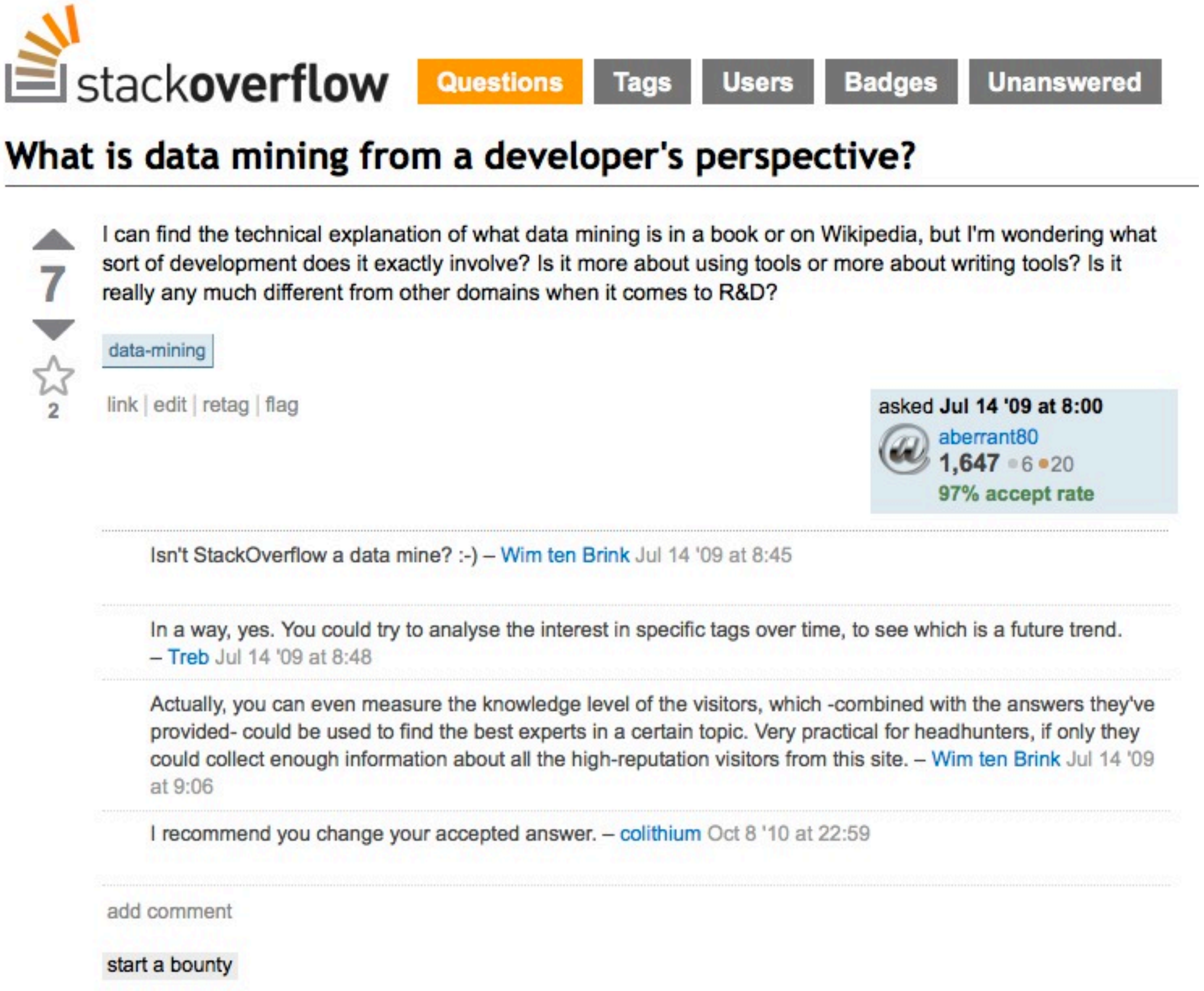
CSCI 568A

Discussion 01: Data Mining

Hi.

Yong Joseph Bakos
ybakos@mines.edu
http://mines.humanoriented.com/568

DATA
WHAT?



What is data mining from a developer's perspective?

I can find the technical explanation of what data mining is in a book or on Wikipedia, but I'm wondering what sort of development does it exactly involve? Is it more about using tools or more about writing tools? Is it really any much different from other domains when it comes to R&D?

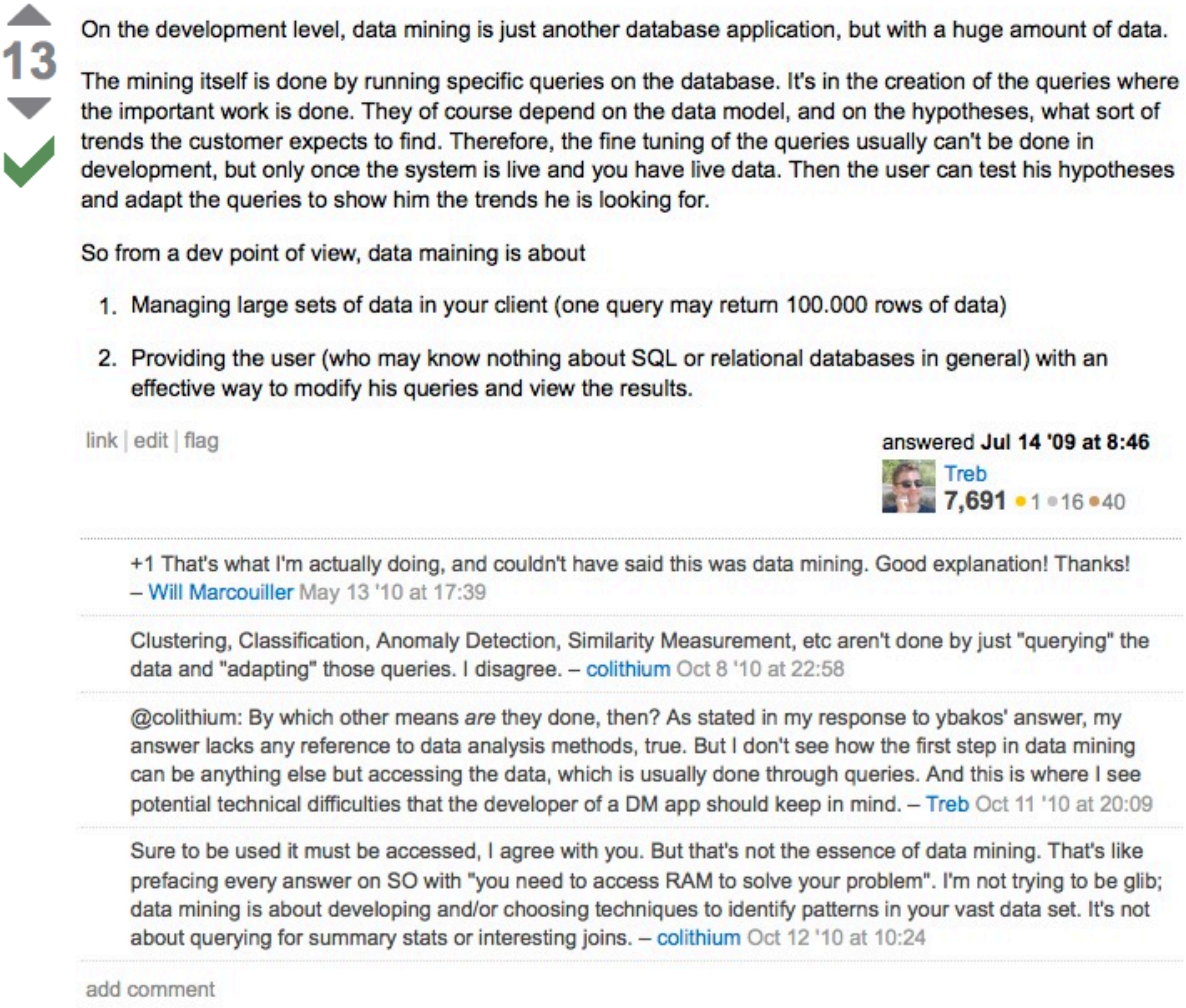
asked **Jul 14 '09 at 8:00**
aberrant80
1,647 ● 6 ● 20
97% accept rate

Isn't StackOverflow a data mine? :-) – Wim ten Brink Jul 14 '09 at 8:45

In a way, yes. You could try to analyse the interest in specific tags over time, to see which is a future trend. – Treb Jul 14 '09 at 8:48

Actually, you can even measure the knowledge level of the visitors, which -combined with the answers they've provided- could be used to find the best experts in a certain topic. Very practical for headhunters, if only they could collect enough information about all the high-reputation visitors from this site. – Wim ten Brink Jul 14 '09 at 9:06

I recommend you change your accepted answer. – colithium Oct 8 '10 at 22:59



13

✓

On the development level, data mining is just another database application, but with a huge amount of data.

The mining itself is done by running specific queries on the database. It's in the creation of the queries where the important work is done. They of course depend on the data model, and on the hypotheses, what sort of trends the customer expects to find. Therefore, the fine tuning of the queries usually can't be done in development, but only once the system is live and you have live data. Then the user can test his hypotheses and adapt the queries to show him the trends he is looking for.

So from a dev point of view, data maining is about

1. Managing large sets of data in your client (one query may return 100.000 rows of data)
2. Providing the user (who may know nothing about SQL or relational databases in general) with an effective way to modify his queries and view the results.

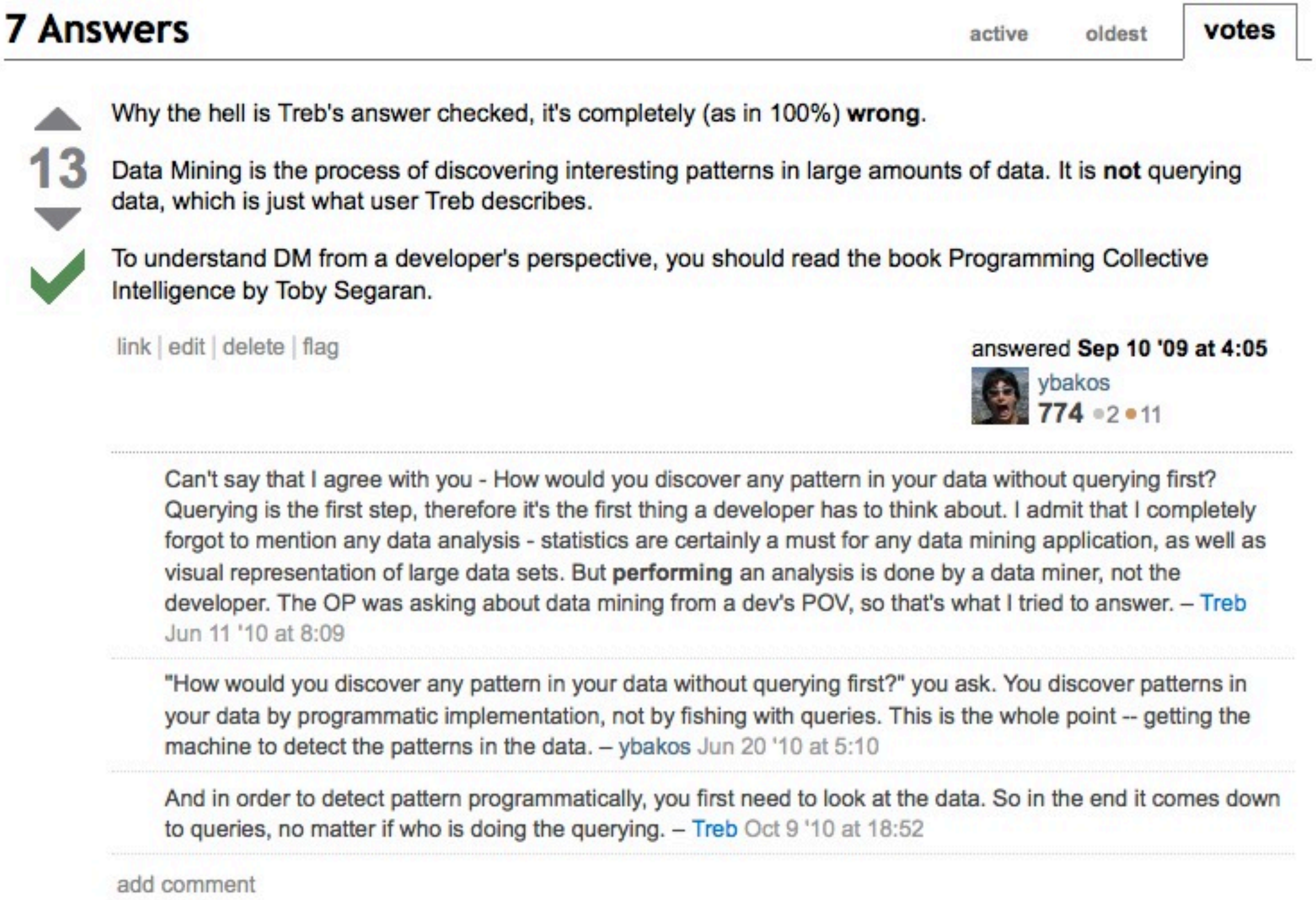
answered **Jul 14 '09 at 8:46**
Treb
7,691 ●1 ●16 ●40

+1 That's what I'm actually doing, and couldn't have said this was data mining. Good explanation! Thanks! – Will Marcouiller May 13 '10 at 17:39

Clustering, Classification, Anomaly Detection, Similarity Measurement, etc aren't done by just "querying" the data and "adapting" those queries. I disagree. – colithium Oct 8 '10 at 22:58

@colithium: By which other means *are* they done, then? As stated in my response to ybakos' answer, my answer lacks any reference to data analysis methods, true. But I don't see how the first step in data mining can be anything else but accessing the data, which is usually done through queries. And this is where I see potential technical difficulties that the developer of a DM app should keep in mind. – Treb Oct 11 '10 at 20:09

Sure to be used it must be accessed, I agree with you. But that's not the essence of data mining. That's like prefacing every answer on SO with "you need to access RAM to solve your problem". I'm not trying to be glib; data mining is about developing and/or choosing techniques to identify patterns in your vast data set. It's not about querying for summary stats or interesting joins. – colithium Oct 12 '10 at 10:24



7 Answers

Why the hell is Treb's answer checked, it's completely (as in 100%) **wrong**.

Data Mining is the process of discovering interesting patterns in large amounts of data. It is **not** querying data, which is just what user Treb describes.

To understand DM from a developer's perspective, you should read the book Programming Collective Intelligence by Toby Segaran.

answered **Sep 10 '09 at 4:05**
ybakos
774 ●2 ●11

Can't say that I agree with you - How would you discover any pattern in your data without querying first? Querying is the first step, therefore it's the first thing a developer has to think about. I admit that I completely forgot to mention any data analysis - statistics are certainly a must for any data mining application, as well as visual representation of large data sets. But **performing** an analysis is done by a data miner, not the developer. The OP was asking about data mining from a dev's POV, so that's what I tried to answer. – Treb Jun 11 '10 at 8:09

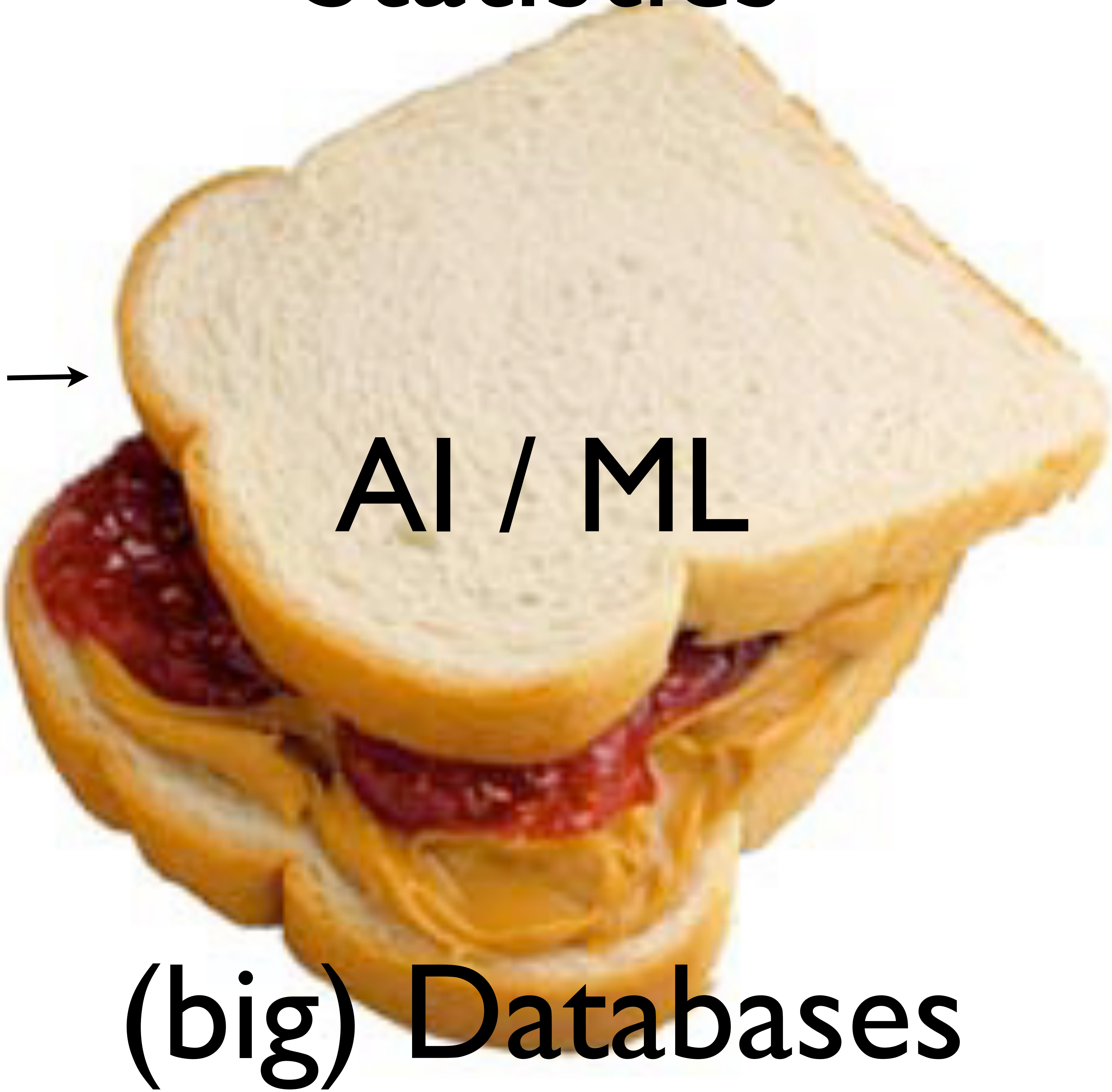
"How would you discover any pattern in your data without querying first?" you ask. You discover patterns in your data by programmatic implementation, not by fishing with queries. This is the whole point -- getting the machine to detect the patterns in the data. – ybakos Jun 20 '10 at 5:10

And in order to detect pattern programmatically, you first need to look at the data. So in the end it comes down to queries, no matter if who is doing the querying. – Treb Oct 9 '10 at 18:52

What Data Mining *Isn't*

- Crawling / harvesting / screen scraping
- Querying (fishing)
- Collecting
- Drinking

Statistics



data mining sandwich →

AI / ML

(big) Databases

"The process of discovering [useful] patterns in large amounts of data."

Fry, B. Visualizing Data. 2008.

What patterns do we see?

milk	cereal	diapers	beer
1	1	0	0
1	1	0	0
1	1	1	1
0	0	1	1

6 Core Topics

- Data & Big Data
- Classification
- Association Analysis
- Clustering
- Anomaly Detection
- Data Visualization & Interaction

Homework

- Project 1
- Reading 1