# 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

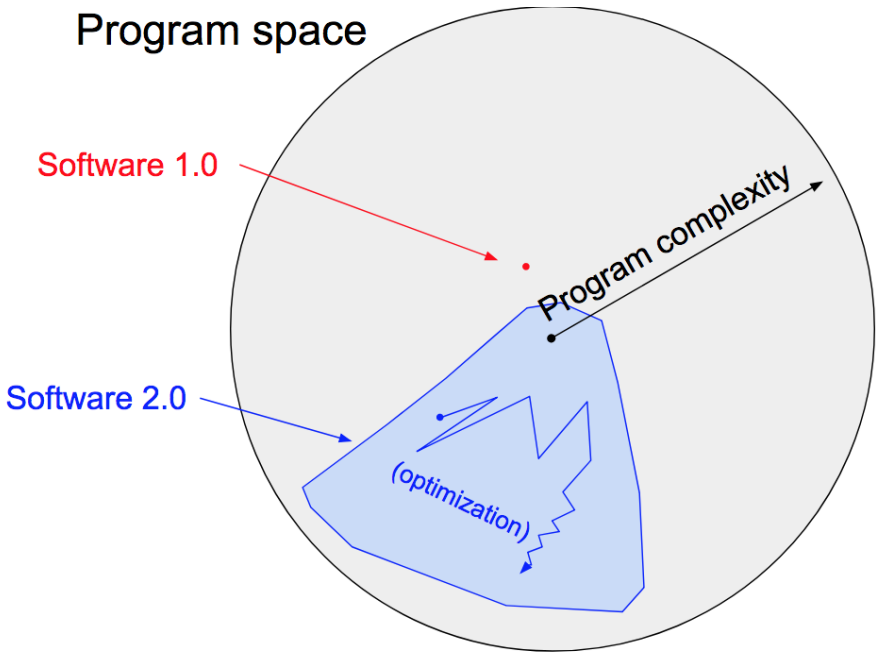





Kristoffer L Nielbo
kln@cas.dk
knielbo.github.io

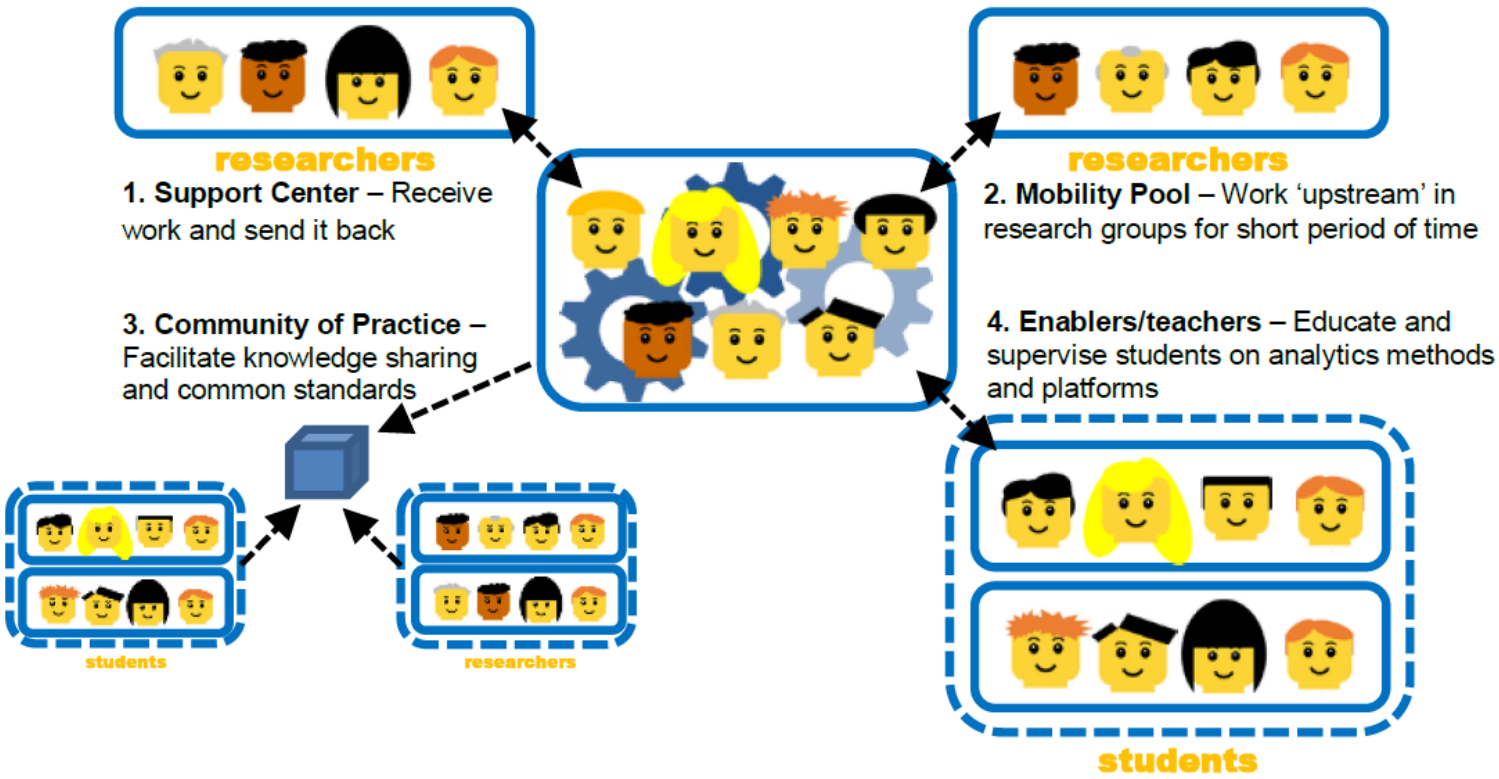



# Udvikling af computerbaseret forskning






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




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**



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

## 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 $\rightarrow$ 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

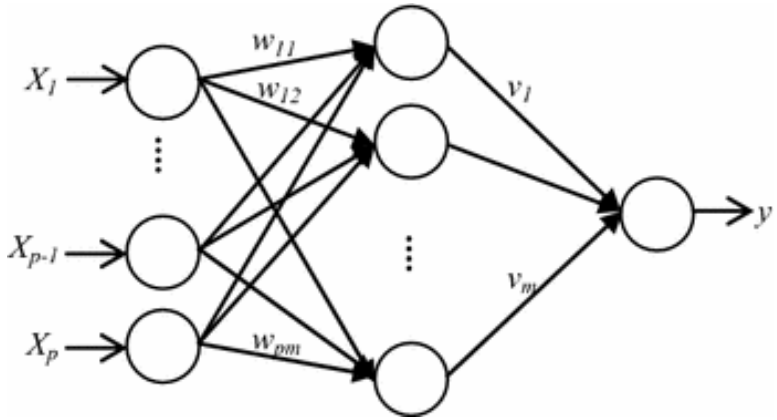
Problemer som maskinlæring løser:

– maskinlæring som mål for et forskningsproblem

– maskinlæring som middel, der automatiserer resourcekrævende hjælpeopgaver

# Afmystificering af neurale netværk

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.

# NFSG Parser

```
%%%xml
<p rend="firstIndent">
<persName key="pe35">Schiller</persName>
siger i Fortalen:
<seg type="comStart" n="com46" />
den høieste Nydelse
er
<seg type="com" n="com208">Gemyttets</seg>
Frihed under alle Kræfternes levende Spil
<seg n="com46" type="comEnd" />
, og er dette saa, da maa dette Drama
næsten heelt igiennem unde os den høieste Nydelse;
thi uagtet den stærkeste
Sysselsættelse af enhver
<seg type="com" n="com47">Kraft</seg>
, uagtet den høie
Interesse, vi føle for de
fremtrædende Figurer,
see vi dog deres Undergang med samme Rolighed, som de
selv.
</p>
```

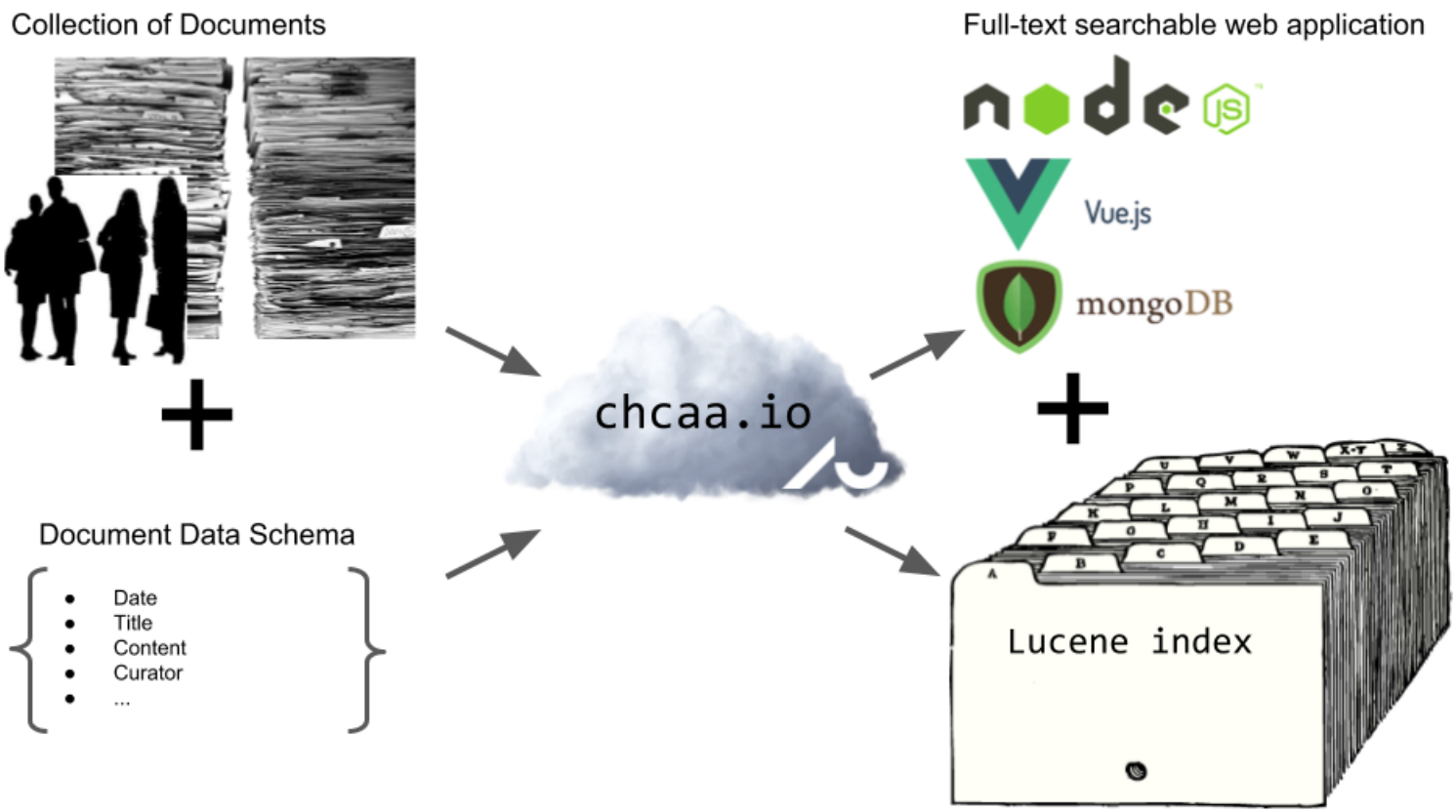
"Schiller siger i Fortalen: den høieste Nydelse er Gemyttets Frihed under alle Kræfternes levende Spil , og er dette saa, da maa dette Drama næsten heelt igiennem unde os den høieste Nydelse; thi uagtet den stærkeste Sysselsættelse af enhver Kraft, uagtet den høie Interesse, vi føle for de fremtrædende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



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

# NFSG DB

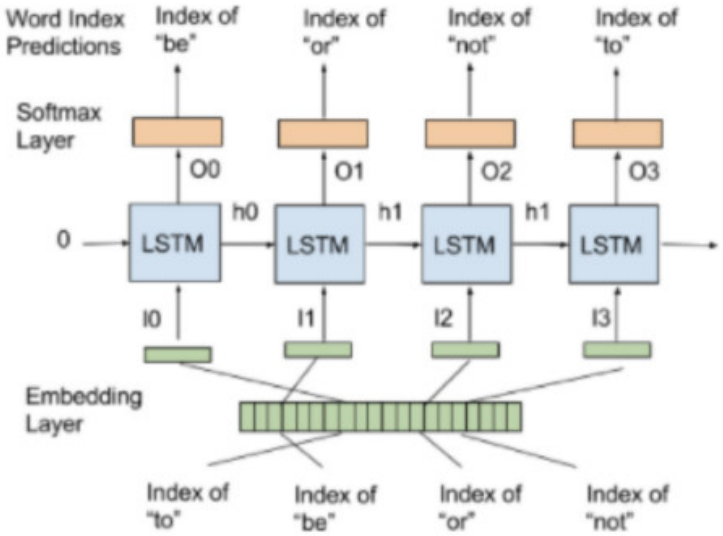
`NFSG Parser` er en del af `NFSG DB`, der vil gøre *Grundtvigs Værker* lettere tilgængelige for computerbaseret forskning. Parseren tilbyder pt. også subkorpusgeneration via eksisterende metadata.




`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

$\rightarrow$

er

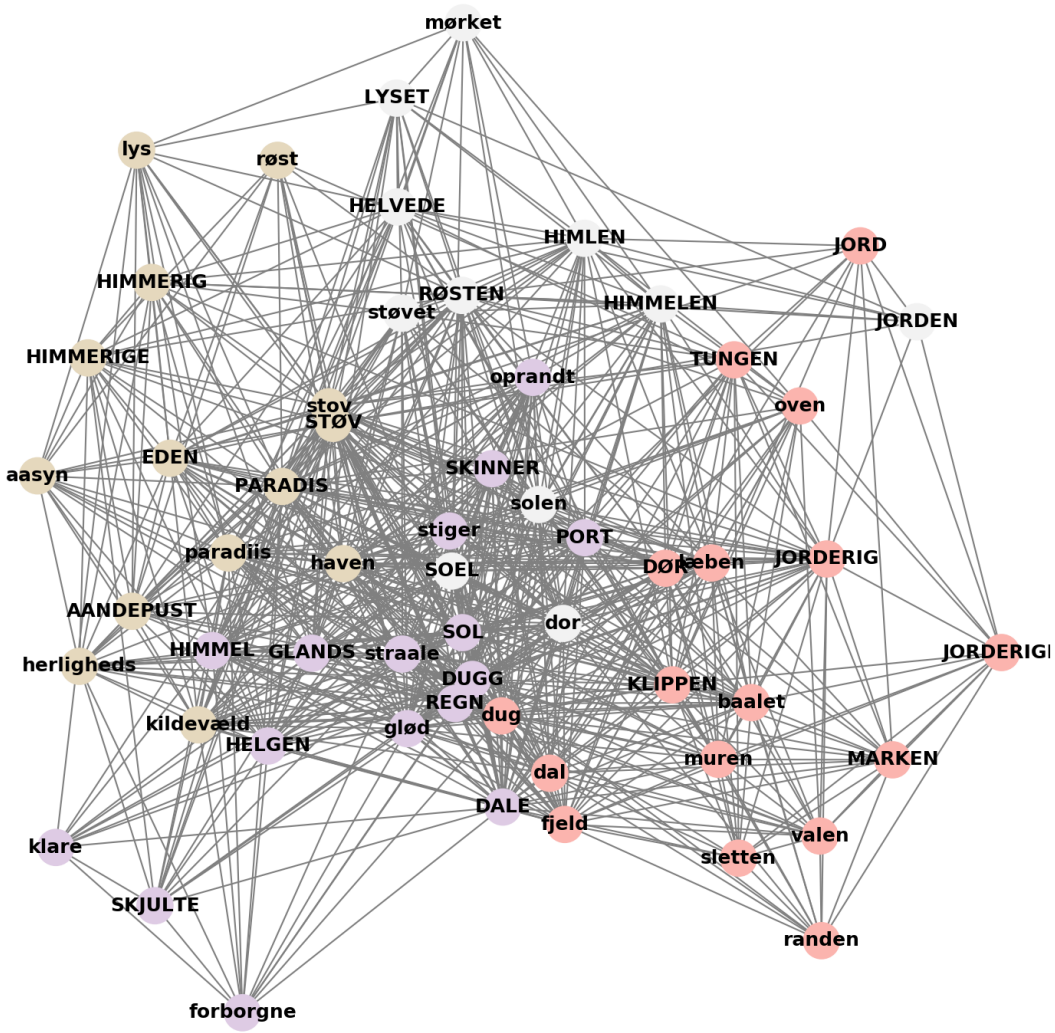
den høieste Nydelse er Gemyttets Frihed $[x_1, x_2, x_3, x_4, x_5]$

$\rightarrow$

under alle Kræfternes levende Spil

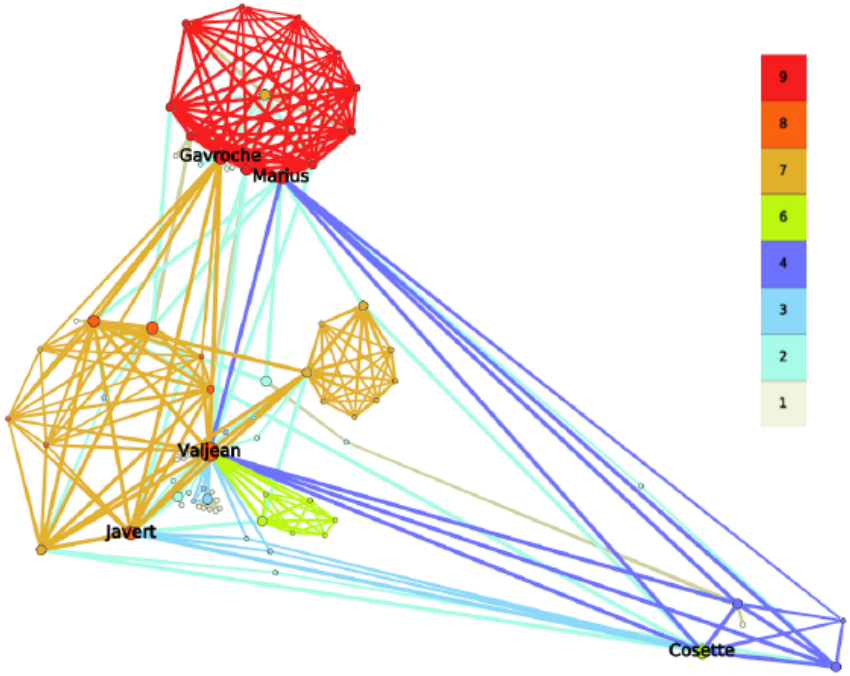
Efterfølgende kan vi bruge repræsentationerne til at løse natursprogsopgaver, fx. spørgsmål-svar, ordassociationer og tekstgenerering

# NFSG Concept Nucleus




Concept Nucleus er den første applikation baseret på `QA Modellen`

# NFSG Hyper-graf




Decompostion of the co-appearance network Les Miserables, 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

---

Abello J. & F. Queyroi (2014) Network decomposition into fixed points of degree peeling. Social Network Ananlysis and Mining

```
if questions:
    try:
        answer()
    except RunTimeError:
        pass
    else:
        print("thank you")
```



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



## TAK

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

slides: http://knielbo.github.io/files/kln_nfsgsandbox.pdf

**& tak til**

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