

Towards More Usable Dataset Search: From Query Characterization to Snippet Generation

Jinchi Chen¹, **Xiaxia Wang**¹, Gong Cheng¹, Evgeny Kharlamov^{2,3}, Yuzhong Qu¹

¹ National Key Laboratory for Novel Software Technology, Nanjing University, China

² Bosch Center for AI, Bosch GmbH, Germany

³ University of Oslo, Norway



BOSCH



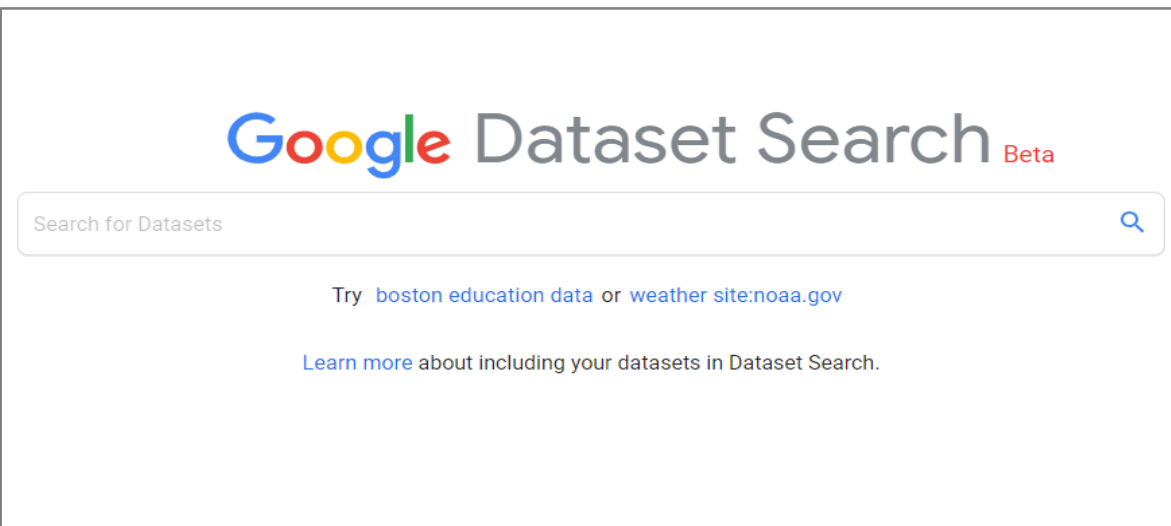
Dataset search engines emerge for conveniently finding useful datasets.



The screenshot shows the Google Dataset Search Beta interface. At the top, the text "Google Dataset Search" is displayed in the multi-colored Google font, with "Beta" in red to the right. Below this is a search bar with the placeholder text "Search for Datasets" and a magnifying glass icon on the right. Underneath the search bar, there is a suggestion: "Try [boston education data](#) or [weather site:noaa.gov](#)". At the bottom of the interface, there is a link: "[Learn more](#) about including your datasets in Dataset Search."

Google Dataset Search

Dataset search engines emerge for conveniently finding useful datasets.



Google Dataset Search

Existing efforts:

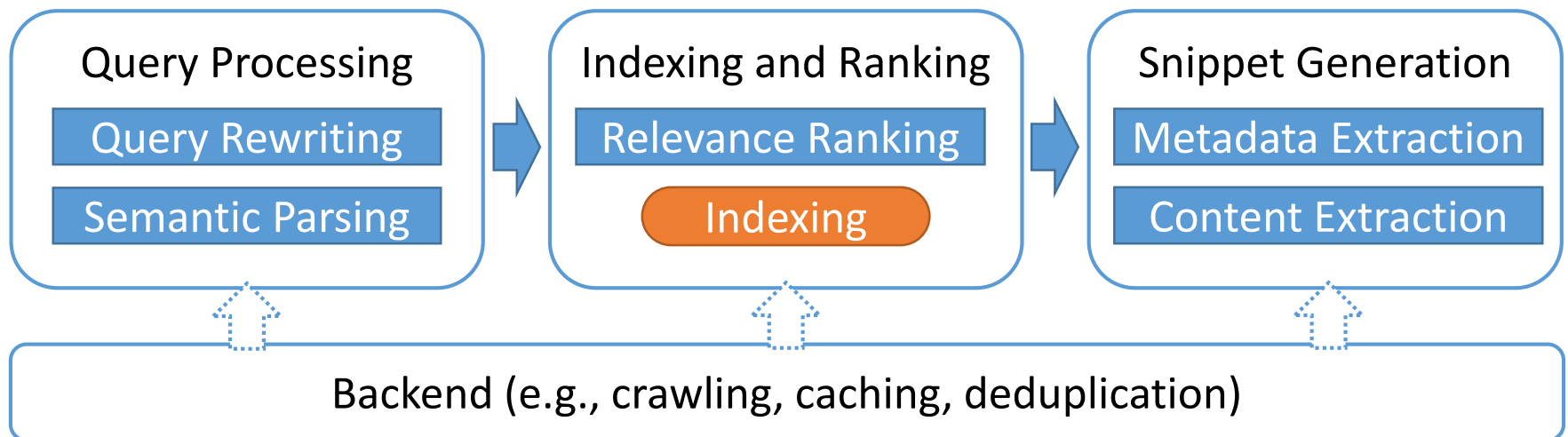
- Metadata management
- Dataset browsing
- ...

Query processing? Result displaying?

What should a **usable dataset search engine** be like?

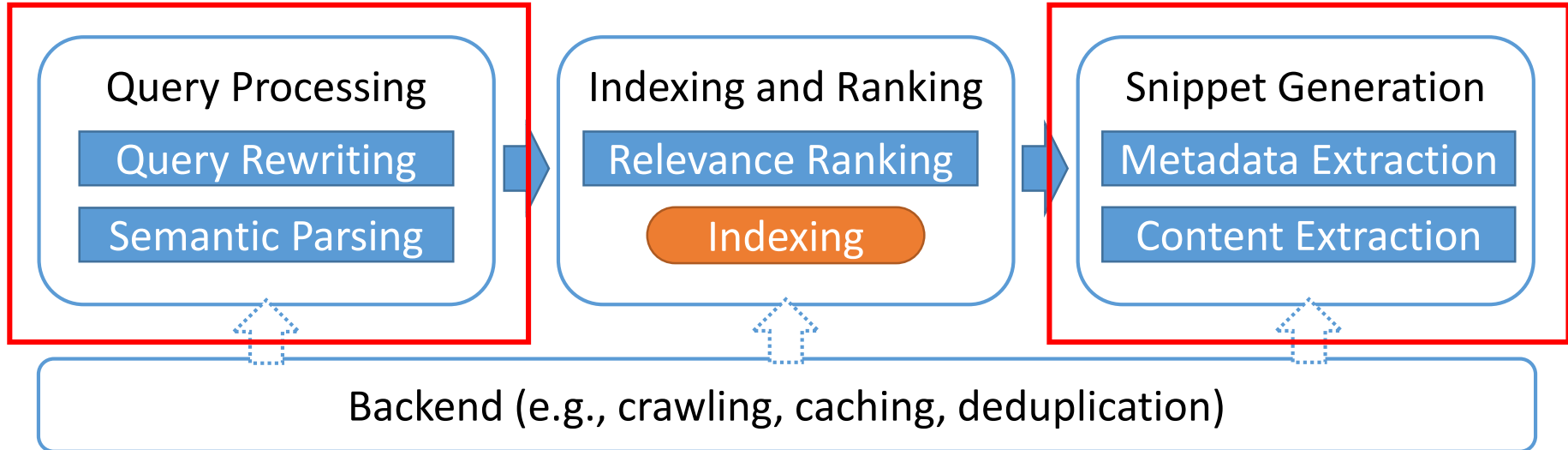
A query-centered framework for dataset search

- 4 main **components**



A query-centered framework for dataset search

- 4 main **components**



1. Characterizing data needs

- Any **special characters**?
- How to **analyze and use** them?

Home

PUBLIC

Stack Overflow

Tags

Users

Jobs

TEAMS

What's this?

First 25 Users Free

looking for dataset containing longitude and latitude [closed]

Asked 7 years, 11 months ago Active 6 years, 8 months ago Viewed 9k times

▲

1

<http://www.geonames.org/export/>

▼

★

1

Hi I have a been set an assignment in university for which i have to use datasets with longitude and latitude data in but im struggling to find some so far i have looked at

this website offers it but not in a efficient way to read,

but the csv forms they offer are proving very hard to import into a mysql database

and the last site is

<http://linkeddata.org/data-sets>

which doesnt seem to contain any datasets with longitude and latitude data,

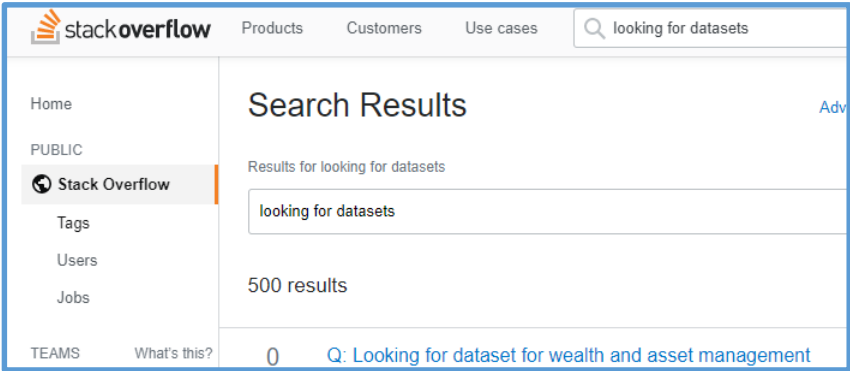
i was wondering if anyone had any ideas of where i could find datasets that meet the criteria that i am looking for and are also free.

Thank you in advance

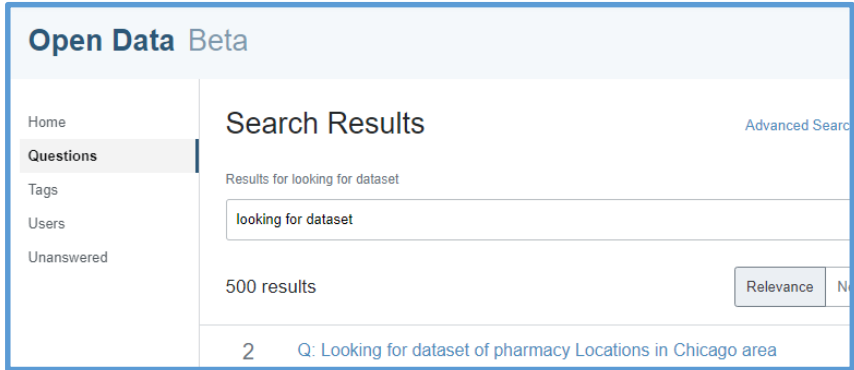
*/*A Data Need Example*/*

1. Characterizing data needs

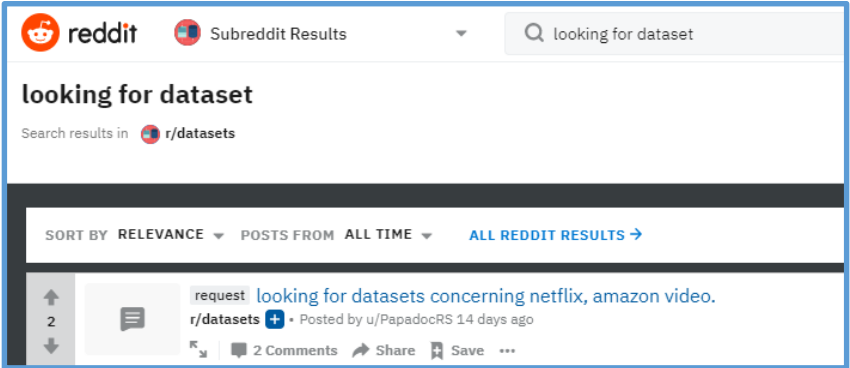
Collect



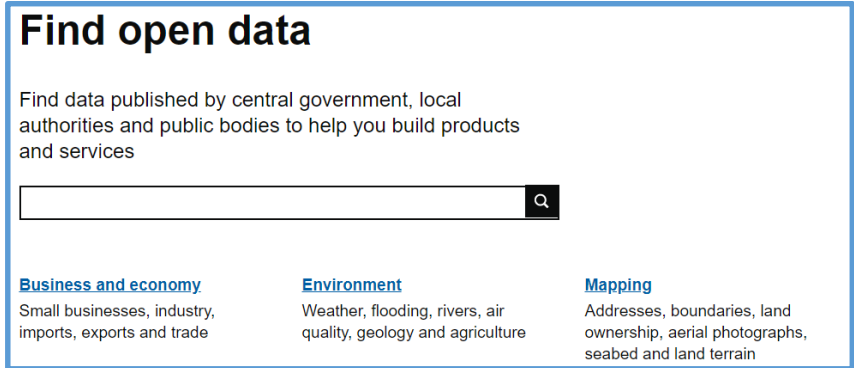
Stack Overflow



Open Data Stack Exchange



Reddit



data.gov.uk

1. Characterizing data needs



Data need description:

"I am looking for datasets that lists the location of accidents or traffic (latitude and longitude) with date and time in many countries. I found datasets for USA and UK, now looking for datasets for other countries. Any type of road accident would be great."

10 Human
Experts



Rewrite

Dataset search query:

"location of accidents or traffic with date and time in many countries."

An example of data need and its corresponding query

1. Characterizing data needs

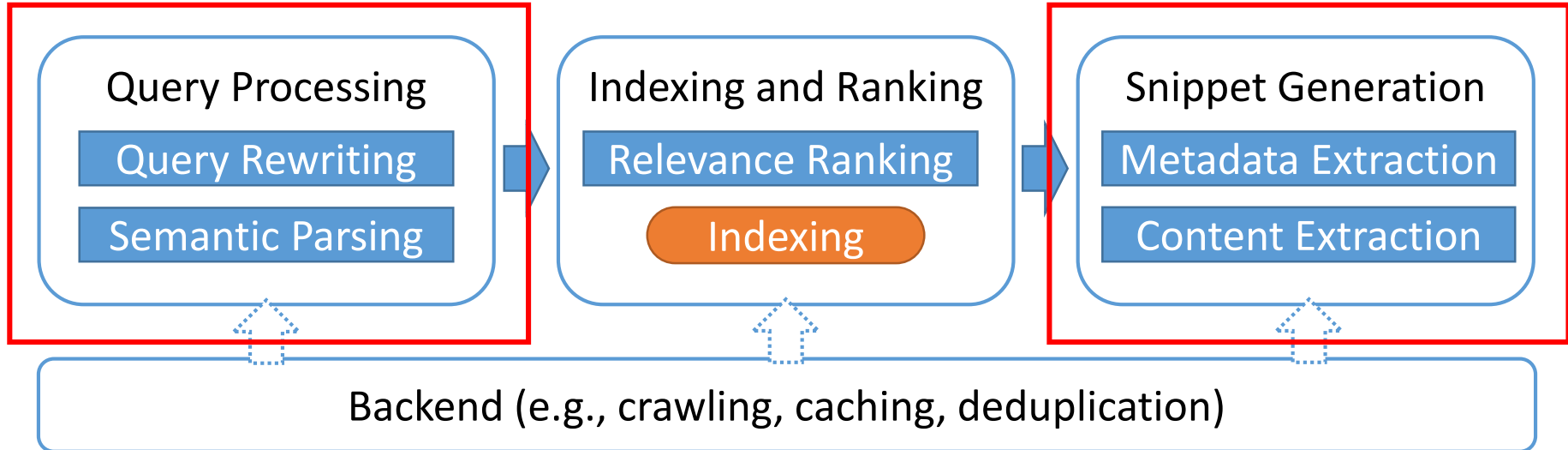


Annotation scheme and its distribution in dataset search queries

Category	% of Queries	Example Query
Metadata	Name	3.54% <i>HUST-ASL</i> Dataset
	Domain/Topic	94.45% <i>weather</i> dataset with solar radiance and solar energy production
	Data Format	16.23% <i>jpg images</i> for all unicode characters
	Language	3.90% annotated moive review dataset in <i>German</i>
	Accessibility	7.40% <i>open source</i> handwritten English alphabets dataset
	Provenance	0.21% <i>FDA datasets</i> about medicine name and the result has adverse events
	Statistics	2.98% dataset contains at least <i>1000</i> examples of opinion articles
	Overall	96.05%
Content	Concept	50.59% dataset about people, include <i>gender, ethnicity, name</i>
	Geospatial	19.21% judicial decisions in <i>France</i>
	Other Entities	0.41% datasets with nutrition data for many commercial food products (i.e., <i>Lucky Charms, Monster Energy, Nutella</i> , etc.)
	Temporal	9.35% <i>2011–2013</i> MoT failure rates on passenger cars
	Other Numbers	1.59% businesses that employ over <i>1000</i> people in Yorkshire region
	Overall	63.79%

A query-centered framework for dataset search


- 4 main **components**



2. Generating dataset snippets

Existing systems serve only metadata:

- No detailed information of dataset content
- **Limited utility** for users

 Met Office

Mountain Weather Forecasts

[Explore at data.gov.uk](#) [Explore at data.wu.ac.at](#)

37 scholarly articles cite this dataset ([View in Google Scholar](#))

Dataset updated Dec 19, 2013

Dataset provided by
Met Office

License
Fair Usage - <http://www.metoffice.gov.uk/about-us/legal/fair-usage>

Available download formats from providers
html

Description
Mountain forecasts are provided for the main mountain areas and those which have daylight hours.

Google Dataset Search system

2. Generating dataset snippets

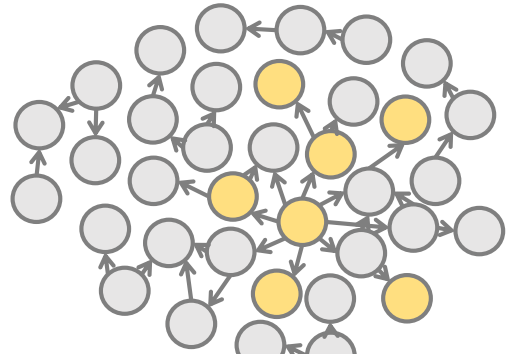
Dataset snippet (for RDF datasets): A **subset** of RDF triples to exemplify the **dataset content** and its **relevance to the query**.

2. Generating dataset snippets

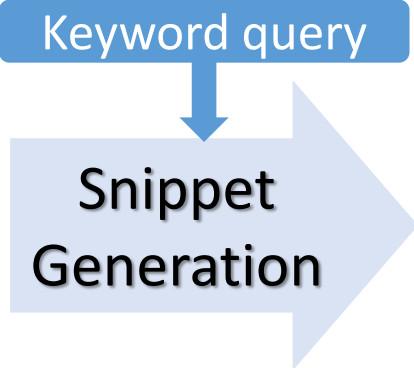
Dataset snippet (for RDF datasets): A **subset** of RDF triples to exemplify the **dataset content** and its **relevance to the query**.

- <Beijing-locatedIn-China>
- <Beijing-neighboringCity-Tianjin>
- <Beijing-TYPE-Capital>
- <Beijing-TYPE-City>
- <China-TYPE-Country>
- <China-isPartOf-EastAsia>
- <Tianjin-TYPE-City>
- <Tianjin-locatedIn-China>

Dataset: A set of subject-predicate-object triples

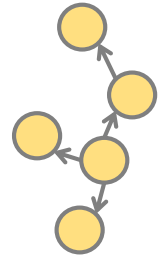


An RDF graph



- <Beijing-locatedIn-China>
- <Beijing-TYPE-City>
- <China-TYPE-Country>

Snippet: A size-limited subset of triples



A subgraph ¹³

2. Generating dataset snippets

➤ Methods

- Query-biased Snippets

[Li et al., Efficient and progressive group steiner tree search. SIGMOD, 2016]

- Cover all the query keywords
- Connected and compact

- Illustrative Snippets

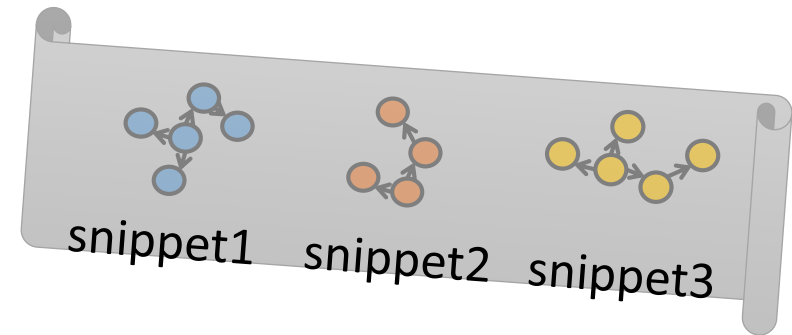
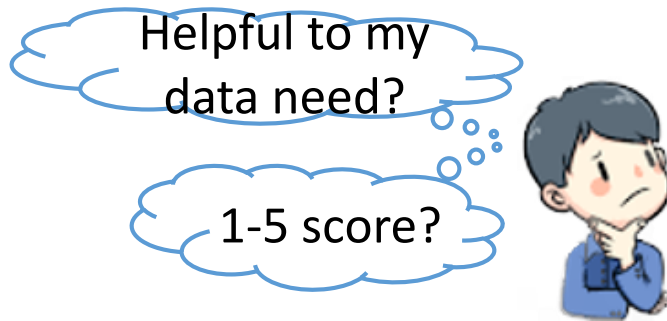
[Cheng et al., Generating illustrative snippets for open data on the web. WSDM, 2017]

- Cover frequent classes and properties
- Cover central entities

2. Generating dataset snippets

Experiment: user study

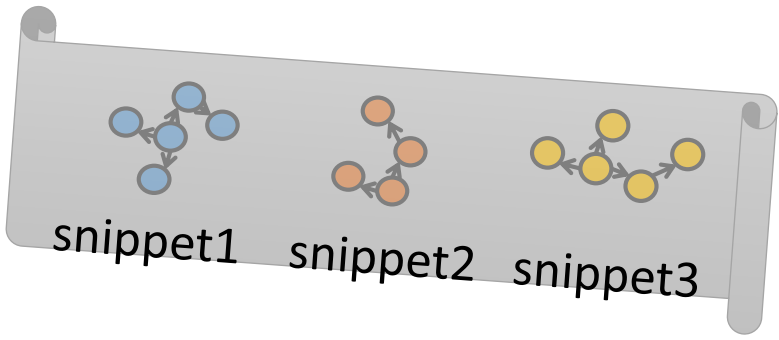
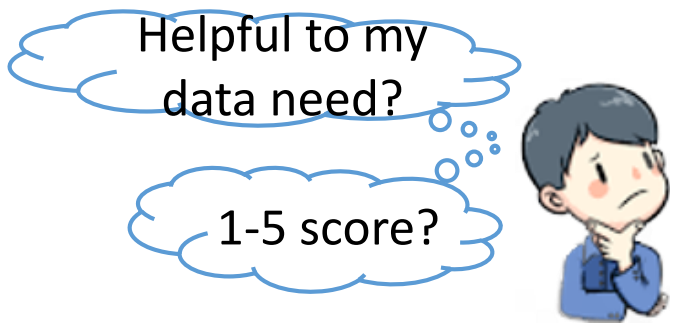
- 311 real datasets: From *Datahub*
- 15 participants, rating on a 1 - 5 scale



2. Generating dataset snippets

Experiment: user study

- 311 real datasets: From *Datahub*
- 15 participants, rating on a 1 - 5 scale



Result

- Two snippets have different preferences
- But neither of them performed well for users

Human-rated Usefulness of Snippets

	Score (1 - 5)
Query-biased	1.91 ± 1.22
Illustrative	3.04 ± 1.23

Summary

➤ Our contribution

- **Collect** real data needs, **derive** 1,947 dataset search queries, semantically **annotate** them using a fine-grained scheme.
- **Present** a query-biased snippet and an illustrative snippet, **compare** them in a user study.

➤ Future Work

- Automatic query **parsing**
- Better snippet **generation**
- Dataset search engine **construction**

Thanks for your time!

Q&A

Contact: xxwang@smail.nju.edu.cn