LaMachine

a meta-distribution for NLP software

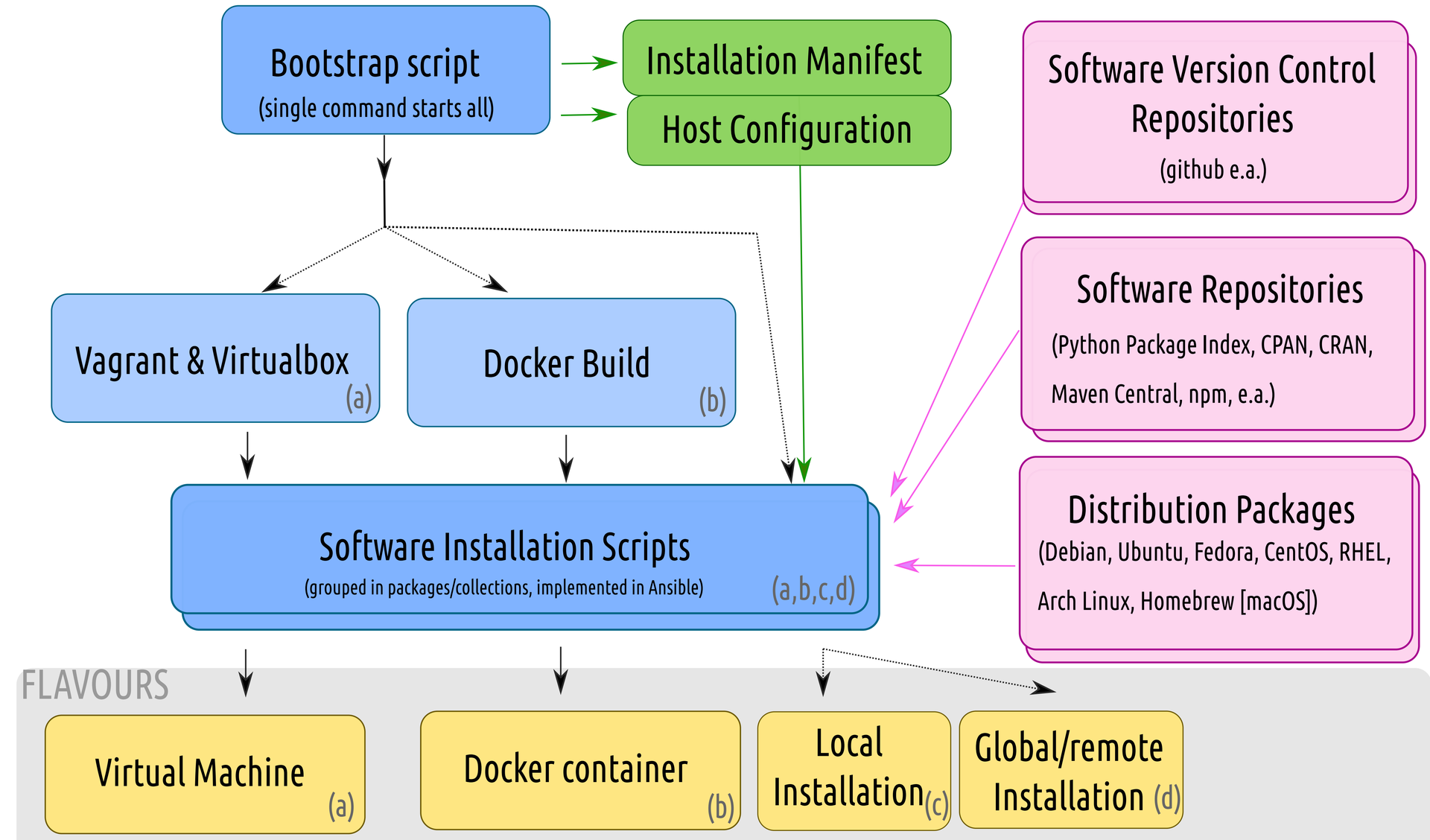
Maarten van Gompel & Iris Hendrickx
Centre for Language and Speech Technology
Radboud University, Nijmegen, the Netherlands

LaMachine is a unified Natural Language Processing (NLP) open-source software distribution to facilitate the installation and deployment of a large amount of software projects, many of which were developed in the scope of CLARIN-NL/CLARIAH. LaMachine provides a kind of Virtual Laboratory for data scientists.

https://proycon.github.io/LaMachine

```
$ bash <(curl -s https://raw.githubusercontent.com/proycon/LaMachine/master/bootstrap.sh)
```

Architecture



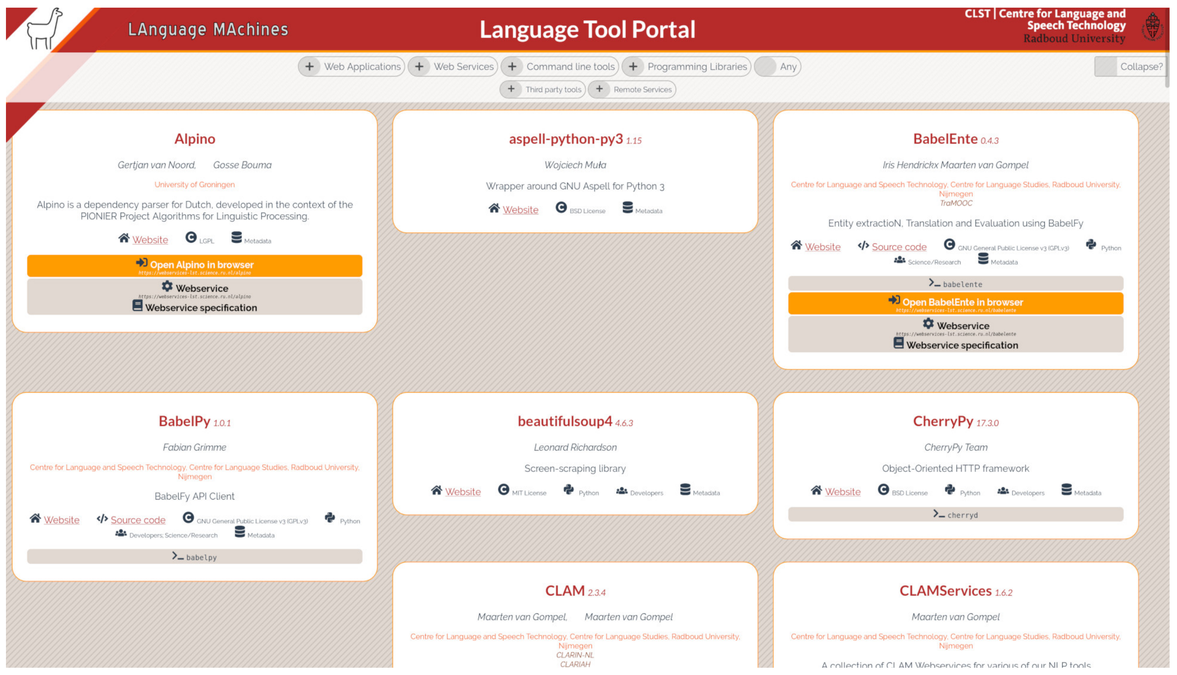
Software Metadata & Reproducibility

- LaMachine collects and converts software metadata from various sources to codemeta
  https://codemeta.github.io
- Software metadata is made accessible to users and software.
- Allow installation of specific custom versions; provides limited reproducibility.

Included Software

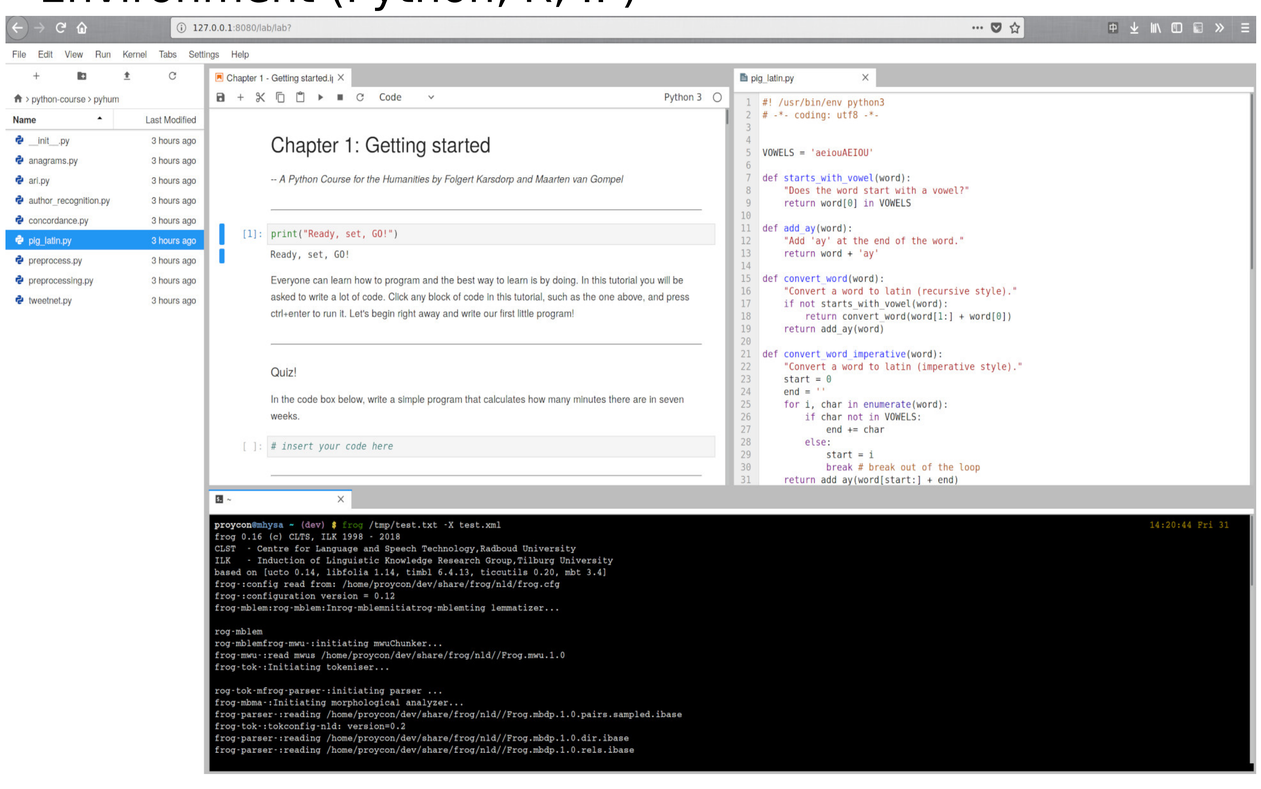
(exact software selection is always user configurable!)

- Major scientific python software (*scikit-learn, scipy, keras, pytorch, etc...*)
- Software developed in/around CLARIN/CLARIAH: *Frog, FLAT, ucto, CLAM, FoLiA, Alpino, PICCL, ...*
- Third party NLP software: Core*NLP, spacy, nltk, kaldi, ...*
- Includes "*Python Course for the Humanities*"



Interfaces

- Command-line terminal (e.g. ssh)
- Portal website on built-in webserver to access webservices and webapplications.
- Jupyter Labs: Notebooks, web-based terminal, scripting environment and Integrated Development Environment (Python, R, ..)



Case Study: Text Mining

LaMachine provided the stand-alone research environment for text mining experiments on highly sensitive data in the Dutch project: "*Text mining for Inspection: an exploratory study on automatic analysis of health care complains*". We experimented with grouping complaints, discovering topics and patterns and aimed to predict the severity of the complaints.

LaMachine offered:

- Use as a self-contained VM on an isolated Windows machine without internet access.
- Data sharing to handle access to sensitive data.
- Rich availability of linguistic preprocessing tools, text mining software and Python environment.
- Access via command line and web-based Python notebooks.