PBN Presentation

Low Value Added to Data-driven Business

Content

- Introduction of Pannon Business Network (PBN)
- Introduction of am-LAB
- am-LAB activities
- Natural Language Processing (NLP)
  - Aim of the method
  - Tasks to do with NLP
  - Artificial Neural Networks (ANN)
  - Steps and applications of NLP
  - Experiences at am-LAB with NLP related tasks
  - Challenges

PBN Key Data

2006 April foundation

25 running projects (digitalisation, SME development)

1.5 mio € average yearly turnover

70 international development projects

450+ EU partner

PBN Main Results – 2019

10 modular digitalisation curricula development

200+ businesses' international study tour in 2019

200 participants at international AI conference

Artificial Intelligence Working Group Coordination of Danube Region

150 manufacturing companies complex development in 2019

4 mio € and 74 staff average client size

am-LAB Key Data

2018 Foundation

Engineering staff

8 Digitalisation technologies

300 m² of training and research space

12 Digitisation training material

am-LAB results – 2019

10 integrated digitisation applications

10 modular curricula developed

DIH accredited by the EU
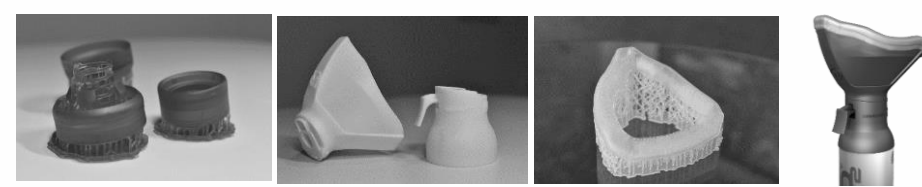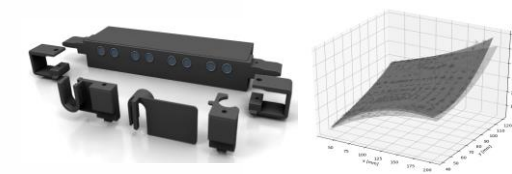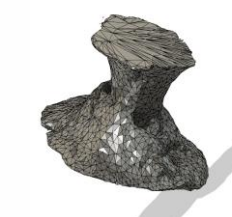
50+ businesses trained in one year

200+ 12-18 year old students per year

20+ schools cooperation with

350 Central European SMEs' current level as well as their needs were analysed in correlation with the nine technologies of I.4.0.

Linking Natural Language proccess and Data science

355 fully operational Digital Innovation Hub's analysis based on a DiH database

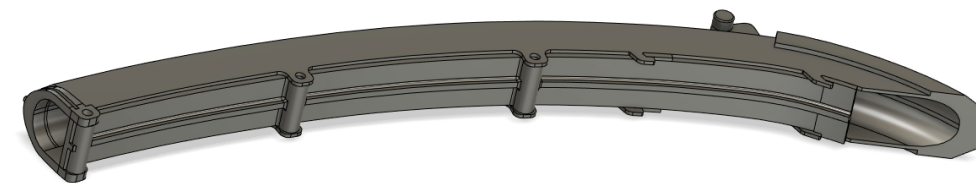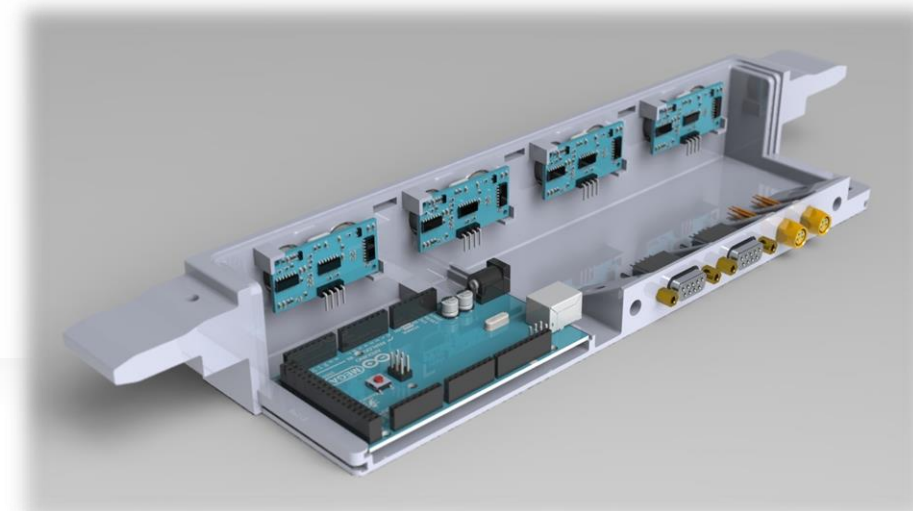Toby and Lily: am-LAB's promo video with CGI methodology:

Toby and Lily: Behind the scenes:
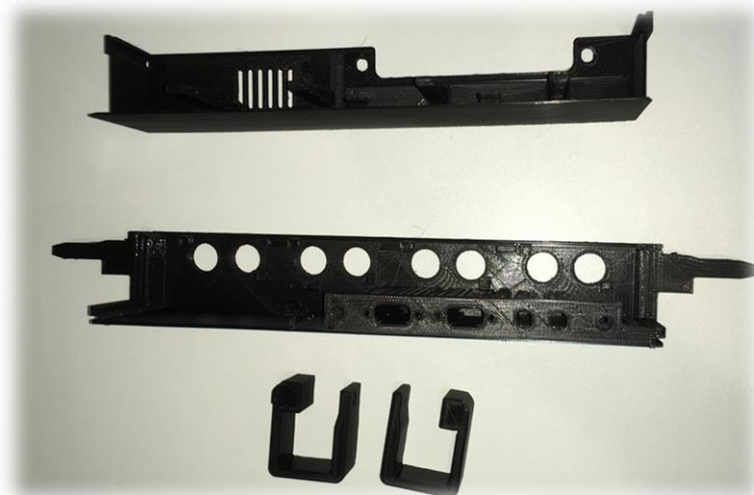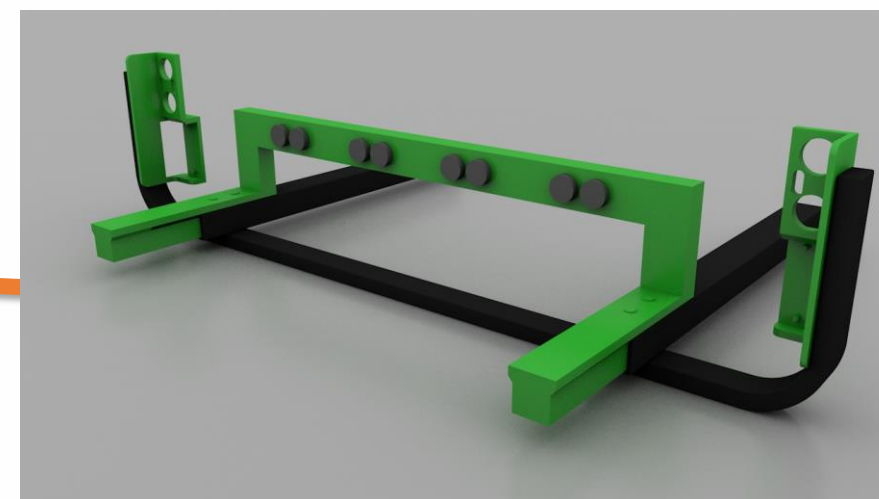
am-LAB reverse engineering

am-LAB product development

am-LAB product development
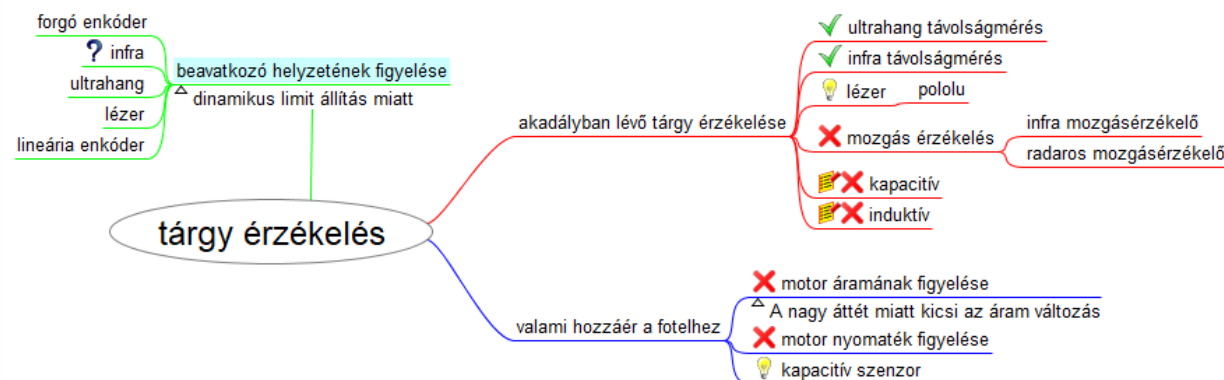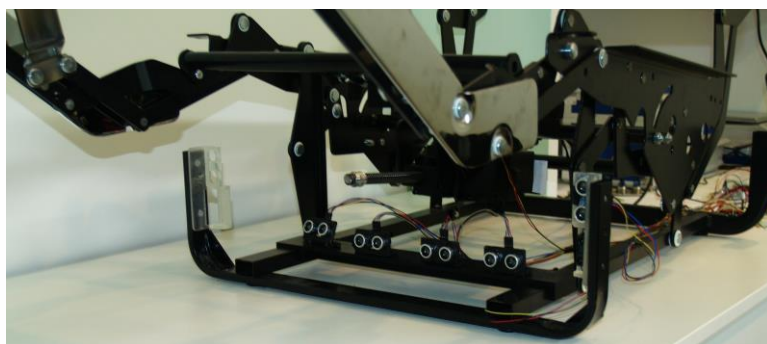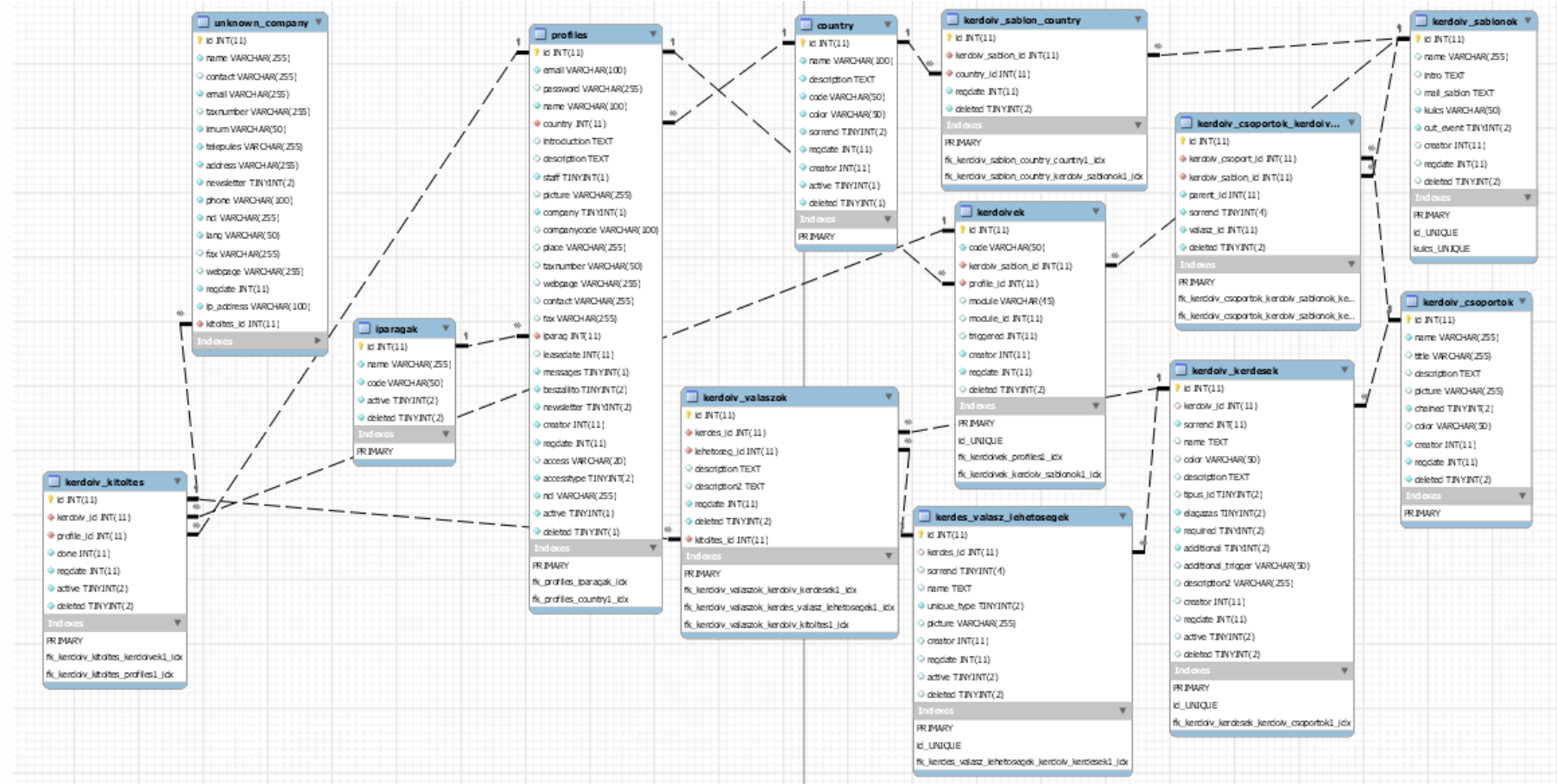
am-LAB product development

Data Analysis

- Data preparation
- Error correction
- Data extension (e.g. geographical information)
- Statistical information extraction
- Hypothesis testing
- ML algorithms (classification, clustering, regression, time-series analysis)

Data Analysis

- Simple graphs
- Heatmaps
- Distribution plots
- Geographical data
- Correlations
- Associations
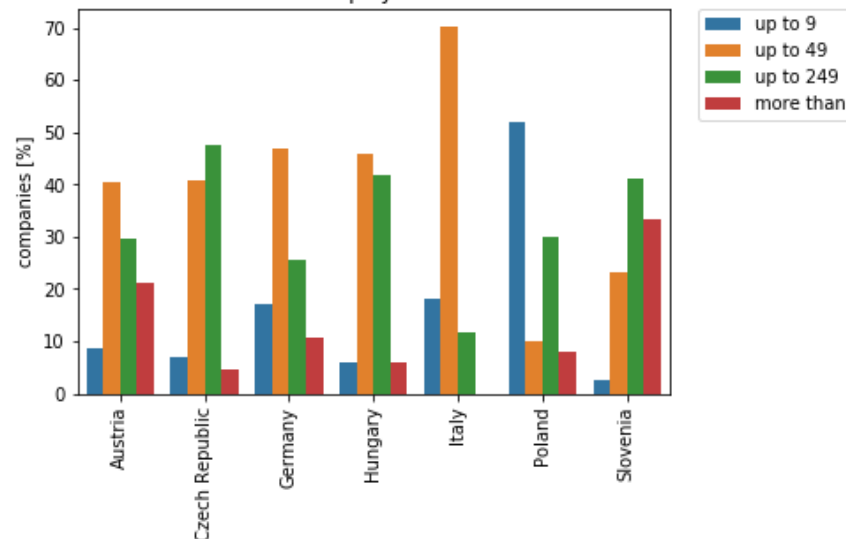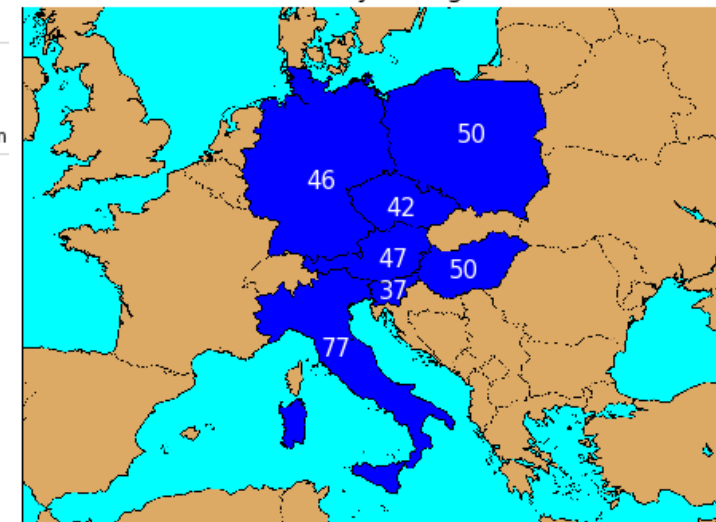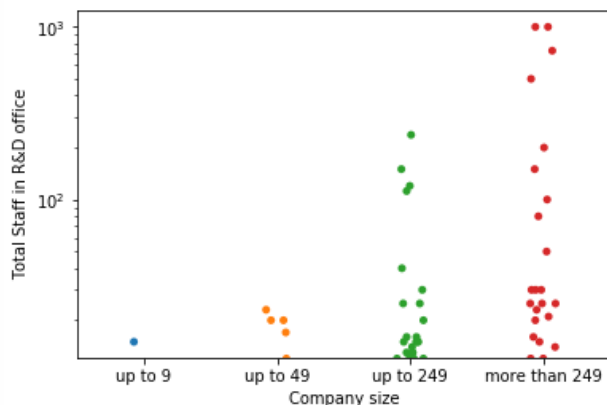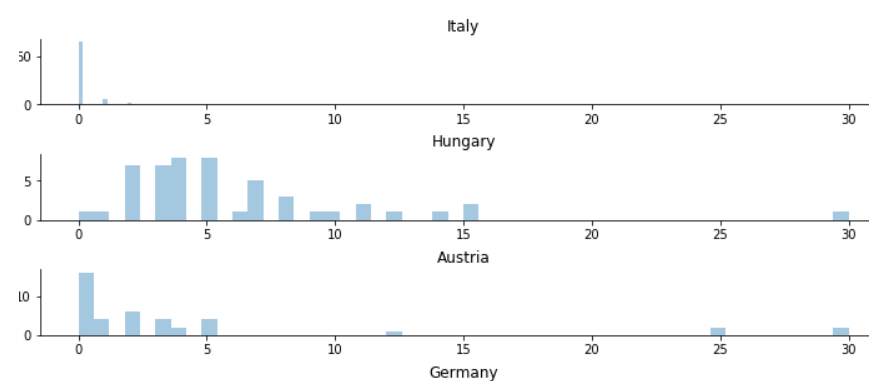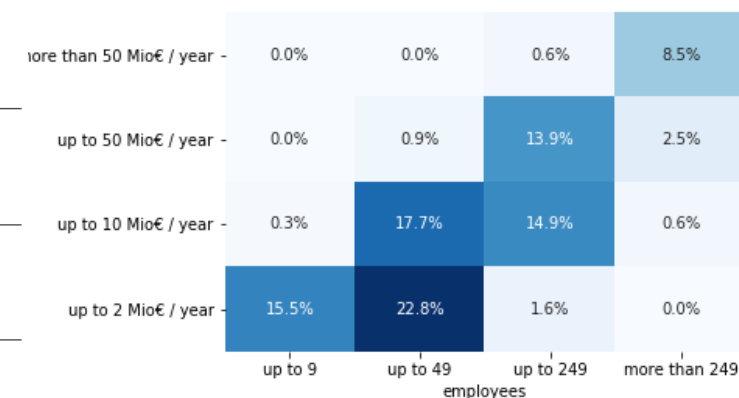
Natural Language Processing

- **Goal:** enable computers to process and analyze *large* amount of natural language data

- Subfield of *linguistics*, computer science and artificial intelligence

- **Challenge:** three completely different competency areas

- Main areas [1]:

  - speech recognition

  - natural language understanding

  - natural language generation

-

NLP

ML

DL

AI

- Translation Software
- Concept Learning

- Natural Language Generation
- Chatbots

- Regular Expression
- Statistics

- Automated Learning
- Non-Text Models
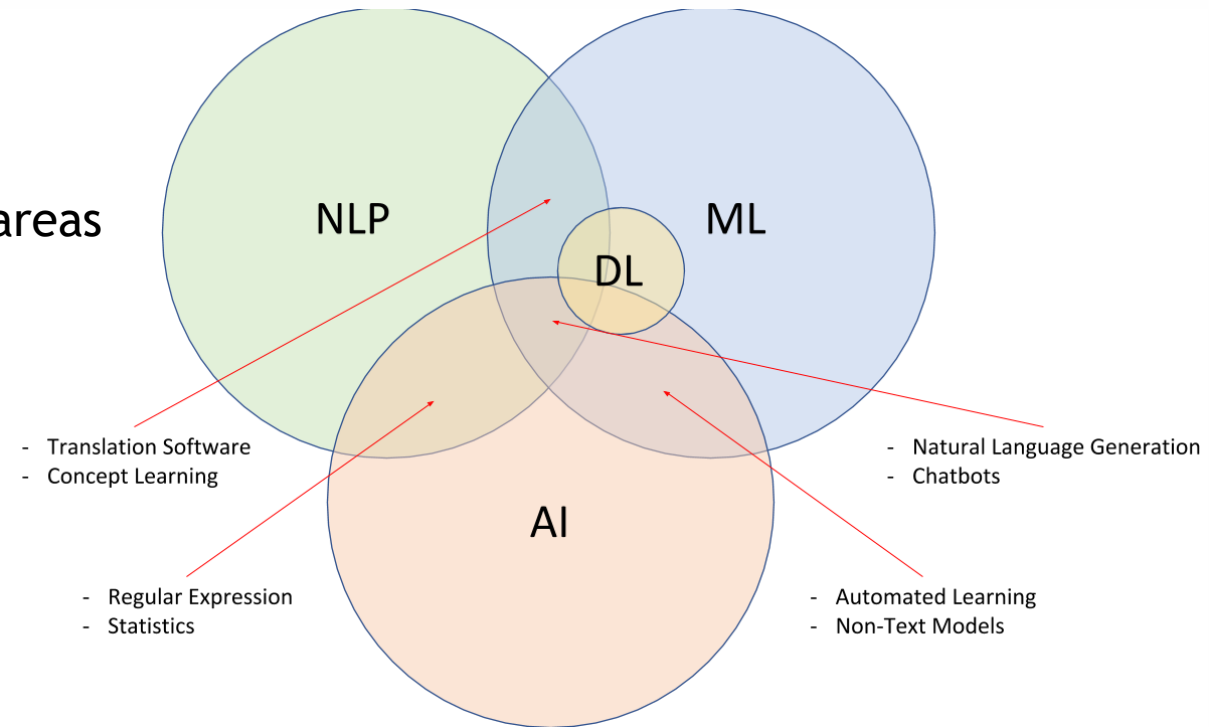
Natural Language Processing

- Input: textual data in digital form

- **Challenge:** collect and structure textual data (websites, .pdf, .png, audio files), which is in many cases

  - Laborious

  - Needs task specific queries and scripts

  - Requires creation of large databases

  - Problematic due to legal questions of information collection and extraction

NLP

ML

DL

AI

- Translation Software
- Concept Learning

- Natural Language Generation
- Chatbots

- Regular Expression
- Statistics

- Automated Learning
- Non-Text Models
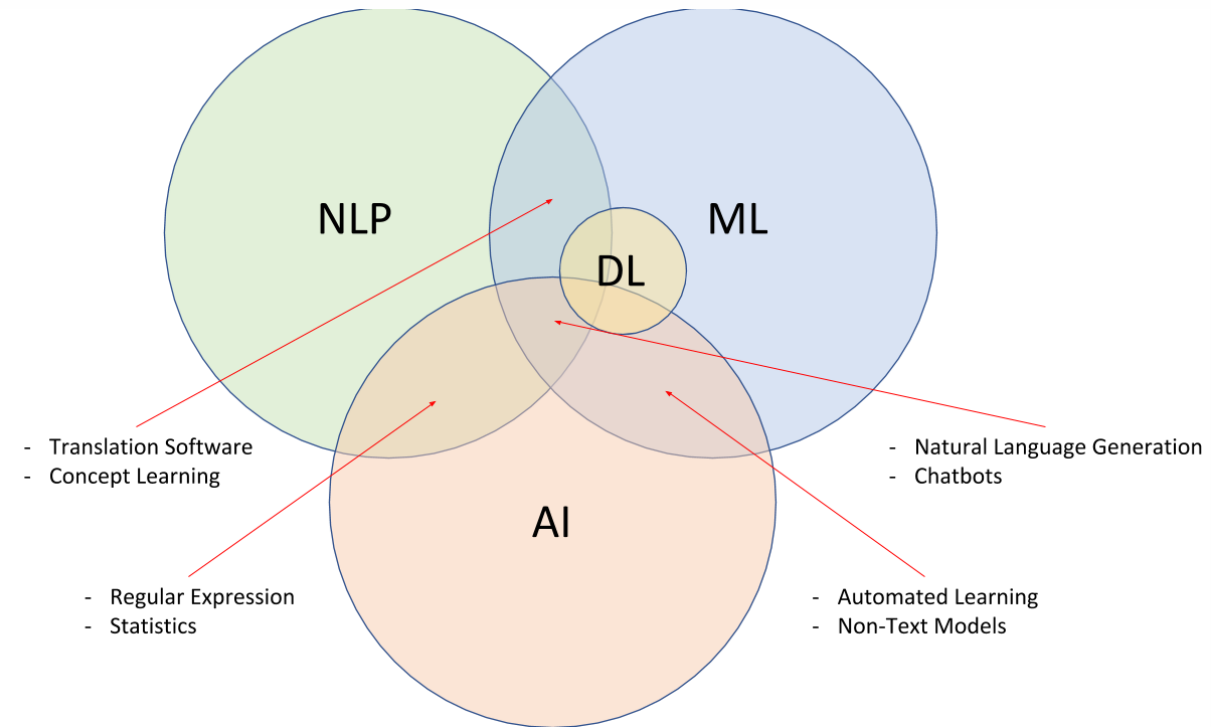
Natural Language Processing

- Up to 1980s mainly a rule-based decision system specified the NLP systems

- From 1980s a huge increase in computer power helped the existing machine learning algorithms to gain more importance

- Development and new theories in understanding grammar (transformational grammar)

- Taxonomy of current NLP research ares [2]:

PBN
advanced management

PANNON BUSINESS NETWORK
advanced management

am·LAB
advanced manufacturing
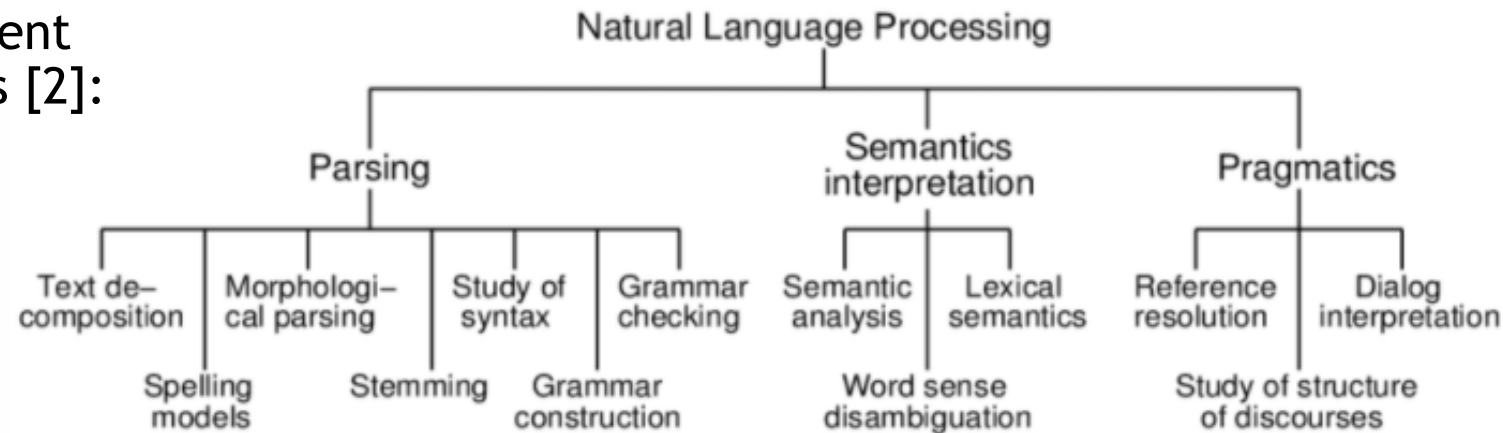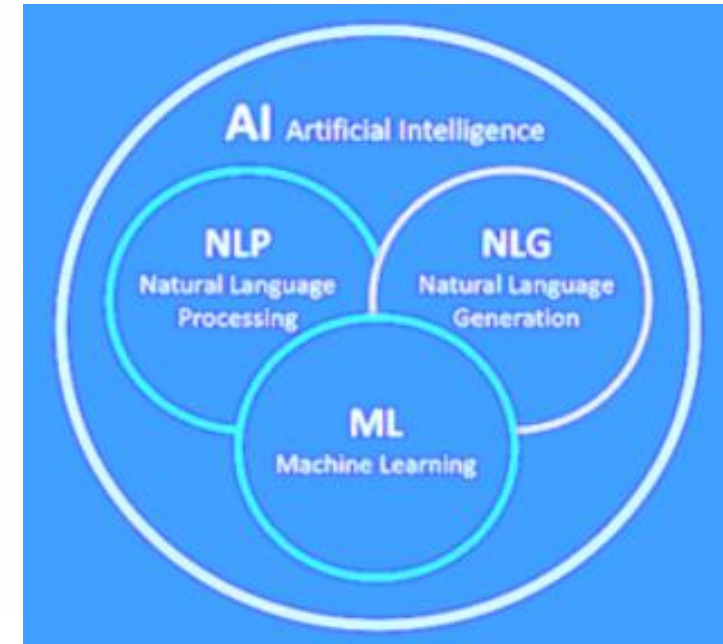
Natural Language Processing

- The NLP methods apply statistical approaches in order to:
  - Recognize
  - Understand
  - Categorize
  - Generate
  natural language texts

- The name, NLP may seem general

- It differentiates itself from other conventional fields of linguistics by:
  - the usage of statistical methods
  - the automatized fulfilment of the above tasks

Natural Language Processing → Analysis Tasks, Generation Tasks

Analysis Tasks → Syntactic, Semantic, Pragmatic

Syntactic → Tokenization

Semantic → Sentence Classification, Name Entity Recognition

Pragmatic → Word Sense Disambiguation

Part of Speech Tagging

Generation Tasks → Machine Translation, Question Answering, Language Modelling
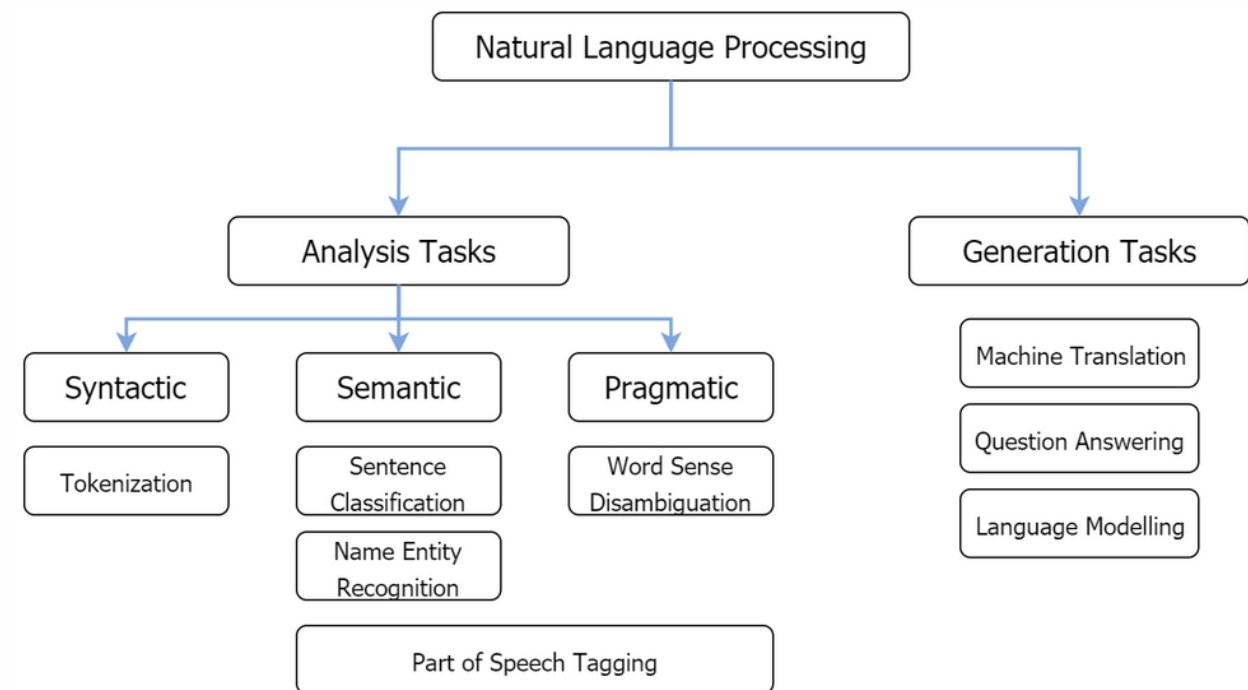
Natural Language Processing

- Nowadays (approximately from the 2010s) this area tends to use _neural networks_ so as to achieve its various goals

- Fortunately there are already numerous software solutions and even open-source possibilities to utilise
  - script development
  - concept testing
  - creation of complex applications
  - etc.
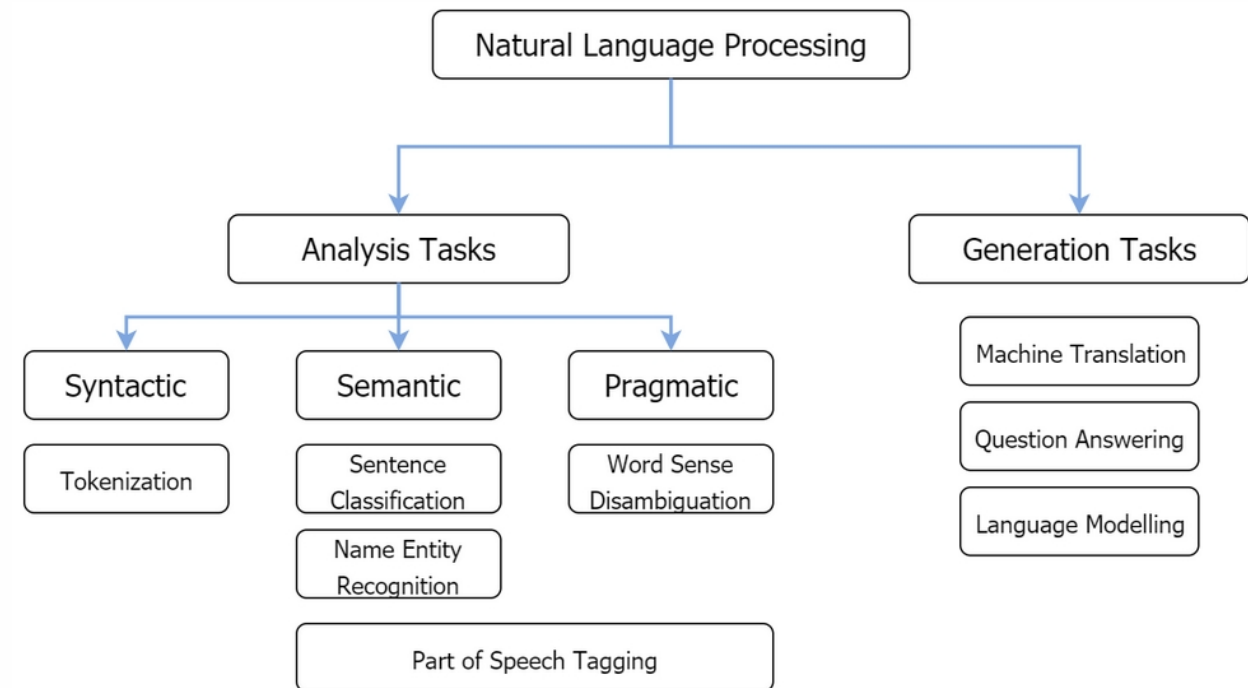
Natural Language Processing

- **Artificial Neural Networks** (ANN) are computing systems that are inspired by the biological neural networks (constitutional part of brain)

- Tries to replicate the functions of the human brain

- ANN is a model of a mathematical problem (mainly optimization problems) that is realized by computational algorithms in a way that resembles biological neuron function behavior

Artificial Neural Network
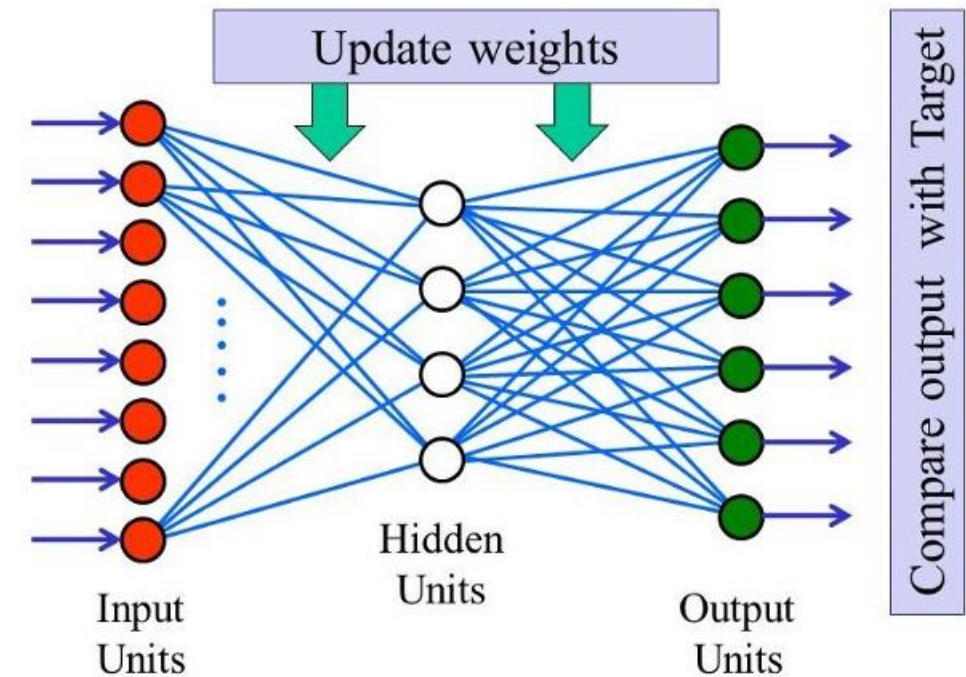
Natural Language Processing

- Simple model:

  - Input layer: receives data (sensors, images, audio etc.)

  - Hidden layer: processes the data (there might be several hidden layers)

  - Output layer: serves with one or more data points with respect to the function of the ANN model at hand (cat detection   detection of humans, cars animals etc.)

Artificial Neural Network

Update weights

Input Units

Hidden Units

Output Units

Compare output with Target
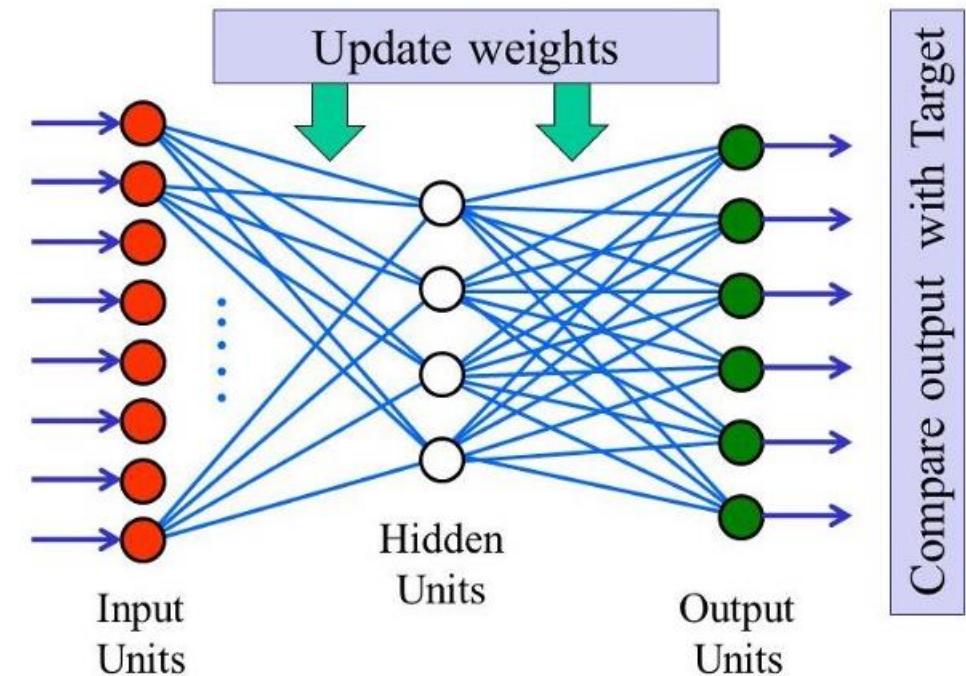
Natural Language Processing

- ANN-s conceptually belong under the hood of Artificial Intelligence methods and in a more restricted manner under Machine Learning methods

- The construction of ANN-s comprises training by real-life examples, where the results to a given input is given and well known (e.g.: labeled photos, tagged texts etc.)

- The efficiency / strength of the training of such a network is usually measured by the difference of the processed output (prediction) of the neural network and the well-known target output

Artificial Neural Network

Update weights

Input Units

Hidden Units

Output Units

Compare output with Target
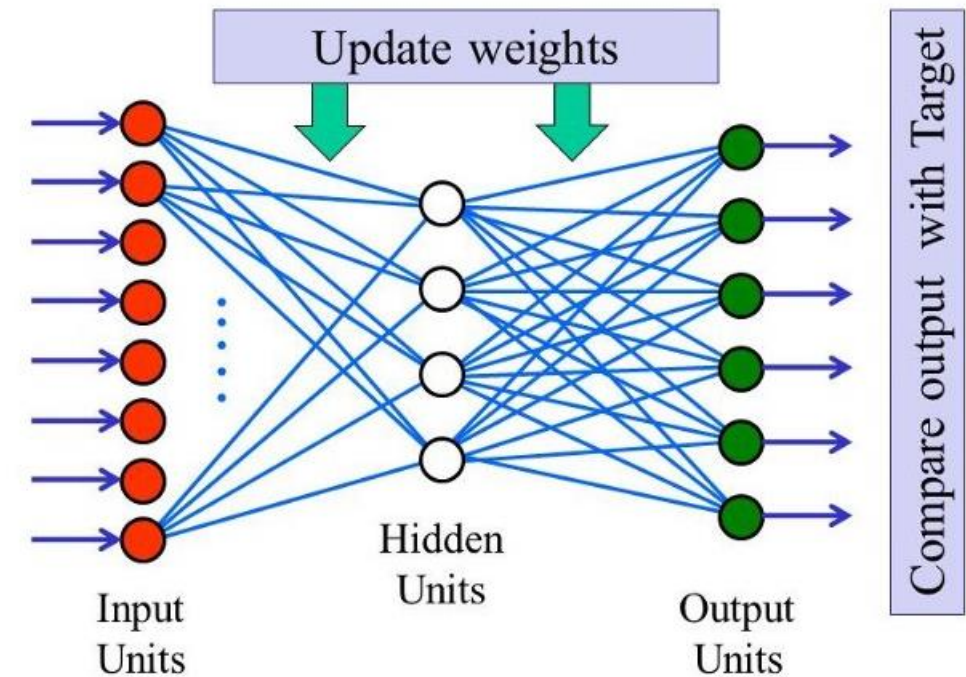
Natural Language Processing

- Given the error term the weights in the model are adjusted in a way that the network for the same input serves with an always better (more similar) prediction to the target output by reducing this error ➡️ *supervised learning of ANN-s*

- The learning is finished in case a new observation does not reduce the error rate significantly
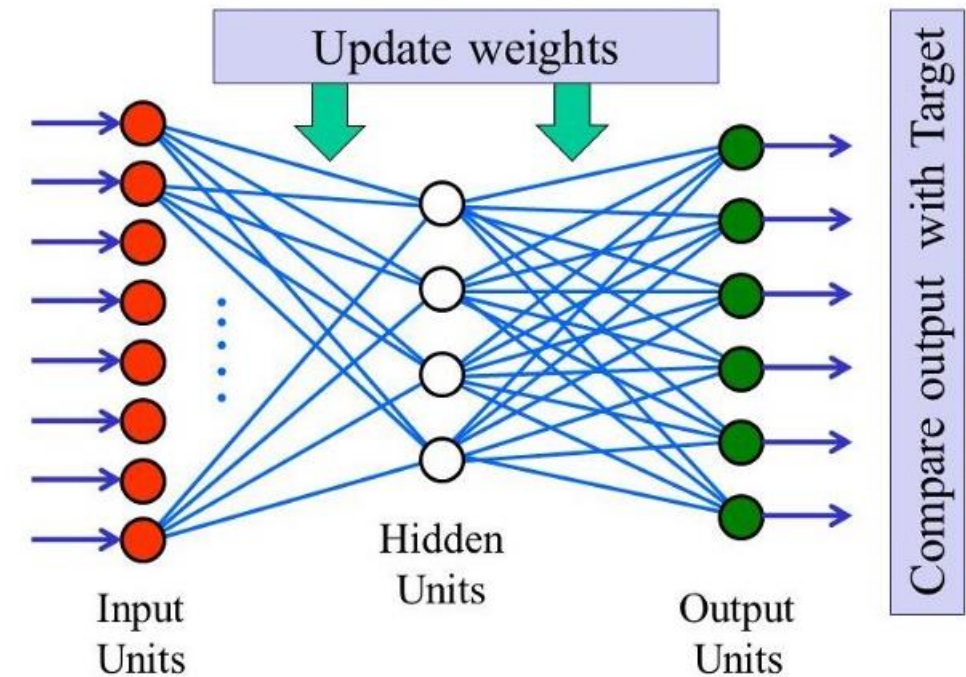
Natural Language Processing

- Still widely and heavily researched area
- State-of-the-art
- Has enjoyed a huge attention recently due to the numerous important application examples
- In many cases the reason why a neural network brings a suitable solution is still not completely understood

Artificial Neural Network

Update weights

Compare output with Target

Input Units

Hidden Units

Output Units
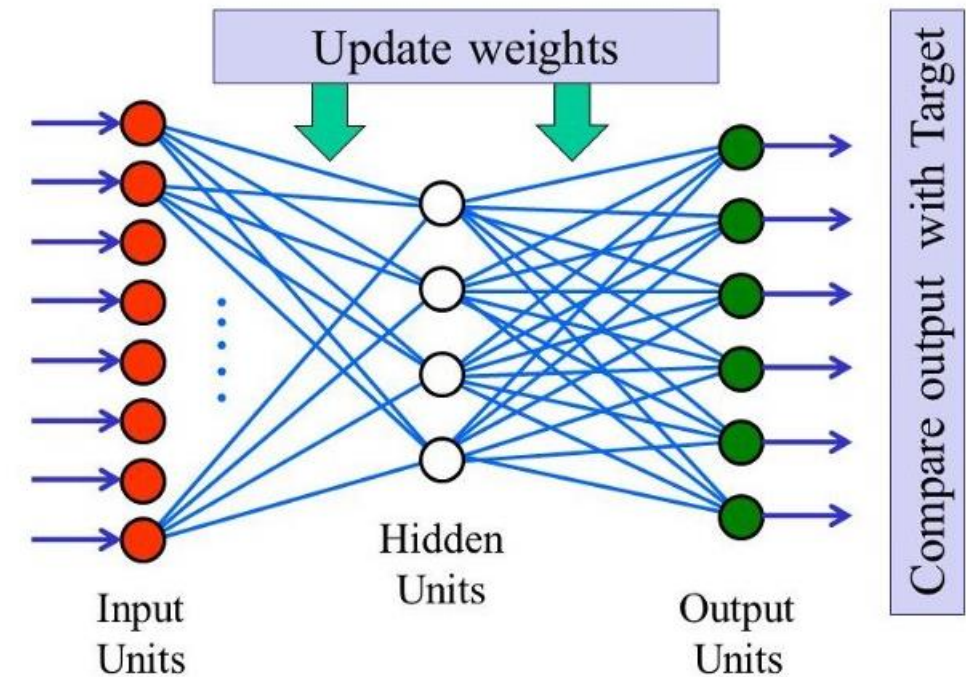
Natural Language Processing

- In practice:
  - many researcher/user usually apply „well-working" ANN-s
  - try to modify/develop them further in order to gain better and better results for a specific problem [5]
- This is also widespread due to the vast amount of computing resources needed to train a model

- (Transfer learning: select a pre-trained model and fine-tune it with new domain-specific examples [6])

Artificial Neural Network
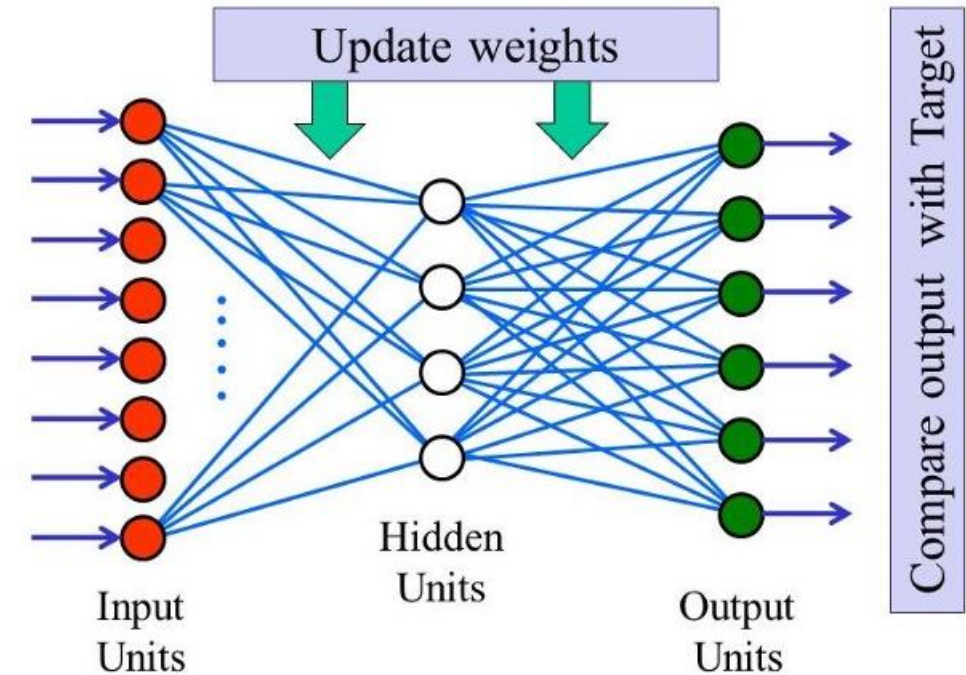
Natural Language Processing

- ANNs rely on probabilistic models that integrates threshold logic

- Each layer consists of mathematical model-elements that function similar to a biological neuron

- Each node in the representation corresponds to a neuron that receives a linear combination of weighted inputs to produce a single outcome through a (non)linear activation function

- ANNs are algorithms that are based on probabilities

- They use mathemtical functions to realize stochastic learning

- Note: ➡ ANNs do not „think" like human brain [3]
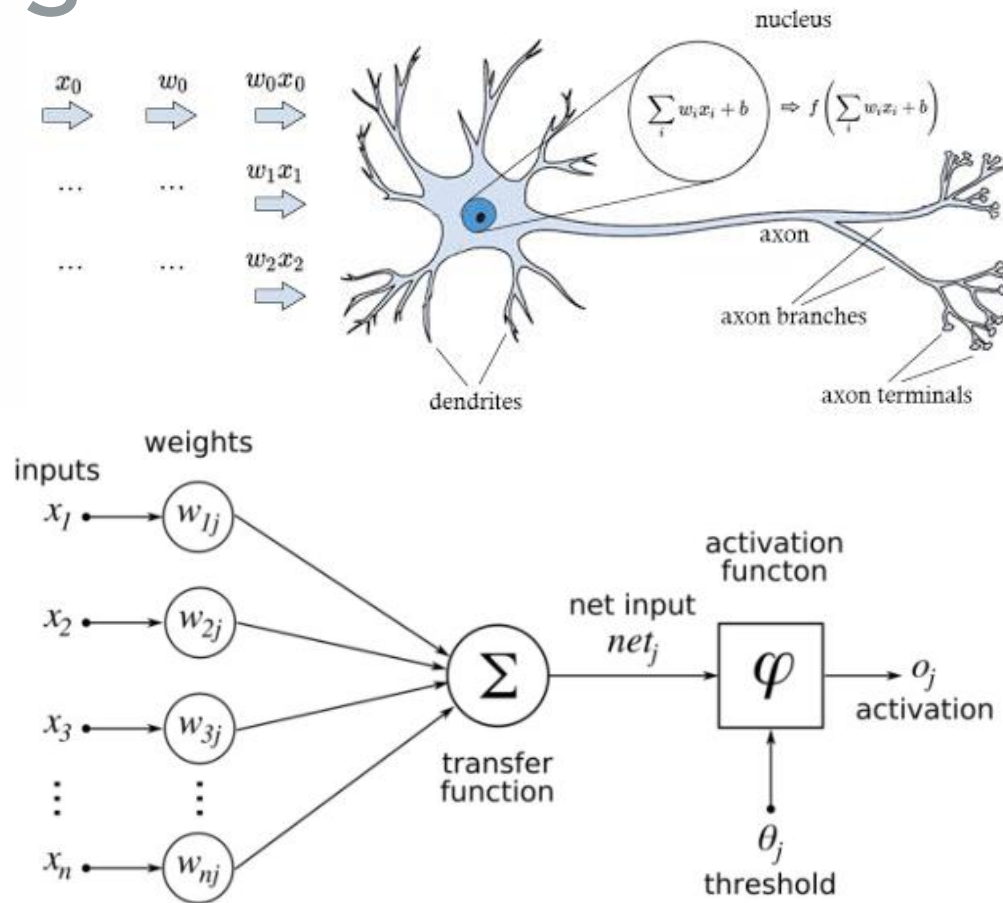
Natural Language Processing

- ANNs learn progressively through examples rather than through pre-defined commands and clear rules (neither sequential nor deterministic)

- ANNs do not run instructions

- The logic is dispersed across the many neurons

- Mathematical operations are being performed on each "neuron" inputs

- The collective operation of these "neurons" that develop the overall behaviour of the model [6]
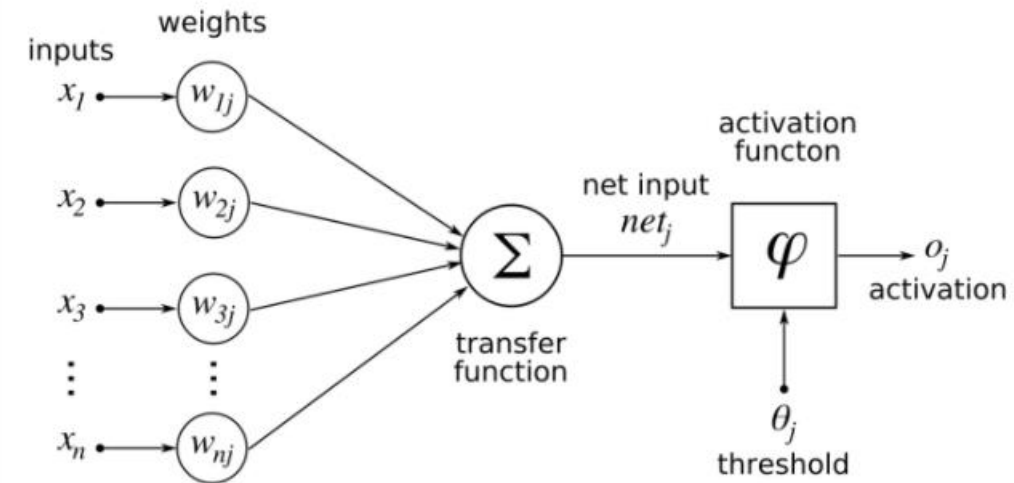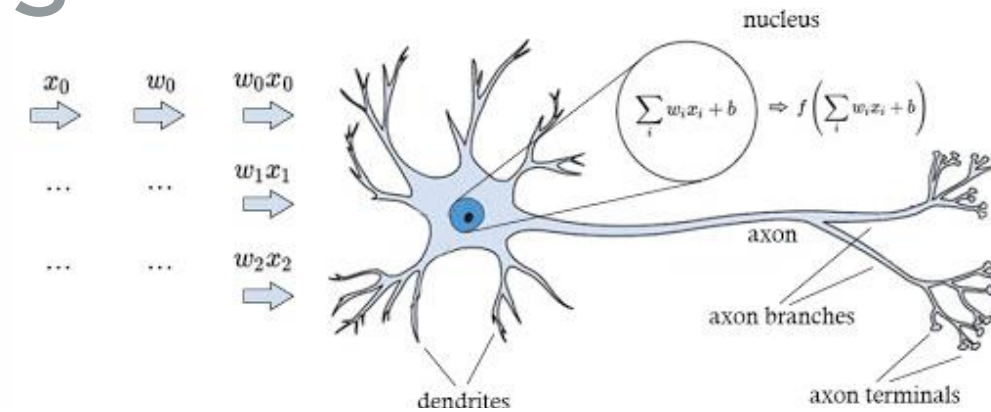
Natural Language Processing

- Difficult problems can be decomposed to small subproblems

- By iterative weight adjustments an accurate representation can be gained of the task

- This structure seems a set of trivial math operations

- Nonetheless hunderds of thousands of such artificial "neurons" connected in complex structures can perform very complicated tasks
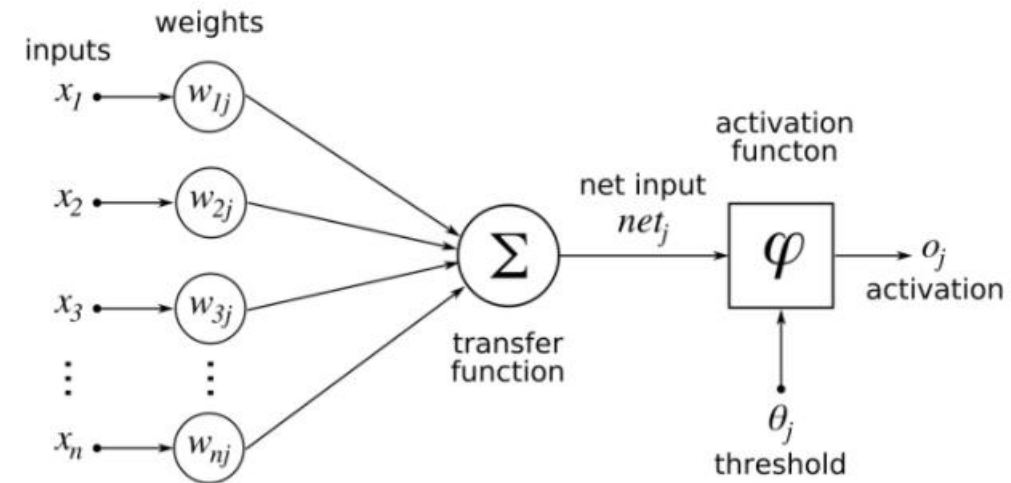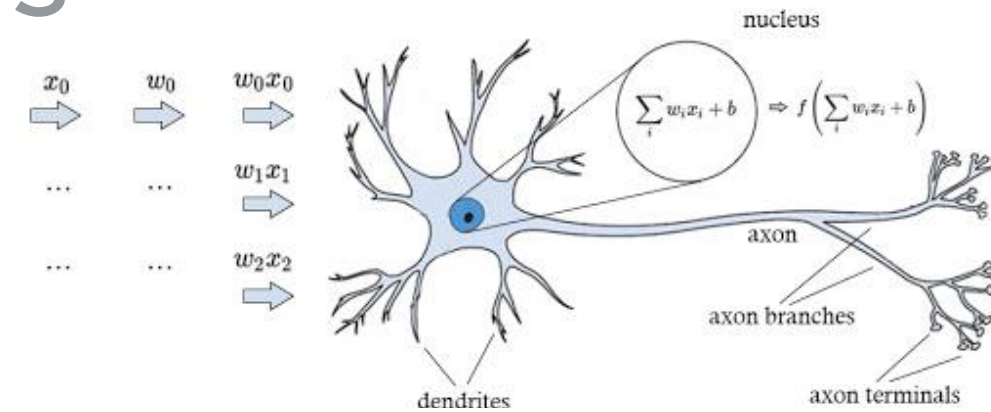
Natural Language Processing

- ANNs have gained a huge influence in various ways of life

- However they are often misunderstood or misinterpreted by users

- In applications where interpretability is more important than accuracy other machine learning methods might be preferred
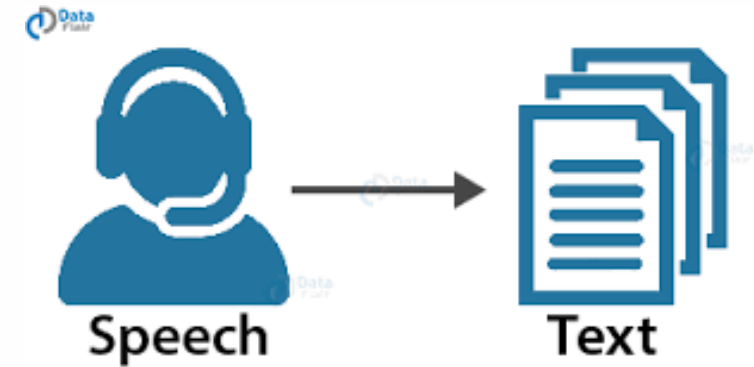
Natural Language Processing

- Despite of generally used terms ANNs are far from functioning as a human brain and AI is far from real intelligence:

  - Lot of data (examples) is needed

  - Bad at generalization (mainly good at a specific task but bad in anything else, ANNs cannot develop knowledge like humans)

  - ANNs are opaque (hard to determine the logic behind the scenes, even wrong factors can lead to an adequate solution, black box models)

Speech

Text
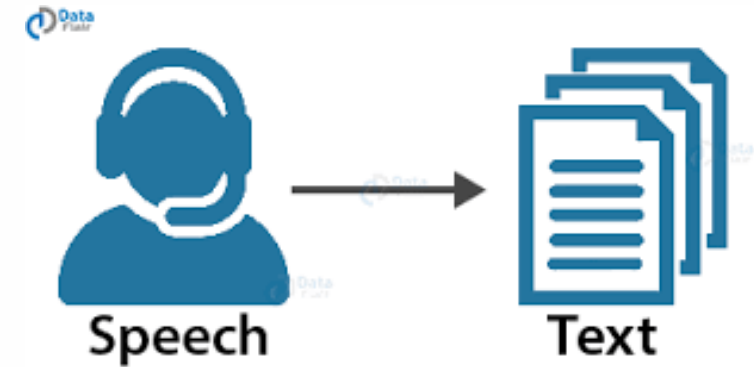
Natural Language Processing

- As the result of an exhausting training process the ANN-s can be usually applied to:

  - Classification of data

  - Classification of images (cat/dog, medical images etc.)

  - Data filtering

  - Data clustering

  - System identification and control (vehicle control, trajectory prediction)

  - Handwritten and printed text recognition

  - Speech-to-text conversions

  - e-mail spam detection

  - *and to many more...*

Natural Language Processing

- ANN based NLP applications:
  - Sentiment analysis (opinion, viewpoint)
  - Question answering → Chatbot implementation
  - Speech recognition (Google assistant, Cortana)
  - Machine Translation (Google translate)
  - Spell checking
  - Keyword searching
  - Advertisement matching
  - Text summarization and infromation extraction
  - Identification of relations and correlations
  - Prediction (e.g.: Does mentioning symptoms imply illness or threat of it?)
  - Topic categorization in large amount of texts etc.

Natural Language Processing

- Steps in NLP:
    1) **Tokenization:** breaking strings to smaller structures (e.g.: words of a sentence)
    2) **Stemming:** normalization of words to their roots (e.g.: usage, using, used → use, cutting of common pre- and suffixes, does not work always)
    3) **Lemmatization:** based on a detailed dictionary it links back to its root form (lemma), it considers the morphology of the word and the output shall be a proper word (e.g.: going, went, gone → go),
    4) **POS (parts of speech) tagging:** grammatical role (noun, verb, adjective, adverb etc.), based on context more meanings can be associated to the same word
    5) **Named entity recognition:** names of people, locations, organizations, quantities etc. done by noun phrase identification, phrase classification and entity disambiguation
    6) **Chunking:** putting together larger meaningful parts out of small tokens (e.g. nice red carpet)

Natural Language Processing

- Python has several open source NLP libraries
  - am-LAB has experience with python and a number of these libraries
    - NLTK, gensim, Stanford CoreNLP etc
  - Tokenize and tag text elements
  - Stemming and lemmatization
  - Display most frequently used words or tokens
  - Identify topics and their relations to each other
  - Perform text classification based on identified topics
  - Necessary programming skills for textual data query and database building
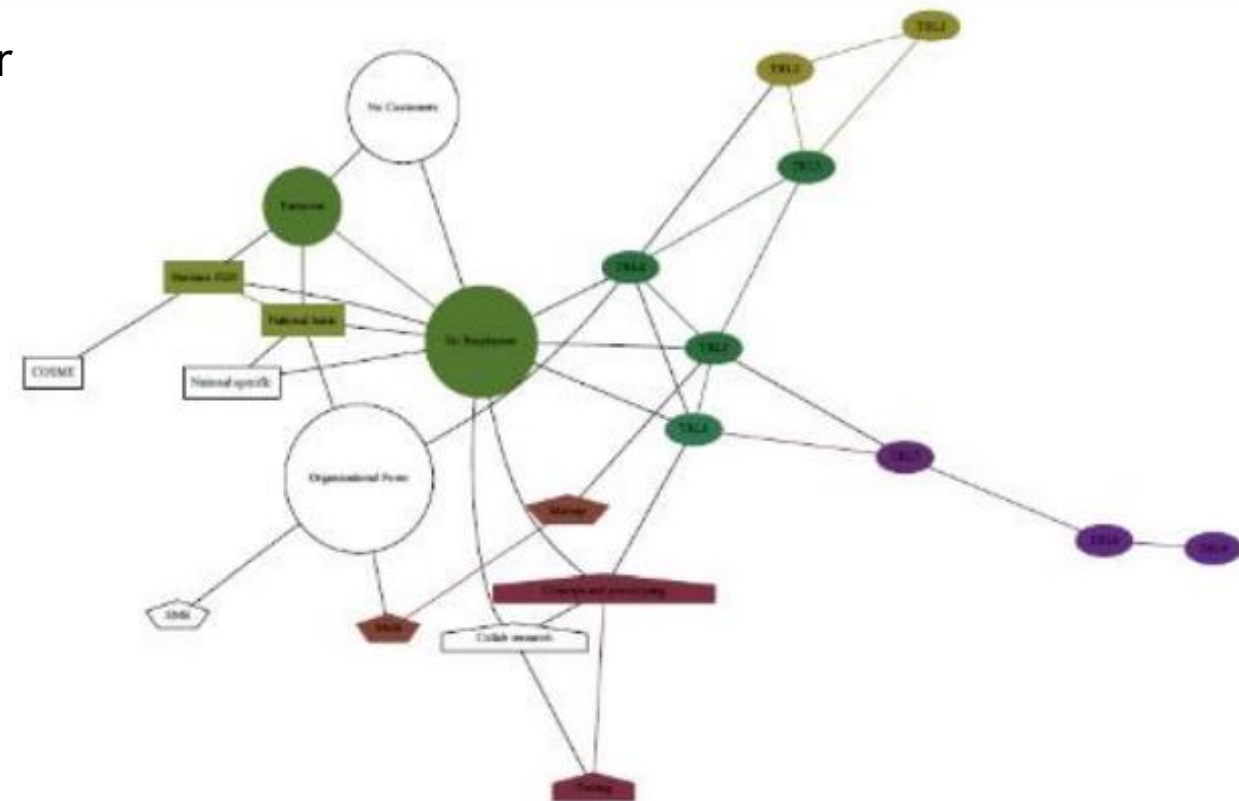
Natural Language Processing

- **Challenges:**
  - In case of textual data processing large amount of data has to be looked through

  - The data should be digital (websites etc.)

  - Different types of data formats (.pdf, audio, .html, pictures etc.)

  - To train neural networks, to perform various tasks the data needs to be tagged by experts

  - English language is well researched and understood in terms of grammar and structure; Hungarian is morphologically complex and its processing poses special challenges

  - In our specific case further cooperation with Hungarian research institutes might be necessary

Natural Language Processing

- **Challenges:**
  - Depending on particular problem more serious hardware might be needed than that is currently available at am-LAB
  - Due to the morphological complexity of Hungarian language a significant amount of language specific work may be necessary
  - A huge amount of unstructured language data is expected (social media sites, forums, news sites etc.)
  - The media consumed by the target groups need to be surveyed and mapped to discover the most important areas in terms of fake news and hate speech
  - Despite of state-of-the-art problem a lot of learning has to be done even in application of available tools and possibilites

Thank you for your attention.

Ferenc Tolner

ferenc.tolner@am-lab.hu

References

- [1]: https://en.wikipedia.org/wiki/Natural_language_processing

- [2]: https://www.researchgate.net/publication/279501910_Computationally_intelligent_methods_for_qualitative_data_analysis

- [3]: https://groupfuturista.com/blog/artificial-neural-networks-man-vs-machine/

- [4]: https://en.wikipedia.org/wiki/Artificial_neural_network

- [5]: https://voices.uchicago.edu/willett/teaching/fall-2019-mathematical-foundations-of-machine-learning/

- [6]: https://thenextweb.com/neural-basics/2020/05/27/everything-you-need-to-know-about-artificial-neural-networks/

- [7]: http://www.datascienceassn.org/sites/default/files/Natural%20Language%20Processing%20with%20Python.pdf