Data Analysis

Data / population → sample → data set X → Combination of multiple queries over X → analysis result

discover / predicate

Data Analysis
some queries rely on the results of other queries

Adaptive Data Analysis

\[ q_1(X) \]
\[ q_2(X) \]
\[ q_3(X) \]
\[ \ldots \]
\[ q_k(X) \]

output \( R \)

longest chain / adaptivity depth \( k \)
Adaptive Data Analysis - example

$X$ input

$\text{query}_1(X)$

$\text{query}_2(X)$

$\text{query}_3(X)$

$\text{query}_4(\cdot, X)$

$\text{query}_5(\cdot, X)$

$\cdots$

$\cdots$

$\text{query}_n(\cdot, X)$

output $R$

adaptivity depth $\max(k_1, k_2)$

depth $k_1$

depth $k_2$
Motivation
– Generalization Error / Overfitting

Adaptivity in analysis will propagate the overfitting
Existing Methods – 1 Query Guarantee

Guarantee generalization error

fits well

Data / Population

sample

data set X

Mechanisms

query over X

Gaussian Mechanism
Laplace Mechanism
Threshold out
...

analysis result

Data Analysis

Guarantee will lose in multiple adaptive queries
Motivation – Multiple Queries Guarantee

Adaptivity
Depth

Where to apply mechanisms

How many mechanisms needed

Light or intense mechanism

... 

Guarantee Generalization Error for Multiple adaptive queries
OUR WORK

analyze the adaptivity depth for data analysis program.
Novelty

use mechanism to encapsulate queries and combine them.
Challenges in Language Design

- adaptivity depends on the **Runtime Information**
- represent the **probabilistic computing**
- provide precise **Upper Bound** for adaptivity depth
Refinement Types

singleton type: \texttt{int[ I ]}

index term: indicate the value of an integer

domain: \mathbb{N}

representing the run time information
Expressions

\[ \delta(q) \]

represent the mechanism \( \delta \) applied over a query \( q \)

**uniform** \( v_1 \ v_2 \)

the range \([v_1, v_2] \), where sample from

parameterized const represents values sampled from distributions
Typing Judgement

Annotated typing judgement: approximates an upper bound on adaptivity depth of expressions

\[ \Gamma \vdash e : \tau \]

upper bound on adaptivity depth of \( e \)

expression
Typing Judgement

Important rule for calculating the adaptivity:

\[ \vdash_z q : \tau \]

\[ \vdash_{z+1} \delta(q) : \tau \]
Soundness

step-indexed logical relations

fundamental theorem

\[\text{if } \vdash_z e : \tau \text{ then } e \in [\tau]^Z_e\]
System Overview

prog → input → type system → output → adaptivity depth
System Overview – Simple Example

```
let prog = \x. 
let q = \x.x in 
Mech( q(x) )
```

- Gaussian Mechanism
- Laplace Mechanism
- Threshold out
- ...

input: data set

query

input

output

1
THANKS