IN4080 – 2022 FALL

NATURAL LANGUAGE PROCESSING

Jan Tore Lønning

Today

- Part 1: Course overview
  - What is this course about?
  - How will it be organized?

- Part 2: ”Looking at data”:
  - Descriptive statistics
  - Some language data

Name game

- Computational Linguistics
  - Traditional name, stresses interdisciplinarity
- Natural Language Processing
  - Computer science/AI/NLP
  - ”Natural language” a CS term
- Language Technology
  - Newer term, emphasize applicability
  - LT today is not SciFi (AI), but part of everyday app(lication)s
- The terms have different historical roots
  - Today: NLP = Computational Linguistics, restricted to written language
  - LT = NLP + speech (No speech in this course)

Megatrends

Natural Language Processing

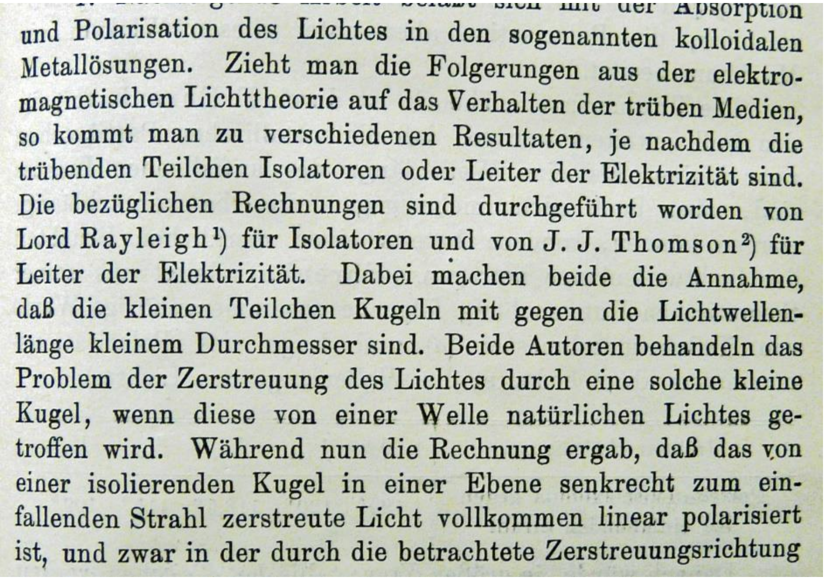
Artificial Intelligence AI

- Machine learning
  - Deep learning

"Data science"
Big data
(WWW)

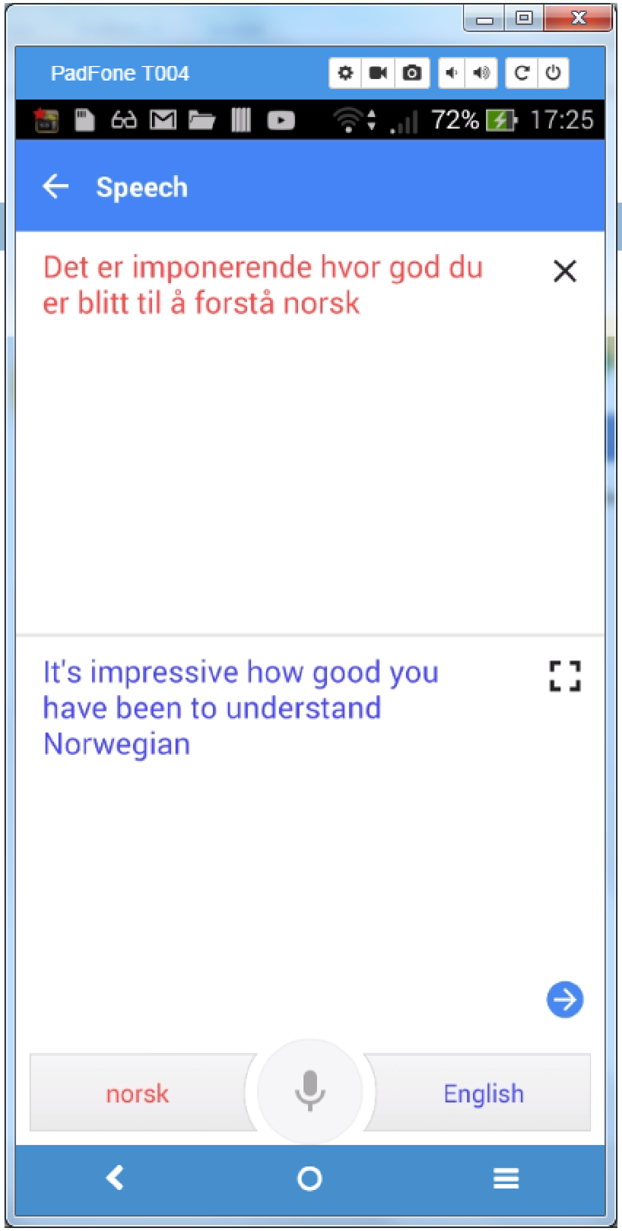
Language technology: examples

1. Speech ↔ text



1. Die vorliegende Arbeit befaßt sich mit der Absorption und Polarisation des Lichtes in den sogenannten kolloidalen Metallösungen. Zieht man die Folgerungen aus der elektromagnetischen Lichttheorie auf das Verhalten der trüben Medien, so kommt man zu verschiedenen Resultaten, je nachdem die trübenden Teilchen Isolatoren oder Leiter der Elektrizität sind. Die bezüglichen Rechnungen sind durchgeführt worden von Lord Rayleigh[1]) für Isolatoren und von J. J. Thomson[2]) für Leiter der Elektrizität. Dabei machen beide die Annahme, daß die kleinen Teilchen Kugeln mit gegen die Lichtwellenlänge kleinem Durchmesser sind. Beide Autoren behandeln das Problem der Zerstreuung des Lichtes durch eine solche kleine Kugel, wenn diese von einer Welle natürlichen Lichtes getroffen wird. Während nun die Rechnung ergab, daß das von einer isolierenden Kugel in einer Ebene senkrecht zum einfallenden Strahl zerstreute Licht vollkommen linear polarisiert ist, und zwar in der durch die betrachtete Zerstreuungsrichtung

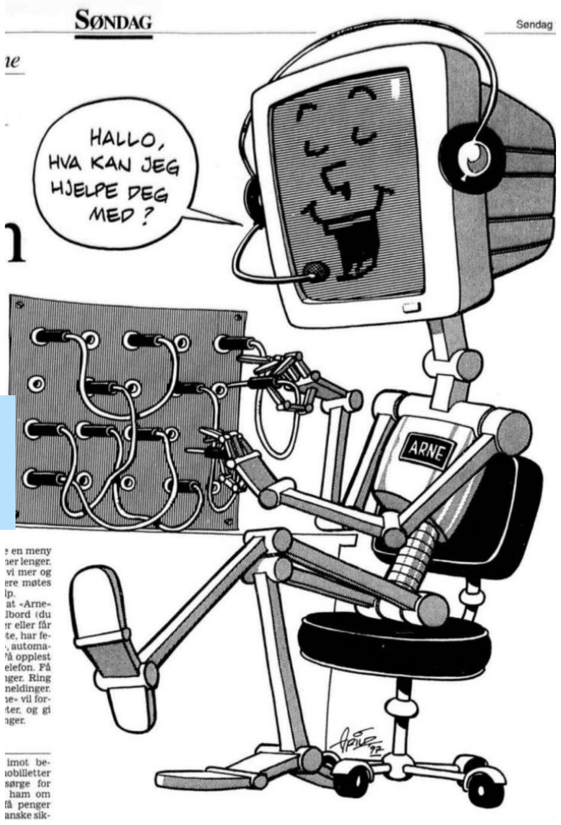
2. Machine translation



Det er imponerende hvor god du er blitt til å forstå norsk

It's impressive how good you have been to understand Norwegian

3. Dialogue systems



Aftenpoſten Søndag 1. juni 1997

4. Sentiment analysis and opinion mining

Sentiment/opinion mining:

- Do consumers appreciate more sugar in the soda?
- Do (my core voters) like my last Twitter outburst?
- How will the stock prices develop?
- Is there a danger of a revolt in country X?

- Personalization:
  - Adds
  - News

5. Text analytics

- Goal, example IBM's Watson system:
- Read medical papers + records:
  - Propose diagnoses
  - Propose treatments

- Similarly in other domains:
  - Oil & Gas
  - Legal domain:

+

6. NLP applications – more examples

- Intelligence
- Surveillance:
  - How does NSA manage to read all those e-mails?
- User content moderation
- Election influence

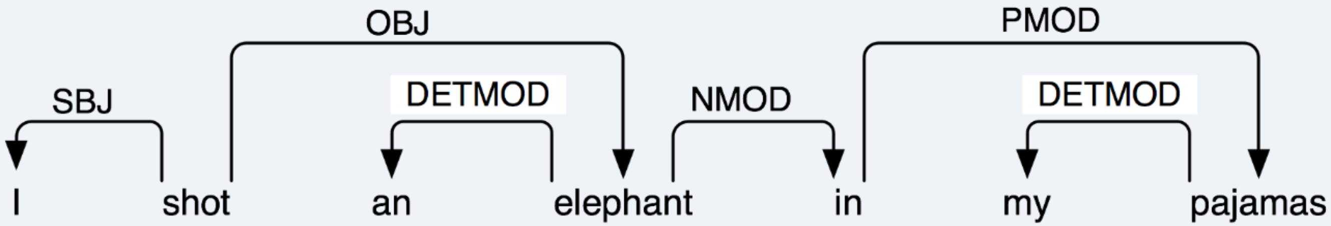
Cambridge Analytica

What?

What

- https://www.uio.no/studier/emner/matnat/ifi/IN4080/index.html
- Consider some of the main tasks in a bottom-up NLP system
- Consider some of the main methods
  - Starting with the simpler ones
  - Machine-learning, experiments
- Dialogue systems (October-November)
  - "…in-depth knowledge of at least one [NLP] application…"
- Ethics in NLP (November)

Some steps when processing text

<table>
<tr><th>Split into sentences</th><th>Obama says he didn't fear for 'democracy' when running against McCain, Romney.</th></tr>
<tr><td>Tokenize (normalize)</td><td>| Obama | says | he | did| not | fear | for | ‘ | democracy | ‘ | when | running | against | McCain | , | Romney | .</td></tr>
<tr><td>Tag</td><td>Obama_N says_V he_PN did_V not_ADV fear_V …</td></tr>
<tr><td>Lemmatize</td><td>Says_V → say_V, did_V → do_V, running_V → run_V …</td></tr>
<tr><td>Parsing (dependency)</td><td>
</td></tr>
<tr><td>Coreference resolution</td><td>Obama says he did not …..</td></tr>
<tr><td>Semantic relation detect.</td><td>Fear(Obama, Democracy) Run_against(Obama, McCain),..</td></tr>
<tr><td>Negation detection</td><td>… did not fear … → Not(Fear(Obama, Democracy))</td></tr>
</table>

The two cultures (up to the 1980s)

Symbolic

- 1956 →
- Sub-cultures
  1. AI (NLU)
     - McCarthy, Minsky → SHRDLU ('72)
  2. Formal Linguistics/Logic
     - Chomsky
       - automata, formal grammars
     - + Logic in the 80s
     - LFG, HPSG
  3. Discourse, pragmatics

Stochastic

- Information theory, 1940s
- Statistics
- Electrical engineering
- Signal processing

Trends the last 30 years

- 1990s: combining the cultures
  - methods from speech adopted by NLP
    - division of labor between methods
    - stochastic components in symbolic models, e.g., statistical parsing
  - (larger) text corpora
  - Jurafsky and Martin, SLP, 2000

- 2000s:
  - More and more machine learning in NLP, at all levels
  - Examples and corpora
  - Rethinking the curriculum and the order in which it is taught
  - J&M, 2. ed, 2008

Example:
machine translation systems that are trained on earlier translated texts

Last 10 years

- 2010s Deep learning
  - ML with multi-layered Neural Networks
  - Revolution, in particular for
    - Image recognition
    - Speech
  - Entered all parts of NLP
    - Key: "Word embeddings"

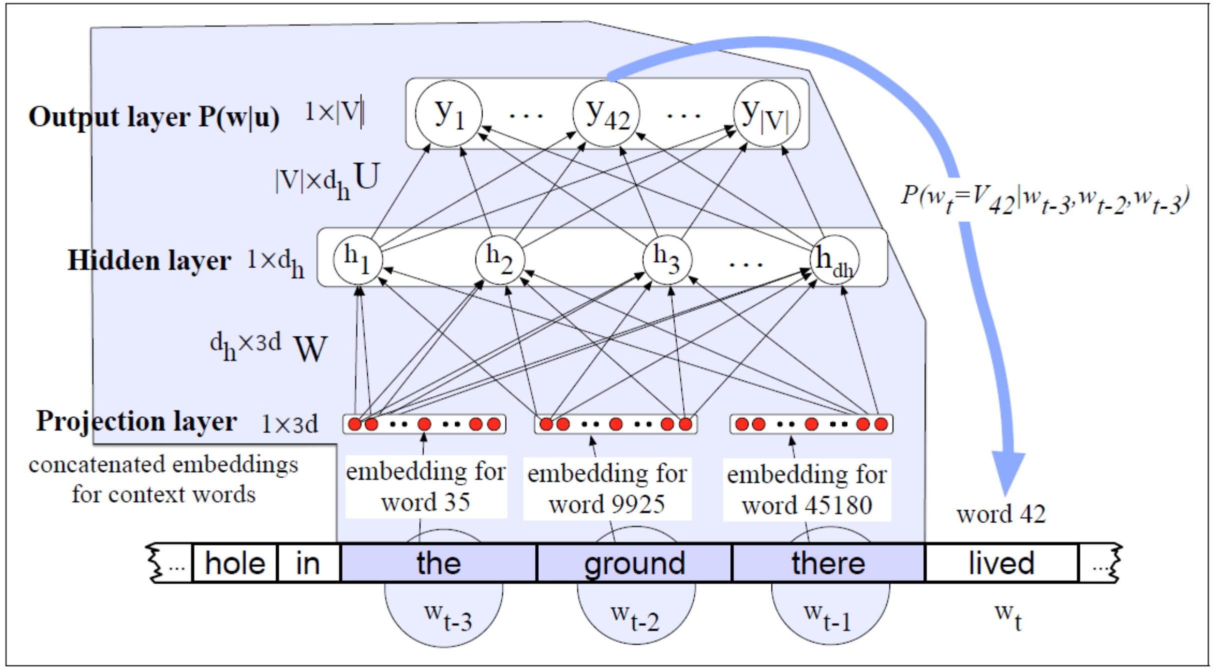
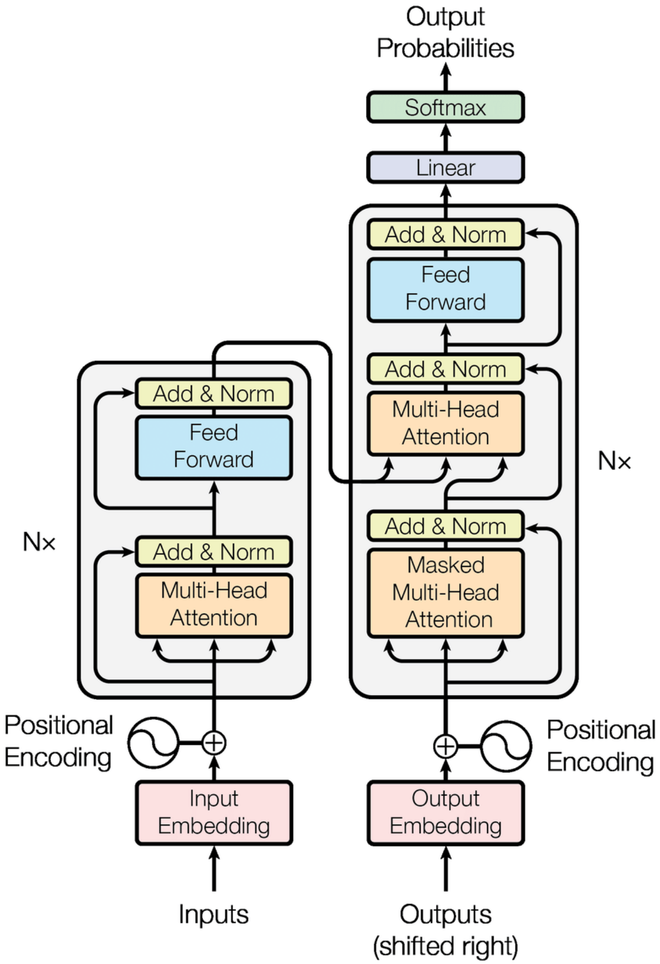


Figure 9.1 A simplified view of a feedforward neural language model moving through a text. At each time step *t* the network takes the 3 context words, converts each to a *d*-dimensional embedding, and concatenates the 3 embeddings together to get the 1 × *Nd* unit input layer *x* for the network.

Currently: Transformers

- BERT, 2018

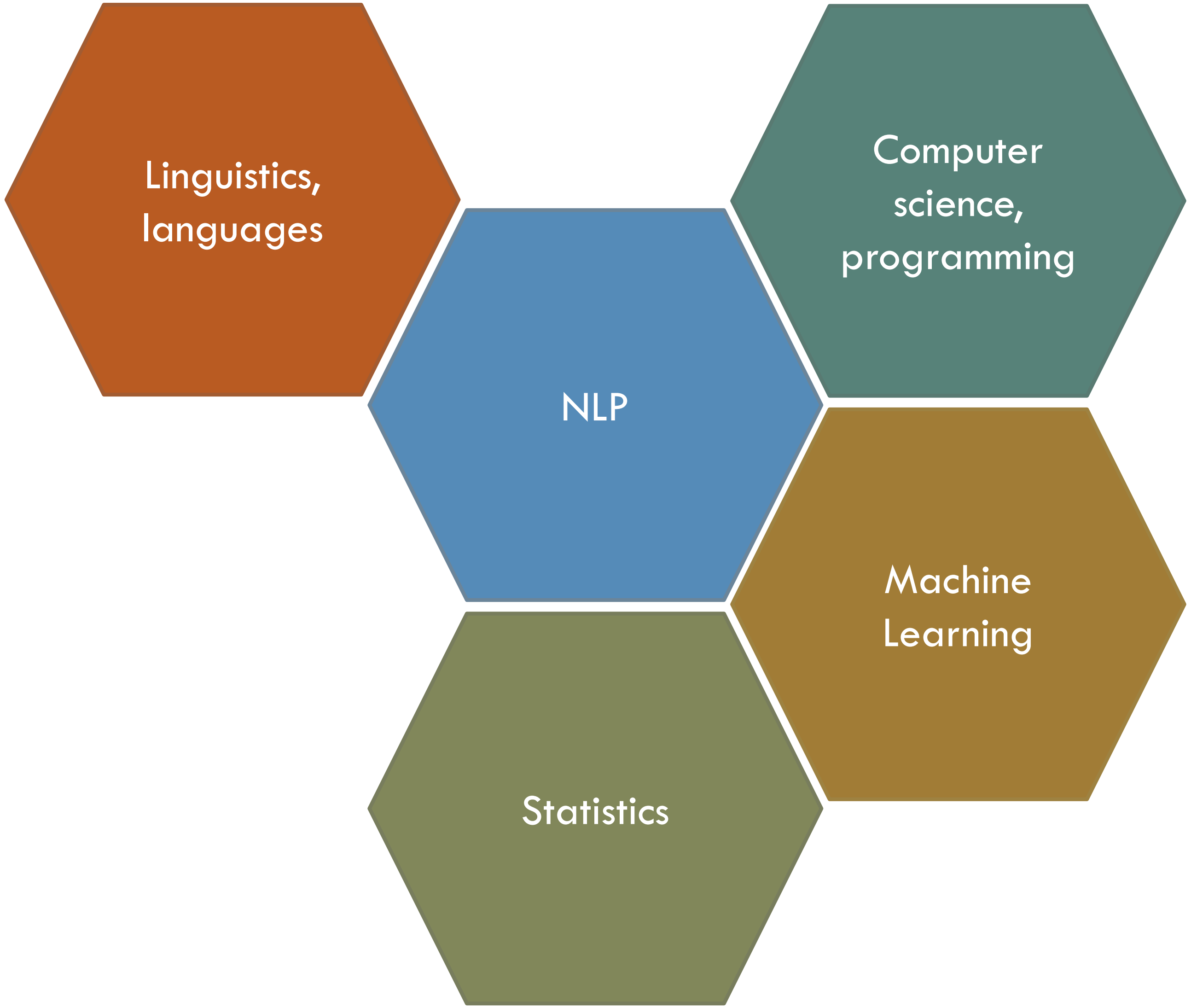


DL and IN4080

- Should we jump directly to deep learning?
- We will initially focus on simpler models.
- Most tasks are independent of learning algorithm, and can be easier understood using simpler models
- For several tasks, traditional ML is still compatible

- We will eventually consider some of the principles of deep learning in NLP
- We will, however, in the practical part stick to what can be done on a PC/CPU
  - This limits what we can do
- "IN5550 Neural Methods in NLP“, spring 2023
  - Deep learning in NLP
  - HPC

NLP is based on



Linguistics, languages

NLP

Computer science, programming

Machine Learning

Statistics

Why statistics and probability in NLP?

1. "Choose the best"
(=the most probable given the available information)

- *bank* (Eng.) can translate to b.o. *bank* or *bredd* in No.
  - Which should we choose?
  - What if we know the context is "*river bank*"?
- *bank* can be Verb or Noun,
  - which tag should we choose?
  - What if the context is *they bank the money* ?
- A sentence may be ambiguous:
  - What is the most probable parse of the sentence?

Use of probabilities and statistics, ctd.:

2. In constructing models from examples (ML):
   - What is the best (most likely) model given these examples?

3. Evaluation:
   - Model1 is performing slightly better than model 2 (78.4 vs. 73.2), can we conclude that model 1 is better?
   - How large test corpus do we need?

How?

Syllabus (online)

- Lecture slides put on the web
- Jurafsky and Martin, *Speech and Language Processing, 3.ed.*
  - In progress, edition of Dec 2021/Jan 2022
- Articles from the web
- In addition
  - Some selections from
    - S. Bird, E. Klein and E. Loper: *Natural Language Processing with Python*
    - available on the web, python 3 ed.
  - Probabilities and statistics (some book or)
    - https://www.openintro.org/book/os/

Computational “Work Bench”

- Python
  - Well-suited for text
  - Readable, structured code
- Python-packages
- Widely used for "data science" and machine learning
- General scientific computing
  - NumPy
  - Scipy (stats)
  - Matplotlib
  - Pandas

- Machine learning:
  - scikit-learn
  - Pytorch
- NLP
  - NLTK
  - spaCy
  - gensim
- and more
- (https://www.anaconda.com/open-source)

Challenges for a master's course like this

- You have different backgrounds:
  - Some are familiar with some NLP from e.g., IN2110
  - Some are familiar with simple probabilities and statistics, some are not
  - Some are familiar with Machine Learning
  - Some are familiar with Language and linguistics
- For you:
  - Some of the parts will be a pure repetition for you
  - In other parts, you may experience a steep learning curve
  - Concentrate on the parts with which you are less familiar

Schedule

- Lectures: Thursdays 12.15-14
  - Jan Tore Lønning,
  - Pierre Lison
  - Room Smalltalk
  - Screencasts distributed after lecture
- Lab sessions:
  - Tuesdays 12.15-14
  - Huiling You
  - Room: Sed
  - No screencast
- 3 mandatory assignments (oblig.s)
  - Weeks 38, 41, 45
- Written exam
  - Wednesday 13 December 9AM

First weeks

- Observe:
  - The week starts on Thursday
  - Ends the following Wednesday

Week	Thursday Lecture	Tuesday Group/Lab
1	25.8 Introduction, Language data, Descr. Stat.	30.8: Set-up, Python, NLTK, Frequencies, Plots, Language data
2	2.9 Words, Morphology, Tokenization, Tagged text	7.9:
3	9.9 Lecture: Text classification	16.9: Lab

- Tutorial on probabilities: when?

Background knowledge

- Please fill in:
- https://nettskjema.no/a/279606