



# VL Deep Learning for Natural Language Processing

---

## 2. DL Programming

*Prof. Dr. Ralf Krestel*  
*AG Information Profiling and Retrieval*

# Lerning Goals for this Chapter

---

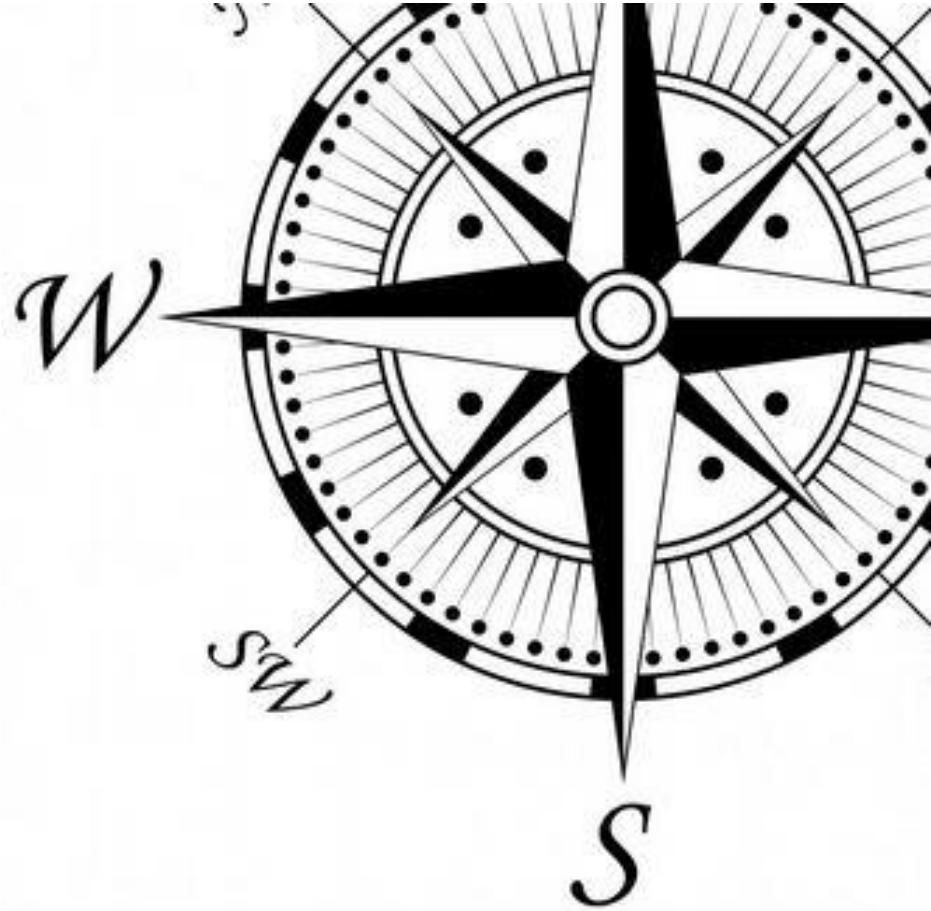


- Know about deep learning frameworks
- Have an idea what Keras is
- Know how Colab works
- Be able to use Jupyter Notebooks

# Topics Today

---

1. **Frameworks & Libraries**
2. Keras
3. Jupyter Notebook & Google Colab



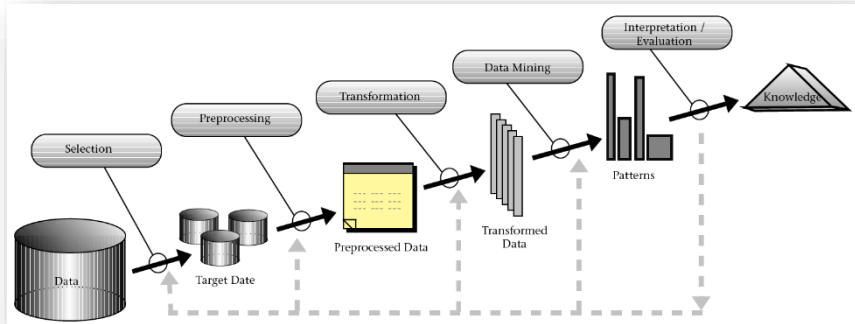
# Motivation



U. Fayyad, G. P.-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. AI Magazine, 17(3):37-54, Fall 1996.

- Nobody wants to reinvent the wheel everytime from scratch
  - Libraries
  - Frameworks

What's the difference?



- Machine learning has many standardized procedures and methods/algorithms
  - Data preparation
  - Data splitting (train, val, test)
  - Training
  - Evaluation/Inference
  - Visualization
  - ...
- Decision trees
  - Support vector machines
  - Naive Bayes
  - Linear regression
  - Conditional random fields
  - ...

# Important Python Libraries for DM/ML

---




- Pandas
    - <https://pandas.pydata.org/>
    - Data analysis library for python, providing high-performance, easy-to-use data structures and data analysis tools.
  - Numpy
    - <https://numpy.org/>
    - NumPy is the fundamental package for scientific computing with Python. Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined.
  - Scikit-learn:
    - <https://scikit-learn.org/>
    - Machine learning in python (built on numpy, scipy, matplotlib): preprocessing, classification, regression, dim reduction, clustering, model selection
-

# Important Frameworks for Deep Learning



- Overview of deep learning software
  - [https://en.wikipedia.org/wiki/Comparison\\_of\\_deep\\_learning\\_software](https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software)
- Popular frameworks (majority written in C++ with Python interface)
  - TensorFlow (Google)
  - Theano (U Montreal)
  - PlaidML (Intel)
  - CNTK (Microsoft)
  - MXNet (Apache)
  - Torch
  - Deeplearning4j
  - Caffe (Berkeley)
  - Matlab+DL (MathWorks)



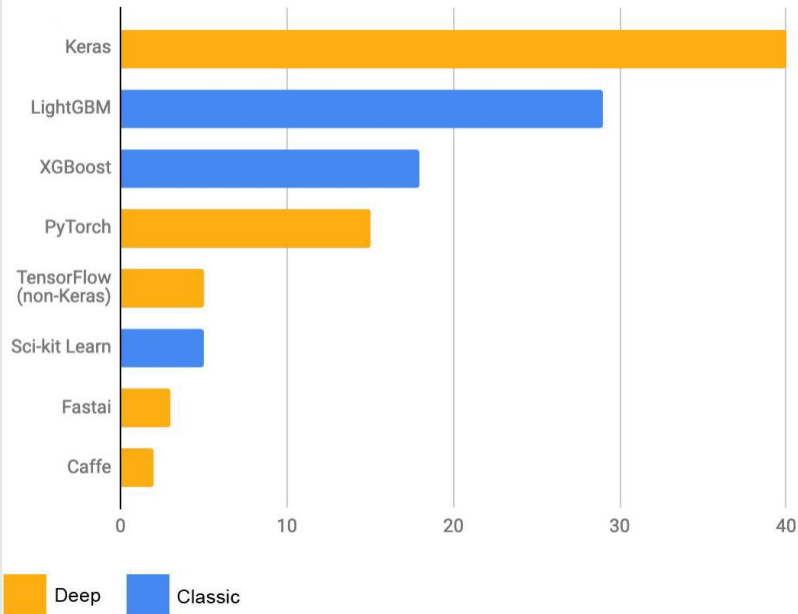
- Belongs to Google
  - Online community of data scientists and machine learning practitioners
  - Retrieve and publish data sets
  - Explore, build, and share models
- 
- Organizes competitions to solve data science challenges
    1. The competition host prepares the data and a description of the problem.
    2. Participants experiment with different techniques and compete against each other to produce the best models. Work is shared publicly through Kaggle Kernels to achieve a better benchmark and to inspire new ideas.
    3. After the deadline passes, the competition host pays the prize.

<https://en.wikipedia.org/wiki/Kaggle>

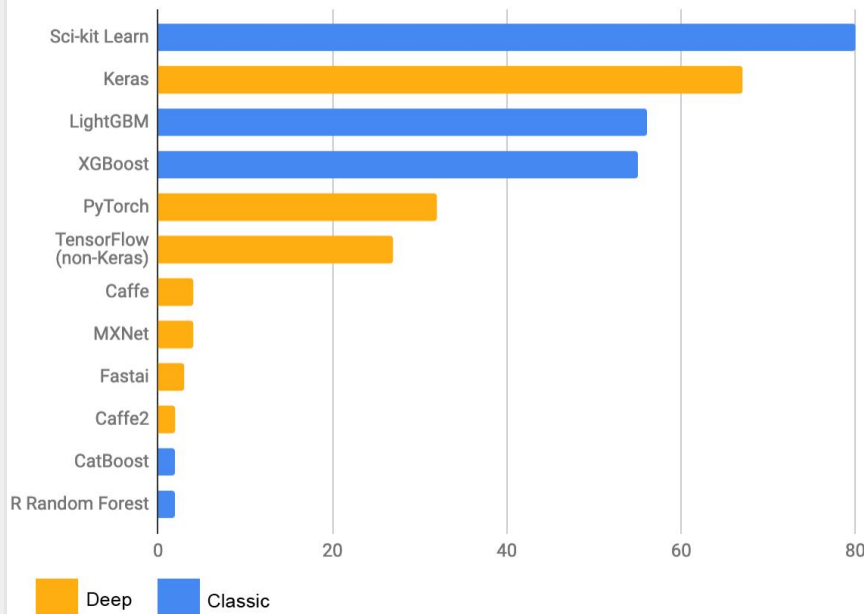
# Software Used in Kaggle Competitions



Primary ML software tool used by top-5 teams on Kaggle in each competition (n=120)



All (primary + auxiliary) ML software tools used by top-5 Kaggle teams in each competition (n=120)



[https://keras.io/why\\_keras/](https://keras.io/why_keras/)



# Survey: Prior Experience



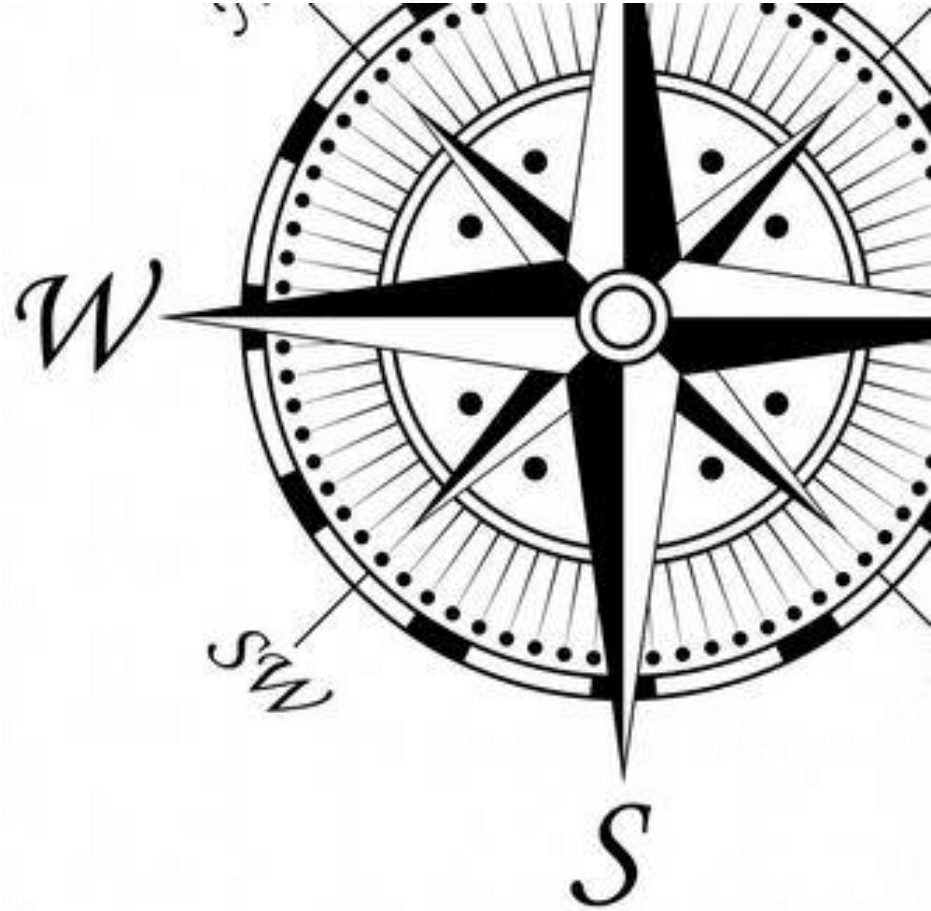
- <https://elearn.informatik.uni-kiel.de/mod/choice/view.php?id=4206>



# Topics Today

---

1. Frameworks & Libraries
2. **Keras**
3. Jupyter Notebook & Google Colab

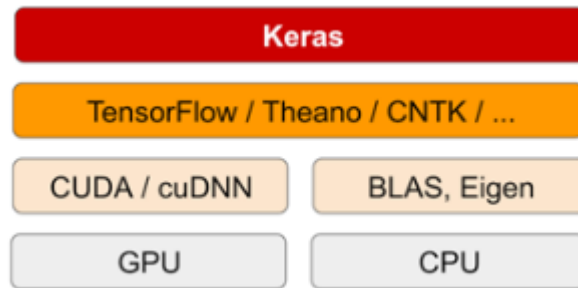


- Deep learning framework for Python that provides a convenient way to define and train almost any kind of deep-learning model (MIT license)
- [www.keras.io](http://www.keras.io)
- Compatible with Python 3.6–3.9
  - Ubuntu 16.04 or later; Windows 7 or later; macOS 10.12.6 (Sierra) or later.
- **Simple**: but not simplistic. Keras reduces developer *cognitive load* to free you to focus on the parts of the problem that really matter.
- **Flexible**: Keras adopts the principle of *progressive disclosure of complexity*: simple workflows should be quick and easy, while arbitrarily advanced workflows should be *possible* via a clear path that builds upon what you've already learned.
- **Powerful**: Keras provides industry-strength performance and scalability: it is used by organizations and companies including NASA, YouTube, or Waymo.

# TensorFlow 2 + Keras



- TensorFlow 2 is **low-level, an end-to-end, open-source machine learning platform**. You can think of it as an infrastructure layer for differentiable programming. It combines four key abilities:
  - Efficiently executing low-level tensor operations on CPU, GPU, or TPU.
  - Computing the gradient of arbitrary differentiable expressions.
  - Scaling computation to many devices, such as clusters of hundreds of GPUs.
  - Exporting programs ("graphs") to external runtimes such as servers, browsers, mobile and embedded devices.
- Keras is the **high-level API of TensorFlow 2** for solving machine learning problems, with a focus on modern deep learning. It provides essential abstractions and building blocks for developing and shipping machine learning solutions.



# More about Keras



## What is a Backend?

- A backend in Keras is a library/platform that performs all low-level computation such as tensor products, convolutions, and many other things.
- The “backend engine” will perform the computation and model building.
- Tensorflow is the default backend engine, others are possible.

Parameters	Keras	Tensorflow
Type	High-Level API Wrapper	Low-Level API
Complexity	Easy to use if you Python language	You need to learn the syntax of using some of Tensorflow function
Purpose	Rapid deployment for making model with standard layers	Allows you to make an arbitrary computational graph or model layers
Tools	Uses other API debug tool such as TFDBG	You can use Tensorboard visualization tools
Community	Large active communities	Large active communities and widely shared resources

# Installing TensorFlow 2 + Keras

---

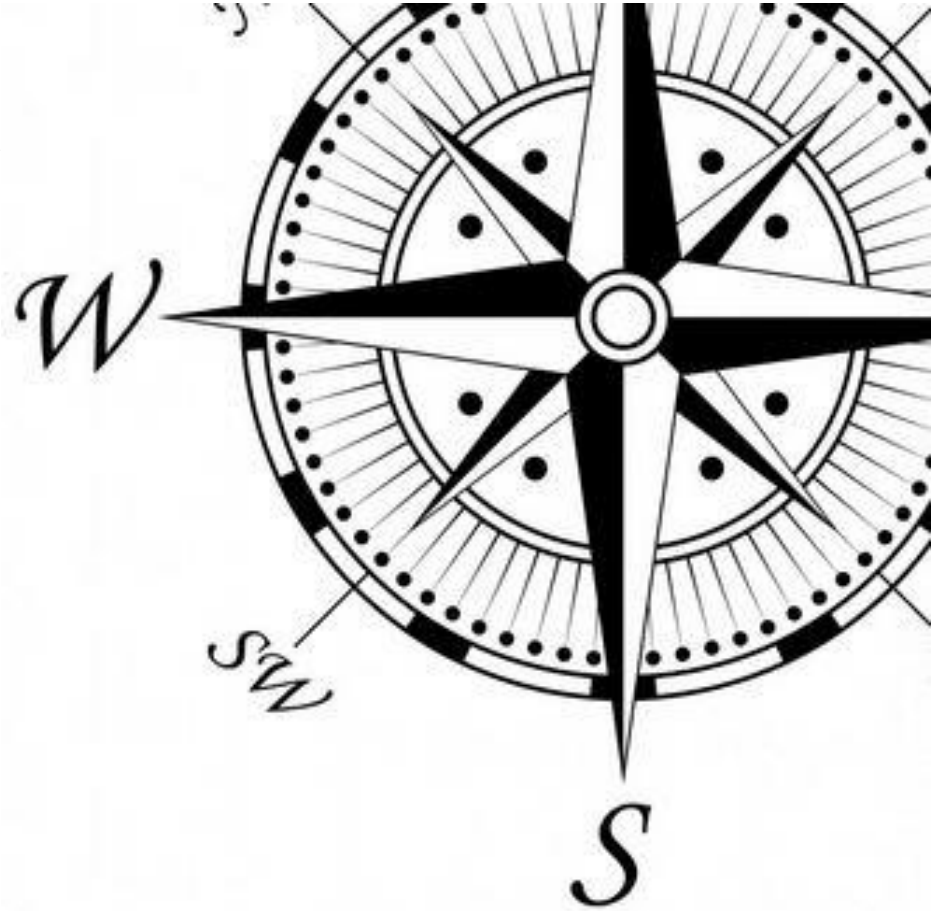


- More like a homework...

# Topics Today

---

1. Frameworks & Libraries
2. Keras
3. Jupyter Notebook & Google Colab



# Jupyter Notebook

---



- <https://jupyter.org/>
- The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
  - Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.



- <https://colab.research.google.com/>
- Colaboratory is a free Jupyter notebook environment that requires minimum or no setup and runs entirely in the cloud.
- With Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all for free from your browser.
- <https://colab.research.google.com/notebooks/>



- Train a text classifier for the 20 newsgroups dataset using scikit.learn on Google Colab after exploring (understanding) the dataset
- Hints
  - The 20 newsgroups dataset is directly available within scikit.learn
  - It is easy to transform text to tfidf vectors

```
from sklearn.datasets import fetch_20newsgroups  
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
vectorizer = TfidfVectorizer()
```

```
newsgroups_train = fetch_20newsgroups(subset='train', remove=('headers', 'footers', 'quotes'))  
vectors_train = vectorizer.fit_transform(newsgroups_train.data)  
newsgroups_test = fetch_20newsgroups(subset='test', remove=('headers', 'footers', 'quotes'))  
vectors_test = vectorizer.transform(newsgroups_test.data)
```

# Lerning Goals for this Chapter

---



- Know about deep learning frameworks
- Have an idea what Keras is
- Know how Colab works
- Be able to use Jupyter Notebooks

# Literature

---



- <https://keras.io/>
- <https://docs.jupyter.org/en/latest/>
- <https://colab.research.google.com/>
- <https://docs.python.org/3/>