Real Time Data Analysis

Subject main topics

- Basics of (py)Spark
- Data Analysis with pyspark
- ML wtih pyspark
- Streaming data & APIs
- (Web Scraping)

What is the key element to distinguish "big data" from other data fields?

Mentimeter

the processing methodologies used

very large information

Big data starts when the data no longer fits in memory and needs to be computed in a distributed manner.

Handling very large amounts of Data

High Velocity, Volume, Variety

requires massive computation power and needs to be processed in real time ...also its big

Big data is structured\unstructured data on large scale that requires additional effort in storage, processing and retrieval

The 3vs: Volume, Velocity and Variety

output, and huge info

12

What is the key element to distinguish "big data" from other data fields?

Mentimeter

Amount of data enough big to be difficult for humans to process.

huge Data

data visualitzation

12

Which are the subjects you liked the most so far?

Mentimeter

12

Which is the most important feature of the cloud platforms (AWS, Azure...)?

Mentimeter

scalability

scalability

Cloud security ?

consistency

Extendibility, Availability, Robustness

Scalability, Focus on development, instead of infrastructure, take advantage of different levels of abstractions.

cost efficient.

Horizontally distribute computing work

incremental iterative development

11

Which is the most important feature of the cloud platforms (AWS, Azure...)?

Mentimeter

able to access everywhere

- Provide a single and secure access point to data, in theory single source of truth (SSOT).- Data governance- Horizontal scaling

11

How would you define Spark?

Mentimeter

Cluster programming interface

It is an Apache "open source" product used for processing Big Data

framework to process large amounts of data quickly and over different devices.

A java framework based on Hadoop for working with distributed data.

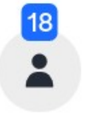
Open source engine that performs processing tasks on very large data sets

Fast engine for large-scale data processing. The secret for being fast is that it runs on RAM.

A data streaming platform that allows you to easily distribute work amongst clusters of machines in a scalable manner

Spark is a general-purpose distributed data processing engine that is suitable for use in a wide range of circumstances. on top of spark there is many programming languages for different needs

streaming, querying and analytical programming language using distributed processing in memory



How would you define Spark?

Mentimeter

useful to do batch and streaming data

scala offers better performance and it more stable and less prone to produce bugs if you change the code

Scala is less difficult to learn than Python and it is difficult to write code in Scala

Is an analytics engine for large-scale data (SQL, batch processing, stream processing, and machine learning)It's Fast, flexible, and developer-friendly

Scala is a static-typed language, and Python is a dynamically typed language. Type-safety makes Scala a better choice for high-volume projects because its static nature lends itself to faster bug and compile-time error detection.

- Support concurrency or multithreading- Use JVM (statically type) - Type safe- Better for testing

Enables concurrency

Scala is a statically typed language

for data analyzing, Scala is faster than python. Also, Scala is easy to use

18

What makes Scala a good language compared to Python?

Mentimeter

Scala is 10 times faster than python for data analysis and processing

Scala is a static-typed language, and Python is a dynamically typed language. Type-safety makes Scala a better choice for high-volume projects because its static nature lends itself to faster bug and compile-time error detection.

Scala allows writing of code with multiple concurrency primitives whereas Python doesn't support concurrency or multithreading.

Spark is written in Scala

Scala would be more beneficial in order to utilize the full potential of Spark

Scala is less difficult to learn than Python and it is difficult to write code in Scala

parallel processing, fast

scala is less prone to produce bugs .. scala is faster

- Scala allows devs to make good use of standard JVM features and Java libraries (scala source code -> Bytecode -> JVM - > Target Platform).

What is a RDD?

Mentimeter

Resilient Distributed Dataset

It is a data structure in Spark. It represents an immutable, partitioned collection of elements that can be operated on in parallel

it is an immutable distributed collection of objects

At a high level, every Spark application consists of a driver program that runs the user's main function and executes various parallel operations on a cluster. The main abstraction Spark provides is a resilient distributed dataset (RDD).

a dataset distributed across clusters without a single point of failure

Fundamental spark data structure. It is an immutably distributed collection of elements from a dataset.

fault-tolerant collection of elements that can be operated on in parallel

Resilient Distributed Dataset (RDD), a programming abstraction that represents an immutable collection of objects that can be split across a computing cluster (At the heart of Apache Spark).

Immutable distributed collection of elements of your data partitioned across nodes in your cluster.

What is a RDD?

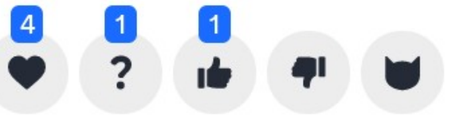
Mentimeter

Resillient Distributed Dataset-- API for spark

10

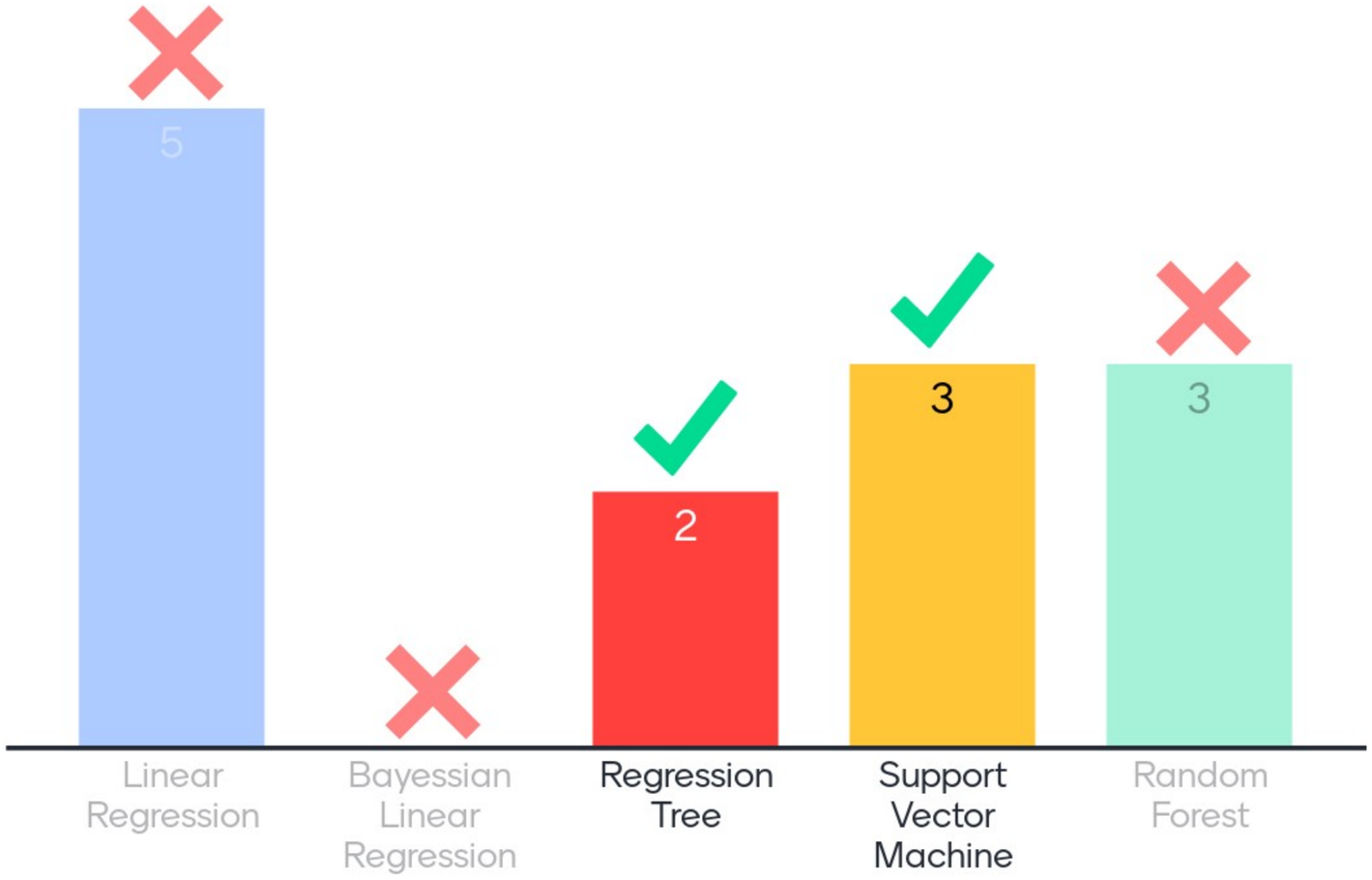
First Day Contest

Let's have some fun!



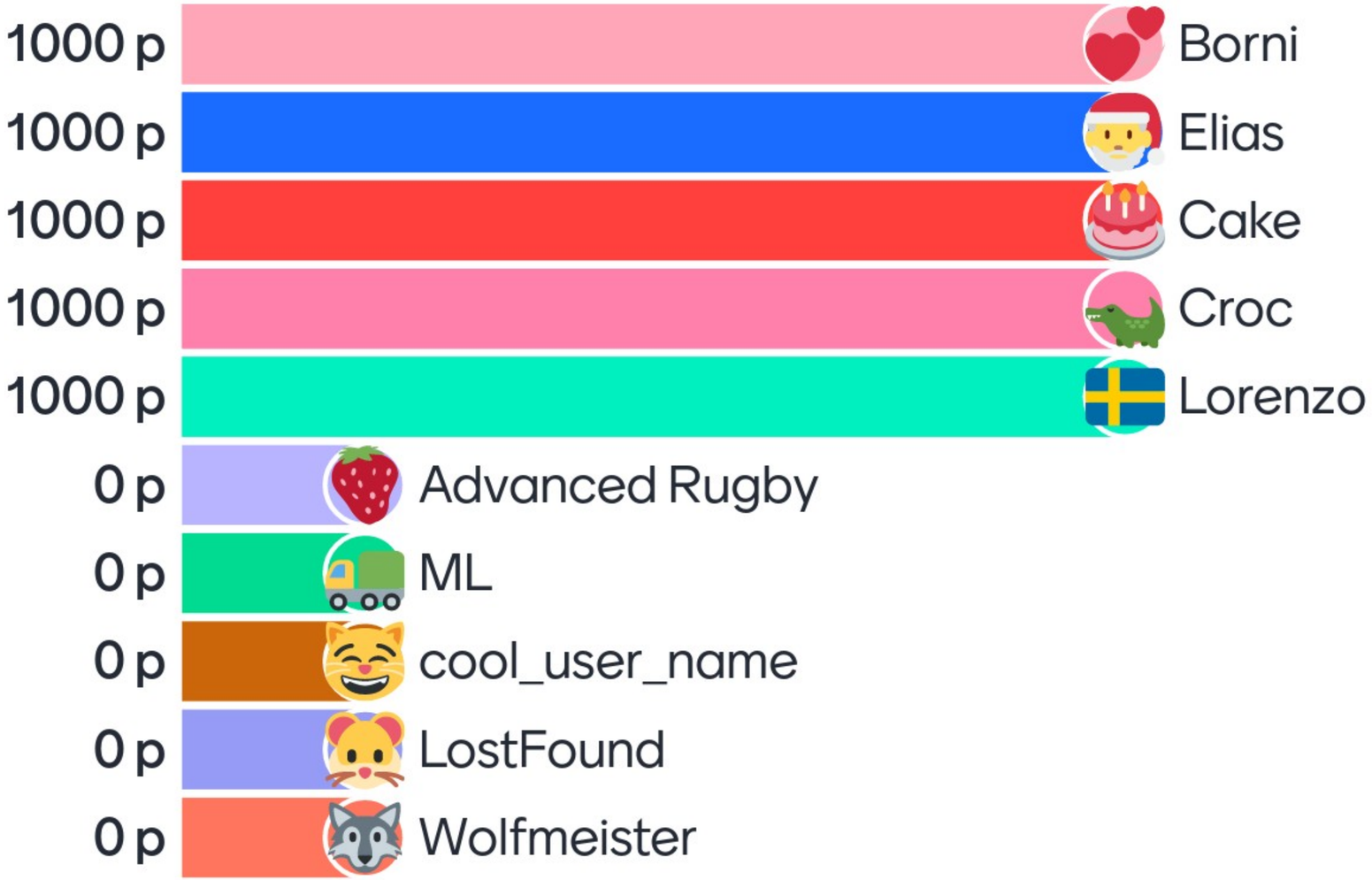
Which of these algorithms is more likely to overfit the train data?

Mentimeter



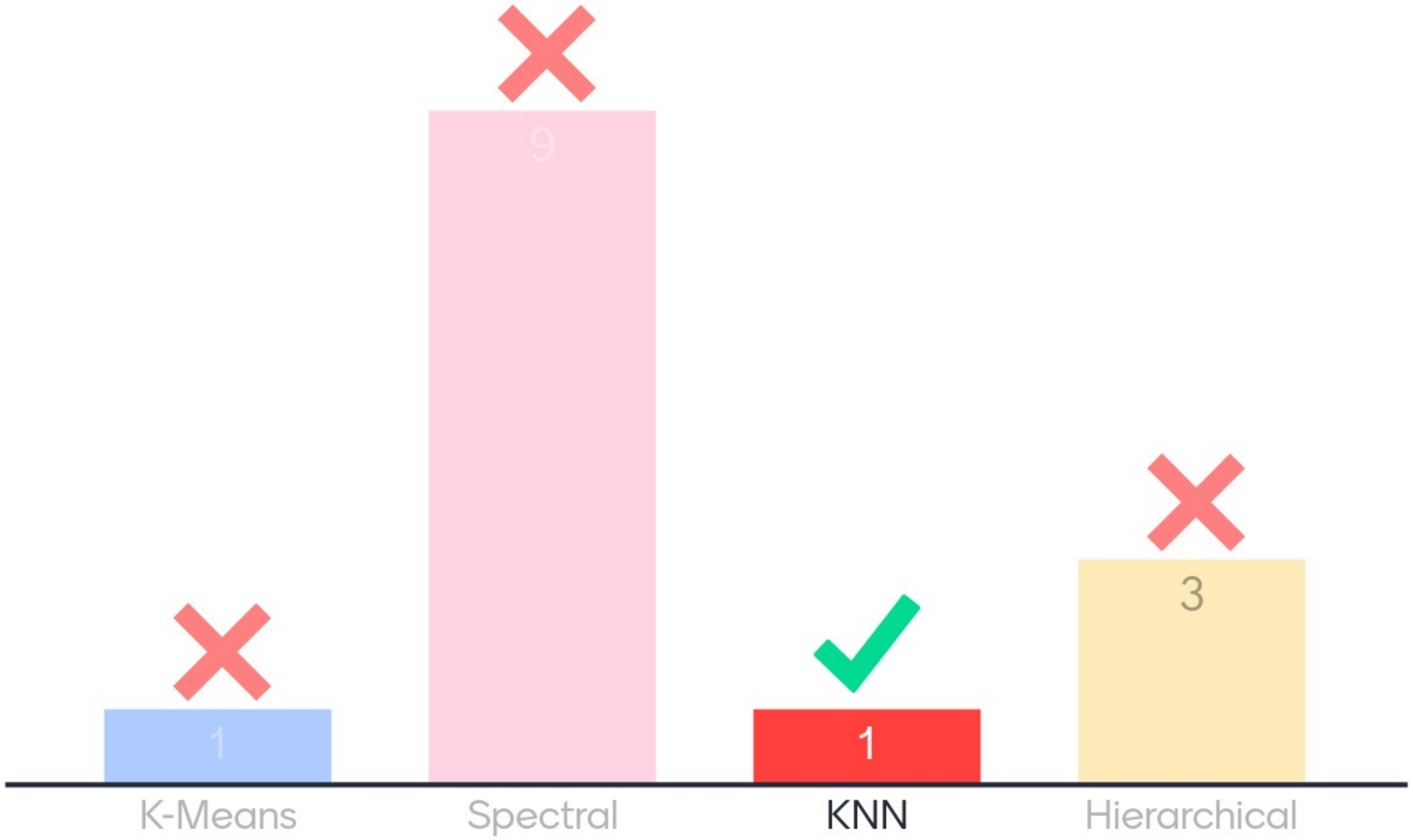
13

Leaderboard



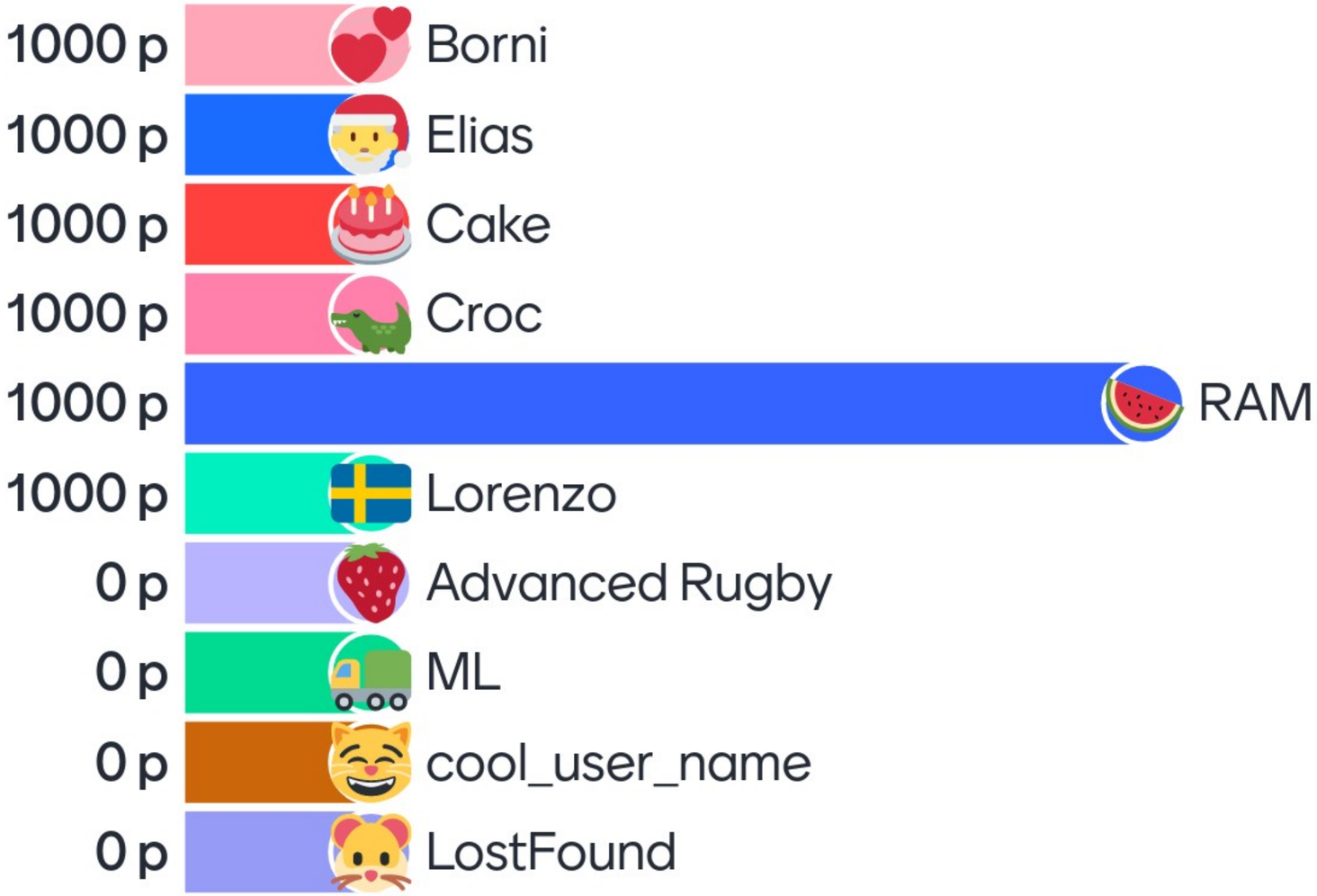
Which of these IS NOT a clustering algorithm?

Mentimeter



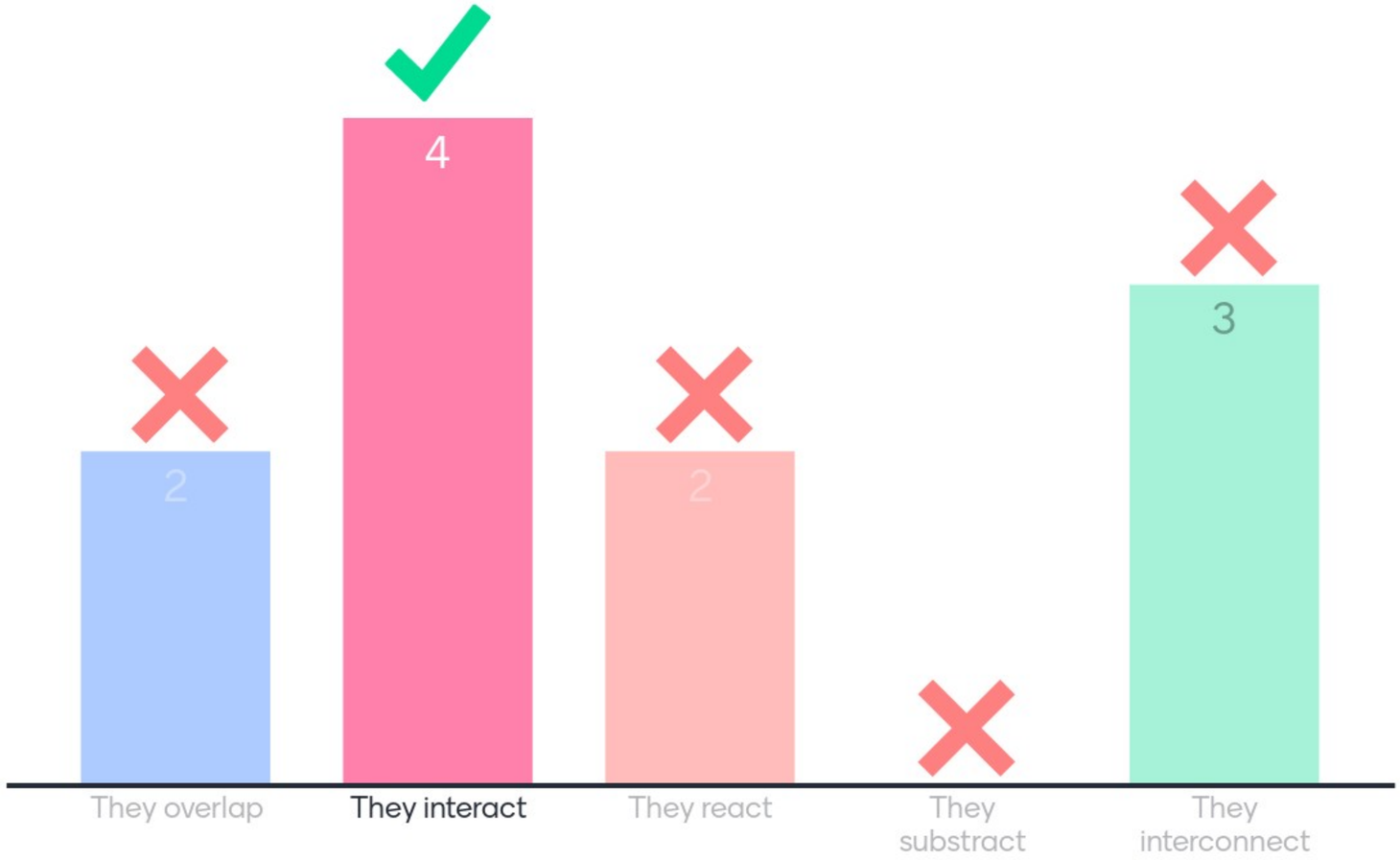
14

Leaderboard

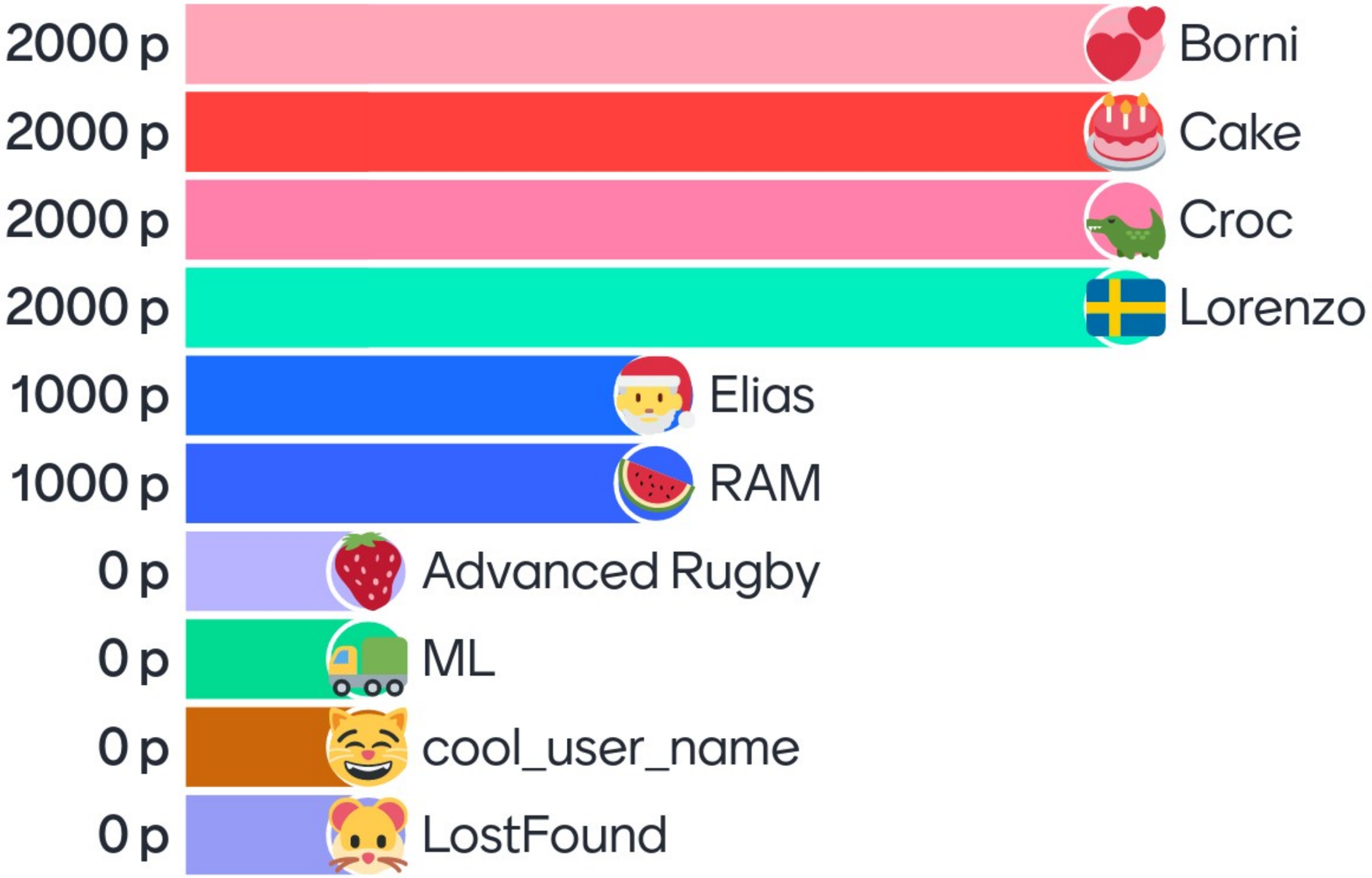


If the effect of two variables combined is more than the isolated effects of them separately, we say...

Mentimeter

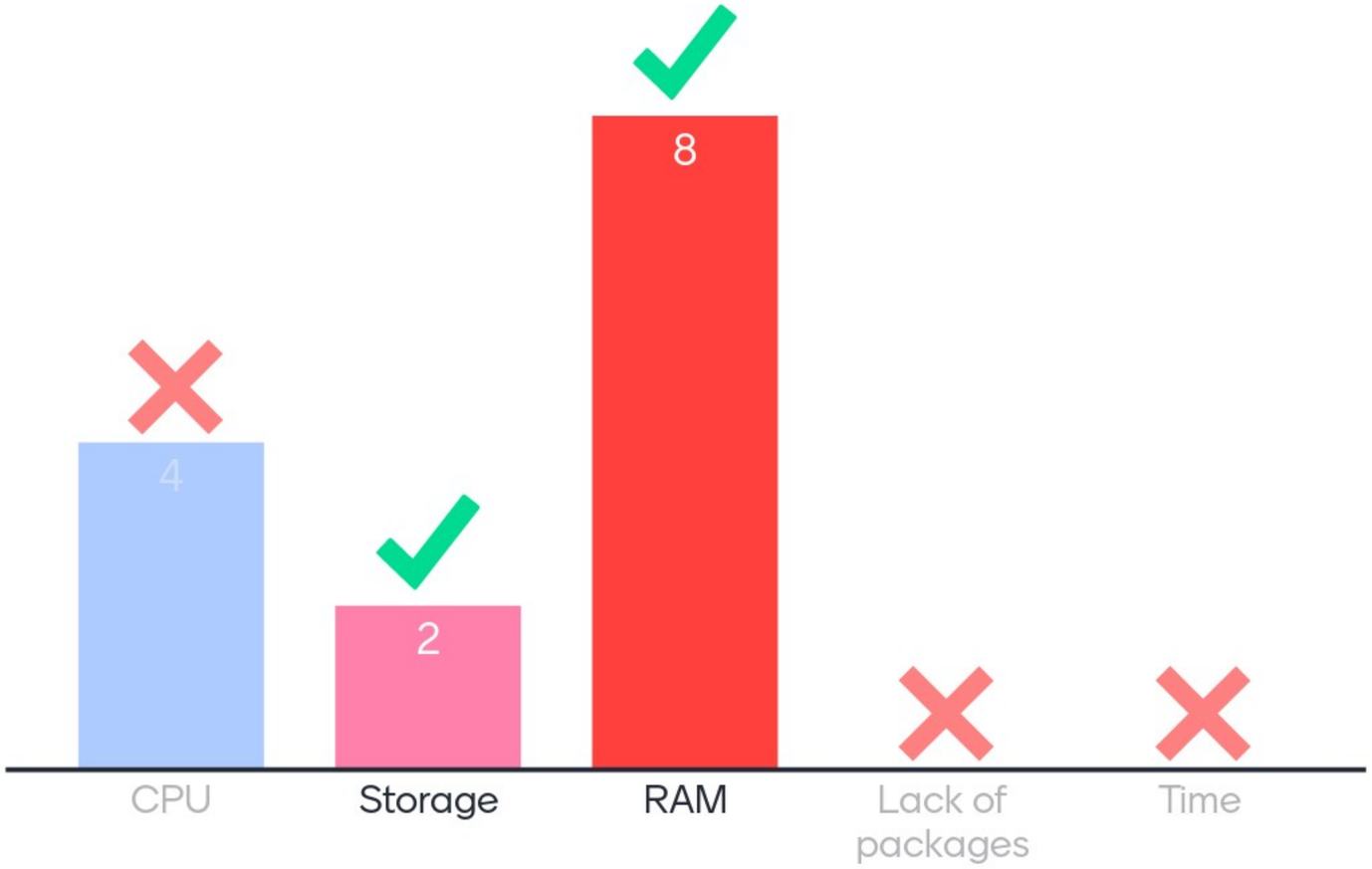


Leaderboard



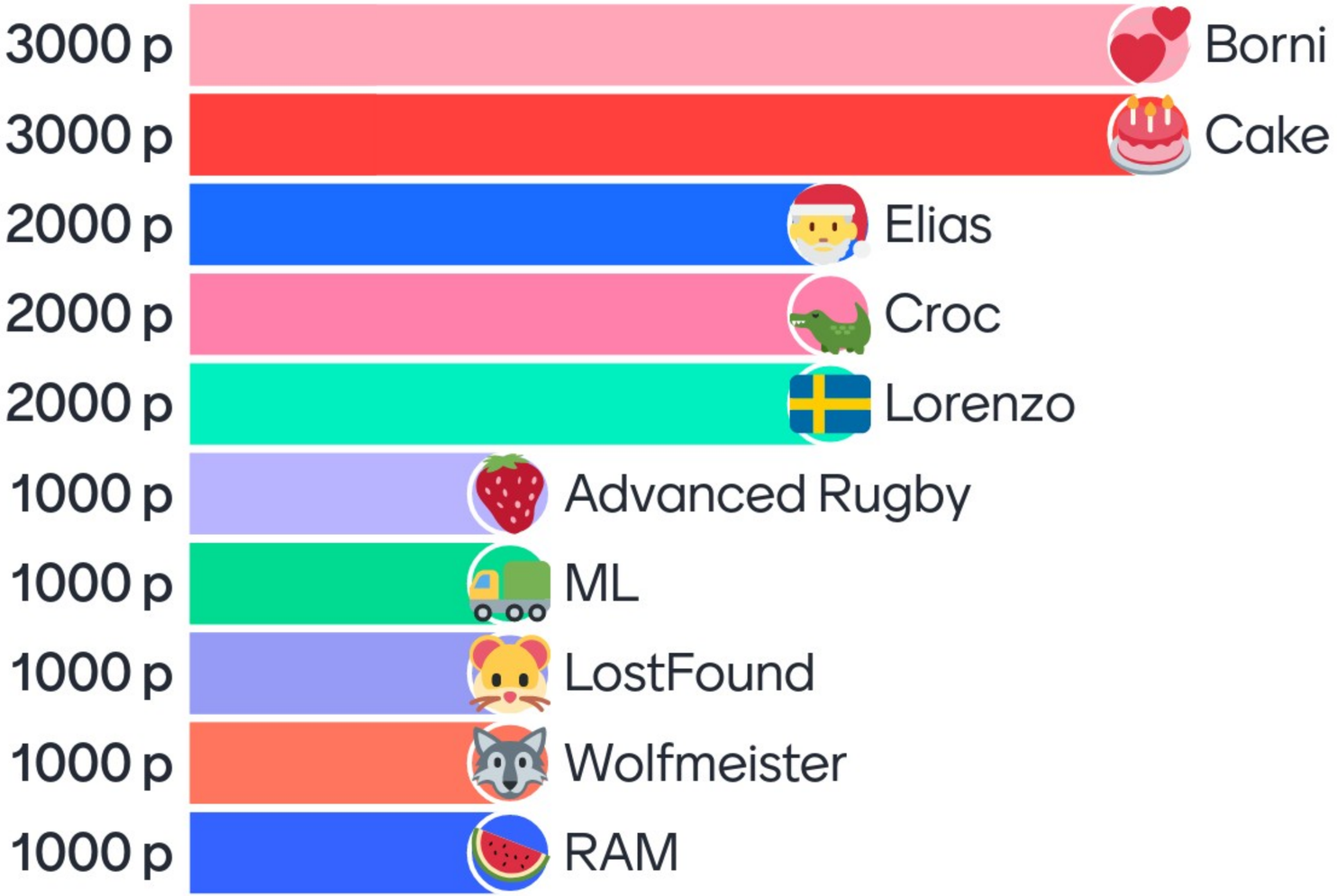
When using regular Python, our main limitation is...

Mentimeter



14

Leaderboard

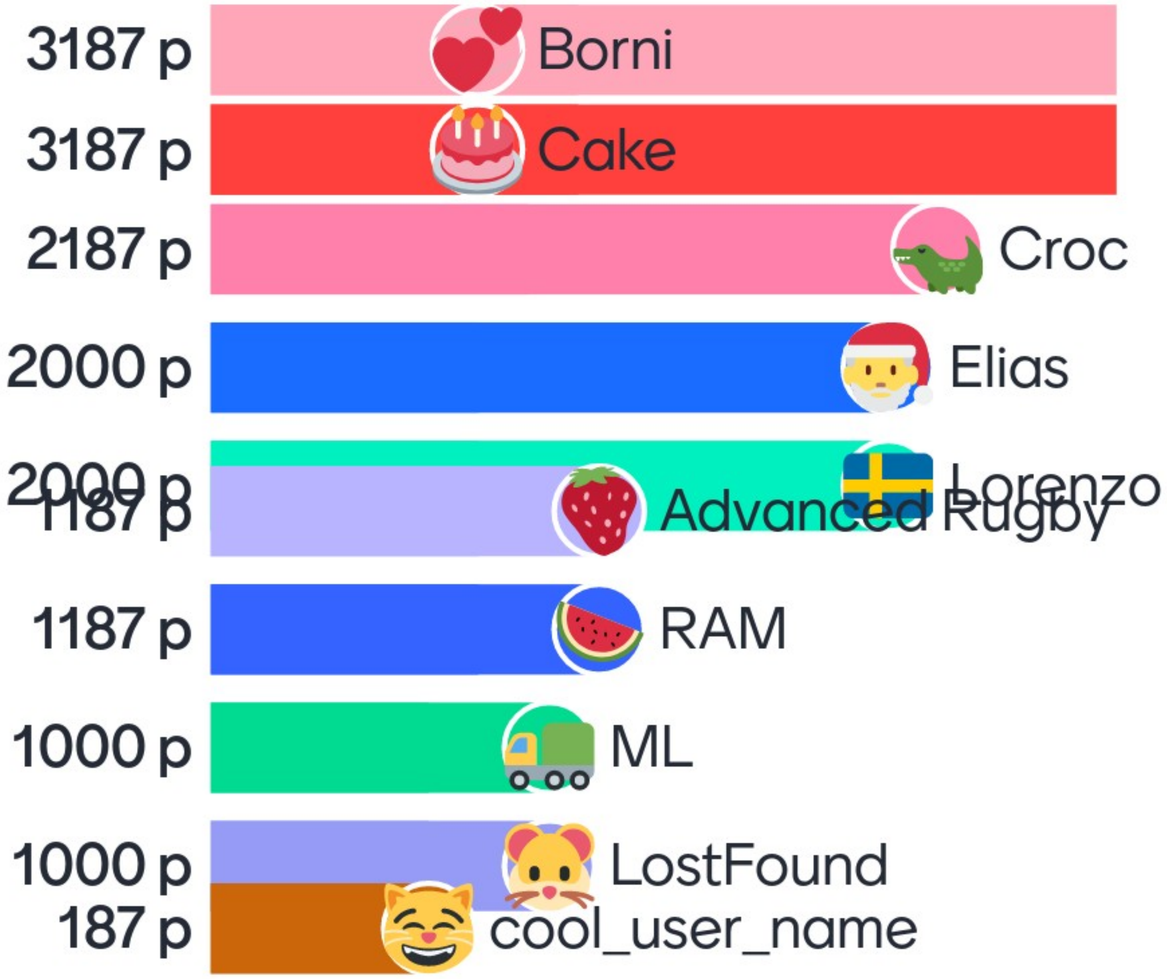


If you want to rank some explanatory variables based on their importance, you should use...

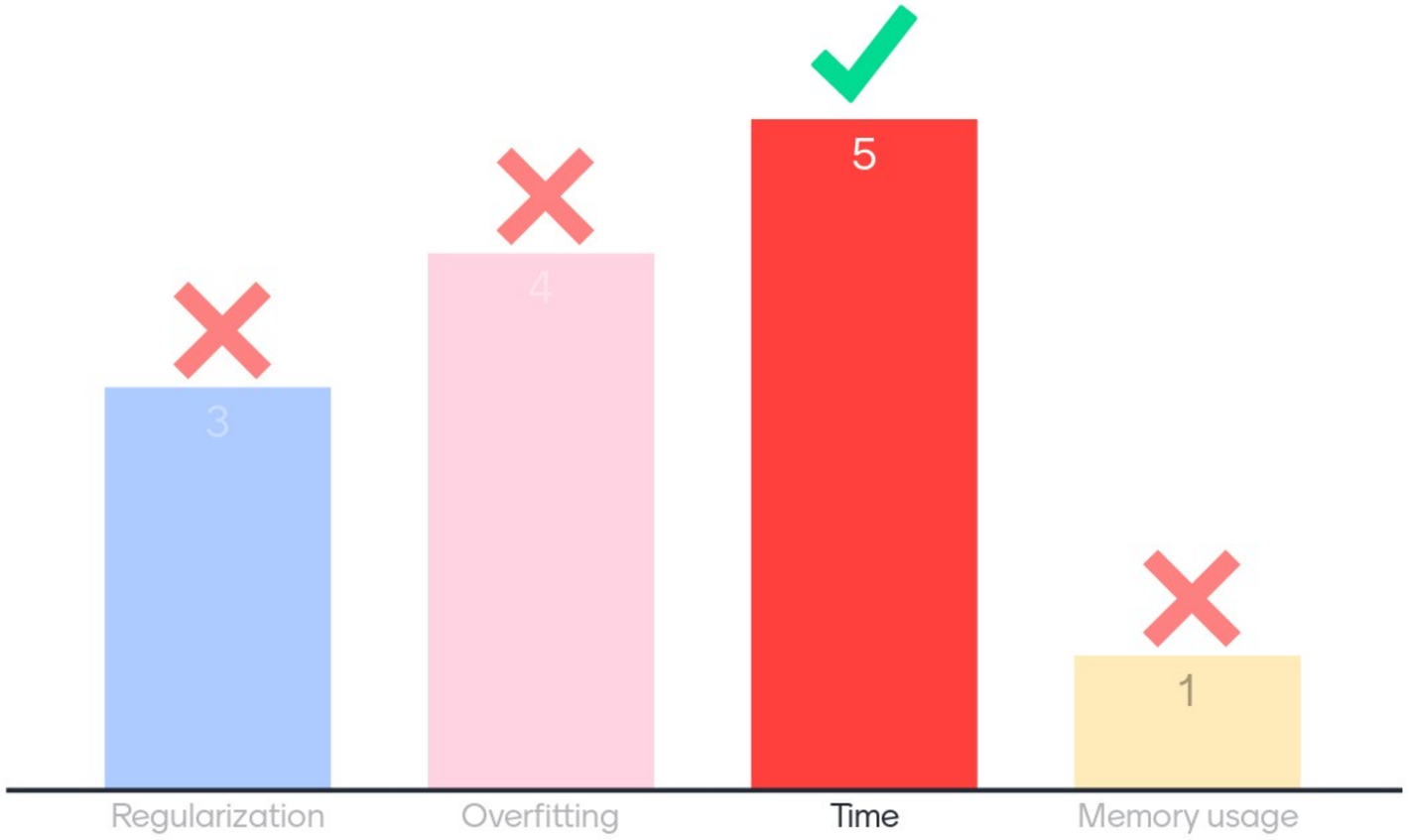
Mentimeter

14

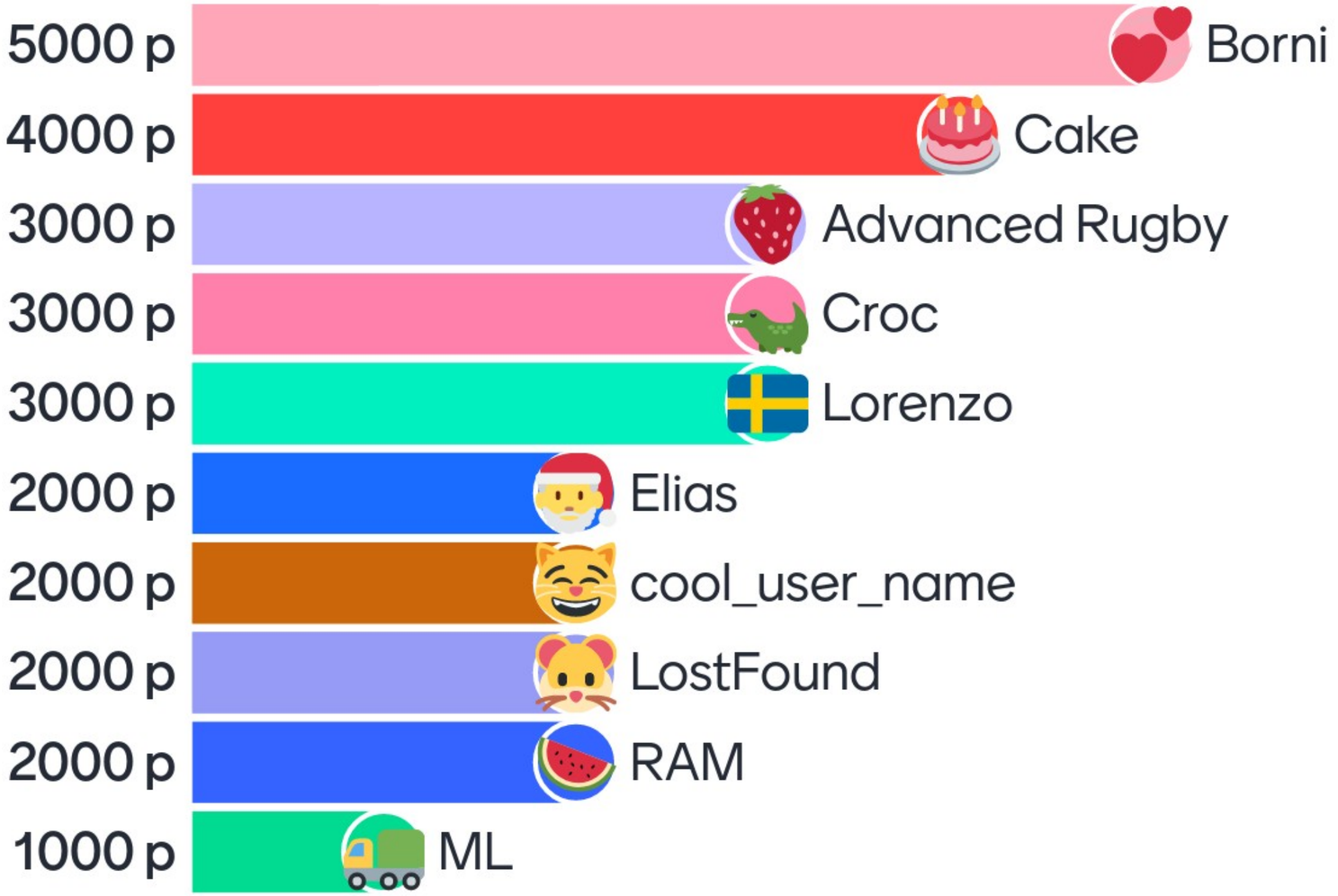
Leaderboard



The main problem of using a Leave One Out validation process is...

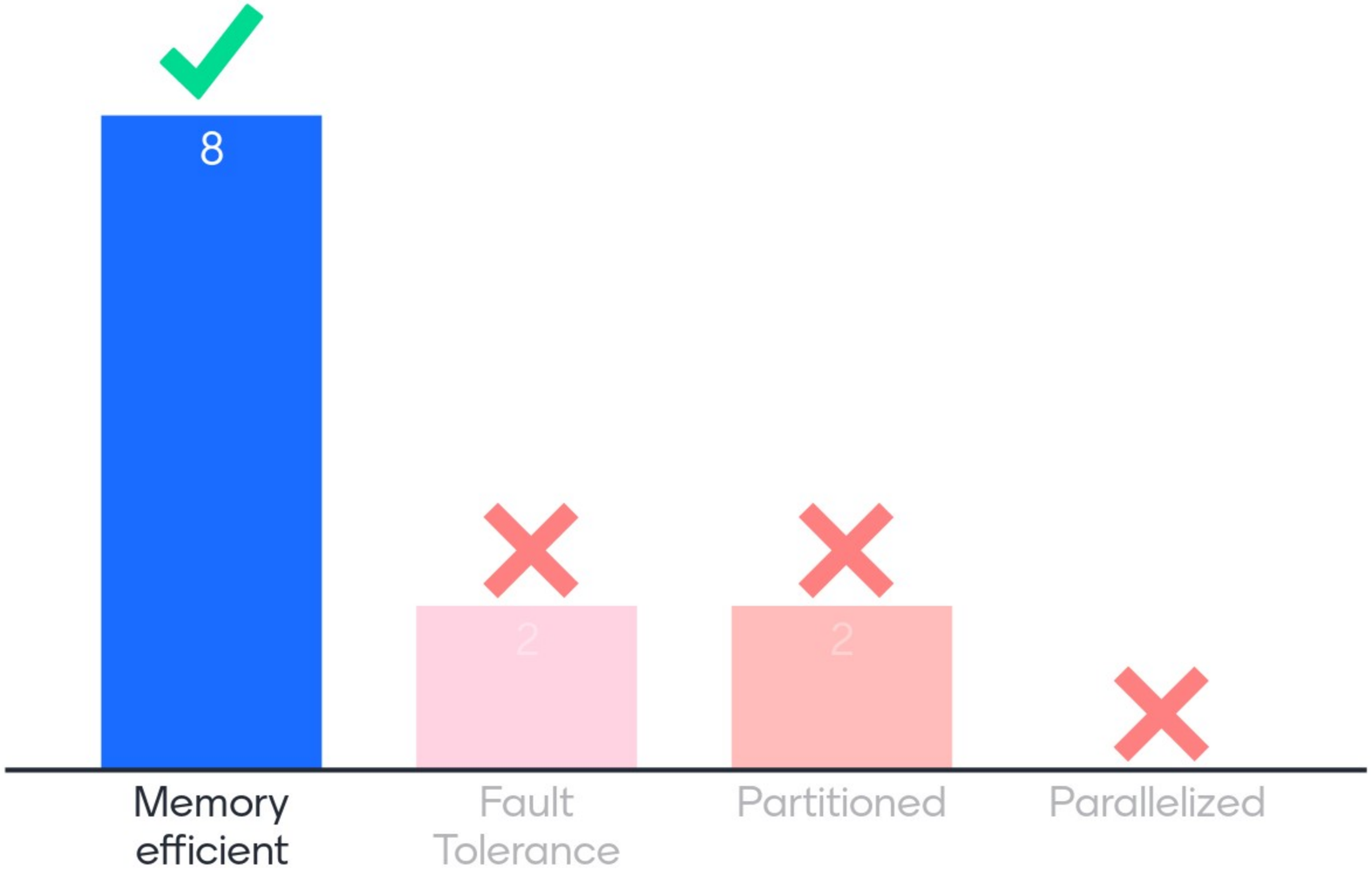


Leaderboard



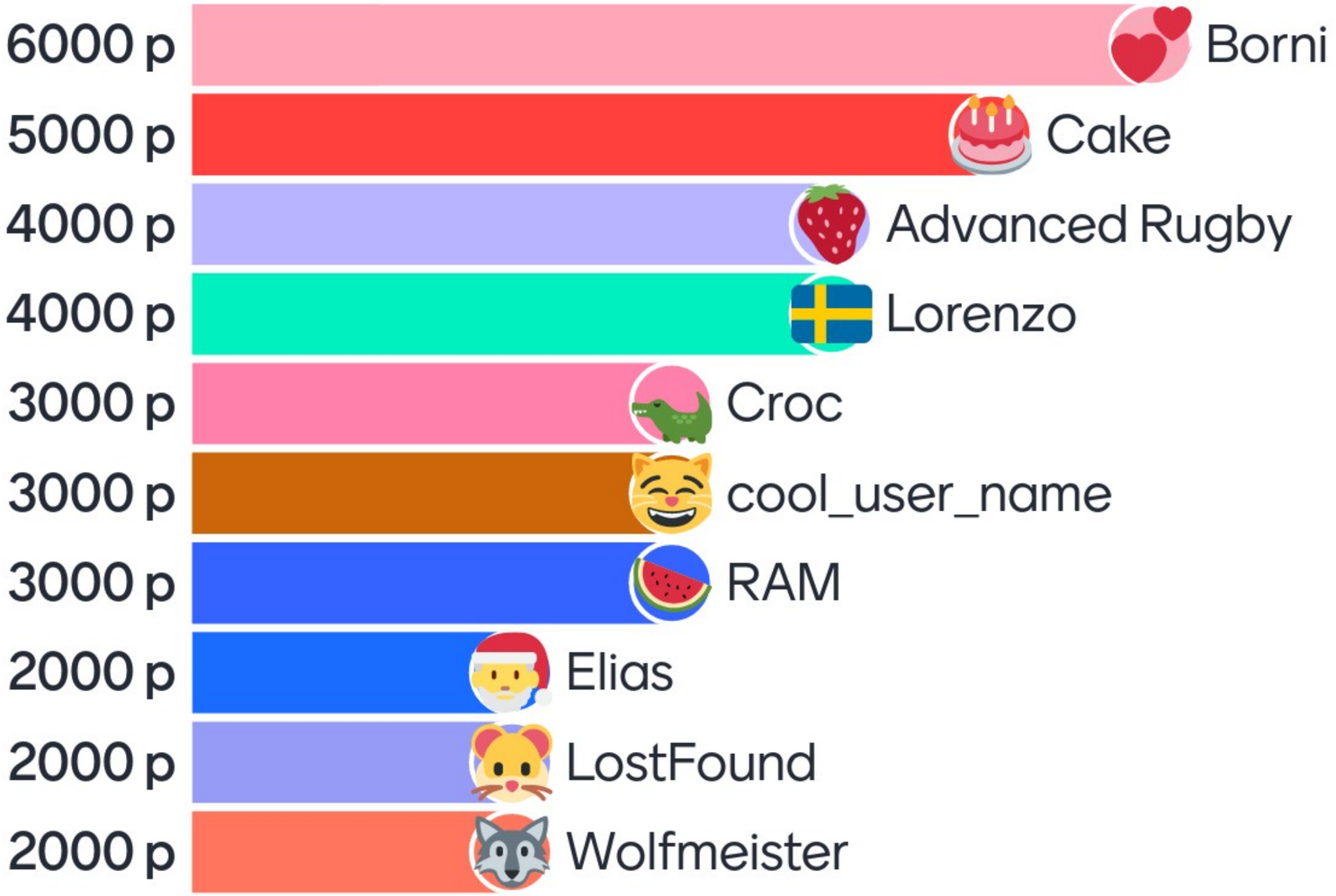
Which of these IS NOT a property of the RDDs?

Mentimeter



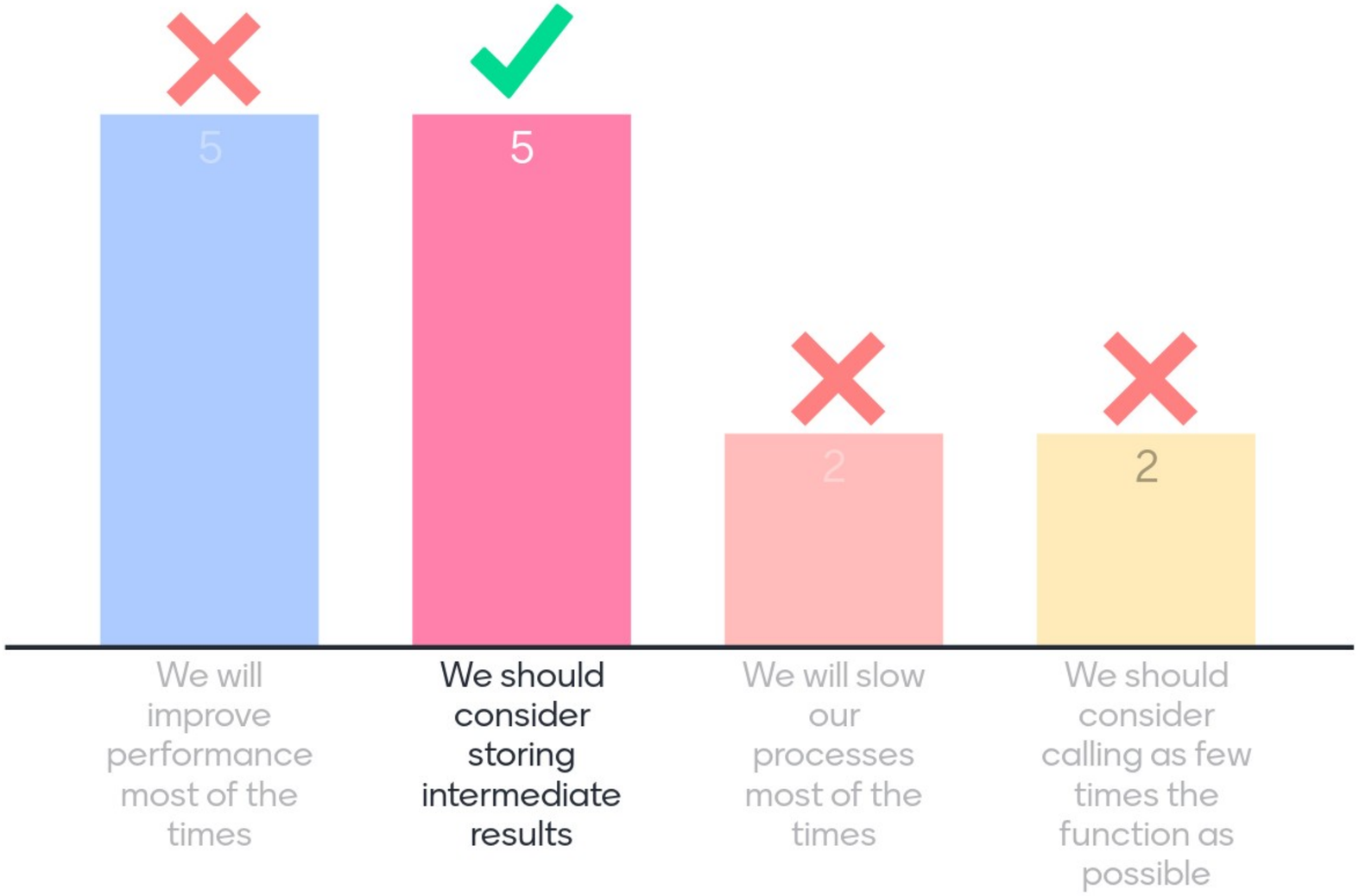
12

Leaderboard

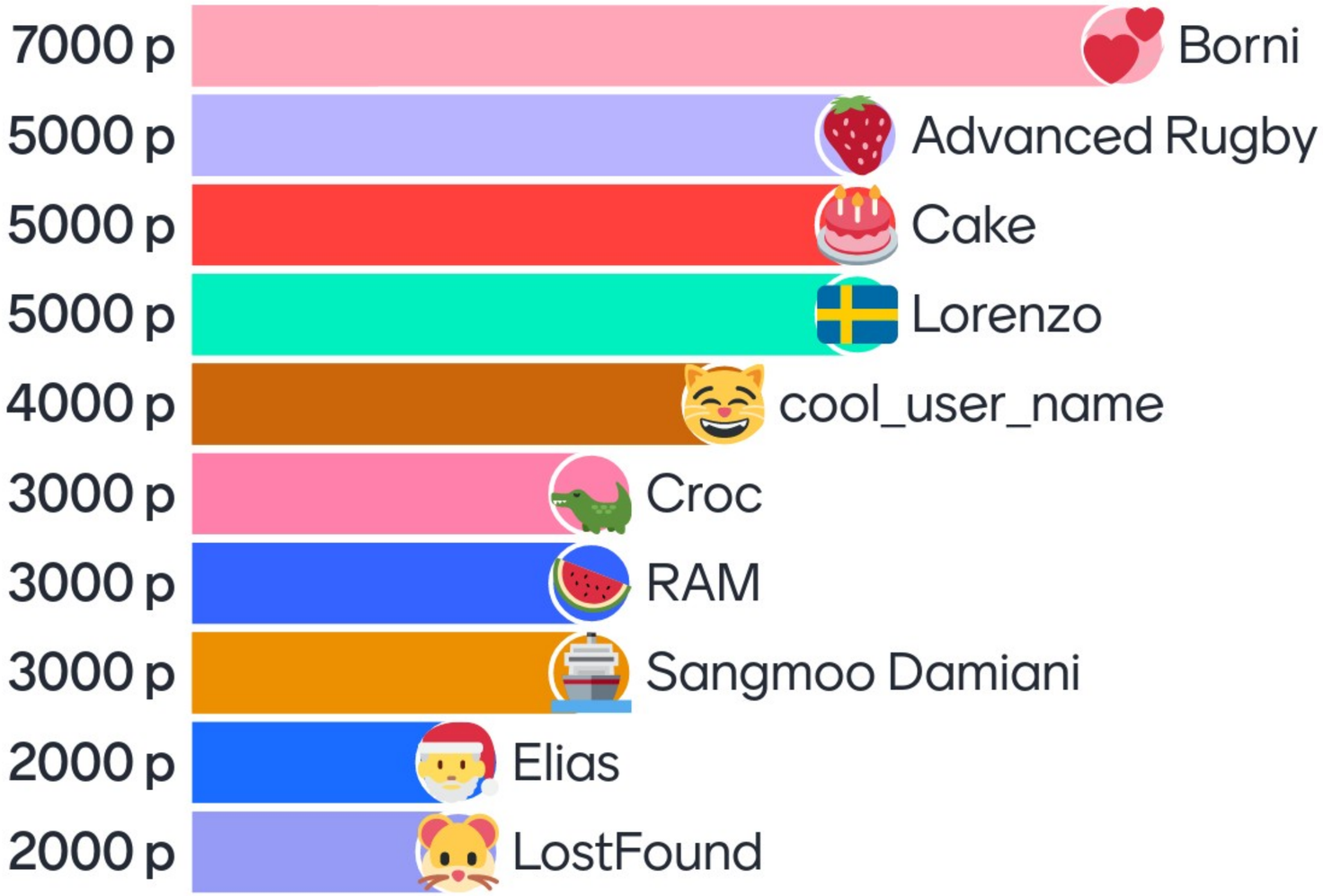


If we code a function in a recursive way...

Mentimeter

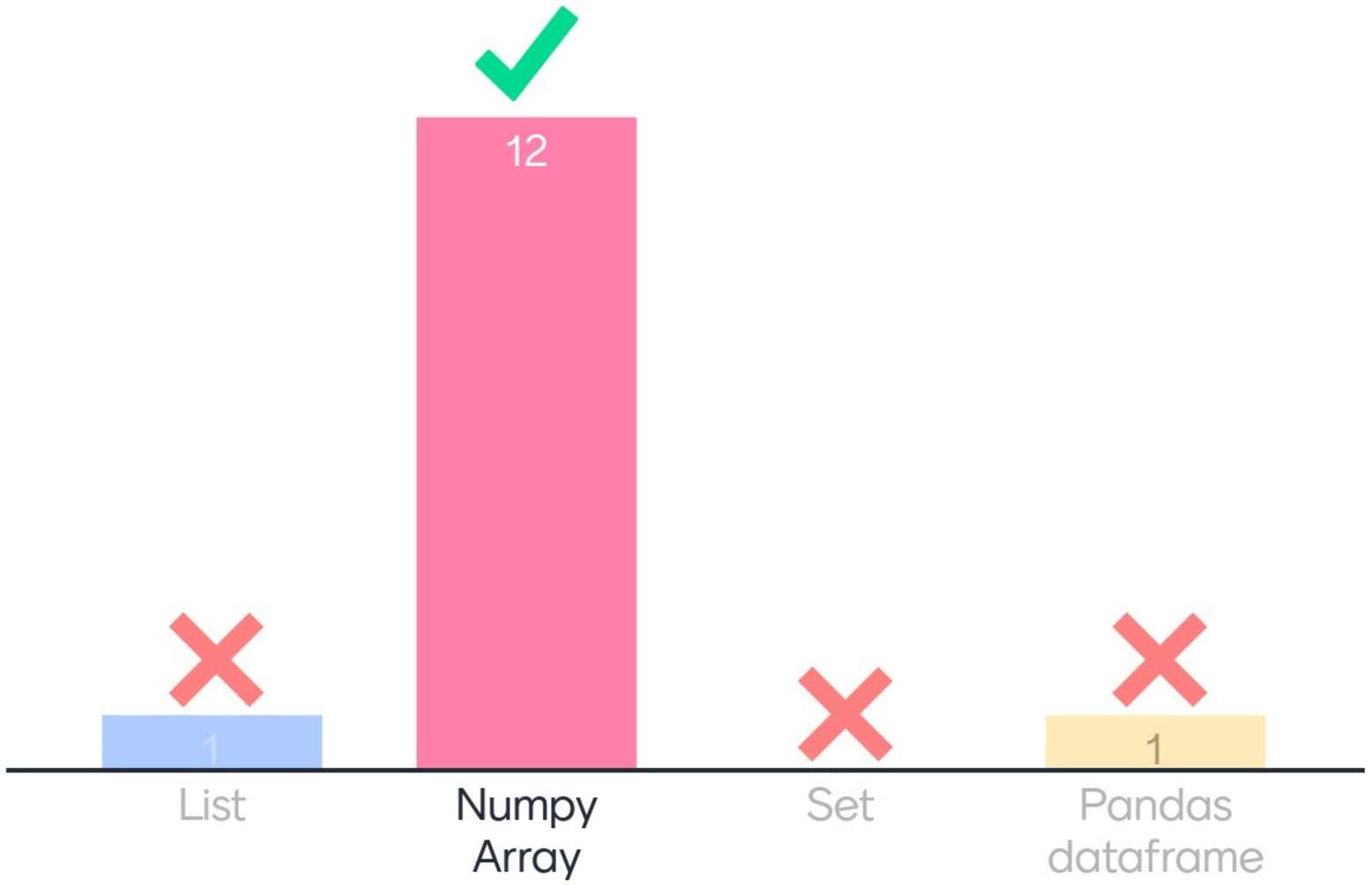


Leaderboard



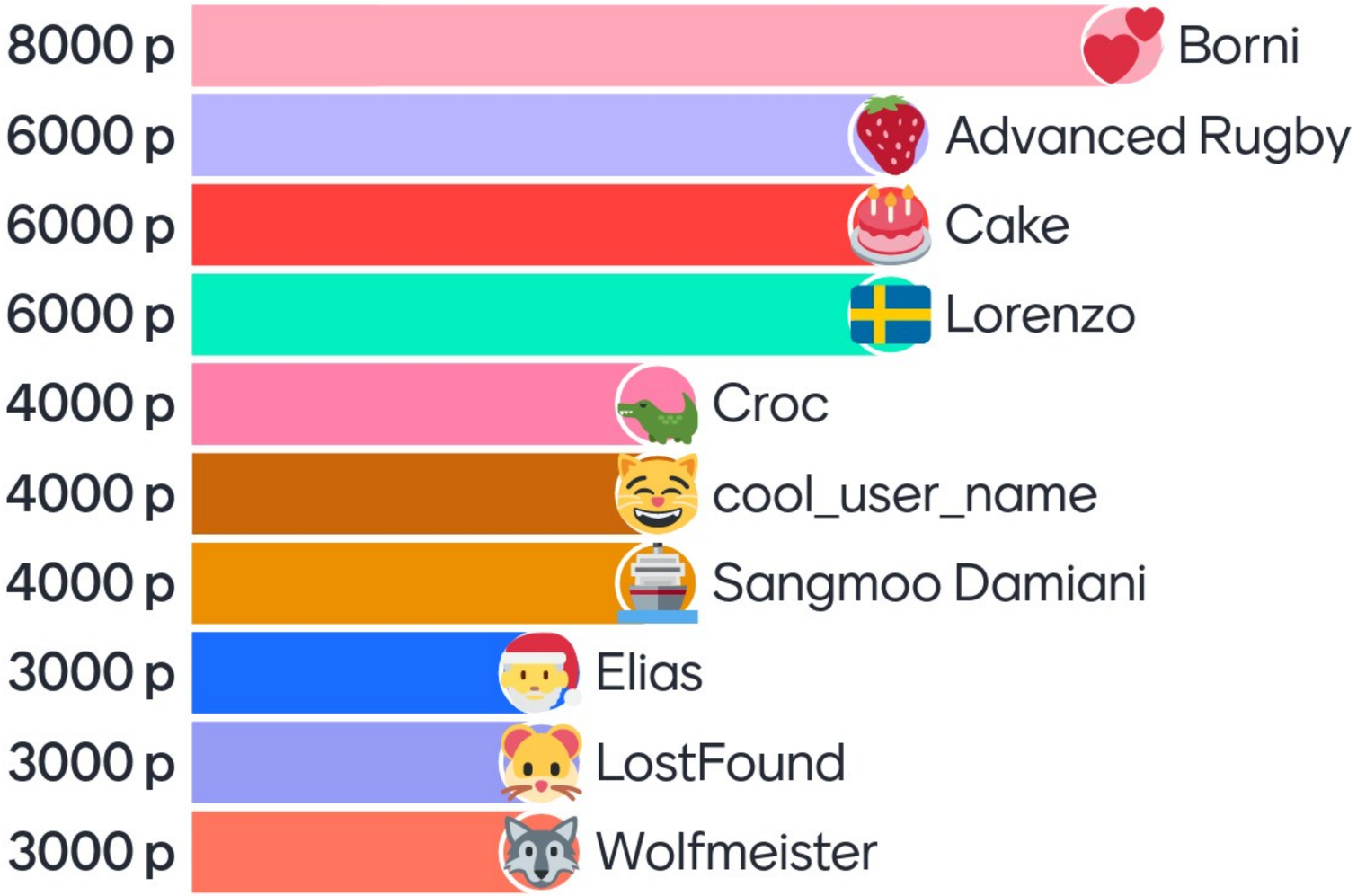
Which object is processed faster?

Mentimeter



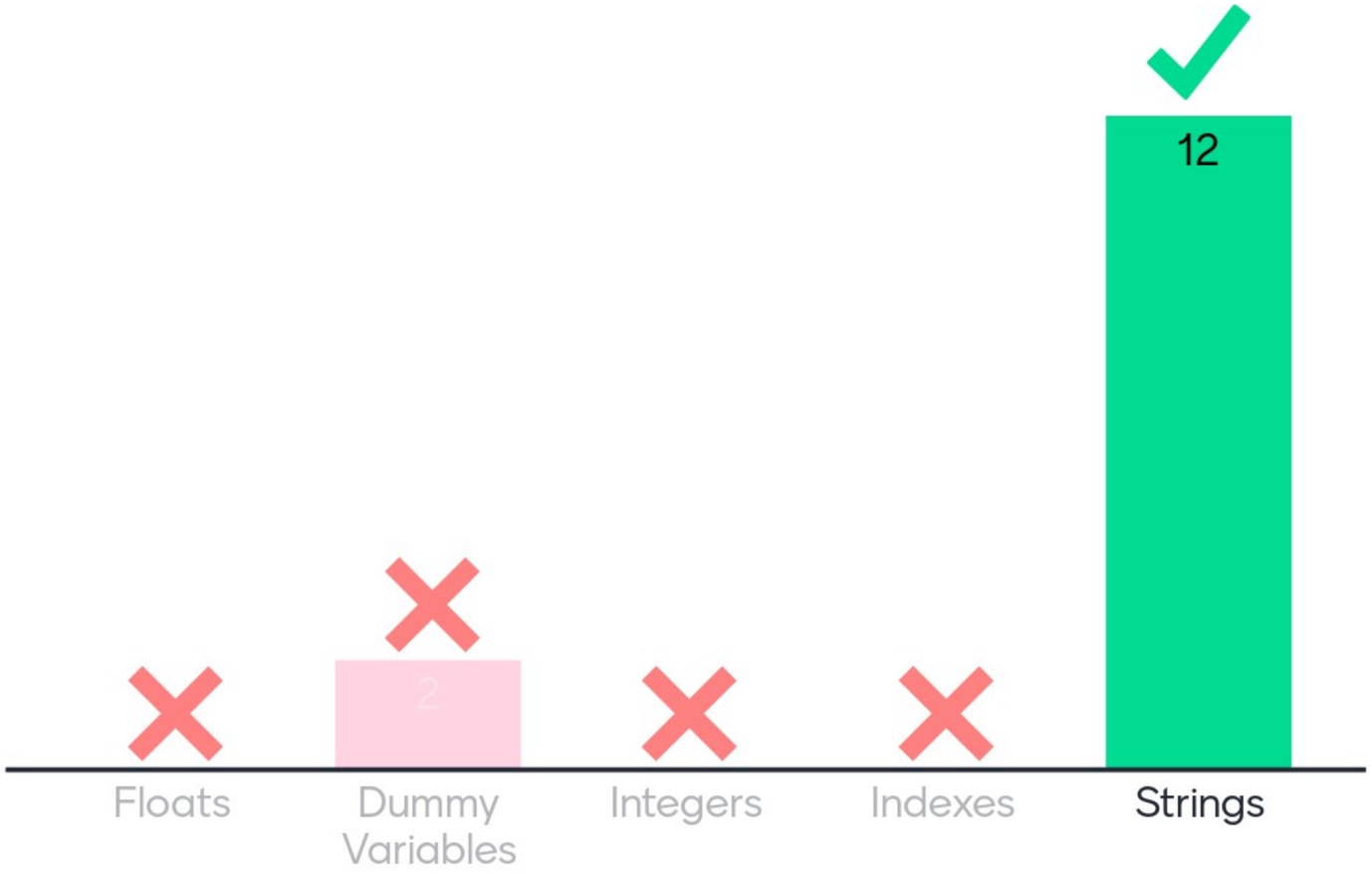
14

Leaderboard



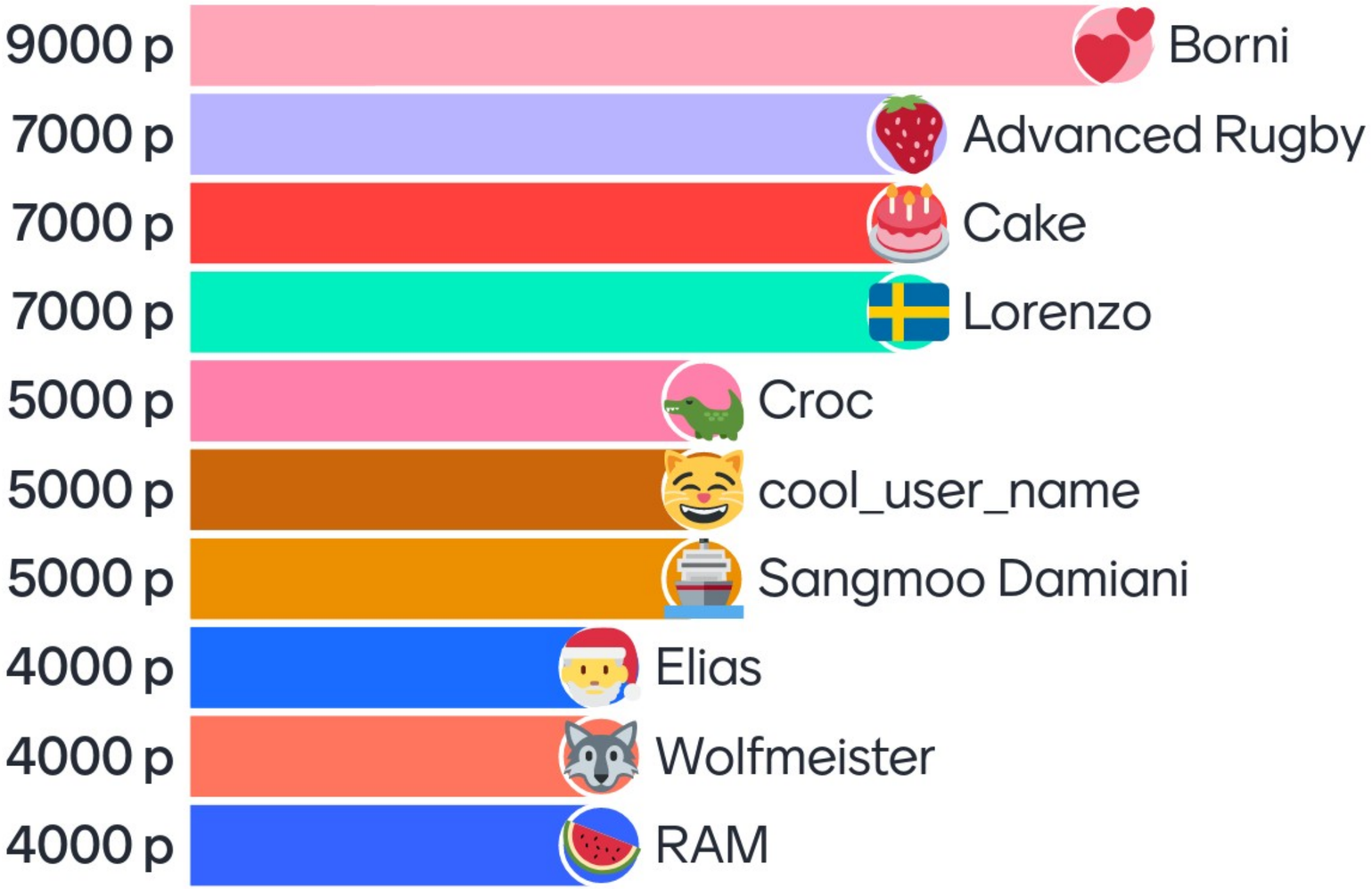
Which type of variable needs to be converted when doing ML?

Mentimeter



14

Leaderboard



TieBreaker

Be fast!

How many ways do we have of ordering "A", "B", "C" and "D"?

Mentimeter

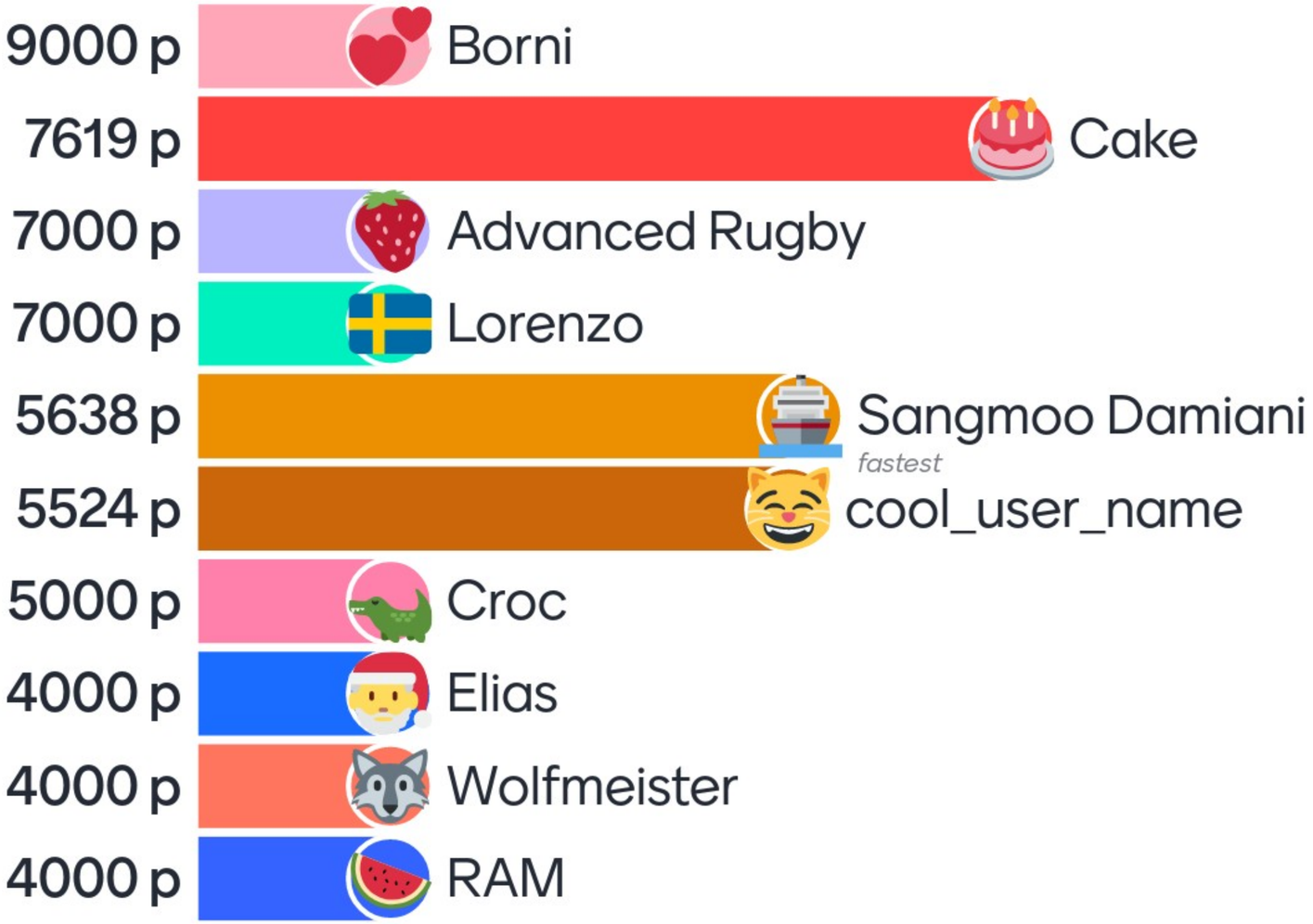
24 ✓ 3x	4 ✗	7 ✗
12 ✗	16 ✗ 4x	4 x 4 ✗

The correct answer is: 24

11

Leaderboard

Mentimeter



11

What is the Map Reduce model?

Mentimeter

MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.

A model on Hadoop to read or store information.

processing large data sets with a parallel , distributed algorithm on a cluster

Algorithm for processing data in parallel, distributed way

A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks.

is a processing technique and a program model for distributed computing. MapReduce program work in two phases, namely, Map and Reduce. Map tasks deal with splitting and mapping of data while Reduce tasks shuffle and reduce the data.

its a processing approach that utilizes parallel processing and HDFS

Parallel computing

Map phase divides the data into for analysis and reduce phase uses results from map tasks as input to a set of parallel reduce tasks.

What is a lambda infrastructure?

Mentimeter

Divides datastream into speed/batch layers for querying

Splits data processing into batch layer (most computaiotns are done here) and speed layer (streaming) to later integrate and serve the results.

Lambda architecture is a pivotal architecture in big data systems (by taking advantage of both batch and stream-processing)

Way of processing big data that provices access to batch and stream processing methods.

It is an infrastructure allowing processing data either using batch or streaming

No tengo ni idea

lambda is a structure of databases that enables using data in real time .. its uses datalake layers - real time layers and combines it into one serving layer ... its very complex and prone to errors

A way of storage data using batch layer and a permanent layer

8

Which skills would you like to improve the most in the upcoming months (in all the master)?

Mentimeter

ML

read only data in memory

Deploying real-time data applications

Data Ingestion, Streaming Processing, Data Processing, Model Selection, Model Deployment.

Immutable dataset

ML models

ML

big data manipulation

Skills most needed in big data industry :-)

16

Which skills would you like to improve the most in the upcoming months (in all the master)?

Mentimeter

ML without any libraries like sklearn

learn about actual use cases for big data

learn how to trick interviewers to hire me

Big data, model deployment.

Designing Big Data Architectures for different uses cases

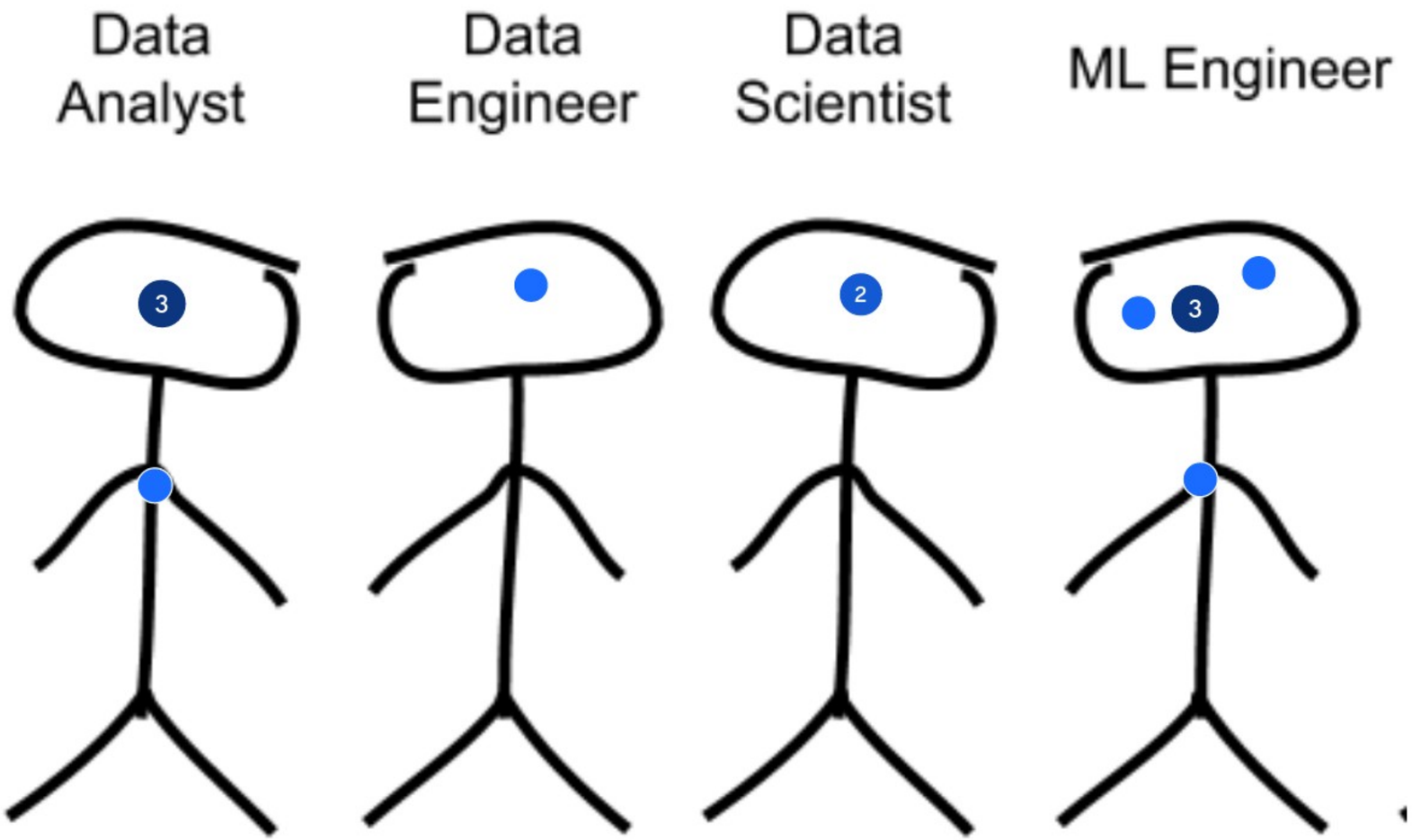
Use several vms to train ML models

- big data ecosystem (Infrastructure, Analytics, Applications) - ML Models

16

Which role fits you better?

Mentimeter



13

Implementing Spark

https://azure.microsoft.com/en-us/services/hdinsight/

Examples of stream data

Mentimeter

Supply Chain Security - Internet of things

Any data from IOT devices

Website interactions

JSON for weather

tweeter feeds

citizen surveillance cameras in china

Satellite imagery analysis over time

IoT device data that need real time processing

Location-based marketing.

Examples of stream data

Mentimeter

click counter

Amazon Stores where you can walk in and out without paying at a till

in-game player activity, e-commerce purchases, social networks, etc

gambling for Rugby

13

Which APIs have you worked with?

Mentimeter

yahoo finance, weather and crypto currency

Te develop an API: Flash, FastAPI, to connect to an API: requests,

Twitter

some weather API .. I have very little experience

NOAA, NREL

building apis using Asp.net WebApi

google maps api

7

Implementing Streaming Analytics

https://azure.microsoft.com/en-us/services/stream-analytics/

Time to install

How to install pyspark in Windows/Ubuntu/Mac...