Web Information Systems @TU Delft

Delft University of Technology

Sparrows and Owls:

Characterisation of Expert Behaviour in StackOverflow

Jie Yang, Alessandro Bozzon

Ke Tao, Geert-Jan Houben

OUTLINE

- **Expertise in Social Networking Systems**
  - Need and Challenges
- **Expertise Identification**
  - Expertise Metric
- **Expertise Characterisation**
  - Contribution
  - Preference
  - Temporal Evolution

OUTLINE

- **Expertise in Social Networking Systems**
  - Need and Challenges
- **Expertise Identification**
  - Expertise Metric
- **Expertise Characterisation**
  - Contribution
  - Preference
  - Temporal Evolution

EXPERTISE AS A USER PROPERTY

Sentiment

Culture

Location

Personality

Expertise

Quora

ANSWERS

stackoverflow

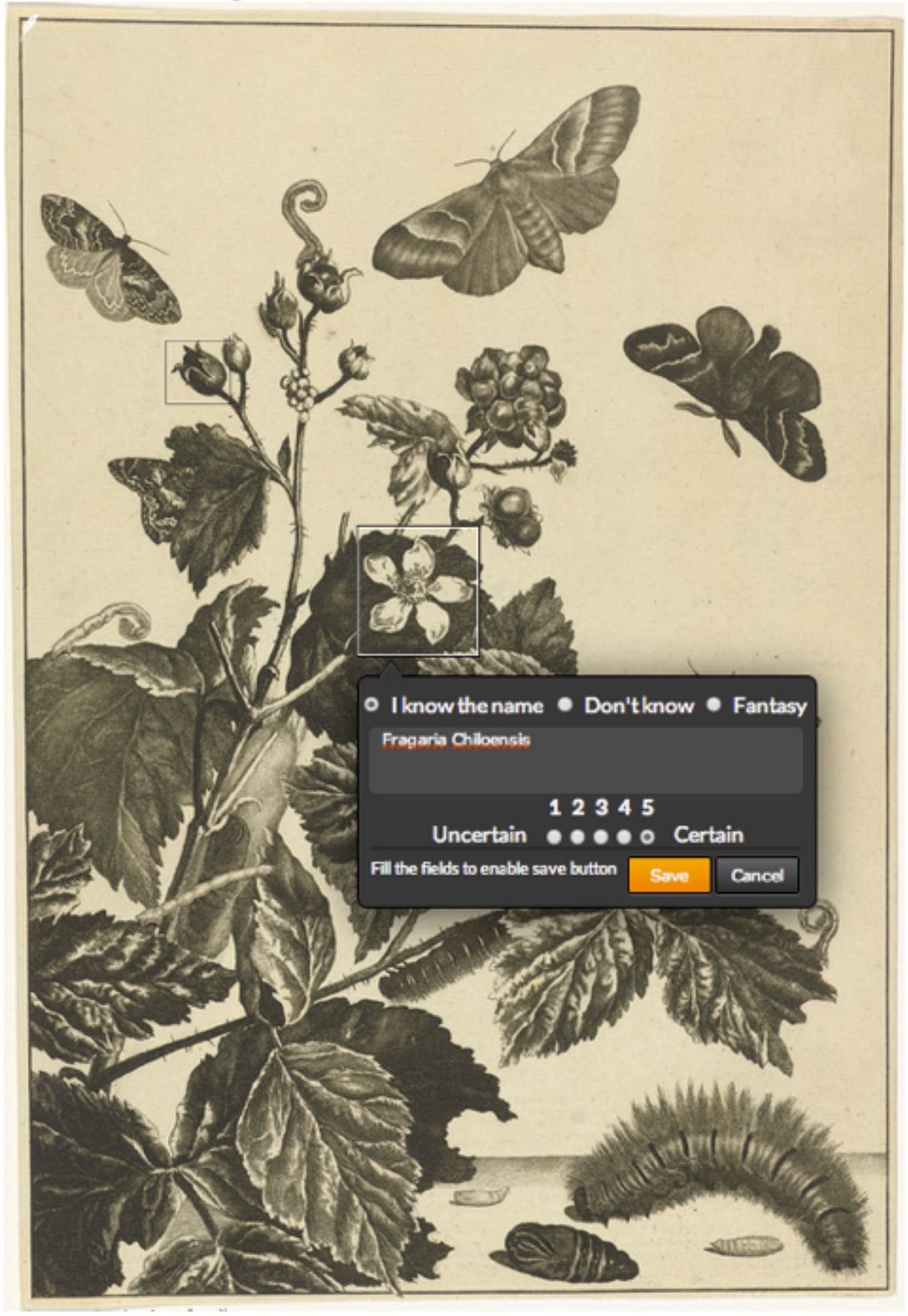
reddit

Why Expertise

APPLICATIONS

- Expert finding system
- Knowledge Intensive Task
  - Recommendation
  - Sourcing

Click for the full-size image



How many FLOWERS are in this image?*

10

Count every flower and flower bud you see on the image. Click for the full-size image if needed.

Tag each individual FLOWER by drawing a bounding box around it.

For each box fill in the fields of the popup.

Number of bounding boxes: 2

For how many of the FLOWERS you identified on this image did you provide a FLOWER NAME?*

- 0
- 1
- 2
- 3
- 4
- 5
- 6
- More

Please provide a reference on how you got the answer to these questions*

Describe the search strategy you used, or the website you used to find the flower name, or the reason you already knew the names of the flowers.

stackoverflow

Questions | Tags | Users | Badges | Unanswered | Ask Question

All Questions newest | 446 featured | frequent | votes | active | unanswered

7,488,117 questions

8006 votes · 9 answers · 442k views

Why is processing a sorted array faster than an unsorted array?

Here is a piece of C++ code that seems very peculiar. For some strange reason, sorting the data miraculously makes the code almost six times faster: #include <algorithm> #include <ctime> ...

java | c++ | performance | optimization | branch-prediction

asked Jun 27 '12 at 13:51 GManNickG 152k ● 19 ● 258 ● 39

53 votes · 22 answers · 784k views

...I change the message? I have not yet pushed the

git | git-rewrite-history

4643 votes · 25 answers · 1.3m views

Undo the last Git commit?

I accidentally added the wrong directory ... directory containing the .class file. How can ...

git

...10 users 30%

Peter Mortensen

Commun... Bulletin

Jobs near you

.NET Web Developer
Clipster
Breda, Netherlands
asp.net | c#

Big Data Scientist / Engineer at

Domain specific

Highly active

Engagement: "Reputation"

Why StackOverflow

ACTIVENESS = EXPERTISE?

- Existing Metrics
  - #answers
  - reputation (mostly got from voting's for answers)
- Zscore (#answers-#questions)

All biased to user activeness

ACTIVENESS VS. EXPERTISE

A Motivating Example

According to #votes

Activeness of an answerer

Question: C# to C++ ‘Gotchas’

Rank 1	C++ has so many gotchas…	2 answers
Rank 2	Garbage Collections!	26 answers
Rank 3	There are a lot of differences	175 answers
…	…	
Rank 14	The following isn’t meant…	24 answers

Best answer is provided by an inactive user

OUTLINE

- **Expertise in Social Networking Systems**
  - Need and Challenges
- **Expertise Identification**
  - Expertise Metric
- **Expertise Characterisation**
  - Contribution
  - Preference of Experts
  - Temporal Evolution

DATASET

Retrieved at September 2013

- Global: 5.6M questions, 10.3M answers, 2.3M users
- Topic C# related
- 472K questions, 1M answers, 117K answerers
- #answers per question： 2.27±1.74
- #answers per user: 9.15±76.66. (Power Law)

EXPERTISE METRIC:
MEAN EXPERTISE CONTRIBUTION (MEC)

what to consider

- Answer Utility
  - 1/(rank position) of an answer
  - measure the usefulness of answer to a question.
- Question Debatableness
  - #answers to a question
  - consider difficulty of the question

MEAN EXPERTISE METRIC

Factor1: Answer Utility

$$\mathtt{MEC}_{u,t} = \frac{1}{|Q_t^u|} \sum_{\forall q_i \in Q_{u,t}} \mathcal{AU}(u, q_i) * \frac{\mathcal{D}(q_i)}{\mathcal{D}_t^{avg}}$$

Answer Utility = 1/2

Question: C# to C++ ‘Gotchas’		
Rank 1	C++ has so many gotchas…	2 answers
Rank 2	Garbage Collections!	26 answers
Rank 3	There are a lot of differences	175 answers
…	…	
Rank 14	The following isn’t meant…	24 answers

MEAN EXPERTISE CONTRIBUTION

Factor2: Question Debatablenss

$$\mathtt{MEC}_{u,t} = \frac{1}{|Q_t^u|} \sum_{\forall q_i \in Q_{u,t}} \mathcal{AU}(u, q_i) * \frac{\mathcal{D}(q_i)}{\mathcal{D}_t^{avg}}$$

Debatableness = 14

Question: C# to C++ ‘Gotchas’		
Rank 1	C++ has so many gotchas…	2 answers
Rank 2	Garbage Collections!	26 answers
Rank 3	There are a lot of differences	175 answers
…	…	
Rank 14	The following isn’t meant…	24 answers

MEAN EXPERTISE CONTRIBUTION

Overall metric

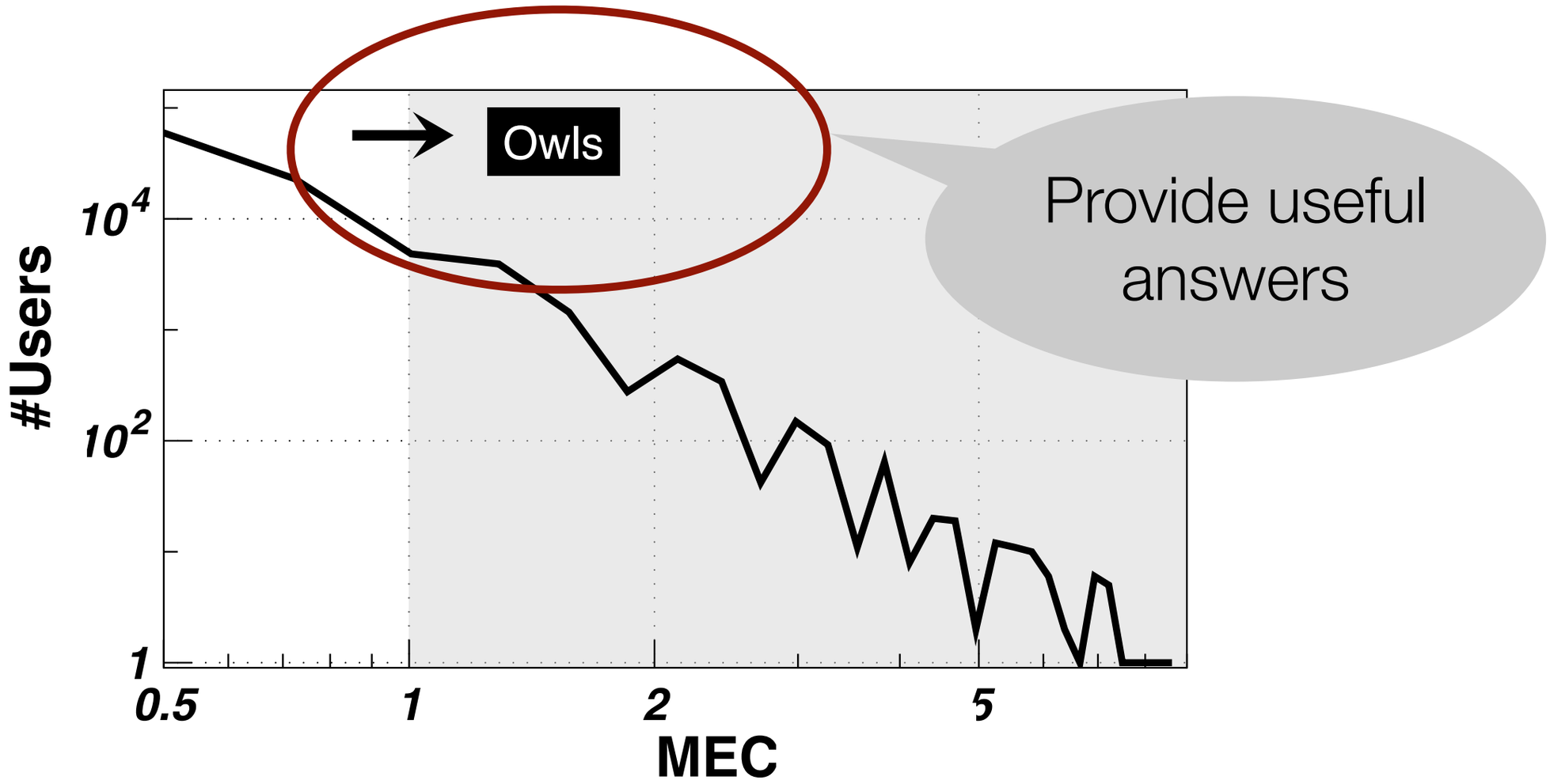
$$\text{MEC}_{u,t} = \frac{1}{|Q_t^u|} \sum_{\forall q_i \in Q_{u,t}} \mathcal{AU}(u, q_i) * \frac{\mathcal{D}(q_i)}{\mathcal{D}_t^{avg}}$$

Answer Utility * Debatableness = 7

Question: C# to C++ ‘Gotchas’		
Rank 1	C++ has so many gotchas…	2 answers
Rank 2	Garbage Collections!	26 answers
Rank 3	There are a lot of differences	175 answers
…	…	
Rank 14	The following isn’t meant…	24 answers

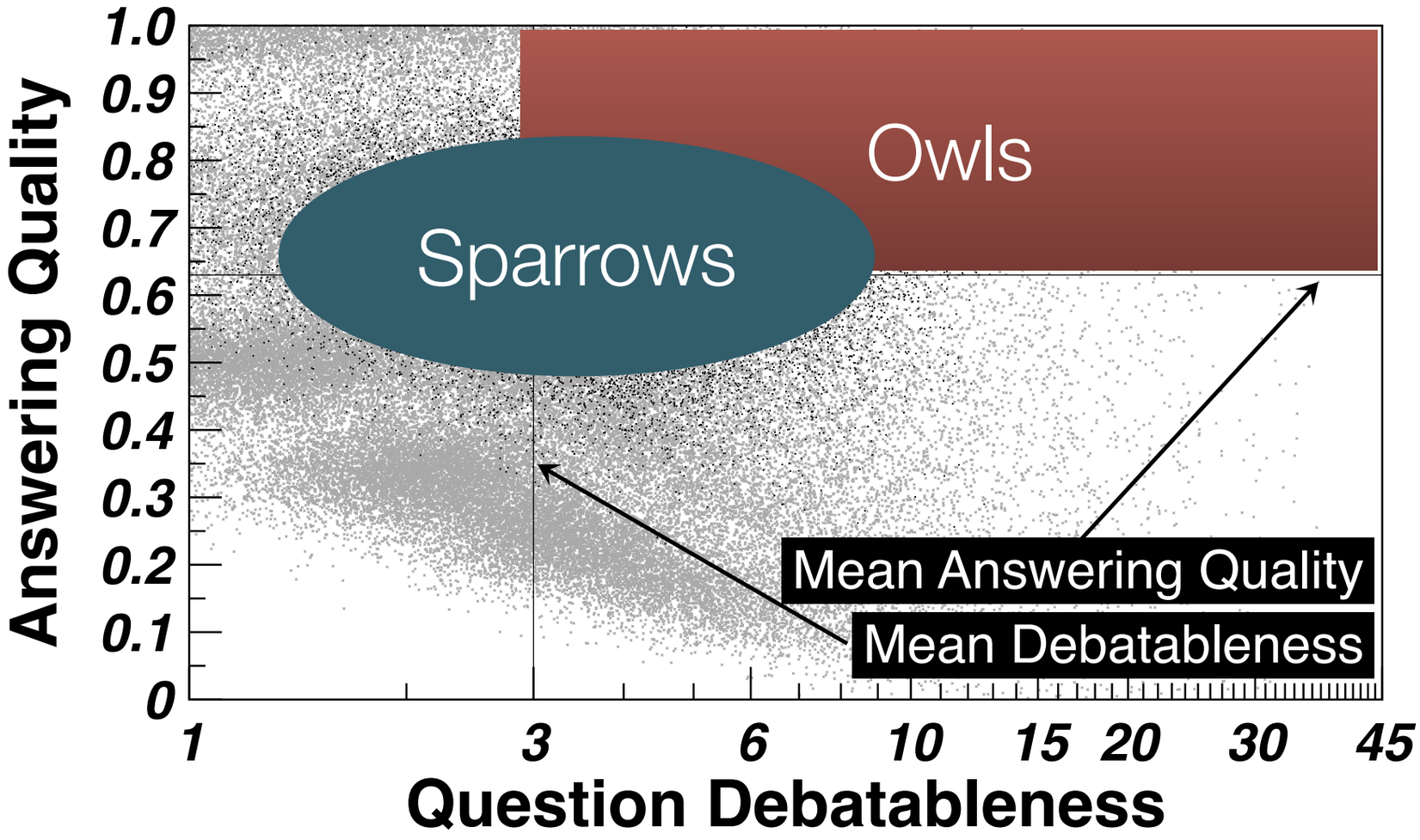
POWER LAW DISTRIBUTION OF MEC

Just like other statistics in StackOverflow



A small number of users have high MEC (provide useful answers), while others do not.

Owls: MEC >= 1
Sparrows: #answers >= 10
Both important!



Sparrows and Owls
9.9% Overlapping

OUTLINE

- **Expertise in Social Networking Systems**
  - Need and Challenges
- **Expertise Identification**
  - Expertise Metric
- **Expertise Characterisation**
  - Contribution
  - Preference
  - Temporal Evolution

A COMPARATIVE STUDY

3 Research Questions

RQ1. How do CONTRIBUTIONS from Sparrows and Owls differ?

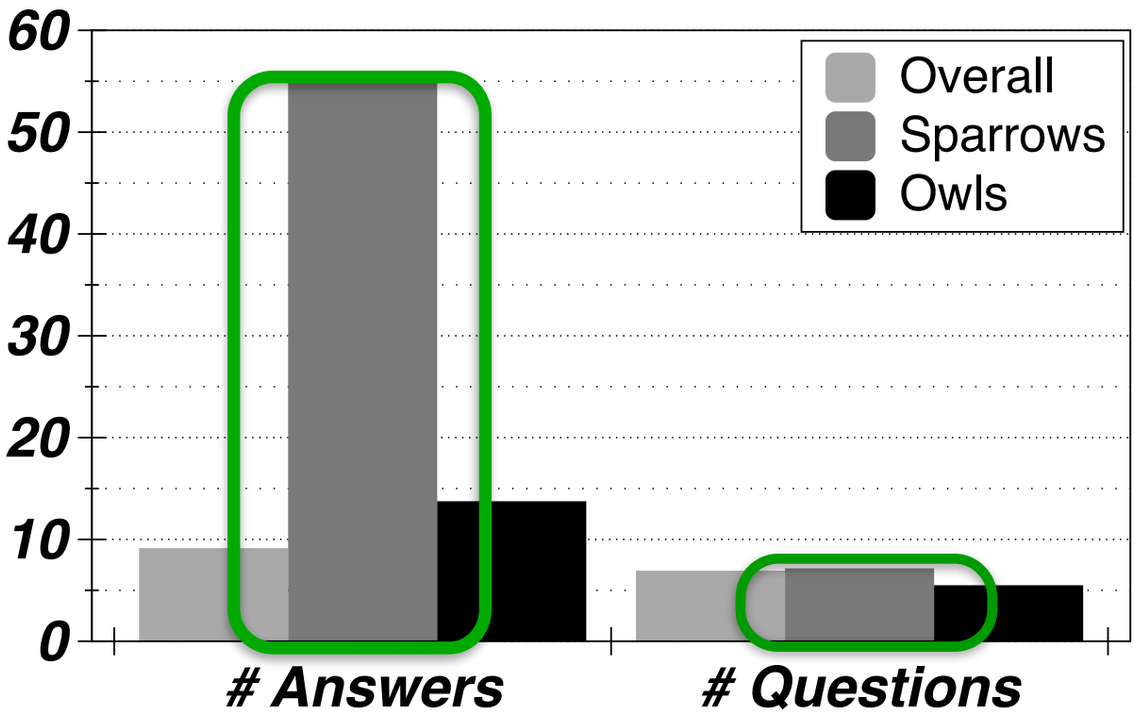
RQ2. Do Sparrows and Owls show different PREFERENCES in knowledge creation?

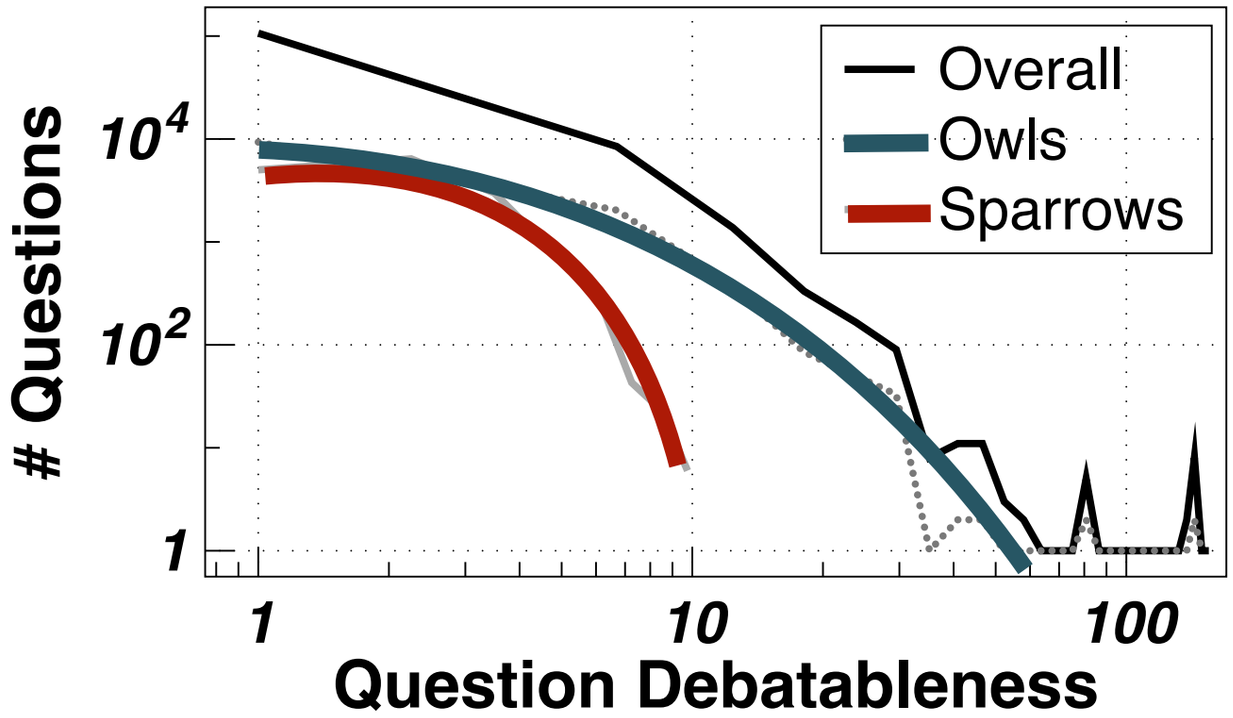
RQ3. Are INCENTIVISING mechanism equally effective on sparrows and owls?

RQ1. How do CONTRIBUTIONS from Sparrows and Owls differ?

PARTICIPATION ACTIVENESS

#question,answers, distribution of debatableness of the questions they answer to

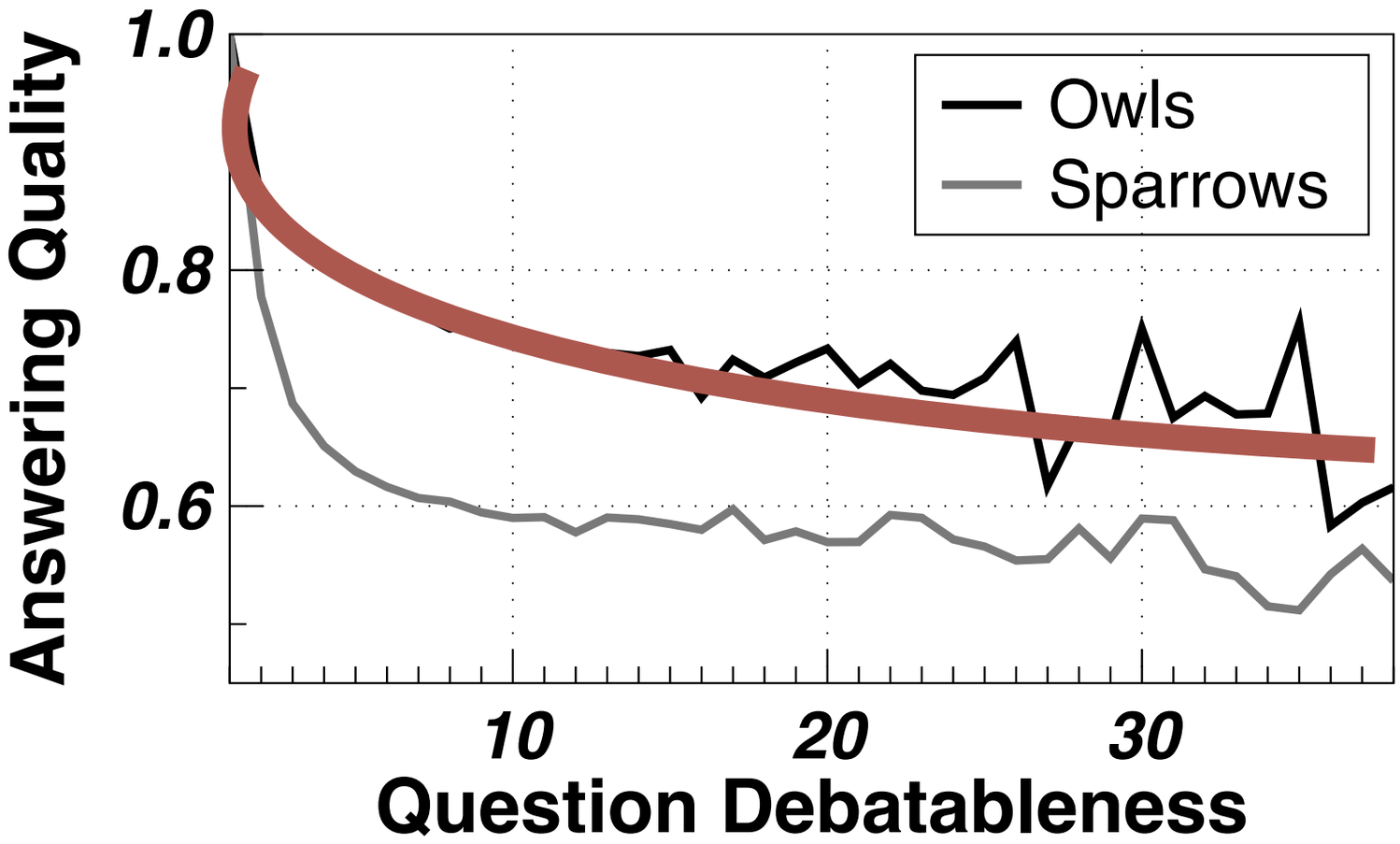




Sparrows answer much more, and more selective in answering less debatable questions.

ANSWERING QUALITY

Relative ranking of answers.

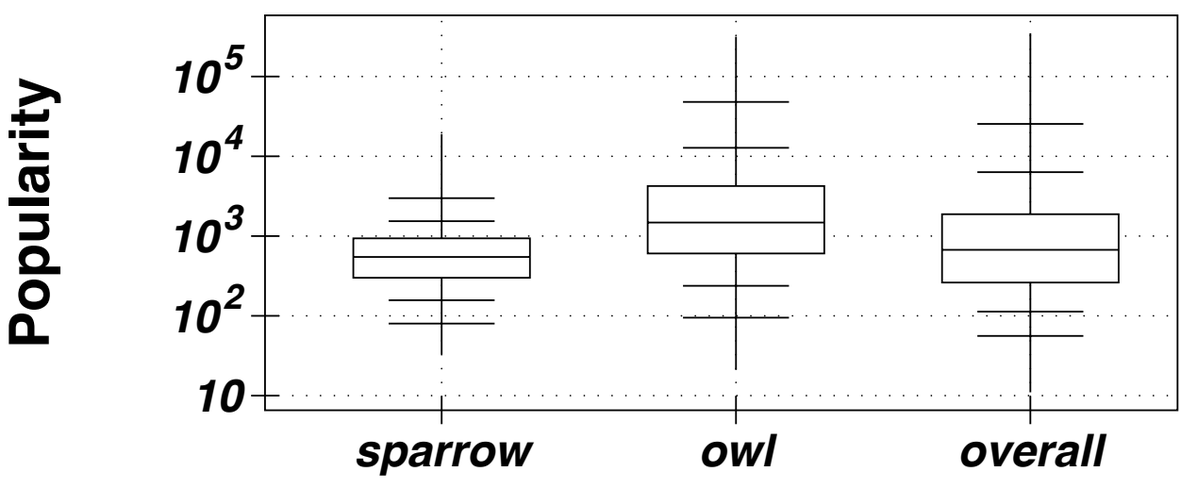


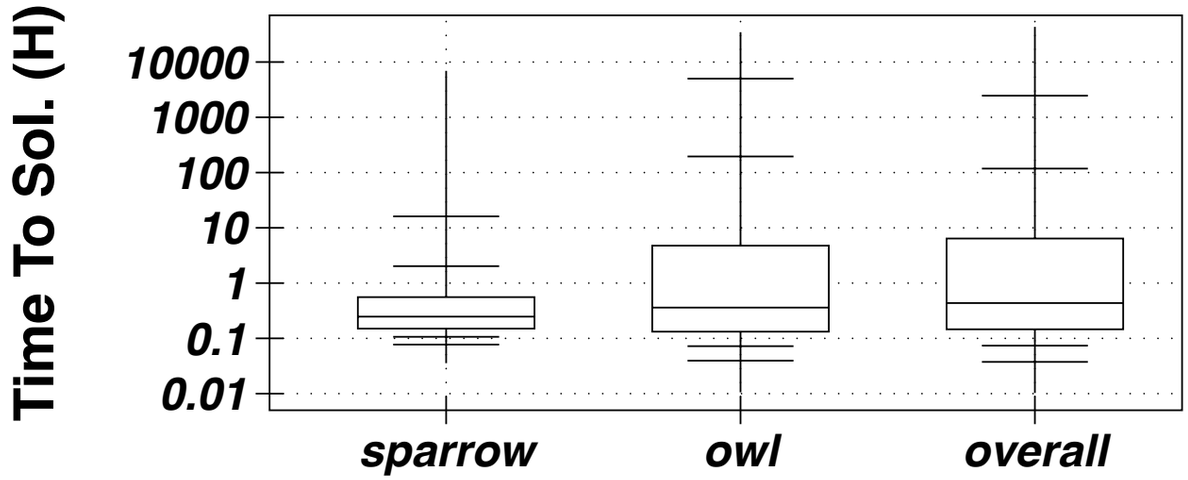
Owls give better answers than Sparrows for questions of all different debatableness.

RQ2. Do Sparrows and Owls show different PREFERENCES in knowledge creation?

QUESTIONS THEY ANSWER TO

Popularity = #views
Difficulty = Time to Solution = Taccept - Tpost





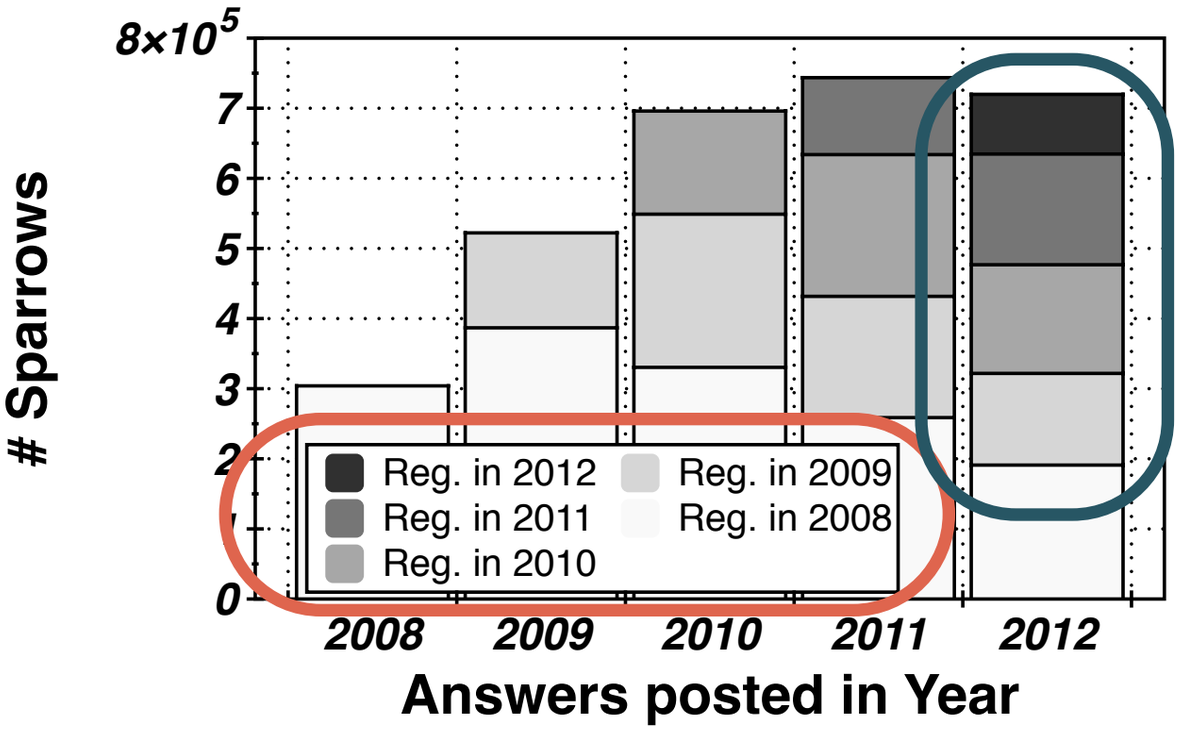
Owls ANSWER to questions that are more popular, and more difficult.

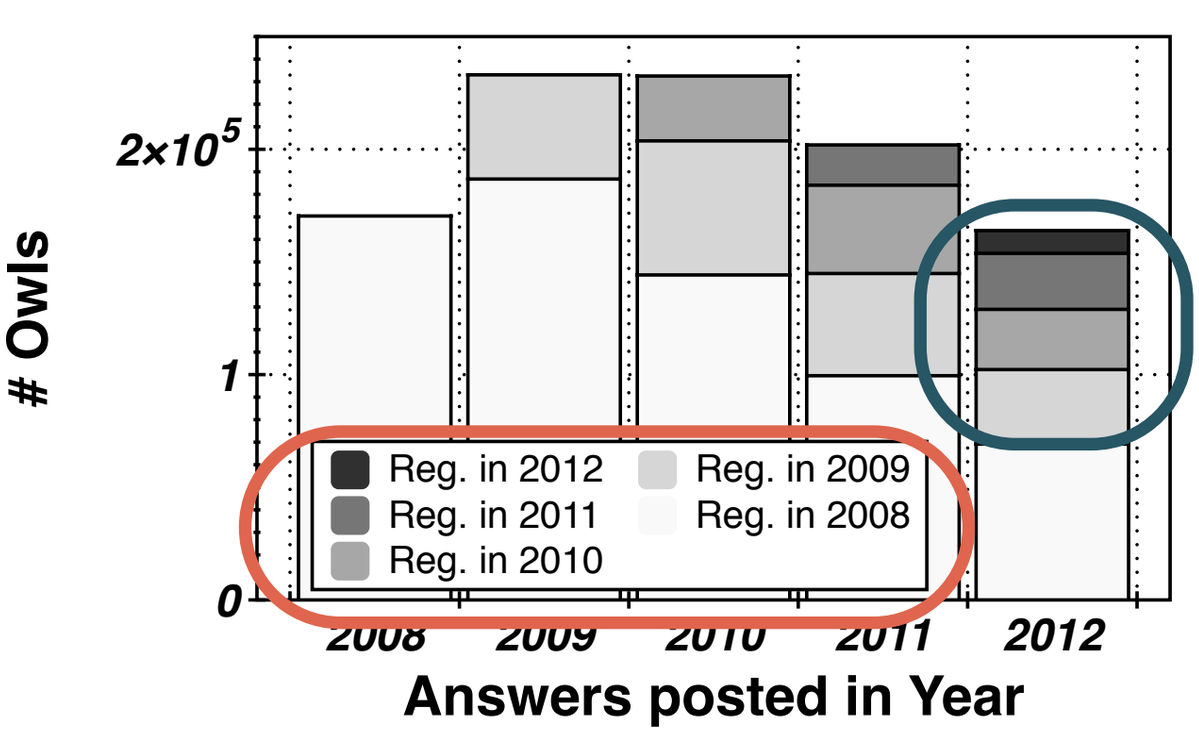
Similarly: Owls POST questions that are more popular, and more difficult.

RQ3. Are incentivising mechanisms equally effective on sparrows and owls?

ANSWERS POST BY EACH GROUP

NOTE: Comparable #registrations

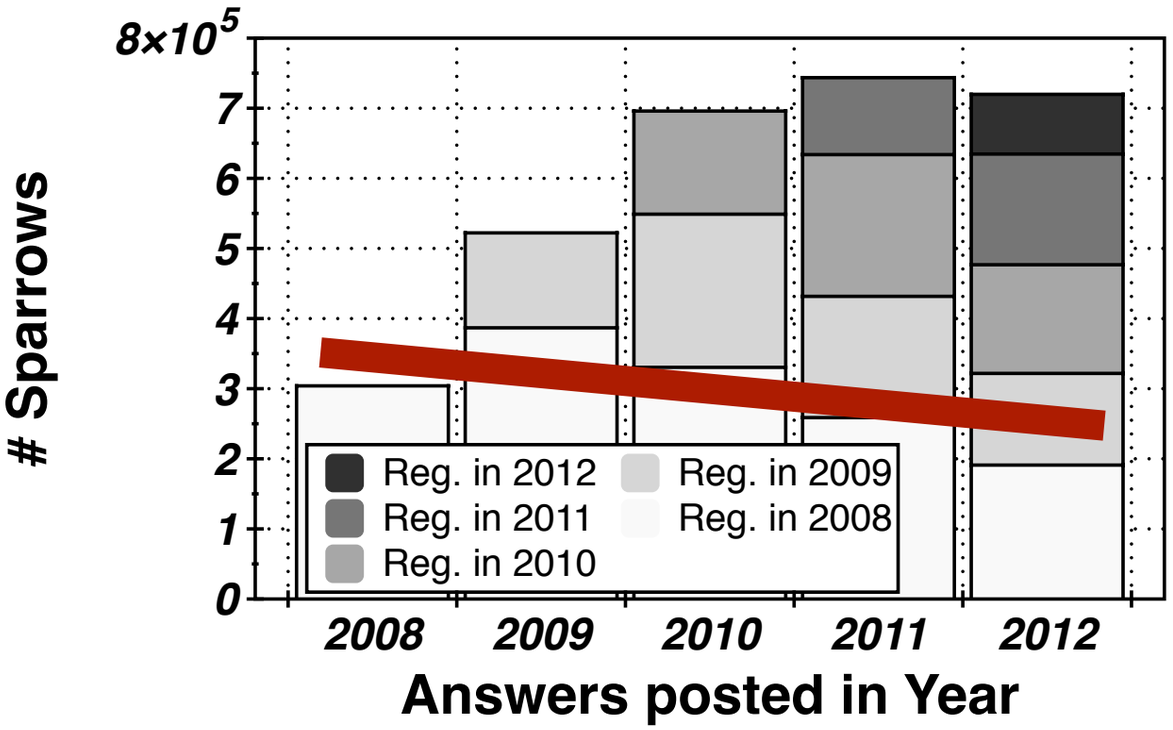


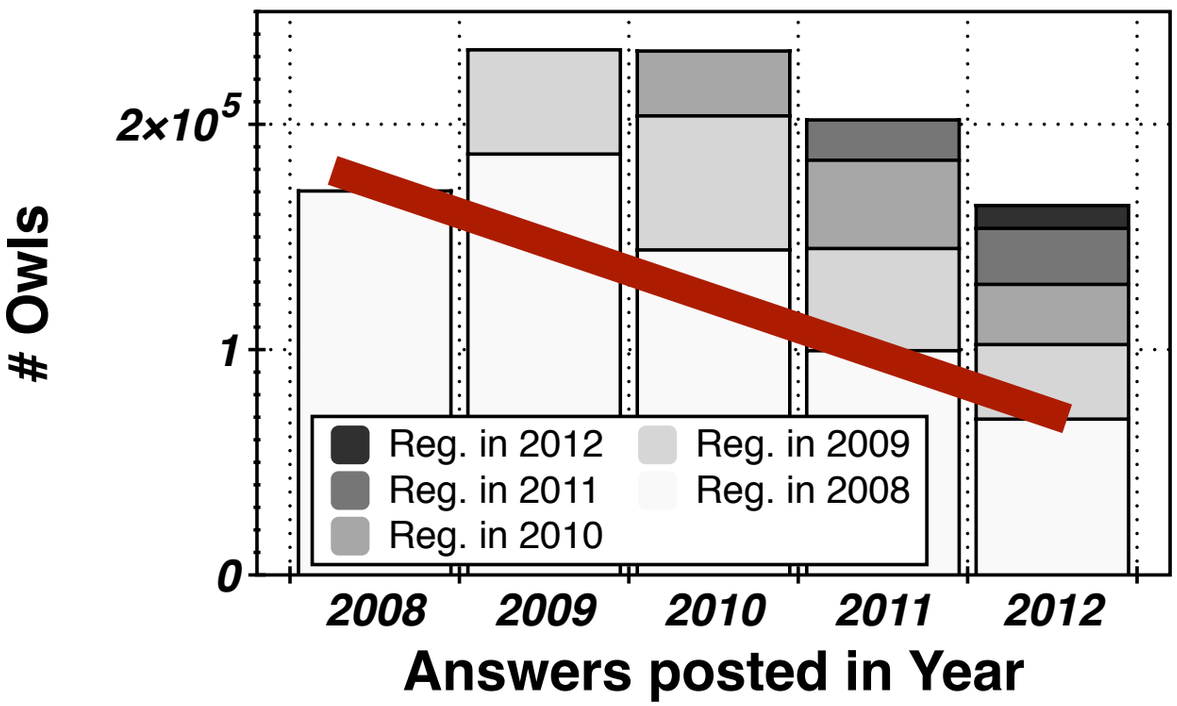


Newly registered sparrows contribute much more than newly registered owls

ANSWERS POST BY EACH GROUP

NOTE: Comparable #registrations





Gamification incentives can more effectively retain Sparrows than Owls

Insights

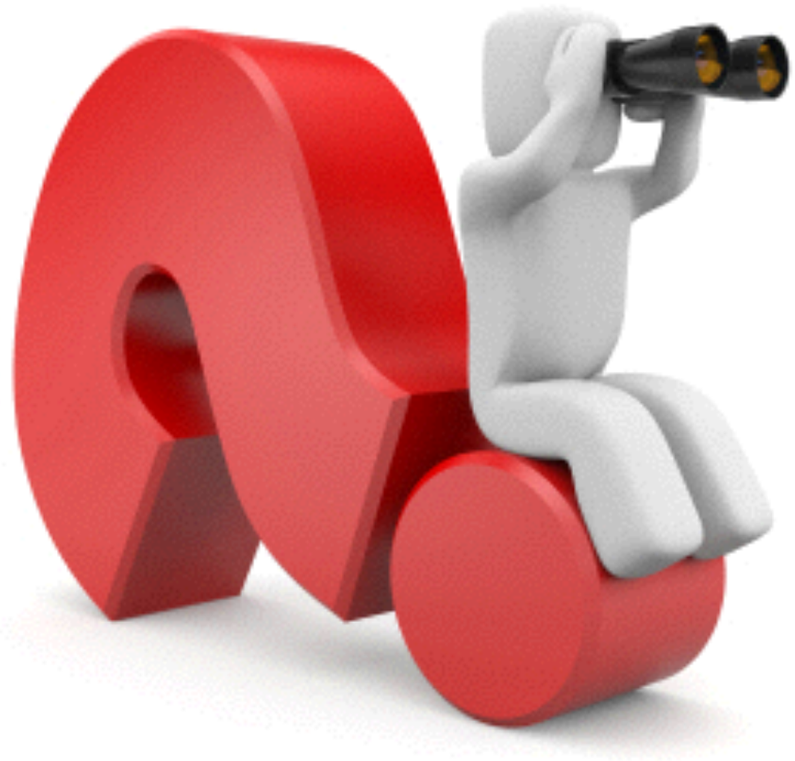
Q&A systems are important,

modelling their users can be useful.

Expertise might be there, but you need a right way to find it.

We provide an expertise metric, which can be a good start!

Thanks



Jie Yang
j.yang-3@tudeflt.nl