Query-Ads Matching in Sponsored Search: Challenges and Solutions

Alibaba Corporation

Speaker：Yunhua Hu

Self Introduction

- Expertise
  - Computational Advertising, text mining, Natural Language Processing, Machine Learning
- Experience
  - From 2012：Alibaba Corp.
  - 2007~2012：Work in MSRA
  - 2003~2006：Internship in MSRA
- Contact
  - Sina microblog：胡云华
  - Email：wugou.hyh@taobao.com

What is Query-Ads Matching in Sponsored Search

Sponsored Search in Taobao

Query-Ads Matching

Consumer

fish

Advertiser

Fresh fish

Shark
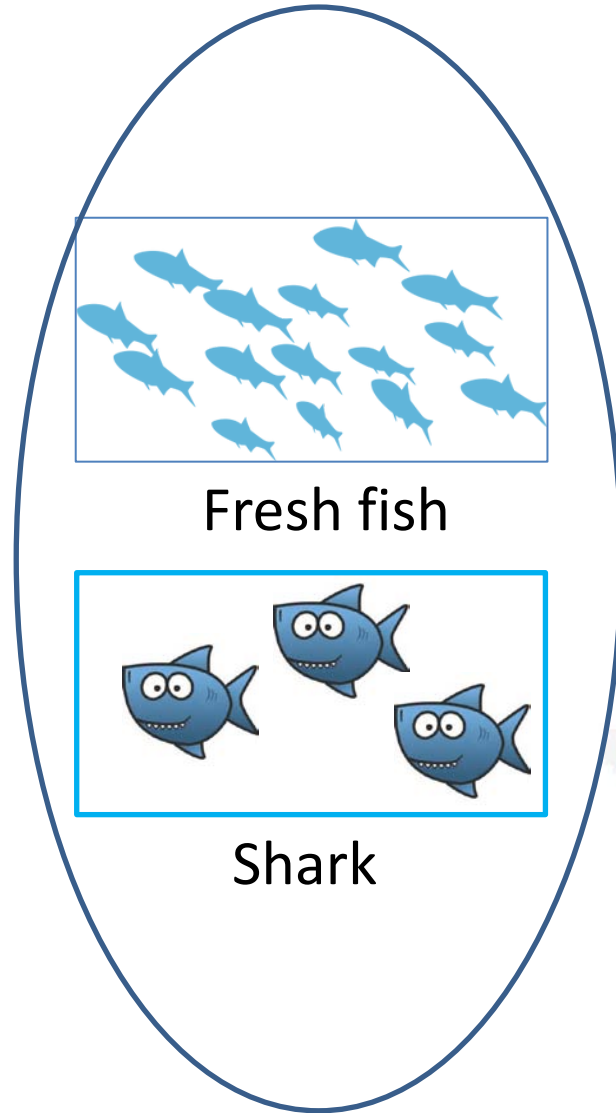
fish

Three-Stage Query-Ads Matching

- Process of query-ads matching

Fresh fish

fish

Shark

Sea fish

Query Rewriting

Bid-Ads Ranking
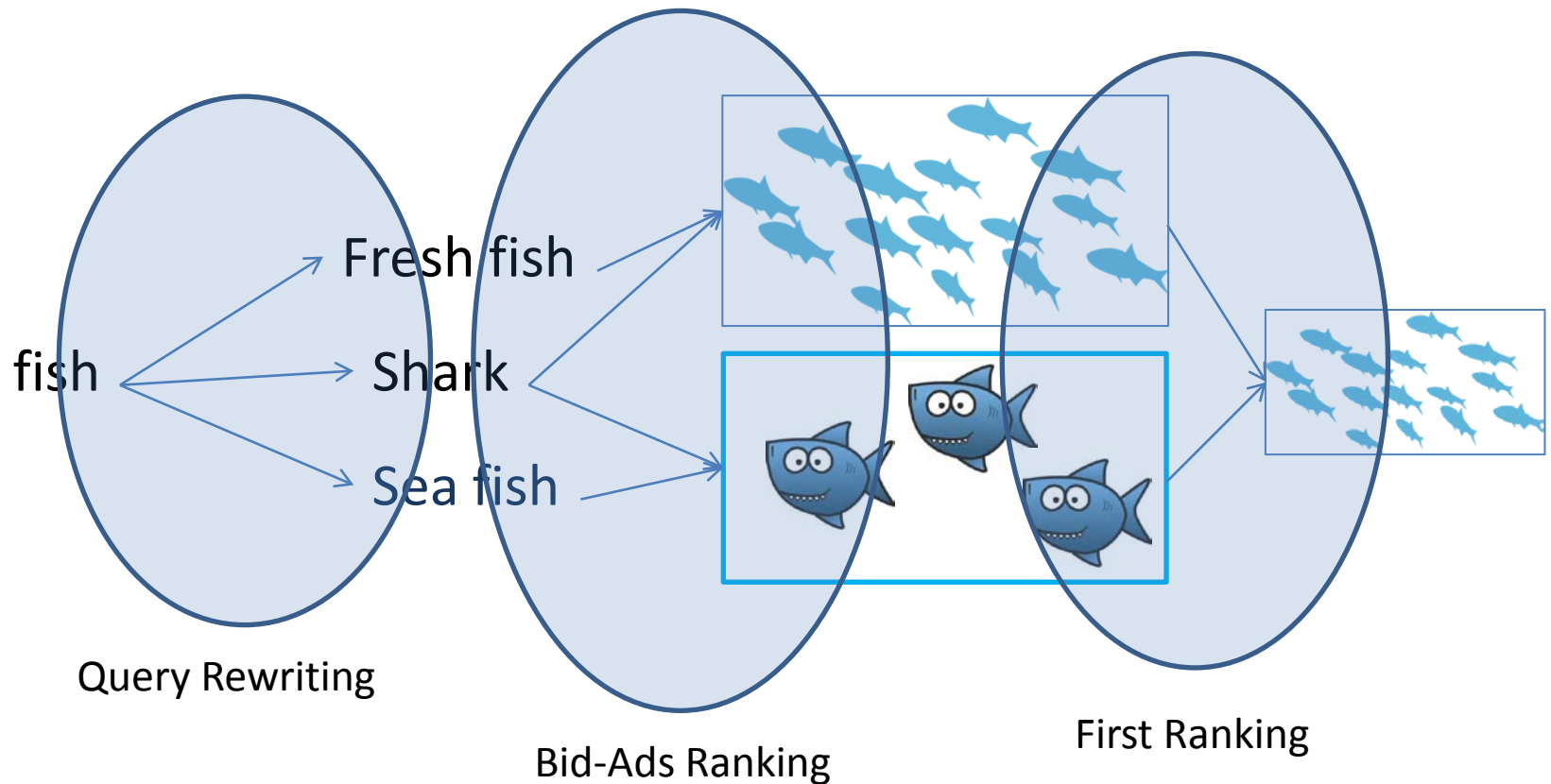
First Ranking

Three-Stage Query-Ads Matching

- Three-stage matching：

Query → [Query Rewriting] → Bid Words → [Bid-Ads Ranking] → Ads Candidates → [First Ranking] → Ads
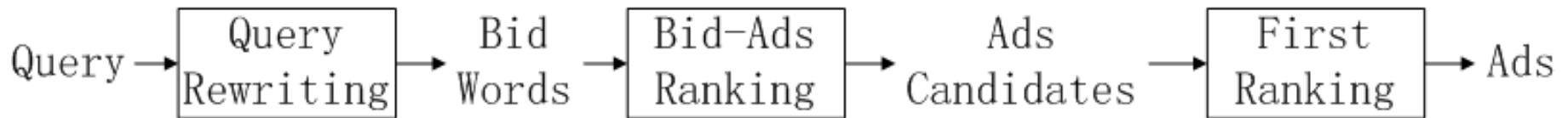
- Query Rewriting：
  - Matching query to related Bids
- Bid-Ads Ranking：
  - Finding high quality Ads for each Bid
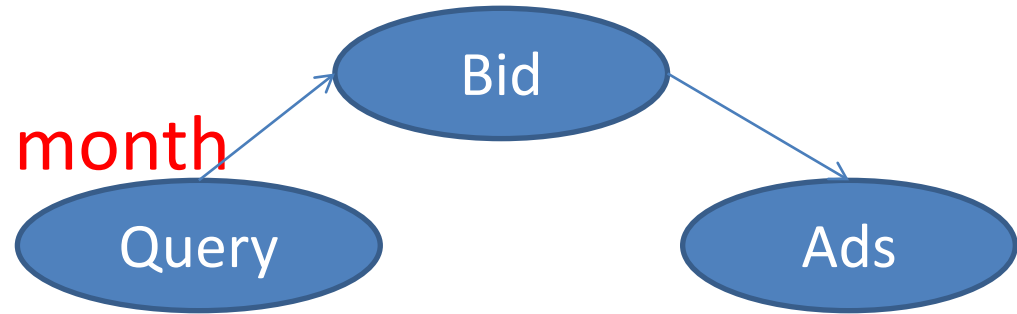- First Ranking：
  - Finding candidate Ads set for CTR prediction

Problems in Three-Stage Model

Lessons We Learnt

- Invest：>5 people*6 month
- Main idea：
  - Optimizing each stage then the whole process
  - Query Rewriting：query clustering + Bid Ranking
  - Bid-Ads Ranking：CTR prediction
  - First Ranking：Scoring and filtering
- Output：CTR improvement <2%，far less than expected
- Difficulties：
  - Not each to evaluate and optimization

Bid

Query

Ads

Three-Stage Query-Ads Matching
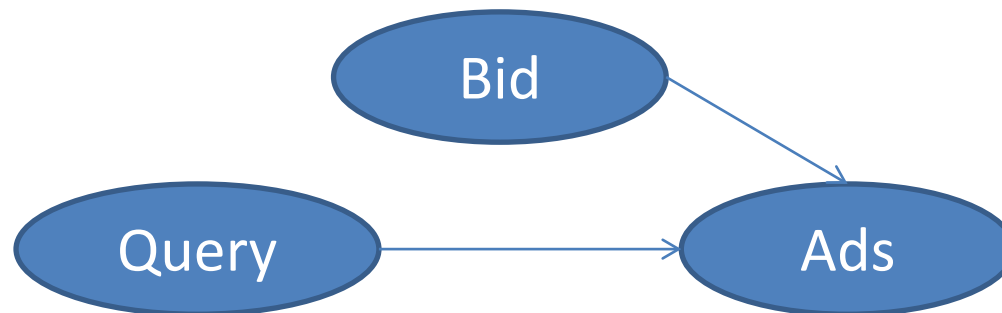
First Ranking

Query Rewriting

Bid-Ads Ranking

- Formulation of three-stage matching

$$P(a \mid q) = \sum_i P(b_i \mid q) P(a \mid b_i, q)$$

q : $input\,query$, b : $Bid\,words\,providing\,to\,advertiser$, a : $related\,ads\,of\,the\,query$

- Explanation of three-stage matching
  - Need to search every related b_i
  - The probability of ads depended both on bid b_i and query q

Problems in Three-Stage Model
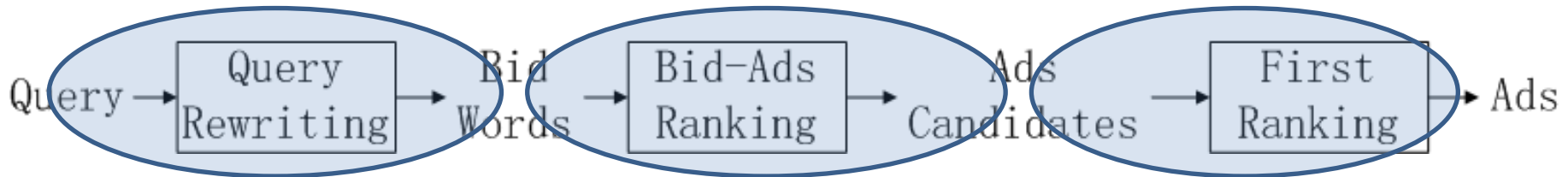
- Problem 1: Top n bids set the limitation!

$$P(a \mid q) = \sum_i P(b_i \mid q) P(a \mid b_i, q)$$

$$= \sum_1^n P(b_i \mid q) P(a \mid b_i, q) + \sum_{n+1} P(b_i \mid q) P(a \mid b_i, q)$$

- Due to the limitation, the number of bid will set to a small number, for example, let *n=10*
- Hot bid word has been over competed
- "High price bid + bad ads" win "Low price bid + good ads"

Problems in Three-Stage Model

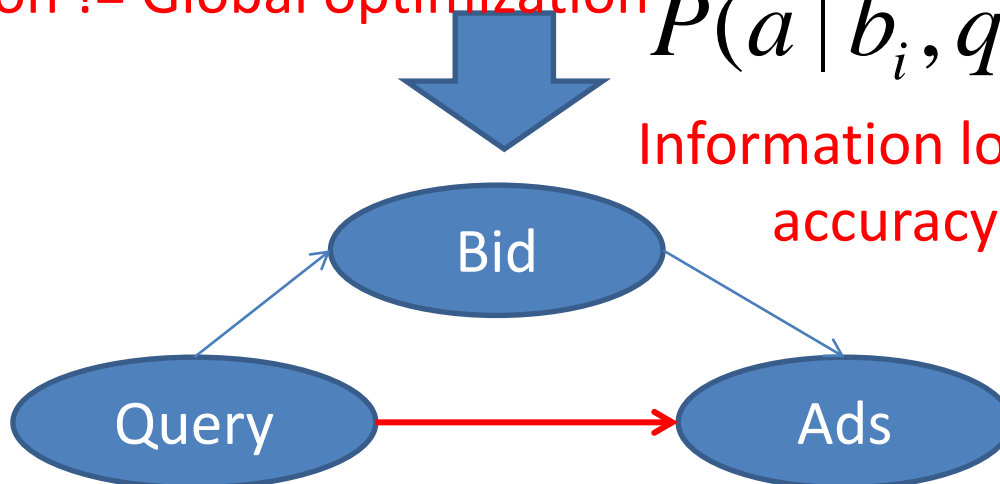- Problem 2: Linear combination is not reasonable!

$$P(a \mid q) = \sum_i P(b_i \mid q) P(a \mid b_i, q)$$

Local optimization != Global optimization

$$P(a \mid b_i, q) \neq P(a \mid b_i)$$

Information loss causes accuracy loss!

Any Better Solutions？
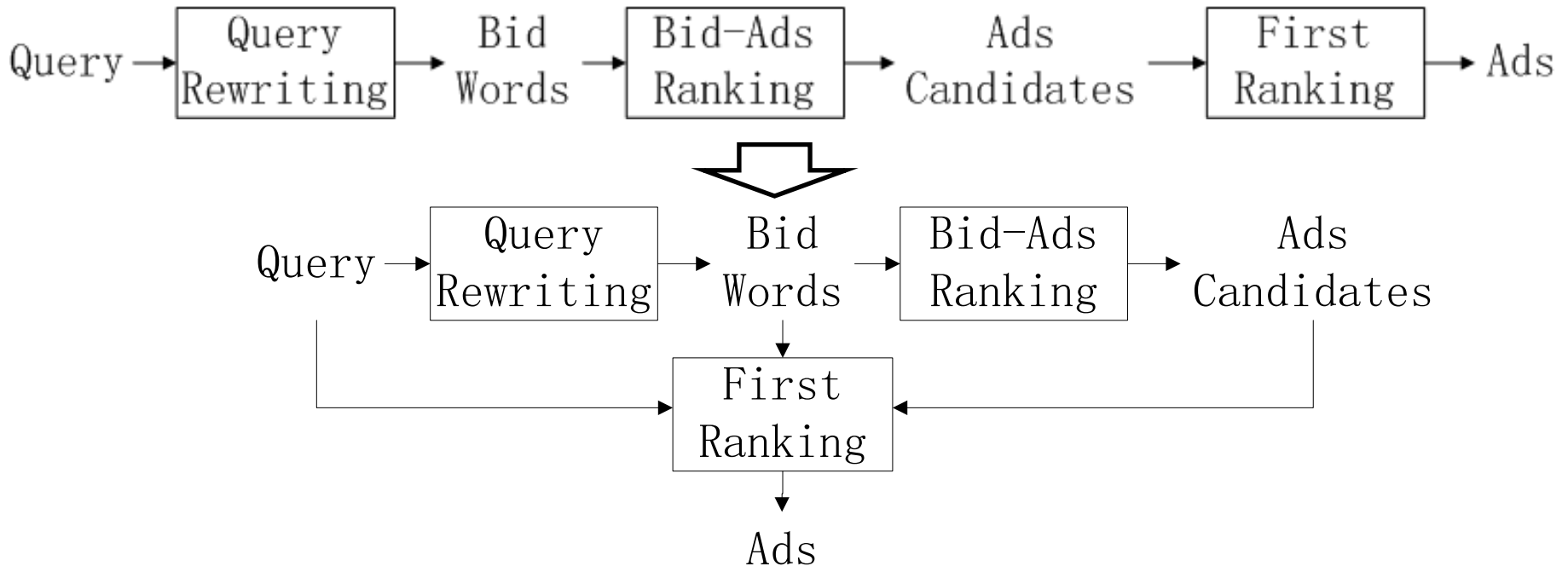
Yes !

Problem 1: Top *n* bids set the limitation
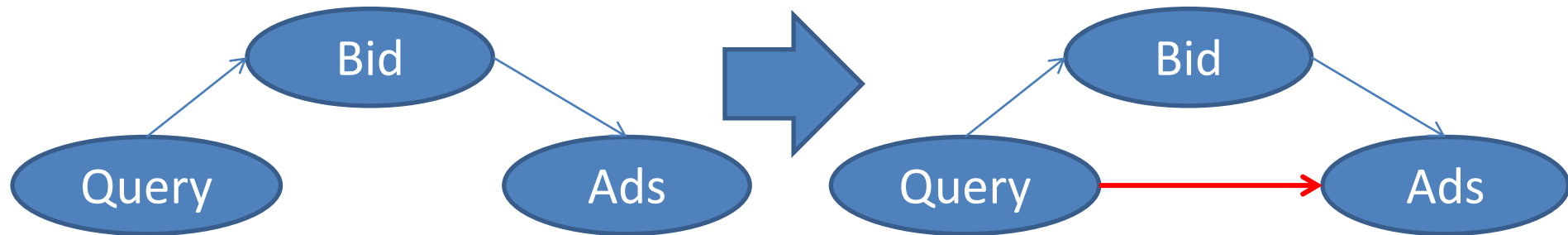
- Solution: Reduce the limitation by increasing *n*!

$$P(a \mid q) = \sum_i P(b_i \mid q) P(a \mid b_i, q)$$

$$= \sum_1^n P(b_i \mid q) P(a \mid b_i, q) + \sum_{n+1} P(b_i \mid q) P(a \mid b_i, q)$$
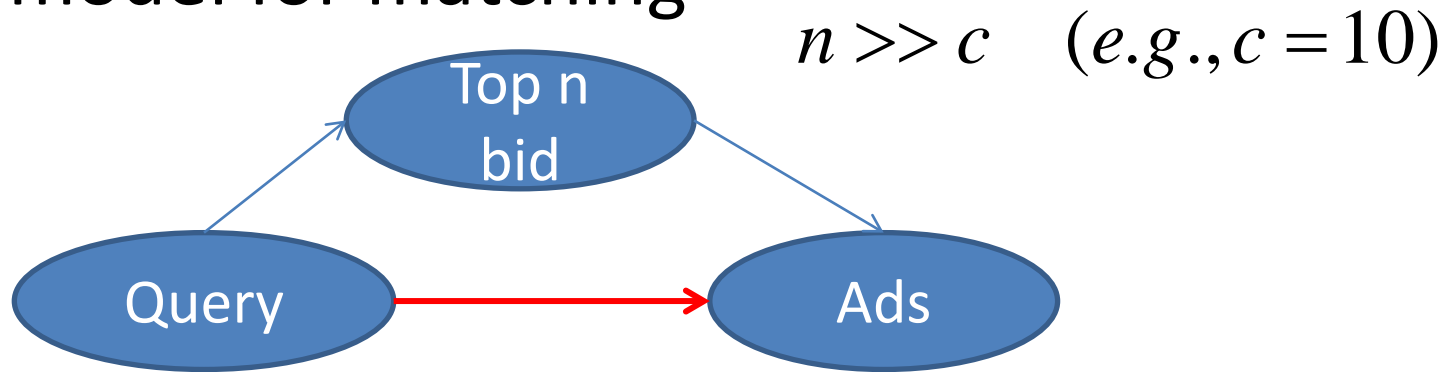
Increase n from 10 to 200!

Problem 2: Linear combination is not reasonable

- Solution: Change the linear structure!

New Three-Stage Mode

- New model for matching

$$n >> c \quad (e.g., c = 10)$$

- Ads
  - More bids will bring more efforts to the system
    - How to set n?
  - Non-linear structure is difficult to optimize too!
    - How to get the probability of ads?
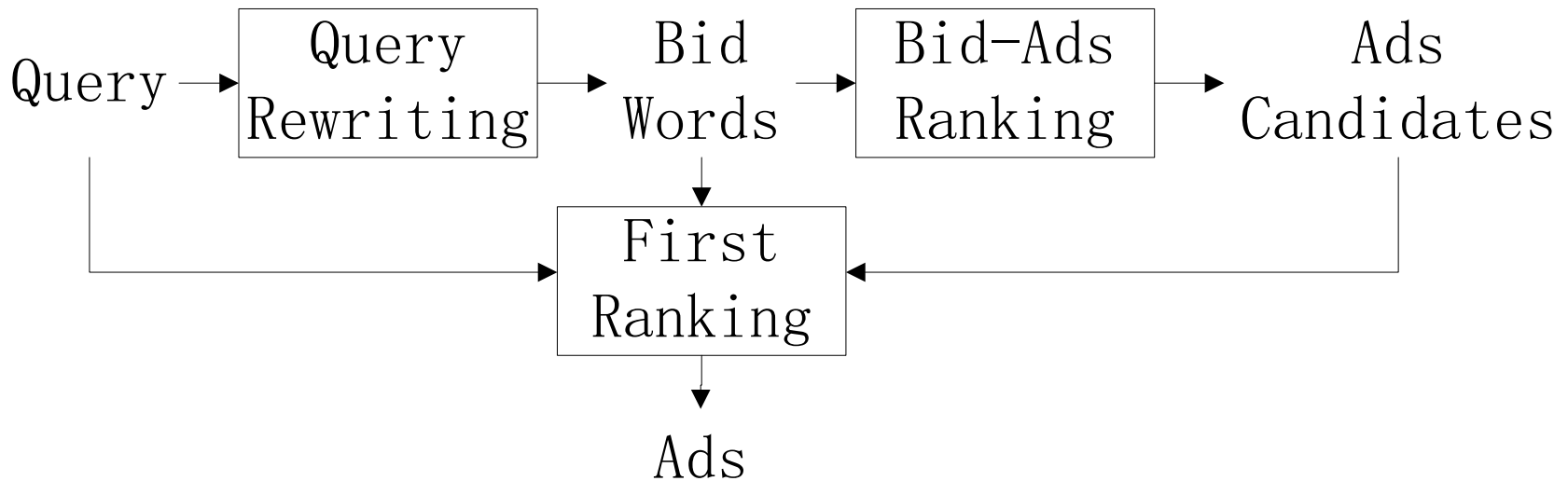
How to Tackle the Challenges？

Construct Data Set for Direct Optimization

- How to optimize the three-stage model？
  - Given a set of queries, find **all ads** for each query by considering all related bids about the query
  - Scoring each ads by using **CTR prediction** on each query-ads pair
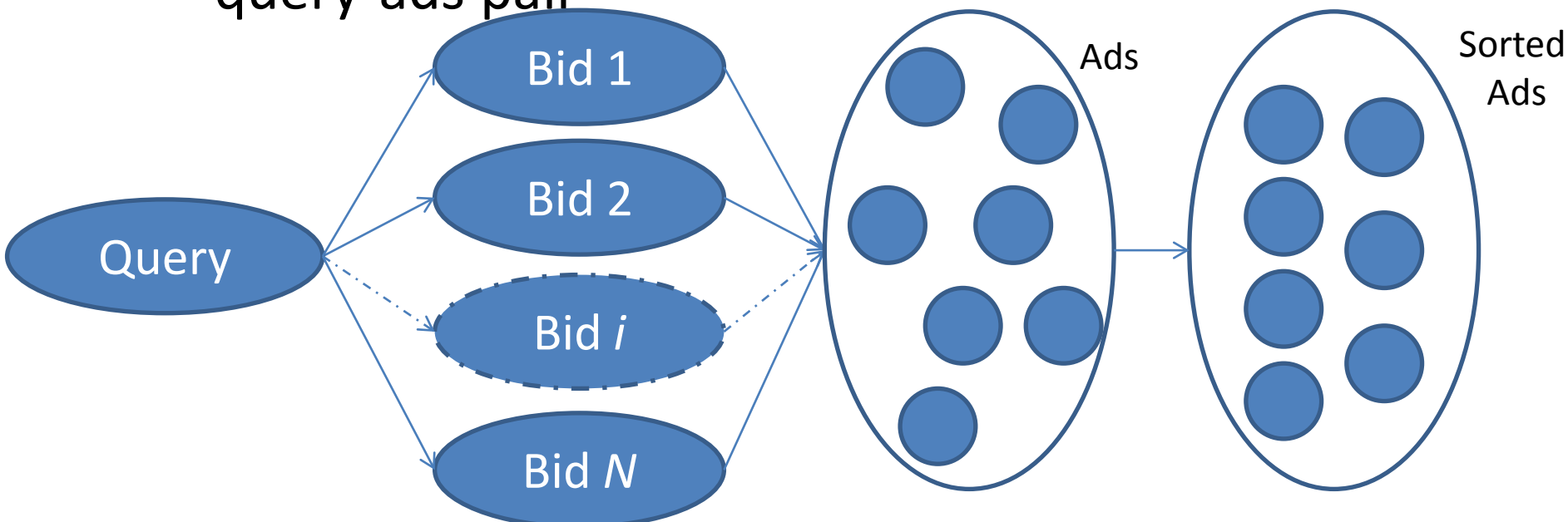
Direct Optimization for the New Three Stage Model

- The New Three Stage Model

```
Query →  ┌──────────┐  →  Bid   →  ┌──────────┐  →  Ads
         │  Query   │      Bid     │ Bid-Ads  │      Ads
         │ Rewriting│     Words    │ Ranking  │   Candidates
         └──────────┘              └──────────┘
    │                      ↓                           │
    │                 ┌──────────┐                     │
    └────────────────→│  First   │←────────────────────┘
                      │ Ranking  │
                      └──────────┘
                           ↓
                          Ads
```

- The idea of direct optimization
  - Convert the problem to a Supervised Learning problem
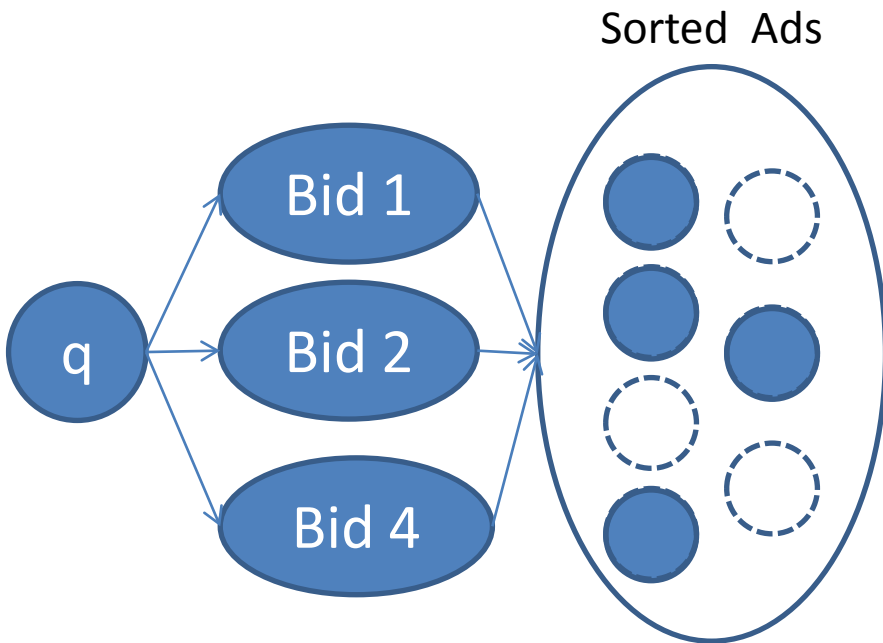
How to Optimize Query Rewriting Stage?

- Fix query rewriting algorithm, try different *n* to see the coverage of best ads set
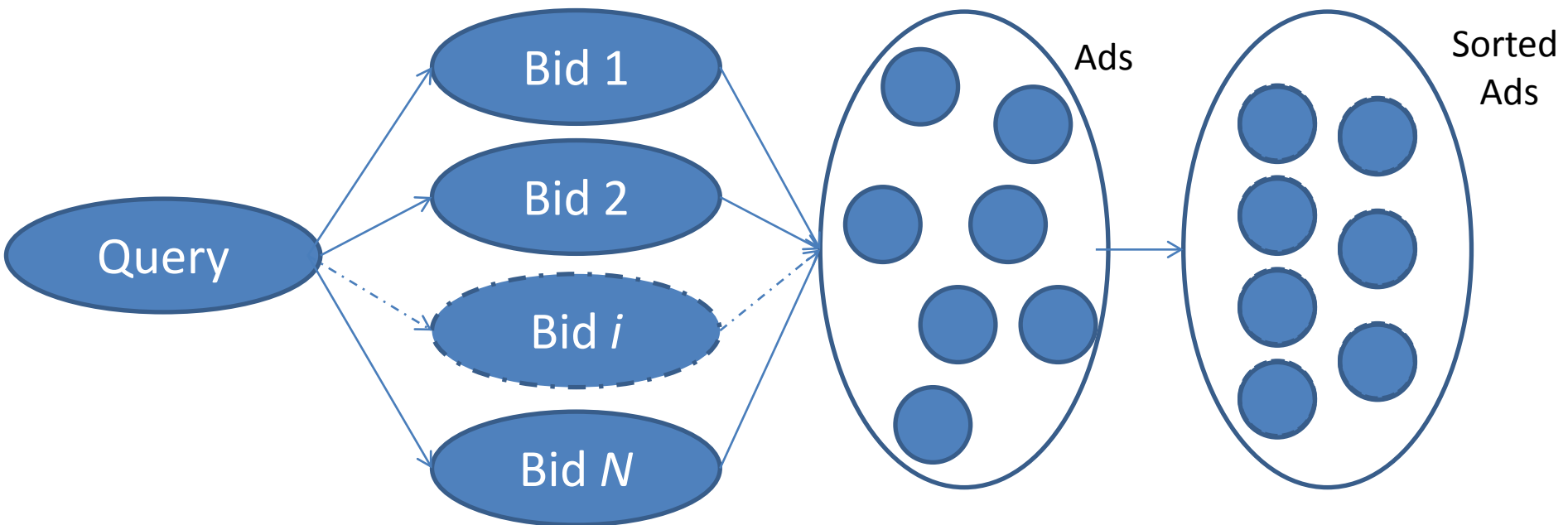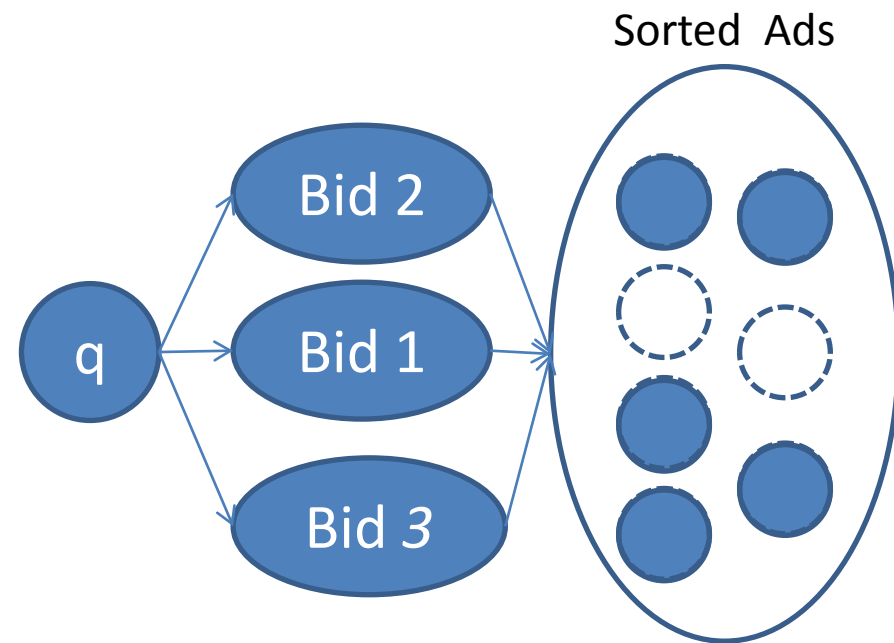
How to Optimize Query Rewriting Stage?

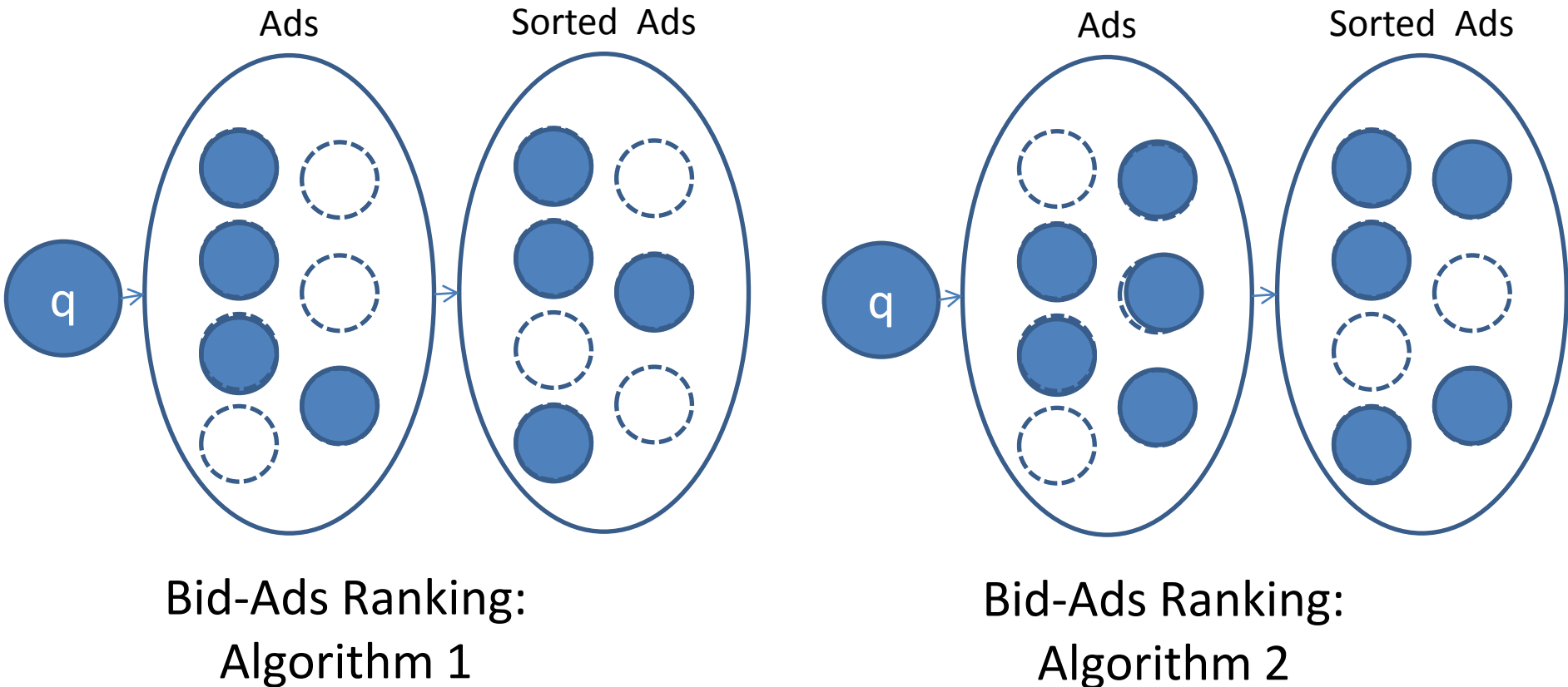- Fix parameter *n*, try different rewriting algorithms

Query Rewriting: Algorithm 1

Query Rewriting: Algorithm 2
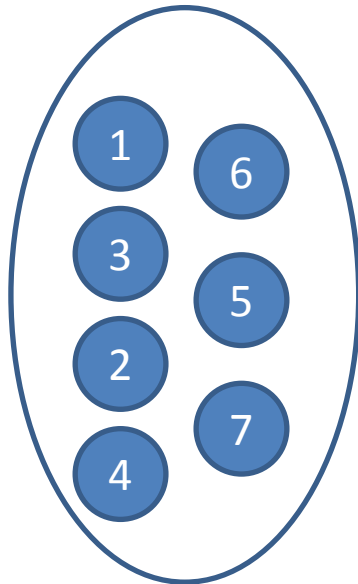
How to Optimize Bid-Ads Ranking Stage?

- Try different Bid-Ads ranking algorithms (ads filtering) to see the coverage of best ads set

Ads   Sorted Ads   Ads   Sorted Ads

q

Bid-Ads Ranking:
Algorithm 1

q

Bid-Ads Ranking:
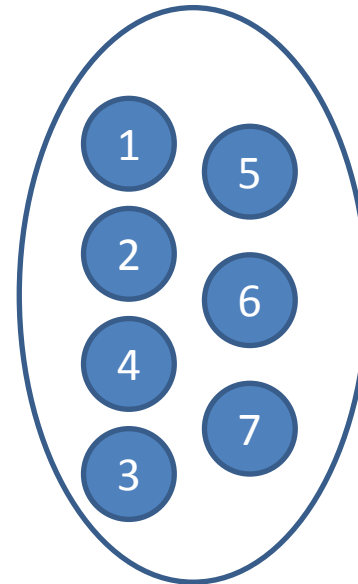Algorithm 2

How to Optimize First Ranking Stage?

- Try different first ranking algorithms (light weight model to get better results) to see the ranking list of best ads set

Sorted Ads

First Ranking:
Algorithm 1

Sorted Ads

First Ranking:
Algorithm 2

Experimental Results

Direct Optimization for the New Three-Stage Matching Model

- Experiment Setting
  - Query rewriting：SimRank + Log Mining
  - Bid-Ads ranking: Using a CTR prediction model
  - First ranking: light weight CTR prediction model
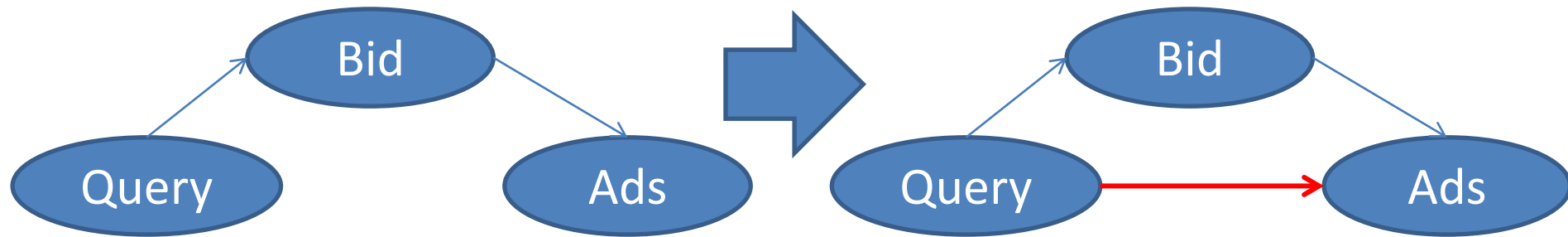 - Parameters：Top *n=200* query，top 8000 ads

Experimental Results

- Preliminary results

	Top	Tail	All
CTR	+4.4%	+5.2%	+4.7%
PPC	+0.0%	+1.1%	+0.3%
RPM	+4.4%	+6.3%	+5.0%

Conclusion

- A framework to help optimizing matching process

Thanks!

Q & A