Subjective Databases: Enabling Search by Experience

Wang-Chiew Tan
Megagon Labs

EDBT 2019

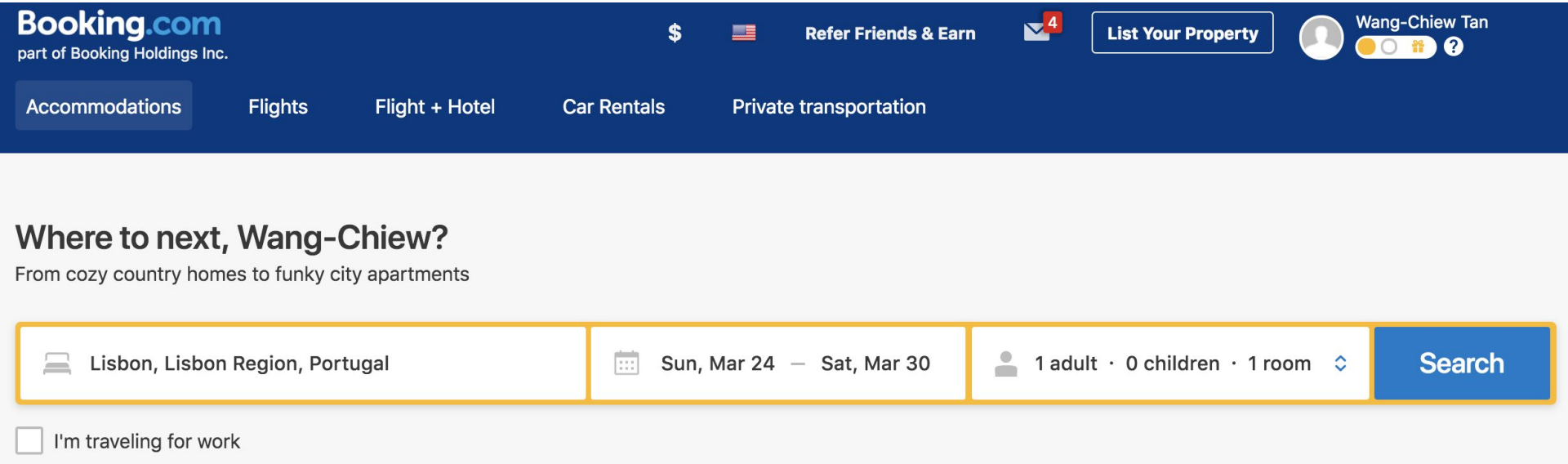
Megagon Labs

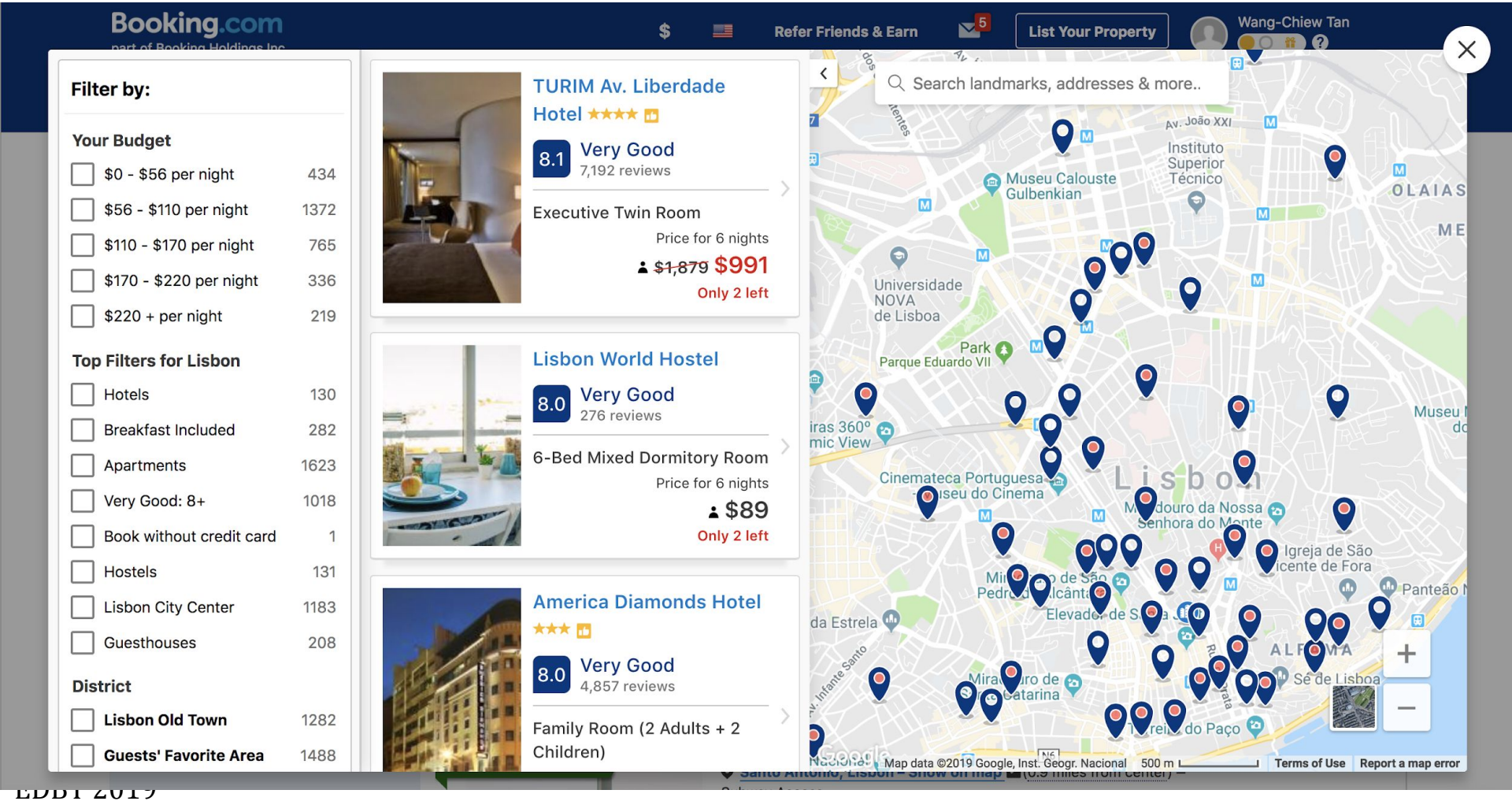
Recruit Holdings:
A human resources and lifestyle company, 200+ online services.

An example hotel query

“Hotels with clean rooms near IST congress center in Lisbon, Portugal.”

Today's hotel websites




Booking.com
part of Booking Holdings Inc
$
Refer Friends & Earn
5
List Your Property
Wang-Chiew Tan
Filter by:
Your Budget
$0 - $56 per night 434
$56 - $110 per night 1372
$110 - $170 per night 765
$170 - $220 per night 336
$220 + per night 219
Top Filters for Lisbon
Hotels 130
Breakfast Included 282
Apartments 1623
Very Good: 8+ 1018
Book without credit card 1
Hostels 131
Lisbon City Center 1183
Guesthouses 208
District
Lisbon Old Town 1282
Guests' Favorite Area 1488
TURIM Av. Liberdade Hotel ★★★★
8.1 Very Good
7,192 reviews
Executive Twin Room
Price for 6 nights
$1,879 $991
Only 2 left
Lisbon World Hostel
8.0 Very Good
276 reviews
6-Bed Mixed Dormitory Room
Price for 6 nights
$89
Only 2 left
America Diamonds Hotel
★★★
8.0 Very Good
4,857 reviews
Family Room (2 Adults + 2 Children)
Search landmarks, addresses & more..
Av. João XXI
Instituto Superior Técnico
Museu Calouste Gulbenkian
Universidade NOVA de Lisboa
Park
Parque Eduardo VII
Cinemateca Portuguesa
Lisbon
Igreja de São Vicente de Fora
Panteão
da Estrela
Sé de Lisboa
Terreiro do Paço
Map data ©2019 Google, Inst. Geogr. Nacional
500 m
Terms of Use
Report a map error

EDBT 2019

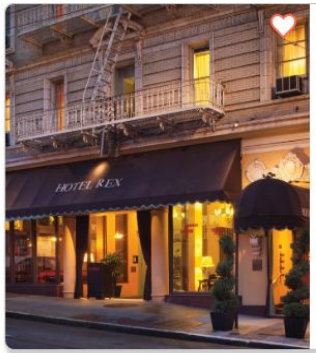
Voyageur: An Experiential Travel Search Engine.
WWW 2019 demonstration screenshot.

- Powered by our Subjective Database engine.

Hotel Recommendations

has really clean room × less than $150 × search for hotels Add

Hover on subjective filters to see people's real reviews.



Hotel Rex

- Good value though and definitely recommended as a base for seeing downtown San Francisco.
- Very well priced and a thumbs up if you're looking for a good value boutique hotel!
- If you like small, unique hotels, the Hotel Rex is a good bet.

See More

61% of reviews are related to this query

"clean room"

"nice decorated rooms"

"very comfortable lounge lobby"

"immaculate room"

Columbus Motor Inn

- I would recommend this property for someone with a car looking for a good value.
- I would definitely recommend it to a friend, best value in town!!!
- The rates offer excellent value - especially if you are using a car.
- If you' re looking for clean, spacious rooms, and a good location- the Columbus Motor Inn is a good choice.

See More

30% of reviews are related to this query

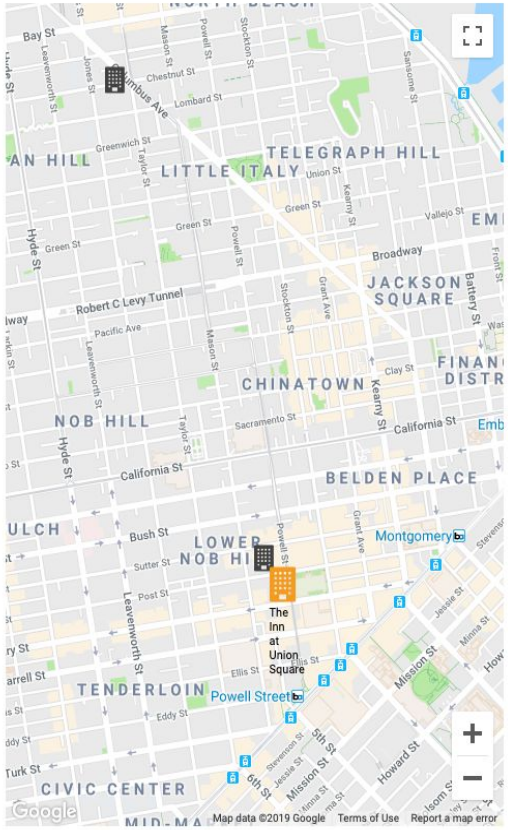
"very nice rooms"

"immaculate room"

"clean motel"

"clean place"

"immaculate everything"



Today's hotel search systems

- Exposes as many attributes as they think important.
- Schema is fixed a priori.
- Results are objective:
  - A hotel either satisfies the objective criteria or not.

Example subjective queries in different domains

Hotels: "*Hotels with clean rooms near IST congress center in Lisbon, Portugal.*"

Restaurant: "*Restaurants which are romantic and decently priced*."

Jobs: "*Companies working on cutting edge AI tech. and offers good benefits.*"

Criteria for search are subjective

- ***Subjective***: based on or influenced by personal feelings, tastes, or opinions.

- J. McAuley and A. Yang. *Addressing Complex Subjective Product Related Queries with Customer Reviews*. WWW 2016.

  "*around 20% of [product] queries were labeled as being 'subjective' by workers.*"

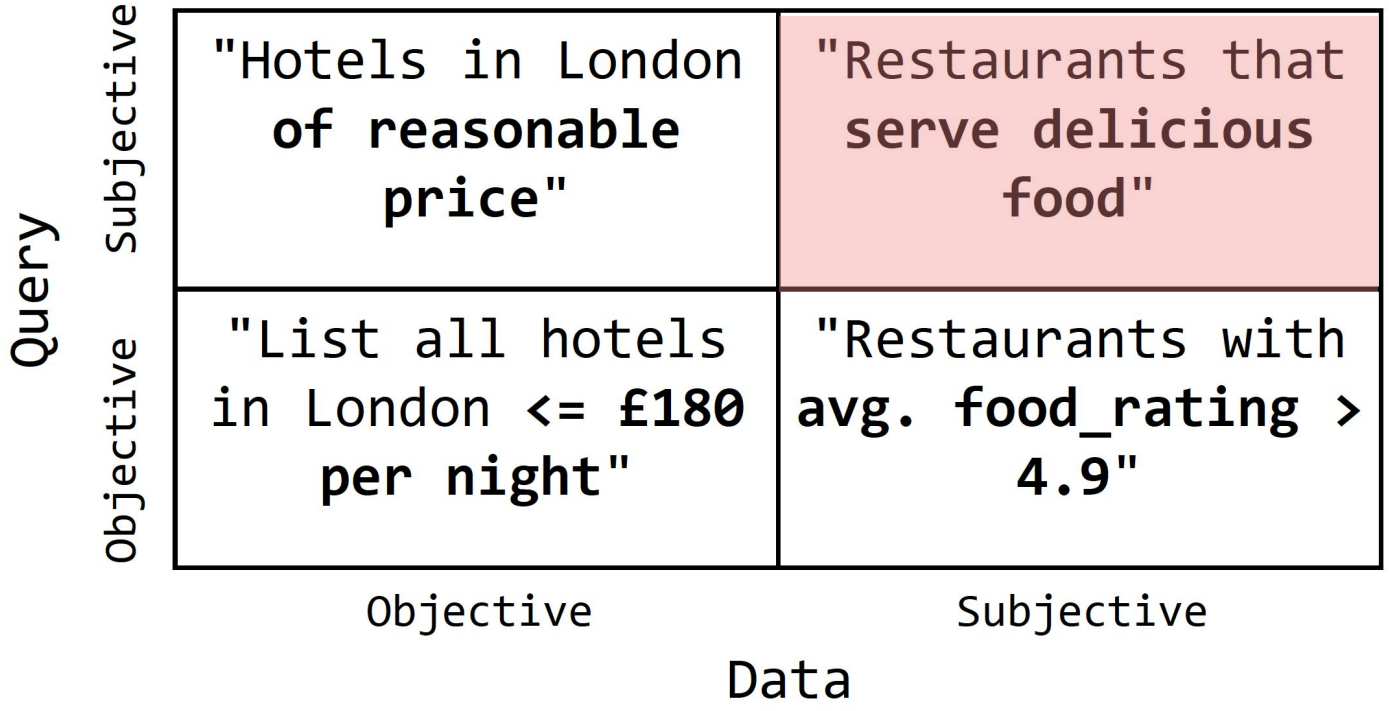
Criteria for search are subjective

Y.Li, A.Feng, J.Li, S.Mumick, A.Halevy, V.Li, T. *Subjective Databases*, ArXiv 2019.

Domain	%Sub Attr	Some examples
Hotel	69.0%	cleanliness, food, comfortable
Restaurant	64.3%	food, ambiance, variety, service
Vacation	82.6%	weather, safety, culture, nightlife
College	77.4%	dorm quality, faculty, diversity
Home	68.8%	space, good schools, quiet, safe
Career	65.8%	work-life balance, colleagues, culture
Car	56.0%	comfortable, safety, reliability

A.Halevy. *The Ubiquity of Subjectivity.* IEEE DEB 2019.

Subjective/objective data and queries



Subjective queries against subjective data

Why is this a hard problem?

- Experiences are subjective and personal.
- Specified in a variety of ways.
  - Often in text, not in a database.
  - Their meanings are often imprecise.
  - Hard to model in a database.

Subjective Data: Examples

Convenient location|Pleasant staff|Good ac in room|Clean room|Comfy beds||But **Breakfast** average||The 'mini gym' is a disgrace |4 machines |Runner ancie

Had a great time overall. **Breakfast** was very delicious and had many options. Slept pretty well. Only issue we had was that the air conditioning needs to be impeoved in

Front desk very helpful. Food was great for **breakfast**. Staff in restaurant were amazing efficient super friendly, remember you. Beds were very comfortable. The pillo

"Nice, but no more. For a bit more money you can move to a Hilton on Union Square. Lobby feels modern, room a little less so, but pleasant. Standard king on upper floor is not spacious - 1 or 2 night max. Staff is pleasant. Comfortable bed, quiet. Location good but not great. On the edge of walkability, on the edge of sketch (not that SF ever really gets sketchy in main areas, just … disheveled). Decent breakfast. Good for a fly in / fly out biz trip."

Beach is too sandy. But clear water.

Review of **Tumon Beach**

 Reviewed August 31, 2012

It's a great beach, just too sandy. Very nice for swimming. Water was so clean and crystal clear. Overall nice experience on Tumon Beach. Recommend travelers to visit anyways.

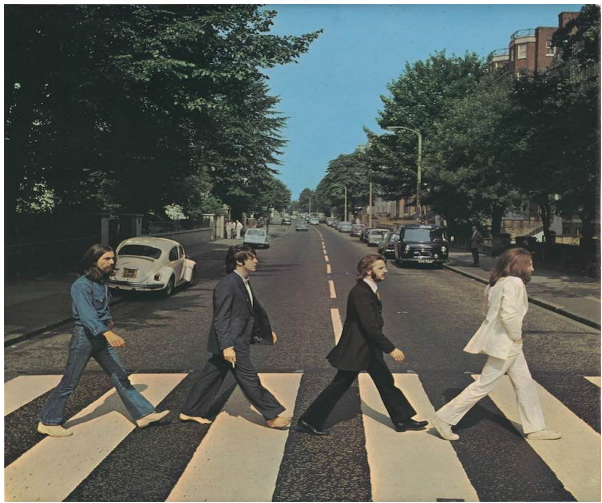
Date of experience: May 2012

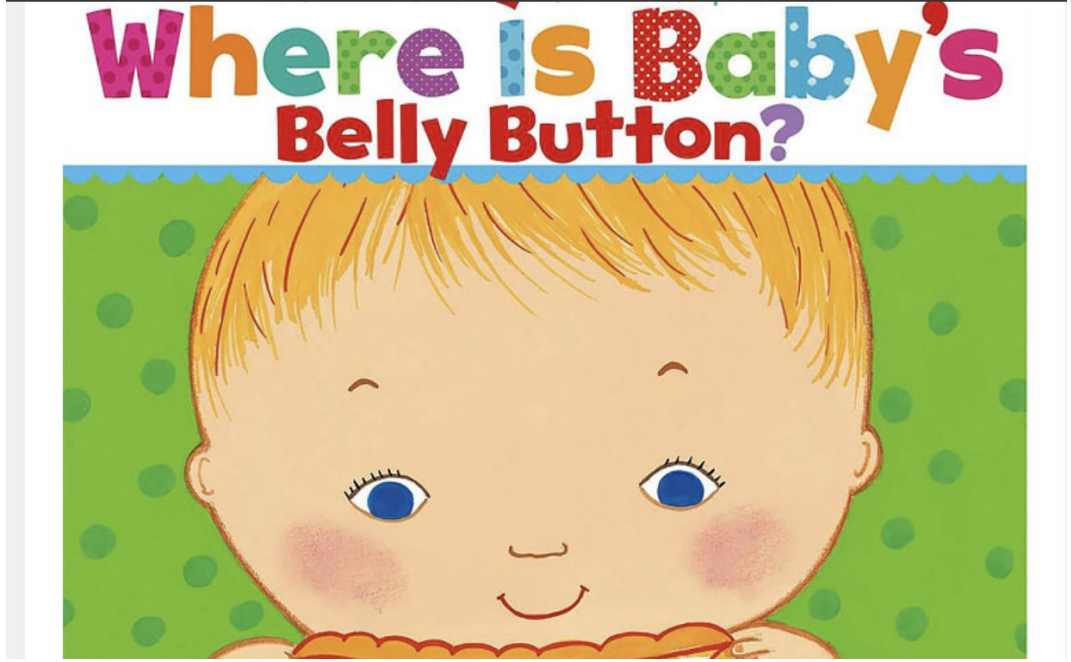
"Disappointed by the lack of zebras"

Reviewed 12 March 2015

I was told to check out the zebras crossing Abbey Road whilst in London but didn't see any. Perhaps you have to come at night to see them. There were lots of people there looking for them but didn't see a single one! I'd recommend London Zoo at Regent's Park over Abbey Road for zebras.

Was this review helpful? Yes 9

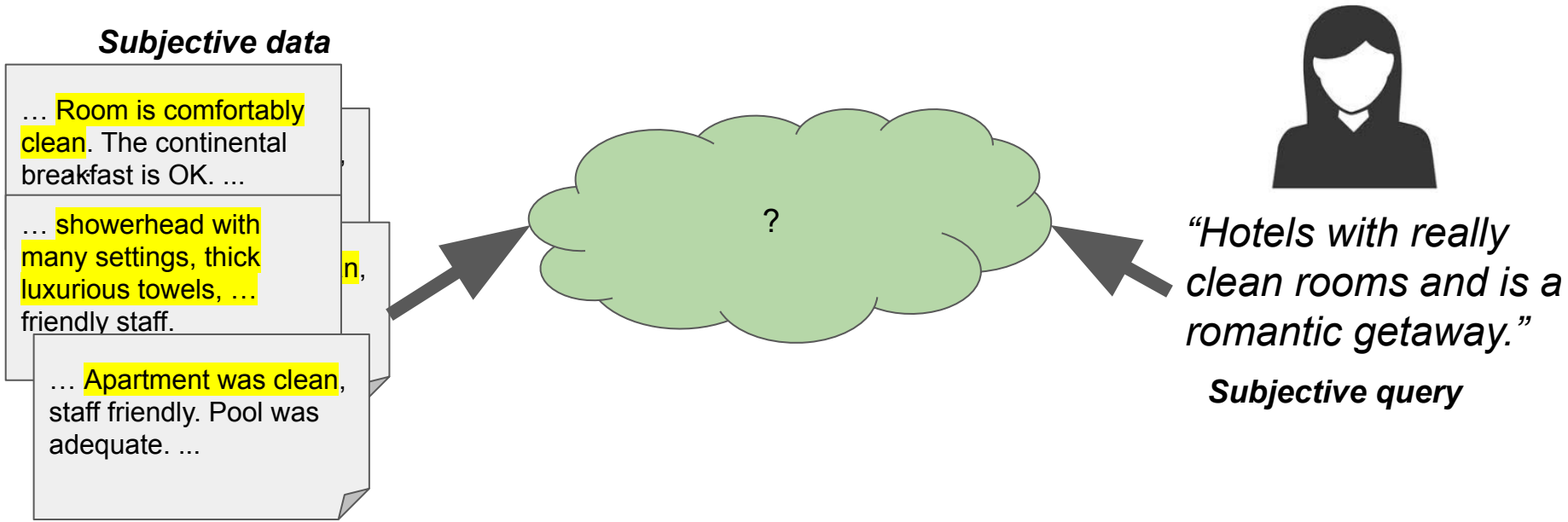




★☆☆☆☆ **DO NOT buy this book, you can SEE the ending right on the cover!**, April 19, 2012

Subjective queries against subjective data

Why is this a hard problem?



The remainder of this talk

OpineDB

Y.Li, A.Feng, J.Li, S.Mumick, A.Halevy, V.Li, T. *Subjective Databases*, ArXiv 2019.

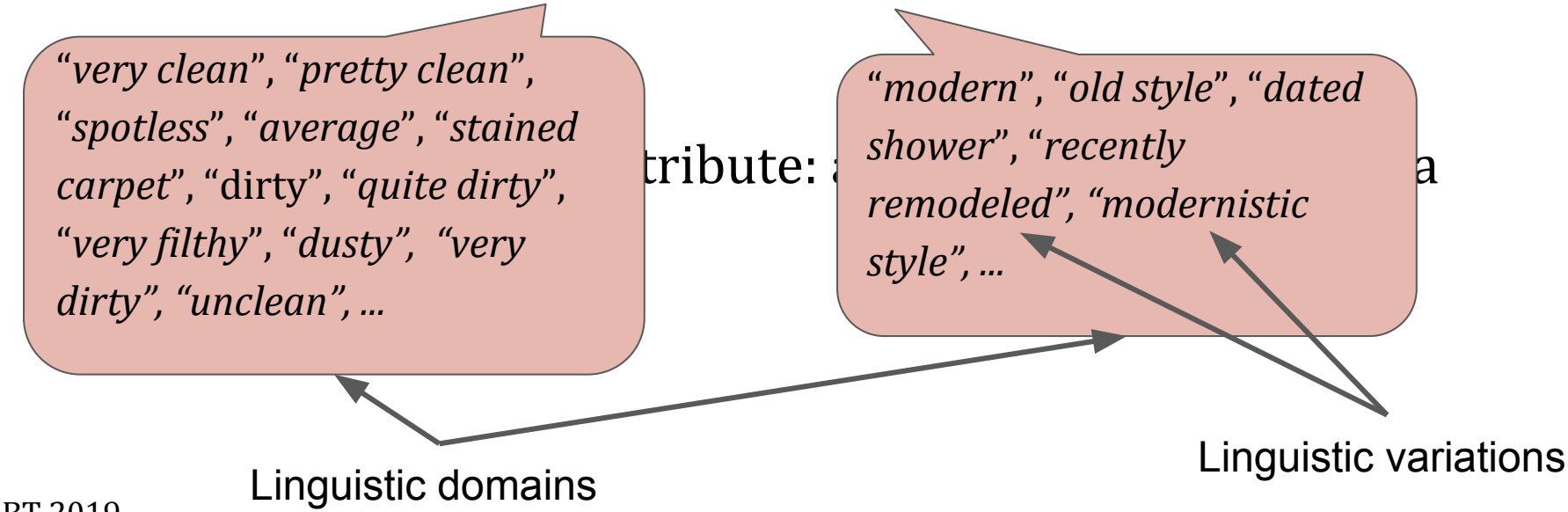
- Subjective database model
- Processing subjective database queries
- Building subjective databases
- Concluding remarks
- Demonstration screenshots

Subjective database schema

- Relation schemas $R(K, A_1, \ldots, A_n)$.
- Objective attributes and subjective attributes
  - values are based on facts, indisputable
  - values are influenced by personal beliefs or feelings

Subjective attributes

Hotel (hotelname, capacity, address, price_pn,
*room_cleanliness, *bathroom, *service, *comfort)

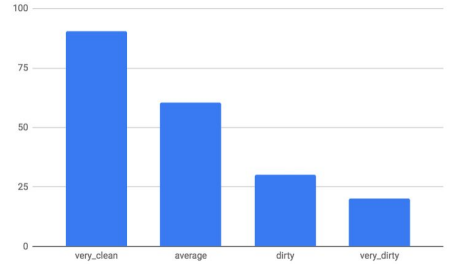


Linguistic domain and marker summaries

- Linguistic domain (LD) of an attribute
  - a set of short linguistic variations that describe the attribute.
- Marker
  - a word in the LD
- Marker summary:
  - a set of markers in the LD representative of the LD
- Room_cleanliness[“*very clean*”, “*average*”, “*dirty*”, “*very dirty*”]

Marker Summaries

- Linearly-ordered
  - Markers form a linear-scale.
  - Room_cleanliness[“*very clean*”, “*average*”, “*dirty*”, “*very dirty*”]



“*rooms are pretty clean*”

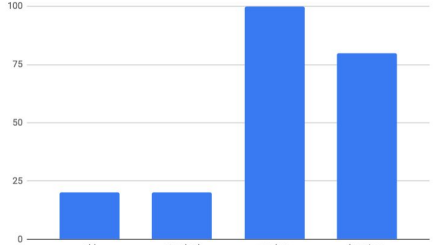
0.5

0.5

“*extravagant old-fashioned bathrooms*”

1

1



- Categorical
  - No two markers of the marker summary form a linear scale.
  - Bathroom[“*old-fashioned*”, “*standard*”, “*modern*”, “*luxurious*”]

Subjective queries against subjective data

Subjective queries against subjective data

Subjective data

… Room is comfortably clean. The continental breakfast is OK. ...

… showerhead with many settings, thick luxurious towels, … friendly staff.

… Apartment was clean, staff friendly. Pool was adequate. ...

Hotel (hotelname, capacity, address, price_pn,
**room_cleanliness*, **bathroom*,
**service*, **comfort*)

Marker summaries

Room_cleanliness
[*very_clean, average, dirty, very_dirty*]
Bathroom
[*old, standard, modern, luxurious*]
Service
[*exceptional, good, average, bad, very_bad*]
Bed
[*very_soft, soft, firm, very_firm, ok, worn_out*]

Linguistic domains ...

"Hotels with really clean rooms and is a romantic getaway."

Subjective query

Subjective database queries

“Find hotels with cost less than $150 per night, has really clean rooms and is a romantic getaway.”

select * ***from*** *Hotels*
where *price_pn < 150 and*
“ has really clean rooms ” and
“ is a romantic getaway ”

Lots of related work (NLP and DB)

- Natural language interfaces to databases
  - Parse natural language into semantic structure (SQL).
  - Parsing objective queries.

V. Zhong, C.Xiong, R.Socher. Seq2SQL: *Generating structured queries from natural language using reinforcement learning*. arXiv 2017.
F.Li, H.V.Jagadish. *Understanding Natural Language Queries over Relational Databases*. SIGMOD Record 2016.
A.Simitsis, G.Koutrika, Y. Ioannidis. *Précis: from unstructured keywords as queries to structured databases as answers*. VLDBJ 2008.
Yael Amsterdamer, Anna Kukliansky, Tova Milo: *A Natural Language Interface for Querying General and Individual Knowledge*. PVLDB 2015.
S. Iyer, I. Konstas, A. Cheung, J. Krishnamurthy, L. Zettlemoyer. *Learning a neural semantic parser from user feedback*. ACL 2017.
A.Popescu, O.Etzioni, H.Kautz. *Towards a theory of natural language interfaces to databases*. IUI 2003.

And more!

Subjective database queries

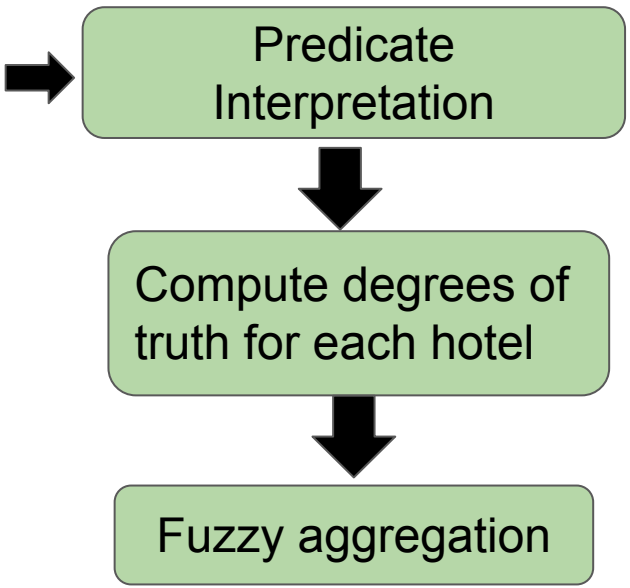
“Find hotels with cost less than $150 per night, has really clean rooms and is a romantic getaway.”

select * ***from*** *Hotels*
where *price_pn < 150 and*
“ has really clean rooms ” and
“ is a romantic getaway ”

Processing subjective database queries

select * **from** *Hotels*
where *price_pn < 150 and*
"has really clean rooms" and
"is a romantic getaway"

" has really clean rooms ",
" is a romantic getaway "



"has really clean rooms" → **0.7** room_cleanliness["very clean"]

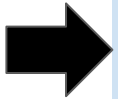
"is a romantic getaway" →
0.63 Service["exceptional"] ⊕
0.82 Bathroom["luxurious"]

Query result:
1. Holiday Hotel 2. Inn Hotel ...

Predicate interpretation

Interpret each predicate into a fuzzy logic expression over attribute markers.

select * **from** *Hotels h*
where *price_pn < 150*
and
“has really clean rooms”
and
“is a romantic getaway”



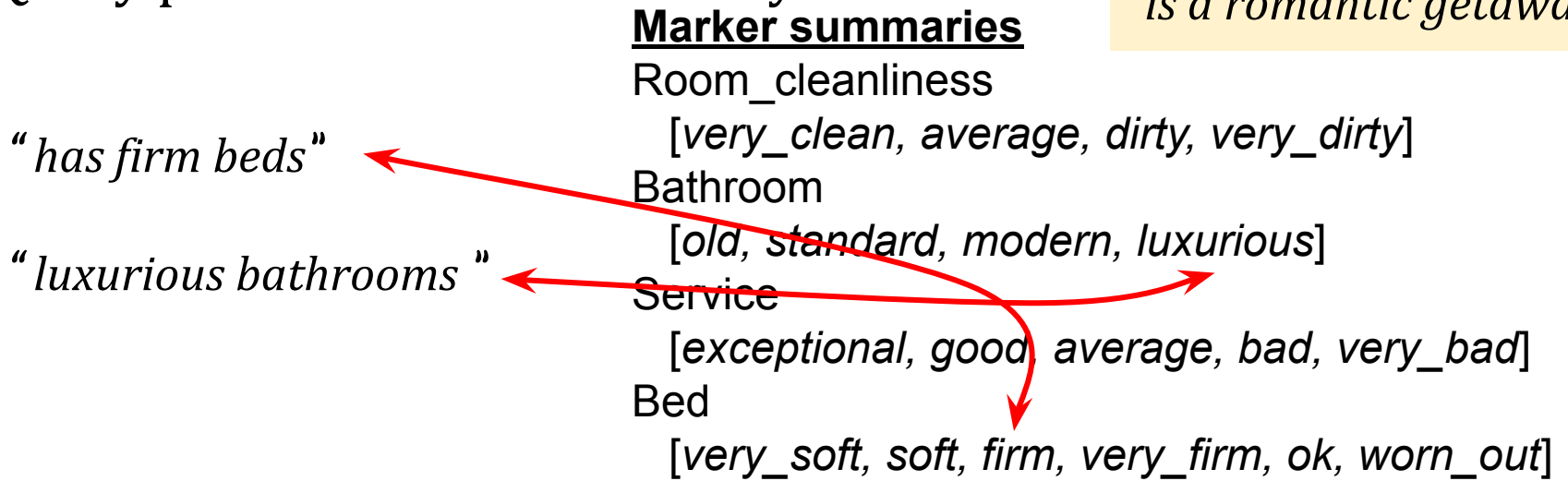
select * **from** *Hotels h*
where *price_pn < 150*
⊗
h.room_cleanliness ≋ “really clean”
⊗
(h.service ≋ “exceptional” ⊕
h.bathroom ≋ “luxurious”)

Predicate interpretation: The easy case

- **Problem**: Given a query predicate *p*, find the marker(s) that best represent *p*.

Query predicates match directly to markers.

"has really clean rooms" ?
"is a romantic getaway" ?



Predicate interpretation: The harder case

Query predicates have arbitrary phrases.

- Word embedding method:
  - Find variations similar to *p* based on its word embedding.
- Co-occurrence method:
  - Find a marker whose linguistic variations frequently co-occur with *p* in the reviews.
- When all else fails … text-retrieval method.

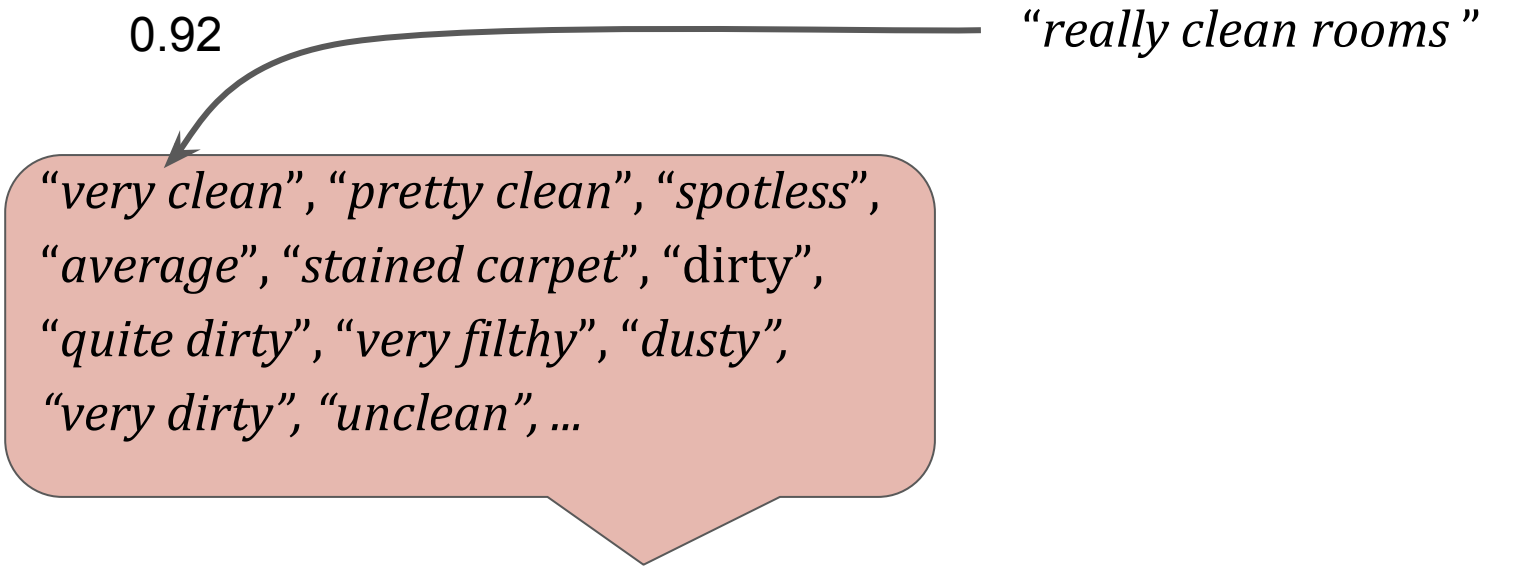
Predicate interpretation: word embedding method

- Find best semantically matching variations to *p*.

$$\text{rep}(p) = \sum_{w \in p} \text{w2v}(w) \cdot \text{idf}(w) \qquad \text{similarity}(q, p) = \cos(\text{rep}(q), \text{rep}(p))$$

  - *p* = query predicate, w2v(*w*) = word vector of *w*,
  - idf(*w*) = inverse document frequency of *w* in the review corpus.
  - Interpretation: corresponding marker of *q* with highest similarity score to *p* above a certain threshold.

Word embedding method



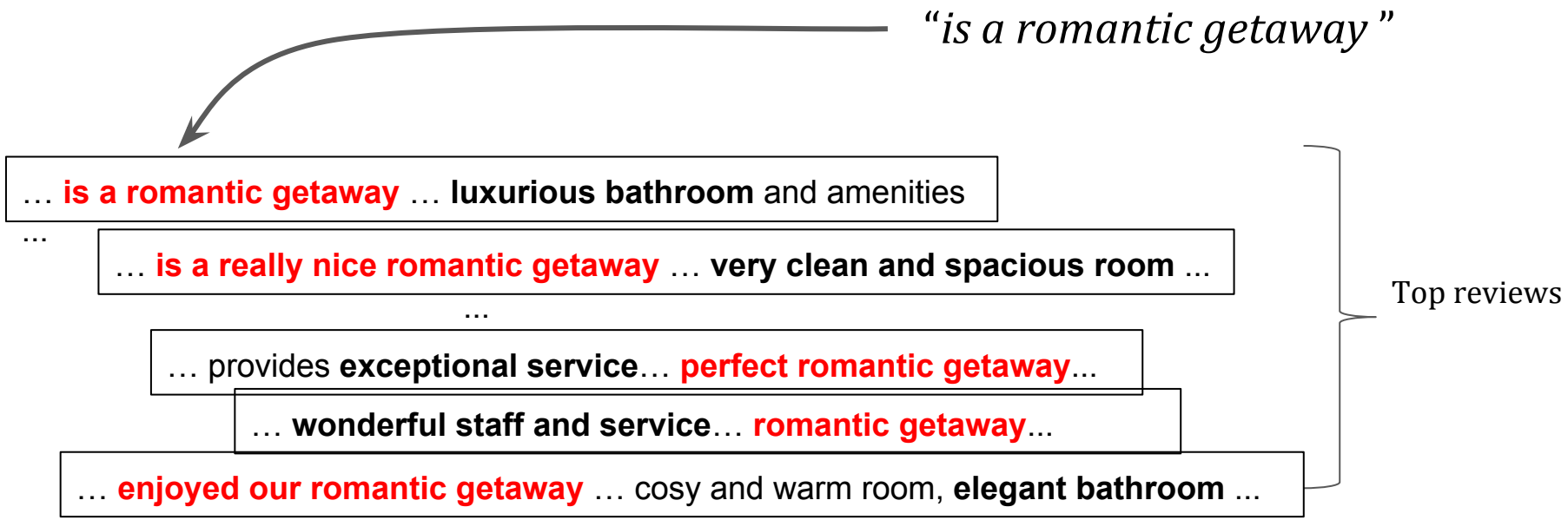
Predicate interpretation: co-occurrence method

- "*is a romantic getaway*"
  - does not match any linguistic variation well.
  - frequently co-occurs with "*excellent service*" or "*five-star bathrooms*".
- "*is a romantic getaway*" →
  Service["*exceptional*"] OR Bathroom["*luxurious*"]

Predicate interpretation: co-occurrence method

- Find top-k positive reviews where p occurs.
  - rankscore(d) = BM25(d,p) * senti(d)
- Find most correlated attributes A_1, …, A_n.
  - freq(A)*idf(A), highest TF-IDF scores.
  - freq(A): # linguistic variations of A_i that occur in top-k reviews.
  - $A_i.m_i$: m_i has highest # linguistic variations in top-k reviews.
- Build a disjunctive expression out of $A.m$.

Co-occurrence method



Example output of co-occurrence method

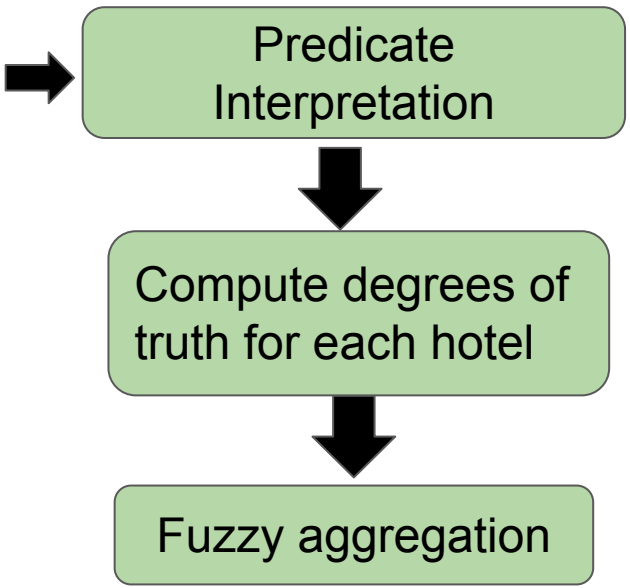
Predicate	**Top-1 interpretation**
"*for our anniversary*"	Staff["*great staff*"]
"*multiple eating options*"	Food["*great food*"]
"*close to public transportation*"	Location["*great location*"]
"is a romantic getaway"	**Top-2 interpretations:** Service[*"exceptional"*] *OR* Bathroom[*"luxurious"*]

When all else fails … Text-retrieval method

- Apply traditional IR techniques
  - when both word embedding method and co-occurrence method fail.
- Represent reviews of each hotel by a single document D (concatenate all reviews).
- Compute BM25(D, p).

Processing subjective database queries

```
select  *
where   price_pn < 150 and
        " has really clean rooms " and
        " is a romantic getaway "
```



" has really clean rooms ",
" is a romantic getaway "

"has really clean rooms" → **0.7** room_cleanliness["very clean"]

"is a romantic getaway" →
0.63 Service["exceptional"] ⊕
0.82 Bathroom["luxurious"]

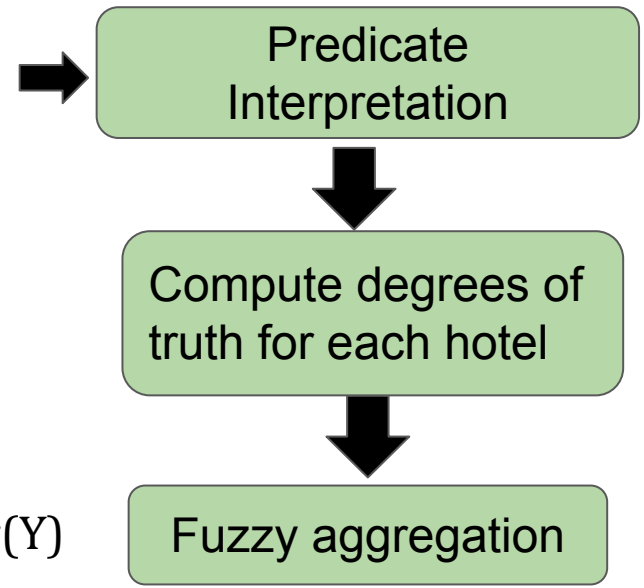
Query result:
1. Holiday Hotel 2. Inn Hotel ...

Compute degrees of truth

- Computes a degree of truth for each interpreted predicate.
  - How well does the marker summary represent the query predicate?
- Train a Logistic Regression model on triples:
  - (*room_cleanliness*, “room is really clean”) → 0/1
  - plus other features
  - Loss function used as degree of truth.

Processing subjective database queries

```
select  *
where   price_pn < 150 and
        " has really clean rooms " and
        " is a romantic getaway "
```



" has really clean rooms ",
" is a romantic getaway "

"has really clean rooms" → **0.7** room_cleanliness["very clean"]

"is a romantic getaway" →
0.63 Service["exceptional"] ⊕
0.82 Bathroom["luxurious"]

- Multiplication variant
  - X ⊗ Y = deg(X) * deg(Y)
  - NOT X = 1-deg(X)
  - X ⊕ Y = (1-(1-deg(X)*(1-deg(Y))

Query result:
1. Holiday Hotel 2. Inn Hotel ...

Fuzzy logic versus thresholds

(h.price < $150) > 0.9 ⊗
(h.room_cleanliness ≋ “really clean” > 0.7) ⊗
(h.style≋”luxurious” > 0.6)

- extremely clean but not so luxurious?
- really clean and very luxurious but costs $159 per night?

Subjective queries against subjective data

Building subjective databases

- Construct linguistic domains from reviews.
  - Extract aspects + opinions.
  - High-performing DL systems require a lot of training data.
    - Repeated for each domain.
  - Use pre-trained BERT [DCLT18] on less training data.
    - F1 score of 75.6%. Better than 73.3% [WPDX16-17].

Lots of related work (NLP/Data Mining/DB)

- Aspect extraction, opinion mining, sentiment analysis, identifying/extracting subjective expressions.

J.Wiebe.++ (since 1999)

B.Liu *Sentiment Analysis and Opinion Mining*" Morgan Claypool, 2012.

W.Wang, S.J.Pan, D.Dahlmeier, and X.Xiao. *Recursive neutral conditional random fields for aspect-based sentiment analysis*. EMNLP 2016

W.Wang, S.J.Pan, D.Dahlmeier, and X.Xiao. *Coupled Multilayer attentions for co-extraction of aspect and opinion terms*. AAAI 2017.

L.Zhang,S.Wang, and B.Liu. *Deep learning for sentiment analysis: A survey*. Wiley Interdiscip. Rev. Data Mining Knowledge Discovery. 2018

H. Xin, R. Meng, L. Chen. *Subjective Knowledge Base Construction Powered By Crowdsourcing and Knowledge Base*. SIGMOD 2018.

:

Building subjective databases

- Schema designer designs subjective attributes.
- Map linguistic variations to subjective attributes.
  - Text classification.
  - Labeled data obtained by seed expansion.
    - E = {room, bedroom} **+ suite, apartment**
    - P = {clean, dirty, very clean, very dirty, stained} **+ filthy, dusty**
      - Every (e,p) maps to room_cleanliness

Building subjective databases

- Define markers.
  - Linearly-ordered domains.
    - Sort linguistic variations by sentiment analysis.
  - Categorical domains.
    - *k*-means clustering.
- Compute marker summaries.
  - Aggregate linguistic variations from reviews to markers.

Key takeaways

- Language, by nature, is subjective and imprecise.
- Lots of work on extracting subjective expressions and opinions etc. from NLP/IR/Data Mining community.
- Novelty in OpineDB :
  - Manage subjectivity on both ends: data and queries.
  - Need to aggregate and join.
  - We have a schema! Linguistic domains, marker summaries.

Future work

- Consider user profiles and preferences.
- Point out interesting facts, summarize, and explain observations.

Voyageur: An Experiential Travel Search Engine. WWW 2019 demonstration screenshot.

- Powered by our Subjective Database engine.

Hotel Recommendations

has really clean room × less than $150 × search for hotels Add

Hover on subjective filters to see people's real reviews.

Hotel Rex

- Good value though and definitely recommended as a base for seeing downtown San Francisco.
- Very well priced and a thumbs up if you're looking for a good value boutique hotel!
- If you like small, unique hotels, the Hotel Rex is a good bet.

See More

61% of reviews are related to this query

"clean room"

"nice decorated rooms"

"very comfortable lounge lobby"

"immaculate room"

Columbus Motor Inn

- I would recommend this property for someone with a car looking for a good value.
- I would definitely recommend it to a friend, best value in town!!!
- The rates offer excellent value - especially if you are using a car.
- If you' re looking for clean, spacious rooms, and a good location- the Columbus Motor Inn is a good choice.

See More

30% of reviews are related to this query

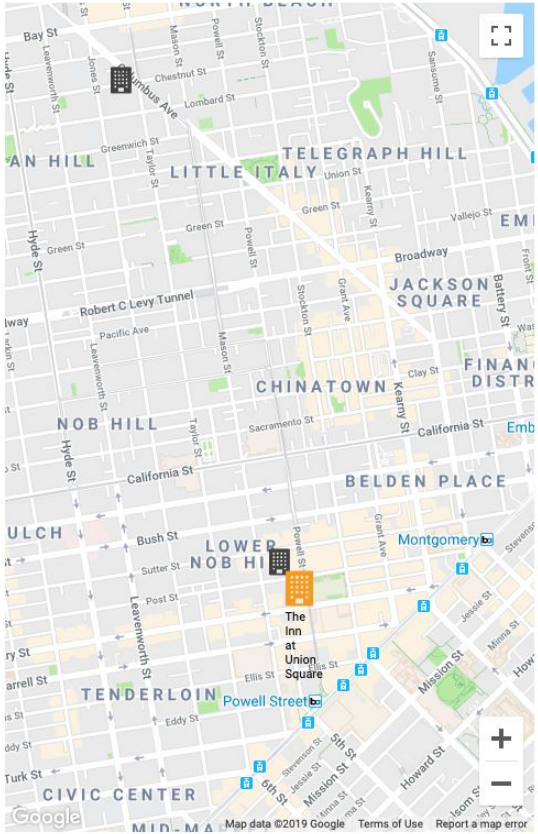
"very nice rooms"

"immaculate room"

"clean motel"

"clean place"

"immaculate everything"



Hotel Recommendations

has really clean room × | **is a romantic getaway** × | less than $150 × | search for hotels | Add

Hover on subjective filters to see people's real reviews.

Castle Inn

- But if you need friendly and really good value then look no further.
- Excellent value.
- We recommend Castle Inn as a best value stay in SF and would choose it again on our next visit to San Francisco.
- We really recommend this hotel!

See More

73% of reviews are related to this query

"great service"

"great stay"

"luxurious ambiance"

"perfect hotel"

"perfect inn"

Hotel Rex

- Good value though and definitely recommended as a base for seeing downtown San Francisco.
- Very well priced and a thumbs up if you're looking for a good value boutique hotel!
- Good inexpensive food, free wireless... can highly recommend it.

See More

49% of reviews are related to this query

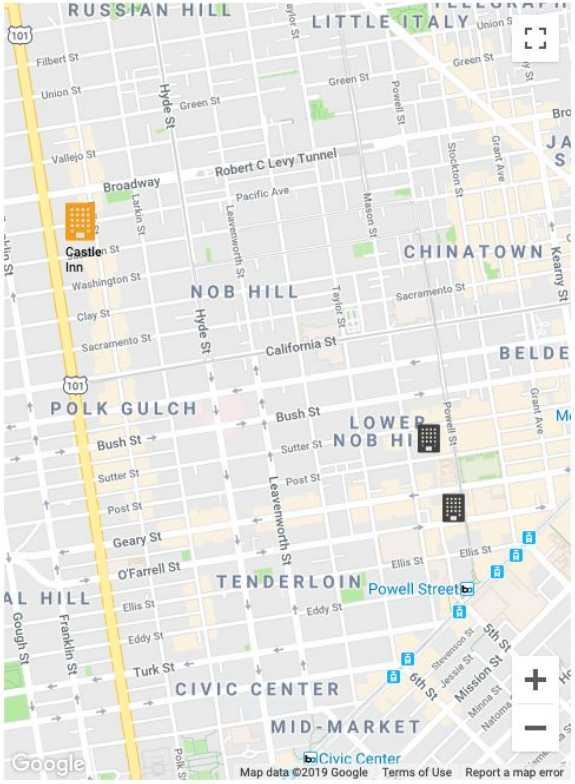
"great hotel"

"great time"

"great experience"

"very nice place"

"wonderful everything"



Ultimate search experience

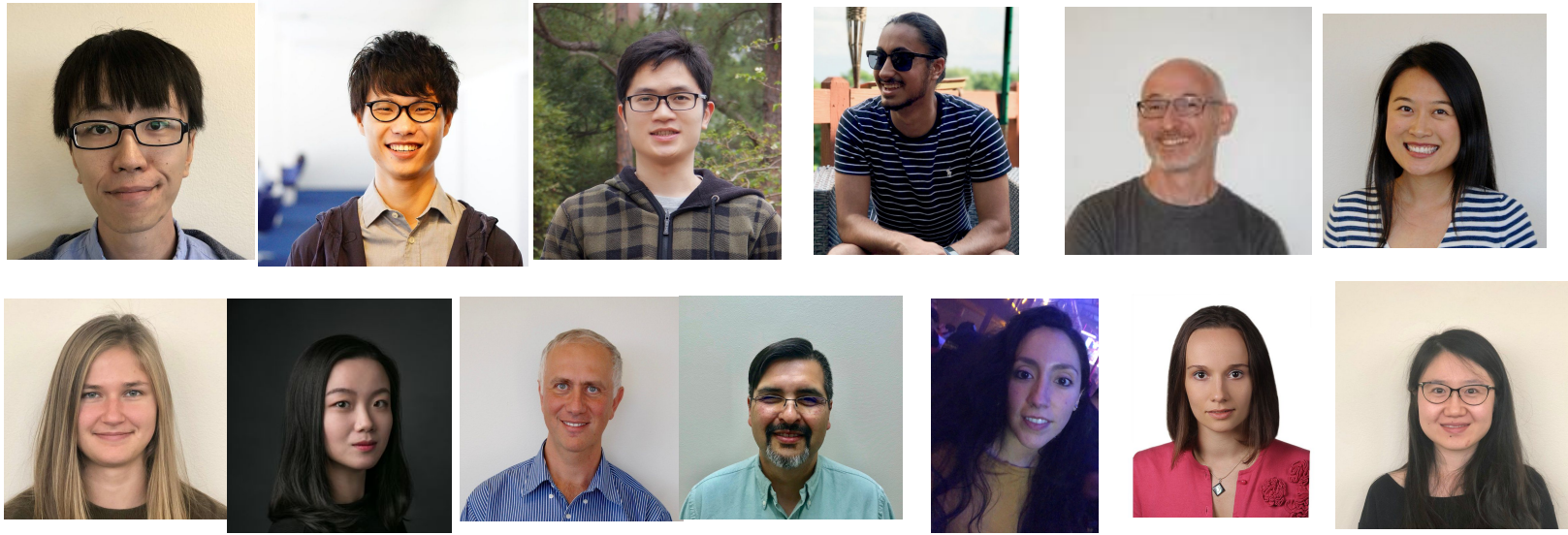
Help users make decisions based on their experiential requests.

My kids have a week off on Feb 19. I want to have a good time with them. What should I do?

I like digital design and I am pretty good at Math and Biology. What should I major in college?

Subjective database team

Yuliang Li, Aaron Feng, Jinfeng Li, Saran Mumick, Alon Halevy, Vivian Li
Development & UI: Sara Evensen, Huining Liu, George Mihaila, John Morales, Natalie Nuno, Kate Pavlovic, Xiaolan Wang



END