Thoth

How to recommend the best possible packages for your application

Fridolin Pokorny <fridolin@redhat.com>
2019-May-4

$ whoami

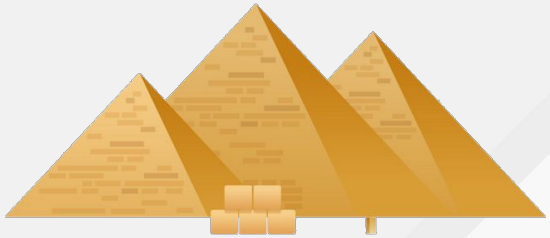
https://fridex.github.io

- Fridolín “fridex” Pokorný
- Senior Software Engineer at Red Hat
- Distributed systems, AI/ML and (of course) Python fan
- Projects:
  - Reverse engineer RetDec (AVG)
  - Linux kernel TLS/DTLS module AF_KTLS
  - Selinon - distributed task flows scheduler on top of Celery
  - Project Thoth

$ whoarewe # project Thoth

https://thoth-station.ninja/

- Project Thoth - https://github.com/thoth-station
- Red Hat - Office of the CTO
  - Emerging technologies
  - AI team - https://github.com/aicoe
- Initially 2 engineers, now growing
  - Christoph Görn <goern@redhat.com>
  - Francesco Murdaca <fmurdaca@redhat.com>
  - Fridolín Pokorný <fridolin@redhat.com>
  - Harshad Reddy Nalla <hnalla@redhat.com>
  - Marek Cermak <macermak@redhat.com>
  - Subin Modeel <smodeel@redhat.com>



What is Thoth?

Why Thoth?

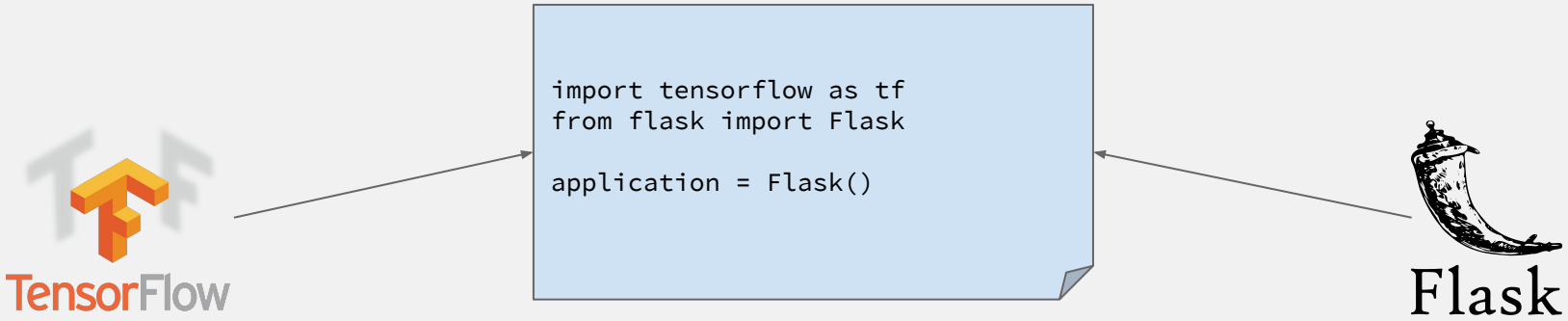
Why Thoth?

- PyPI - Python Package Index
  - https://pypi.org/
  - 178,016 projects
  - 1,303,926 releases (approx. 7 releases per project)

Why Thoth?

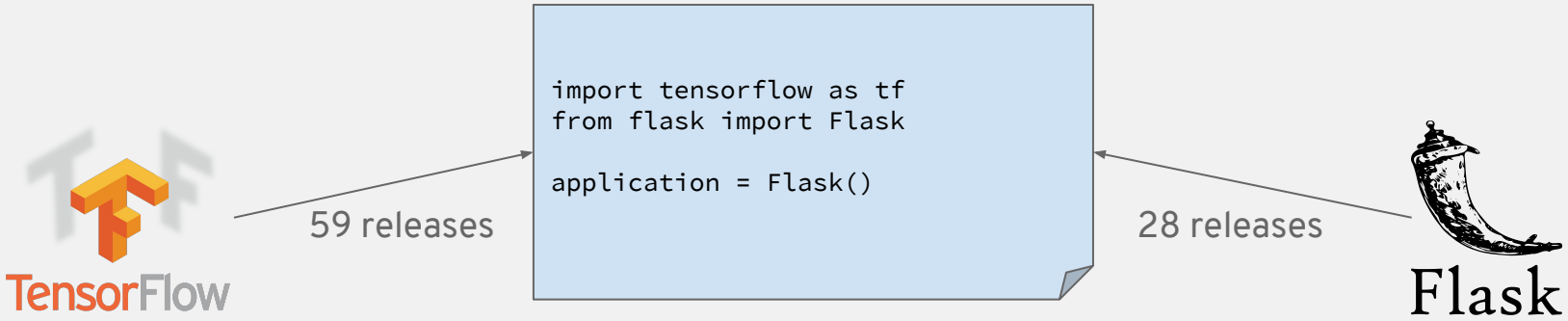
```
import tensorflow as tf
from flask import Flask

application = Flask()
```

Why Thoth?

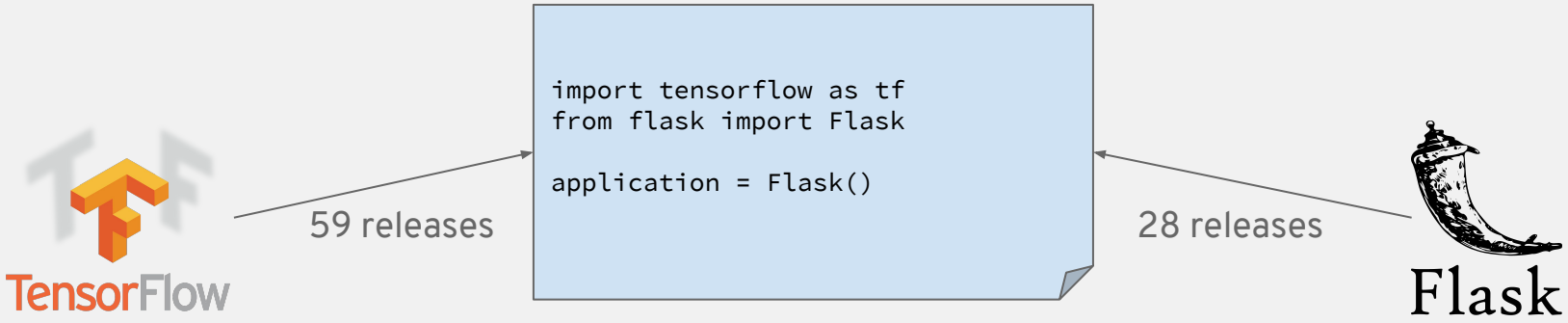


```
$ pip3 install --user tensorflow

$ pip3 install --user flask

$ python3 ./app.py

Error: tensorflow 1.10.1 has requirement
numpy<=1.14.5,>=1.13.3, but you'll have numpy 1.15.1 which
is incompatible.

$
```

Why Thoth?



Why Thoth?



All combinations how to install libraries directly used:

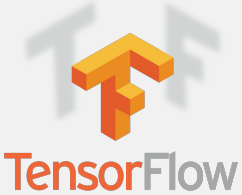
59 * 28 = 1,652

Transitive dependencies

- Flask
  - click, itsdangerous, jinja2, markupsafe, werkzeug

Estimatimated number of combinations: 54,395,000

Transitive dependencies

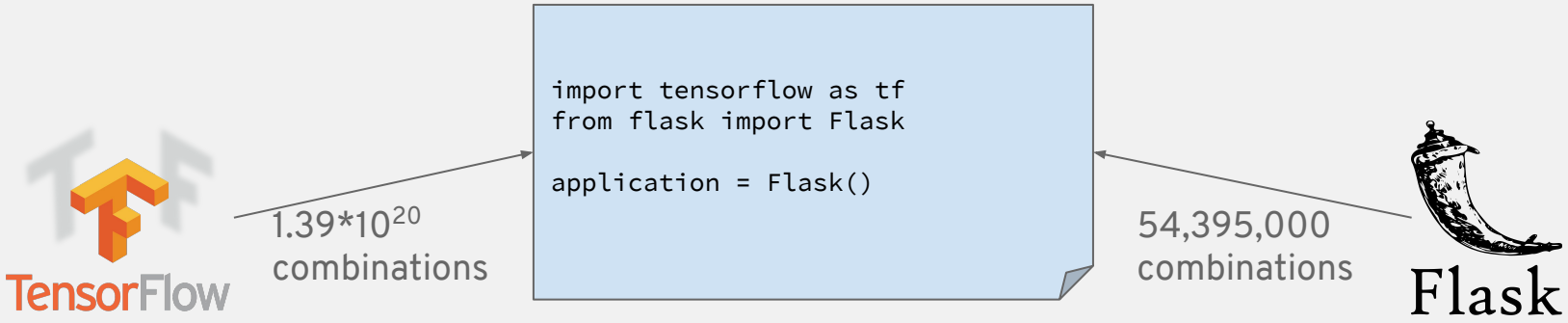


- TensorFlow
  - absl-py, astor, backports-weakref, bleach, enum34, gast, google-pasta, grpcio, h5py, html5lib, keras, keras-applications, keras-preprocessing, markdown, mock, numpy, pbr, protobuf, pyyaml, scipy, setuptools, six, tensorboard, tensorflow-estimator, tensorflow-tensorboard, termcolor, tf-estimator-nightly, werkzeug, wheel

Estimated number of combinations: 139,740,802,927,165,440,000

approx. $1.39*10^{20}$

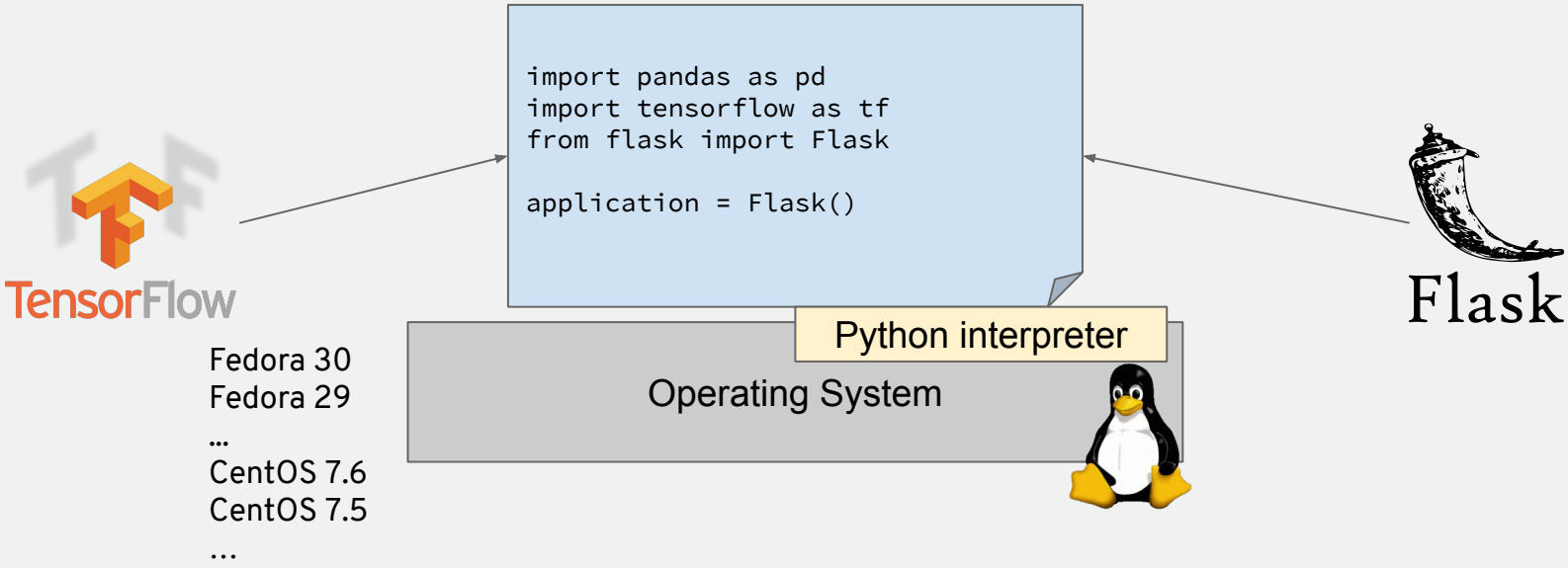
Why Thoth?



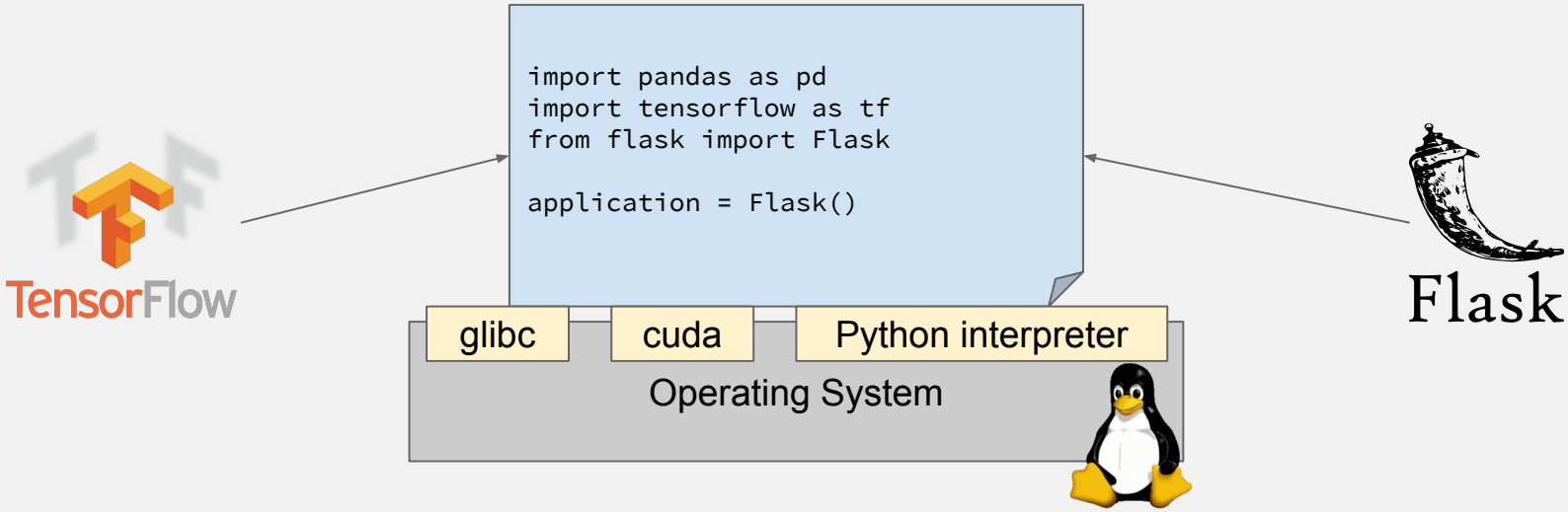
All combinations how to install application stack of libraries directly and indirectly used (estimation):

$$1.39*10^{20} * 54{,}395{,}000 = 7.6*10^{27}$$

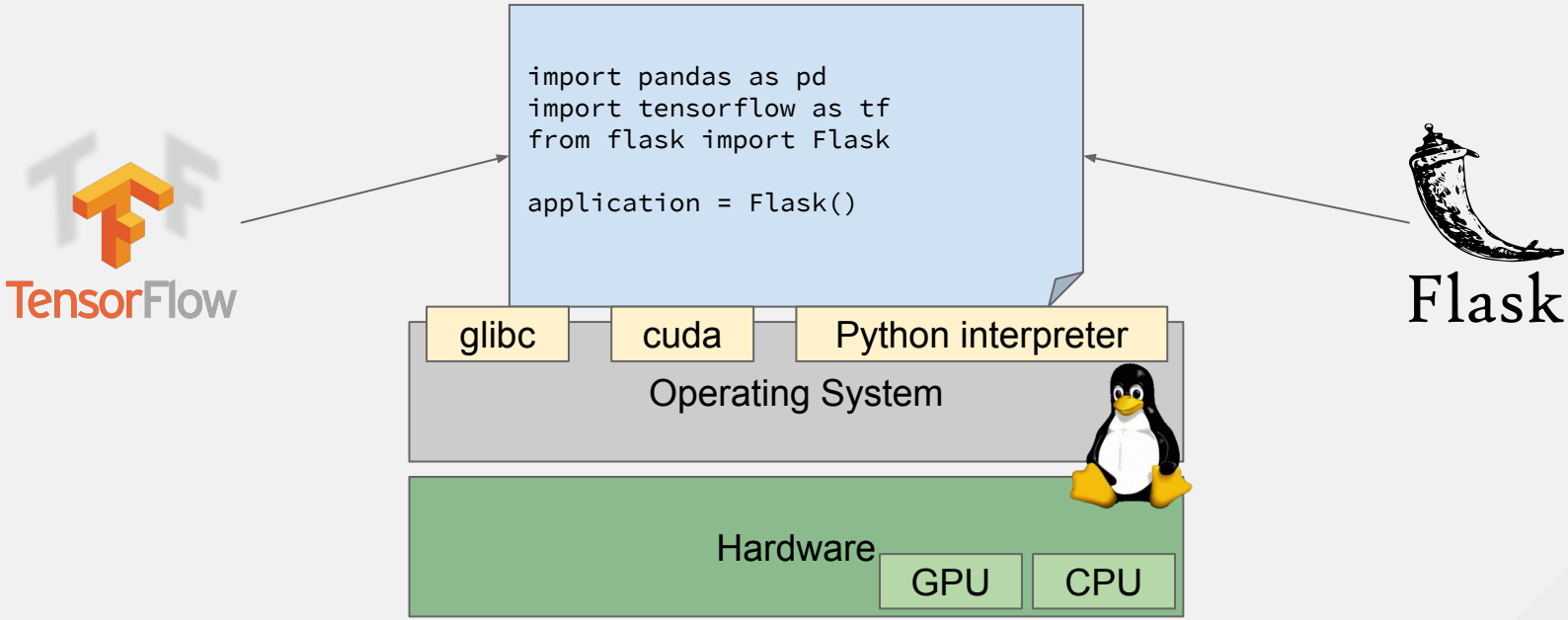
Why Thoth?



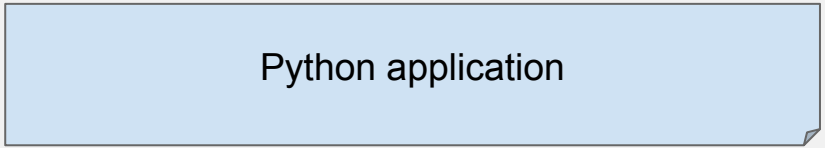
Why Thoth?



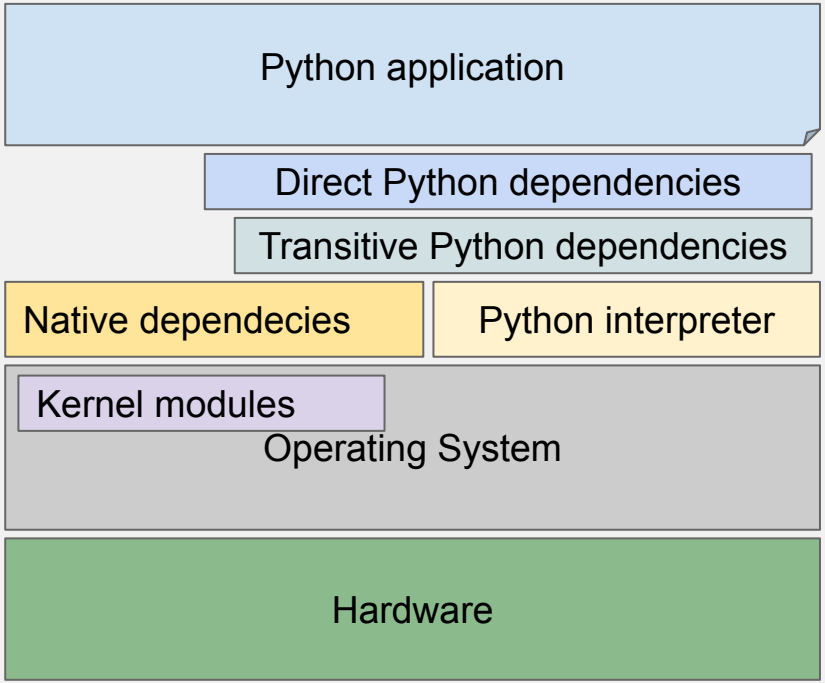
Why Thoth?



Why Thoth?



Why Thoth?



Why Thoth?

- Create knowledge base
  - What packages in which versions should I use?
    - Application builds correctly
    - Application runs correctly
    - Application behaves and performs well
- Create an advanced Python resolver which uses knowledge base to resolve software stacks

Latest versions are not always greatest choices.

Building Thoth’s knowledge base

Gathering data for Thoth's knowledge base

- Resolving software stacks
  - own resolution algorithm
- Analyses of container images
  - JupyterHub images
  - Thoth's container images
- Amun, Dependency Monkey
  - running CI and perf related tests
  - performance related analyses
- ...

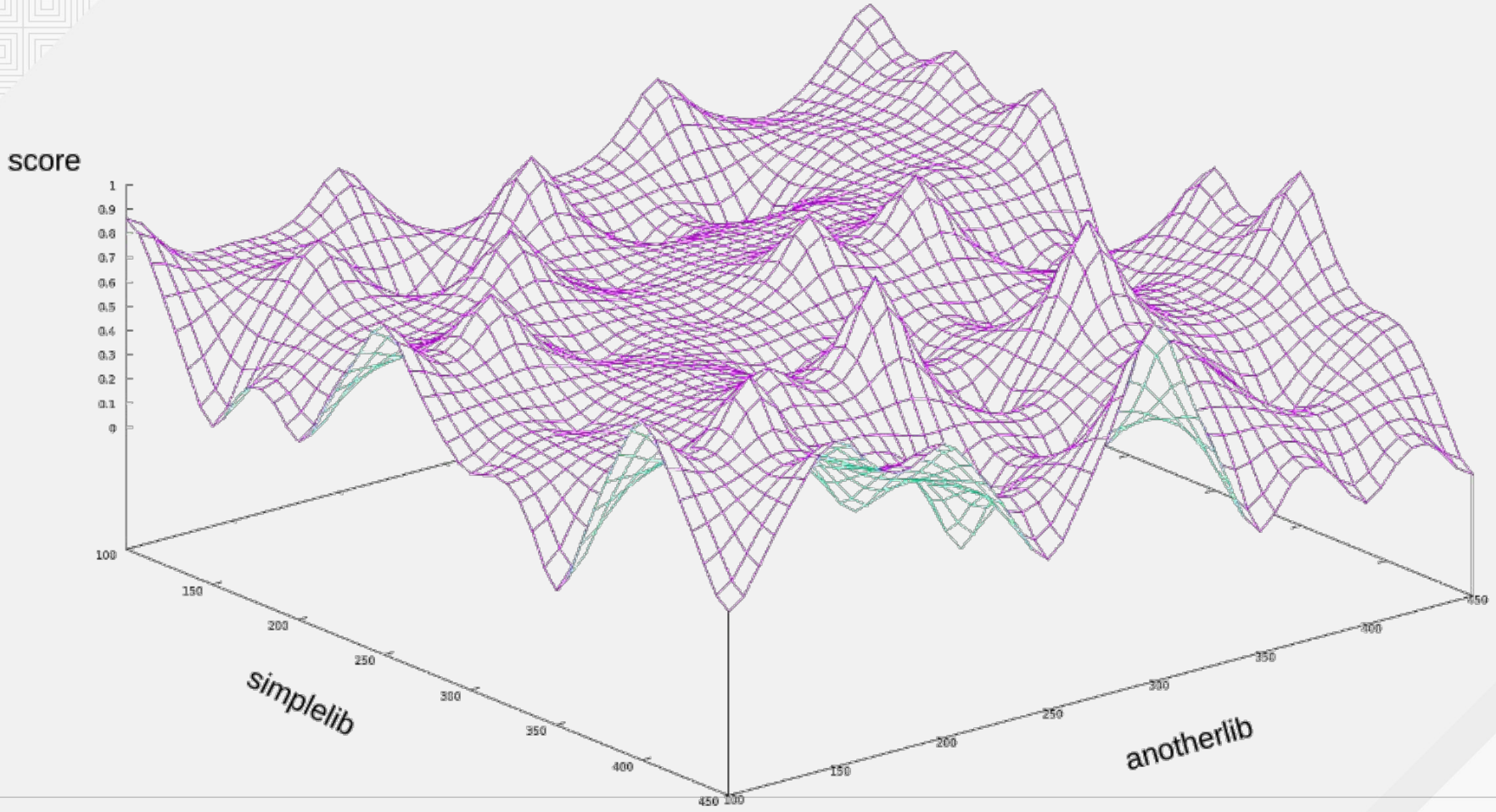
Optimized TensorFlow builds by Thoth team

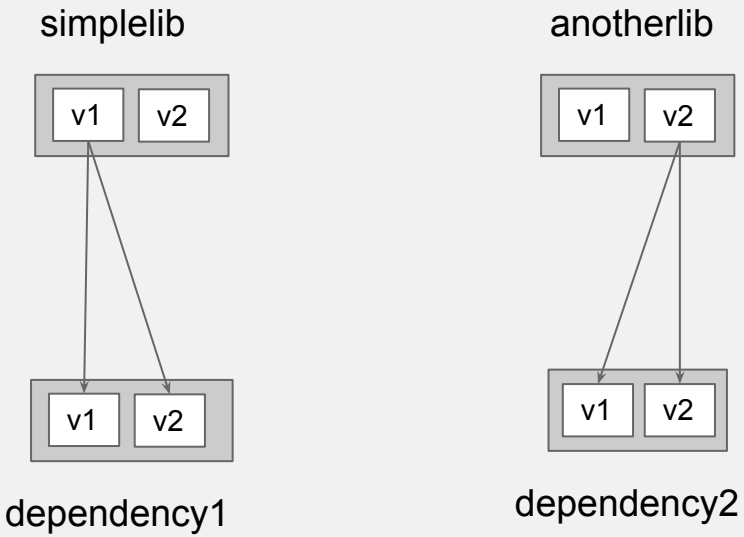
- Automated tests of libraries
- Tests targeting performance
- Optimized TensorFlow builds

https://tensorflow.pypi.thoth-station.ninja/

Recommendations

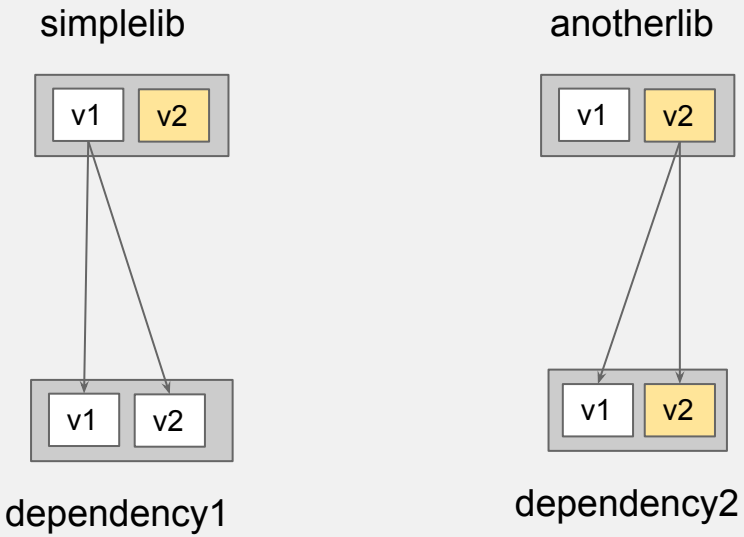
How good is my software stack?

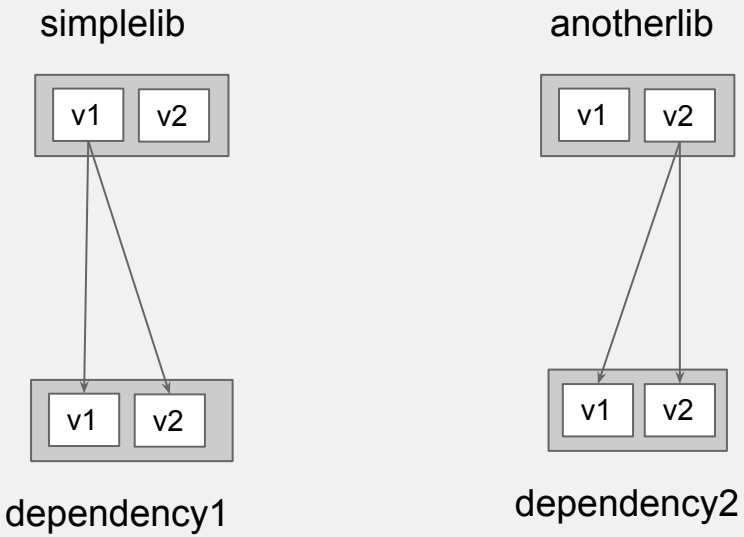

score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
100
150
200
250
300
350
400
450
simplelib
100
150
200
250
300
350
400
450
anotherlib

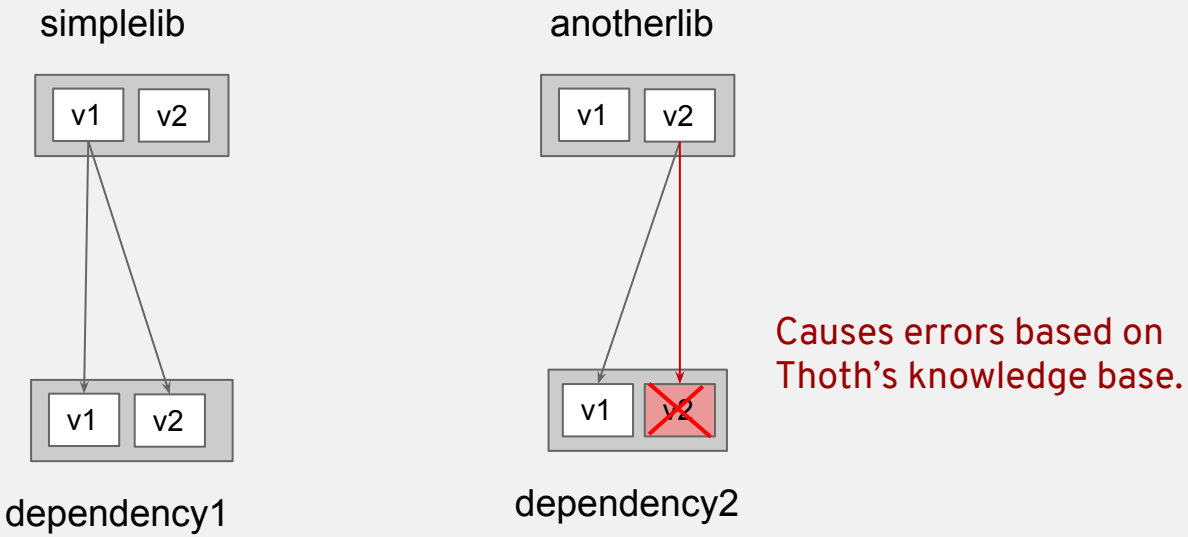

simplelib
v1
v2
v1
v2
dependency1
anotherlib
v1
v2
v1
v2
dependency2

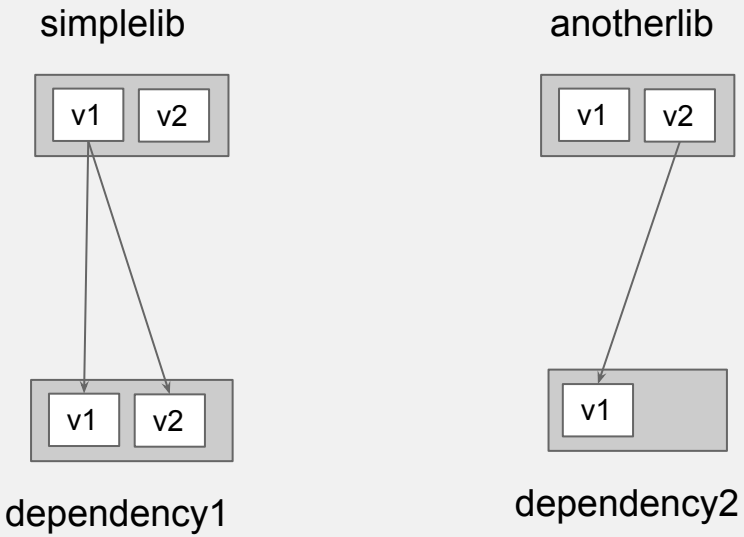
pip/Pipenv (always latest):

```
simplelib     ==v2
anotherlib    ==v2
dependency2   ==v2
```

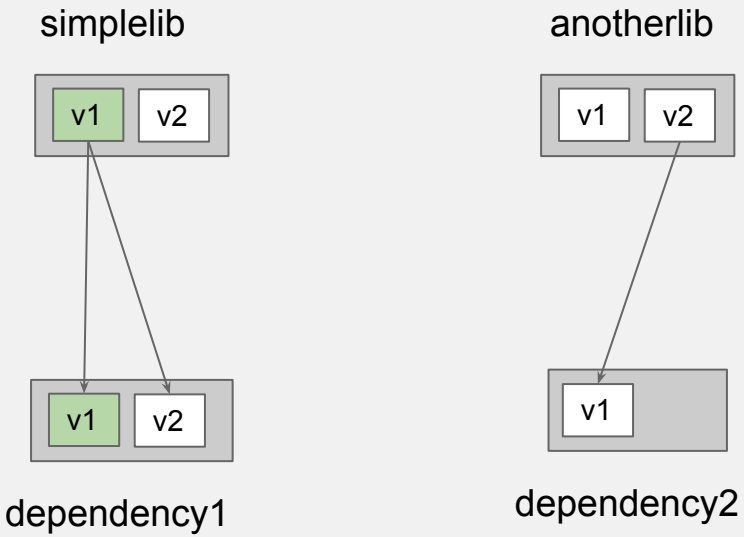


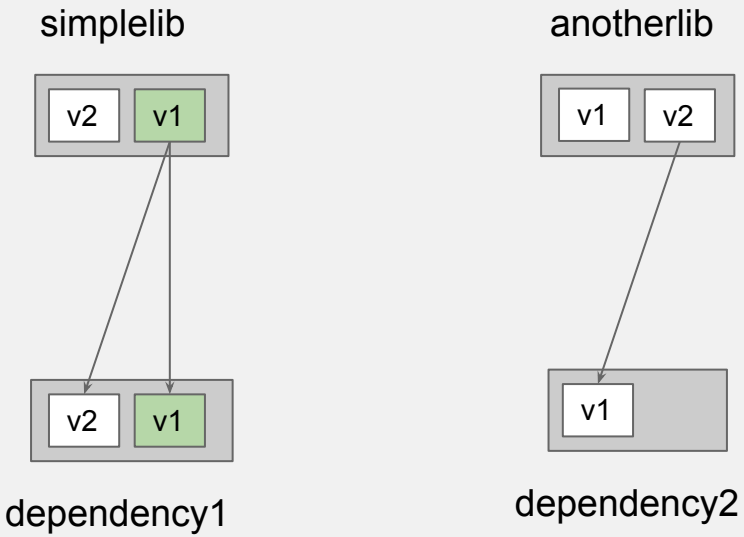

simplelib
v1
v2
v1
v2
dependency1
anotherlib
v1
v2
v1
v2
dependency2


simplelib
v1
v2
v1
v2
dependency1
anotherlib
v1
v2
v1
v2
dependency2
Causes errors based on
Thoth's knowledge base.


simplelib
v1
v2
v1
v2
dependency1
anotherlib
v1
v2
v1
dependency2

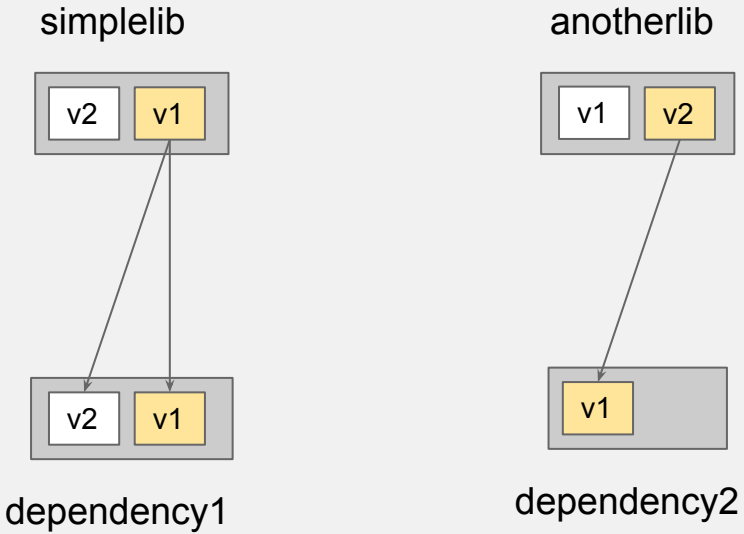
Simplelib in version v1 performs better together with dependency1 in version v1 based on Thoth's knowledge base.



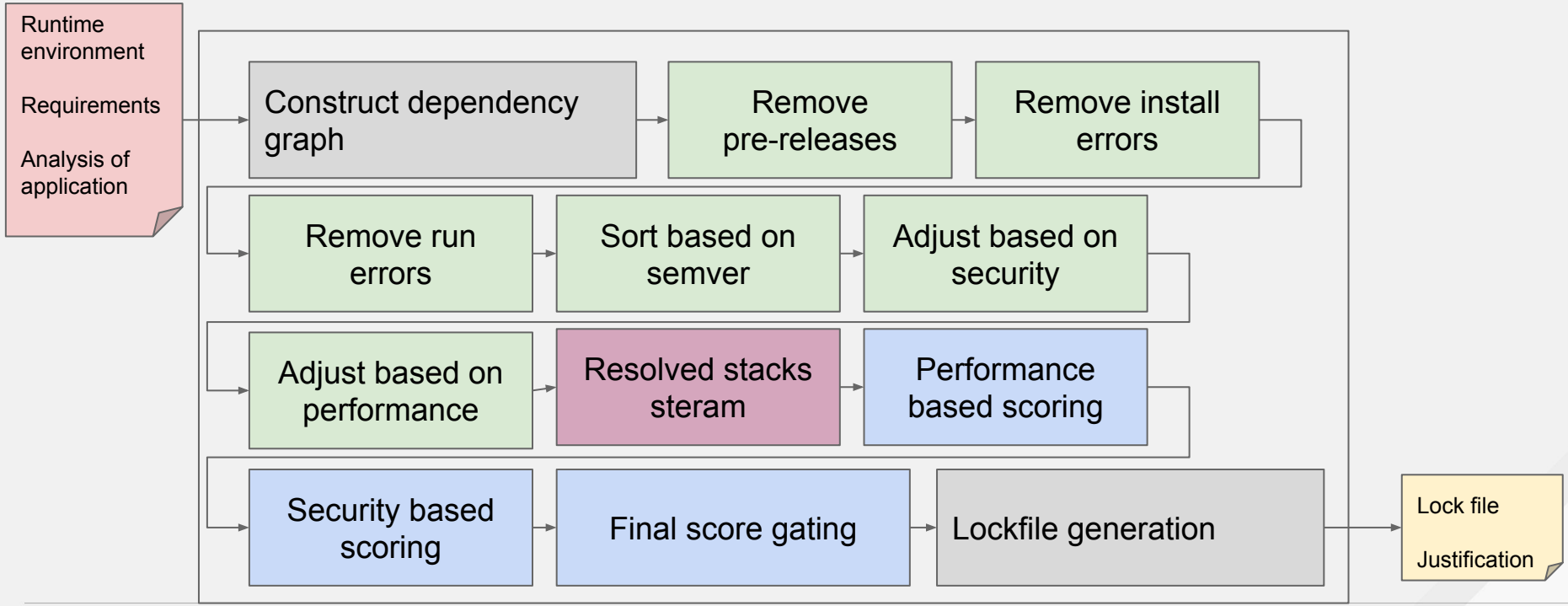

simplelib
v2
v1
v2
v1
dependency1
anotherlib
v1
v2
v1
dependency2

Thoth (always greatest):

```
simplelib     ==v1
anotherlib    ==v2
dependency1   ==v1
dependency2   ==v1
```



Stack generation pipeline



Extending information about Python packages

Python package metadata

- License
- Classifiers
  - `Programming Language :: Python :: 3.6`
  - `Programming Language :: Python :: Implementation :: CPython`
- Package purpose
  - machine learning library
  - plugin
  - ...
- Is the given package affecting performance?

project2vec

- A vector space model
- Each vector in vector space corresponds to a project
- Each item in vector represents a feature
- Allows feature based queries and similar projects search

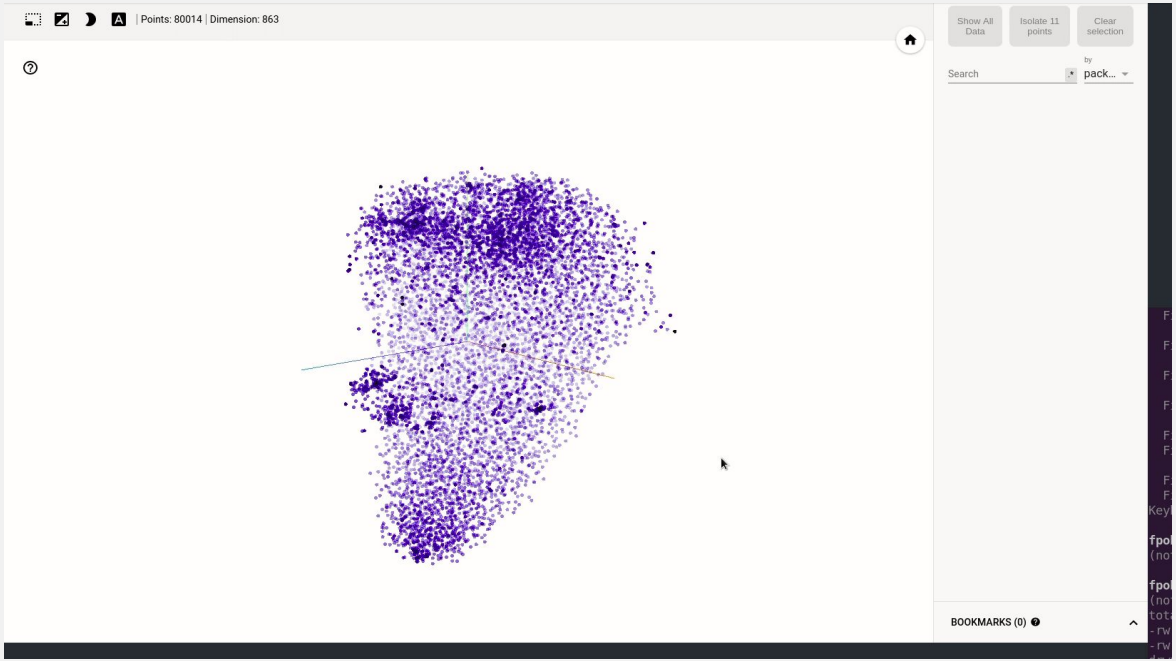
$$F = \{python, machine\text{-}learning, web, django\text{-}framework, webassembly, sql, spark, gpu\text{-}support, Java\}$$

$$F_{tf} = \{1, 1, 0, 0, 0, 0, 0, 1, 0\}$$

F - feature vector

F_{tf} *- feature vector for project TensorFlow*

project2vec



[Image address]

Information about Thoth

- Website:
  - https://thoth-station.ninja/
- Twitter
  - https://twitter.com/thothstation
    - Follow for updates on public availability
- GitHub
  - https://github.com/thoth-station

Features | API (internal Red Hat only now) | Developers | Status

Project Thoth

Using Artificial Intelligence to analyse and recommend Software Stacks for Artificial Intelligence applications.

Get started

It’s all on you

- Community sitting at https://github.com/thoth-station/
- Bot Kebechet
  - https://thoth-station.ninja/kebechet/
- Twitter
  - https://twitter.com/thothstation

```
$ pip3 install thamos
$ cd ~/repositories/my-repo/
$ thamos config
$ thamos advise
```

redhat.

THANK YOU

plus.google.com/+RedHat

facebook.com/redhatinc

linkedin.com/company/red-hat

twitter.com/RedHat

youtube.com/user/RedHatVideos