

# Grundtvigs Sandkasse

## Introduction to the NFSG Sandbox Environment

Kristoffer L Nielbo  
`kln@cas.dk`  
`knielbo.github.io`

Center for Humanities Computing, `chcaa.io`  
Aarhus University, Denmark

April 23, 2019

# Outline

- 1 Center for Humanities Computing Aarhus  
Software 2.0  
CHCAA.io
- 2 Automatisering af humaniora  
Maskinlæring  
... i humaniora  
Neural netværk
- 3 Grundtvigs Sandkasse  
Parser  
Forskningsdatabase  
Concept Nucleus  
Hyper-graf

Grundtvigs Sandkasse

Kristoffer L Nielbo  
kln@cas.dk  
knielbo.github.io

Center for Humanities  
Computing Aarhus

Software 2.0  
CHCAA.io

Automatisering af  
humaniora

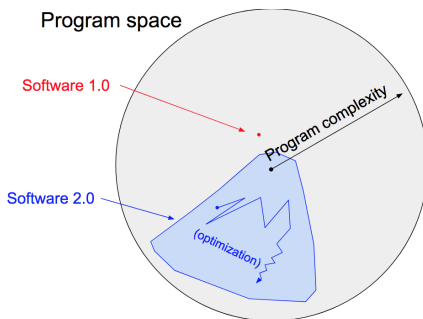
Maskinlæring  
... i humaniora  
Neural netværk

Grundtvigs Sandkasse

Parser  
Forskningsdatabase  
Concept Nucleus  
Hyper-graf



# Udvikling af computerbaseret forskning



Grundtvigs Sandkasse

Kristoffer L Nielbo  
kln@cas.dk  
knielbo.github.io

Center for Humanities  
Computing Aarhus

**Software 2.0**  
CHCAA.io

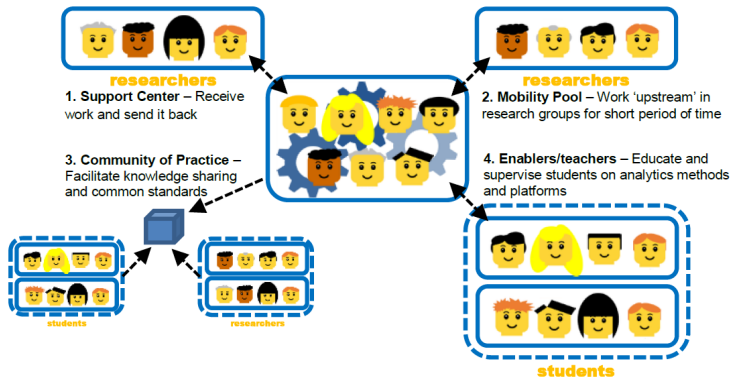
Automatisering af  
humaniora

Maskinlæring  
... i humaniora  
Neural netværk

Grundtvigs Sandkasse

Parser  
Forskningsdatabase  
Concept Nucleus  
Hyper-graf





Aarhus Universitet har besluttet at understøtte denne udvikling ved at etablere **Center for Humanities Computing Aarhus**

- Grundtvig Centeret har indgået et partnerskab med CHCAA, som indebærer udvikling af en række forsknings- og udviklingsprojekter over de næste par år.
- Internt har vi navngivet projektet **The NSFG Sandbox**, som består af en database samt en række applikationer
- Hvis vi har overset emner eller I har forslag til nye applikationer, **send en ticket**

## Når maskiner der lærer

Maskinlæring oprinder i kunstig intelligens (AI)- **bygge et system som automatisk udvikler sig gennem erfaring**

- anvendelsesdomænet er for komplekst til at designe en algoritme manuelt
- applikationen behøver at tilpasses til det operationelle miljø efter den er sat i produktion

### Tom Mitchell's well-posed learning problem

A computer program is said to learn from experience  $E$  with respect to some task  $T$  and some performance measure  $P$ , if its performance on  $T$ , as measured by  $P$ , improves with experience  $E$

Historisk er maskinlæring en del af industrialderens (videre-)udvikling af automatisering



## Humaniora i mødet med maskinlæring

---

Som følge af digitalisering, har humaniora også “jumped the automation bandwagon”

– plus teoretiske innovationer der beror på såkaldt maskin- og dyb læring (e.g., leksikal → kompositionel semantik)

Iboende udfordringer for vores data og brugere

– data er ustrukturerede, mangler normalisering og er relativt små  
– brugere mangler videnskabelige færdigheder med computere, stor afstand mellem teknologi og domæneviden

Problemer som maskinlæring løser:

– maskinlæring som mål for et forskningsproblem  
– maskinlæring som middel, der automatiserer resourcekrævende hjælpeopgaver

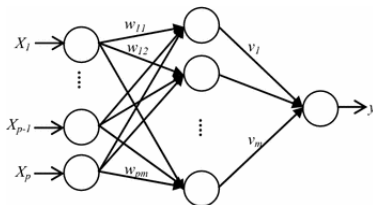


# Afmystificering af neurale netværk

Grundtvigs Sandkasse

Kristoffer L. Nielbo  
kln@cas.dk  
knielbo.github.io

Kunstige neural netværk (“kunstig intelligens”) er mængde simple regneenheder, der er forbundet i en lagdelt struktur.



Et (*feedforward*) neuralt netværk med et skjult lag af størrelse  $m$ . Et skjult lag er et lag mellem *input* og *output*. Skjulte lag udfører transformationer på *input* eller forudgående skjulte lag. Netværket lærer ved at modificere dets vægte  $v$  og  $w$ .

Er neuralt netværk kan bestå af en vilkårlig sammensætning af neuroner og lag. Dyb i dyb læring refererer repræsentationer, der er lært over mange (“dybe”) strukturer. Input propageres fremad gennem transformationer for at genererer et output.

Center for Humanities  
Computing Aarhus

Software 2.0  
CHCAA.io

Automatisering af  
humaniora

Maskinlæring  
... i humaniora

**Neural netværk**

Grundtvigs Sandkasse

Parser

Forskningsdatabase

Concept Nucleus

Hyper-graf



```
1  %%%xml
2  <p rend="firstIndent">
3  <persName key="pe35">Schiller</persName>
4  siger i Fortalen:
5  <seg type="comStart" n="com46" />
6  den høieste Nydelse
7  er
8  <seg type="com" n="com208">Gemyttets</seg>
9  Frihed under alle Kræfternes levende Spil
10 <seg n="com46" type="comEnd" />
11 , og er dette saa, da maa dette Drama
12 næsten heelt igiennem unde os den høieste Nydelse;
13 thi uagtet den stærkeste
14 Sysselsættelse af enhver
15 <seg type="com" n="com47">Kraft</seg>
16 , uagtet den høie
17 Interesse, vi føle for de
18 fremtrædende Figurer,
19 see vi dog deres Undergang med samme Rolighed, som de
20 selv.
21 </p>
```

“Schiller siger i Fortalen: den høieste Nydelse er Gemyttets Frihed under alle Kræfternes levende Spil , og er dette saa, da maa dette Drama næsten heelt igiennem unde os den høieste Nydelse; thi uagtet den stærkeste Sysselsættelse af enhver Kraft, uagtet den høie Interesse, vi føle for de fremtrædende Figurer, see vi dog deres Undergang med samme Rolighed, som de selv.”



NFSG Parser er en del af NFSG DB, der vil gøre *Grundtvigs Værker* lettere tilgængelige for computerbaseret forskning. Parseren tilbyder pt. også subkorpusgeneration via eksisterende metadata.

Collection of Documents



Document Data Schema



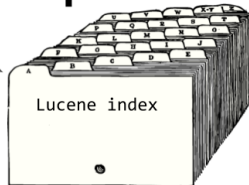
Full-text searchable web application



Vue.js



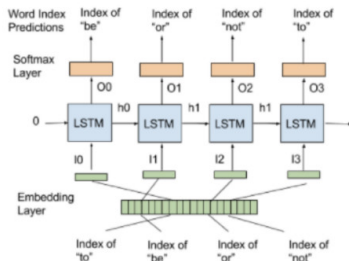
mongoDB



NFSG DB vil tilbyde en række services gennem CHCAAs cluster og data hosting (fx., entitetsgenkendelse, syntaktiske taggers og parsers, samt semantiske modeller).

# NFSG QA Model

NFSG QA Model kombinerer bruger rekurrente neurale netværk til at lære ord og sætningsrepræsentationer af Grundtvig



den høieste Nydelse  $[x_1]$  Gemyttets Frihed

→

er

den høieste Nydelse er Gemyttets Frihed  $[x_1, x_2, x_3, x_4, x_5]$

→

under alle Kræfternes levende Spil

Efterfølgende kan vi bruge repræsentationerne til at løse natursprogsopgaver, fx. spørgsmål-svar, ordassociationer og tekstgenerering

# NFSG Concept Nucleus

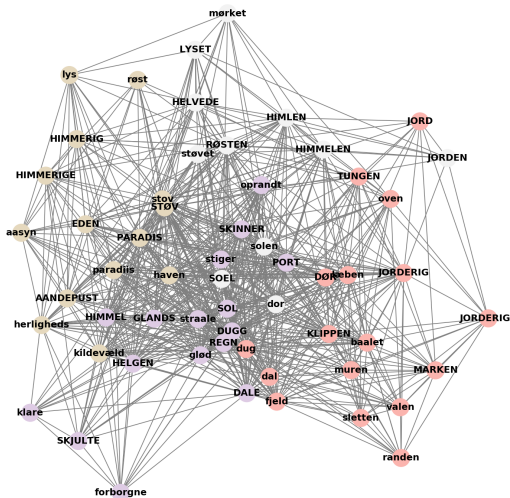
Grundtvigs Sandkasse

Kristoffer L Nielbo  
kln@cas.dk  
knielbo.github.io

Center for Humanities  
Computing Aarhus  
Software 2.0  
CHCAA.io

Automatisering af  
humaniora  
Maskinlæring  
... i humaniora  
Neural netværk

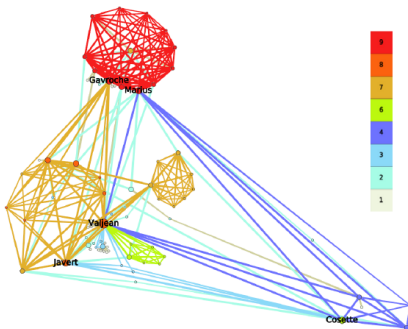
Grundtvigs Sandkasse  
Parser  
Forskningsdatabase  
Concept Nucleus  
Hyper-graf



Concept Nucleus er den første applikation baseret på QA Modellen



# NFSG Hyper-graf



Decomposition of the co-appearance network Les Misérables, the five most diverse characters are explicitly labeled.

- Grafer baseret på entiteters associationer i sætninger, afsnit eller dokumenter er meget populære, men en spatial association er ikke tilstrækkelig
- Vi bygger en hypergraf, hvor hver entitet har tre noder (tid, rum og semantik) for samtlige entiteter i NFSG DB.
- Efterfølgende kan brugere definere eksperimenter på hypergrafen

```
1 if questions:
2     try:
3         answer()
4     except RuntimeError:
5         pass
6     else:
7         print("thank you")
```

**TAK**  
kln@au.dk  
knielbo.github.io  
chcaa.io

slides: [http://knielbo.github.io/files/kln\\_nfsgsandbox.pdf](http://knielbo.github.io/files/kln_nfsgsandbox.pdf)

**& tak til**

Oliver Jarvis, Kenneth Enevoldsen, Ross McLachlan, Max Eckardt, & Peter Vahlstrup

Grundtvigs Sandkasse

Kristoffer L Nielbo  
kln@cas.dk  
knielbo.github.io

Center for Humanities  
Computing Aarhus

Software 2.0  
CHCAA.io

Automatisering af  
humaniora

Maskinlæring  
... i humaniora  
Neural netværk

Grundtvigs Sandkasse

Parser  
Forskningsdatabase  
Concept Nucleus  
Hyper-graf

