Big Data Tutorial Assignment 1 & 2

Marina Ernst
marinaernst@uni-koblenz.de

Institute for Web Science and Technologies
Universität Koblenz

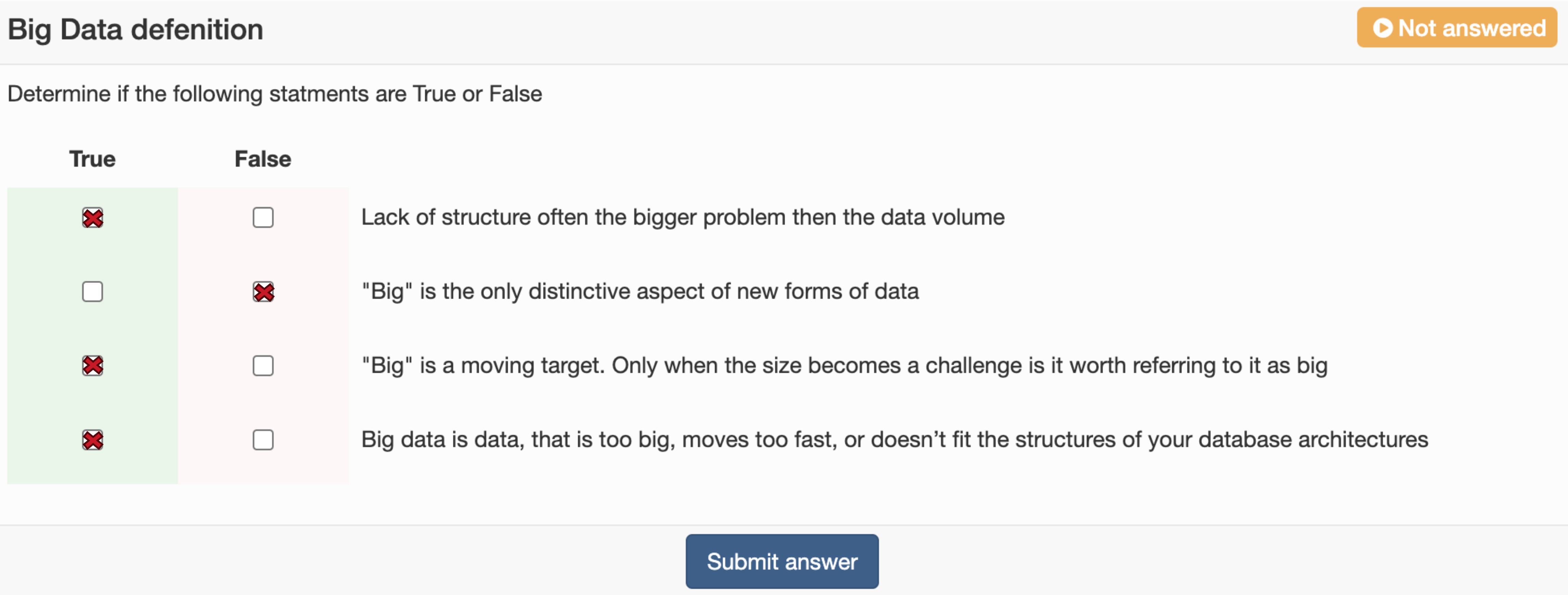
Outline

- Your feedback
- Assignment 1 discussion
- Assignment 2 discussion
- What's next?

Feedback

› Assignment 1

Recall

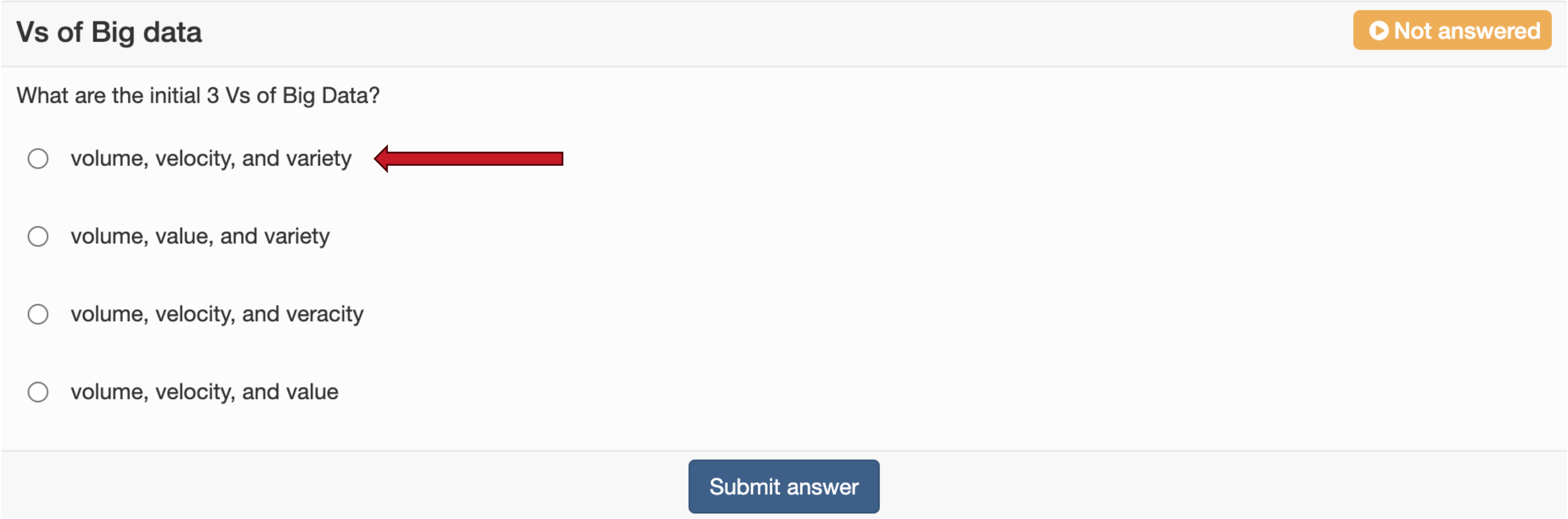


Big Data defenition

Determine if the following statments are True or False

True	False	
☒	□	Lack of structure often the bigger problem then the data volume
□	☒	"Big" is the only distinctive aspect of new forms of data
☒	□	"Big" is a moving target. Only when the size becomes a challenge is it worth referring to it as big
☒	□	Big data is data, that is too big, moves too fast, or doesn't fit the structures of your database architectures

Recall



Vs of Big data

What are the initial 3 Vs of Big Data?

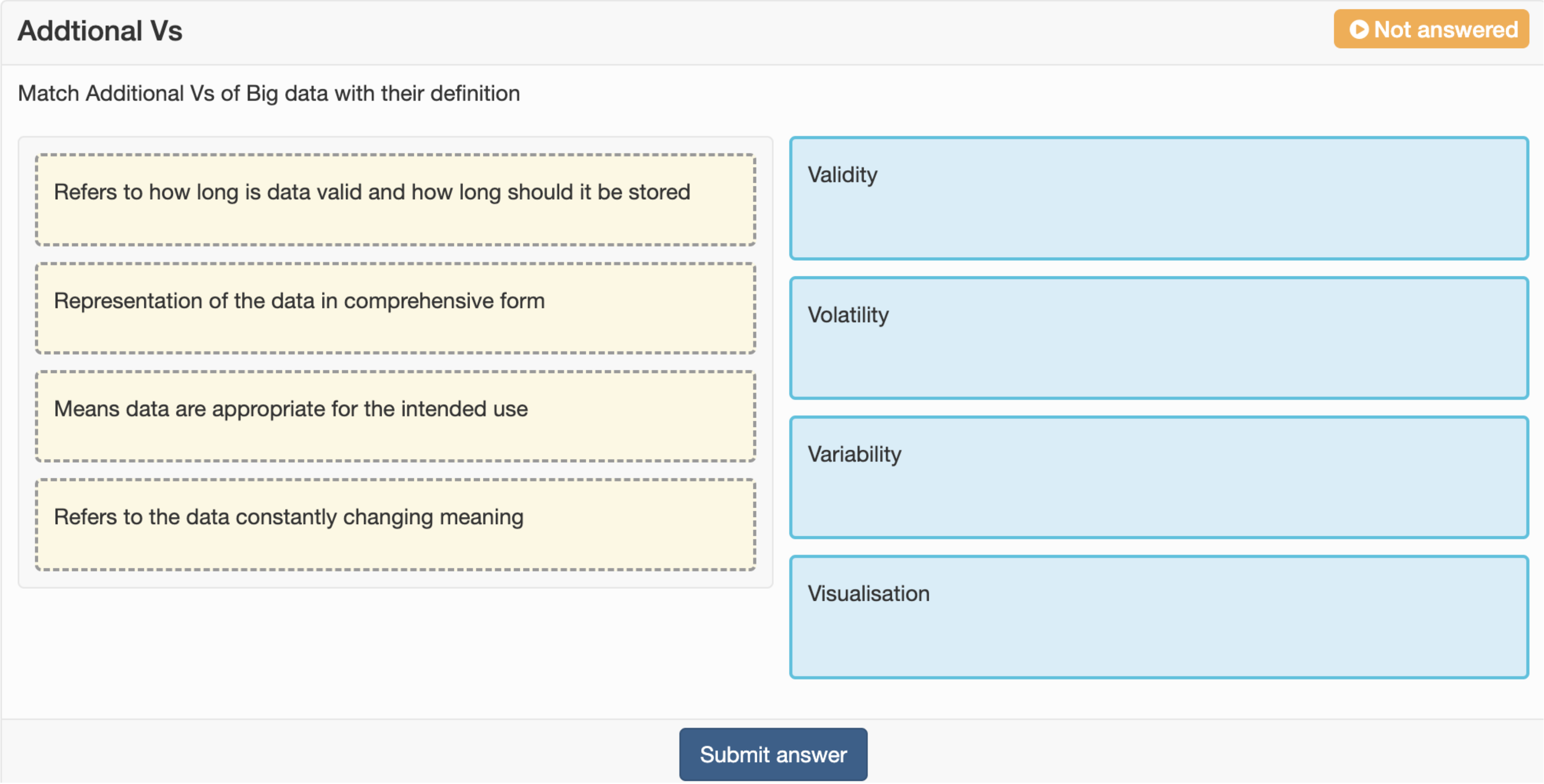
○ [✓ volume, velocity, and variety]

○ volume, value, and variety

○ volume, velocity, and veracity

○ volume, velocity, and value

Recall



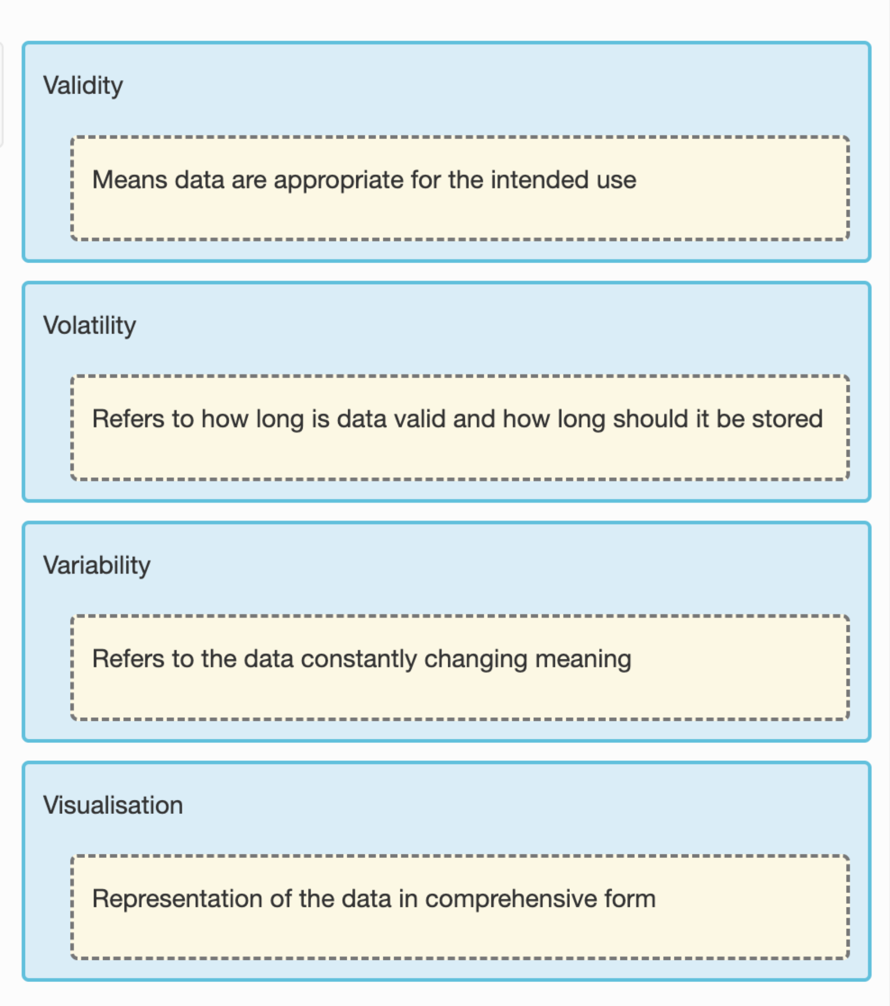
Addtional Vs

Match Additional Vs of Big data with their definition

- Refers to how long is data valid and how long should it be stored
- Representation of the data in comprehensive form
- Means data are appropriate for the intended use
- Refers to the data constantly changing meaning

- Validity
- Volatility
- Variability
- Visualisation

Recall



Validity

Means data are appropriate for the intended use

Volatility

Refers to how long is data valid and how long should it be stored

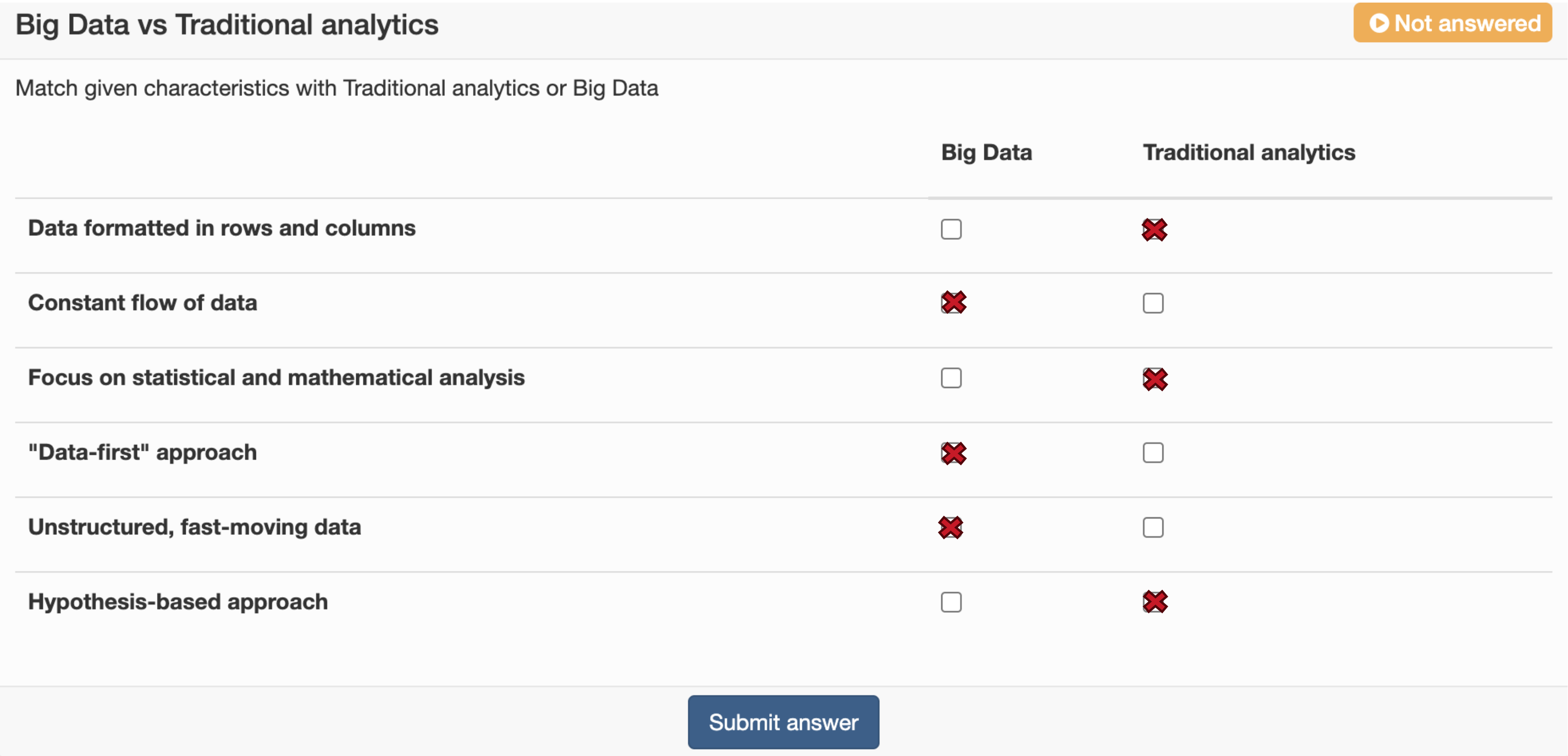
Variability

Refers to the data constantly changing meaning

Visualisation

Representation of the data in comprehensive form

Recall

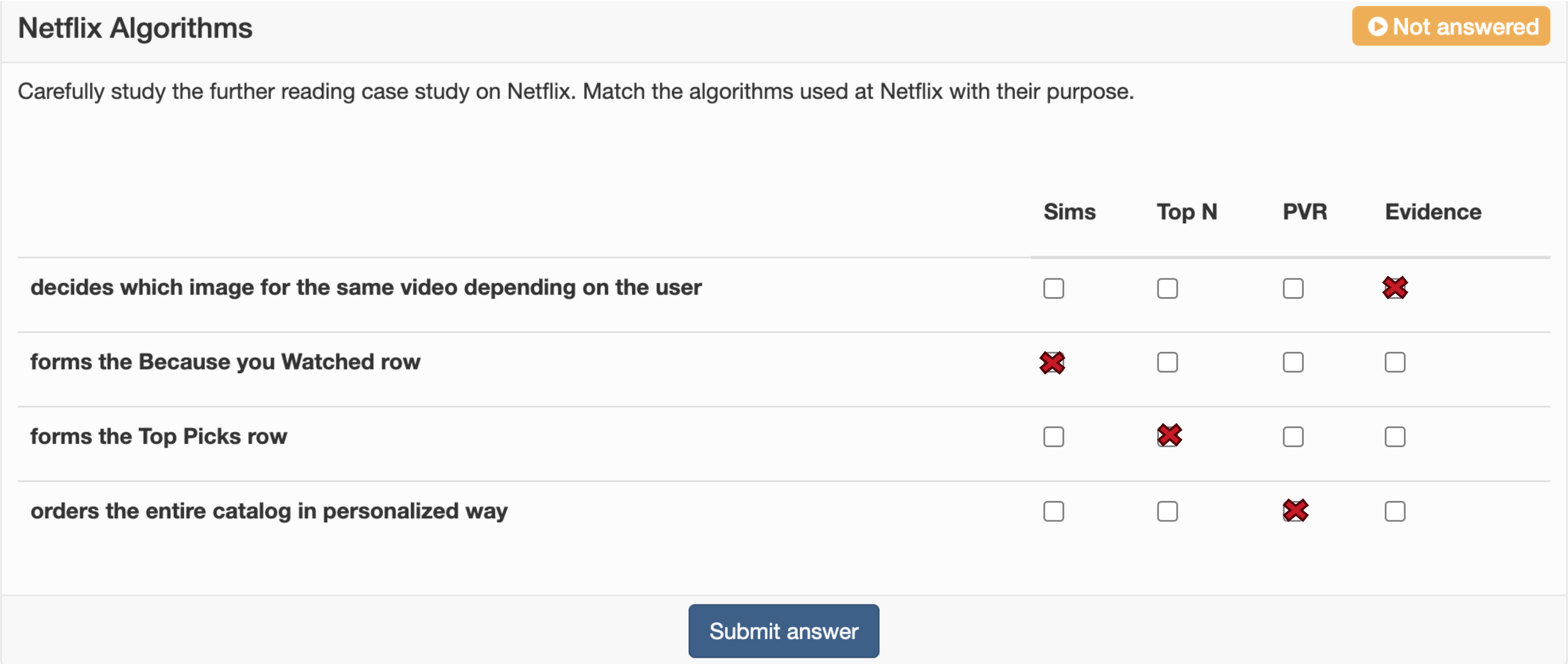


Big Data vs Traditional analytics

Match given characteristics with Traditional analytics or Big Data

	Big Data	Traditional analytics
Data formatted in rows and columns	□	☒
Constant flow of data	☒	□
Focus on statistical and mathematical analysis	□	☒
"Data-first" approach	☒	□
Unstructured, fast-moving data	☒	□
Hypothesis-based approach	□	☒

Knowledge Questions

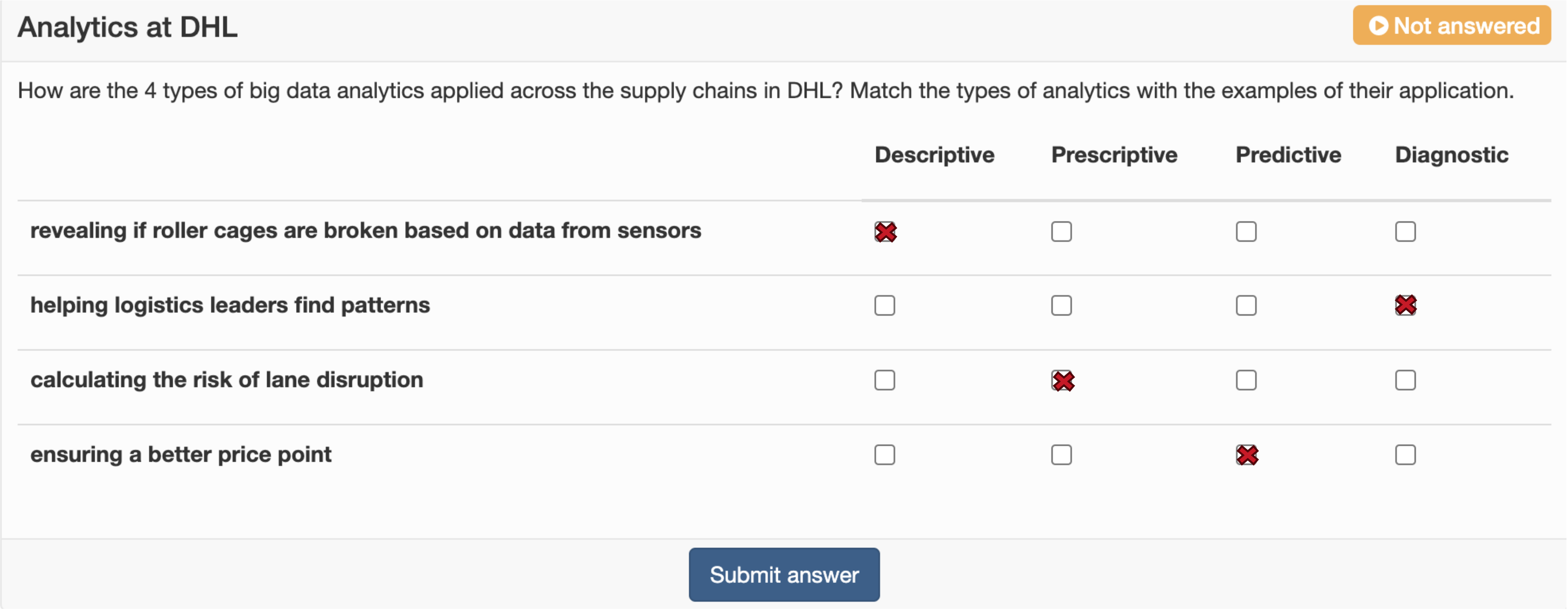


Netflix Algorithms

Carefully study the further reading case study on Netflix. Match the algorithms used at Netflix with their purpose.

	Sims	Top N	PVR	Evidence
decides which image for the same video depending on the user	□	□	□	☒
forms the Because you Watched row	☒	□	□	□
forms the Top Picks row	□	☒	□	□
orders the entire catalog in personalized way	□	□	☒	□

Knowledge Questions



Analytics at DHL

How are the 4 types of big data analytics applied across the supply chains in DHL? Match the types of analytics with the examples of their application.

	Descriptive	Prescriptive	Predictive	Diagnostic
revealing if roller cages are broken based on data from sensors	☒	□	□	□
helping logistics leaders find patterns	□	□	□	☒
calculating the risk of lane disruption	□	☒	□	□
ensuring a better price point	□	□	☒	□

Knowledge Questions

DeepQA

In your own words, explain the concept of DeepQA. To get an understanding of it, you need to use materials from further reading.

Solution:

DeepQA is deep natural language processing, which is sometimes called Deep Question-Answering. Unlike shallow NLP DeepQA focuses on the accuracy rather than precision. To achieve that much more context in incorporated into the process.

Knowledge Questions

A/B testing

Why it is important to work with new users on the platform when implementing the A/B test?

Solution:

Current members already experienced different versions of the application, so any change, even an improvement, might be rejected by them, only because they have already get used to certain flow.

Knowledge Questions

Big Data challenges:

Working with Big Data comes with numerous challenges. In this task, you have to write down **4** of those **challenges**, that you find the most significant. For each challenge, provide an explanation and an example.

You should use **not** only lecture and further reading material, but **external sources** as well.

Knowledge Questions

Big Data challenges:

- Unstructured data
- Storage capacity
- Lack of talent - not enough specialists in the field
- Security
- Privacy
- Legal issues
- Growing data
- Lack of understanding of how algorithms works
- ….

Knowledge Questions

Big Data example:

Provide an example of Big Data application in the industry. Explain how Big Data is used in that context. Do not repeat case studies from further reading.

Examples:

Personalization: Recommendation systems, targeting ads - Spotify, Amazon, Alibaba, etc.

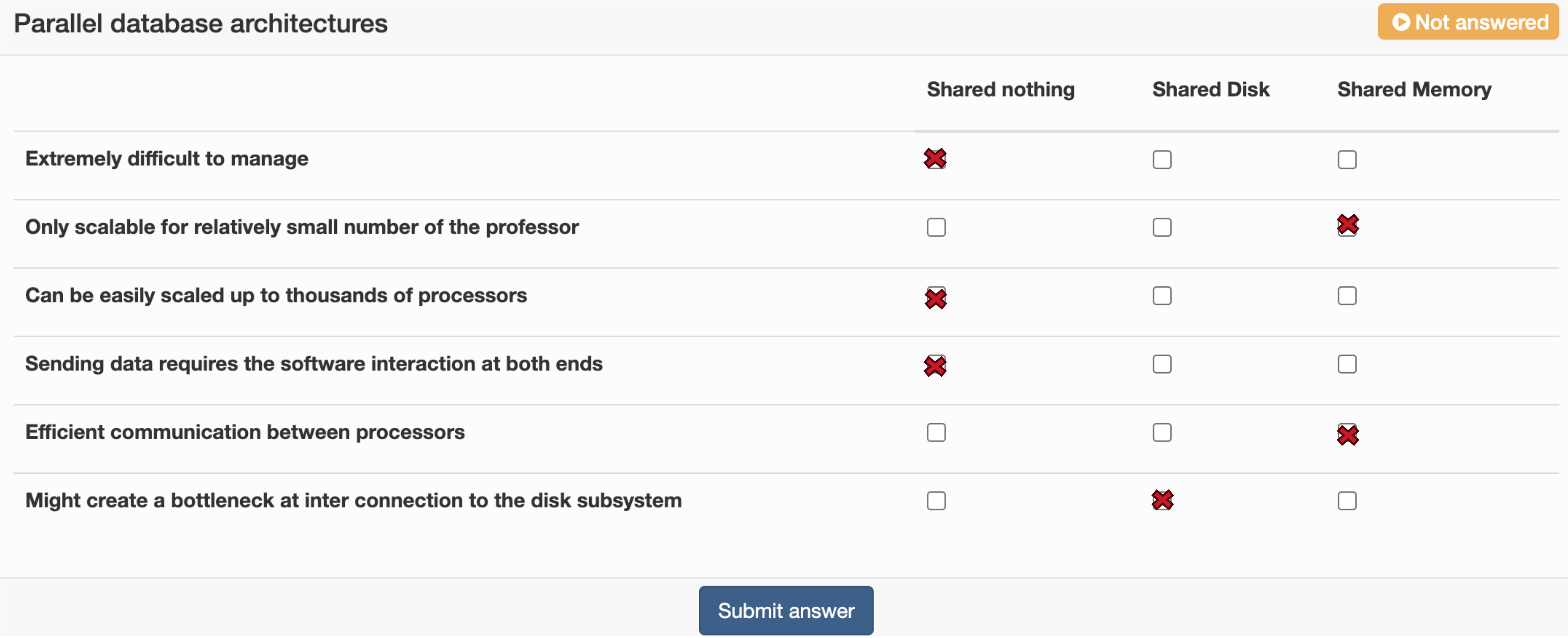
Health care: Electronic Health Records, Google Flu Trends (an example when it did not really work)

...

Q&A

› Assignment 2

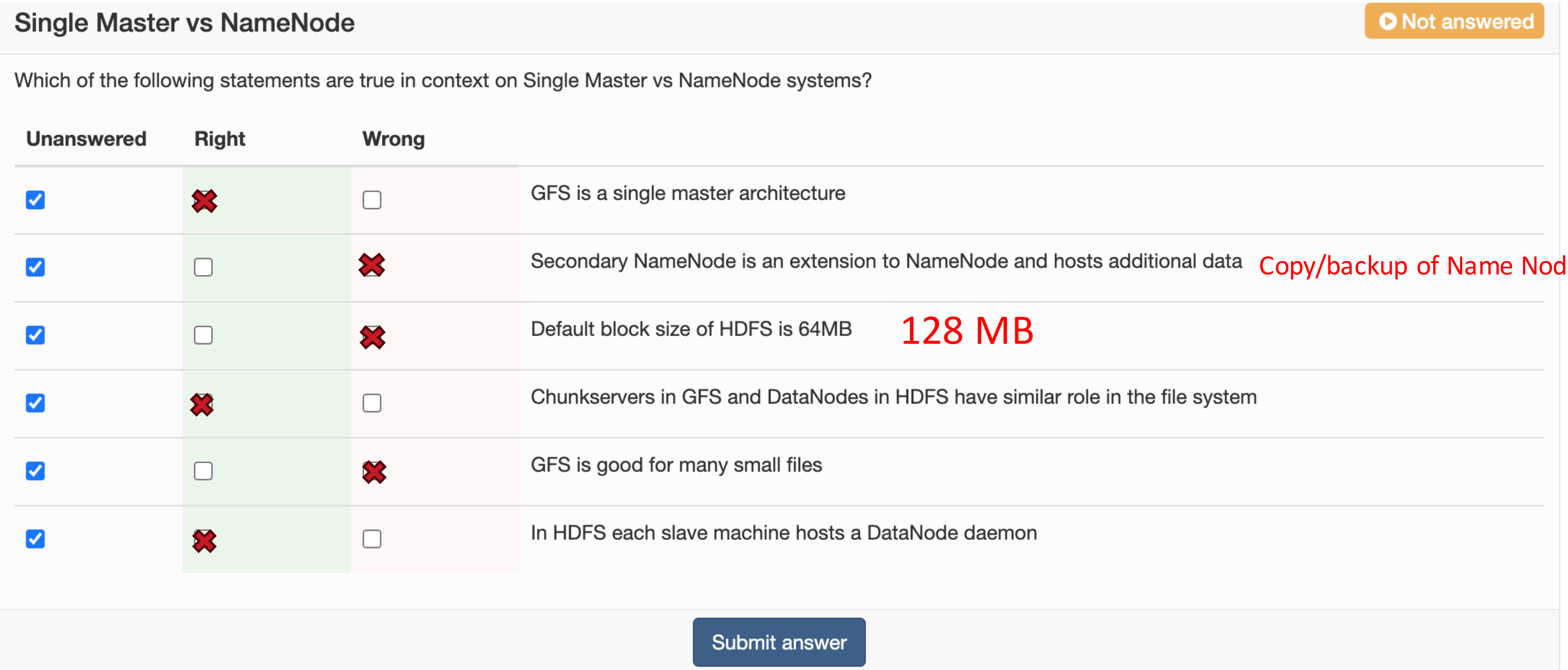
Recall



Parallel database architectures

	Shared nothing	Shared Disk	Shared Memory
Extremely difficult to manage	☒	□	□
Only scalable for relatively small number of the professor	□	□	☒
Can be easily scaled up to thousands of processors	☒	□	□
Sending data requires the software interaction at both ends	☒	□	□
Efficient communication between processors	□	□	☒
Might create a bottleneck at inter connection to the disk subsystem	□	☒	□

Recall

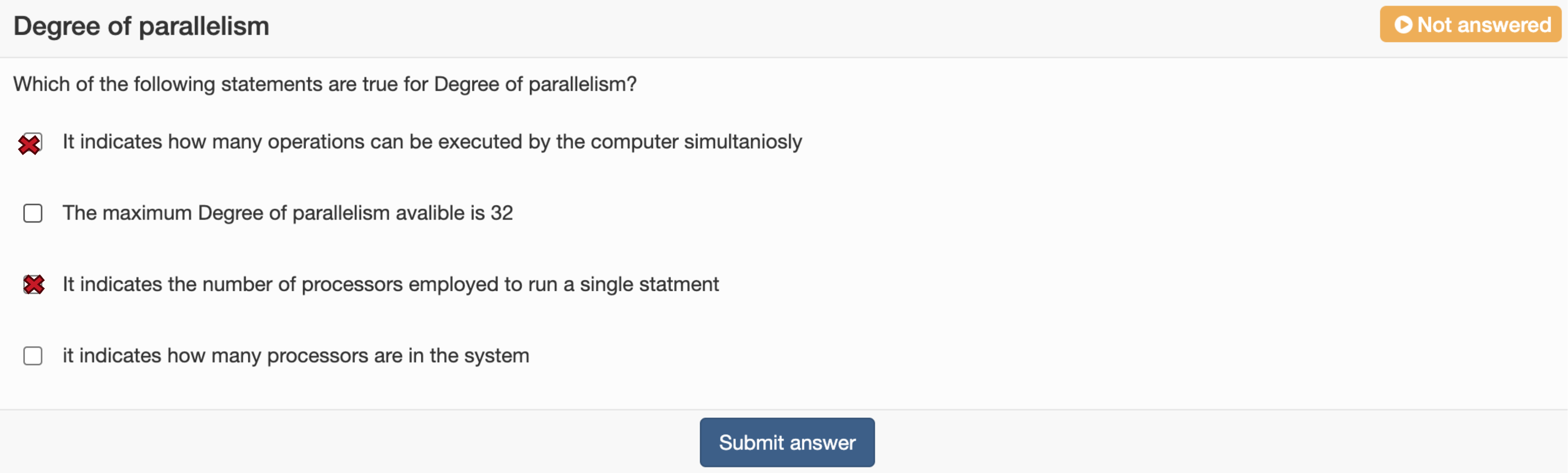


Single Master vs NameNode

Which of the following statements are true in context on Single Master vs NameNode systems?

Unanswered	Right	Wrong	
☑	✖	□	GFS is a single master architecture
☑	□	✖	Secondary NameNode is an extension to NameNode and hosts additional data Copy/backup of Name Node
☑	□	✖	Default block size of HDFS is 64MB 128 MB
☑	✖	□	Chunkservers in GFS and DataNodes in HDFS have similar role in the file system
☑	□	✖	GFS is good for many small files
☑	✖	□	In HDFS each slave machine hosts a DataNode daemon

Recall



Degree of parallelism

Which of the following statements are true for Degree of parallelism?

☒ It indicates how many operations can be executed by the computer simultaniosly

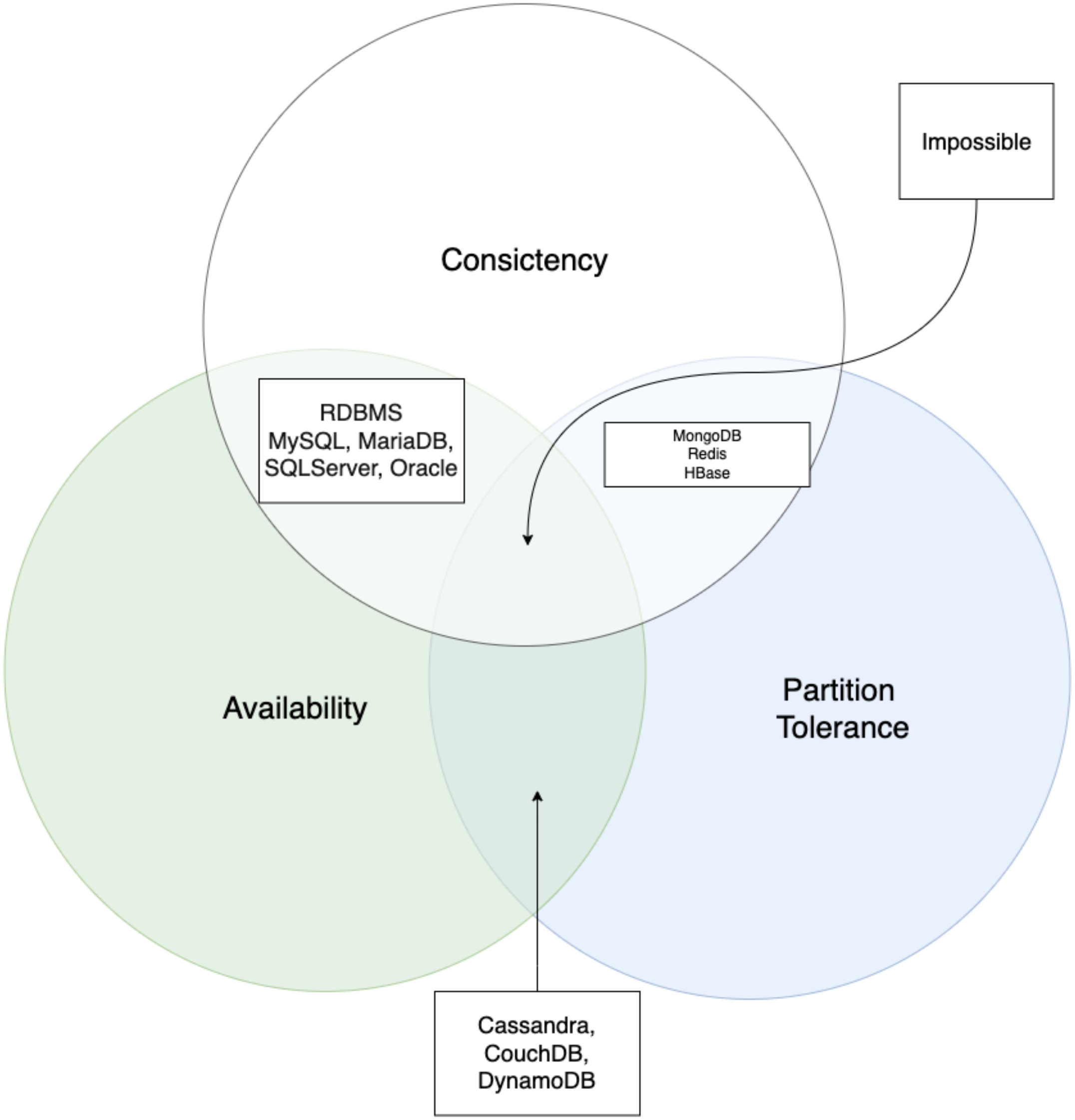
□ The maximum Degree of parallelism avalible is 32

☒ It indicates the number of processors employed to run a single statment

□ it indicates how many processors are in the system

Recall

CAP theorem



Knowledge Questions

Atomic Consistency

Carefully read the further material on CAP. Based on that knowledge, explain what is an Atomic Consistency.

Solution:

Atomic Consistency refers to a property of single request/response operation sequence.

Or

Atomic Consistency mean each operation looks as if it was compiled at a single instance

Knowledge Questions

Single Master

In your own words, explain why GFS uses Single Master and why it is not becoming a bottleneck. Further reading on GFS will help with this task.

Solution:

The master now has global knowledge of the whole system, which drastically simplifies the design.

But the master is not the bottleneck because: Clients never read and write file data through the master; client only requests from master which chunkservers to talk to , Master can also provide additional information about subsequent chunks to further reduce latency, Further reads of the same chunk don’t involve the master.

Q&A

Outline

- 19.05 - Assignment 3 Discussion
- 26.05 - Tutorial from Databricks
- 09.06 - Assignment 4 & 5 discussion
- 16.06 - Tutorial from Databricks
- TBA

Knowledge Questions

Map Reduce and Spark RRD

Explain the difference between Map Reduce and Spark RRD. You may use further reading articles and additional information sources to derive your answer.

Solution:

Spark can do in-memory processing, while Hadoop MapReduce has to read from and write to a disk.

Spark is faster, utilizes RAM not tied to Hadoop's two-stage paradigm, and works well for small data sets that fit into a server's RAM. MapReduce, on the other hand, is more cost-effective for processing large data sets and has more security features and projects.

> **That's all, folks! Happy coding!**