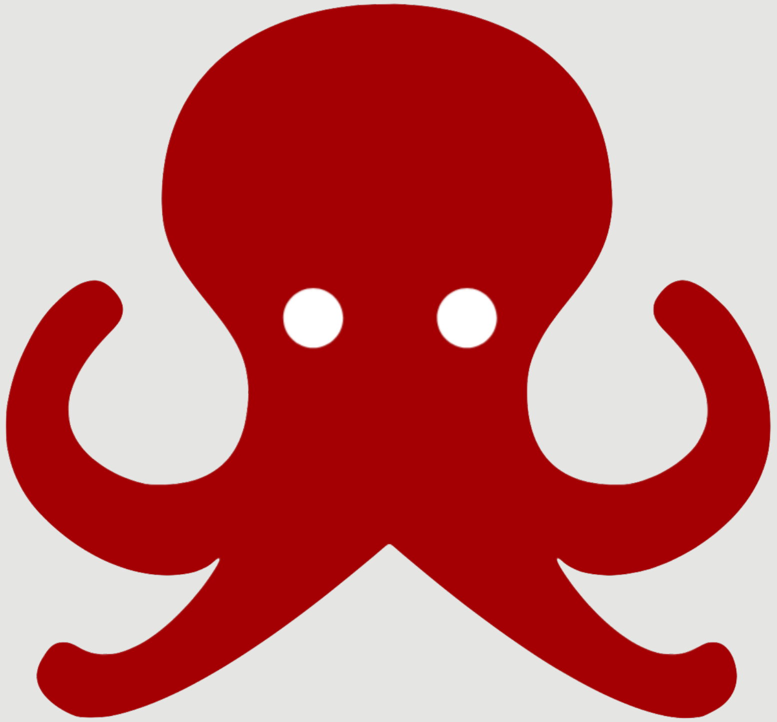
How to express complex task specifications?

Programming by example (PBE)

- A step towards democratization of computational power.
- Enhances usability for both non-experts and experts.

User provides examples → System "guesses" intent → Program synthesis → Result delivery

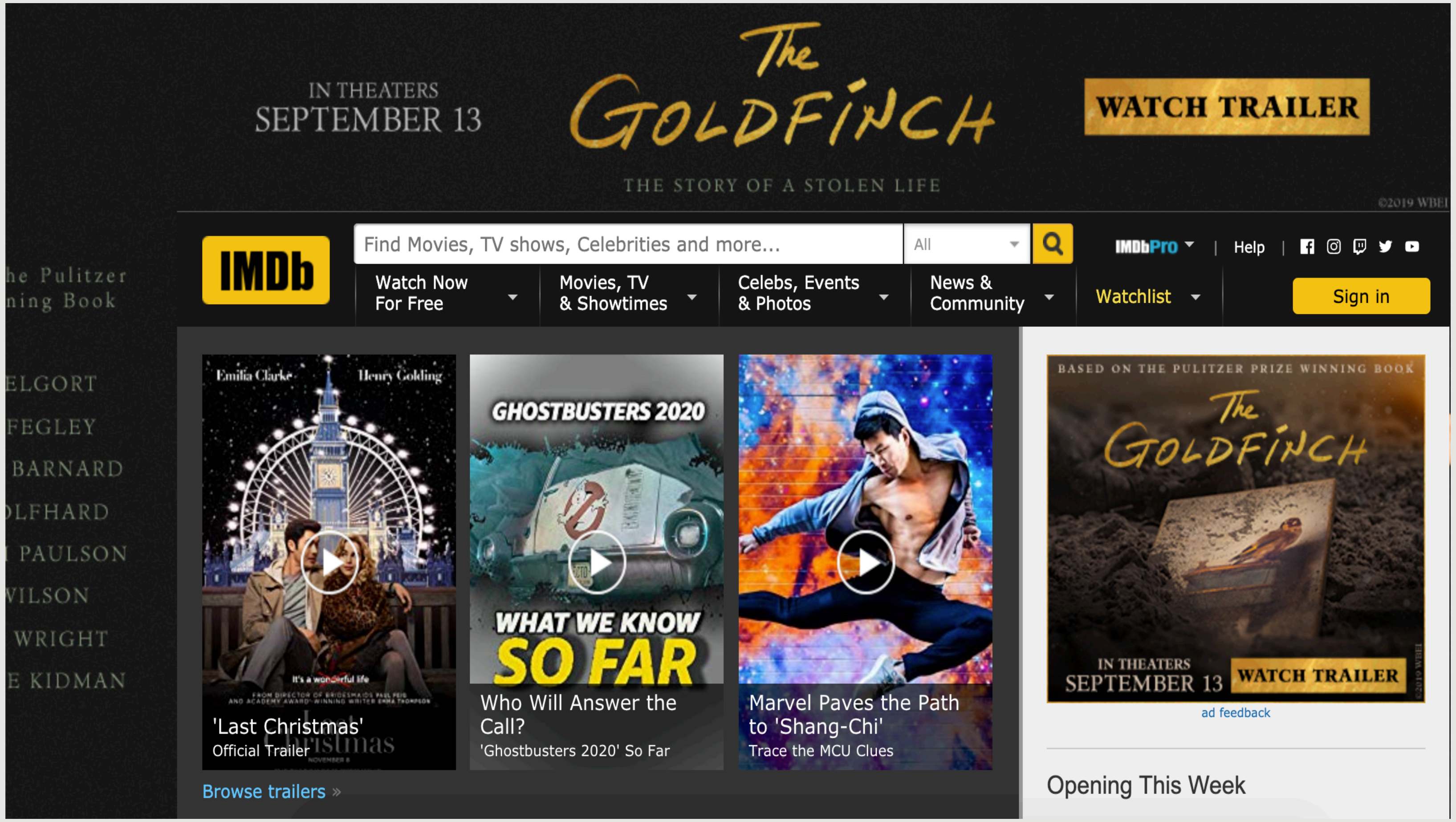
Querying relational databases by example



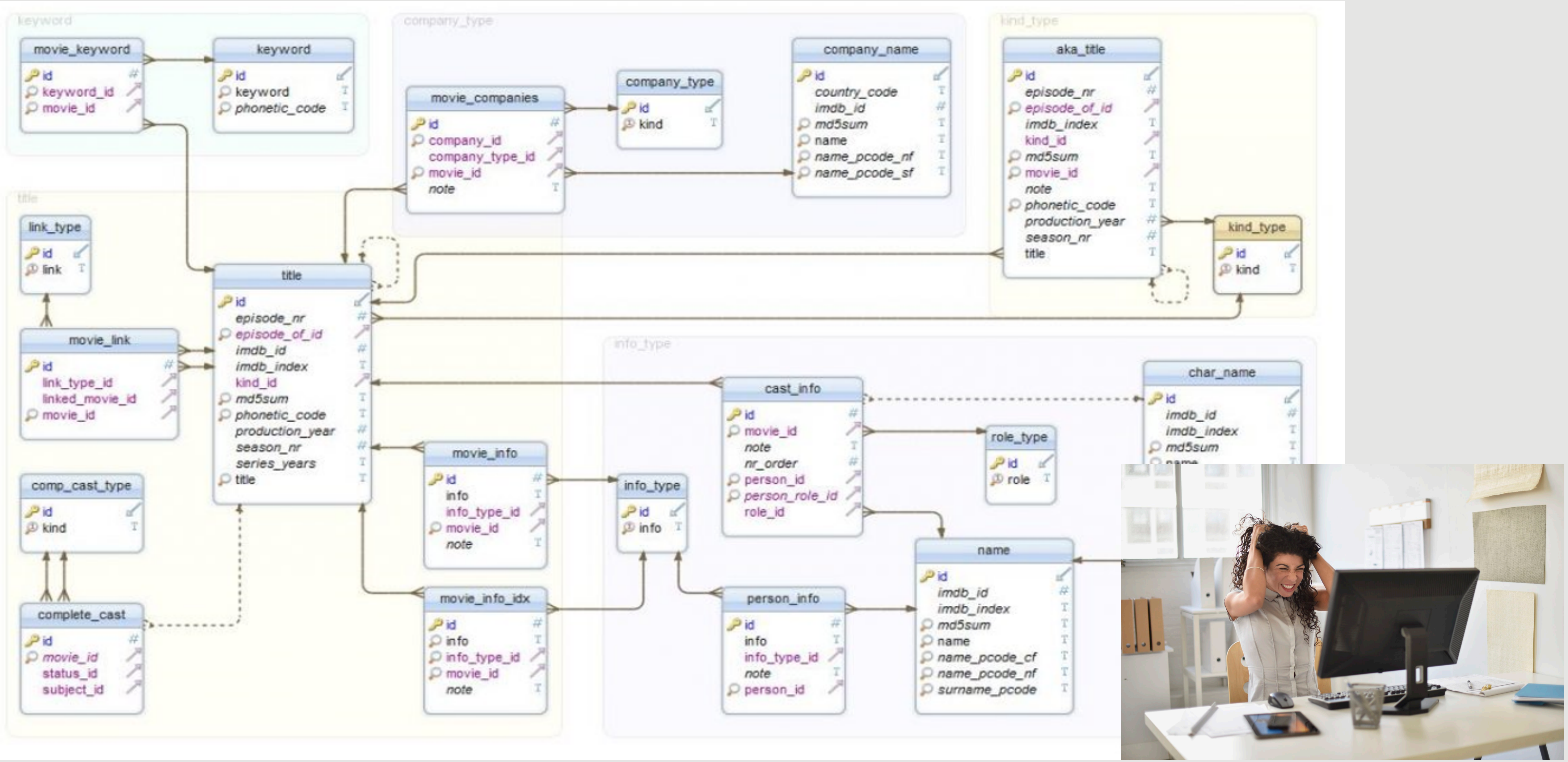
SQuID

Semantic similarity-aware Query Intent Discovery

- Alice wants to find all Funny Actors from the IMDb database.



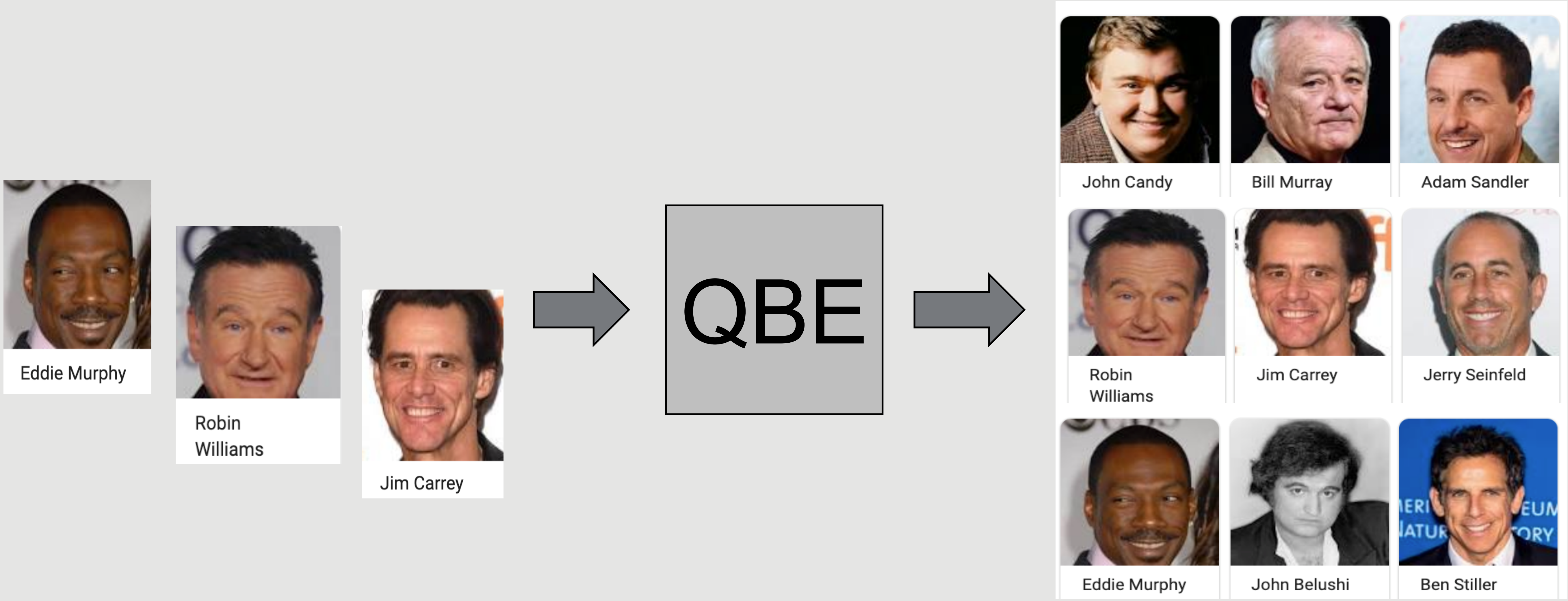
Challenge 1: understanding the schema



Challenge 2: SQL expertise

```
SELECT person.name
FROM person, castinfo, movietogenre, genre
WHERE person.id = castinfo.person_id
  AND castinfo.movie_id = movietogenre.movie_id
  AND movietogenre.genre_id = genre.id
  AND genre.name = 'Comedy'
GROUP BY person.id
HAVING count(*) >= 40
```

Query by example (QBE)



Expectation vs reality

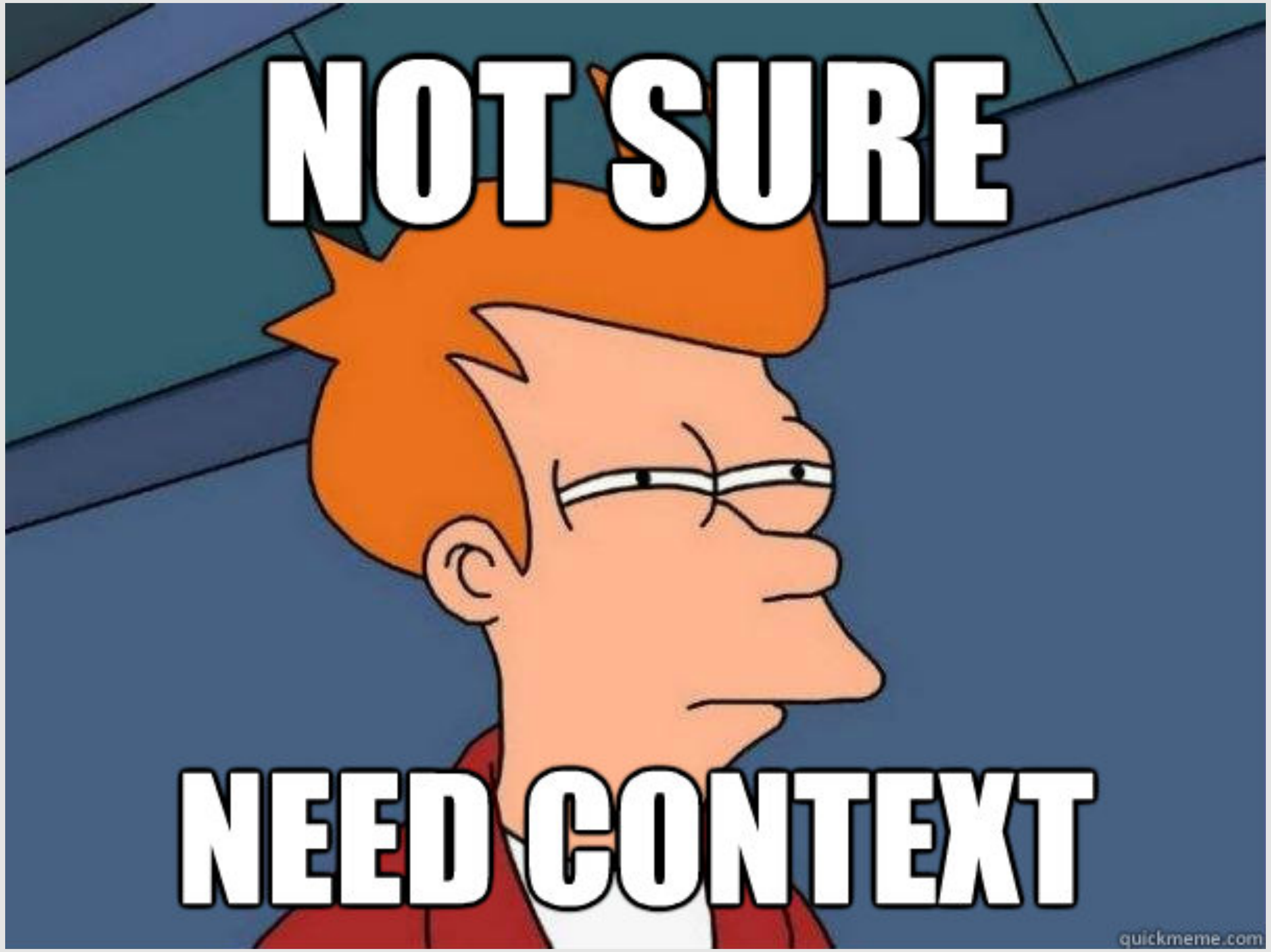
Eddie Murphy

Robin Williams

Jim Carrey

All actors

Humans use context



Discovering semantic similarity

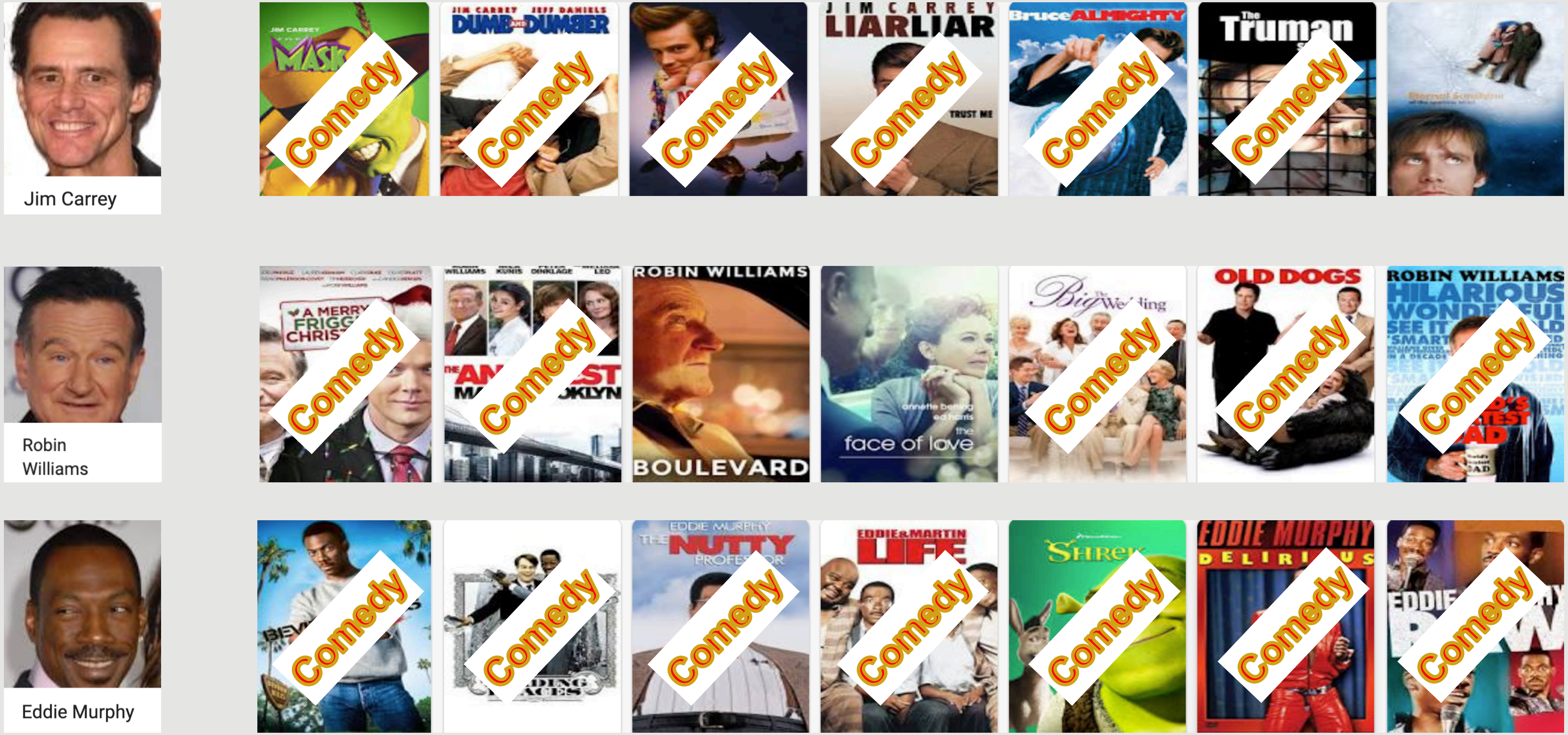

Jim Carrey


Robin Williams

There is no “funny” attribute in the data

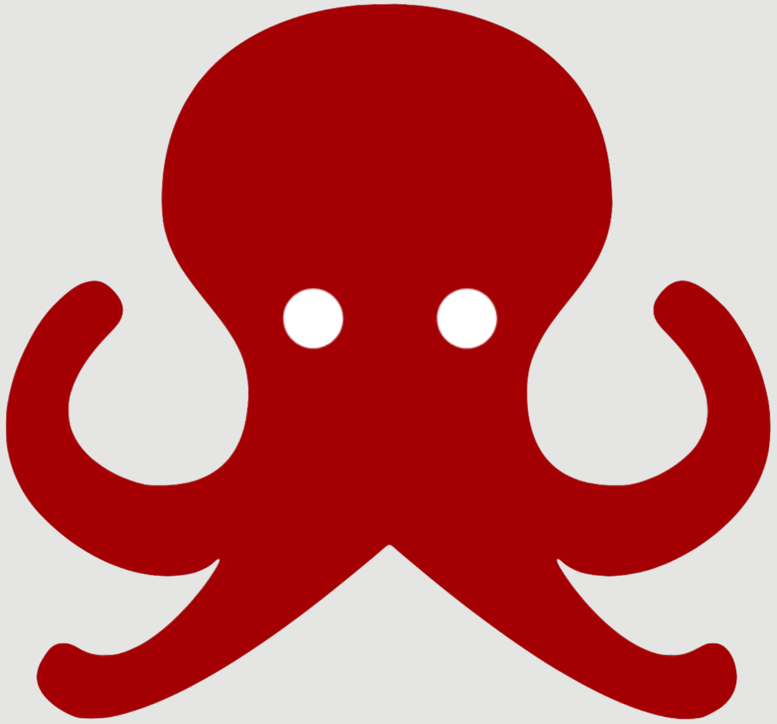

Eddie Murphy

Discovering semantic similarity



SQuID

Semantic Similarity-aware Query Intent Discovery



SQuID Outline

Modeling Semantic Context

Query Intent Discovery

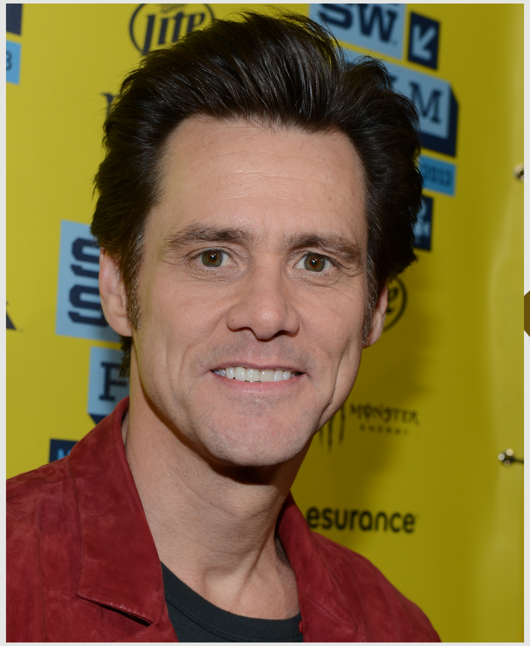
Real-time Performance

Evaluation

Semantic context: basic

- **Directly affiliated with an entity.**

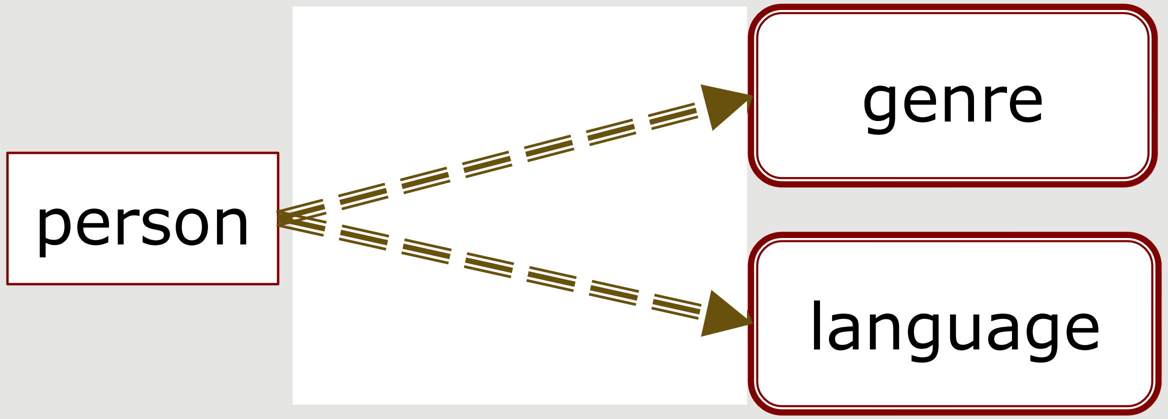
person → birth year, gender, height, age



→ 1962, Male, 6′ 2″, 57

Semantic context: derived

- **Aggregate over a basic property of an associated entity.**
  - number of comedy movies an actor appeared in.



Filters

- Encode semantic context.

```
SELECT
        person
FROM
        people
WHERE
        color = orange
```

Intended or co-incidental?

- Male
- Born in North America
- Appeared in 80+ Hollywood movies
- Appeared in 40+ comedy movies
- Appeared in 20+ drama movies
- Height above 5 feet
- Born after 1940
- …

Abduction

- Most likely **explanation** of an **observation**.
- Most likely **query** given the **examples**.

Maximum likelihood estimation is abduction!

Problem definition

Query intent discovery: given a *Database* and *Example*, find *Query* such that:

$$Example \subseteq Query(Database)$$

$$Query = \arg\max_q P(q|Example)$$

Probabilistic abduction model

semantic context posterior

query prior

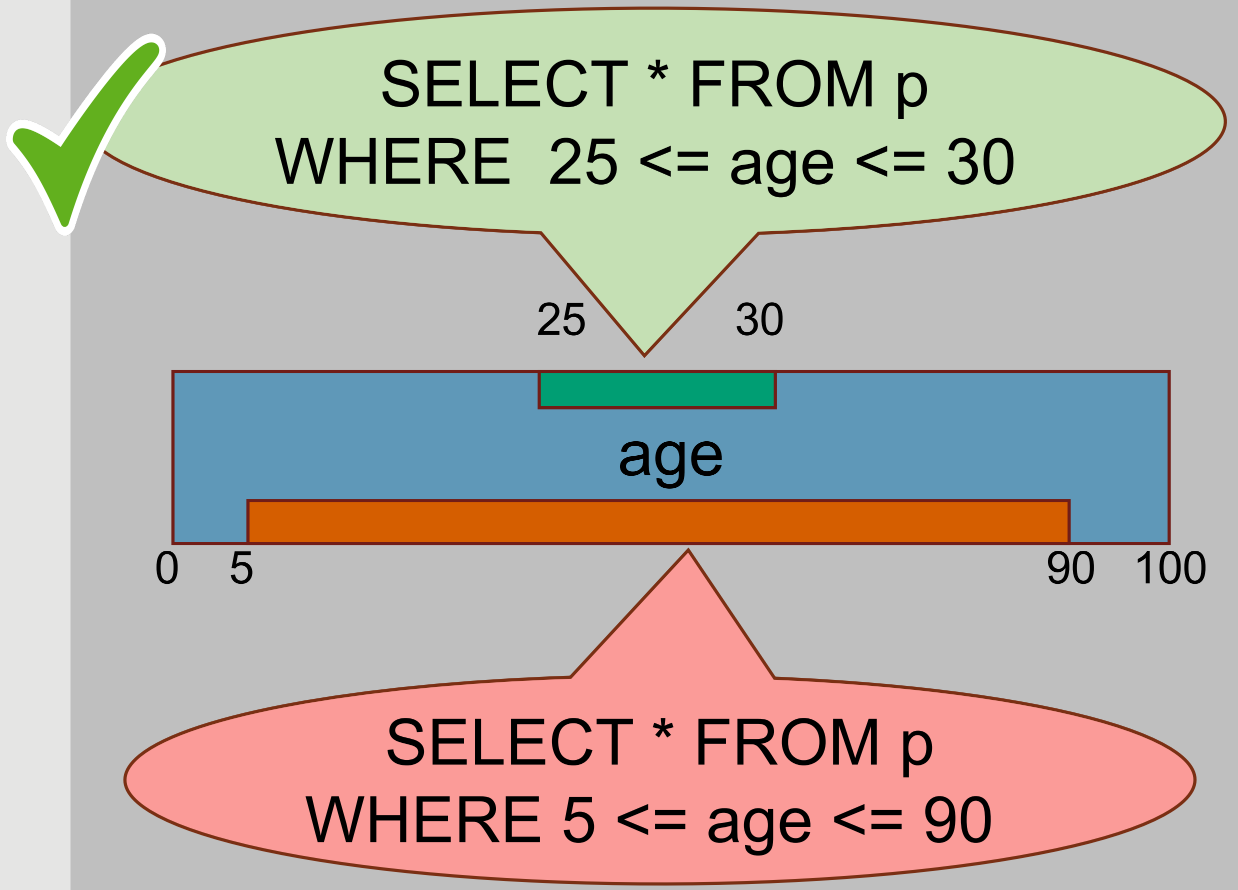
$$P(Query|Example) = \frac{P(Context|Query)P(Query)}{P(Context)}$$

query posterior

semantic context prior

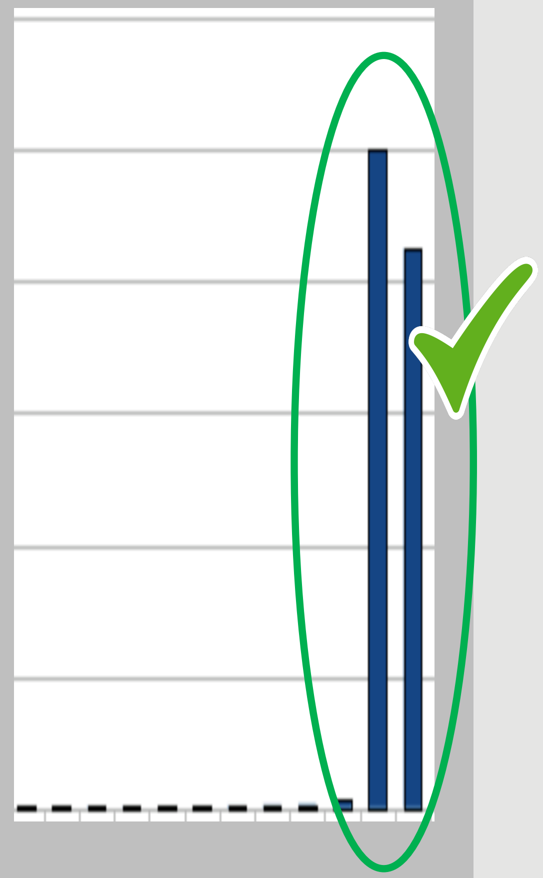
$$\frac{P(Context|Query)\boxed{P(Query)}}{P(Context)}$$

Domain selectivity



Association strength

Outlier



$$\frac{P(Context|Query)P(Query)}{\boxed{P(Context)}}$$

- **Data selectivity**

USA: 80%

	country
…	USA
…	USA
…	USA
…	USA
…	USA
…	CAN
…	USA
…	CAN
…	USA
…	USA

CAN: 20%

$$\frac{\boxed{P(Context|Query)}P(Query)}{P(Context)}$$

Context

Predicate in Query

P(USA| country = USA) = 1

	country
…	USA
…	USA
…	USA
…	USA
…	USA
…	CAN
…	USA
…	CAN
…	USA
…	USA

$$\frac{\boxed{P(Context|Query)}P(Query)}{P(Context)}$$

Context

Predicate in Query

P(USA| country = USA) = 1

P(USA | No Filter) = 0.8

	country
…	USA
…	USA
…	USA
…	USA
…	USA
…	CAN
…	USA
…	CAN
…	USA
…	USA

	country
…	USA

$$\frac{\boxed{P(Context|Query)}P(Query)}{P(Context)}$$

Context

Predicate in Query

P(USA| country = USA)
= 1

P(USA | No Filter)
= 0.8 * 0.8
= 0.64

	country
…	USA
…	USA
…	USA
…	USA
…	USA
…	CAN
…	USA
…	CAN
…	USA
…	USA

	country
…	USA
…	USA

$$\frac{\boxed{P(Context|Query)}P(Query)}{P(Context)}$$

Context

Predicate in Query

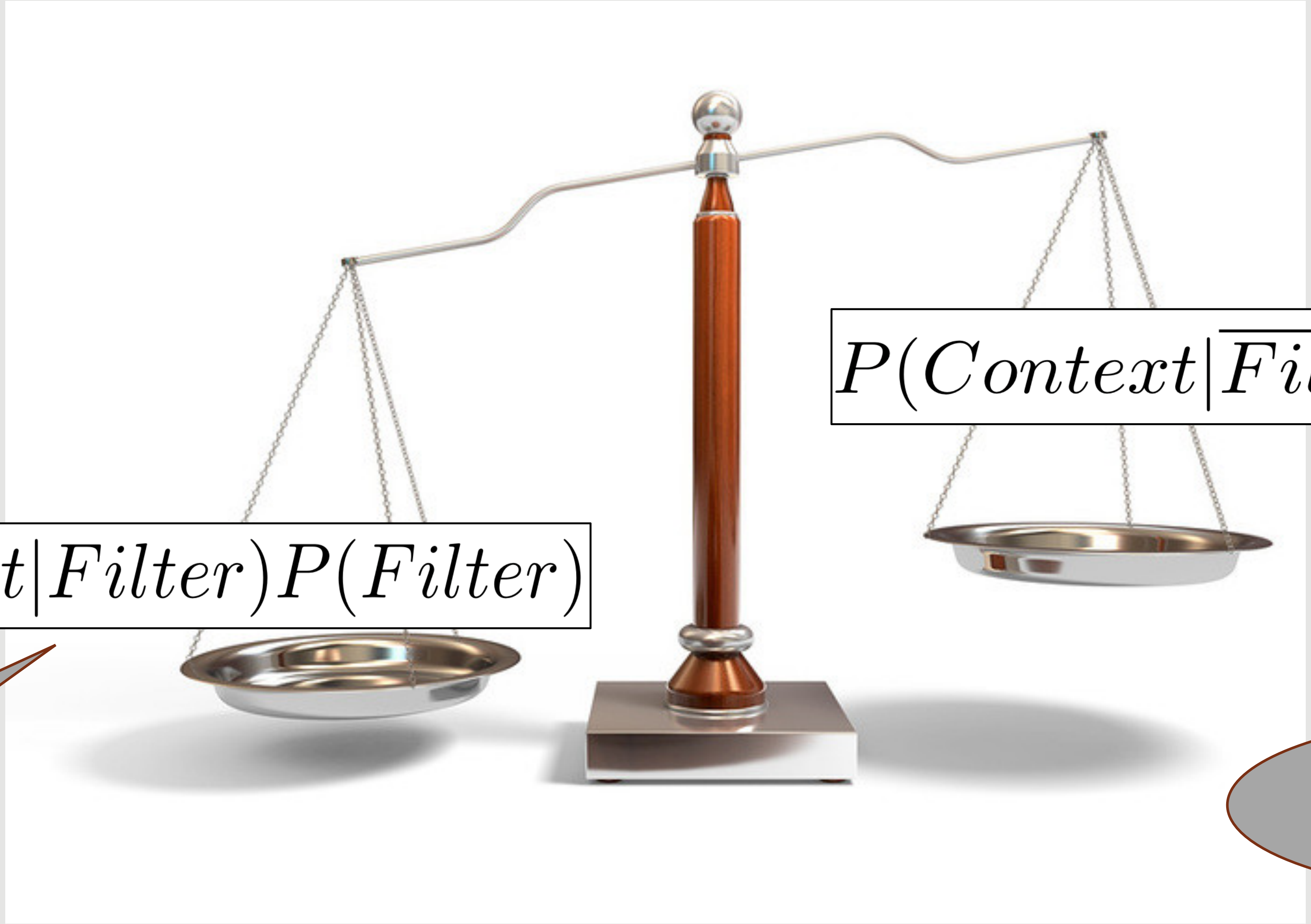
P(USA| country = USA)
= 1

P(USA | No Filter)
= 0.8 * 0.8 * 0.8
= 0.51

	country
…	USA
…	USA
…	USA
…	USA
…	USA
…	CAN
…	USA
…	CAN
…	USA
…	USA

	country
…	USA
…	USA
…	USA

SQuID algorithm: to pick or drop filters?



$$P(Context|\overline{Filter})P(\overline{Filter})$$

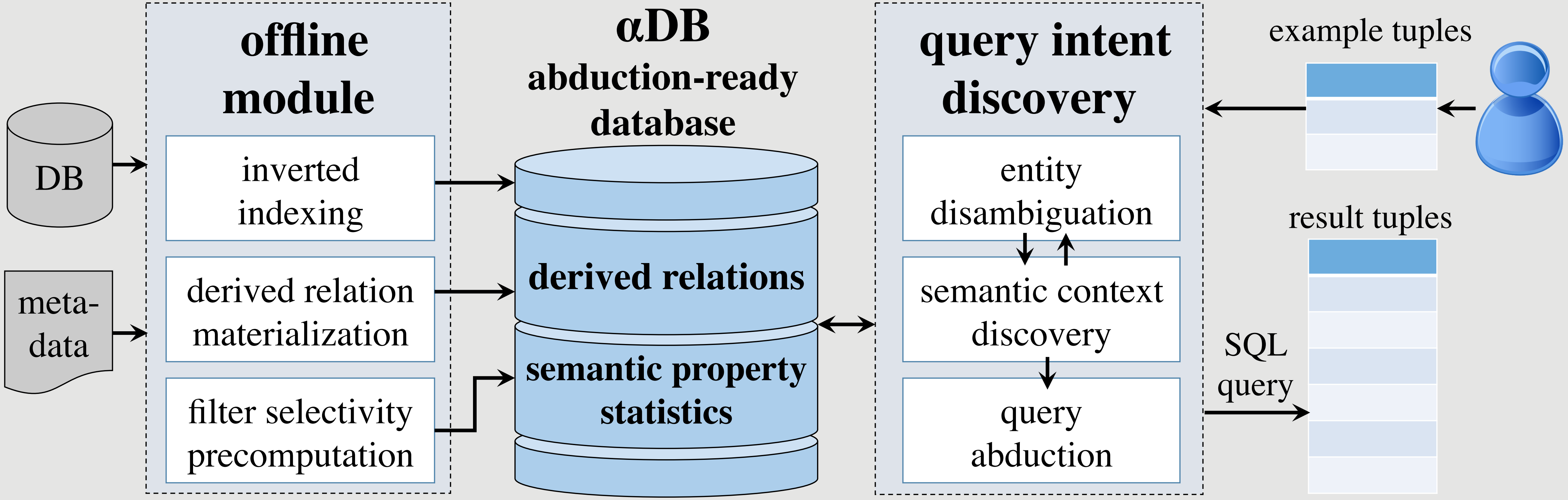
Without filter

$$P(Context|Filter)P(Filter)$$

With filter

Real-time performance

Abduction-ready database



Evaluation

1. **How efficient is SQuID for large datasets and many examples?**
2. **Does SQuID infer the right query?**
3. **Can alternative techniques be effective in intent discovery?**
   - Query Reverse Engineering (TALOS, 2014)
   - Positive and Unlabeled Learning (Elkan et al., 2008)

- **Query run-time comparison**
- **Case studies**

Datasets

633 MB

15 relations
- `person`: 6M rows
- `movies`: 1M rows
- `castinfo`: 14M rows

IMDb

16 benchmark queries

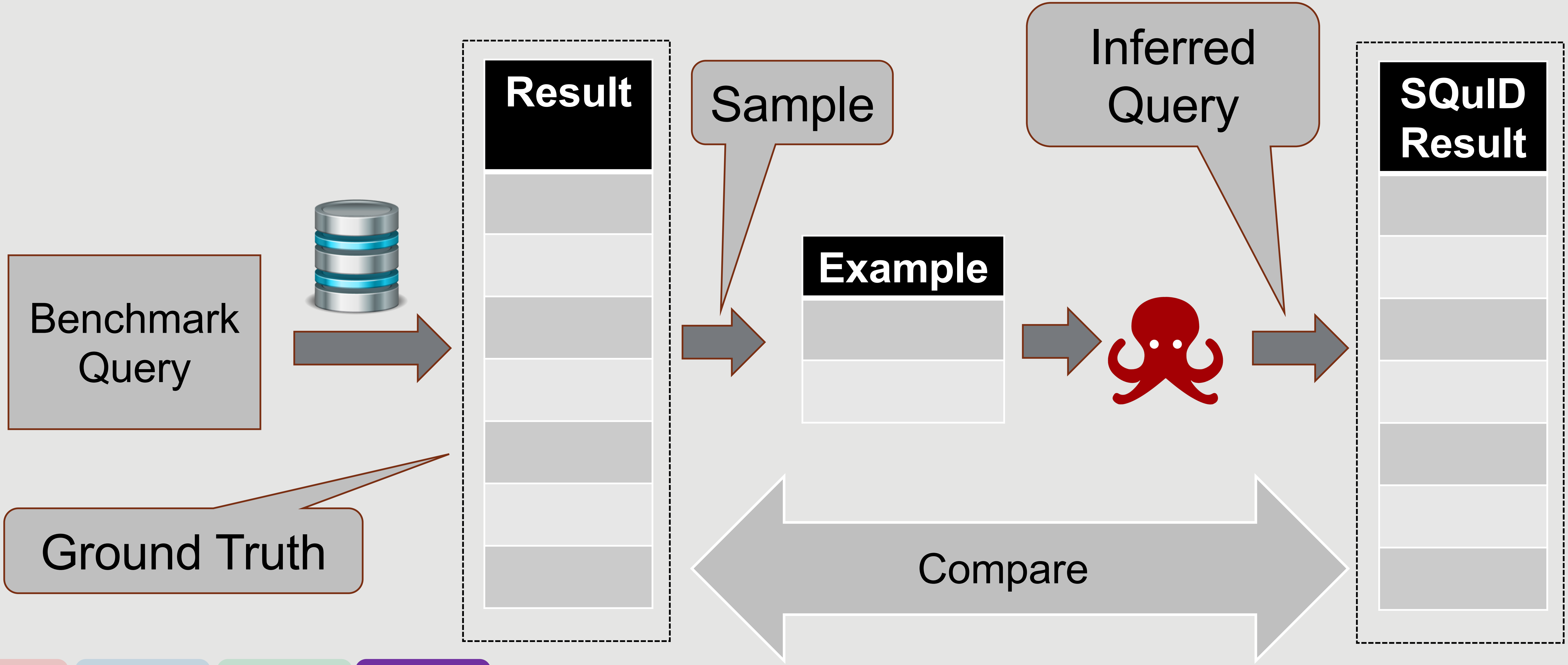
dblp computer science bibliography

5 benchmark queries

UCI Machine Learning Repository

20 benchmark queries

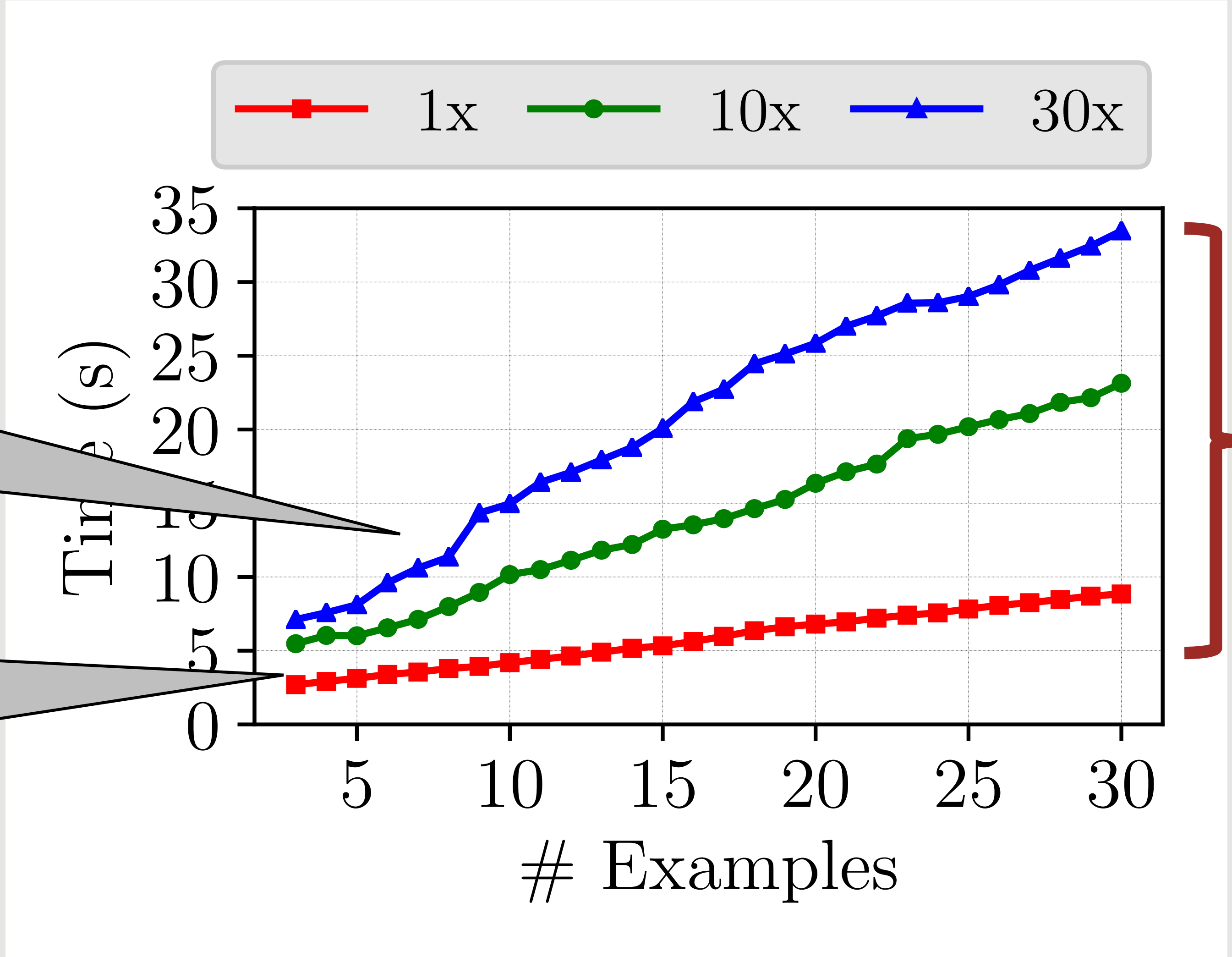
Experiment settings



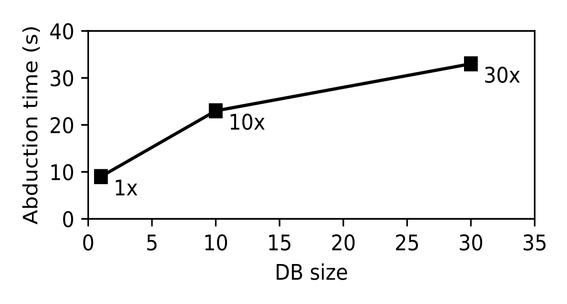
How does SQuID perform with large datasets or many examples?

Abduction time is practical

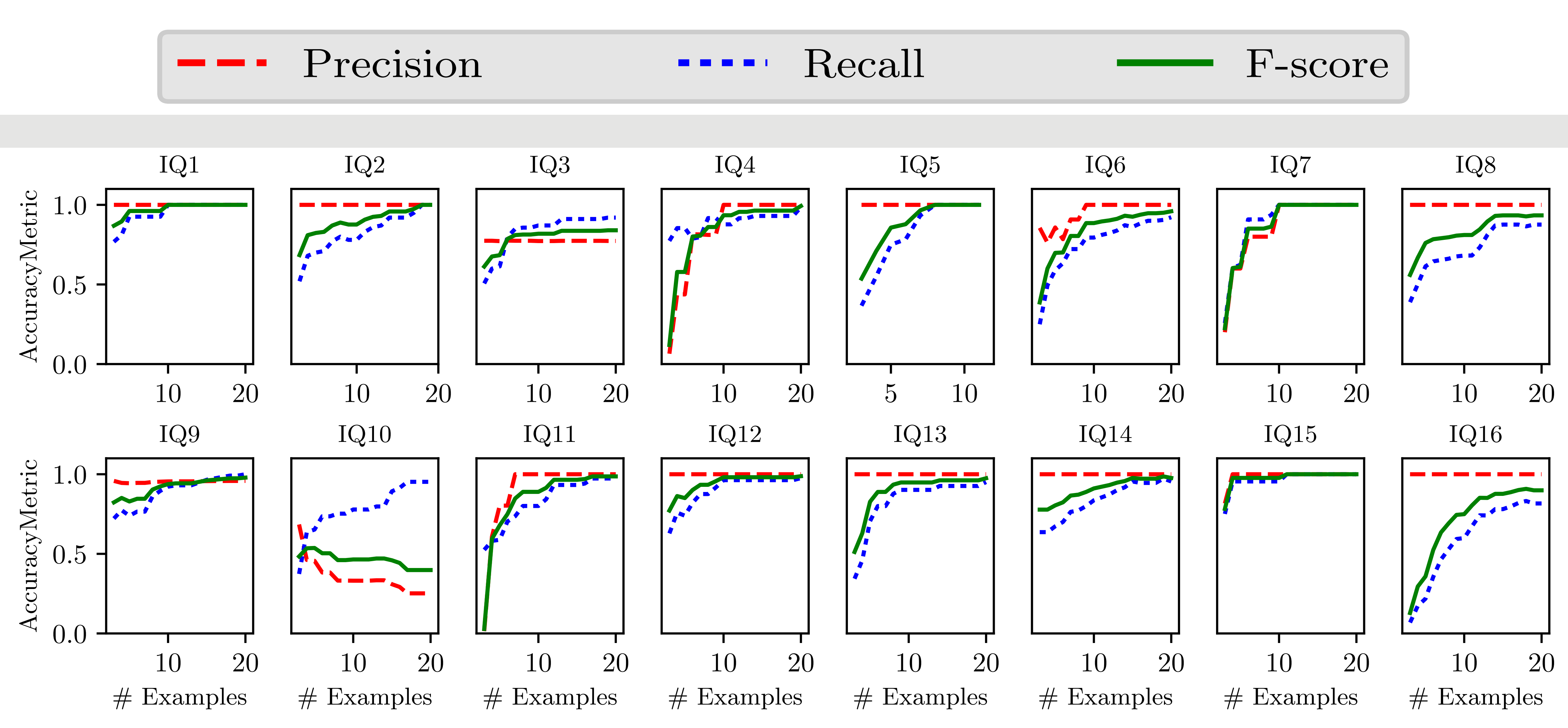
Linear in example size



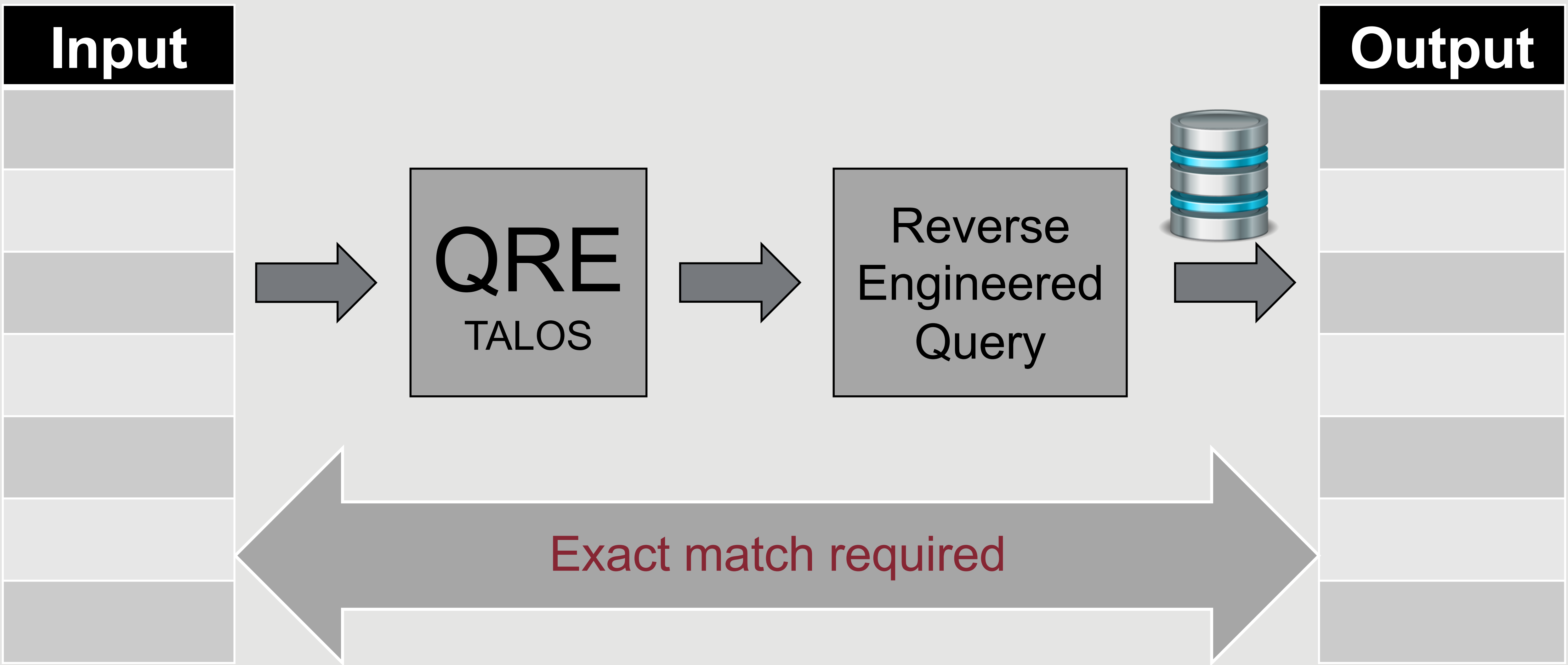
Logarithmic in DB size



SQuID works with few examples



Query reverse engineering (QRE)

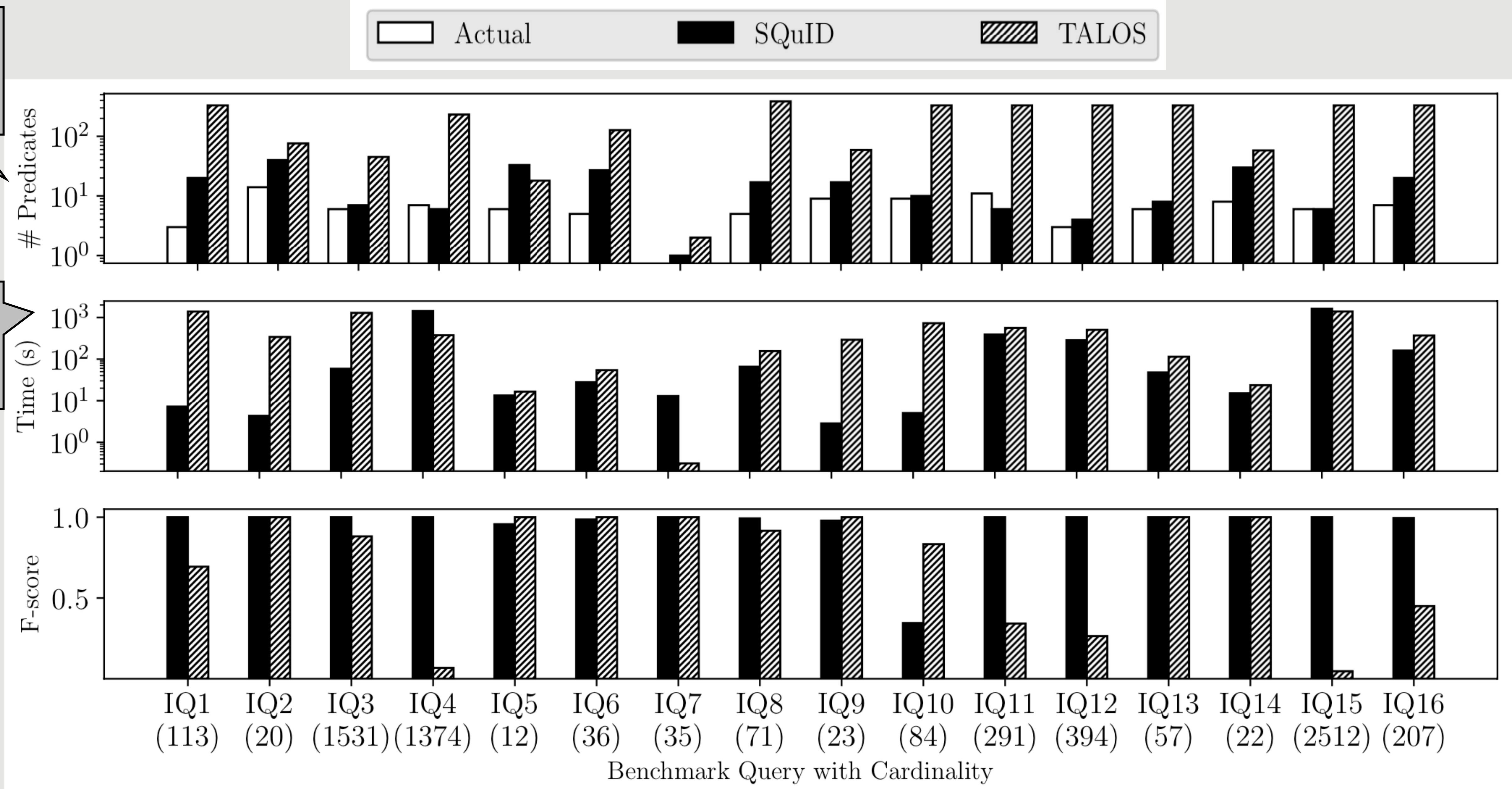


SQuID outperforms QRE

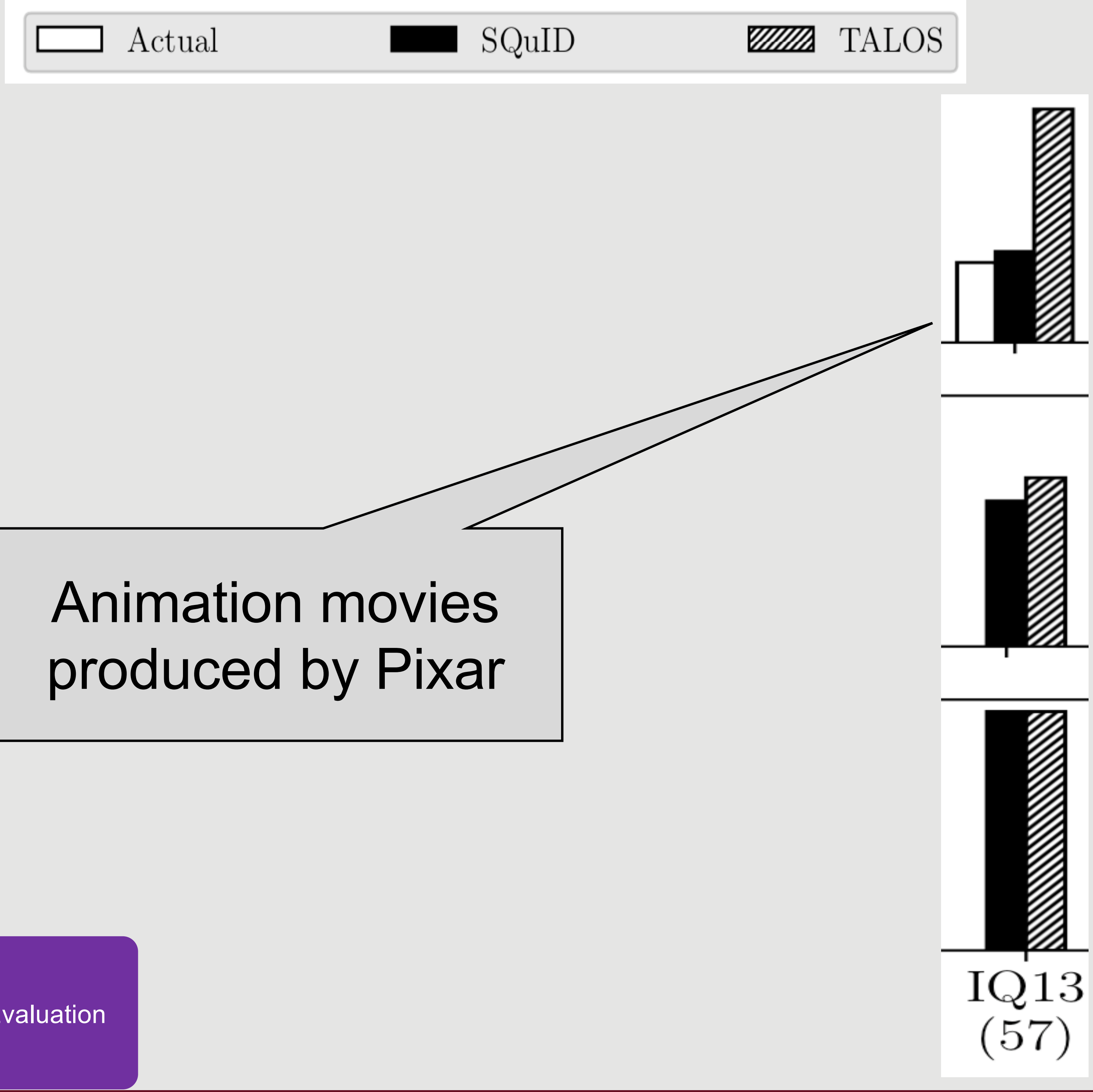
UMass Amherst | College of Information & Computer Sciences

Log scale

Log scale



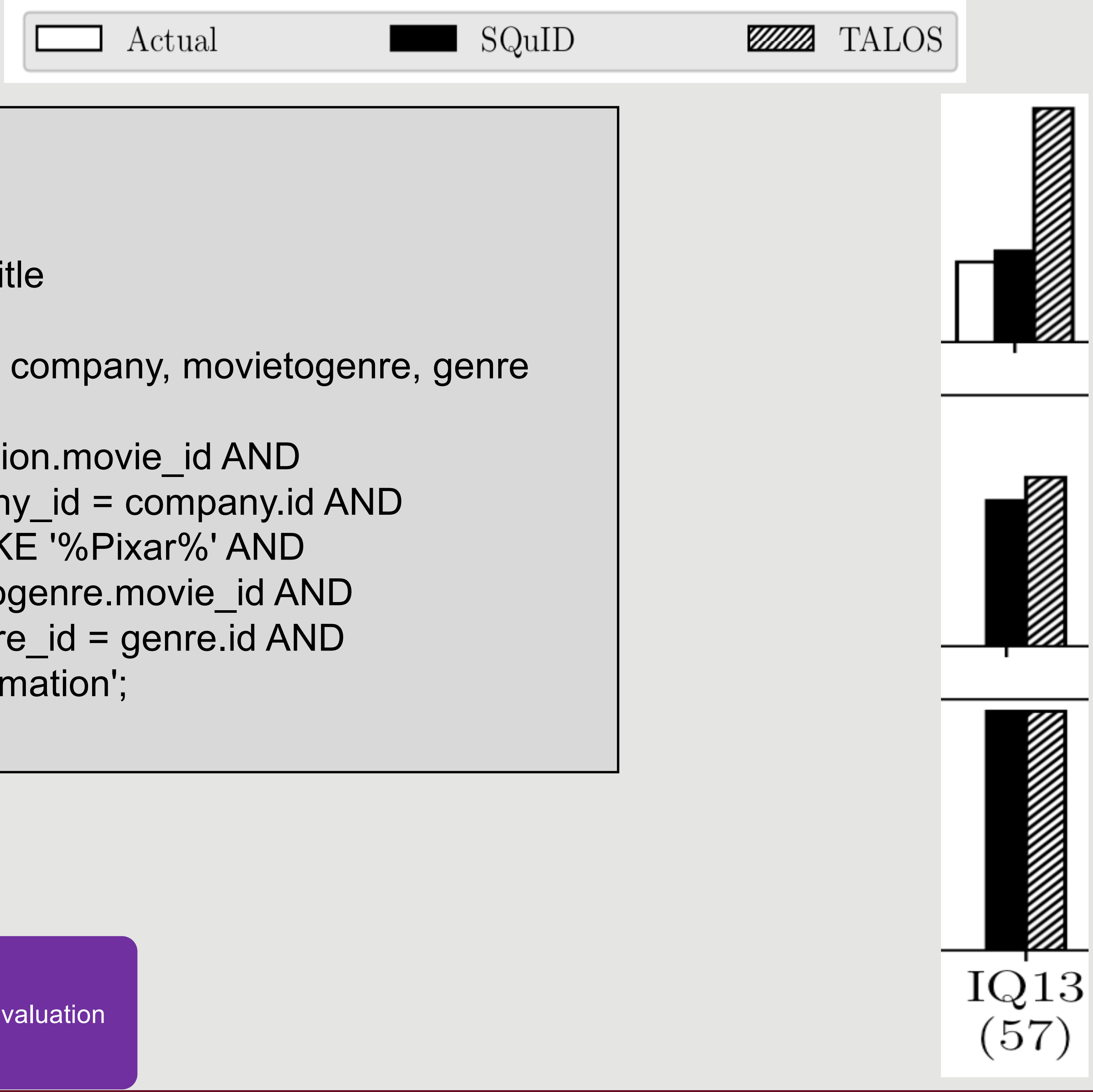
SQuID outperforms QRE



Animation movies produced by Pixar

SQuID outperforms QRE

UMass Amherst | College of Information & Computer Sciences



Original Query

```
SELECT
        DISTINCT movie.title
FROM
        movie, production, company, movietogenre, genre
WHERE
        movie.id = production.movie_id AND
        production.company_id = company.id AND
        company.name LIKE '%Pixar%' AND
        movie.id = movietogenre.movie_id AND
        movietogenre.genre_id = genre.id AND
        genre.name = 'Animation';
```

SQuID outperforms QRE

Original Query

SELECT
DISTINC
FROM
movie, p
WHERE
movie.id
producti
company
movie.id
movietog
genre.na

SQuID Query

SELECT
DISTINCT movie.title
FROM
movie, production, company, movietogenre, genre
WHERE
movie.production_year >= 1984 AND
movie.production_year <= 2021 AND
movie.country = USA AND
genre.name = 'Animation' AND
company.name = 'Pixar' AND
movie.id = movietogenre.movie_id AND
genre.id = movietogenre.genre_id AND
movie.id = movietoproduction.movie_id AND
company.id = movietoproduction.company_id

SQuID outperforms QRE

UMassAmherst | College of Information & Computer Sciences

Actual | SQuID | TALOS

Original Query

SELECT
DISTINC
FROM
movie, p
WHERE
movie.id
producti
company
movie.id
movietog
genre.na

SQuID Query

SELECT
DISTINCT
FROM
movie, pr
WHERE
movie.pro
movie.pro
movie.co
genre.nar
company.
movie.id =
genre.id =
movie.id =
company.

TALOS Query

SELECT
distinct title
FROM
movie
WHERE

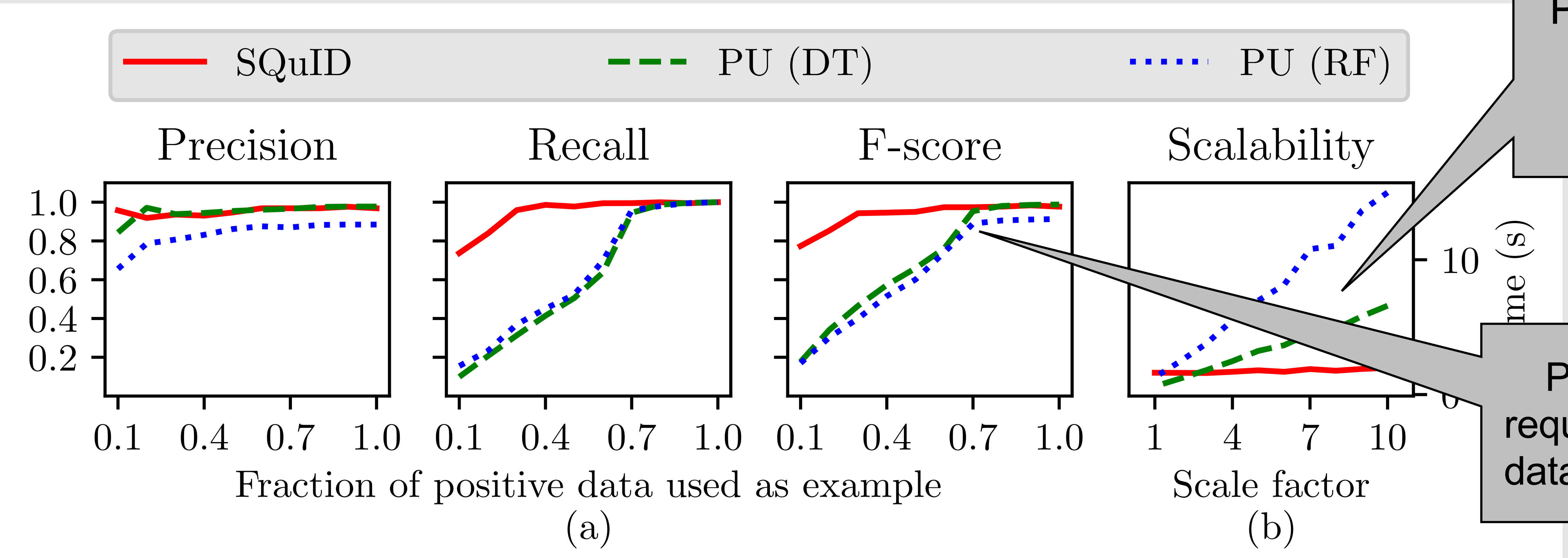
```
((movie.title='Inside Out' OR movie.title='Lifted' OR movie.title='Lou' OR movie.title='For the Birds' OR movie.title='Brave' OR movie.title='Coco')) OR ((movie.title<>'Inside Out' AND movie.title<>'Lifted' AND movie.title<>'Lou' AND movie.title<>'For the Birds' AND movie.title<>'Brave' AND
movie.title<>'Coco') AND (movie.title='A Bug's Life' OR movie.title='André and Wally B.' OR movie.title='Boundin'' OR movie.title='Burn-E' OR movie.title='Cars 2' OR movie.title='Cars 3')) OR ((movie.title<>'Inside Out' AND movie.title<>'Lifted' AND movie.title<>'Lou' AND movie.title<>'For the
Birds' AND movie.title<>'Brave' AND movie.title<>'Coco') AND (movie.title<>'A Bug's Life' AND movie.title<>'André and Wally B.' AND movie.title<>'Boundin'' AND movie.title<>'Burn-E' AND movie.title<>'Cars 2' AND movie.title<>'Cars 3') AND (movie.title='Dante"s Lunch: A Short Tail' OR
movie.title='Finding Dory' OR movie.title='Finding Nemo' OR movie.title='Geri"s Game' OR movie.title='Hawaiian Vacation' OR movie.title='It"s Tough to Be a Bug')) OR ((movie.title<>'Inside Out' AND movie.title<>'Lifted' AND movie.title<>'Lou' AND movie.title<>'For the Birds' AND
movie.title<>'Brave' AND movie.title<>'Coco') AND (movie.title<>'A Bug's Life' AND movie.title<>'André and Wally B.' AND movie.title<>'Boundin'' AND movie.title<>'Burn-E' AND movie.title<>'Cars 2' AND movie.title<>'Cars 3') AND (movie.title<>'Dante"s Lunch: A Short Tail' AND
movie.title<>'Finding Dory' AND movie.title<>'Finding Nemo' AND movie.title<>'Geri"s Game' AND movie.title<>'Hawaiian Vacation' AND movie.title<>'It"s Tough to Be a Bug') AND (movie.title='Knick Knack' OR movie.title='Luxo Jr.' OR movie.title='Luxo Jr. in "Surprise" and "Light & Heavy"'
OR movie.title='Luxo Jr. in "Up and Down"' OR movie.title='Monsters University' OR movie.title='Monsters, Inc.')) OR ((movie.title<>'Inside Out' AND movie.title<>'Lifted' AND movie.title<>'Lou' AND movie.title<>'For the Birds' AND movie.title<>'Brave' AND movie.title<>'Coco') AND
(movie.title<>'A Bug's Life' AND movie.title<>'André and Wally B.' AND movie.title<>'Boundin'' AND movie.title<>'Burn-E' AND movie.title<>'Cars 2' AND movie.title<>'Cars 3') AND (movie.title<>'Dante"s Lunch: A Short Tail' AND movie.title<>'Finding Dory' AND movie.title<>'Finding Nemo'
AND movie.title<>'Geri"s Game' AND movie.title<>'Hawaiian Vacation' AND movie.title<>'It"s Tough to Be a Bug') AND (movie.title<>'Knick Knack' AND movie.title<>'Luxo Jr.' AND movie.title<>'Luxo Jr. in "Surprise" and "Light & Heavy"' AND movie.title<>'Luxo Jr. in "Up and Down"' AND
movie.title<>'Monsters University' AND movie.title<>'Monsters, Inc.') AND (movie.title='Party Central' OR movie.title='Partysaurus Rex' OR movie.title='Red"s Dream' OR movie.title='Riley"s First Date?' OR movie.title='Sanjay"s Super Team' OR movie.title='The Good Dinosaur')) OR
((movie.title<>'Inside Out' AND movie.title<>'Lifted' AND movie.title<>'Lou' AND movie.title<>'For the Birds' AND movie.title<>'Brave' AND movie.title<>'Coco') AND (movie.title<>'A Bug's Life' AND movie.title<>'André and Wally B.' AND movie.title<>'Boundin'' AND movie.title<>'Burn-E' AND
movie.title<>'Cars 2' AND movie.title<>'Cars 3') AND (movie.title<>'Dante"s Lunch: A Short Tail' AND movie.title<>'Finding Dory' AND movie.title<>'Finding Nemo' AND movie.title<>'Geri"s Game' AND movie.title<>'Hawaiian Vacation' AND movie.title<>'It"s Tough to Be a Bug') AND
(movie.title<>'Knick Knack' AND movie.title<>'Luxo Jr.' AND movie.title<>'Luxo Jr. in "Surprise" and "Light & Heavy"' AND movie.title<>'Luxo Jr. in "Up and Down"' AND movie.title<>'Monsters University' AND movie.title<>'Monsters, Inc.') AND (movie.title<>'Party Central' AND
movie.title<>'Partysaurus Rex' AND movie.title<>'Red"s Dream' AND movie.title<>'Riley"s First Date?' AND movie.title<>'Sanjay"s Super Team' AND movie.title<>'The Good Dinosaur') AND (movie.title='The Incredibles' OR movie.title='The Incredibles 2' OR movie.title='Tin Toy' OR
movie.title='Tokyo Mater' OR movie.title='Toy Story' OR movie.title='Toy Story 2')) OR ((movie.title<>'Inside Out' AND movie.title<>'Lifted' AND movie.title<>'Lou' AND movie.title<>'For the Birds' AND movie.title<>'Brave' AND movie.title<>'Coco') AND (movie.title<>'A Bug's Life' AND
movie.title<>'André and Wally B.' AND movie.title<>'Boundin'' AND movie.title<>'Burn-E' AND movie.title<>'Cars 2' AND movie.title<>'Cars 3') AND (movie.title<>'Dante"s Lunch: A Short Tail' AND movie.title<>'Finding Dory' AND movie.title<>'Finding Nemo' AND movie.title<>'Geri"s Game'
AND movie.title<>'Hawaiian Vacation' AND movie.title<>'It"s Tough to Be a Bug') AND (movie.title<>'Knick Knack' AND movie.title<>'Luxo Jr.' AND movie.title<>'Luxo Jr. in "Surprise" and "Light & Heavy"' AND movie.title<>'Luxo Jr. in "Up and Down"' AND movie.title<>'Monsters University'
AND movie.title<>'Monsters, Inc.') AND (movie.title<>'Party Central' AND movie.title<>'Partysaurus Rex' AND movie.title<>'Red"s Dream' AND movie.title<>'Riley"s First Date?' AND movie.title<>'Sanjay"s Super Team' AND movie.title<>'The Good Dinosaur') AND (movie.title<>'The
Incredibles' AND movie.title<>'The Incredibles 2' AND movie.title<>'Tin Toy' AND movie.title<>'Tokyo Mater' AND movie.title<>'Toy Story' AND movie.title<>'Toy Story 2') AND (movie.title='Toy Story 3' OR movie.title='Toy Story 4' OR movie.title='Untitled Pixar Animation Project' OR
movie.title='WALL·E' OR movie.title='Your Friend the Rat' OR movie.title='Cars') AND movie.production_year>2006) OR ((movie.title<>'Inside Out' AND movie.title<>'Lifted' AND movie.title<>'Lou' AND movie.title<>'For the Birds' AND movie.title<>'Brave' AND movie.title<>'Coco') AND
(movie.title<>'A Bug's Life' AND movie.title<>'André and Wally B.' AND movie.title<>'Boundin'' AND movie.title<>'Burn-E' AND movie.title<>'Cars 2' AND movie.title<>'Cars 3') AND (movie.title<>'Dante"s Lunch: A Short Tail' AND movie.title<>'Finding Dory' AND movie.title<>'Finding Nemo'
AND movie.title<>'Geri"s Game' AND movie.title<>'Hawaiian Vacation' AND movie.title<>'It"s Tough to Be a Bug') AND (movie.title<>'Knick Knack' AND movie.title<>'Luxo Jr.' AND movie.title<>'Luxo Jr. in "Surprise" and "Light & Heavy"' AND movie.title<>'Luxo Jr. in "Up and Down"' AND
movie.title<>'Monsters University' AND movie.title<>'Monsters, Inc.') AND (movie.title<>'Party Central' AND movie.title<>'Partysaurus Rex' AND movie.title<>'Red"s Dream' AND movie.title<>'Riley"s First Date?' AND movie.title<>'Sanjay"s Super Team' AND movie.title<>'The Good Dinosaur')
AND (movie.title<>'The Incredibles' AND movie.title<>'The Incredibles 2' AND movie.title<>'Tin Toy' AND movie.title<>'Tokyo Mater' AND movie.title<>'Toy Story' AND movie.title<>'Toy Story 2') AND (movie.title<>'Toy Story 3' AND movie.title<>'Toy Story 4' AND movie.title<>'Untitled Pixar
Animation Project' AND movie.title<>'WALL·E' AND movie.title<>'Your Friend the Rat' AND movie.title<>'Cars') AND (movie.title='La Luna' OR movie.title='Ratatouille' OR movie.title='The Blue Umbrella' OR movie.title='One Man Band' OR movie.title='Presto' OR movie.title='Small Fry') AND
movie.production_year<=2008 AND movie.production_year<=1965) OR ((movie.title<>'Inside Out' AND movie.title<>'Lifted' AND movie.title<>'Lou' AND movie.title<>'For the Birds' AND movie.title<>'Brave' AND movie.title<>'Coco') AND (movie.title<>'A Bug's Life' AND movie.title<>'André
and Wally B.' AND movie.title<>'Boundin'' AND movie.title<>'Burn-E' AND movie.title<>'Cars 2' AND movie.title<>'Cars 3') AND (movie.title<>'Dante"s Lunch: A Short Tail' AND movie.title<>'Finding Dory' AND movie.title<>'Finding Nemo' AND movie.title<>'Geri"s Game' AND
movie.title<>'Hawaiian Vacation' AND movie.title<>'It"s Tough to Be a Bug') AND (movie.title<>'Knick Knack' AND movie.title<>'Luxo Jr.' AND movie.title<>'Luxo Jr. in "Surprise" and "Light & Heavy"' AND movie.title<>'Luxo Jr. in "Up and Down"' AND movie.title<>'Monsters University' AND
movie.title<>'Monsters, Inc.') AND (movie.title<>'Party Central' AND movie.title<>'Partysaurus Rex' AND movie.title<>'Red"s Dream' AND movie.title<>'Riley"s First Date?' AND movie.title<>'Sanjay"s Super Team' AND movie.title<>'The Good Dinosaur') AND (movie.title<>'The Incredibles'
AND movie.title<>'The Incredibles 2' AND movie.title<>'Tin Toy' AND movie.title<>'Tokyo Mater' AND movie.title<>'Toy Story' AND movie.title<>'Toy Story 2') AND (movie.title<>'Toy Story 3' AND movie.title<>'Toy Story 4' AND movie.title<>'Untitled Pixar Animation Project' AND
movie.title<>'WALL·E' AND movie.title<>'Your Friend the Rat' AND movie.title<>'Cars') AND (movie.title='La Luna' OR movie.title='Ratatouille' OR movie.title='The Blue Umbrella' OR movie.title='One Man Band' OR movie.title='Presto' OR movie.title='Small Fry') AND
movie.production_year<=2008 AND movie.production_year>1965) OR ((movie.title<>'Inside Out' AND movie.title<>'Lifted' AND movie.title<>'Lou' AND movie.title<>'For the Birds' AND movie.title<>'Brave' AND movie.title<>'Coco') AND (movie.title<>'A Bug's Life' AND movie.title<>'André and
Wally B.' AND movie.title<>'Boundin'' AND movie.title<>'Burn-E' AND movie.title<>'Cars 2' AND movie.title<>'Cars 3') AND (movie.title<>'Dante"s Lunch: A Short Tail' AND movie.title<>'Finding Dory' AND movie.title<>'Finding Nemo' AND movie.title<>'Geri"s Game' AND
movie.title<>'Hawaiian Vacation' AND movie.title<>'It"s Tough to Be a Bug') AND (movie.title<>'Knick Knack' AND movie.title<>'Luxo Jr.' AND movie.title<>'Luxo Jr. in "Surprise" and "Light & Heavy"' AND movie.title<>'Luxo Jr. in "Up and Down"' AND movie.title<>'Monsters University' AND
movie.title<>'Monsters, Inc.') AND (movie.title<>'Party Central' AND movie.title<>'Partysaurus Rex' AND movie.title<>'Red"s Dream' AND movie.title<>'Riley"s First Date?' AND movie.title<>'Sanjay"s Super Team' AND movie.title<>'The Good Dinosaur') AND (movie.title<>'The Incredibles'
AND movie.title<>'The Incredibles 2' AND movie.title<>'Tin Toy' AND movie.title<>'Tokyo Mater' AND movie.title<>'Toy Story' AND movie.title<>'Toy Story 2') AND (movie.title<>'Toy Story 3' AND movie.title<>'Toy Story 4' AND movie.title<>'Untitled Pixar Animation Project' AND
movie.title<>'WALL·E' AND movie.title<>'Your Friend the Rat' AND movie.title<>'Cars') AND (movie.title<>'La Luna' AND movie.title<>'Ratatouille' AND movie.title<>'The Blue Umbrella' AND movie.title<>'One Man Band' AND movie.title<>'Presto' AND movie.title<>'Small Fry') AND
(movie.title='Day & Night' OR movie.title='Partly Cloudy' OR movie.title='Piper' OR movie.title='Lava' OR movie.title='Up') AND movie.production_year<=1984)
```

Query Reverse Engineering overfits

IQ13
(57)

Modeling Semantic Context | Query Intent Discovery | Real-time Performance | **Evaluation**

SQuID outperforms machine learning

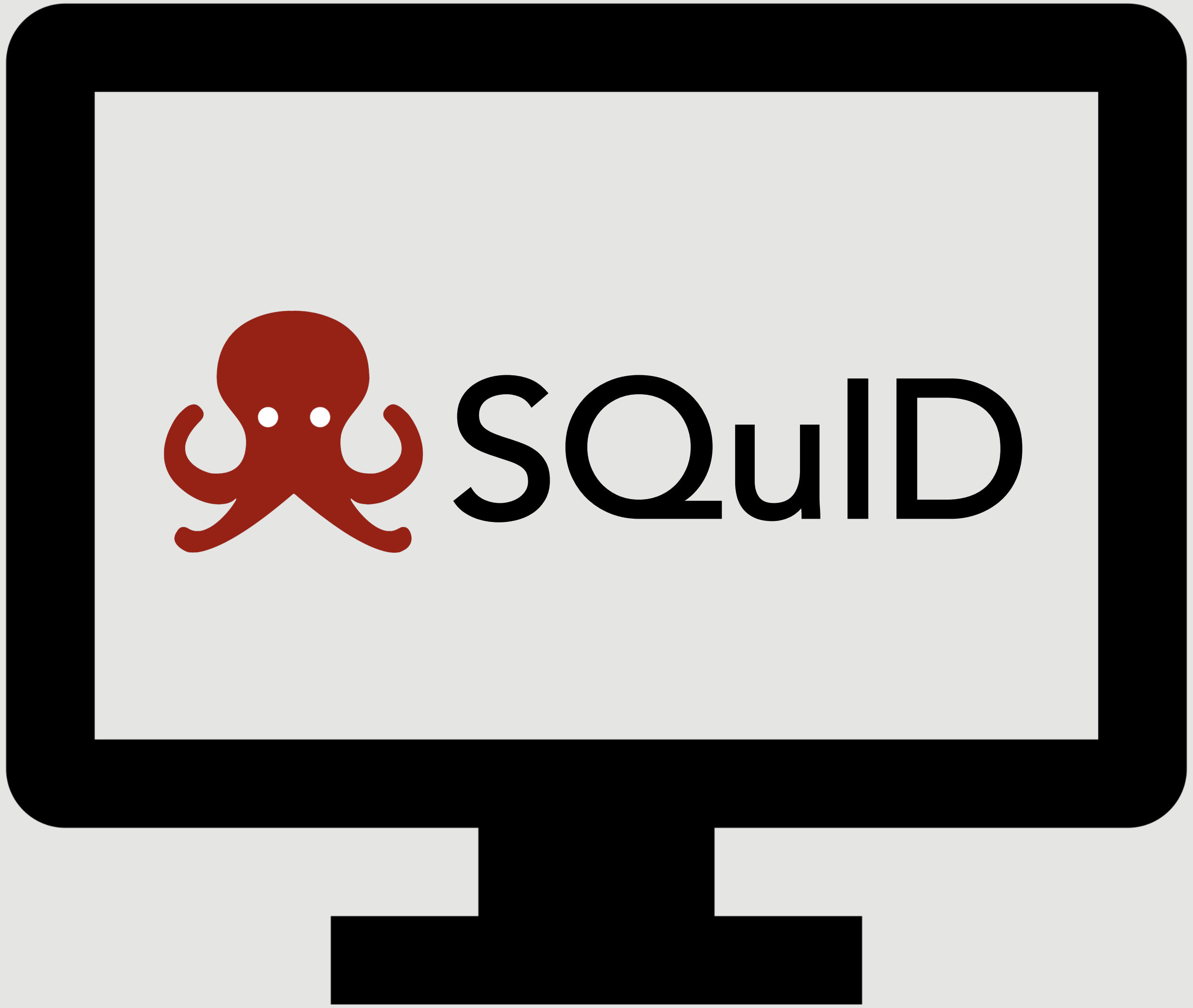


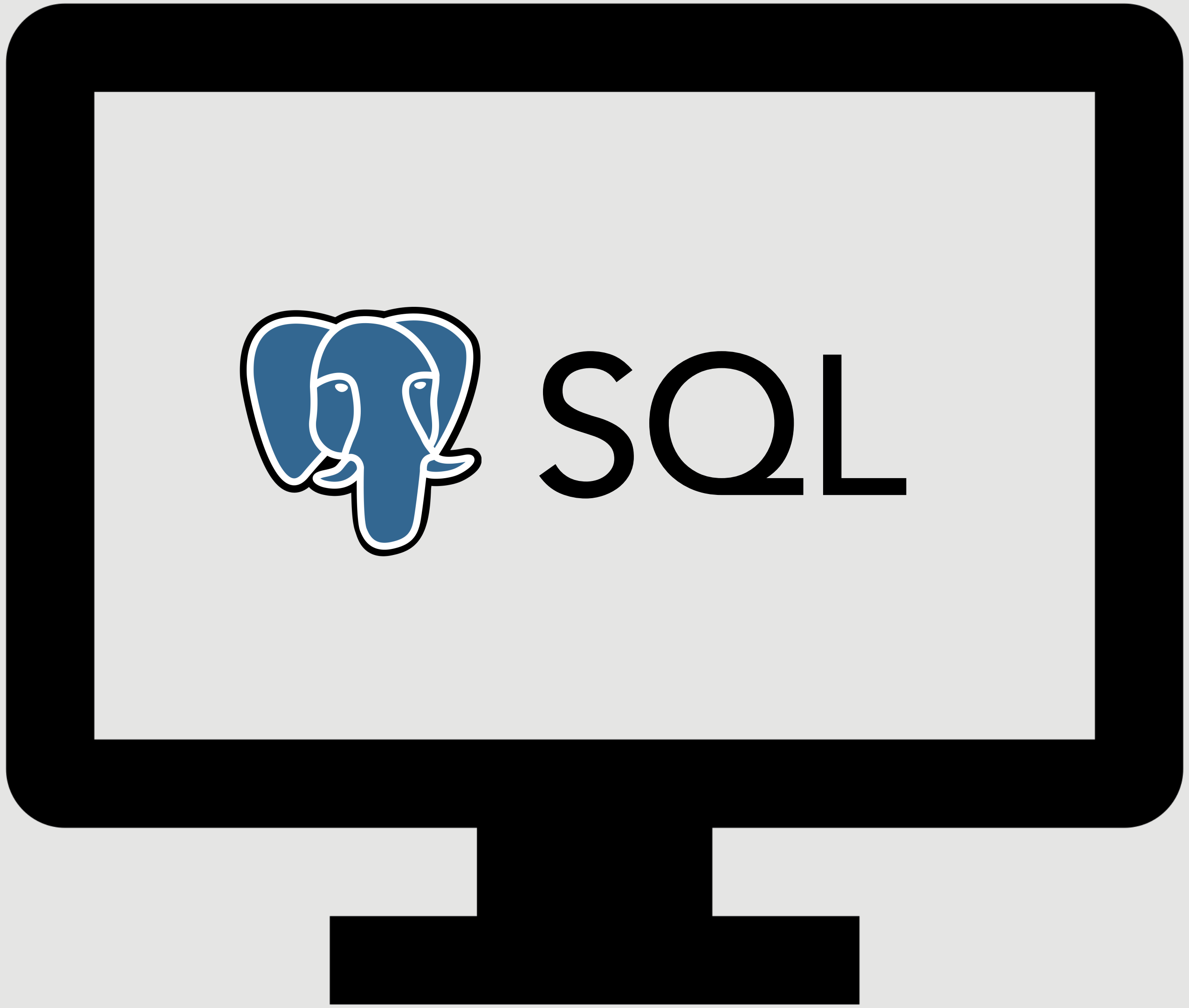
PU learning does not scale

PU learning requires >= 70% data as example

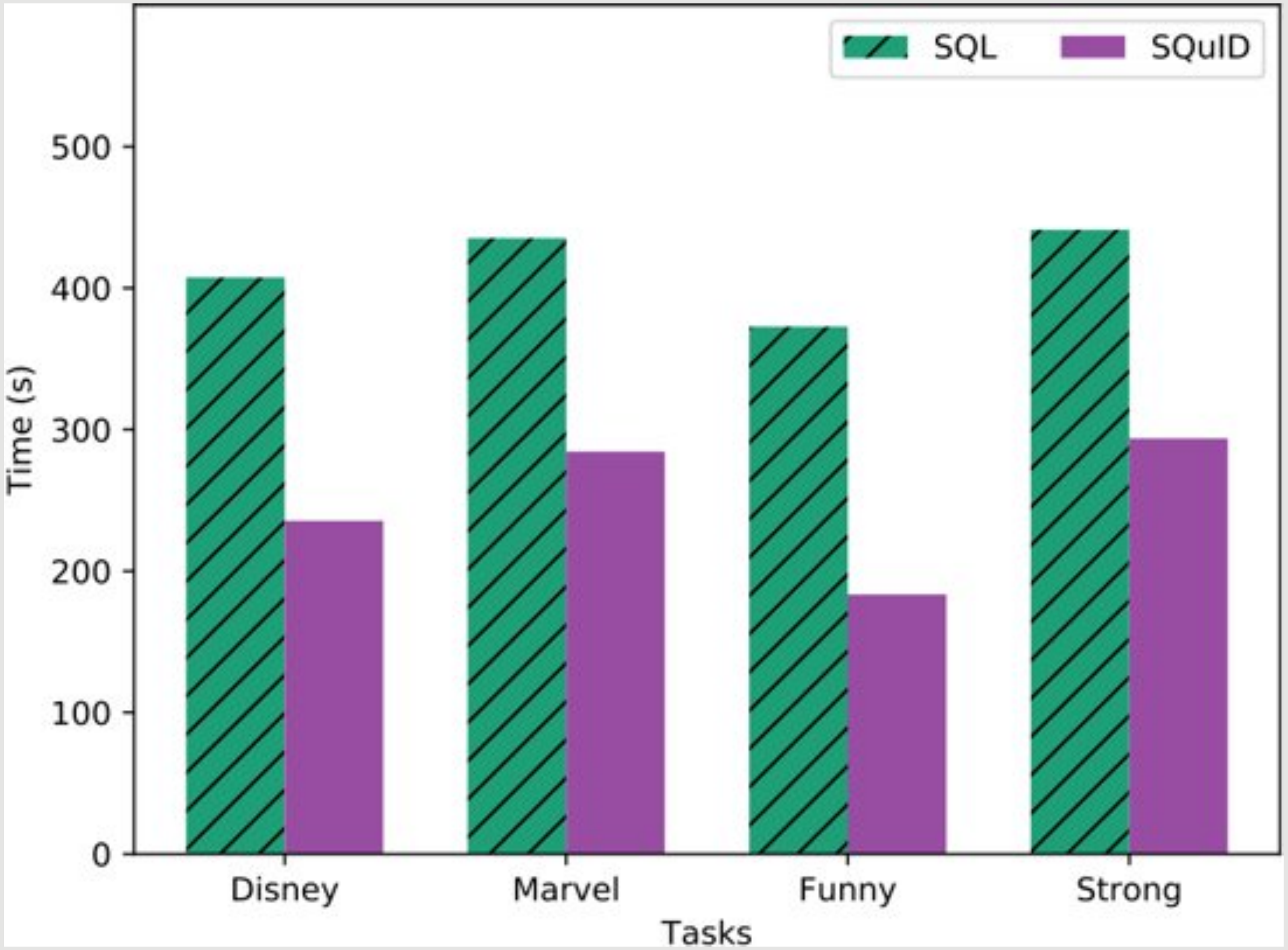
Generic machine learning cannot model RDBMS specific assumptions

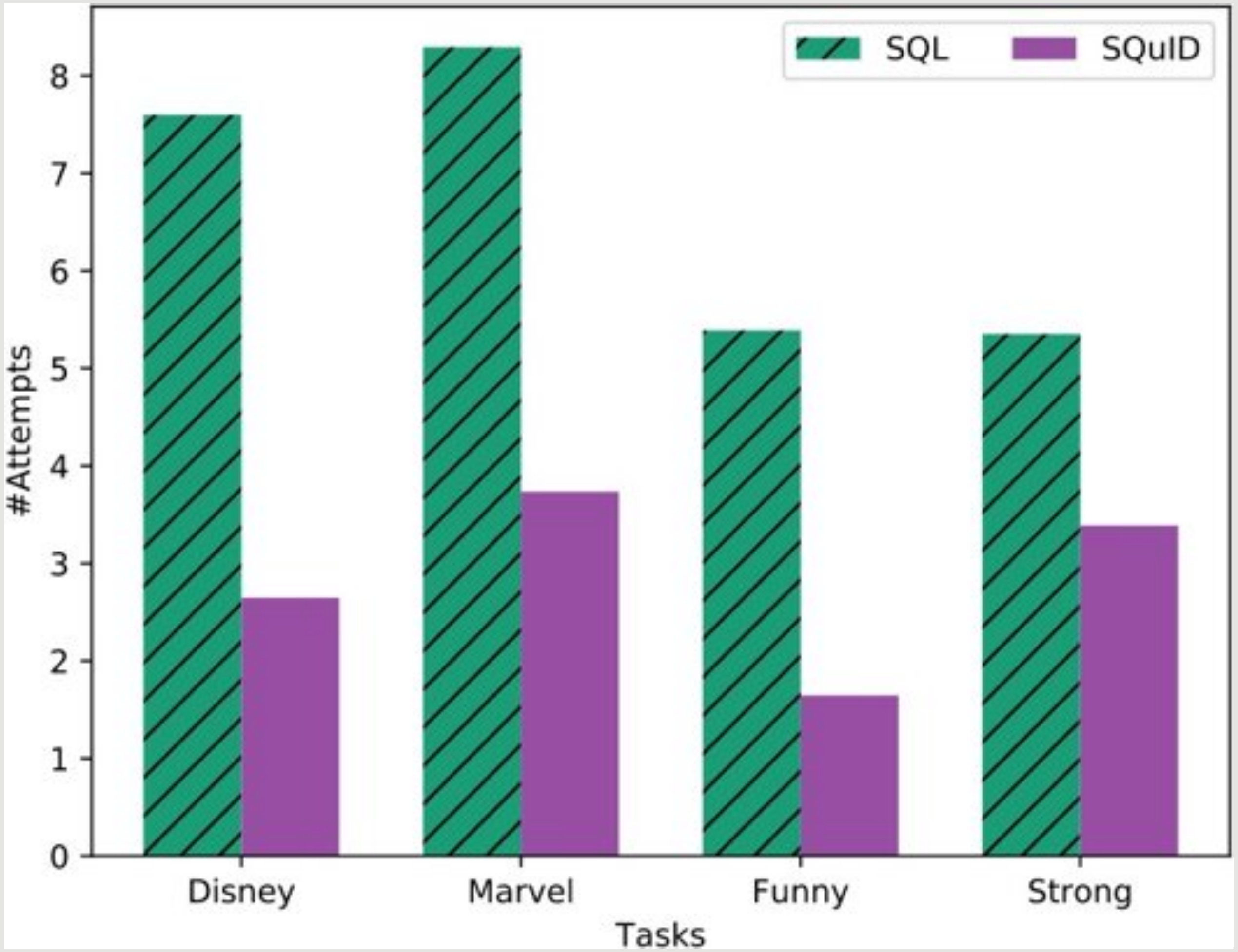
Comparative user studies: QBE vs SQL



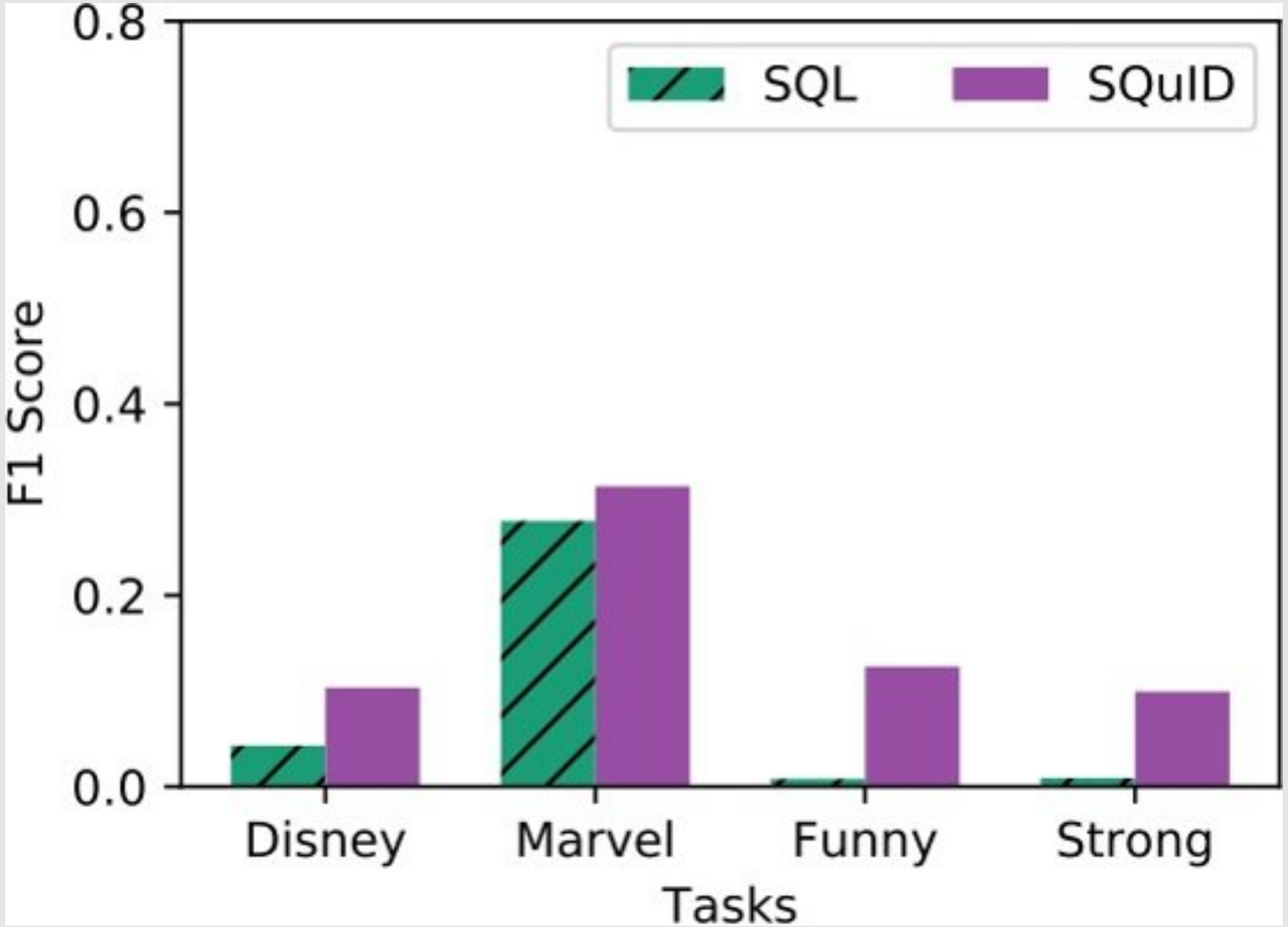


SQuID increased user efficiency

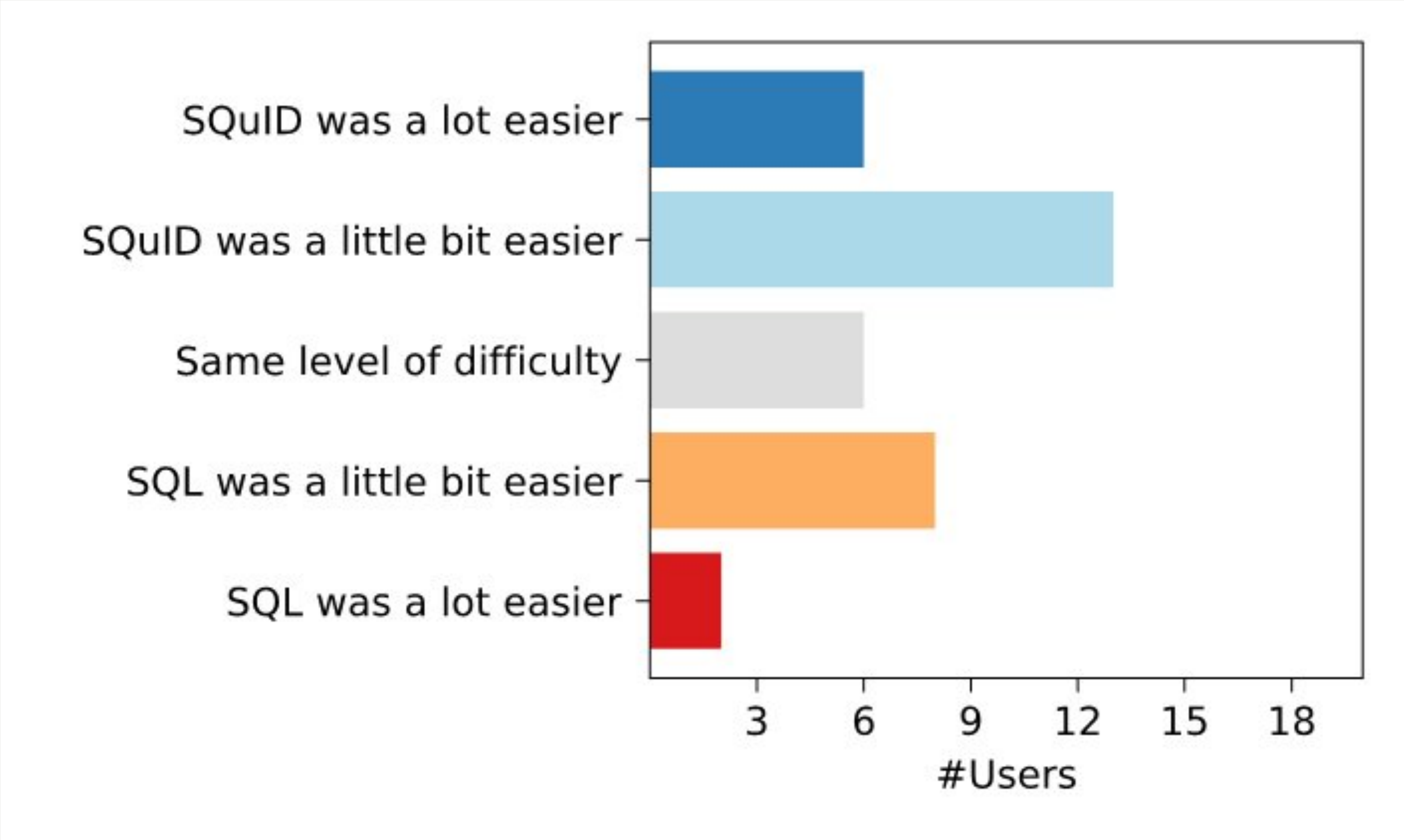




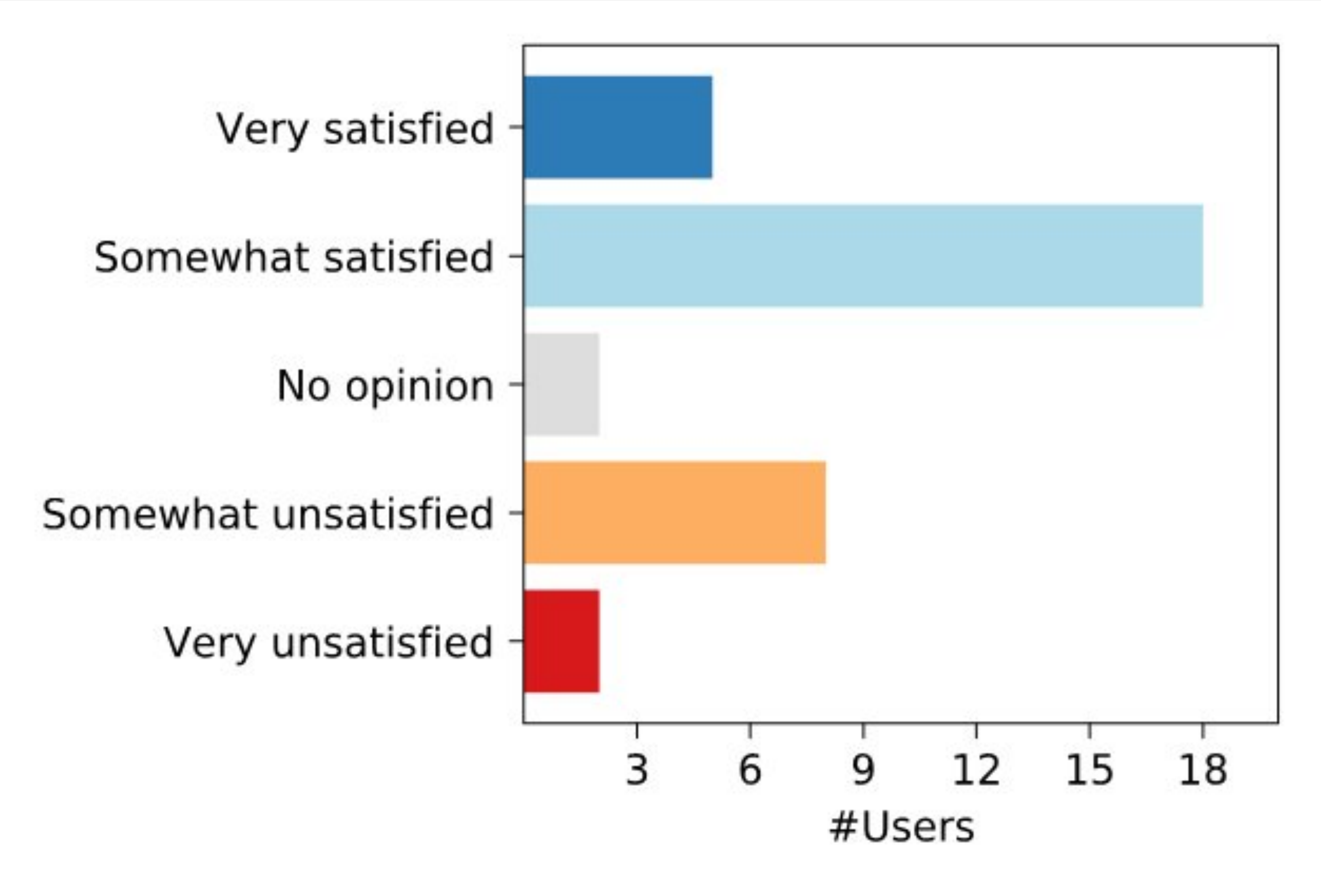
Overall, SQuID generated more accurate results



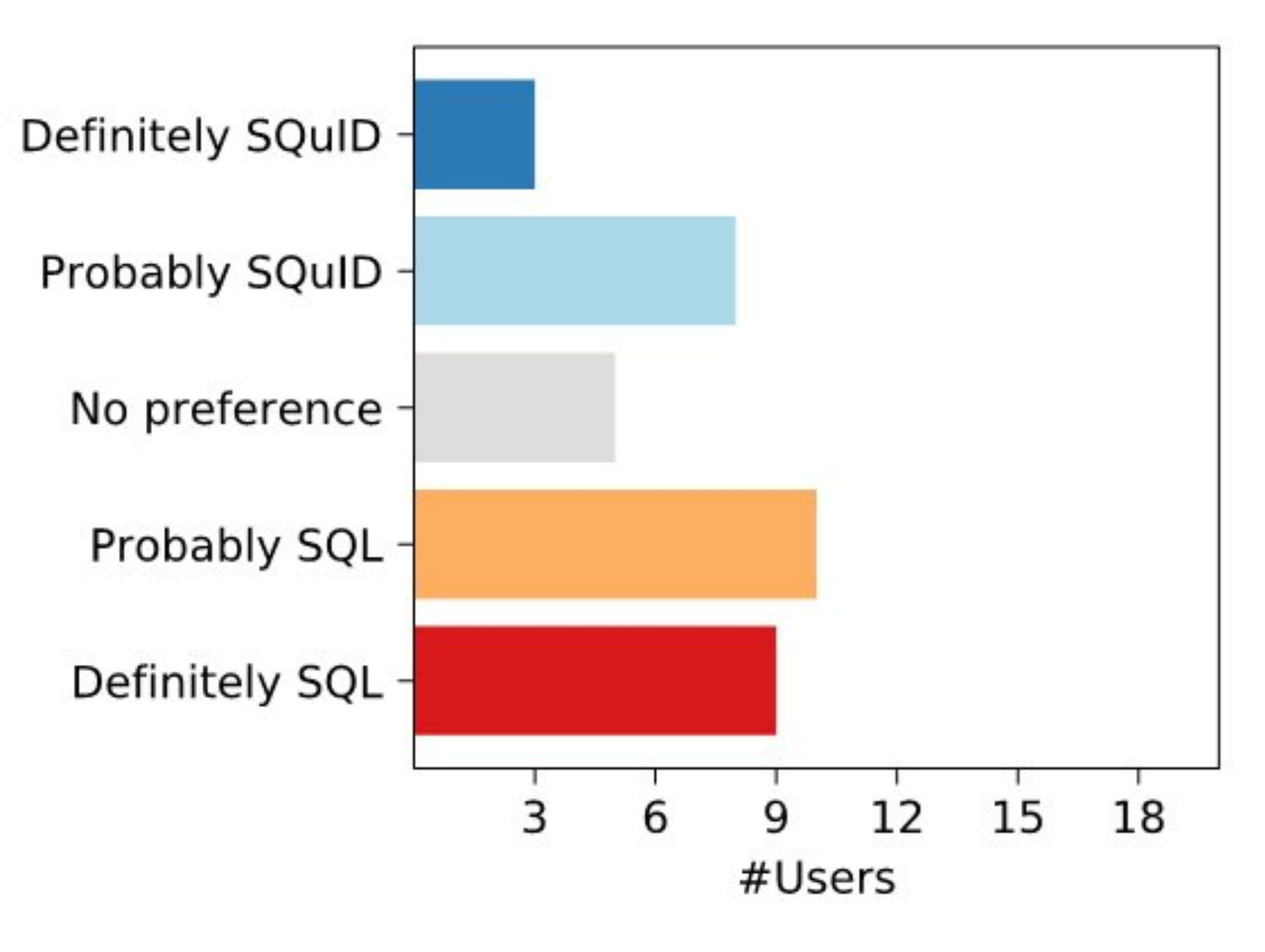
SQuID was easier to use



Participants were satisfied with SQuID results



SQuID or SQL?

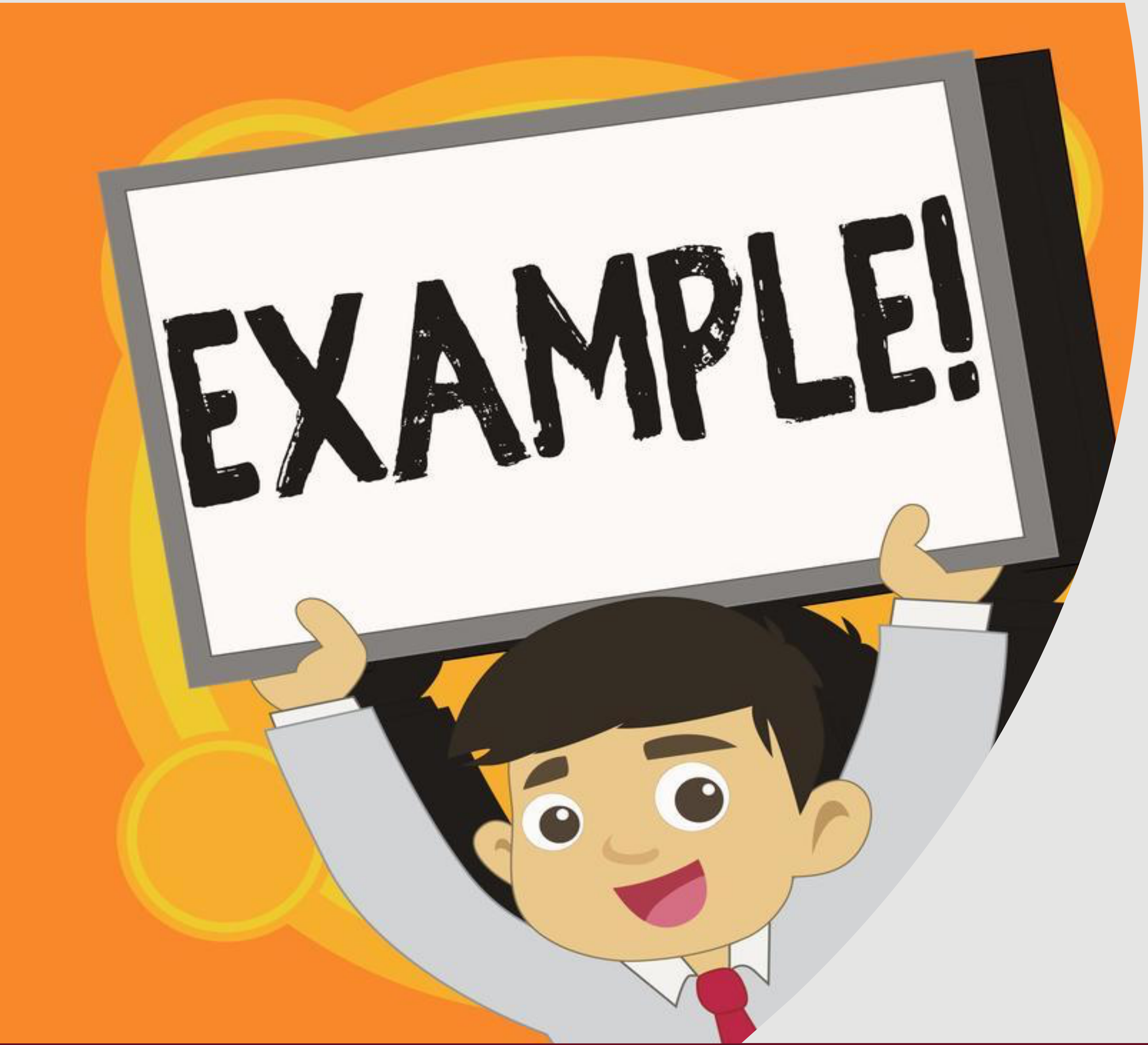

Definitely SQuID
Probably SQuID
No preference
Probably SQL
Definitely SQL
3
6
9
12
15
18
#Users

Anecdotal comments

“Even if I forget about syntax . . . figuring out ***how to go about writing the pseudo-code*** *query for funny actors [is difficult]”*

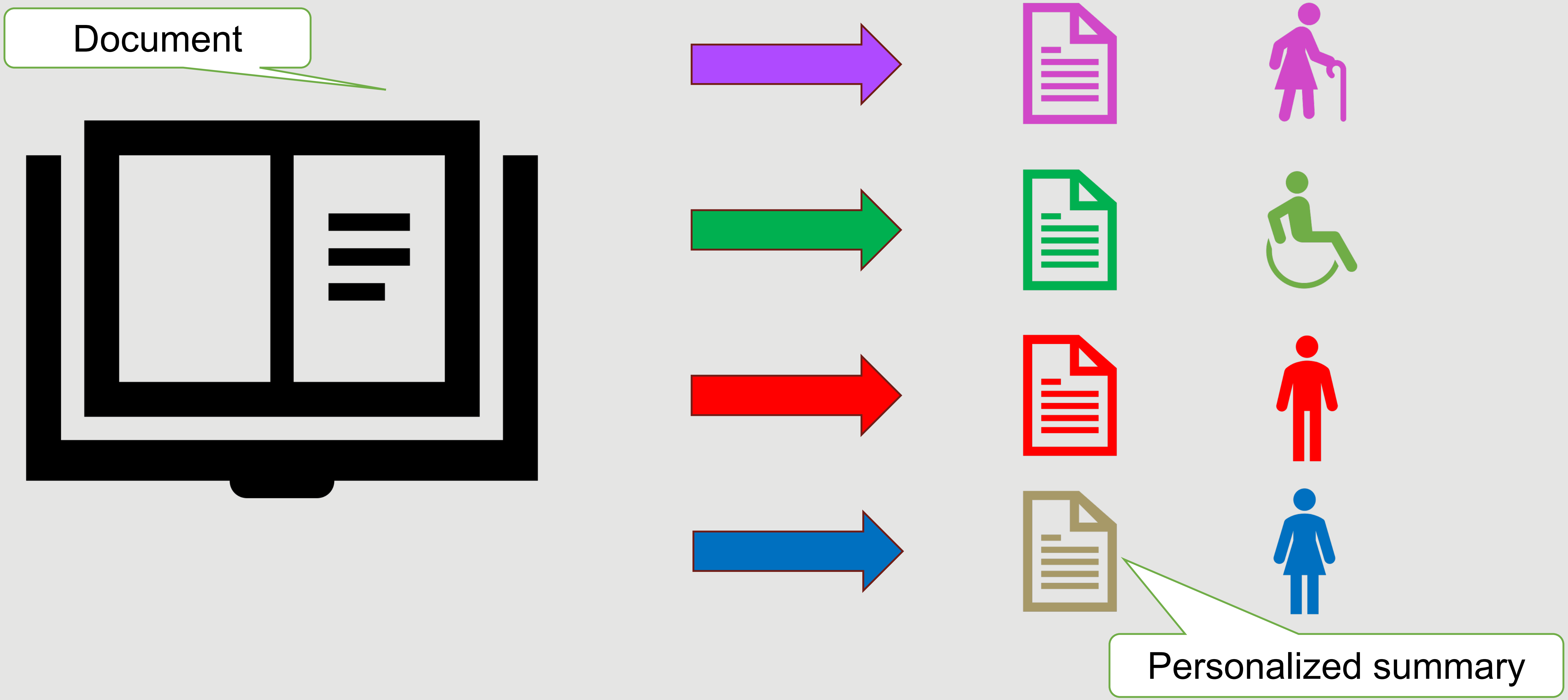
“Vague tasks are generally a lot more open to interpretation. Coding up a query that meets someone’s ***vague specifications*** *[is] hard . . . It was* ***very hard to nail down what the correct definition of funny is****.”*

Personalized text document summarization

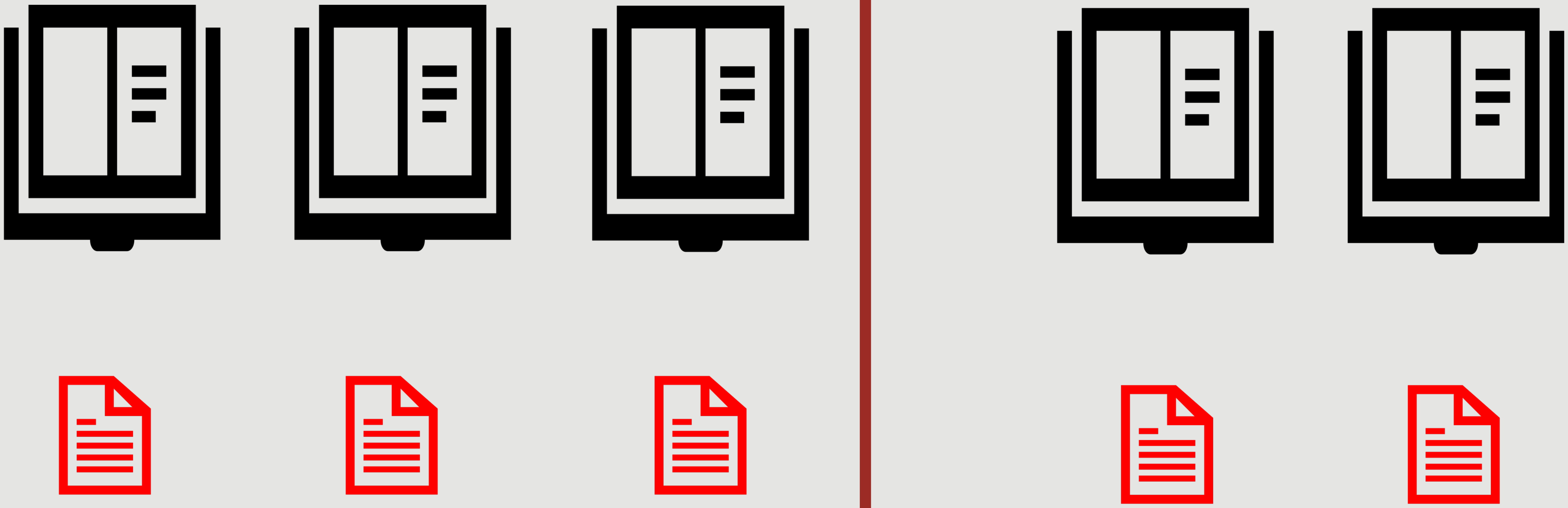


SuDocu: Summarizing Documents by Example

Personalized summarization



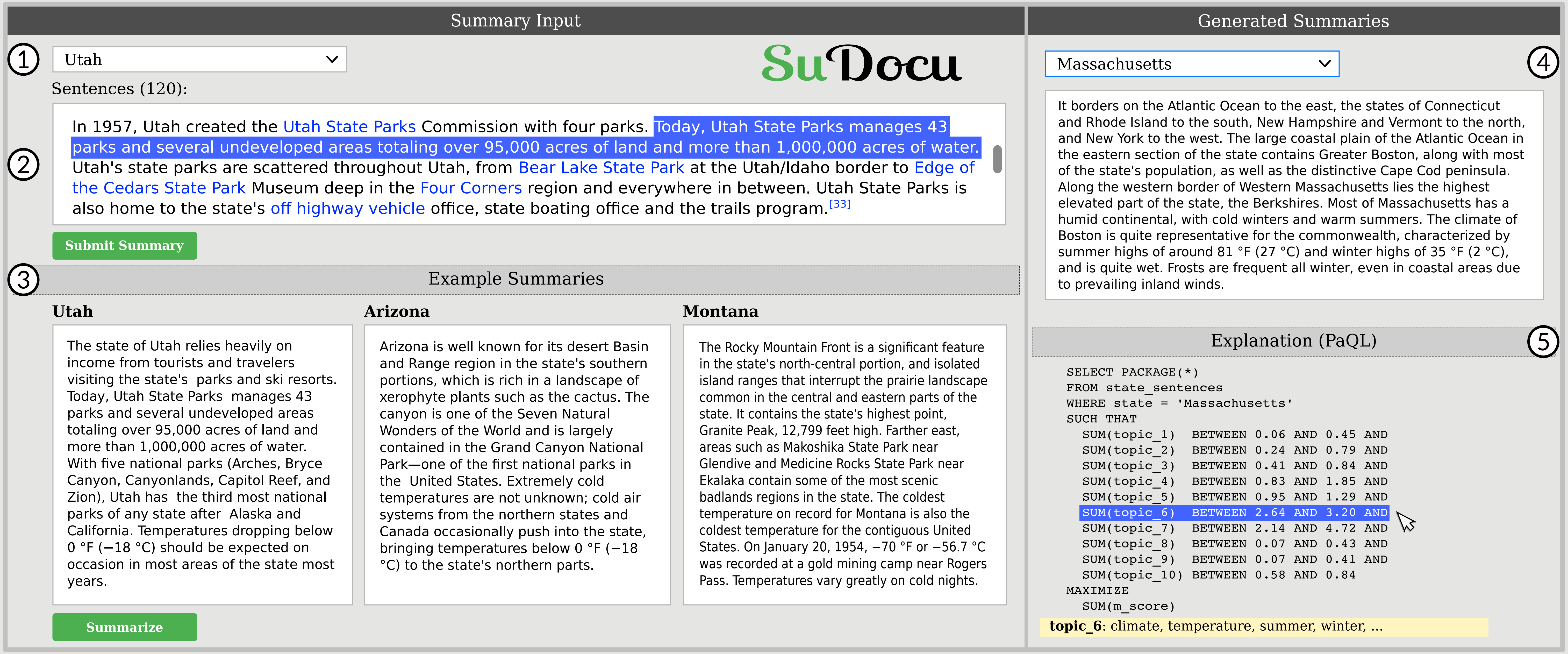
Summarization by example



Example Summaries

Automatic Summaries

SuDocu interface

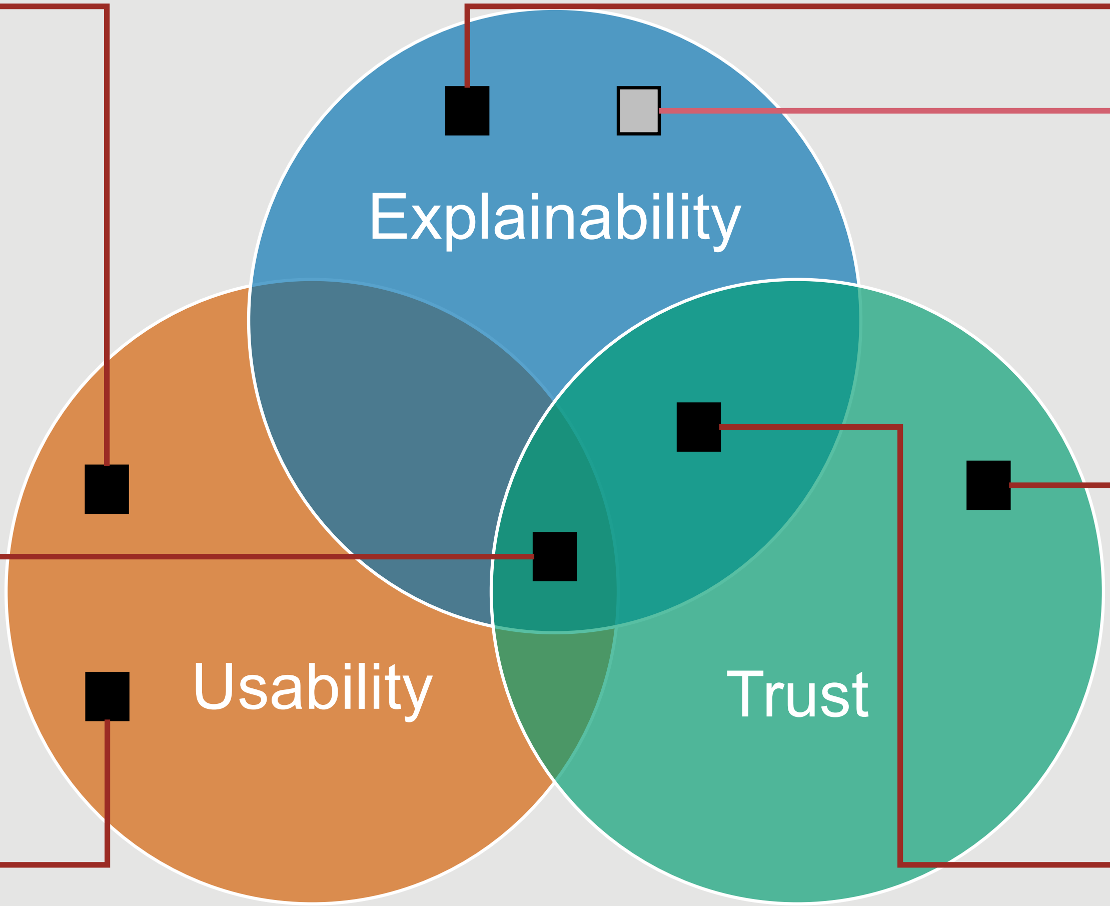


Dissertation outline

Query by Example (QBE)
[VLDB 2019]
[SIGMOD 2018] (demo)

Comparative User Study: QBE vs SQL
[CHI 2020]*

Data Summarization by Example
[VLDB 2020] (demo)



Adaptive Interventional Debugging
[SIGMOD 2020]

Data Change Explanation
WORK IN PROGRESS

Conformance Constraints: Trusted ML
[SIGMOD 2021]*

Explaining Tuple Non-conformance
[SIGMOD 2019] (demo)

* under submission/revision

UMassAmherst | College of Information & Computer Sciences

Part 2:
Trust in Data Systems

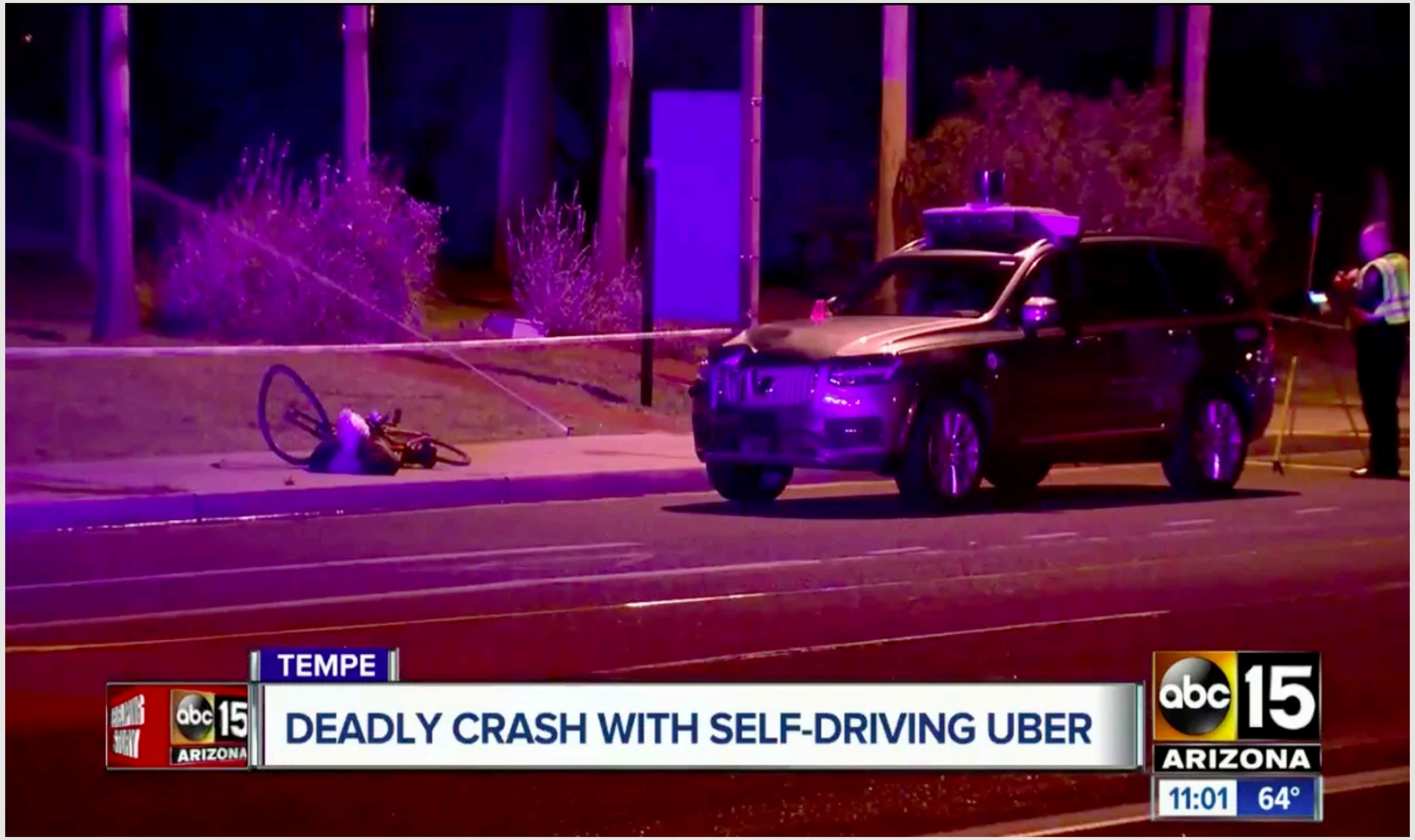
Trust

To trust or not to trust?

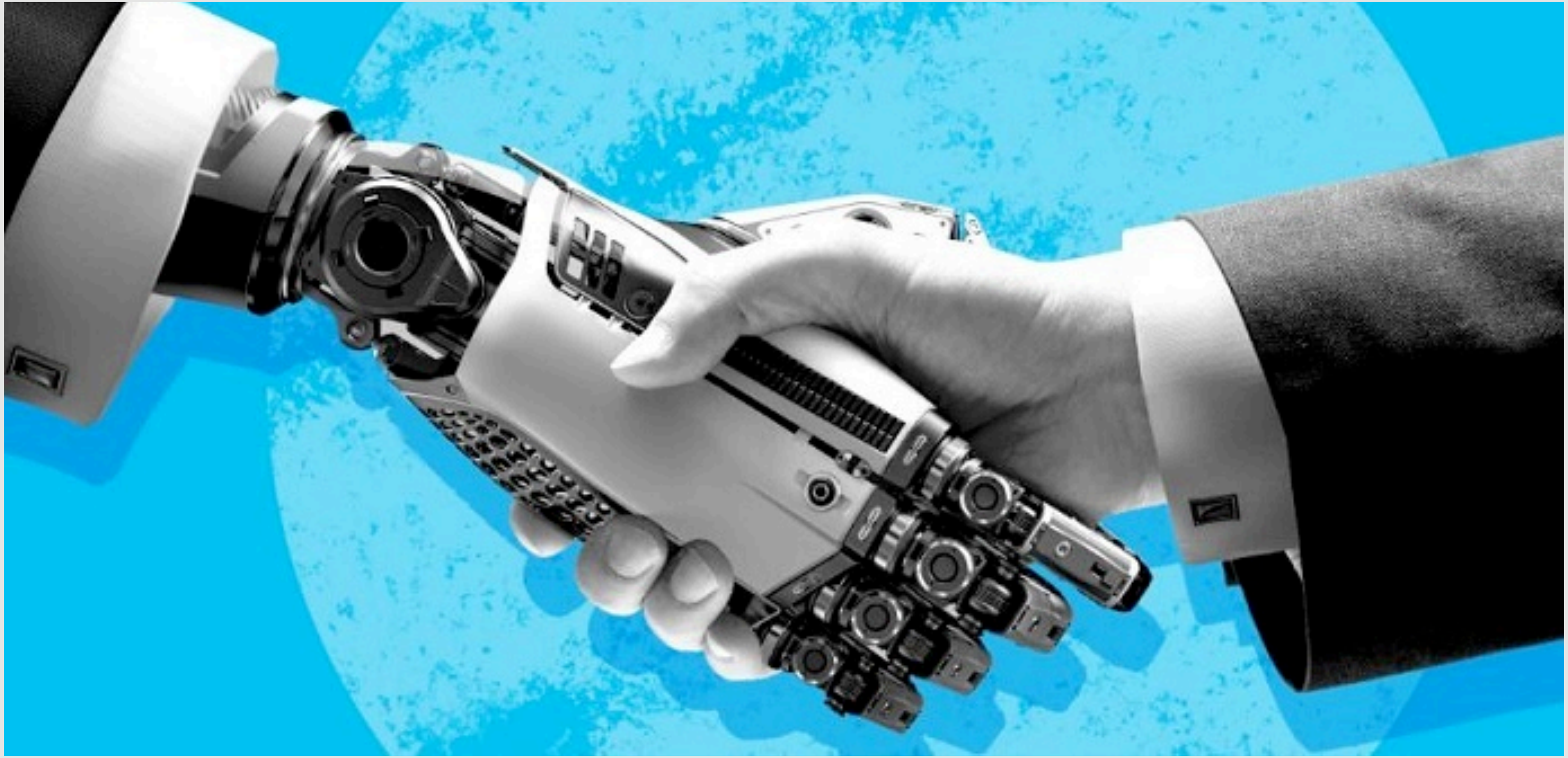
Daily Mail .com Science & Tech

IBM's Watson AI suggested 'often inaccurate' and 'unsafe' treatment recommendations for cancer patients, internal documents show

To trust or not to trust?

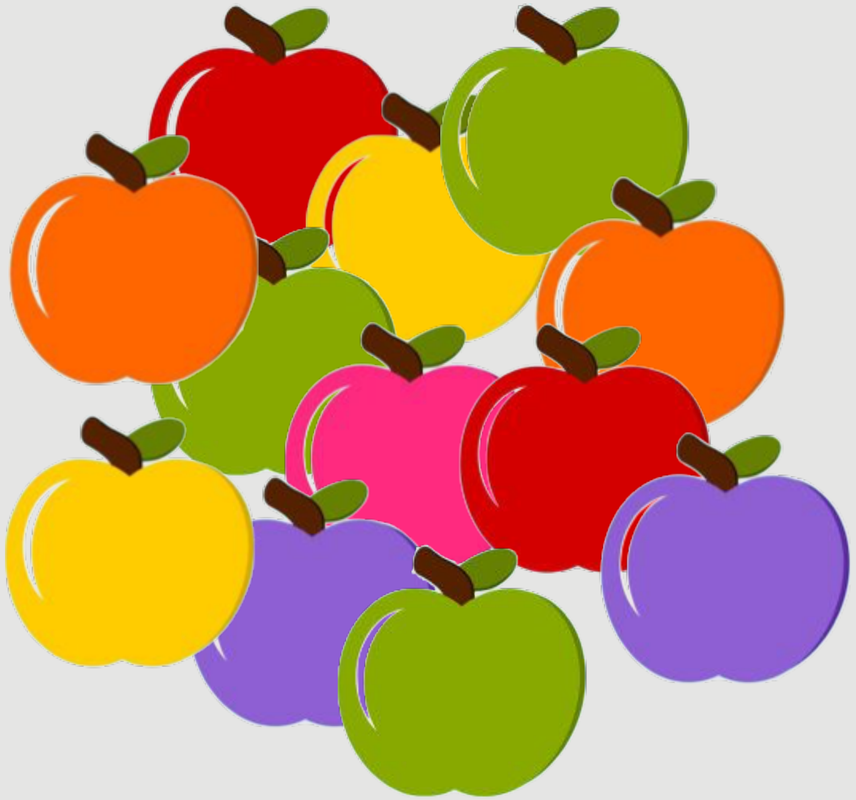


Conformance constraints: trusted machine learning



Trusting ML predictions

Training data



New data


Red


Pink

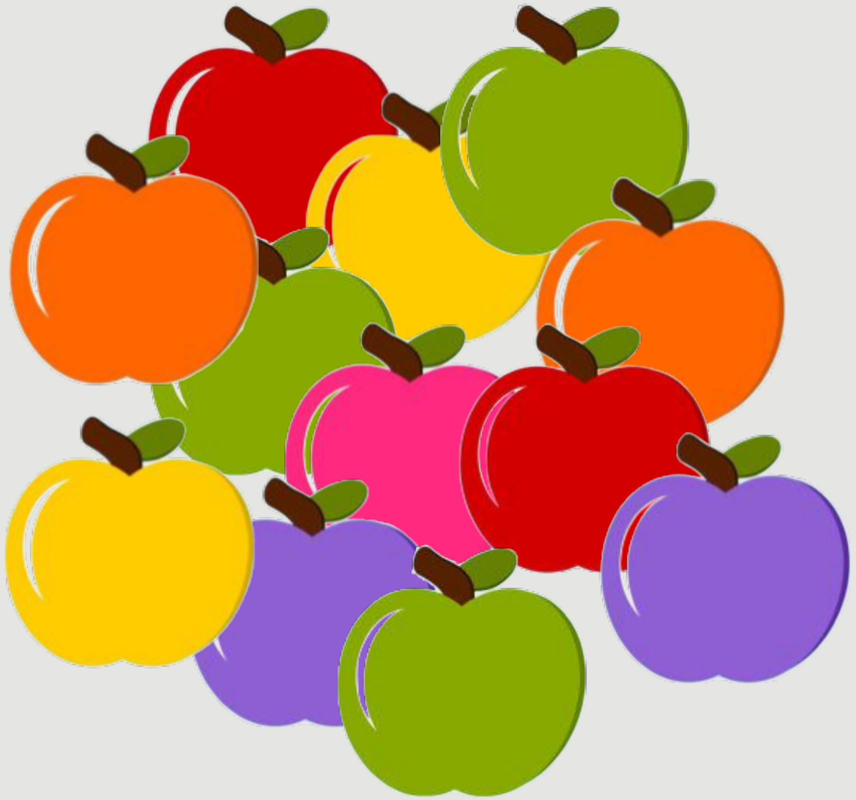

Yellow


Orange


Green

Trusting ML predictions

Training data



New data

Red

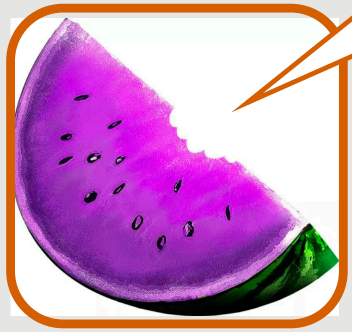
Pink

Yellow

Orange

Green

Non-conforming



Green

Non-conformance = untrustworthy prediction

Detection

Is it
non-conforming?

A real-world example: airlines dataset

Regression task: predict arrival delay

…	dep_date	dep_time	arr_time	duration (minute)
…	May 2	14:30	18:20	230
…	July 22	09:05	12:15	195
…	June 6	10:20	20:00	582
…	May 19	11:10	13:05	117
…	April 7	22:30	06:10	458

DAYTIME flights

OVERNIGHT flight

A real-world example: airlines dataset

- Trained with DAYTIME flights only
- Constraints observed in DAYTIME flights
  - “departure time is earlier than arrival time”
  - “their difference is very close to flight duration”

- OVERNIGHT flights
  - violate DAYTIME flights’ constraints
  - incur high regression error

Constraint violation correlates with high regression error

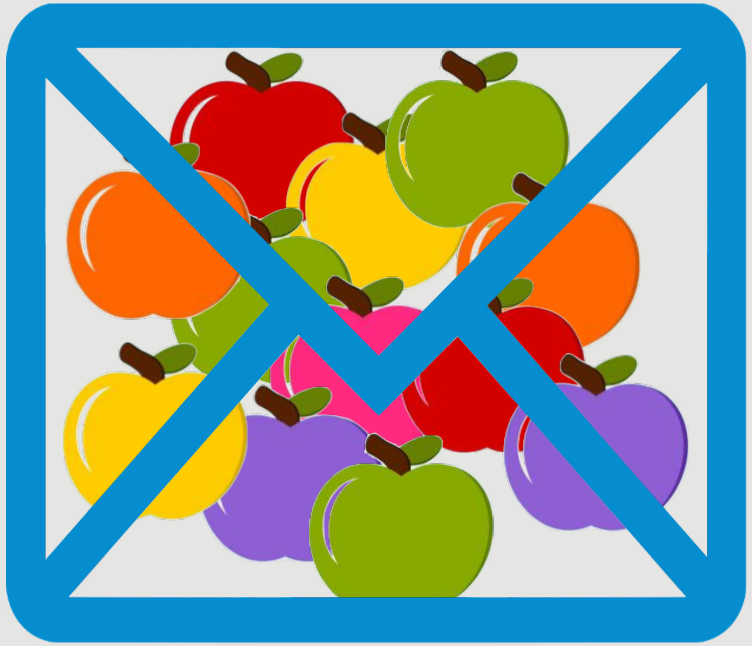
Conformance constraints (CCs)

ML pipelines drop low-variance dimensions to achieve dimensionality reduction.

ML models assume that training data's constraints/properties will continue to hold during serving.

Conformance constraints

- constraints that the data satisfies
- capture the invariants of the data



Conformance constraints

- Encode linear arithmetic relationship over multiple attributes.

$$\underbrace{-\epsilon}_{\text{Lower bound}} \leq \underbrace{(60 \cdot arr\_hour + arr\_min) - (60 \cdot dep\_hour + dep\_min) - duration}_{\text{Projection}} \leq \underbrace{\epsilon}_{\text{Upper bound}}$$

Conformance constraints: example

Height	Weight	BMI
6 feet	142 lbs	19.3
5 feet	170 lbs	33.2
5 feet	130 lbs	25.4

$$\boxed{10 \leq \text{BMI} \leq 40}$$

$$\boxed{-40 \leq (28 \times \text{Height} - \text{Weight}) \leq 30}$$

Violation of conformance constraint

$10 \leq BMI \leq 40$

Height	Weight	BMI
6 feet	142 lbs	19.3
5 feet	170 lbs	33.2
5 feet	130 lbs	25.4
6 feet	170 lbs	231

Degree of violation

$10 \leq BMI \leq 40$

Height	Weight	BMI
6 feet	142 lbs	19.3
5 feet	170 lbs	33.2
5 feet	130 lbs	25.4
6 feet	170 lbs	231
6 feet	170 lbs	20000

Projection

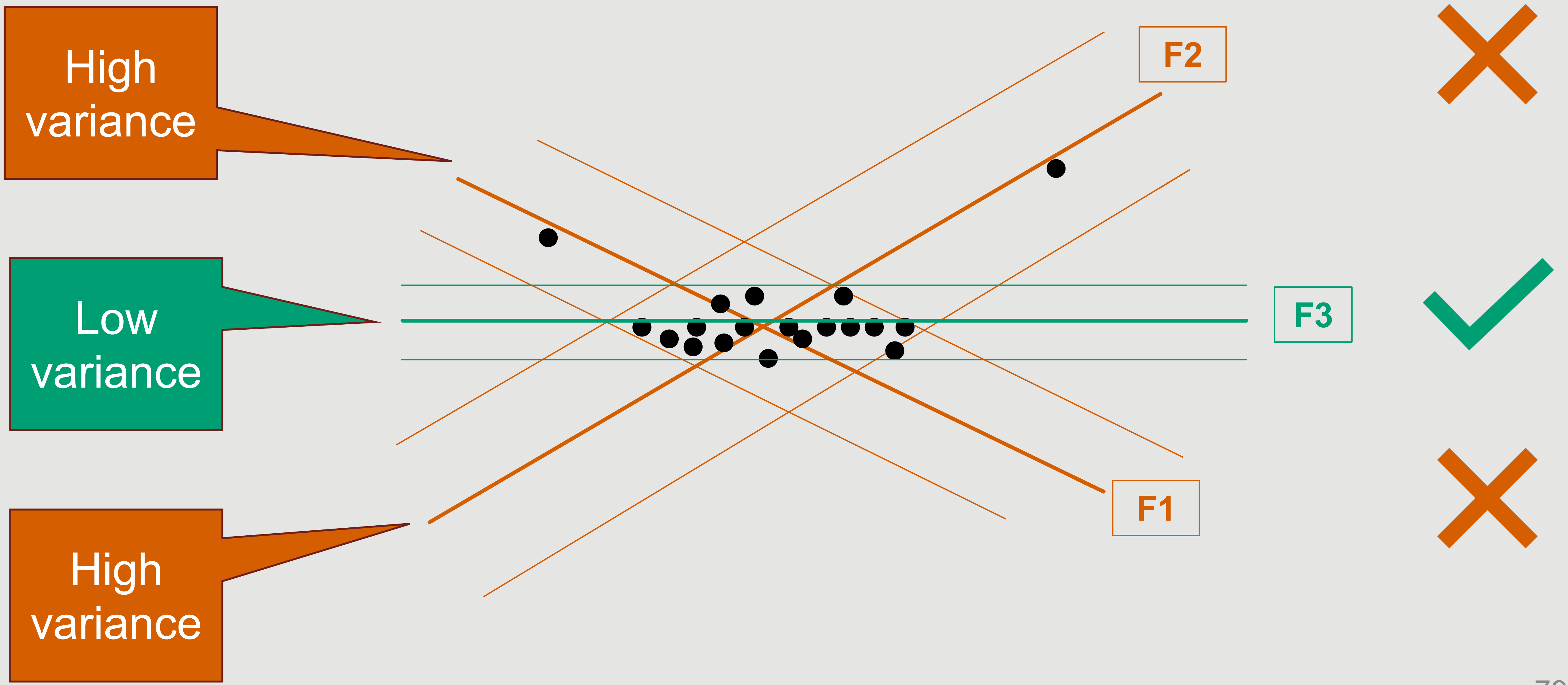
$$\underbrace{-\epsilon}_{\text{Lower bound}} \leq \underbrace{(60 \cdot \text{arr\_hour} + \text{arr\_min}) - (60 \cdot \text{dep\_hour} + \text{dep\_min}) - \text{duration}}_{\text{Projection}} \leq \underbrace{\epsilon}_{\text{Upper bound}}$$

What are "good" projections?

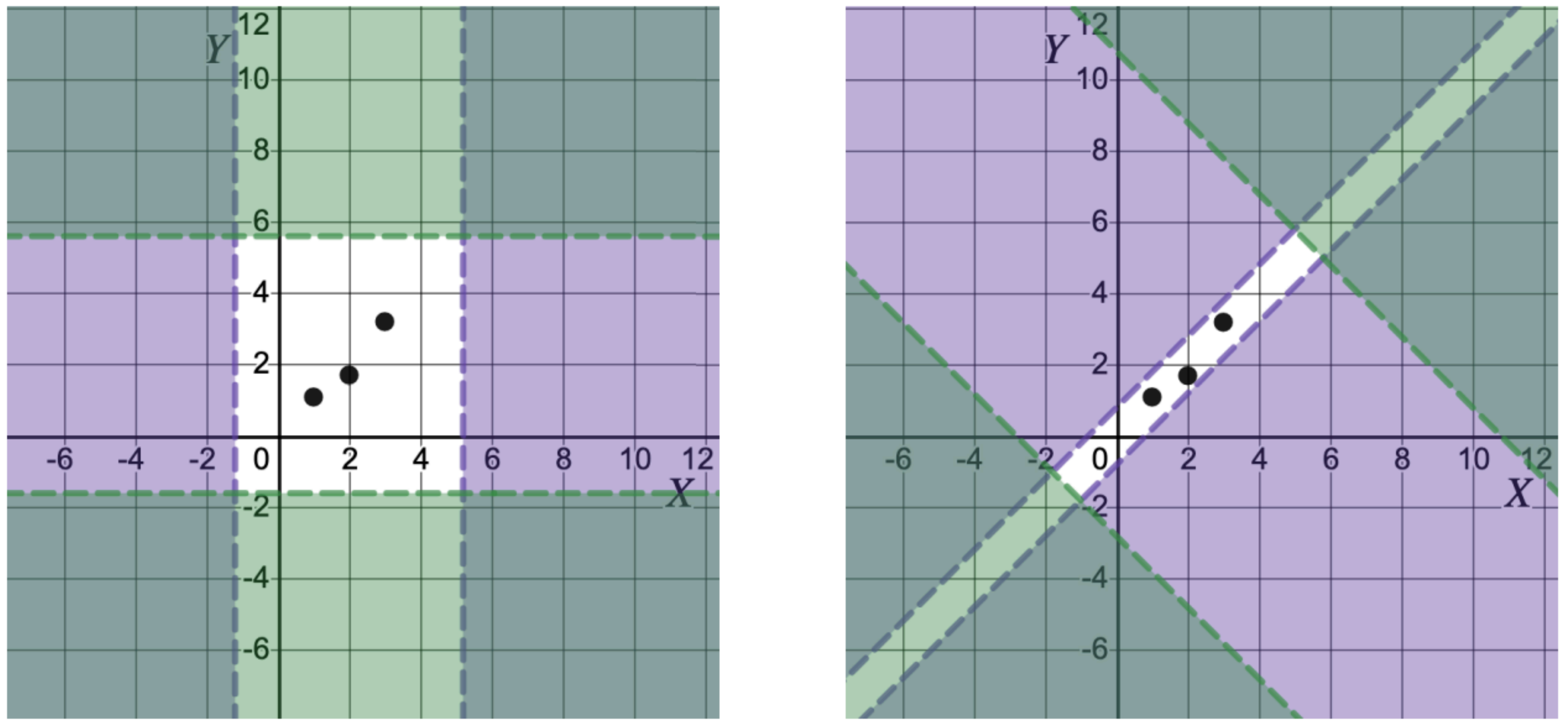
- Infinitely many projections possible
  - Pick the **low-variance** projections.
  - Because?
    - They more useful in detecting trends in the data.

- Do we pick all low-variance projections?
  - Pick a set of projections with **low pair-wise correlations**.
  - Because?
    - They complement each other.

Low-variance projections

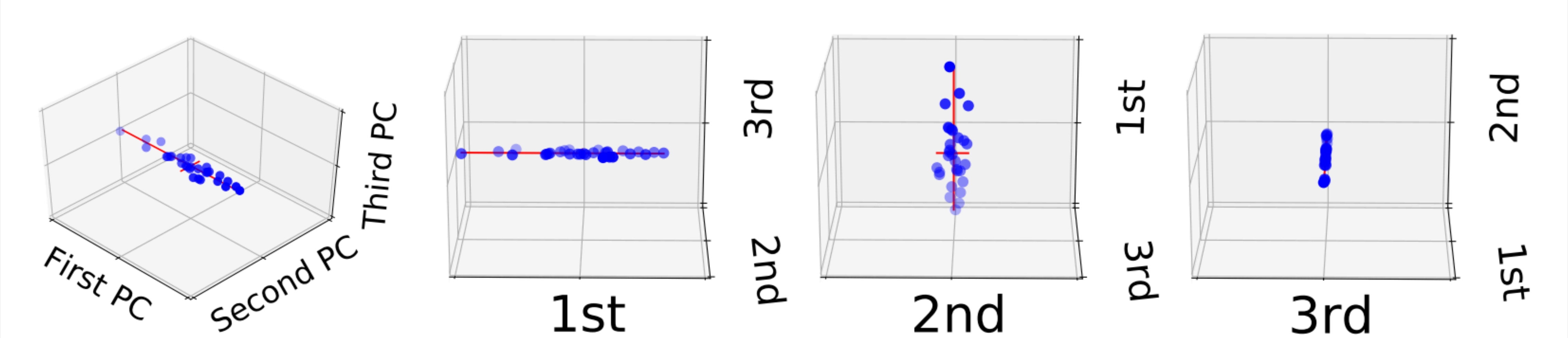


Projections with small mutual correlation



Discovering projections: PCA

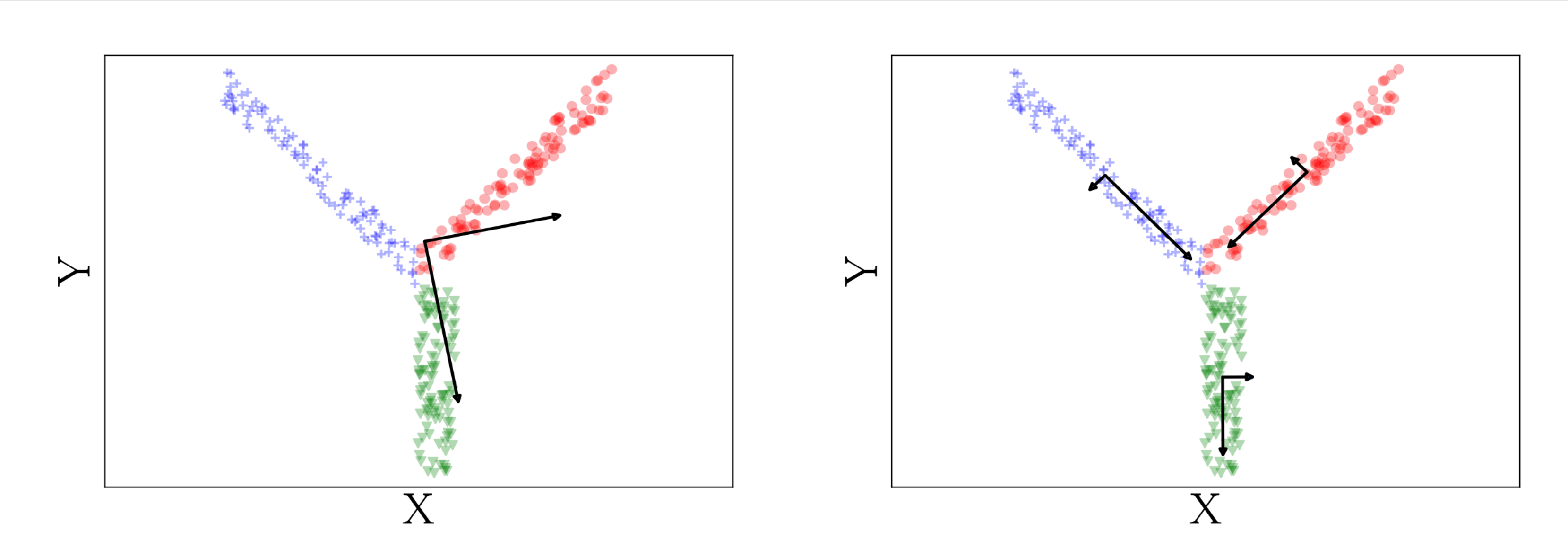
- Principal Component Analysis (PCA)
  - Produces projections with small mutual correlations
    - Intuition: principal components are **orthogonal** to each other
  - Computing violation
    - Weigh CCs with **low** variance projections **more**
    - Weigh CCs with **high** variance projections **less**



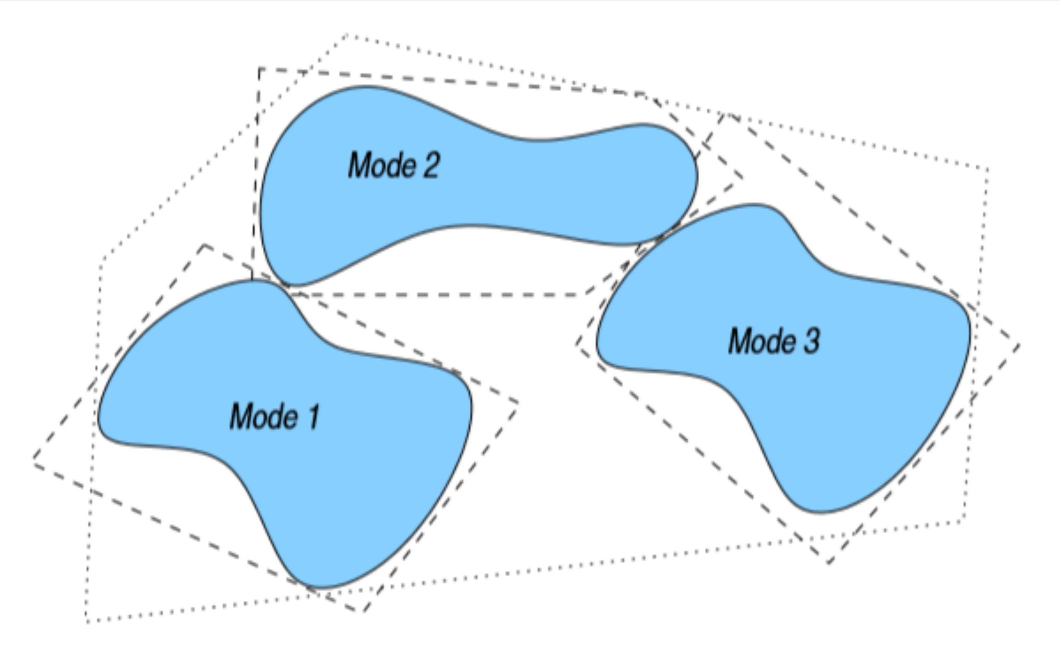
Disjunctive conformance constraints

- Divide the dataset into disjoint partitions.
- Learn CCs for each partition.
- Compute disjunctive CCs.

$$\begin{aligned}\psi_2 : & \; M = \text{“May”} \triangleright -2 \leq AT - DT - DUR \leq 0 \\ \vee & \; M = \text{“June”} \triangleright \; 0 \leq AT - DT - DUR \leq 5 \\ \vee & \; M = \text{“July”} \triangleright -5 \leq AT - DT - DUR \leq 0\end{aligned}$$



(a) PCA (b) Disjoint PCA



Complexity analysis

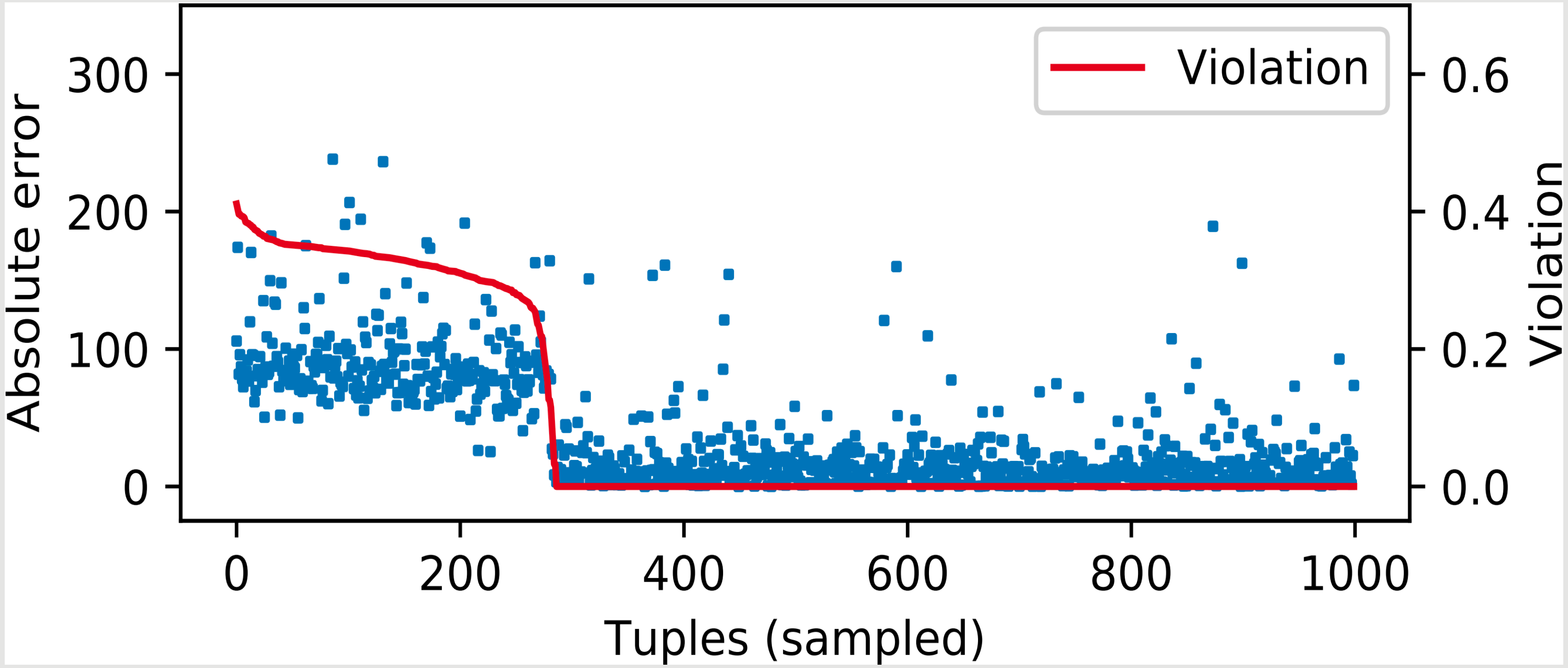
- Runtime
  - Linear in number of tuples in the dataset
  - Cubic in number of attributes
  - Highly parallelizable
- Memory
  - Quadratic in number of attributes

Experimental results: two applications

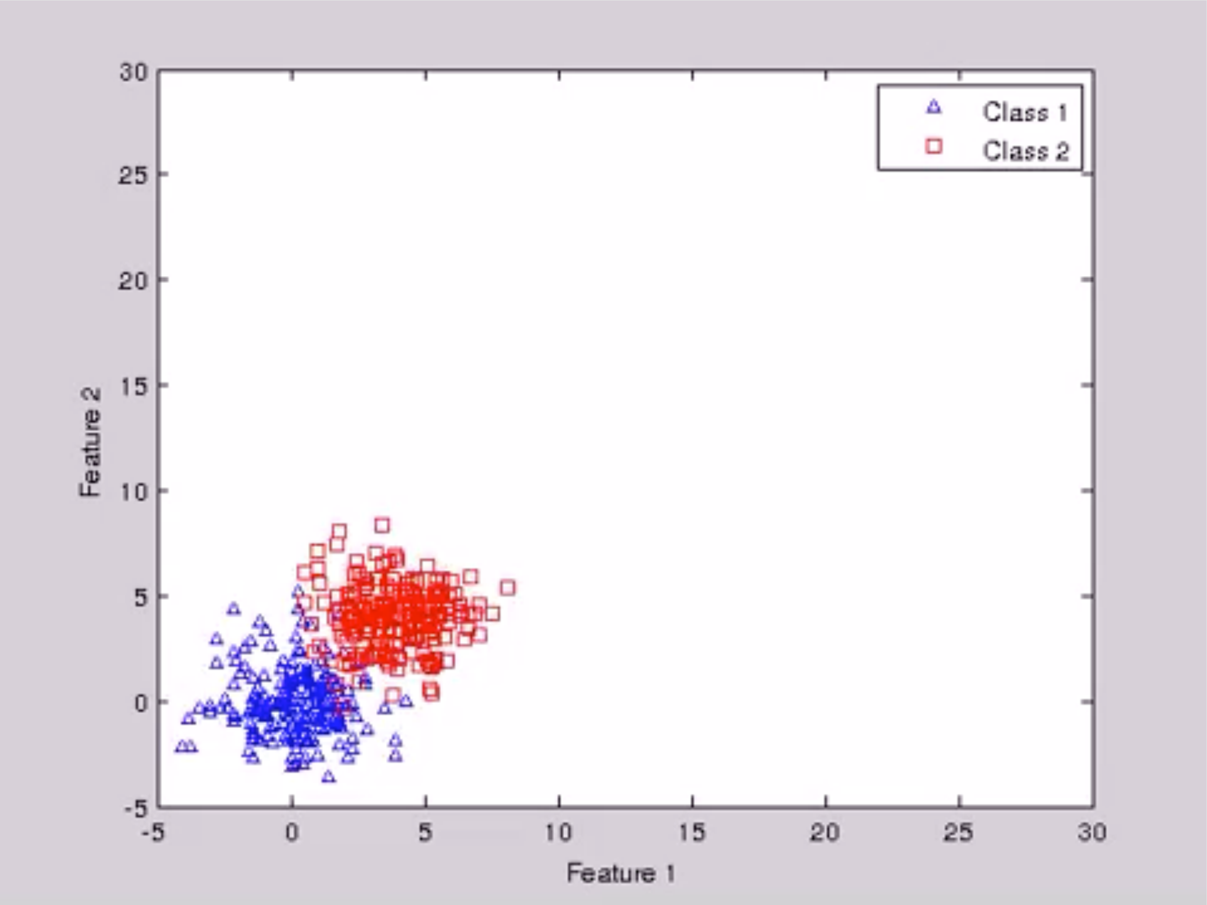
- Trusted Machine Learning
  - Is there a relationship between CC violation and the ML model's prediction accuracy?
- Data-drift
  - Can CCs be used to quantify data drift?

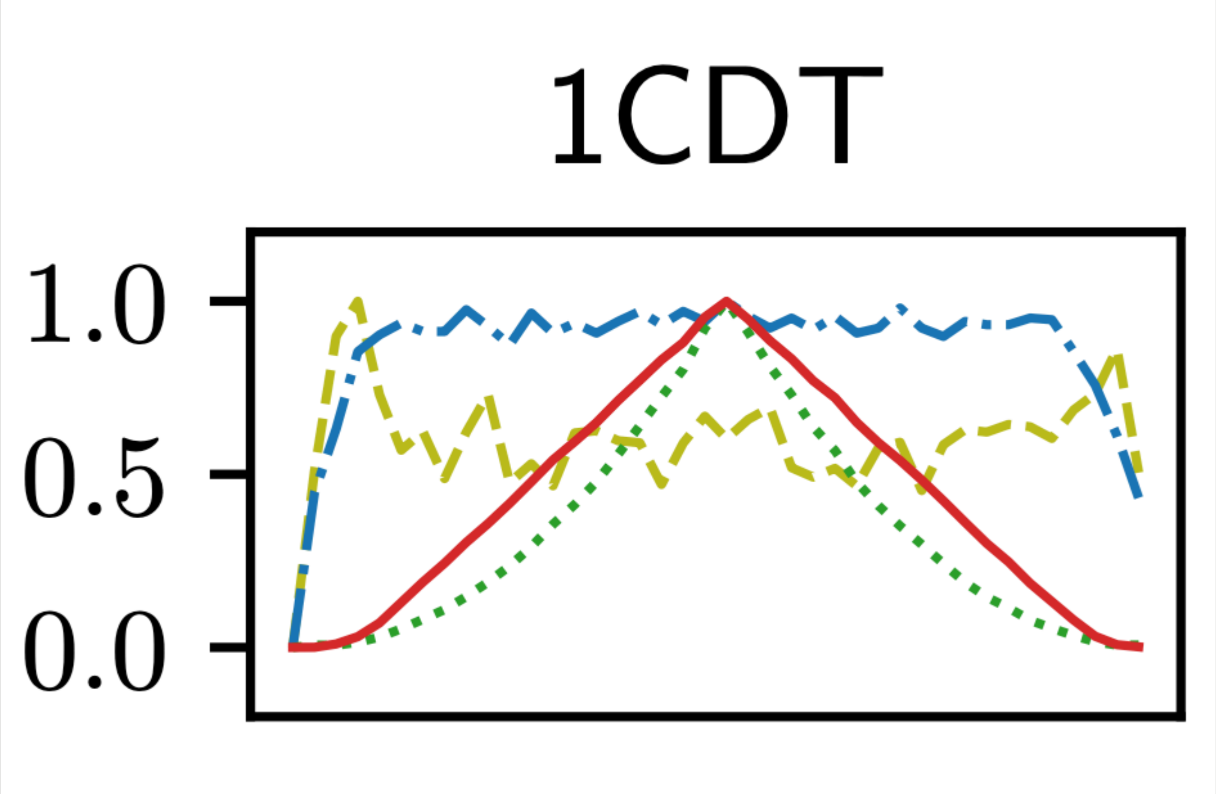
Trusted machine learning: airlines dataset

	Train	Serving		
		Daytime	Overnight	Mixed
Average violation	0.02%	0.02%	27.68%	8.87%
MAE	18.95	18.89	80.54	38.60

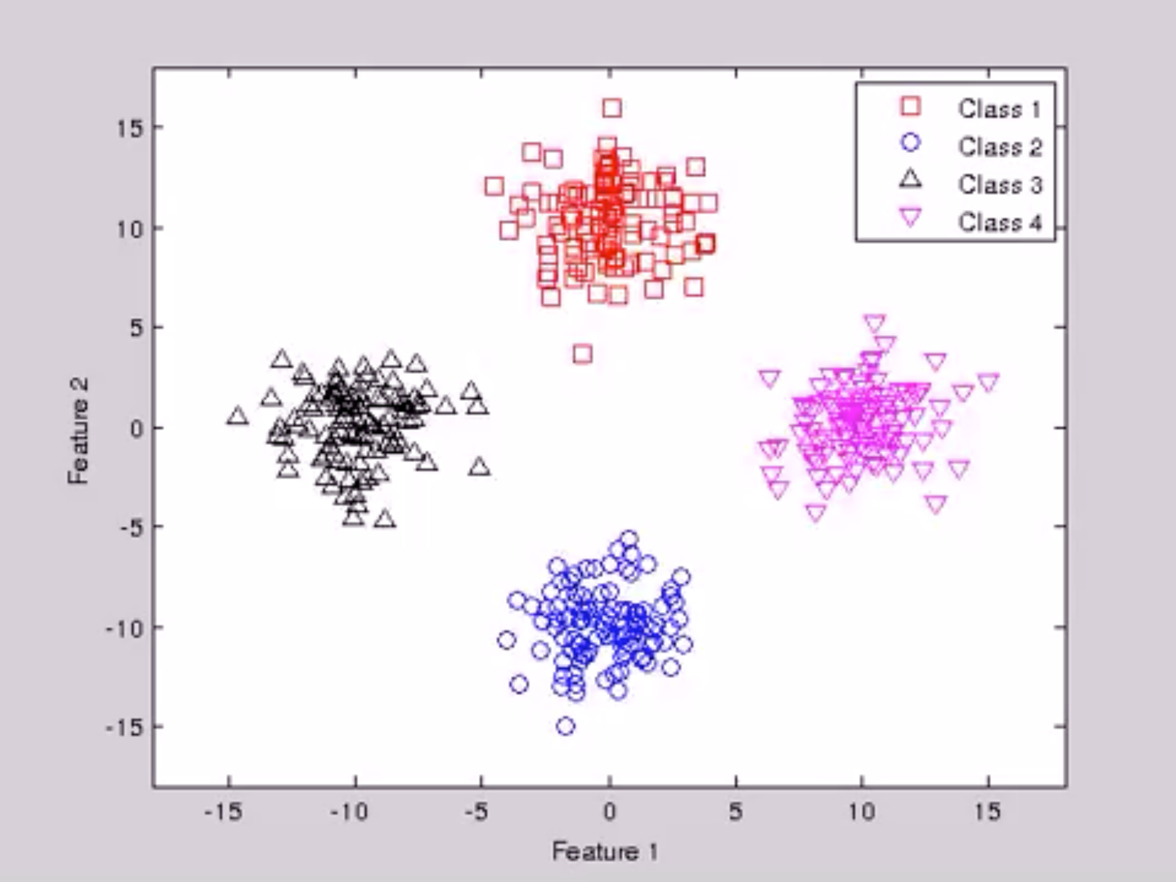


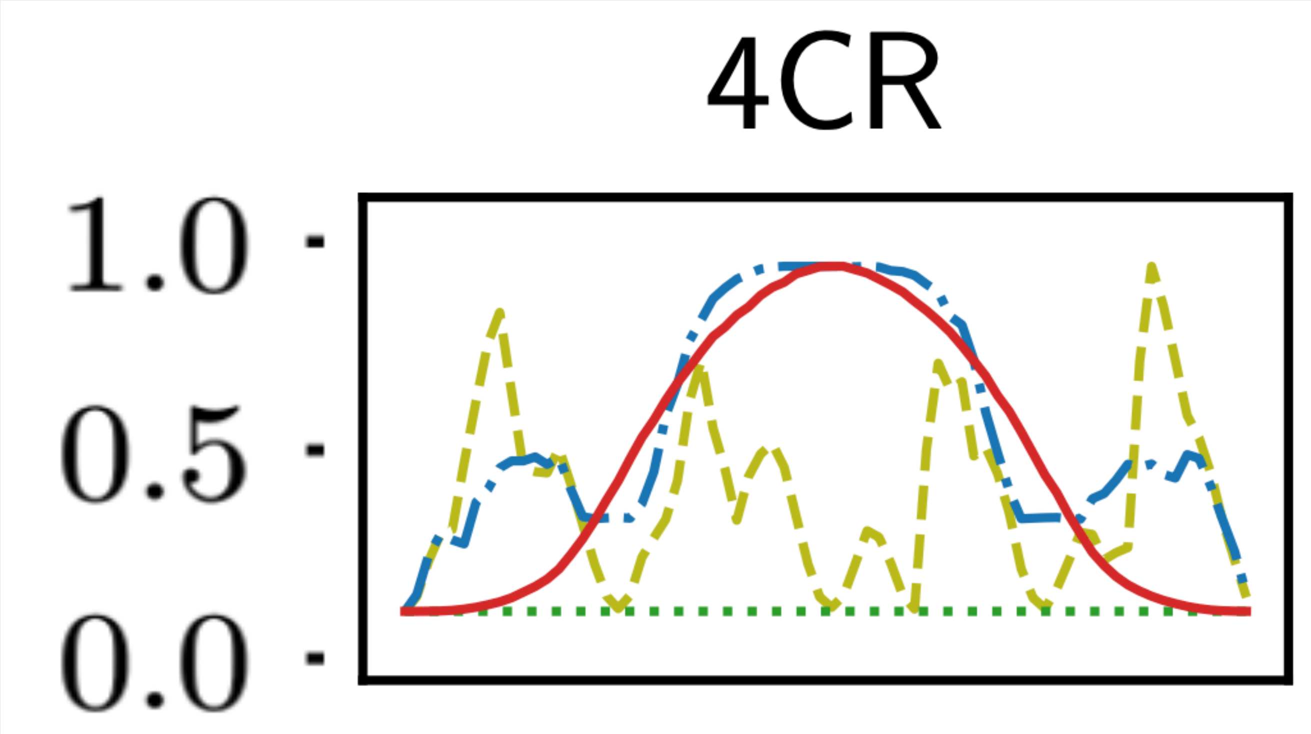
Data drift: EVL benchmark (1/4)



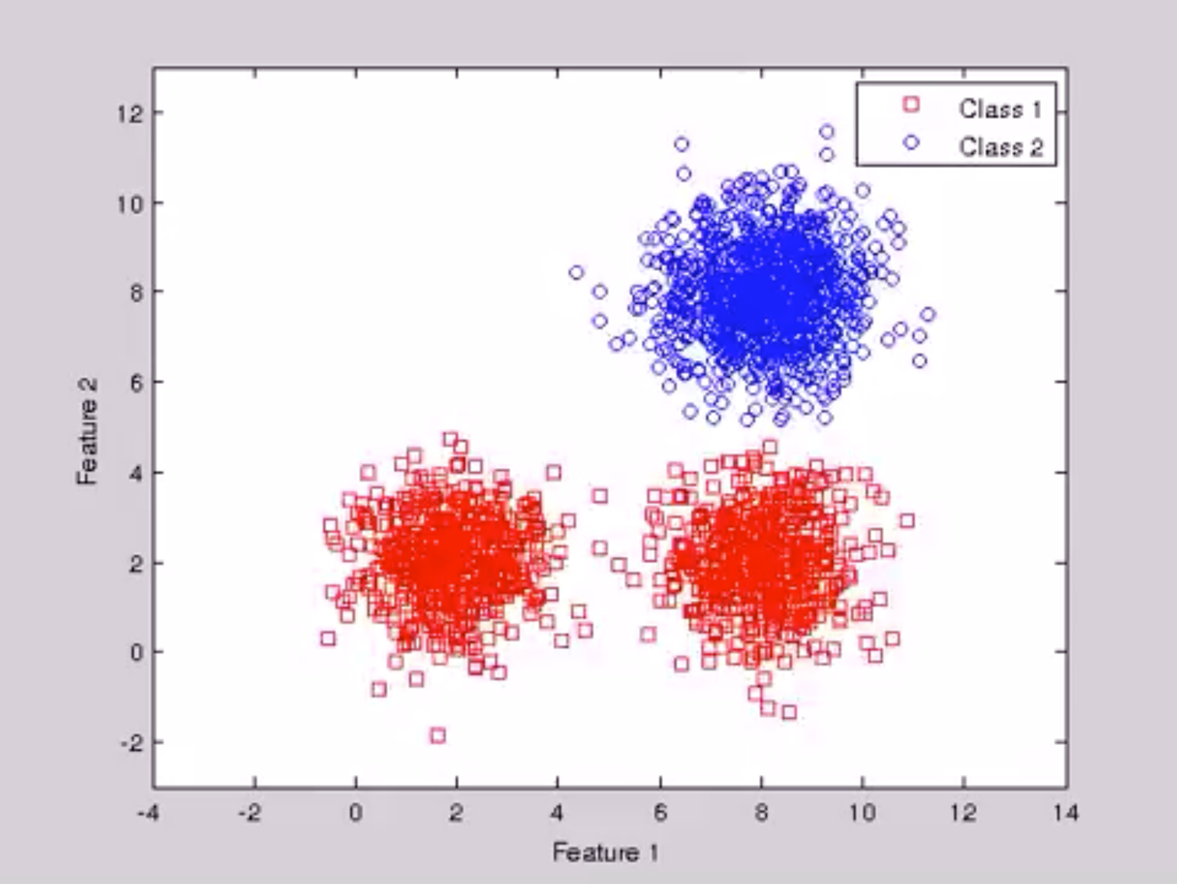


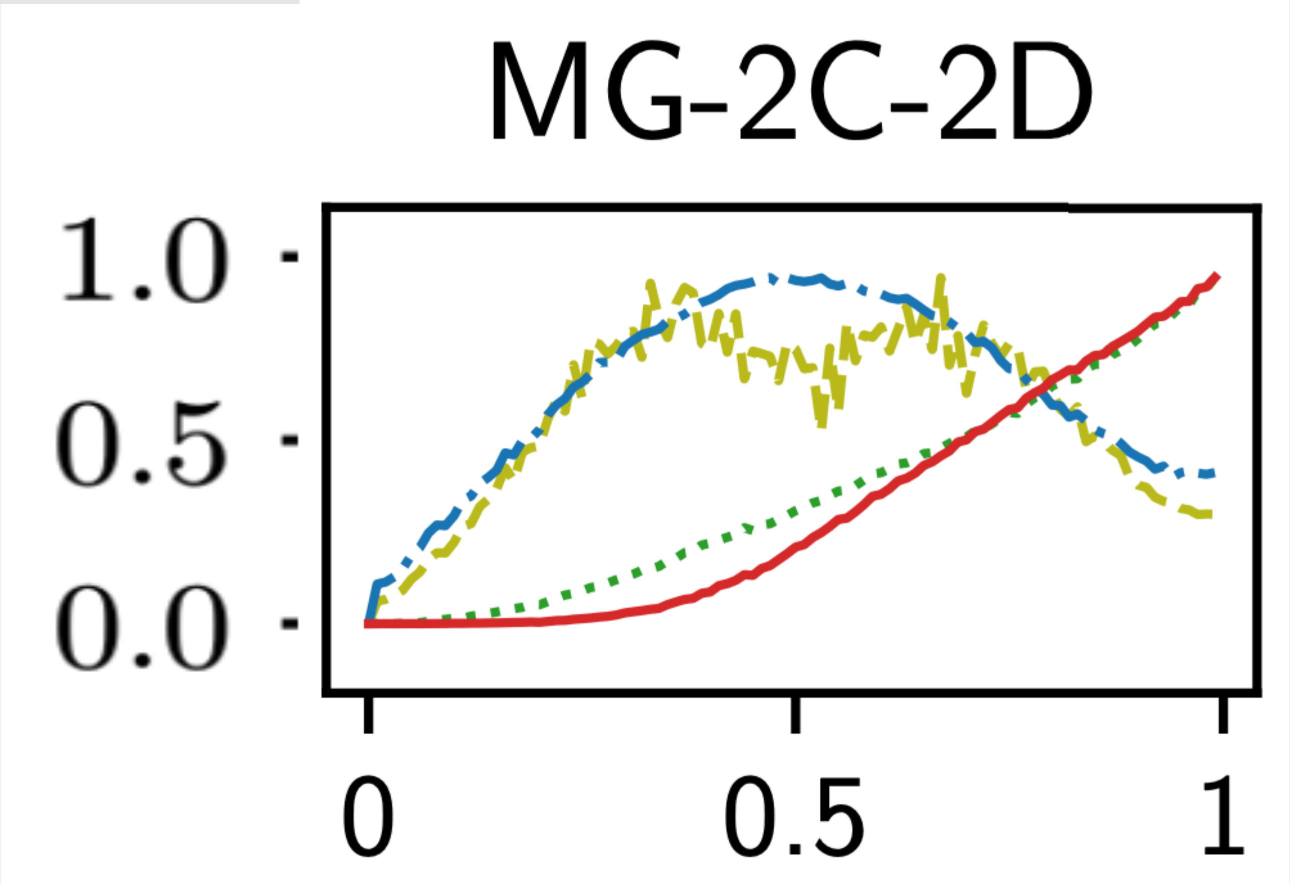
Data drift: EVL benchmark (2/4)





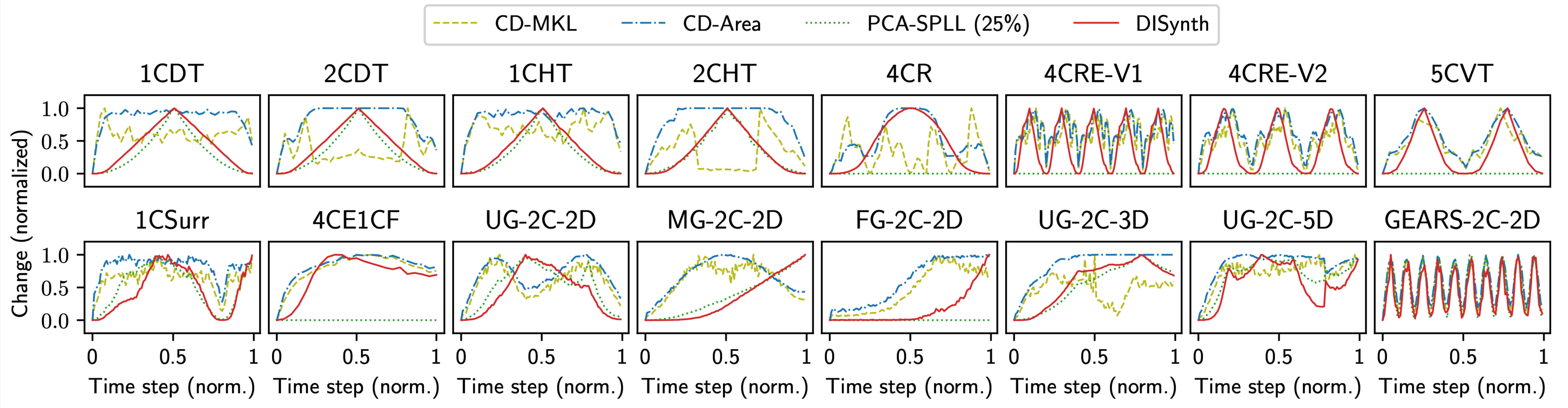
Data drift: EVL benchmark (3/4)





Data drift: EVL benchmark (4/4)

CD-MKL   CD-Area   PCA-SPLL (25%)   CCSynth

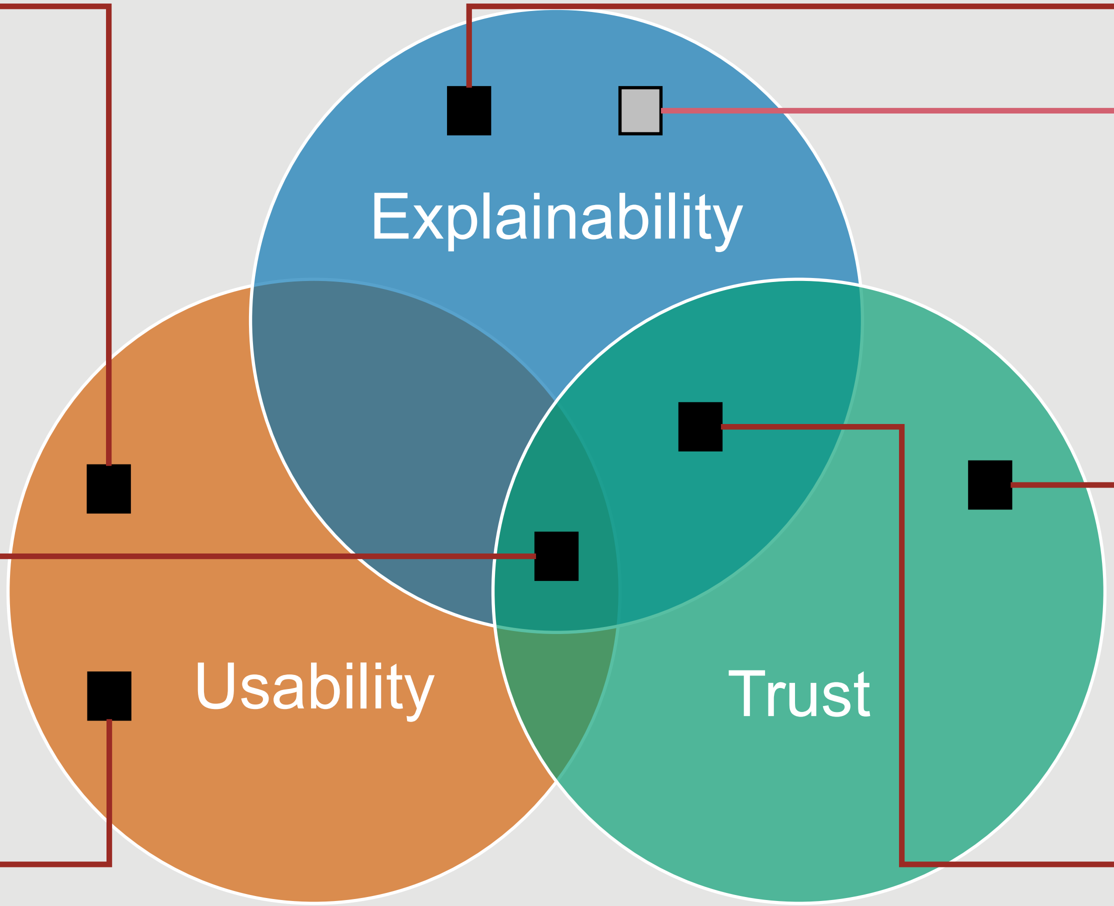


Dissertation outline

Query by Example (QBE)
[VLDB 2019]
[SIGMOD 2018] (demo)

Comparative User Study: QBE vs SQL
[CHI 2020]*

Data Summarization by Example
[VLDB 2020] (demo)



Adaptive Interventional Debugging
[SIGMOD 2020]

Data Change Explanation
WORK IN PROGRESS

Conformance Constraints: Trusted ML
[SIGMOD 2021]*

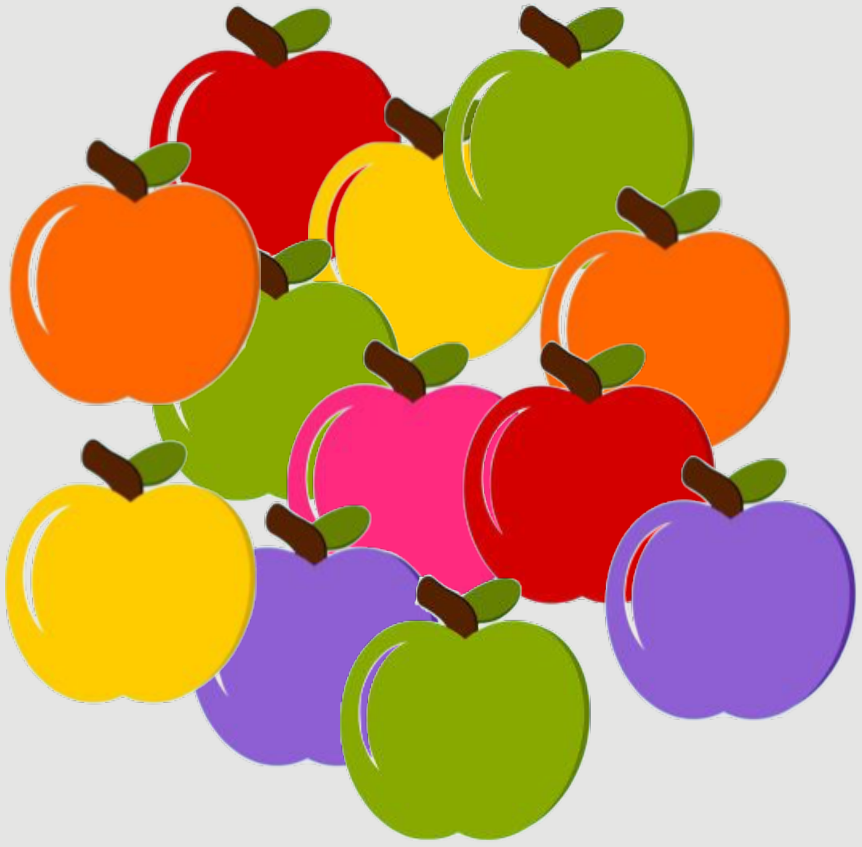
Explaining Tuple Non-conformance
[SIGMOD 2019] (demo)

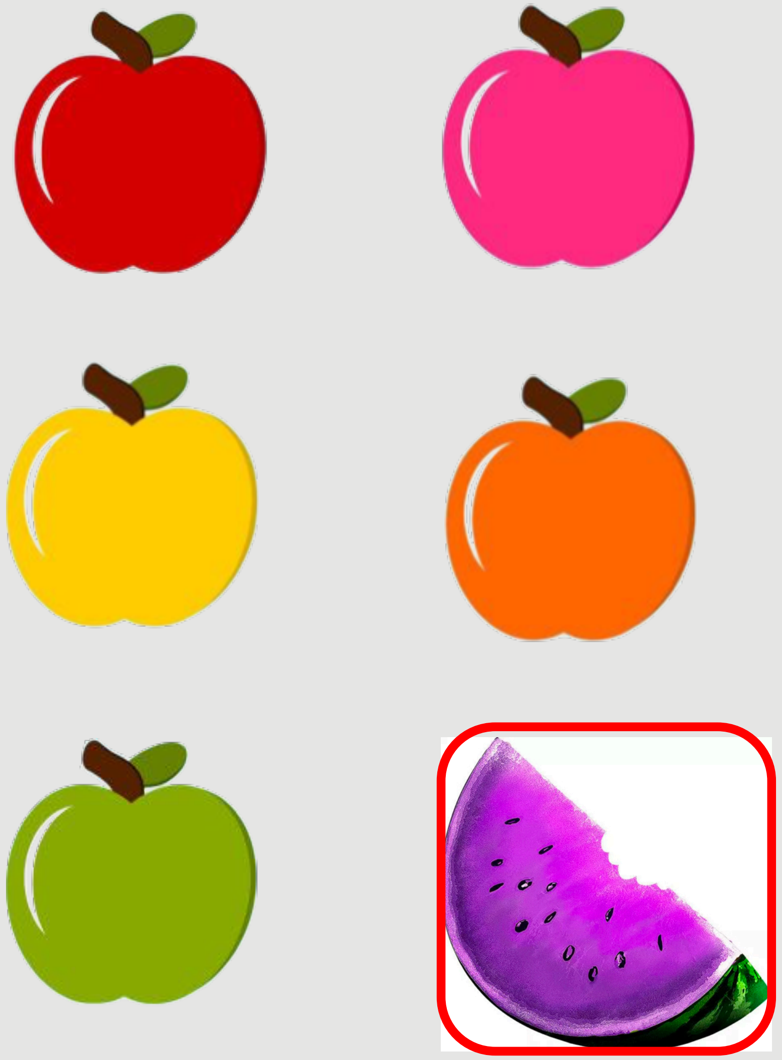
* under submission/revision

Part 3: Explanation Frameworks

Explainability

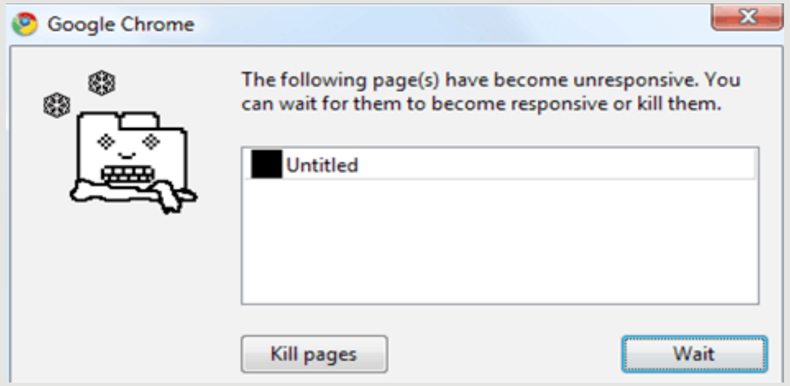
Why ML models fail for certain tuples?





How is this different?

Why do systems (sometimes) behave unexpectedly?



Why did the system crash?

ExTuNe

Explaining Tuple Non-conformance

Conformance constraints

Detection

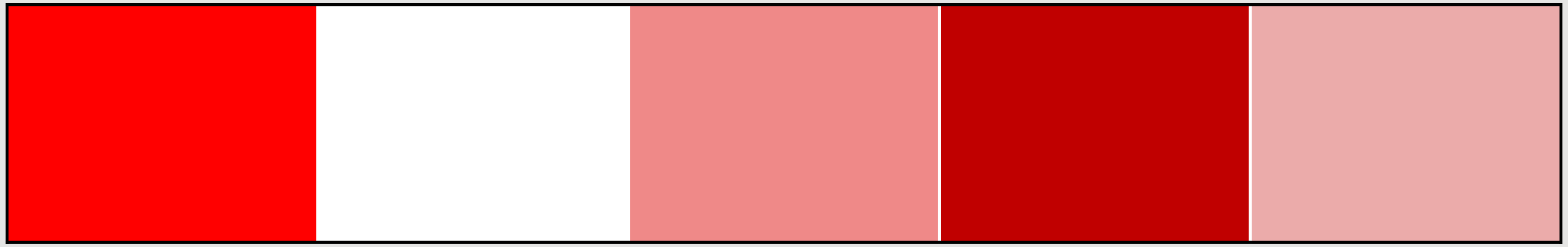
Is it
non-conforming?

ExTuNe

Explanation

Why is it
non-conforming?

Tuple-level explanation



UMass Amherst | College of Information & Computer Sciences

Tuple-level explanation

Height	Weight	BMI
6 feet	142 lbs	19.3
5 feet	170 lbs	33.2
5 feet	130 lbs	25.4
6 feet	170 lbs	231

Tuple-level explanation

Height	Weight	BMI
6 feet	142 lbs	19.3
5 feet	170 lbs	33.2
5 feet	130 lbs	25.4
6 feet	170 lbs	231

Intervention reveals causality

Height	Weight	BMI
6 feet	142 lbs	19.3
5 feet	170 lbs	33.2
5 feet	130 lbs	25.4
6 feet	170 lbs	25.9

Mean = 25.9

change

Intervention reveals causality

Height	Weight	BMI
6 feet	142 lbs	19.3
5 feet	170 lbs	33.2
5 feet	130 lbs	25.4
6 feet	170 lbs	25.9

Blame!

Intervention reveals causality

Blame!

Height	Weight	BMI
6 feet	142 lbs	19.3
5 feet	170 lbs	33.2
5 feet	130 lbs	25.4
16 feet	70 lbs	25.9

Blame!

ExTuNe principles: actual causality

Actual Causality

"When K other events are removed, then C is a counterfactual cause of E"

- C is an actual cause of E
- C's responsibility is 1/(K + 1)

ExTuNe interface

Upload reference data — Select (1) — 1

Learn Data Invariants — Learn — 2

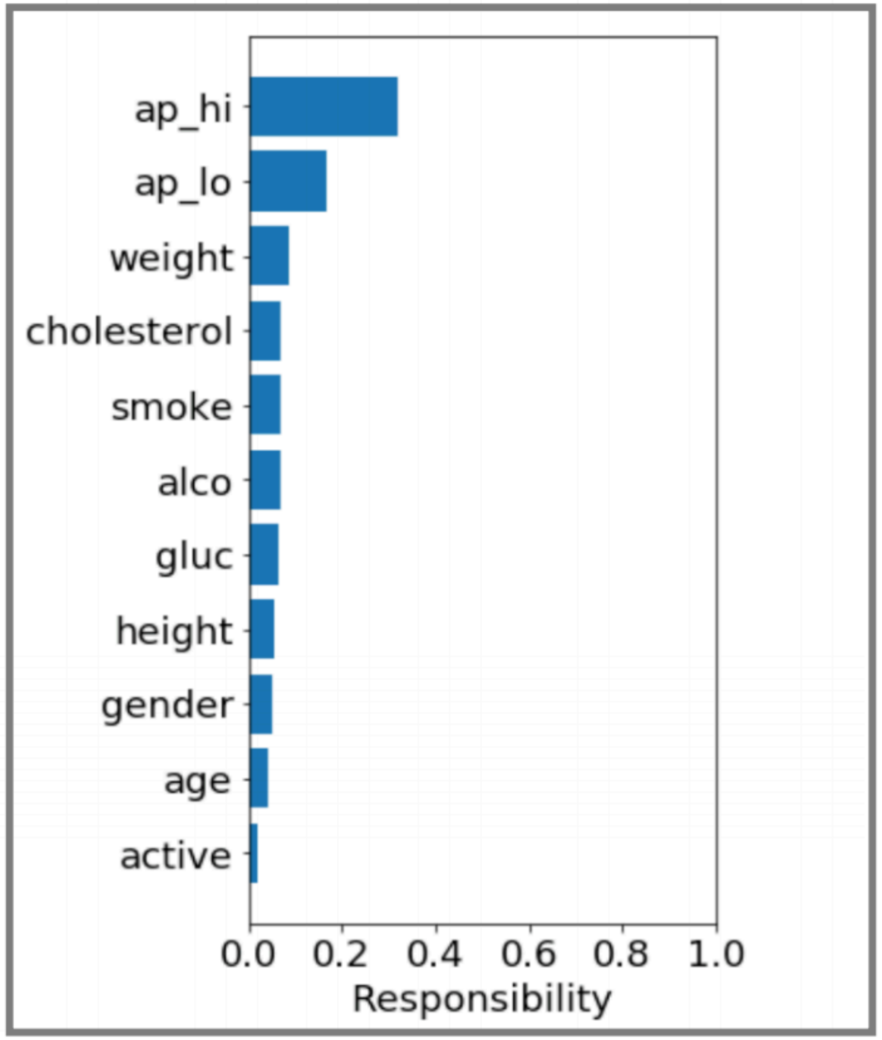
Upload test data: — Select (1) — 3

10 most non-conforming tuples — Preview — 4

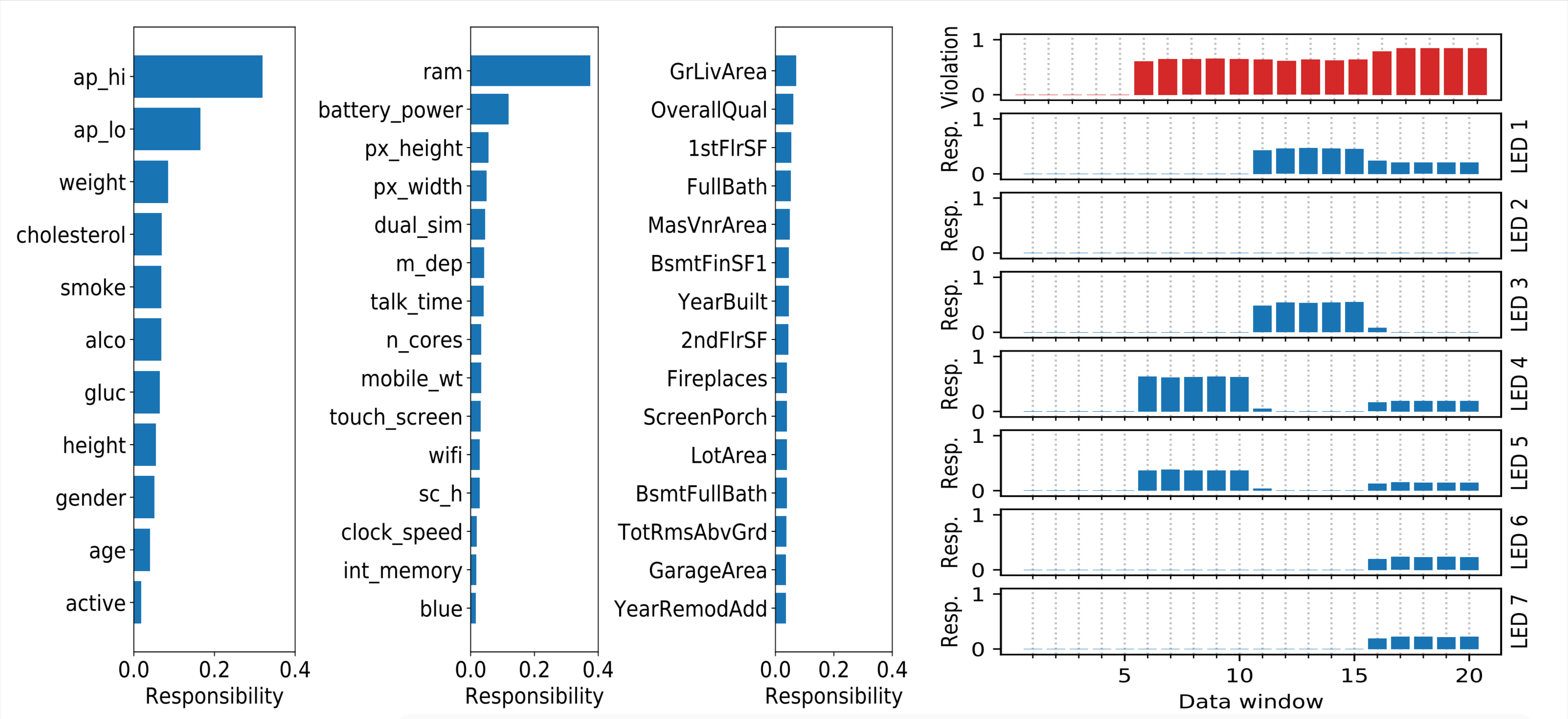
5

id	Violation	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active
21321	0.27	22652	2	163	70	200	180	1	1	0	0	0
18398	0.21	21770	1	161	84	196	182	2	2	0	0	1
2318	0.2	18961	1	158	74	200	170	1	1	0	0	1
8643	0.15	15086	2	190	165	160	60	1	1	0	0	0
15682	0.15	21300	2	176	98	240	110	1	1	1	0	1
29677	0.15	16615	2	196	180	130	80	3	1	0	0	1
14276	0.14	19798	2	174	75	240	120	2	1	0	1	1
24399	0.14	15117	2	180	200	150	90	1	1	0	0	1
17058	0.13	21897	1	152	87	220	110	3	3	0	0	0
10556	0.13	21721	2	173	116	190	80	1	3	1	1	0

Show aggregated attribute-responsibility — Show — 6



ExTuNe evaluation: case studies



Cardiovascular disease

Mobile price

House price

LED malfunctioning

Anomaly in COVID dataset

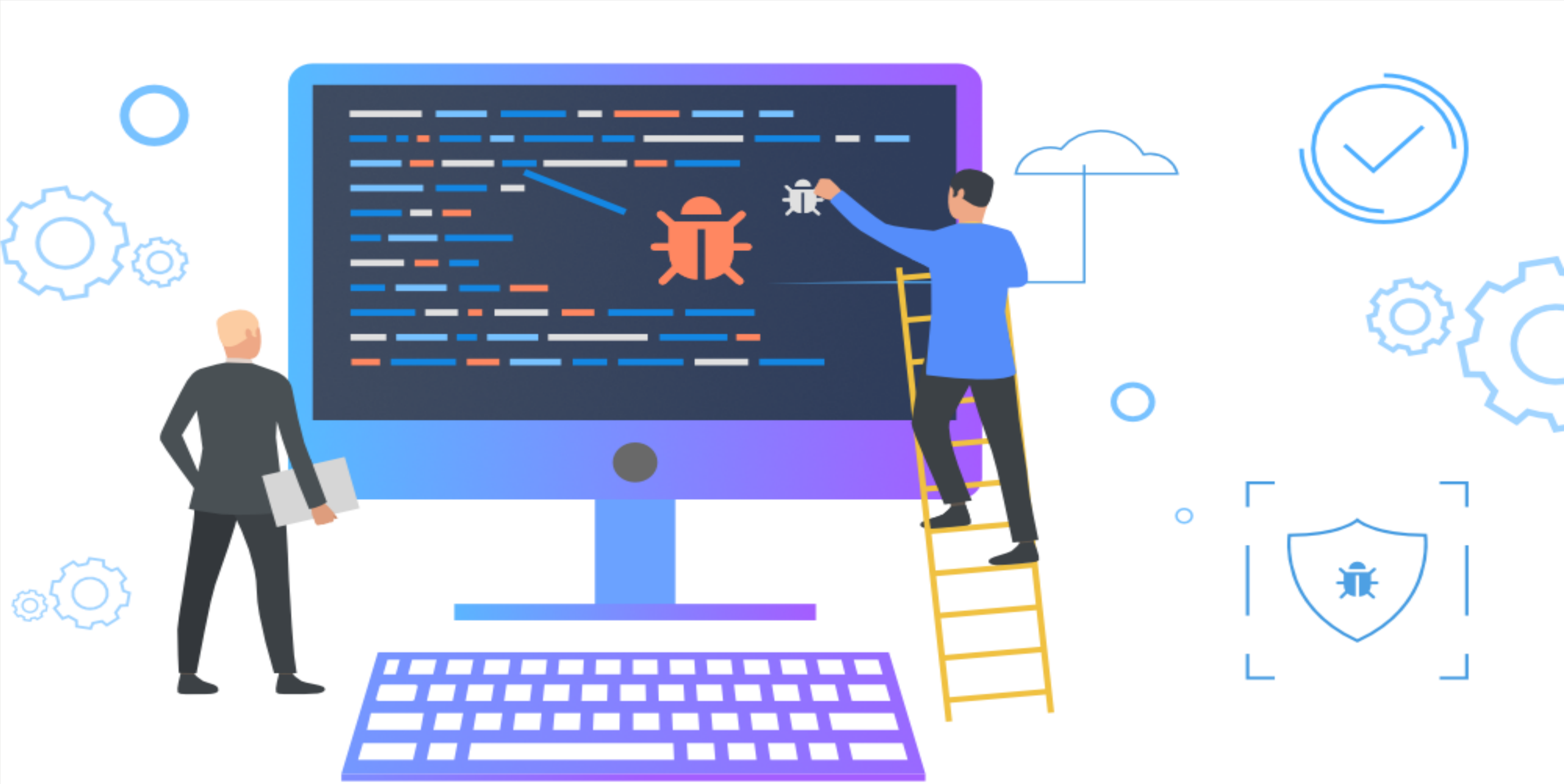
Conformance constraint: #positive + #negative = #total

	Violation	date	state	positive	negative	pending	hospitalized	death	total	population	hospital_beds
201	0.350000	20200321	NY	10356	35081	0	1603	44	45437	19453561	52524
36	0.350000	20200324	NY	25665	65605	0	3234	210	91270	19453561	52524
256	0.290000	20200320	NY	7102	25325	0	0	35	32427	19453561	52524
4	0.240000	20200324	CA	2102	13452	12100	0	40	27654	39512223	71122
59	0.240000	20200323	CA	1733	12567	12100	0	27	26400	39512223	71122
311	0.220000	20200319	NY	4152	18132	0	0	12	22284	19453561	52524
91	0.200000	20200323	NY	20875	57414	0	2635	114	78289	19453561	52524
88	0.190000	20200323	NJ	2844	359	94	0	27	3297	8882190	21317
471	0.130000	20200316	NJ	178	120	20	0	2	218	8882190	21317
389	0.130000	20200317	CA	483	7981	0	0	11	8407	39512223	71122
51	0.130000	20200324	WA	2221	31712	0	0	110	33933	7614893	12945

178 + 120
= 298

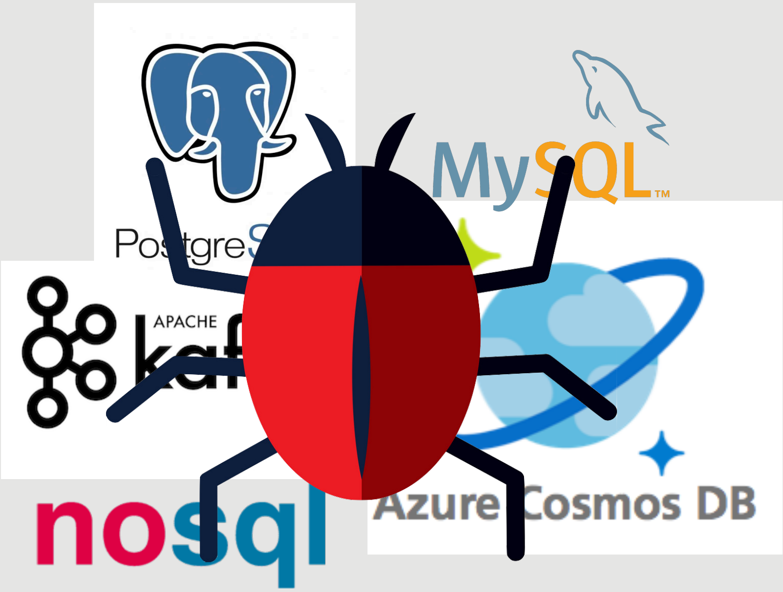
178 + 120
≠ 218

Explaining data systems' failure

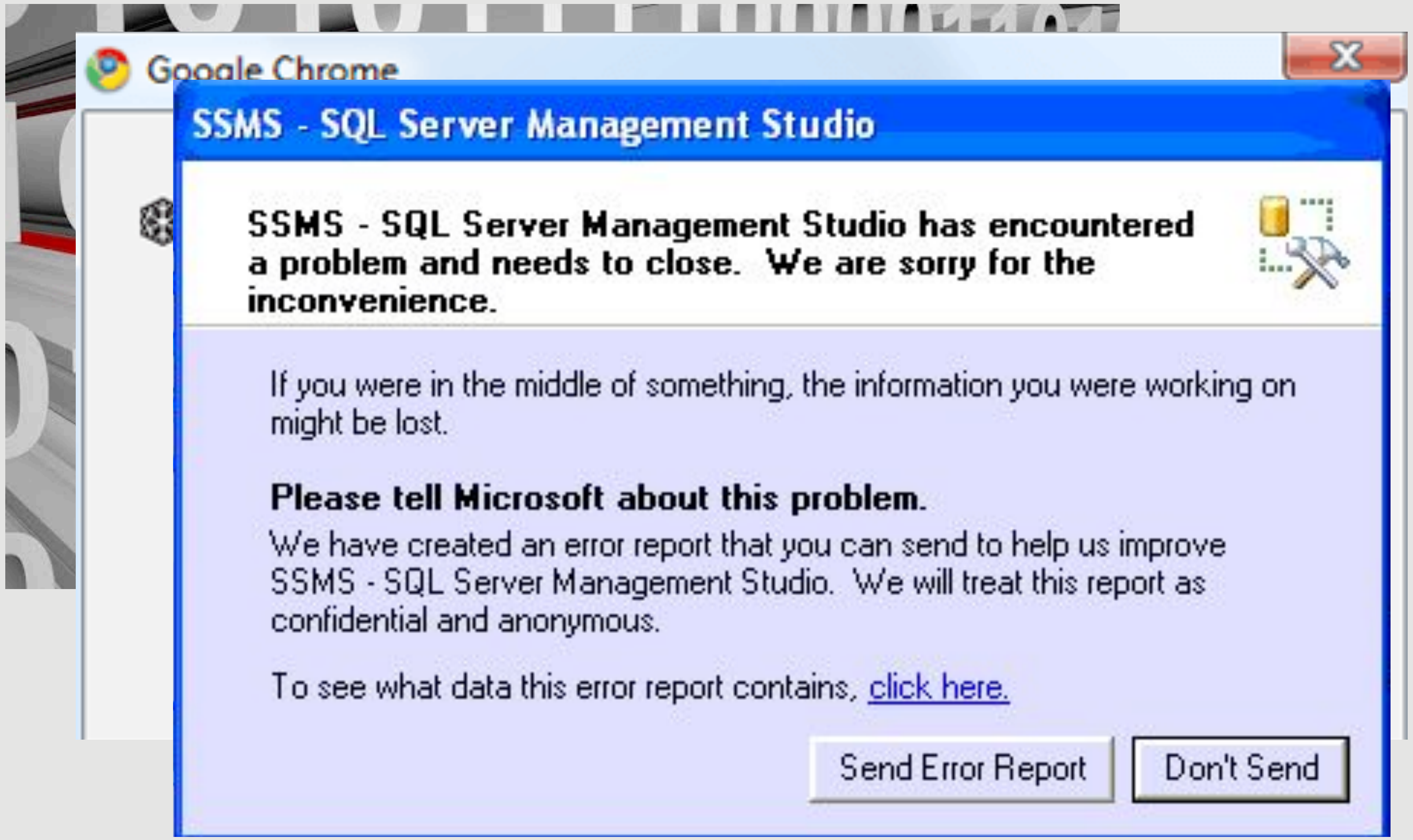


AID: Causality-guided Adaptive Interventional Debugging

DBMS are complex and contain bugs

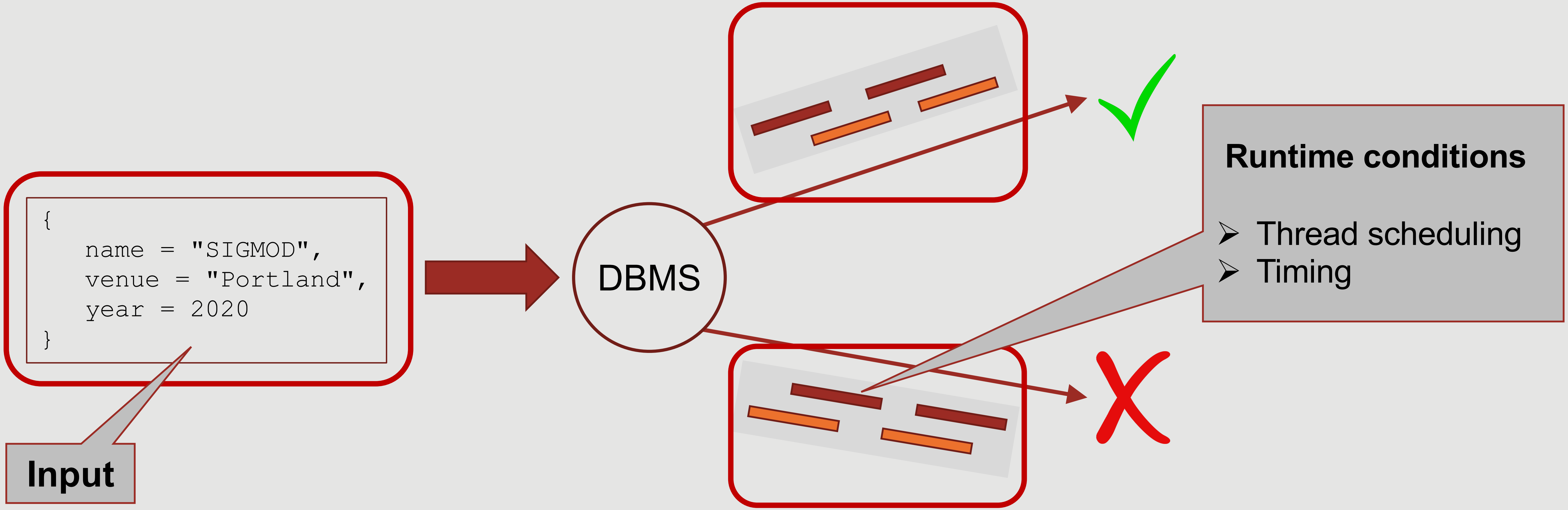


- concurrent
- parallel
- asynchronous



Intermittent failure

sometimes succeeds sometimes fails



Motivation and goal

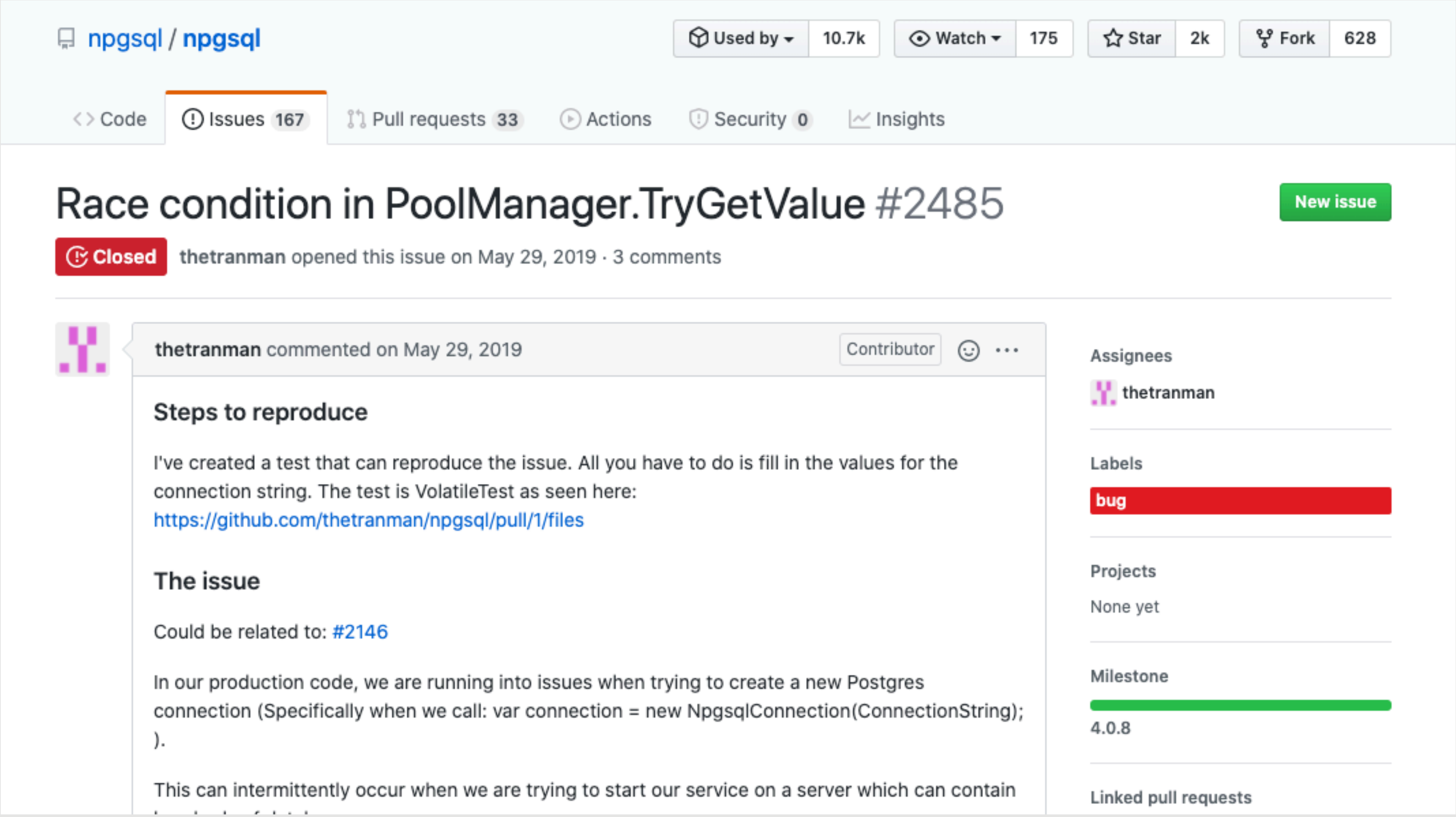
Can't reproduce!

Help me debug!

Investigate root causes of intermittent failure

Npgsql intermittent failure

[ADO.NET data provider for PostgreSQL]


npgsql / npgsql

Used by 10.7k | Watch 175 | Star 2k | Fork 628

Code | Issues 167 | Pull requests 33 | Actions | Security 0 | Insights

Race condition in PoolManager.TryGetValue #2485

New issue

Closed thetranman opened this issue on May 29, 2019 · 3 comments

thetranman commented on May 29, 2019 Contributor

Steps to reproduce

I've created a test that can reproduce the issue. All you have to do is fill in the values for the connection string. The test is VolatileTest as seen here: https://github.com/thetranman/npgsql/pull/1/files

The issue

Could be related to: #2146

In our production code, we are running into issues when trying to create a new Postgres connection (Specifically when we call: var connection = new NpgsqlConnection(ConnectionString);).

This can intermittently occur when we are trying to start our service on a server which can contain

Assignees

thetranman

Labels

bug

Projects

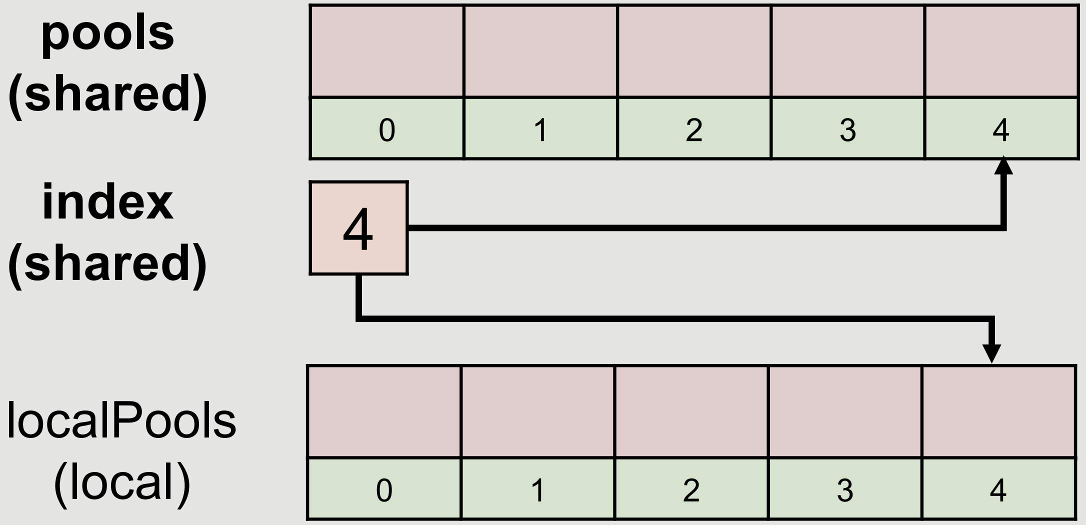
None yet

Milestone

4.0.8

Linked pull requests

Npgsql intermittent failure



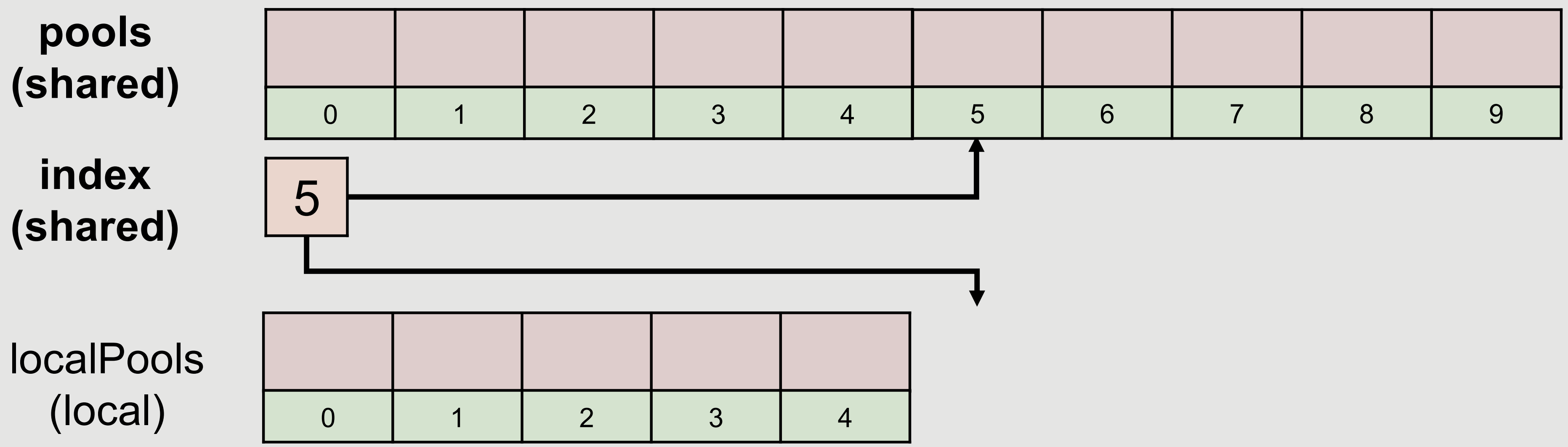
Thread 1

```
Find(key):
1  localPools = pools
```

Thread 2

```
Add(key):
```

Npgsql intermittent failure



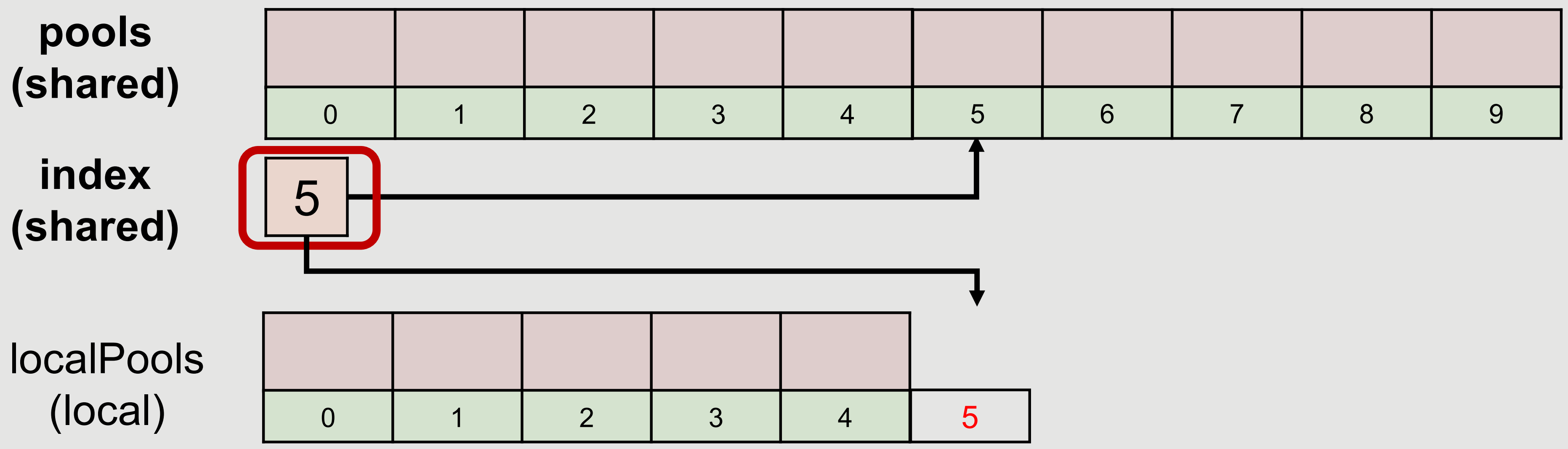
Thread 1

```
Find(key):
1  localPools = pools
```

Thread 2

```
Add(key):
   if pools_is_filled:
       pools = ResizeDouble(pools)

   last_slot ++
   pools[last_slot] = key          2
```

Npgsql intermittent failure



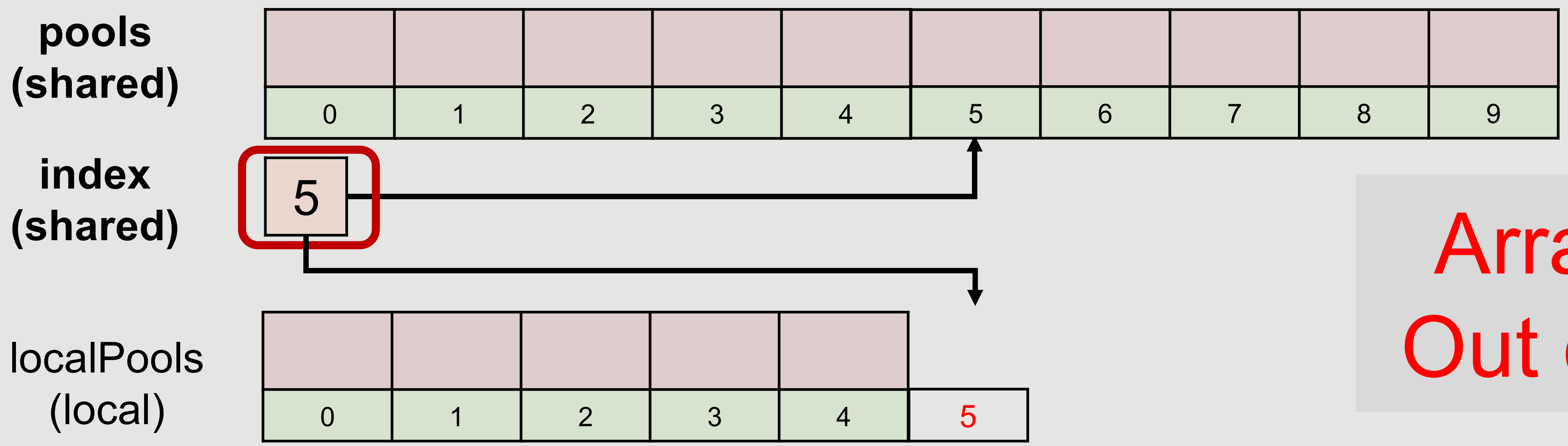
Thread 1

```
Find(key):
1  localPools = pools

3  for i in range(0,last_slot+1):
       if (localPools[i] == key)
           return i

   return null
```

Thread 2

```
Add(key):
   if pools_is_filled:
       pools = ResizeDouble(pools)

   last_slot ++
   pools[last_slot] = key          2
```

Npgsql intermittent failure



Array Index Out of Bound

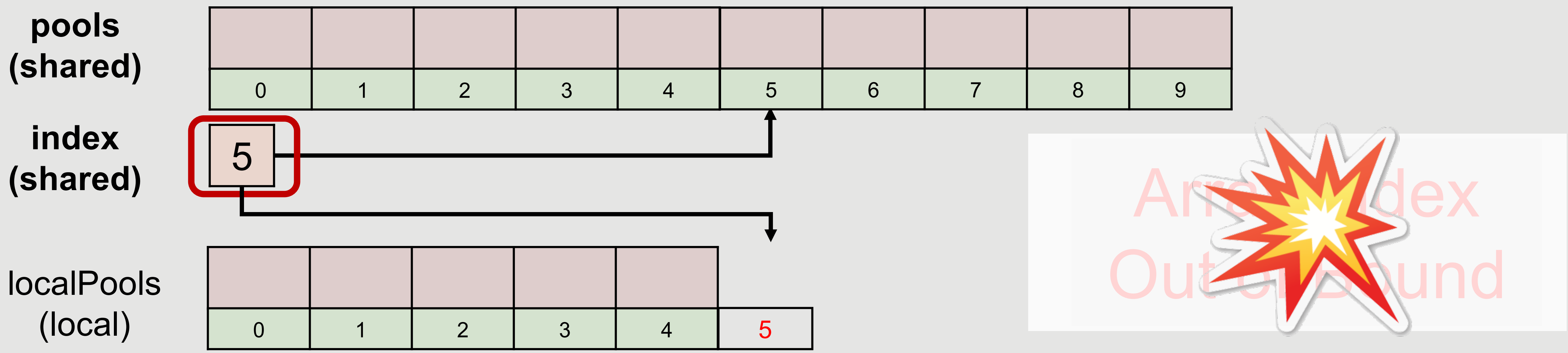
Thread 1

```
Find(key):
1  localPools = pools

3  for i in range(0,last_slot+1):
       if (localPools[i] == key)
           return i

   return null
```

Thread 2

```
Add(key):
   if pools_is_filled:
       pools = ResizeDouble(pools)
                                        2
   last_slot ++
   pools[last_slot] = key
```

Npgsql intermittent failure



Thread 1

```
Find(key):
1  localPools = pools

3  for i in range(0,last_slot+1):
       if (localPools[i] == key)
           return i

   return null
```

Thread 2

```
Add(key):
   if pools_is_filled:
       pools = ResizeDouble(pools)

   last_slot ++
   pools[last_slot] = key              2
```

Npgsql intermittent failure

1

```
localPools = pools
```

```
if pools_is_filled:
    pools = ResizeDouble(pools)

last_slot ++
pools[last_slot] = key
```

2

3

```
for i in range(0,last_slot+1):
    if (localPools[i] == key)
        return i

return null
```

Investigating Npgsql crash

Root cause: `Add()` temporally overlaps with `Find()`

Explanation: `Find()` attempts to access invalid array index

Explanation: `Find()` throws ArrayIndexOutOfBound exception

Failure:

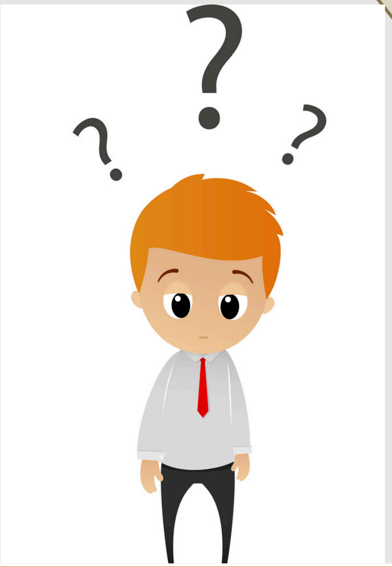
Limitations of statistical debugging

Find() is running too slow

Find() and Get() temporally overlaps

...

Get() is running too fast



crash

Our goals

Root-cause identification

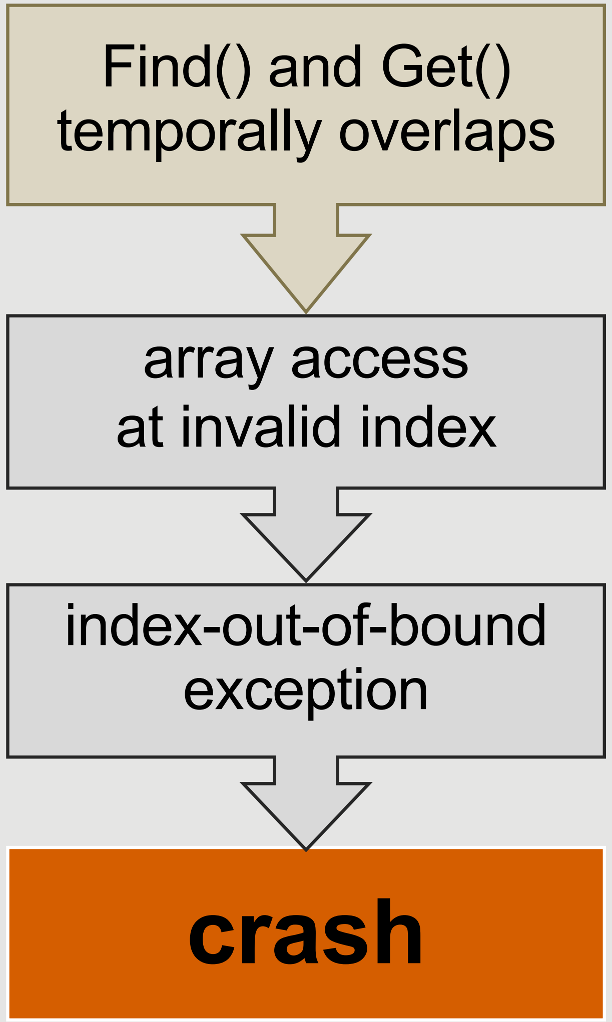
Find() and Get() temporally overlaps

crash

Our goals

Root-cause identification

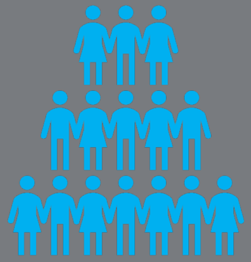
Explanation



statistical debugging

causality

fault injection

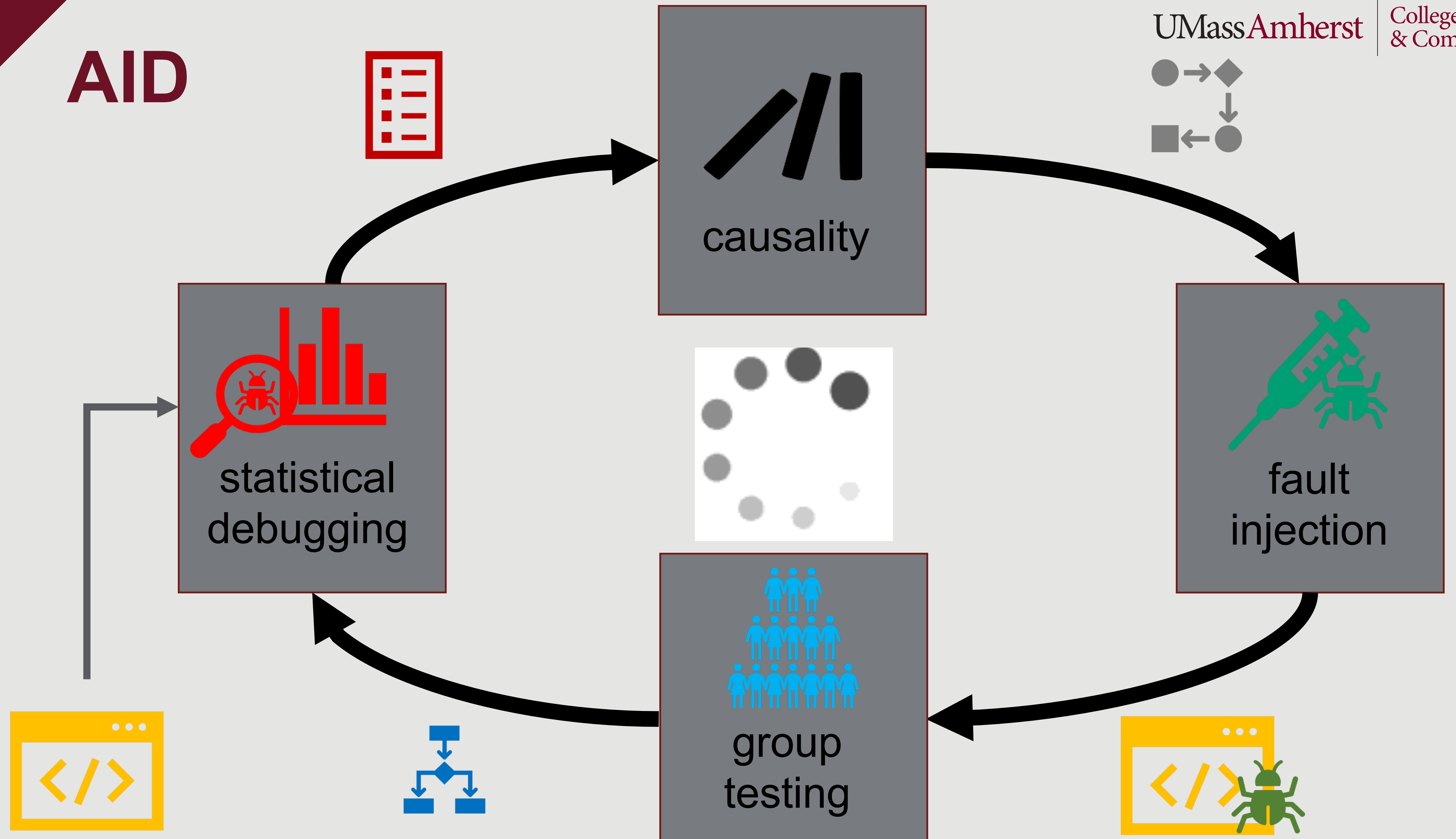


group testing

AID: Adaptive Interventional Debugging

AID

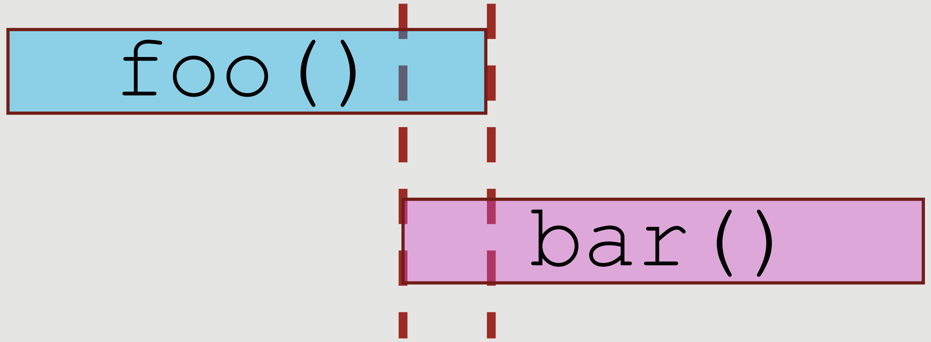
UMass Amherst | College of Information & Computer Sciences



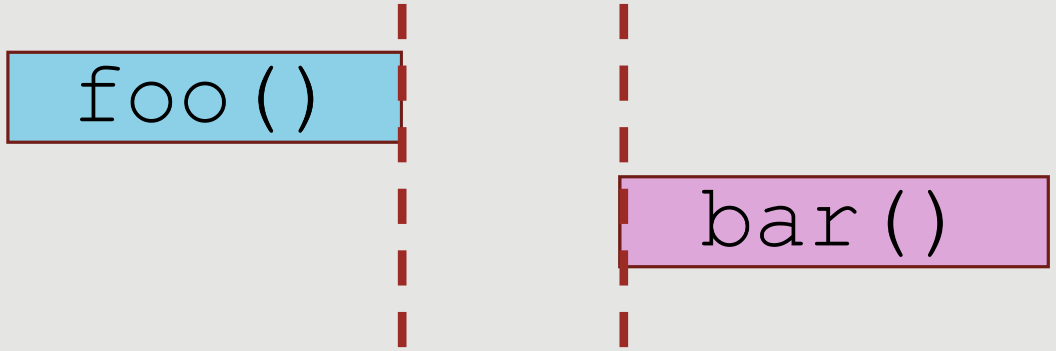
Finding candidate predicates

- Step 1: Program instrumentation finds all predicates
- Step 2: Statistical debugging finds correlated predicates

Always appear in **failed** executions

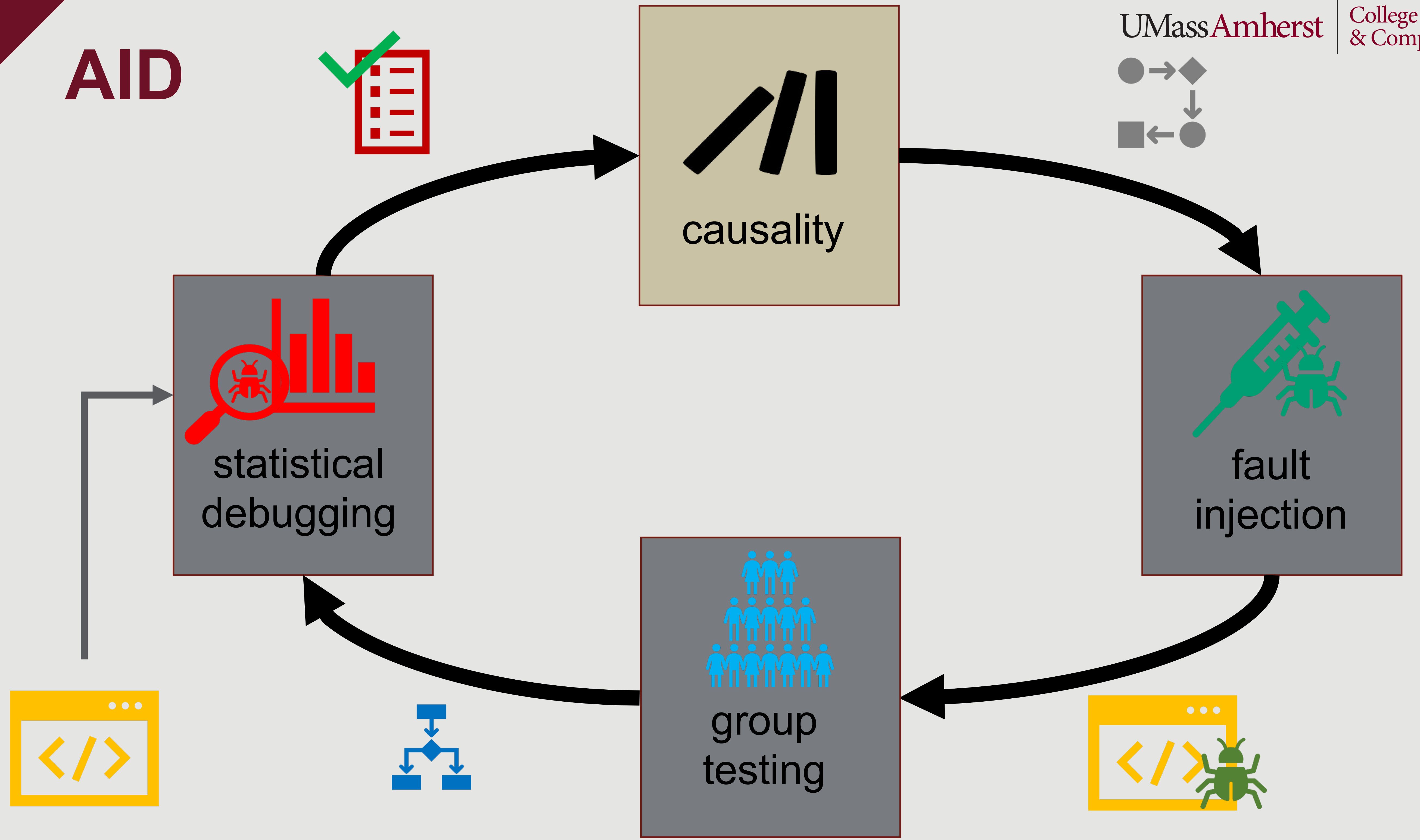


Never appear in **successful** executions



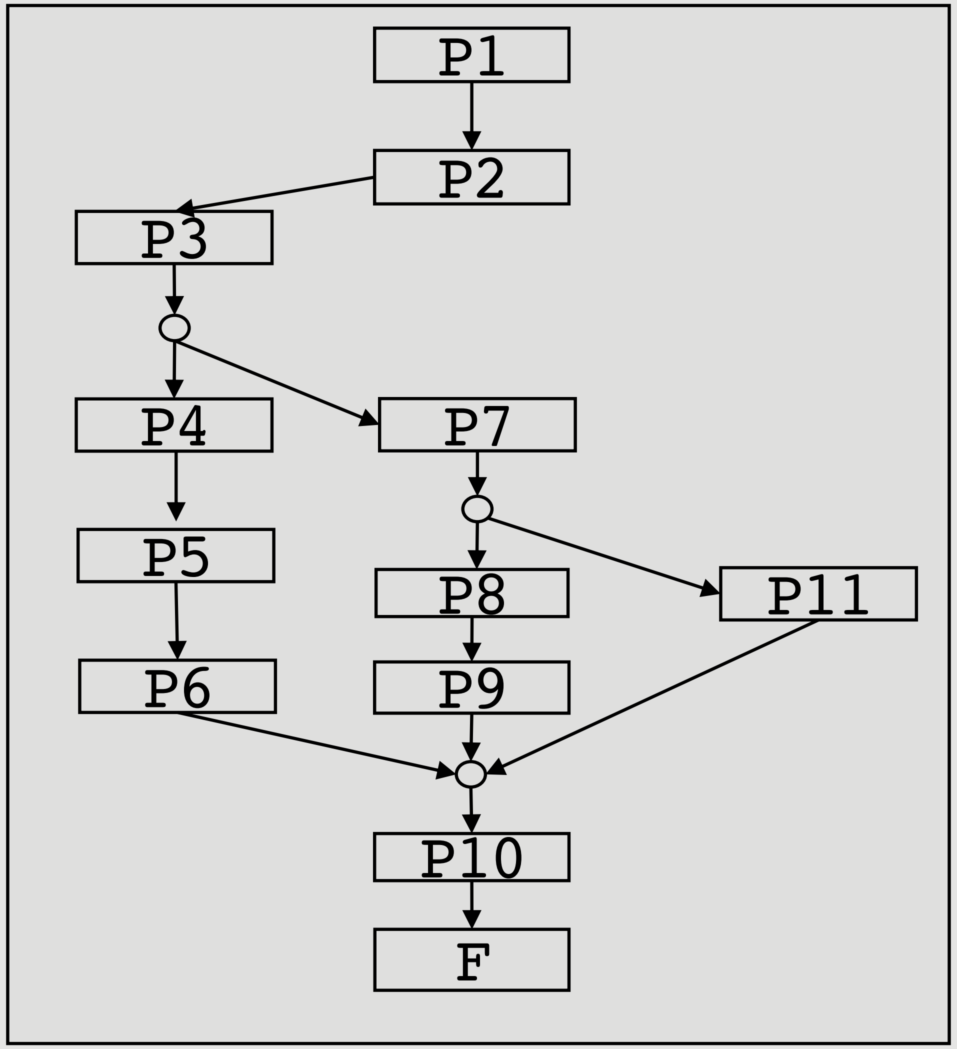
AID

UMass Amherst | College of Information & Computer Sciences

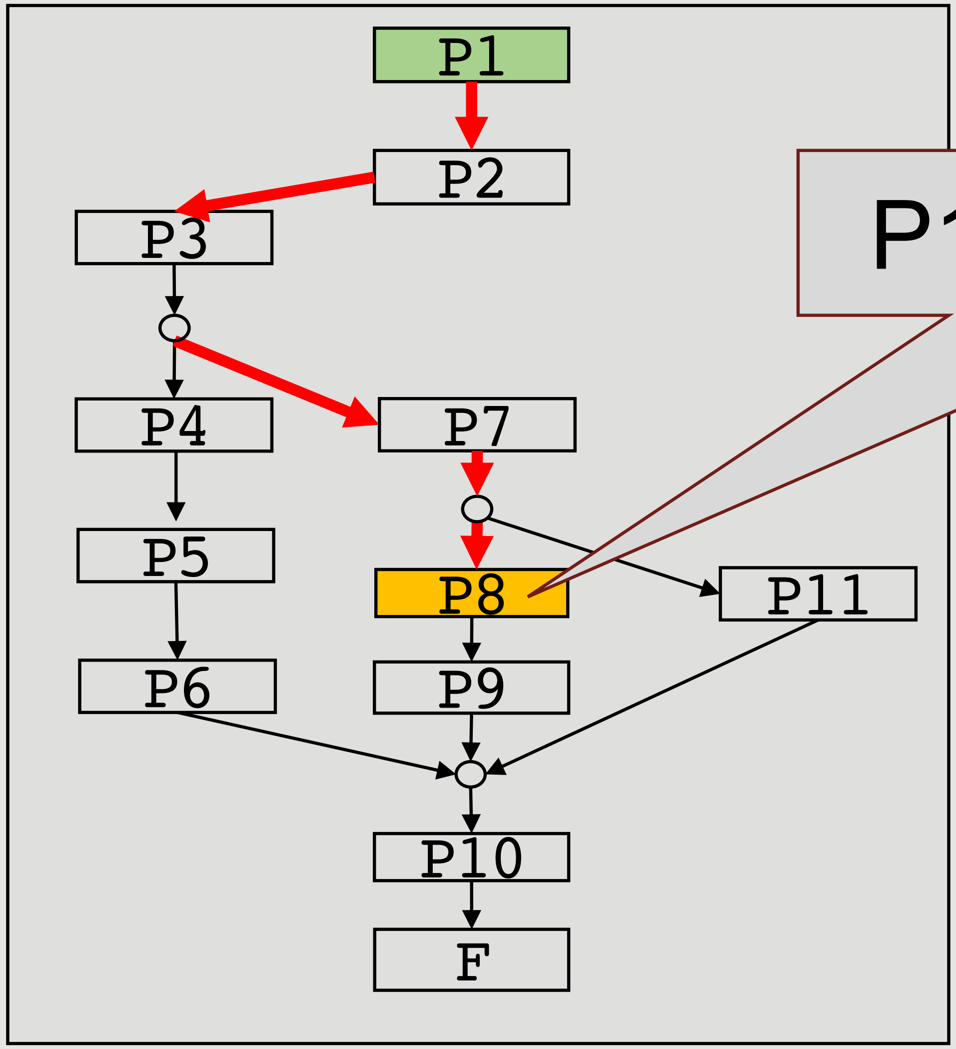


Cause must **temporally precede** effect

Temporal precedence graph

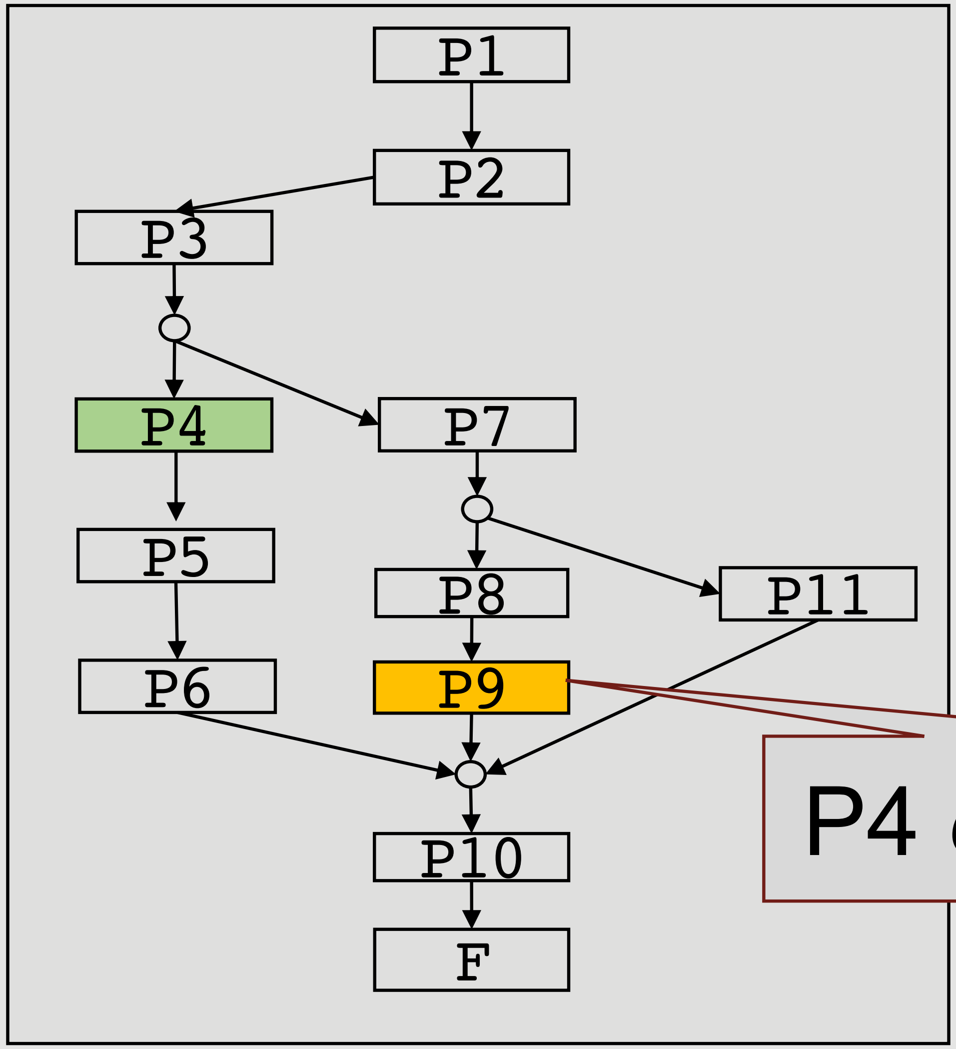


Approximating causality



P1 ***may*** cause P8

Approximating causality



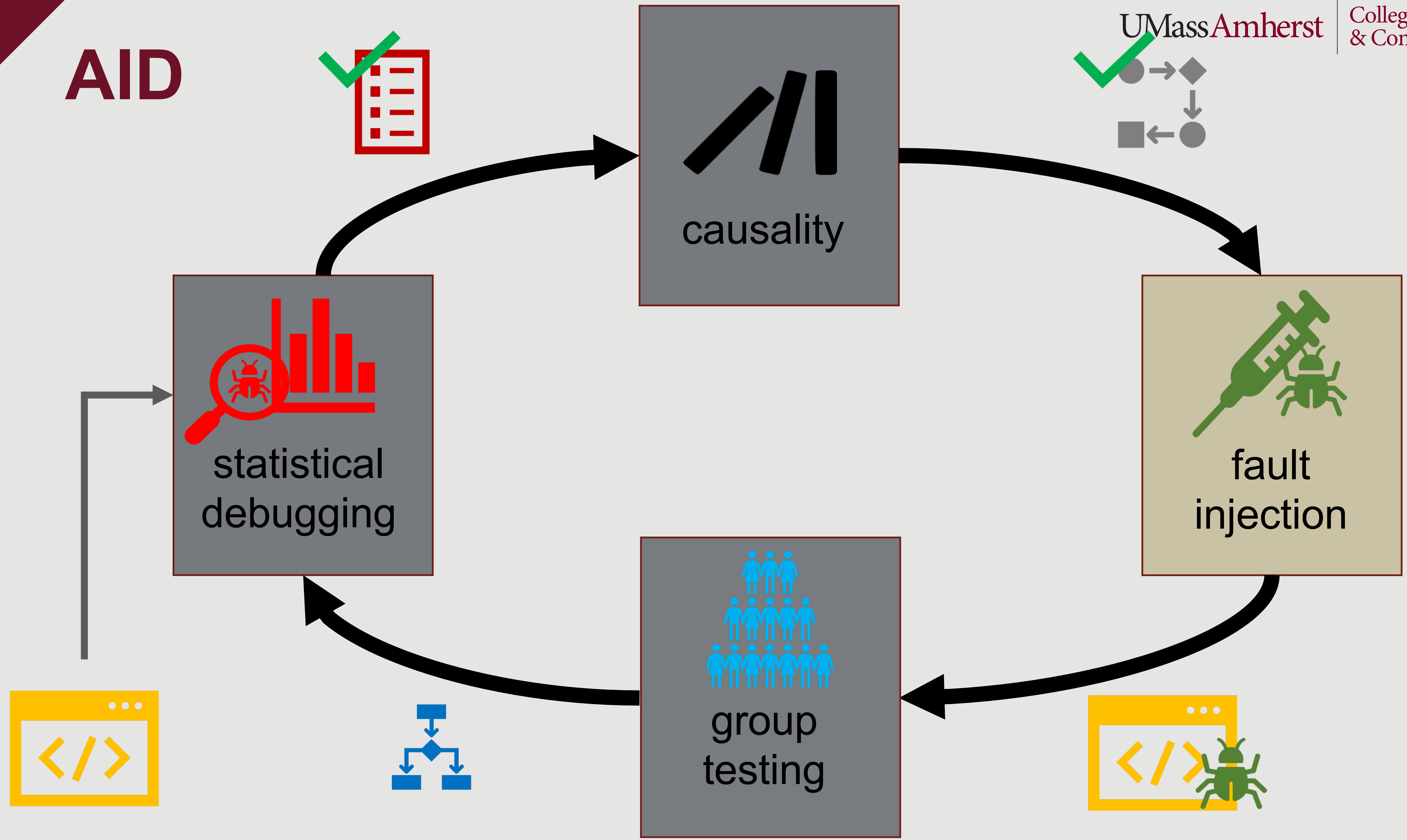
P4 ***cannot*** cause P9

Counterfactual causality

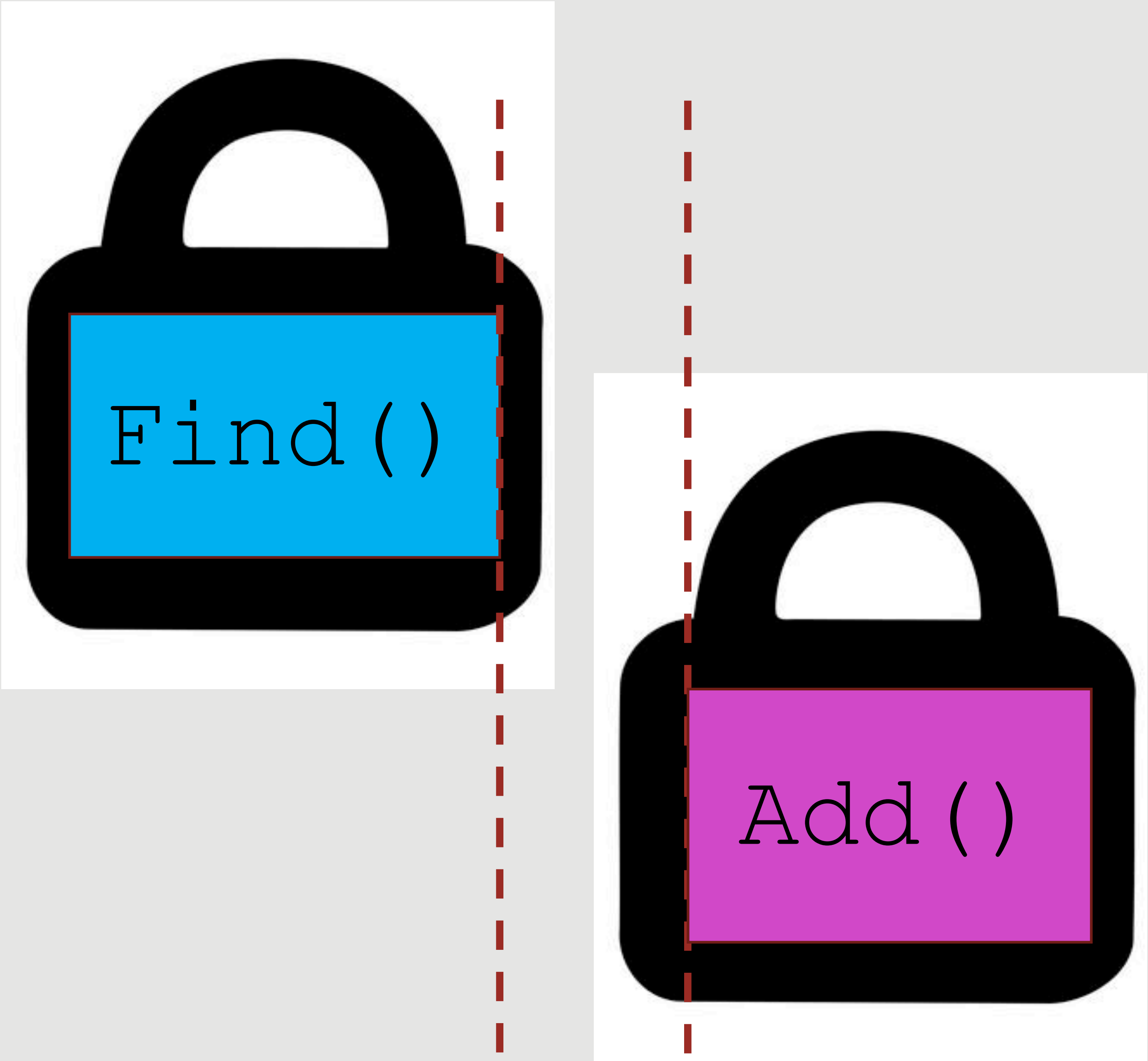
C is a ***counterfactual cause*** of E
If C **had not** occurred
E **would not** have occurred

Intervention

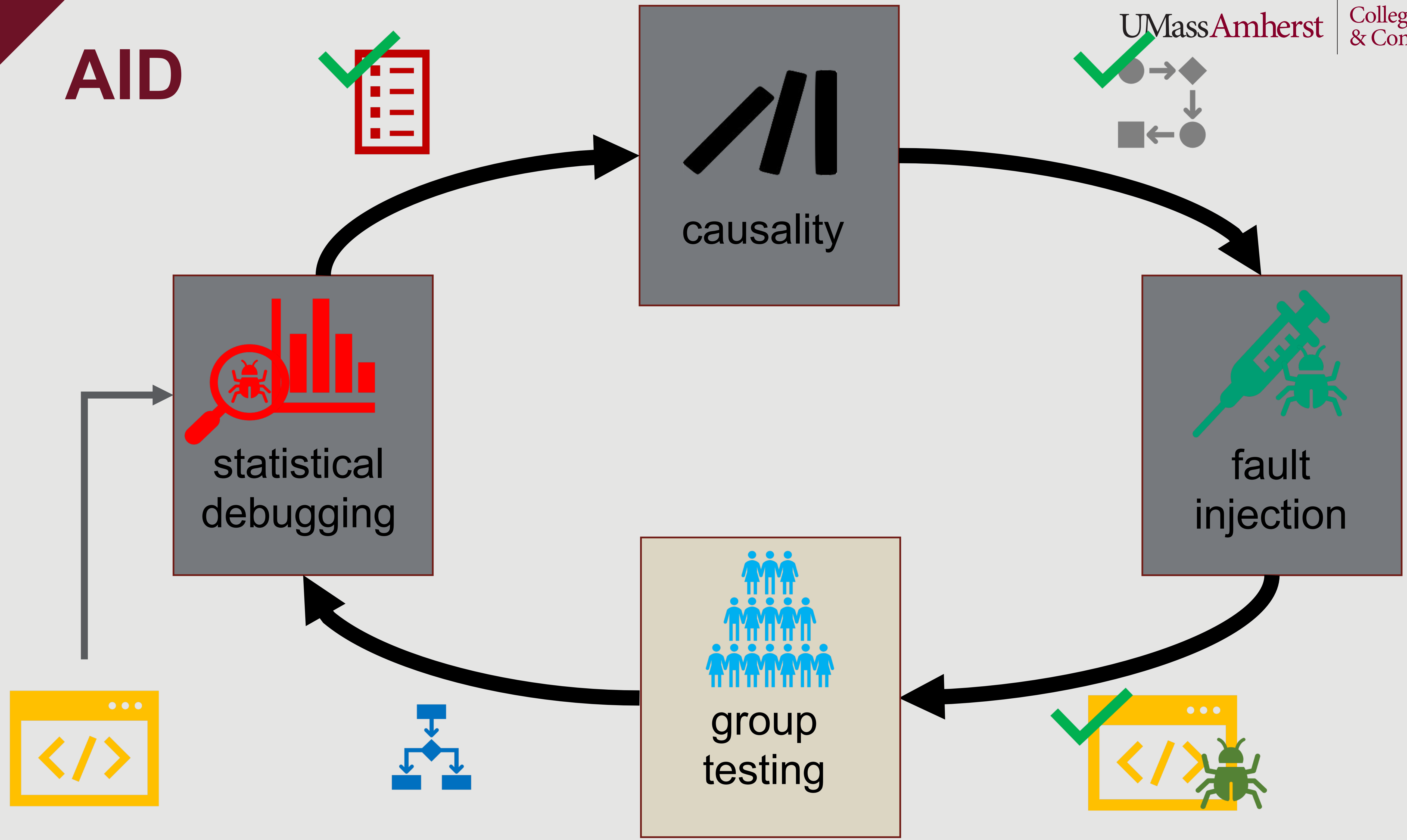
AID


causality
fault injection
group testing
statistical debugging

Fault injection



AID


causality
fault
injection
group
testing
statistical
debugging

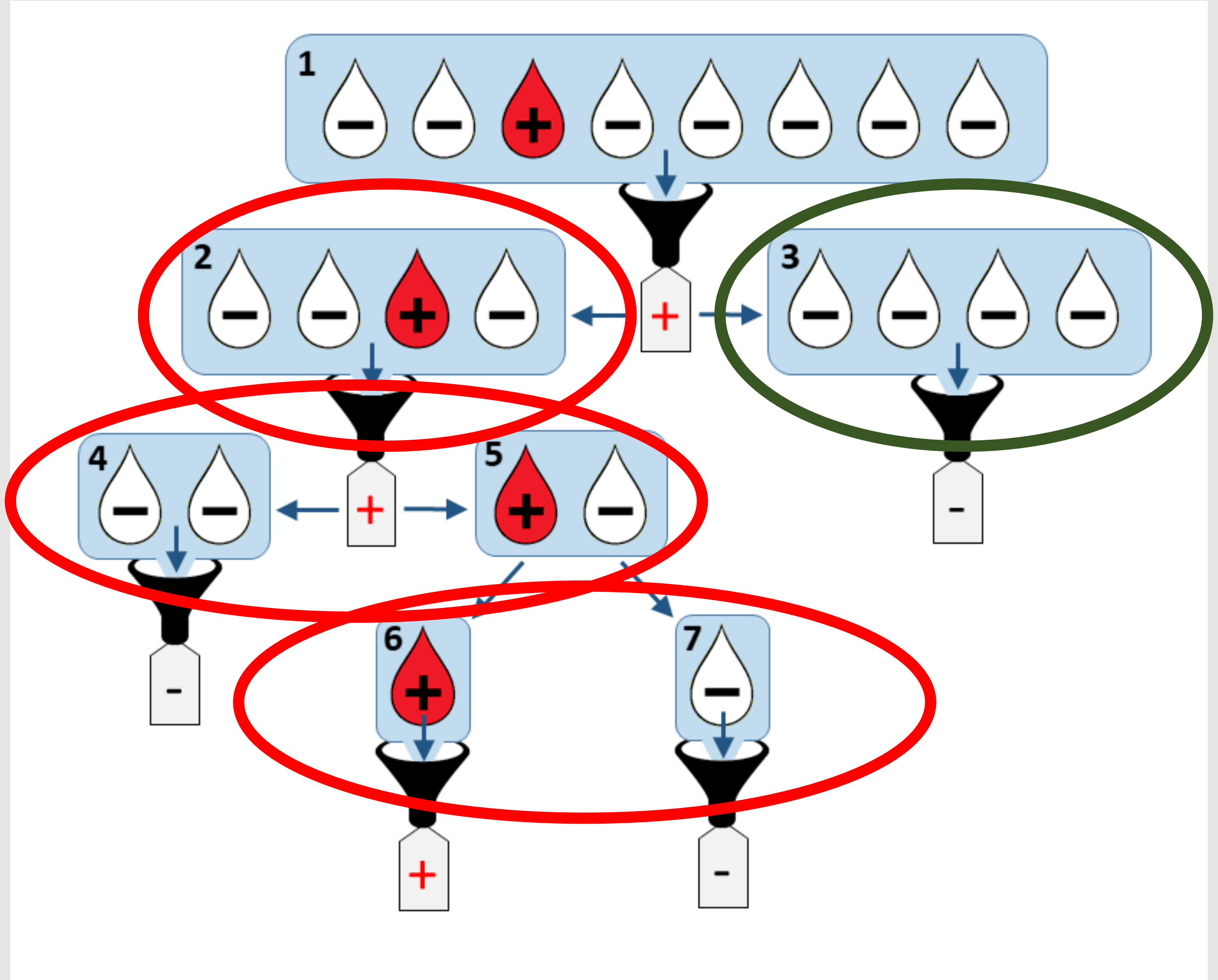
Group testing

PUBLIC HEALTH

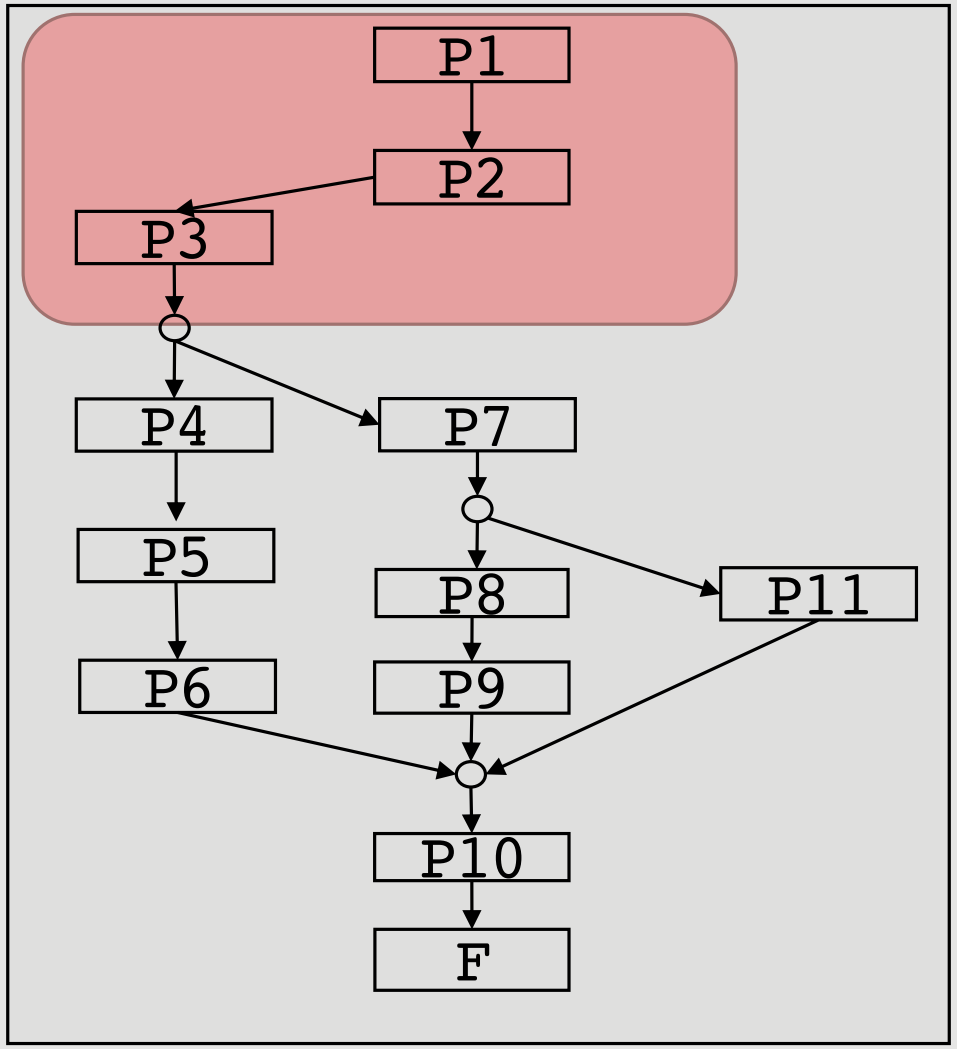
Coronavirus Test Shortages Trigger a New Strategy: Group Screening

Pooling diagnostic samples, and using a little math, lets more people get tested with fewer assays

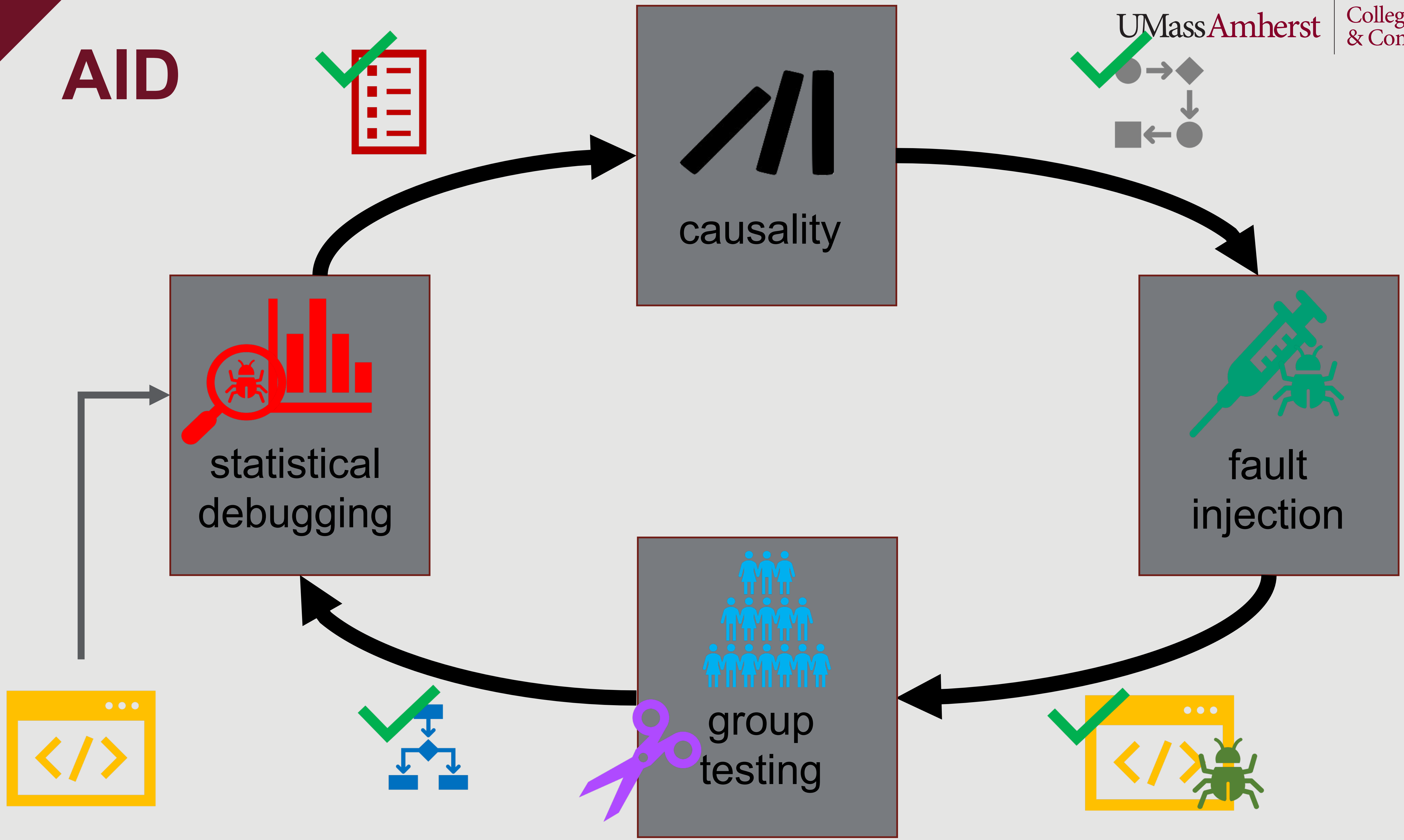
Adaptive group testing



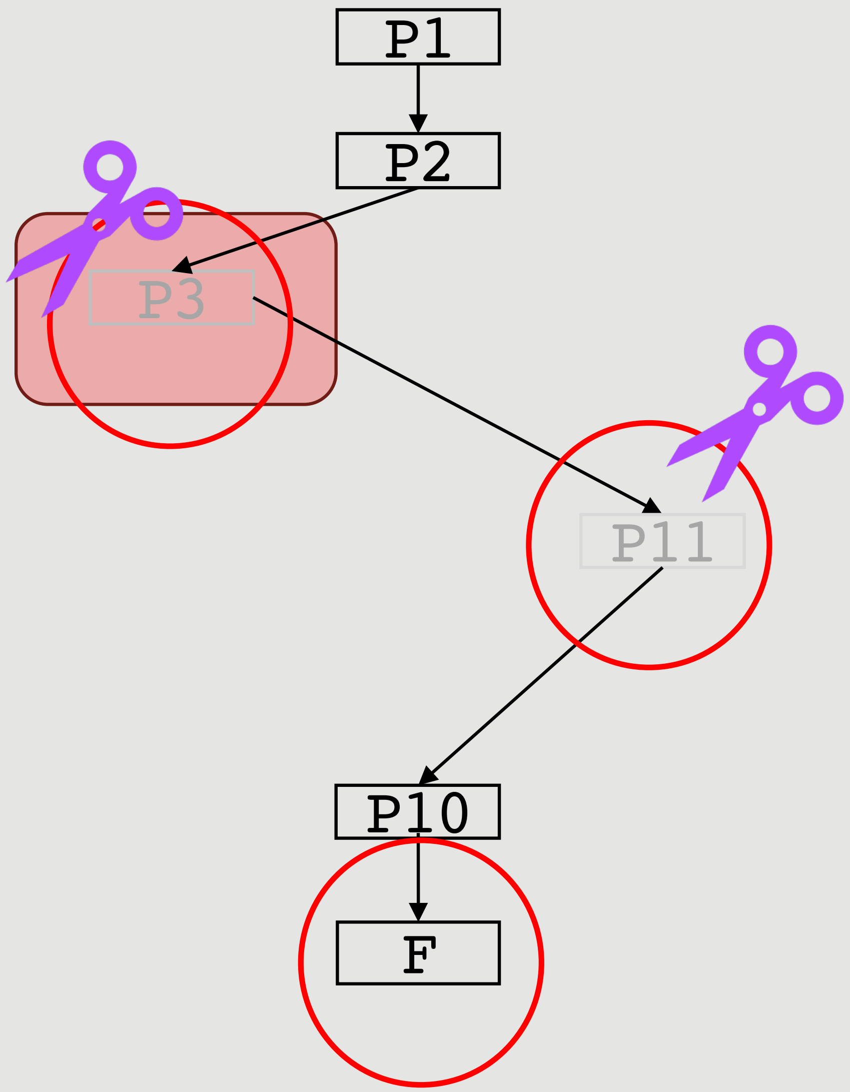
AID applies group intervention



AID


causality
fault
injection
group
testing
statistical
debugging
</>

AID pruning




EVALUATION

Six real-world bugs

Apache Kafka — Data race ✓

Azure Cosmos DB — Use-after-free ✓

Npgsql .NET Data Provider for PostgreSQL — Timing-bug ✓

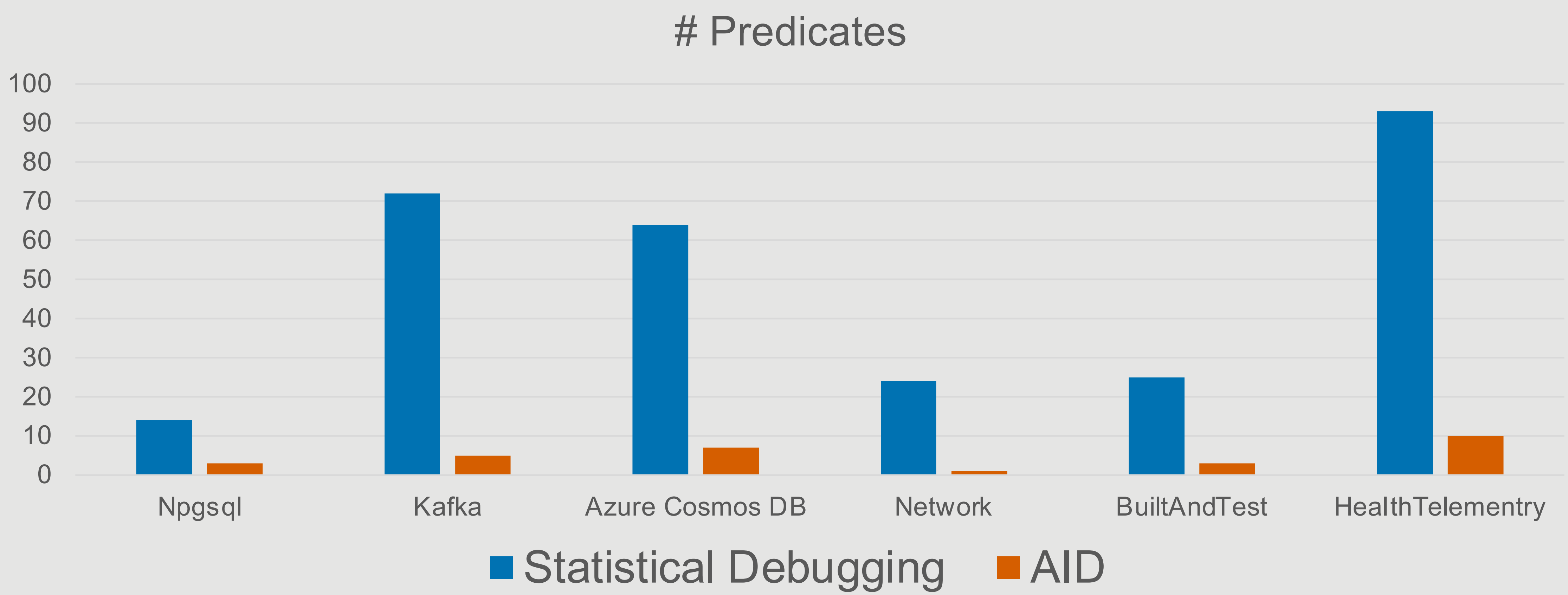
`Network` Microsoft — Random number collision ✓

`BuildAndTest` Microsoft — Order violation ✓

`HealthTelemetry` Microsoft — Race condition ✓

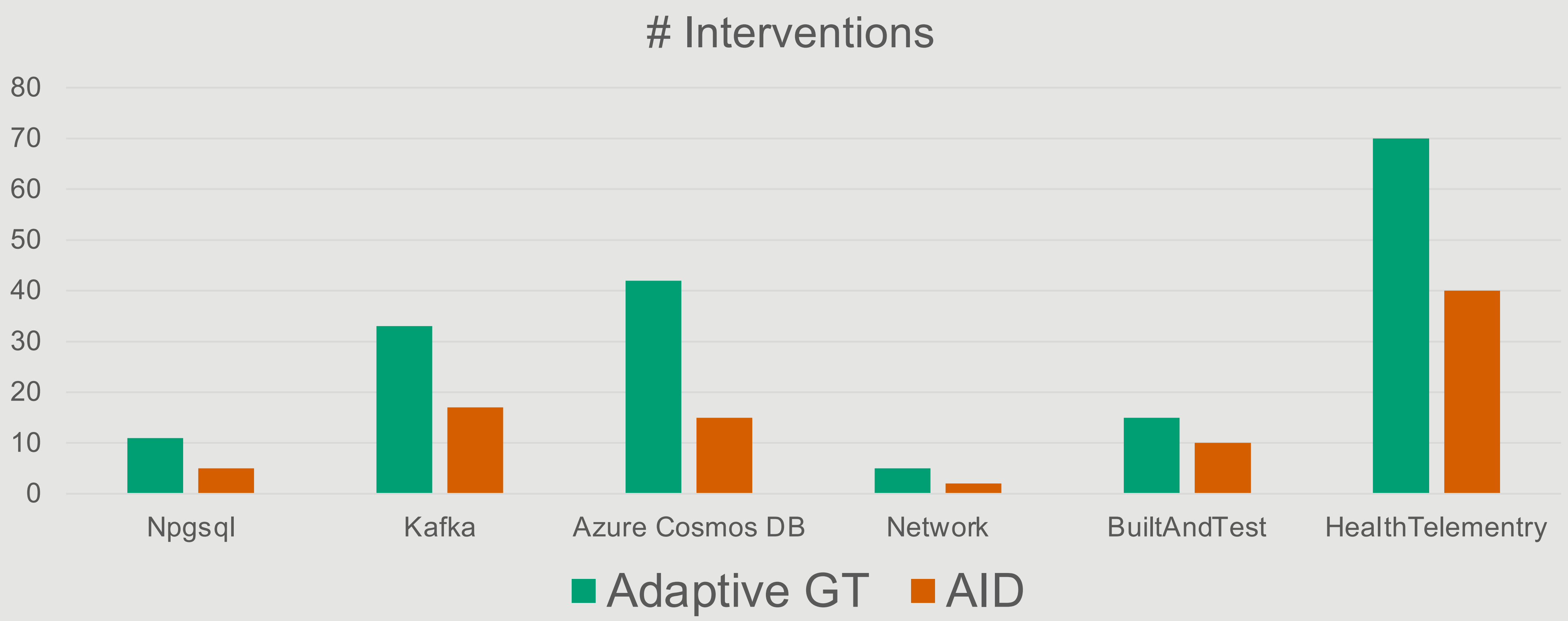
Statistical debugging vs AID

AID produces no false positives



Adaptive group testing vs AID

AID’s pruning reduces #Interventions



Theoretical analyses

CPD: Causal Path Discovery
GT: Group Testing
AID: Adaptive Interventional Debugging
TAGT: Traditional Adaptive Group Testing

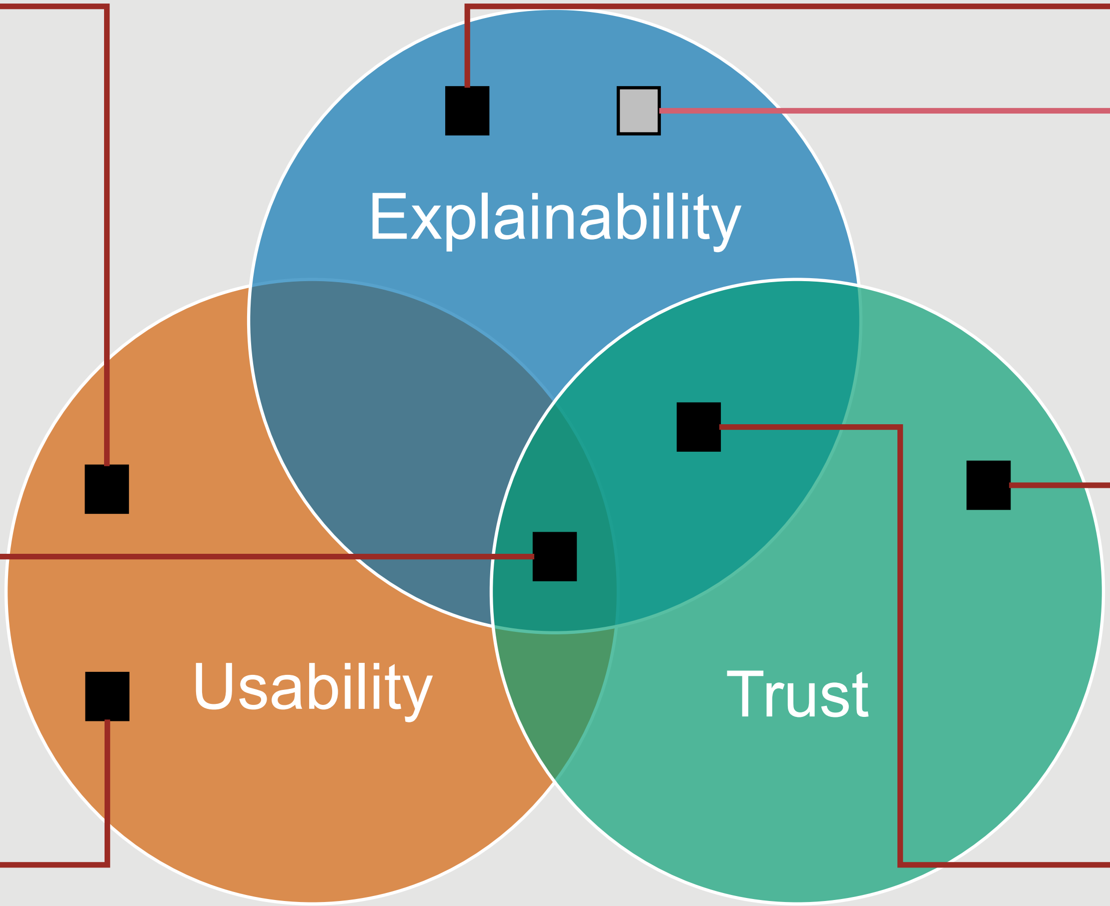
	Search space	#Interventions	
		Lower bound	Upper bound (AID/TAGT)
CPD	$(B(2^n-1)+1)^J$	$\frac{JBn}{JBn+DS_1}\log\binom{JBn}{D}$	$J\log B + D\log(Jn) - \frac{D(D-1)S_2}{2Jn}$
GT	2^{JBn}	$\log\binom{JBn}{D}$	$D\log B + D\log(Jn) - \frac{D(D-1)}{2JBn}$

Dissertation outline

Query by Example (QBE)
[VLDB 2019]
[SIGMOD 2018] (demo)

Comparative User Study: QBE vs SQL
[CHI 2020]*

Data Summarization by Example
[VLDB 2020] (demo)



Adaptive Interventional Debugging
[SIGMOD 2020]

Data Change Explanation
WORK IN PROGRESS

Conformance Constraints: Trusted ML
[SIGMOD 2021]*

Explaining Tuple Non-conformance
[SIGMOD 2019] (demo)

* under submission/revision

Part 4: Proposed Contributions & Tentative Timeline

Data Change Explanation

UMass Amherst | College of Information & Computer Sciences

How did my data change over last couple years?

```
15:50:23,,0.5,69,,11425,,,,"271504218477G",32,,,,,,,,11,,,,,,,
- 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,41,69,32,2018-10-10
15:50:23,,0.5,69,,11428,,,,"271507491568G",32,,,,,,,,11,,,,,,,
- 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,41,69,32,2018-10-10
15:50:23,,0.5,69,,11484,,,,"271508481857G",32,,,,,,,,11,,,,,,,
- 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,77,69,53,2018-10-11
11:35:05,,0.05,134,,11447,,,,"874231098887G",53,,,,,,,,11,,,,,,,
- 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,77,69,53,2018-10-11
11:35:05,,0.05,134,,11448,,,,"874231135374G",53,,,,,,,,11,,,,,,,
- 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,77,69,53,2018-10-11
11:35:05,,0.05,134,,11479,,,,"874231461234G",53,,,,,,,,11,,,,,,,
- 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,87,69,59,2018-10-11
13:43:24,,0.05,34,,11487,,,,"874231676529G",59,,,,,,,,11,,,,,,,
- 73,"FIVE,PERSON",,FIVE,PERSON,,,,,,Y,G,23,73,19,2018-10-08
22:25:59,,0.75,73,,14,,,,"271508486757G",19,,,,,,,,11,,,,,,,
- 73,"FIVE,PERSON",,FIVE,PERSON,,,,,,Y,G,23,73,19,2018-10-08
22:25:59,,0.75,73,,11512,,,,"874231926046G",19,,,,,,,,11,,,,,,,
```

```
15:50:23,,0.5,69,,11425,,,,"271504218477",32,,,,,,,,11,,,,,,,
+ 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,41,69,32,2018-10-10
15:50:23,,0.5,69,,11428,,,,"271507491568",32,,,,,,,,11,,,,,,,
+ 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,41,69,32,2018-10-10
15:50:23,,0.5,69,,11484,,,,"271508481857",32,,,,,,,,11,,,,,,,
+ 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,77,69,53,2018-10-11
11:35:05,,0.05,134,,11447,,,,"874231098887",53,,,,,,,,11,,,,,,,
+ 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,77,69,53,2018-10-11
11:35:05,,0.05,134,,11448,,,,"874231135374",53,,,,,,,,11,,,,,,,
+ 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,77,69,53,2018-10-11
11:35:05,,0.05,134,,11479,,,,"874231461234",53,,,,,,,,11,,,,,,,
+ 69,"FOUR, PERSON",,FOUR,PERSON,,,,,,Y,G,87,69,59,2018-10-11
13:43:24,,0.05,34,,11487,,,,"874231676529",59,,,,,,,,11,,,,,,,
+ 73,"FIVE,PERSON",,FIVE,PERSON,,,,,,Y,G,23,73,19,2018-10-08
22:25:59,,0.75,73,,14,,,,"271508486757",19,,,,,,,,11,,,,,,,
+ 73,"FIVE,PERSON",,FIVE,PERSON,,,,,,Y,G,23,73,19,2018-10-08
22:25:59,,0.75,73,,11512,,,,"874231926046",19,,,,,,,,11,,,,,,,
```

Prior work

- Existing approaches mostly focus on syntactic changes.
- Fail to provide consumable summary of changes.

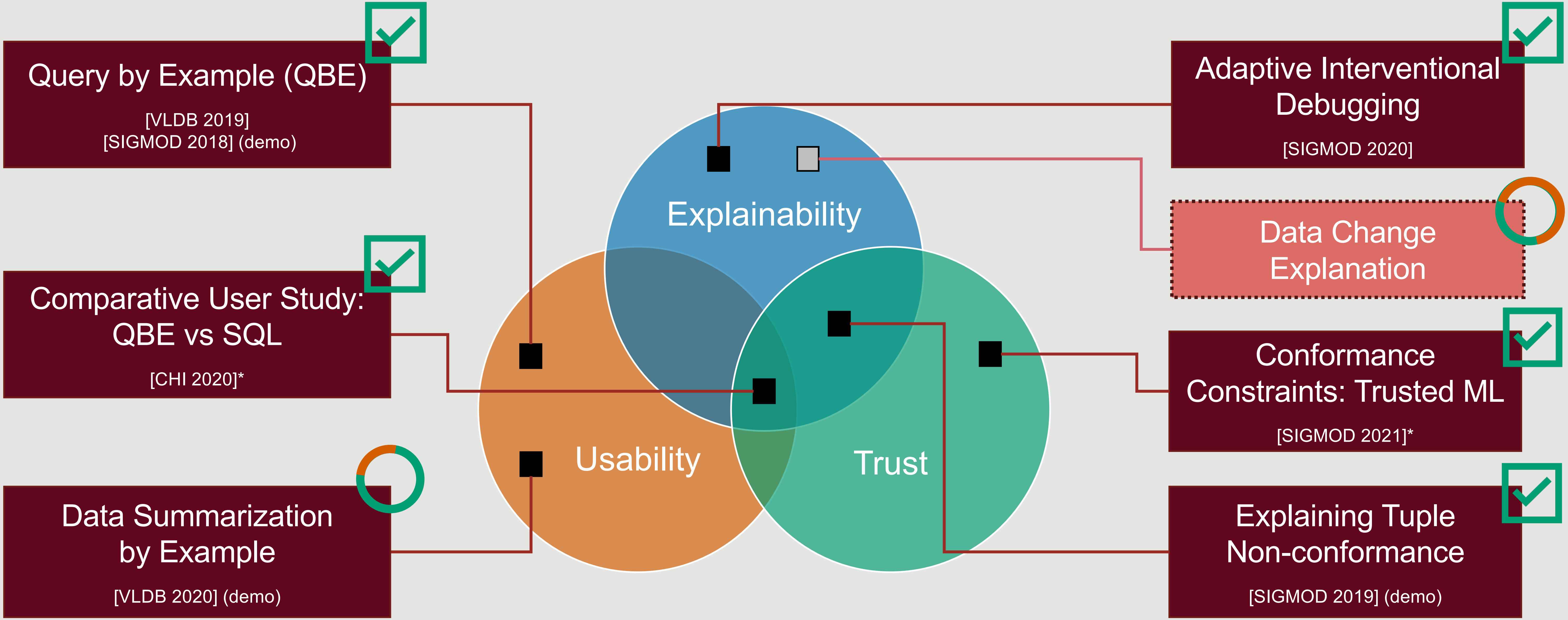
Our goal

- Provide a consumable summary of semantic changes that explains how two databases differ.
- Explains database evolution.
- Reveals patterns in data change.

Evaluating SuDocu

- Data collection
- Tuning SuDocu's learning algorithm
- Evaluation
  - Against ground-truth summaries
  - Comparison with other baselines
  - User study

Current status



* under submission/revision

Tentative timeline

- October 2020: proposal defense

- November – December 2020: evaluating SuDocu
- January 2020: submit to VLDB 2021

- January – June 2021: work on Data Change Explanation Framework
- July 2021: submit to SIGMOD 2022

- June – August 2021: work on dissertation
- August 2021: final defense

Other project affiliations

- Fair classifiers: experiment and evaluation

- Data profile debugger

- Data sampling by example

Acknowledgements

UMass Amherst | College of Information & Computer Sciences

Acknowledgements

- Committee




- Mentors and collaborators






Acknowledgements

- Armand Asnani, UMass
- Lucy Cousins, UMass
- Nischal Dave, UMass
- Larkin Flodin, UMass
- Juliana Freire, NYU
- Sainyam Galhotra, UMass
- Maliha Tashfia Islam, UMass
- Eunice Jun, UW
- Beryl Larson, Wellesley College
- Genglin Liu, UMass
- Raoni Lourenço, NYU
- Raj Kumar Maity, UMass
- Kancha Masalia, UMass
- Sheshera Mysore, UMass
- Tony Ohmann, UMass
- Vincent Pun, UMass
- Sheikh Muhammad Sarwar, UMass
- Michael Satanovsky, Hopkins School
- Divesh Srivastava, AT&T
- Zoey Sun, Smith College
- Nishant Yadav, UMass

UMass Amherst | College of Information & Computer Sciences

COMPUTING FOR THE COMMON GOOD

Anna Fariha

afariha@cs.umass.edu

people.cs.umass.edu/afariha