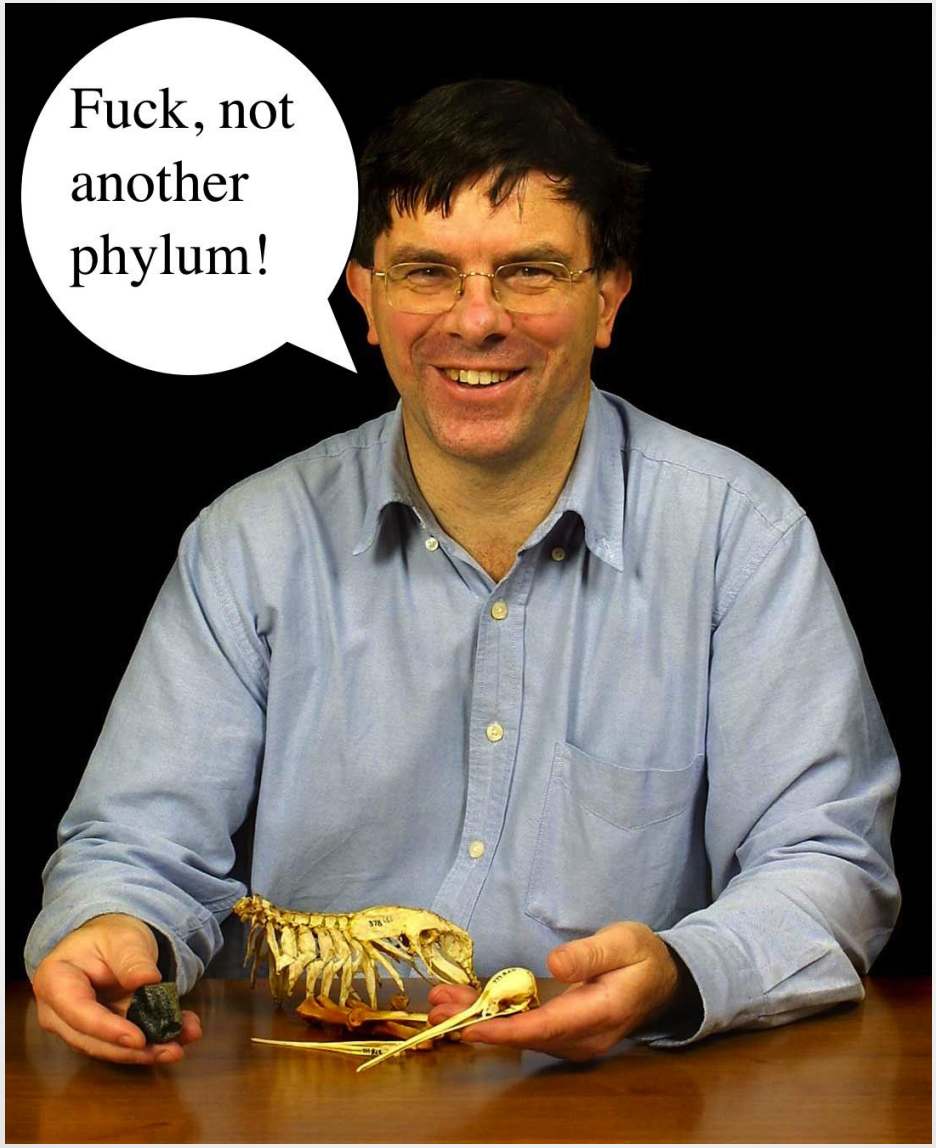
Does Python stand a chance in today’s world of data science?

Radim Řehůřek

YES


hadoop
HIVE
APACHE HBASE
elasticsearch.
cassandra
CouchDB relax
HDFS
cascading
cloudera IMPALA
mongoDB
TITAN
Microsoft Azure
Puppet
CHEF
Amazon Redshift
OpsWorks
Spark
amazon web services
ANSIBLE
redis
STORM
kafka
scikit learn
NumPy
ØMQ
pandas
matplotlib
Celery
RabbitMQ
Open Source Enterprise Messaging
IP[y]: IPython
Interactive Computing


Fuck, not another phylum!

RaRe Technologies Ltd.

AUTODESK

rally

elevate

HEARST

DATASIFT

SPORTS AUTHORITY.

Brainspace

harvest.ai

SUWECO

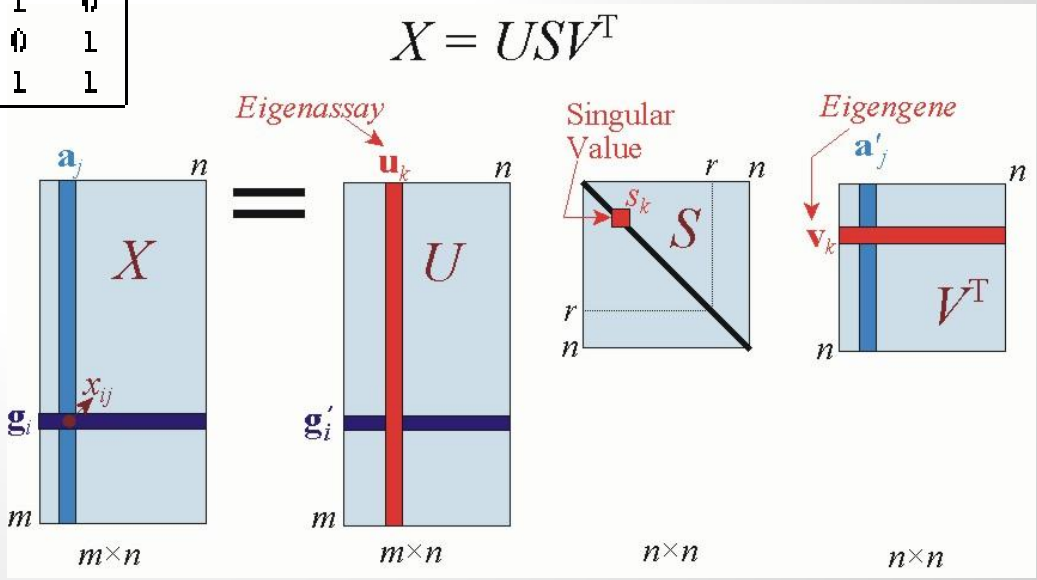
Eva
Expert Virtual Agent

Python vs. rest

- performance?
- deployment?
- logging, debugging?
- workflow, integration?

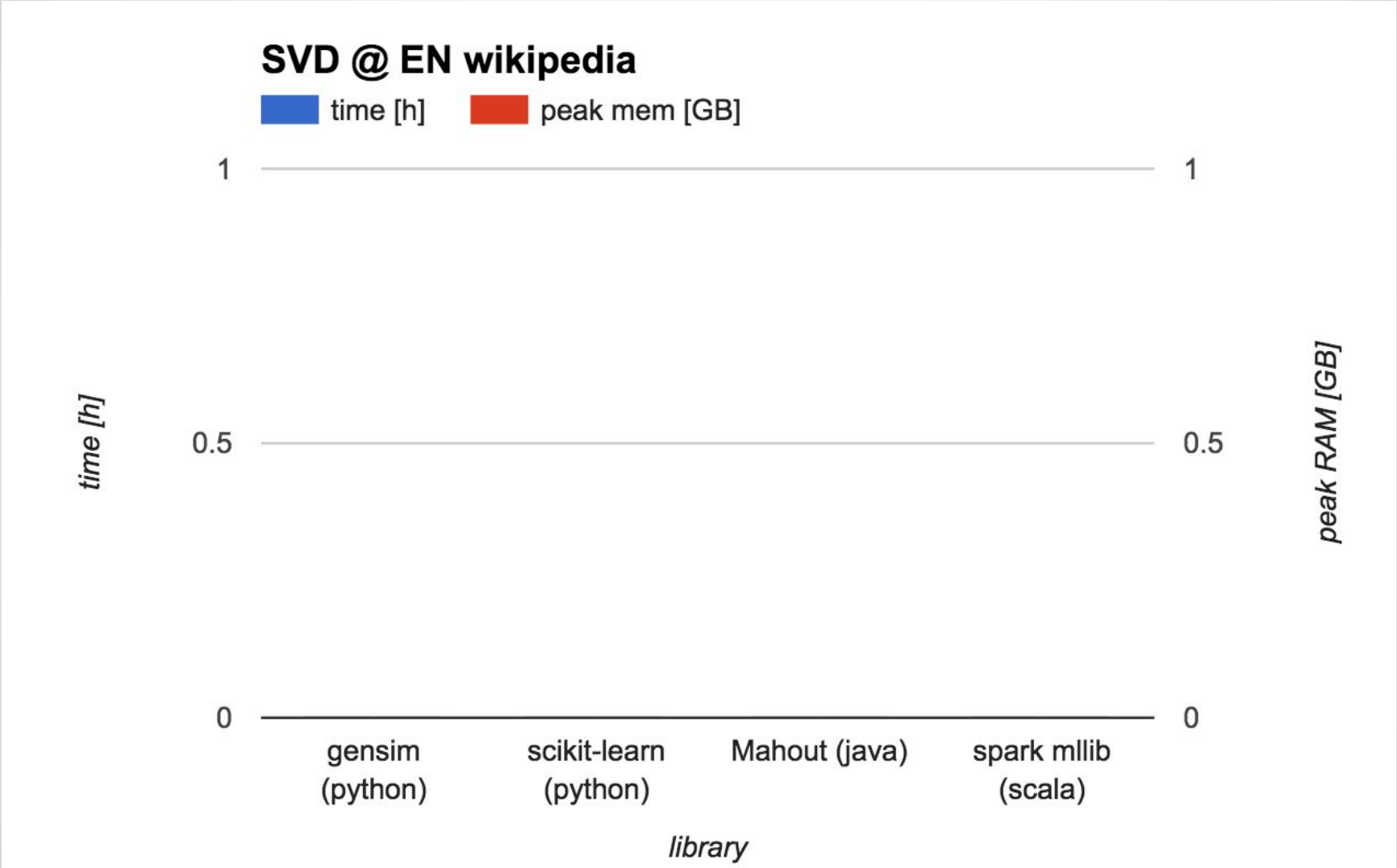
SVD

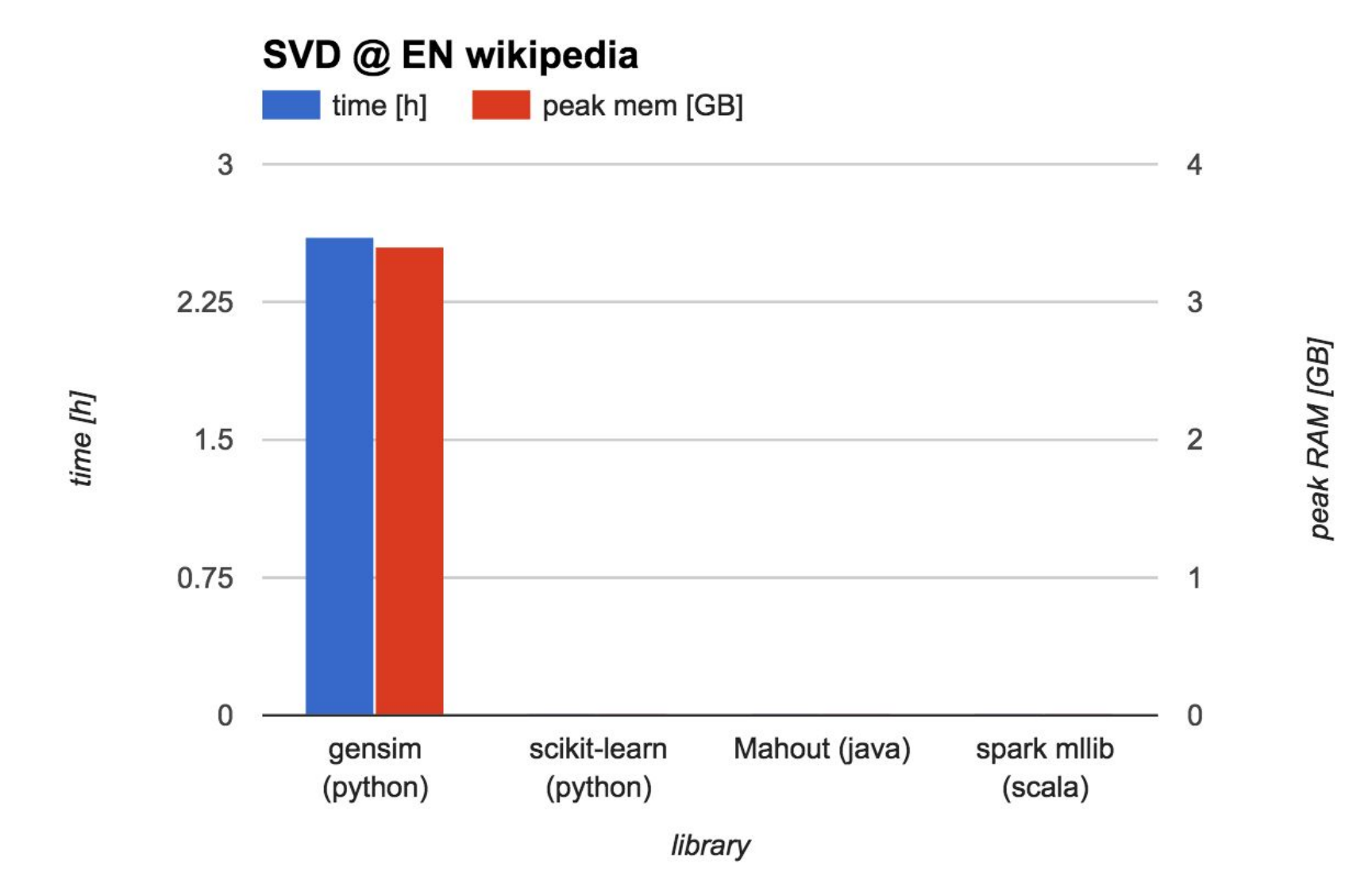
Terms	Documents								
	m1	m2	m3	m4	c2	c5	c3	c4	c1
trees	1	1	1	0	0	0	0	0	0
graph	0	1	1	1	0	0	0	0	0
minors	0	0	1	1	0	0	0	0	0
survey	0	0	0	1	1	0	0	0	0
time	0	0	0	0	1	1	0	0	0
response	0	0	0	0	1	1	0	0	0
user	0	0	0	0	1	1	1	0	0
computer	0	0	0	0	1	0	0	0	1
system	0	0	0	0	1	0	1	2	0
EPS	0	0	0	0	0	0	1	1	0
interface	0	0	0	0	0	0	1	0	1
human	0	0	0	0	0	0	0	1	1

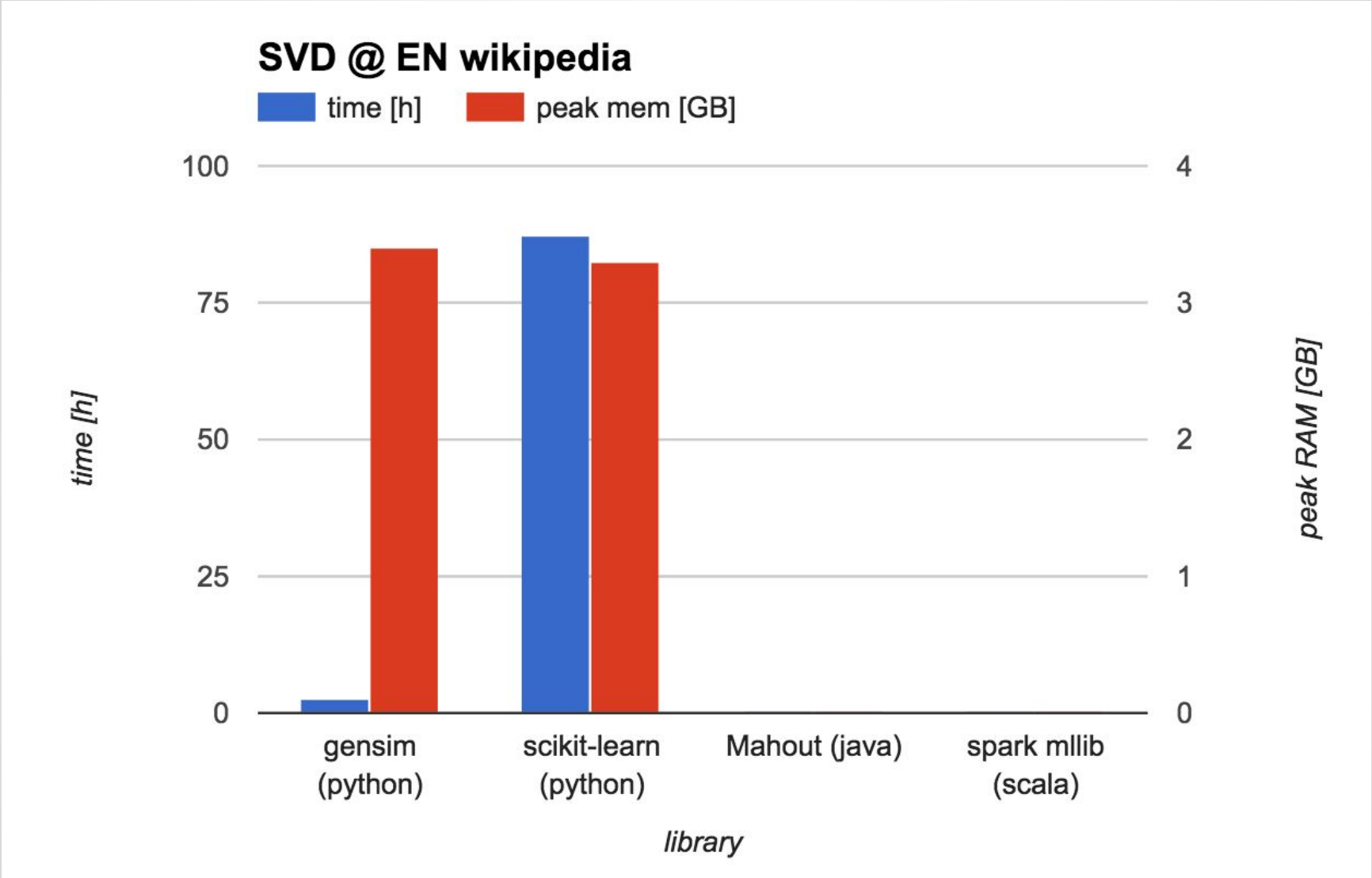


English Wikipedia

- ~3.5M docs
- ~2G words
- with 100K vocab, ~0.5G matrix non-zeros
  - very sparse
- small-ish, but known & accessible and **out -of-core**


SVD @ EN wikipedia
time [h]
peak mem [GB]
1
0.5
0
1
0.5
0
time [h]
peak RAM [GB]
gensim (python)
scikit-learn (python)
Mahout (java)
spark mllib (scala)
library


SVD @ EN wikipedia
time [h]
peak mem [GB]
3
2.25
1.5
0.75
0
4
3
2
1
0
time [h]
peak RAM [GB]
gensim (python)
scikit-learn (python)
Mahout (java)
spark mllib (scala)
library


SVD @ EN wikipedia
time [h]
peak mem [GB]
100
75
50
25
0
4
3
2
1
0
time [h]
peak RAM [GB]
gensim (python)
scikit-learn (python)
Mahout (java)
spark mllib (scala)
library

Spark mllib

- top level Apache project, Scala
- RDDs, Resilient Distributed Datasets
- ~RAM caching + execution engine
- latest Spark 1.3.0 + mllib
- AWS EMR cluster (4x m3.xlarge)

SVD @ mllib

```
15/04/17 21:49:20 INFO scheduler.TaskSetManager: Starting task 17.0 in stage 7.0 (TID 374, ip-172-31-59-4.ec2.internal, PROCESS_L
OCAL, 1107 bytes)
15/04/17 21:49:20 WARN scheduler.TaskSetManager: Lost task 13.0 in stage 7.0 (TID 371, ip-172-31-59-4.ec2.internal): java.lang.Ar
rayIndexOutOfBoundsException: 300000
        at breeze.linalg.operators.DenseVector_SparseVector_Ops$$anon$98.apply(SparseVectorOps.scala:302)
        at breeze.linalg.operators.DenseVector_SparseVector_Ops$$anon$98.apply(SparseVectorOps.scala:282)
        at breeze.linalg.ImmutableNumericOps$class.dot(NumericOps.scala:98)
        at breeze.linalg.DenseVector.dot(DenseVector.scala:50)
        at breeze.linalg.operators.SparseVector_DenseVector_Ops$$anon$58.apply(SparseVectorOps.scala:167)
        at breeze.linalg.operators.SparseVector_DenseVector_Ops$$anon$58.apply(SparseVectorOps.scala:164)
        at breeze.linalg.operators.BinaryRegistry$class.apply(BinaryOp.scala:60)
        at breeze.linalg.VectorOps$$anon$171.apply(Vector.scala:528)
        at breeze.linalg.ImmutableNumericOps$class.dot(NumericOps.scala:98)
        at breeze.linalg.SparseVector.dot(SparseVector.scala:49)
        at org.apache.spark.mllib.linalg.distributed.RowMatrix$$anonfun$5.apply(RowMatrix.scala:96)
        at org.apache.spark.mllib.linalg.distributed.RowMatrix$$anonfun$5.apply(RowMatrix.scala:94)
        at scala.collection.TraversableOnce$$anonfun$foldLeft$1.apply(TraversableOnce.scala:144)
        at scala.collection.TraversableOnce$$anonfun$foldLeft$1.apply(TraversableOnce.scala:144)
        at scala.collection.Iterator$class.foreach(Iterator.scala:727)
        at org.apache.spark.InterruptibleIterator.foreach(InterruptibleIterator.scala:28)
```

Spark / SPARK-3803

ArrayIndexOutOfBoundsException found in executing computePrincipalC

Agile Board

Details

Type:	Bug	Status:	RESOLVED
Priority:	Major	Resolution:	Fixed
Affects Version/s:	1.1.0	Fix Version/s:	1.2.0
Component/s:	MLlib		
Labels:	None		

Description

When I executed computePrincipalComponents method of RowMatrix, I got java.lang.ArrayIndexOutOfBoundsException.

```
14/10/05 20:16:31 INFO DAGScheduler: Failed to run reduce at RDDFunctions.scala:111
org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 31.0 failed 1 times, most
recent failure: Lost task 0.0 in stage 31.0 (TID 611, localhost): java.lang.ArrayIndexOutOfBoundsException:
4878161

org.apache.spark.mllib.linalg.distributed.RowMatrix$.org$apache$spark$mllib$linalg$distributed$RowMatrix$$dsp

        org.apache.spark.mllib.linalg.distributed.RowMatrix$$anonfun$3.apply(RowMatrix.scala:114)
```

Mahout SSVD

- the “scikit-learn” of Hadoop, Java
- originally on MapReduce
- now Mahout Samsara @ Spark, Scala
- newest Mahout 0.10.0

```
15/04/17 15:05:13 INFO metrics.MetricsSaver: Saved 8:22 records to /mnt/var/em/raw/i-4c684163_201
15/04/17 15:05:43 INFO metrics.MetricsSaver: Saved 8:22 records to /mnt/var/em/raw/i-4c684163_201
Exception in thread "main" java.io.IOException: Bt job unsuccessful.
        at org.apache.mahout.math.hadoop.stochasticsvd.BtJob.run(BtJob.java:625)
        at org.apache.mahout.math.hadoop.stochasticsvd.SSVDSolver.run(SSVDSolver.java:433)
        at org.apache.mahout.math.hadoop.stochasticsvd.SSVDCli.run(SSVDCli.java:167)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
        at org.apache.mahout.math.hadoop.stochasticsvd.SSVDCli.main(SSVDCli.java:198)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.hadoop.util.ProgramDriver$ProgramDescription.invoke(ProgramDriver.java:72)
        at org.apache.hadoop.util.ProgramDriver.run(ProgramDriver.java:145)
        at org.apache.hadoop.util.ProgramDriver.driver(ProgramDriver.java:153)
```

+ “local mode” eats up all disk, then fails

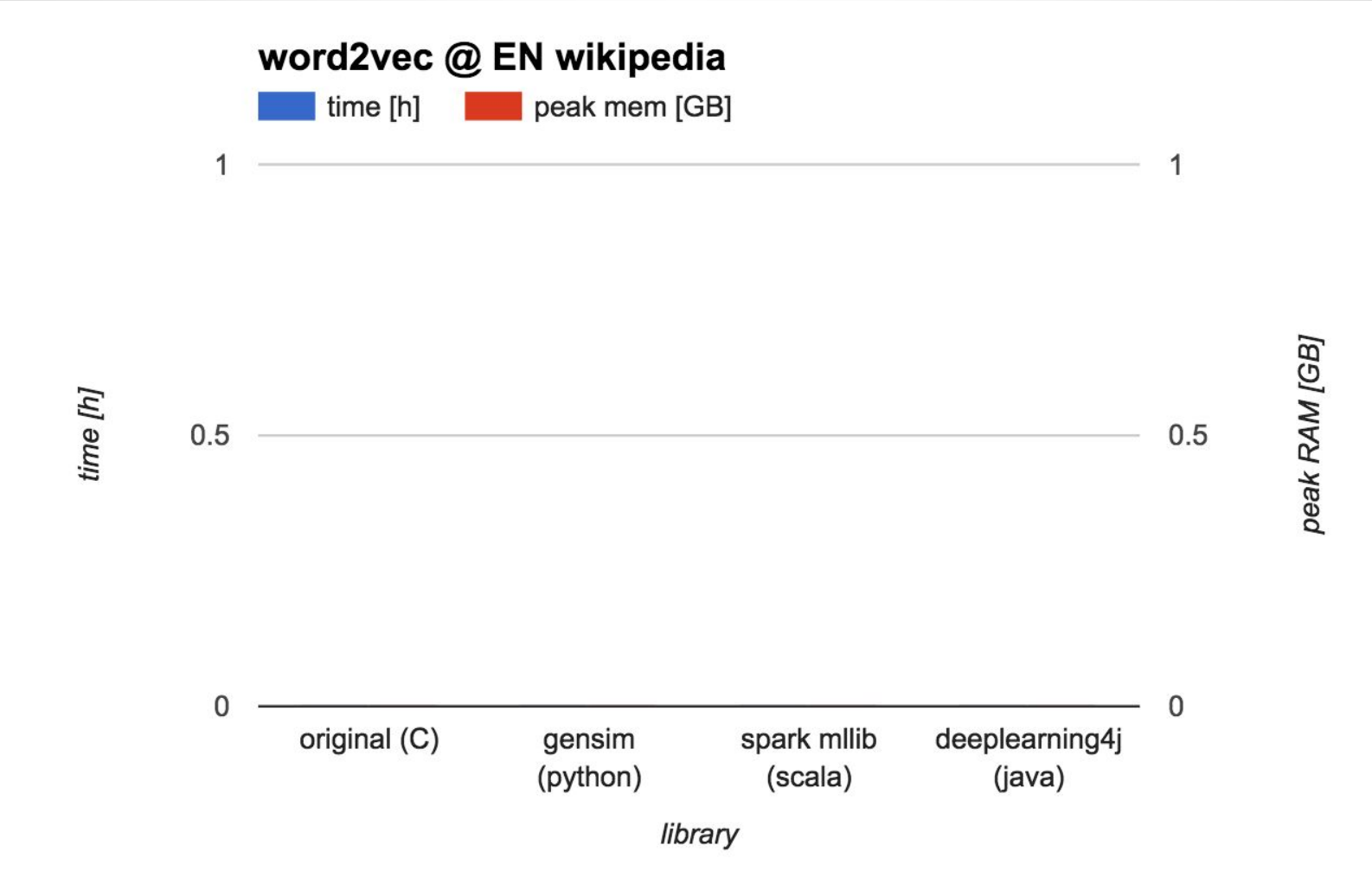
Google's word2vec

unsupervised ML

- *Berlin* is to *Germany* as *Paris* is to …?
- *king* - *man* + *woman* = *queen*
- which word doesn't fit? *"dinner cereal breakfast lunch"*

http://radimrehurek.com/2014/02/word2vec-tutorial/#app

Word2vec @ Wikipedia



C vs. NumPy vs. optimized

optimization	words per second	speed-up
NumPy baseline	1.4k	1.0x
original C word2vec	29.0k	20.7x
Cython	33.3k	23.8x
Cython + BLAS	89.8k	64.1x
Cython + sigmoid table	34.7k	24.8x
Cython + BLAS + sigmoid table	101.8k	72.7x

(+pure Python: **120x slower** than baseline)

Single machine parallelization

	# worker threads (speed/peak RAM/accuracy)			
implementation	1	2	3	4
C word2vec	22.6k / 252MB / 27.4%	42.94k / 252MB / 26.4%	62.04k / 252MB / 26.8%	72.44k / 252MB / 27.2%
gensim word2vec	109.5k / 591MB / 27.5%	191.6k / 596MB / 27.1%	263k / 592MB / 27.3%	**311.7k** / 601MB / **28.2%**

C (1/2/4 workers): 1.0x / 1.9x / 3.2x

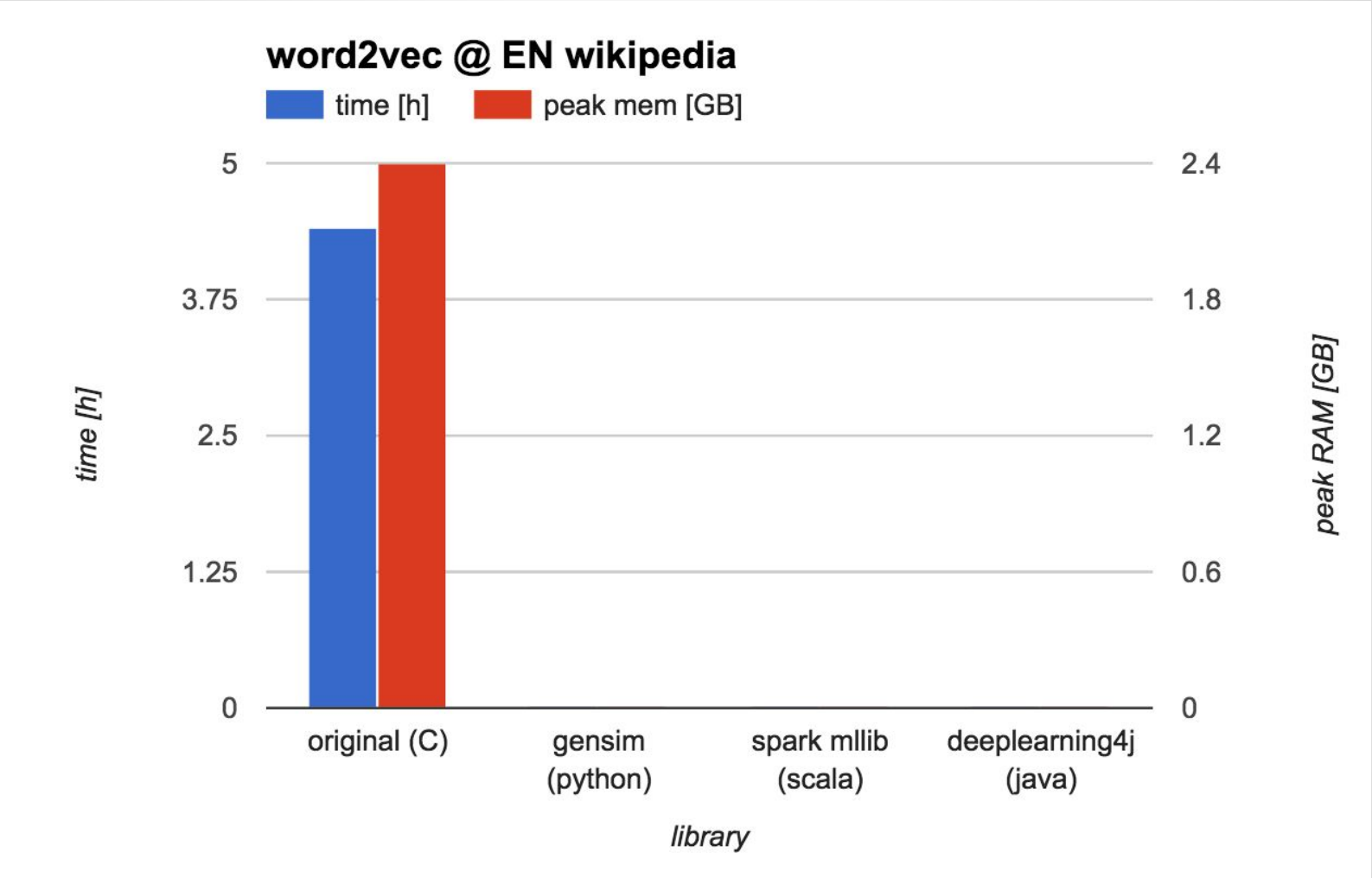
gensim: 1.0x / 1.75x / 2.85x

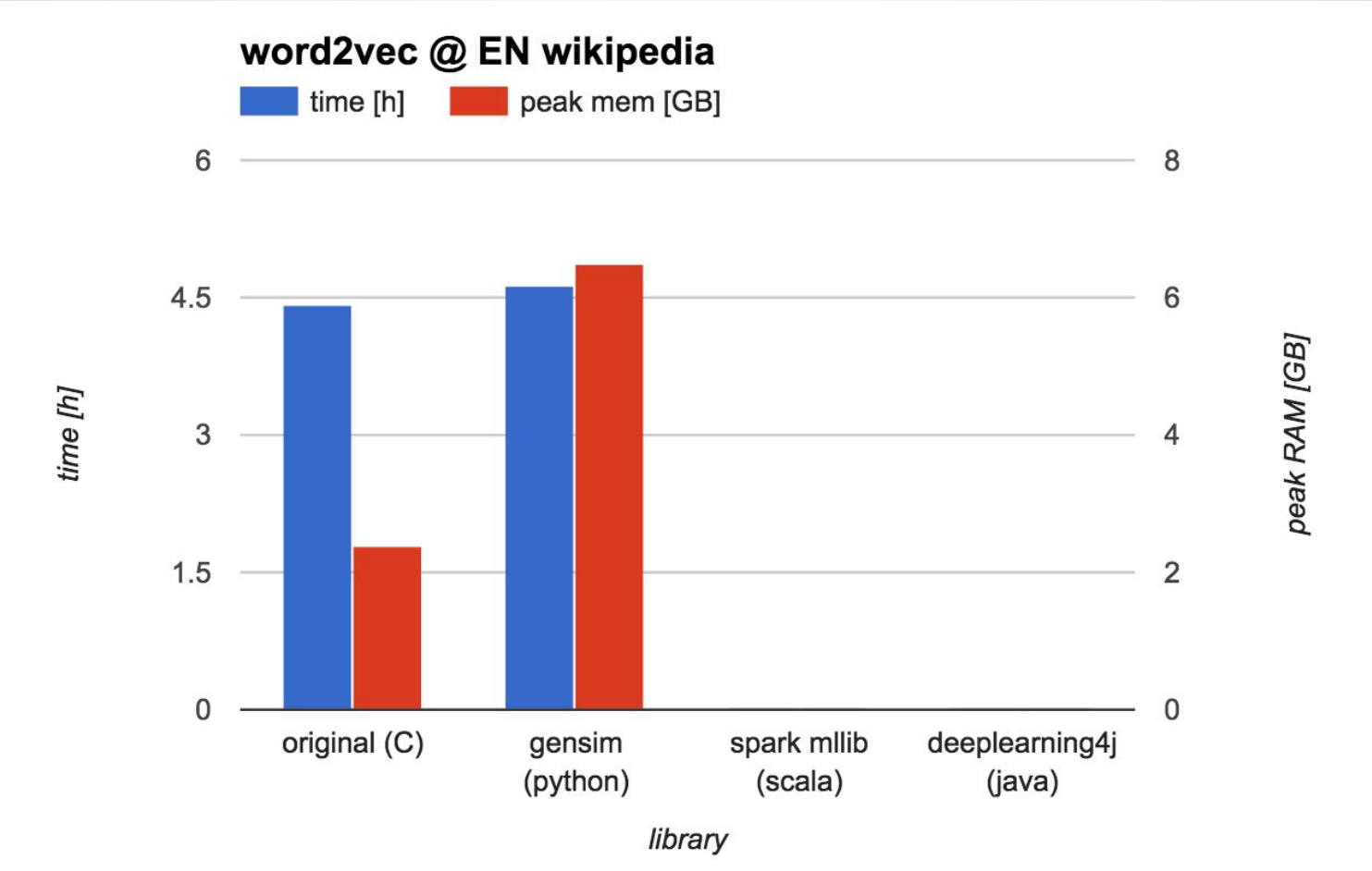
streaming (Python generator) for input

```
>>> for sentence in BrownCorpus('/Users/kofola/nltk_data/corpora/brown'):
...     print sentence
['the/at', 'fulton/np', 'county/nn', 'grand/jj', 'jury/nn', 'said/vb', 'fri
['the/at', 'jury/nn', 'further/rb', 'said/vb', 'in/in', 'term-end/nn', 'pre
['the/at', 'september-october/np', 'term/nn', 'jury/nn', 'had/hv', 'been/be
...
```

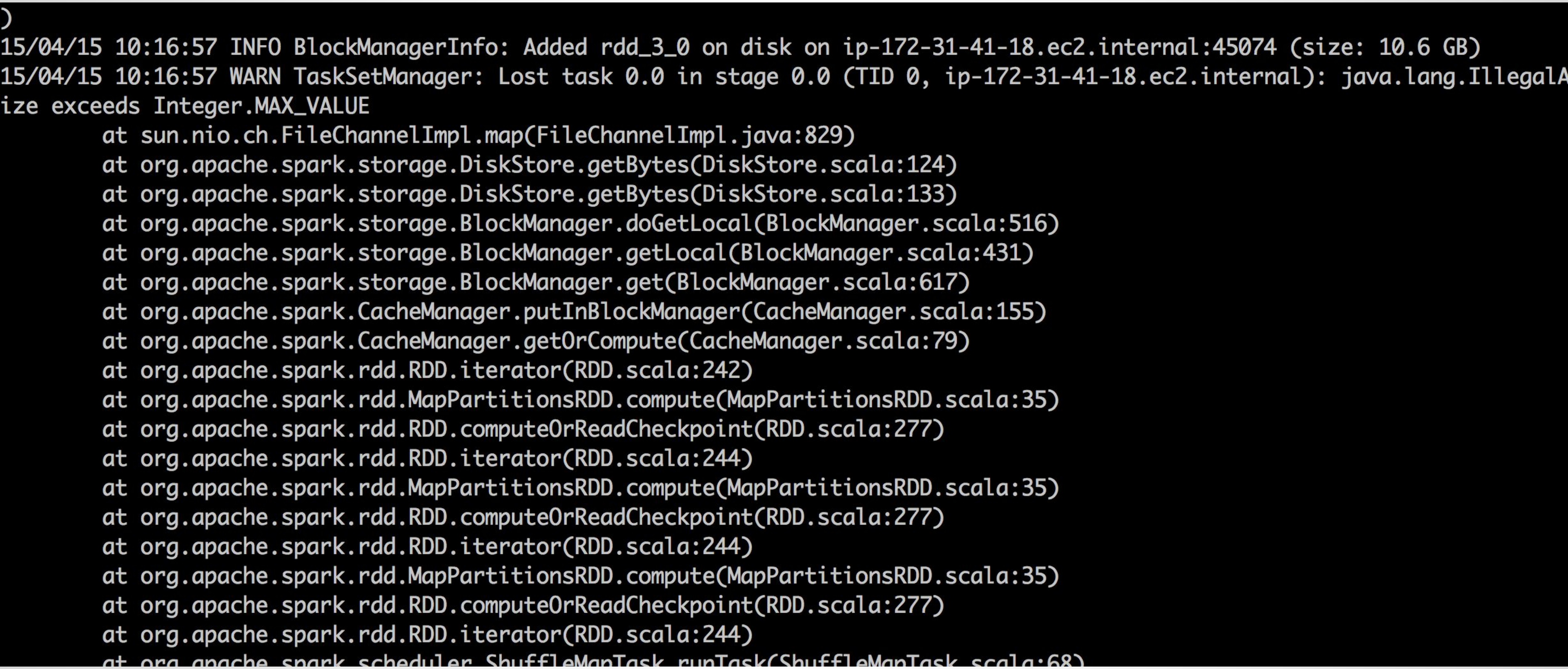
+ amazing Python ecosystem on either end!

```
>>> model.most_similar(positive=['woman', 'king'], negative=['man'], topn=1)
[('queen', 0.50882536)]
>>> model.doesnt_match("breakfast cereal dinner lunch".split())
'cereal'
>>> model.similarity('woman', 'man')
0.73723527
```


word2vec @ EN wikipedia
time [h]
peak mem [GB]
5
3.75
2.5
1.25
0
2.4
1.8
1.2
0.6
0
time [h]
peak RAM [GB]
original (C)
gensim (python)
spark mllib (scala)
deeplearning4j (java)
library


word2vec @ EN wikipedia
time [h]
peak mem [GB]
6
4.5
3
1.5
0
8
6
4
2
0
time [h]
peak RAM [GB]
original (C)
gensim (python)
spark mllib (scala)
deeplearning4j (java)
library

word2vec @ mllib



```
)
15/04/15 10:16:57 INFO BlockManagerInfo: Added rdd_3_0 on disk on ip-172-31-41-18.ec2.internal:45074 (size: 10.6 GB)
15/04/15 10:16:57 WARN TaskSetManager: Lost task 0.0 in stage 0.0 (TID 0, ip-172-31-41-18.ec2.internal): java.lang.IllegalA
ize exceeds Integer.MAX_VALUE
        at sun.nio.ch.FileChannelImpl.map(FileChannelImpl.java:829)
        at org.apache.spark.storage.DiskStore.getBytes(DiskStore.scala:124)
        at org.apache.spark.storage.DiskStore.getBytes(DiskStore.scala:133)
        at org.apache.spark.storage.BlockManager.doGetLocal(BlockManager.scala:516)
        at org.apache.spark.storage.BlockManager.getLocal(BlockManager.scala:431)
        at org.apache.spark.storage.BlockManager.get(BlockManager.scala:617)
        at org.apache.spark.CacheManager.putInBlockManager(CacheManager.scala:155)
        at org.apache.spark.CacheManager.getOrCompute(CacheManager.scala:79)
        at org.apache.spark.rdd.RDD.iterator(RDD.scala:242)
        at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:35)
        at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:277)
        at org.apache.spark.rdd.RDD.iterator(RDD.scala:244)
        at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:35)
        at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:277)
        at org.apache.spark.rdd.RDD.iterator(RDD.scala:244)
        at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:35)
        at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:277)
        at org.apache.spark.rdd.RDD.iterator(RDD.scala:244)
        at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scala:68)
```

Spark / SPARK-4846

When the vocabulary size is large, Word2Vec may yield "OutOfMemoryError: Requested array size exceeds VM limit"

Agile Board

Details

Type:	Bug	Status:	RESOLVED
Priority:	Minor	Resolution:	Fixed
Affects Version/s:	1.1.1, 1.2.0	Fix Version/s:	1.3.0
Component/s:	MLlib		
Labels:	None		
Environment:	Use Word2Vec to process a corpus(sized 3.5G) with one partition. The corpus contains about 300 million words and its vocabulary size is about 10 million.		
Target Version/s:	1.1.2, 1.2.1, 1.3.0		

Description

Exception in thread "Driver" java.lang.reflect.InvocationTargetException
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)

People

Assignee:	Joseph Tang
Reporter:	Joseph Tang
Votes:	0 Vote for this issue
Watchers:	5 Start watching this issue

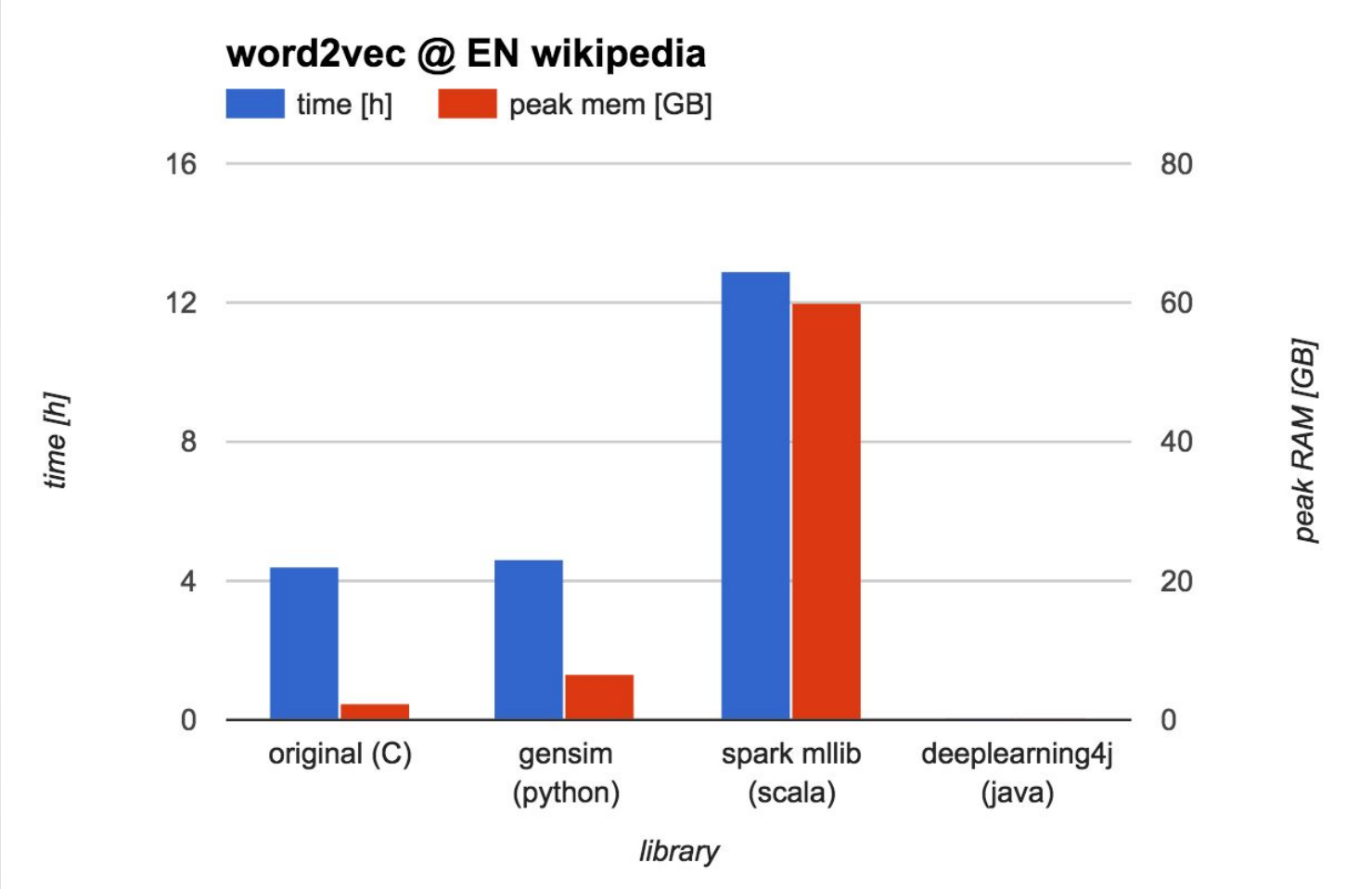
Dates

Created:	15/Dec/14 05:26
Updated:	04/Apr/15 16:47
Resolved:	30/Jan/15 18:07

Agile

View on Board

scaling down for Spark



=> if scaling linearly, Spark needs a cluster of ~12 machines to break even (vs. pySpark)

Deeplearning4j

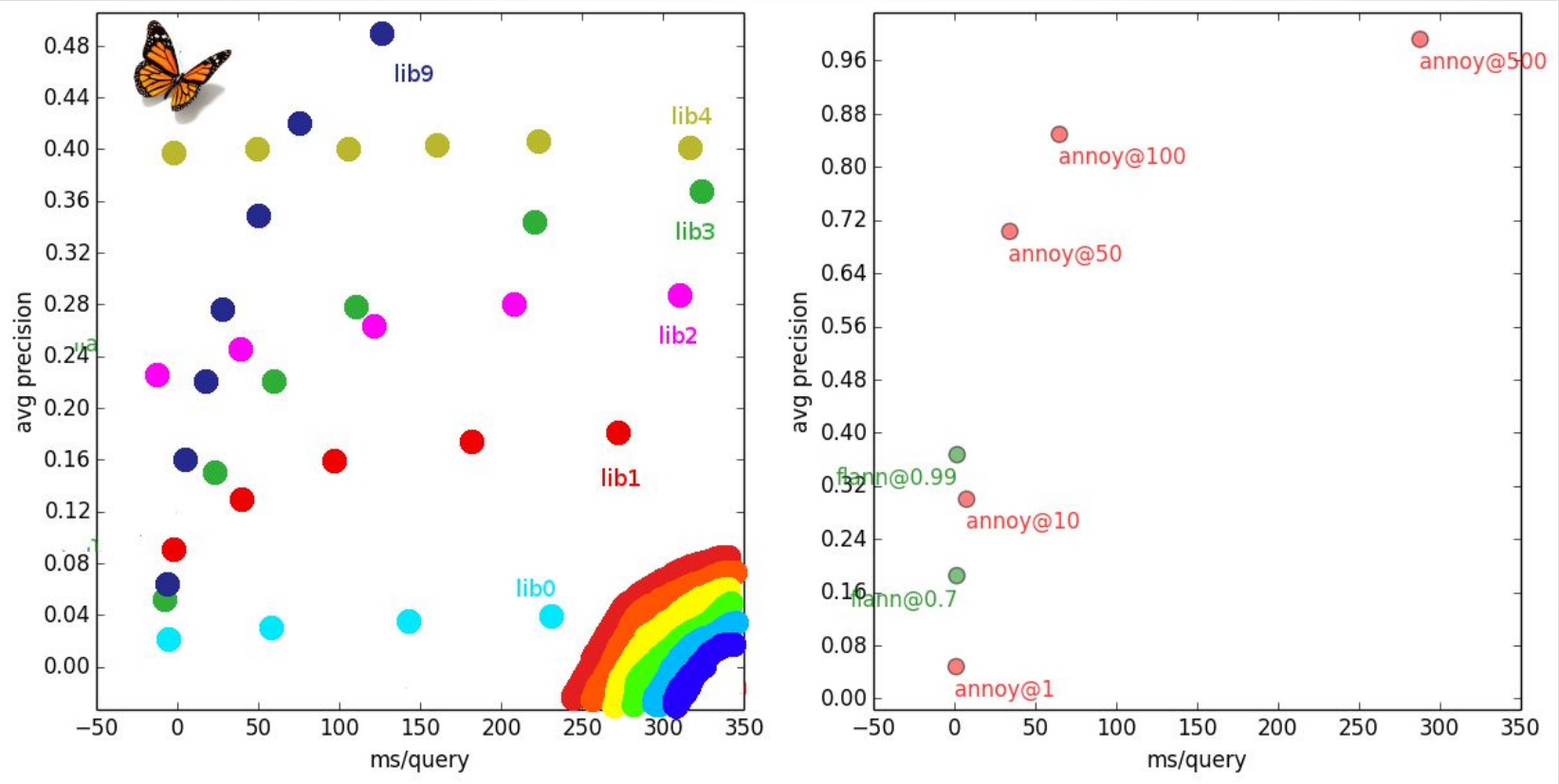
David Przybilla, Idio Ltd.

https://github.com/idio/wiki2vec

Word2Vec tools:

- Gensim
- DeepLearning4j: Feb 2014, Gets stuck in infinite loops on a big corpus
- Spark's word2vec: Feb 2014, `number of dimensions` * `vocabulary si`
  certain value otherwise an exception is thrown. issue

ANN libs @ Wikipedia



Tools

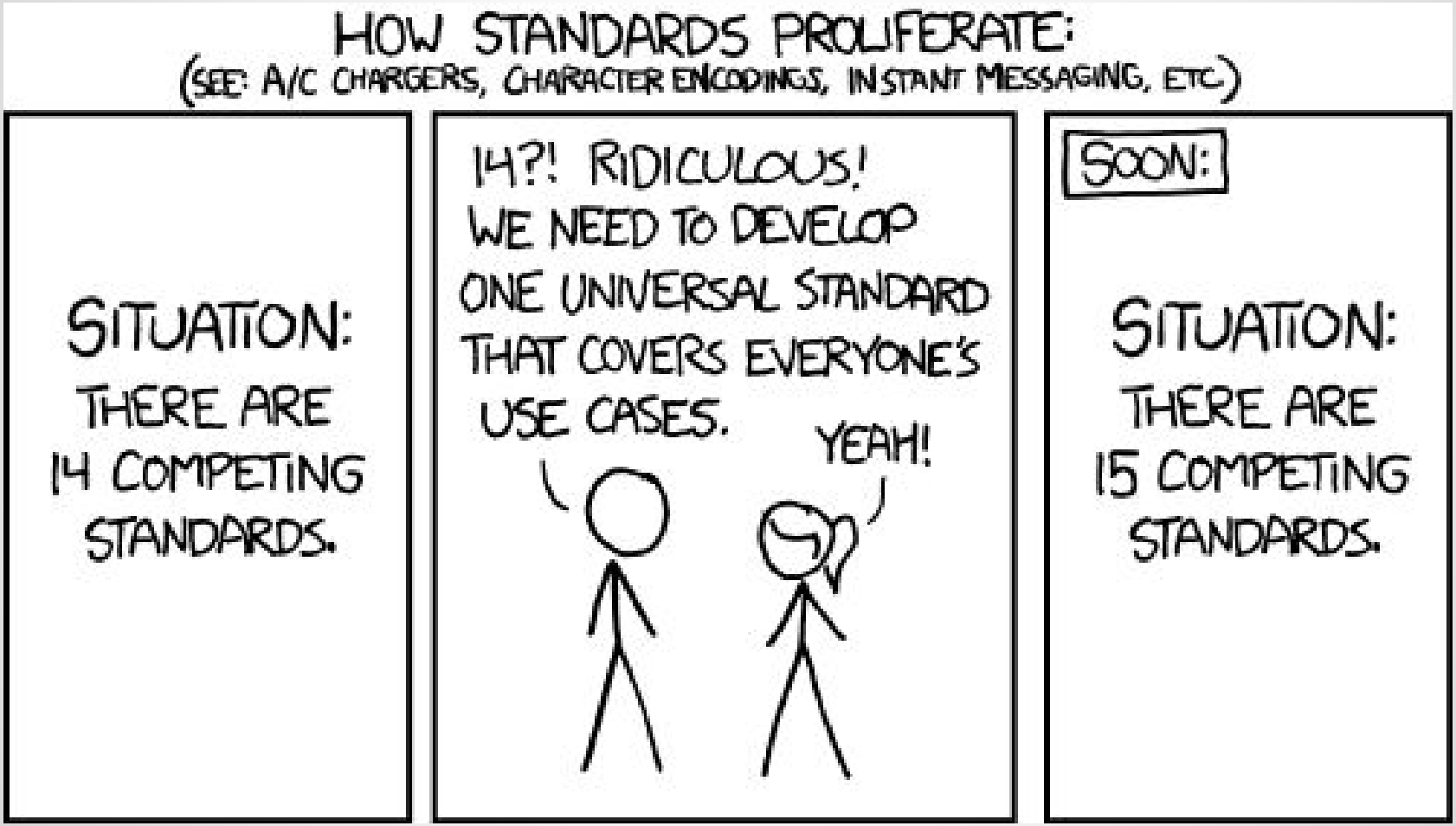
“Do one thing and do it well.”

Doug McIlroy

"Every program attempts to expand until it can read mail. Those programs which cannot so expand are replaced by ones which can."

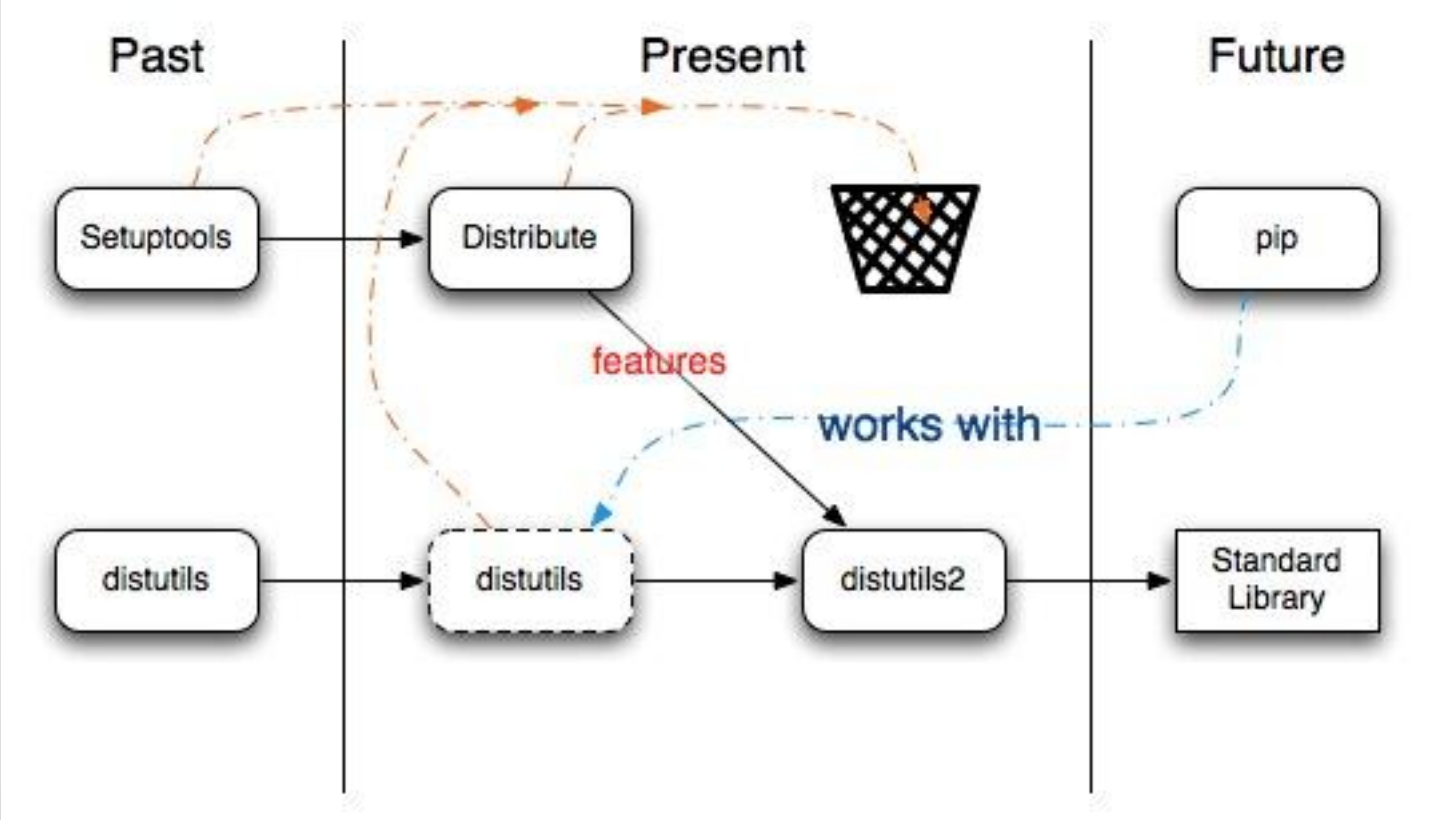
Zawinski’s law of software development

APIs & Abstractions



Configuration, setup, deployment

… the real work!



Python 2 vs Python 3

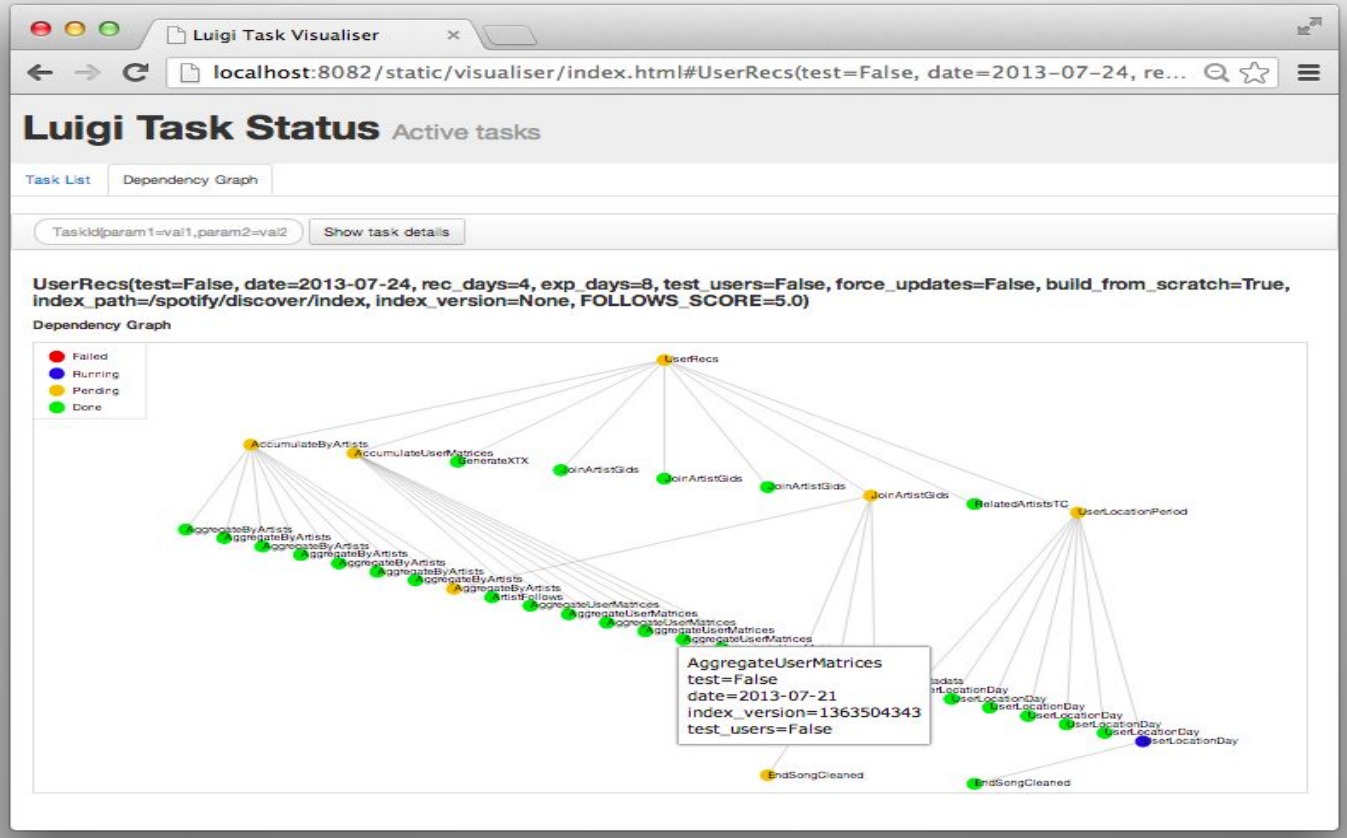
Java


```
Exception in thread main java.lang.NoClassDefFoundError: org/apache/hadoop/hbase/HBaseConfigurati
    at java.lang.Class.forName0(Native Method)
    at java.lang.Class.forName(Class.java:340)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:149)
```

```
java.lang.Exception: java.lang.OutOfMemoryError: Java heap space
at org.apache.hadoop.mapred.LocalJobRunner.runTasks(LocalJobRunner.java:462
at org.apache.hadoop.mapred.LocalJobRunner.run(LocalJobRunner.java:522)
```

Complex pipelines

- Python: Luigi (~Spotify), Pinball (~Pinterest)
- Java: Apache Oozie, Azkaban...



Logging

- UI, job trackers
- tracebacks, continuous
- configurable
- human readable

Navigating the tool landscape

Let it go; if it’s meant to be, it will come back.

Take “progress” easy

mat kelcey @mat_kelcey · May 20

pretty much every paper i've ever read....

method	score
previous approach	good
our approach	almost as good
our approach + last minute hack	slighty better than good!

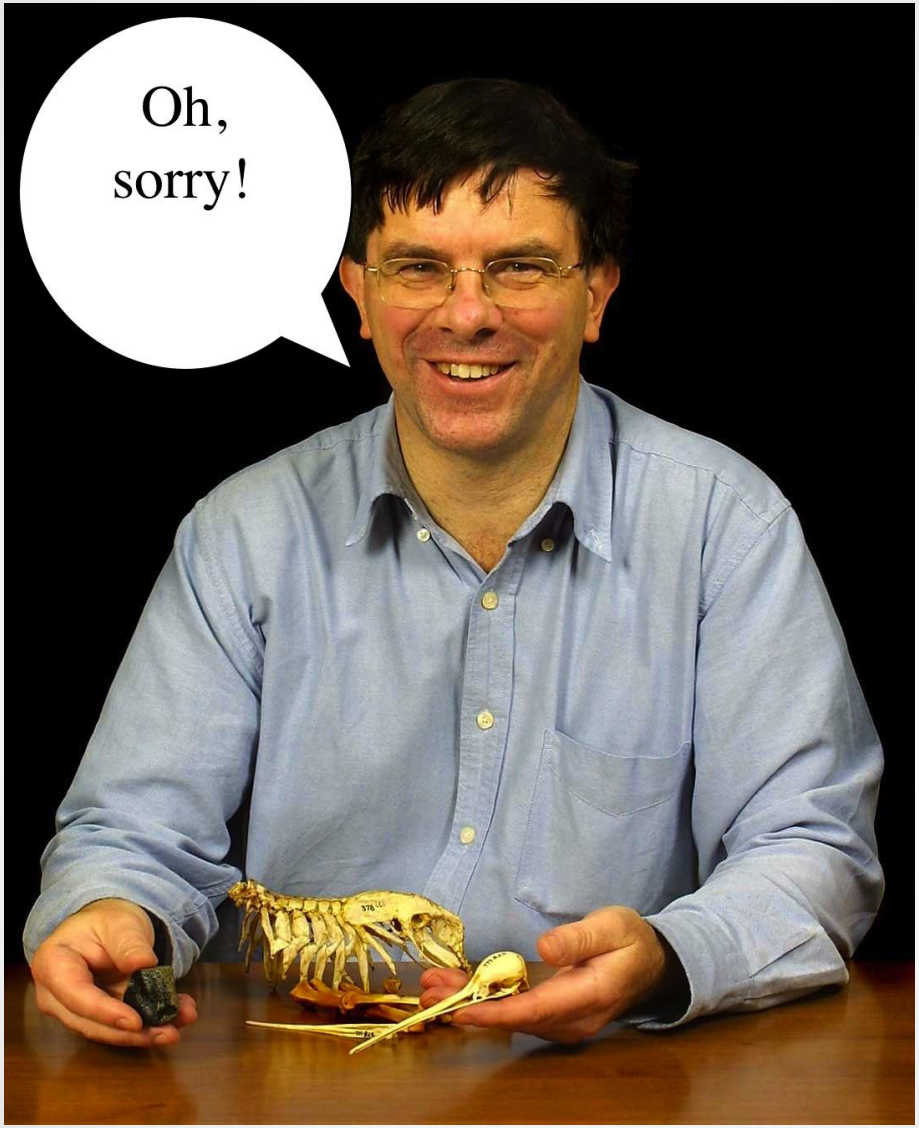
91 82

View photo

Summary

Python’s greatest differentiating factors:

- +experienced full stack engineers
- +pragmatic, mature tools
- +HPC & scientific “baggage”
- -meh deployment, orchestration, packaging
- -not as much enterprise “baggage”


Oh,
sorry!

Radim Řehůřek
http://rare-technologies.com
(formerly radimrehurek.com)
@radimrehurek

Blog (mostly tech): http://radimrehurek.com/blog/

RARE
TECHNOLOGIES