

# IN1140: Introduksjon til språkteknologi

## *Forelesning #1*

Lilja Øvrelid

Universitetet i Oslo

17 august 2020



- ▶ Introduksjon
- ▶ Hva er språkteknologi?
- ▶ Hva er IN1140?
- ▶ Praktiske detaljer
  - ▶ Grupper
  - ▶ Obliger
  - ▶ Lærebøker
  - ▶ Kontakt
  - ▶ m.m.



- ▶ Smittevern er hovedprioritet
  - ▶ Hold avstand
  - ▶ God håndhygiene
  - ▶ Bli hjemme hvis du er syk
  - ▶ **Gode digitale alternativer**
- ▶ Oppdatert informasjon om UiO og korona:  
<https://www.uio.no/om/hms/korona/>

## Forelesere

- ▶ **Lilja Øvrelid** (liljao@ifi.uio.no)
- ▶ **Samia Touileb** (samiat@ifi.uio.no)
- ▶ Fra språkteknologigruppa (LTG)

## Forelesere

- ▶ **Lilja Øvrelid** (liljao@ifi.uio.no)
- ▶ **Samia Touileb** (samiat@ifi.uio.no)
- ▶ Fra språkteknologigruppa (LTG)

## Gruppelærere

- ▶ **Annika Willoch Olstad** (annikaol@student.matnat.uio.no)
- ▶ **Fredrik Aas Andreassen** (98fredrik@live.no)

## Forelesere

- ▶ **Lilja Øvrelid** (liljao@ifi.uio.no)
- ▶ **Samia Touileb** (samiat@ifi.uio.no)
- ▶ Fra språkteknologigruppa (LTG)

## Gruppelærere

- ▶ **Annika Willoch Olstad** (annikaol@student.matnat.uio.no)
- ▶ **Fredrik Aas Andreassen** (98fredrik@live.no)

## Kursassistent

- ▶ **Petter Mæhlum** (pettemae@ifi.uio.no)

## Tid & sted

- ▶ Gruppe 1: fre. 12:15–14:00, Datastue Assembler.
- ▶ Gruppe 2: ons. 14:15–16:00, Datastue Modula.
- ▶ Gruppe 3: tors. 10:15–12:00, IT-seminarrom Sed.
- ▶ Gruppe 4: fre. 14:15–16:00, **Digital gruppe** i Zoom (åpen for alle)
- ▶ Forelesninger: man. 14:15–16:00 i **Simula** (Ole-Johan Dahls hus / IFI).
- ▶ **NB!** Første gruppetime er onsdag **26 august**

- ▶ Tar opp **screencast** for hver forelesning (lyd + foiler).
- ▶ Ment som et supplement, for repetisjon.
- ▶ Alternativ ved sykdom.





- ▶ **Gruppetimene:** Gruppelærerene er der for å hjelpe og veilede.
- ▶ **Padlet** (diskusjonsforum)
- ▶ **in1140-hjelp [at] ifi.uio.no:** Felles adresse til fag-/gruppelærere.



- ▶ Husk å sjekke **UiO-eposten** din og **beskjedlisten** på semestersiden.
- ▶ <http://www.uio.no/studier/emner/matnat/ifi/IN1140/h20/>



# Hva er språkteknologi?

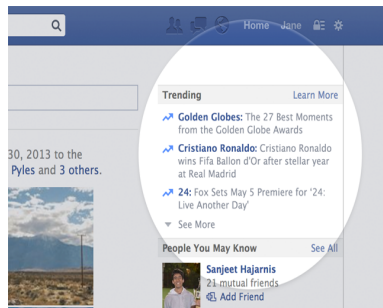
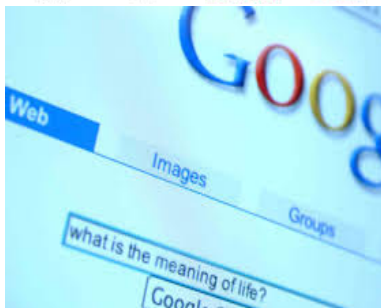
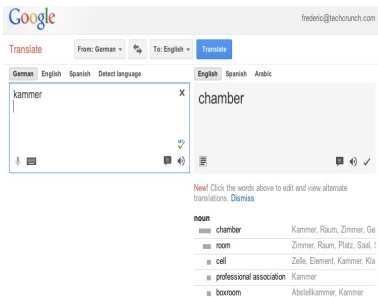


- ▶ Mål: å få datamaskiner til å 'forstå' naturlige språk.
- ▶ Aka:
  - ▶ computational linguistics (datalogistikk)
  - ▶ language technology
  - ▶ language engineering
  - ▶ **natural language processing (NLP)**





# Eksempler på språkteknologi?



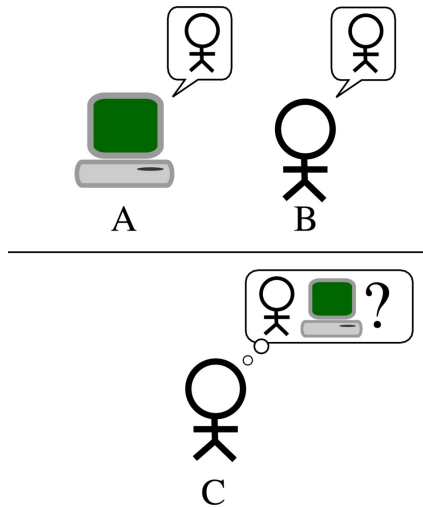
## NLP er et tverrfaglig felt

- ▶ Lingvistikk
- ▶ Informatikk
- ▶ Statistikk
- ▶ Maskinlæring
- ▶ Logikk, Filosofi, Psykologi, ...



- ▶ Del av det bredere feltet **kunstig intelligens** (AI).

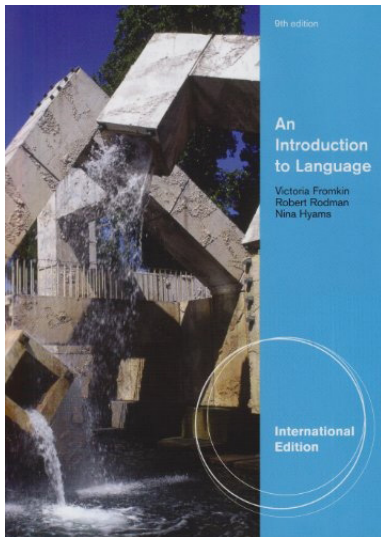
- ▶ Alan Turing i 1950:
- ▶ *I propose to consider the question, 'Can machines think?'*
- ▶ Definisjonsspørsmål. Skulle avgjøres ved Turingtesten.



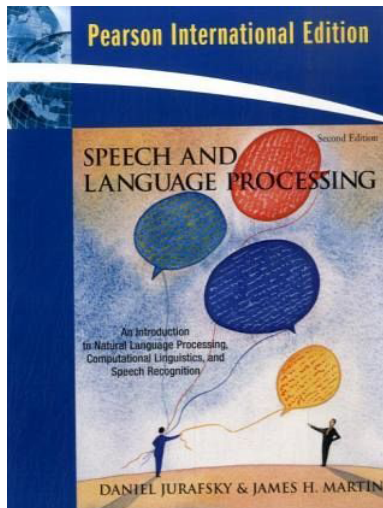
- ▶ Stoffet vi dekker i IN1140 tar også for seg stoff fra flere ulike felt.
- ▶ Innføring i **lingvistikk**,
- ▶ grunnleggende **sannsynlighetsregning**,
- ▶ **programmering**, og
- ▶ språkteknologiske **anvendelser**.



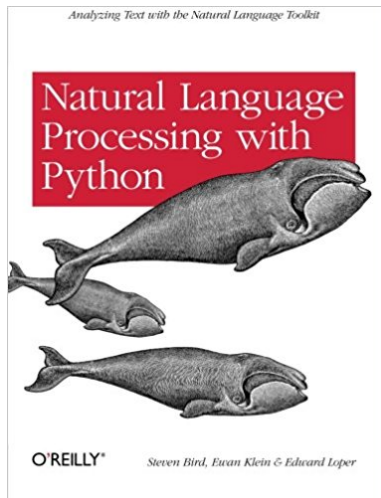
- ▶ Stoffet vi dekker i IN1140 tar også for seg stoff fra flere ulike felt.
- ▶ Innføring i **lingvistikk**,
- ▶ grunnleggende **sannsynlighetsregning**,
- ▶ **programmering**, og
- ▶ språkteknologiske **anvendelser**.
  
- ▶ Gjør deg godt rustet for flere viderekommende emner, f.eks
  - ▶ IN2110 – Språkteknologiske metoder
  - ▶ IN3050 – Kunstig intelligens og maskinlæring
  - ▶ IN3120 – Søketeknologi
  - ▶ og mange flere!



- ▶ *An Introduction to Language*  
av Fromkin, Rodman & Hyams
- ▶ Utvalgte deler (ca 5 kapitler)



- ▶ *Speech and Language Processing* av Jurafsky & Martin
- ▶ Utvalgte deler
- ▶ Gratis nettbok:  
<https://web.stanford.edu/~jurafsky/slp3/>



- ▶ *Natural Language Processing with Python*,  
av Bird, Klein & Loper
- ▶ Oppdatert for Python 3 og NLTK 3 (Natural Language Toolkit)
- ▶ Utvalgte deler
- ▶ Gratis nettbok:  
<http://www.nltk.org/book/>

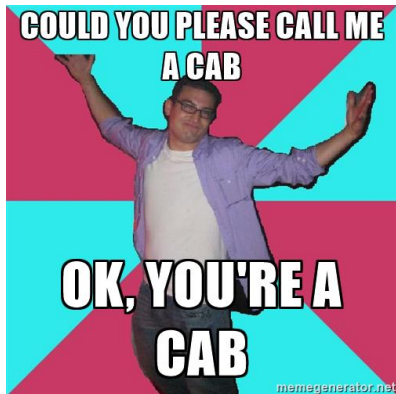
- ▶ Programmering lærer dere først og fremst i **IN1000**, ikke **IN1140**.
- ▶ **Forelesningene** i IN1140 kommer til fokusere på **teori**.
- ▶ Samtidig ønsker vi å implementere stoffet i praksis, i Python.
- ▶ **Implementasjon** blir fokus på **gruppene** og **innleveringene**.
- ▶ Kræsjkurs i Python-programmering på de første gruppetimene.
- ▶ Viktig med en del egeninnsats i starten for å henge med.



# Hvorfor er språkforståelse utfordrende?



- ▶ Språk er vagt, ulike tolkninger mulig.
- ▶ **Flertydighet** overalt.
- ▶ Gir kompakt kommunikasjon:
- ▶ Samme uttrykk kan brukes i ulike kontekster.



- ▶ Flertydighetene er stort sett usynlige for oss, vi finner den intenderte tolkningen nærmest ubevisst.
- ▶ For **maskiner** er det motsatt: **lett** å finne alle mulige tolkninger, men **vanskelig** å se hvilken som er riktig.

# Eksempel: Flertydighet på ordnivå



- ▶ Norsk: *rett*.
- ▶ Engelsk: ?
- ▶ Flertydig ift betydning + ordklasse (verb, subst., adj., adv.).
- ▶ Vi trenger kontekst for å avgjøre.

# Eksempel: Flertydighet på ordnivå



- ▶ Norsk: *rett*.
- ▶ Engelsk: *?*
- ▶ Flertydig ift betydning + ordklasse (verb, subst., adj., adv.).
- ▶ Vi trenger kontekst for å avgjøre.

avgrenset av en *rett* linje tvers over kanalen

*straight*

Hva er *rett* svar?

*correct, right*

lovbestemt *rett* til innsyn

*right*

Denne *rett* avsa enstemmig dom i saken 4. juli 1980

*court*

Norsk *rett* tilpasses EUs regelverk

*law*

Vennligst *rett* disse prøvene!

*grade, correct*

Det bar *rett* i fengsel

*directly, straight*

De spiste en deilig *rett* av grønnsaker.

*meal, dish*

han var *rett* utenfor, *rett* nå

*just*

Slikt skjer *rett* som det er.

må omskrives




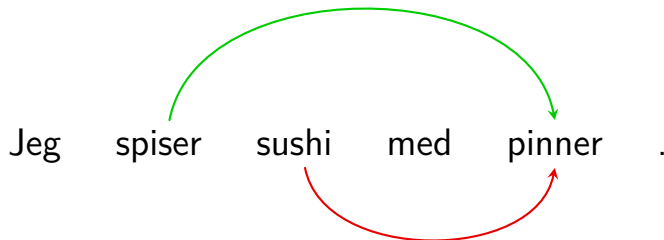
We gave the monkeys<sub>1</sub> the bananas<sub>2</sub>  
... because they<sub>1</sub> were hungry.  
... because they<sub>2</sub> were ripe.



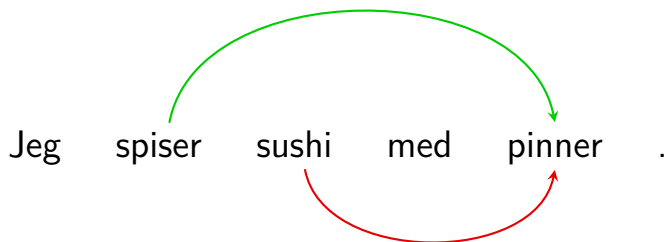
Jeg spiser sushi med pinner .

Jeg spiser sushi med pinner .





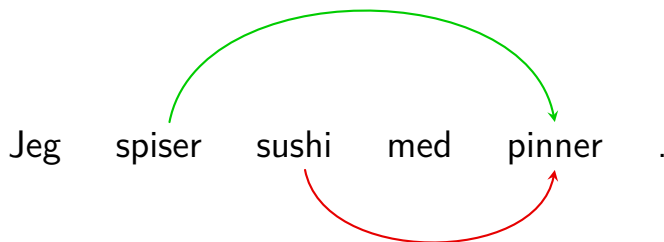
Jeg spiser sushi med pinner .



The diagram illustrates the ambiguity of the sentence "Jeg spiser sushi med pinner ." by using two arcs. A green arc connects the word "spiser" to "pinner", indicating the interpretation "I eat sushi with sticks (pinner)". A red arc connects the word "sushi" to "pinner", indicating the interpretation "I eat sushi with salmon (laks)".

Jeg spiser sushi med laks .

Jeg spiser sushi med pinner .

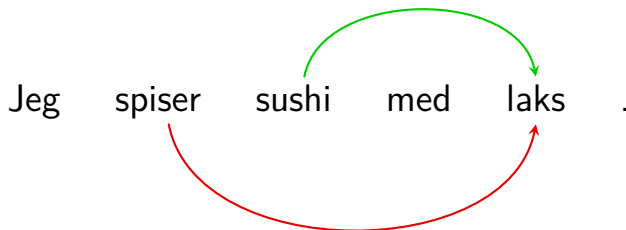
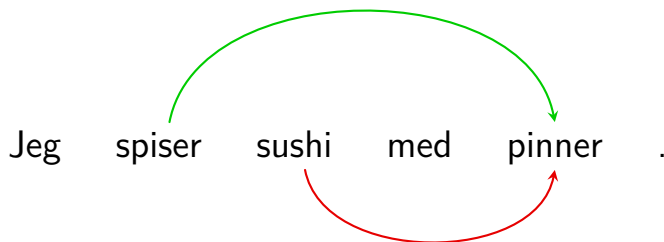


The diagram illustrates the ambiguity of the sentence "Jeg spiser sushi med pinner ." by showing two possible syntactic structures. A green arc connects the word "spiser" to "pinner", representing the structure "Jeg spiser sushi med pinner" (I eat sushi with sticks). A red arc connects the word "sushi" to "pinner", representing the structure "Jeg spiser sushi med pinner" (I eat sushi with sticks).

Jeg spiser sushi med laks .



The diagram illustrates the ambiguity of the sentence "Jeg spiser sushi med laks ." by showing two possible syntactic structures. A green arc connects the word "spiser" to "laks", representing the structure "Jeg spiser sushi med laks" (I eat sushi with salmon).



The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard. The mental abilities of a four-year-old that we take for granted — recognizing a face, lifting a pencil, walking across a room, answering a question — in fact solve some of the hardest engineering problems ever conceived. . . As the new generation of intelligent devices appears, it will be the stock analysts and petrochemical engineers and parole board members who are in danger of being replaced by machines. The gardeners, receptionists, and cooks are secure in their jobs for decades to come.

Steven Pinker, *The language instinct*



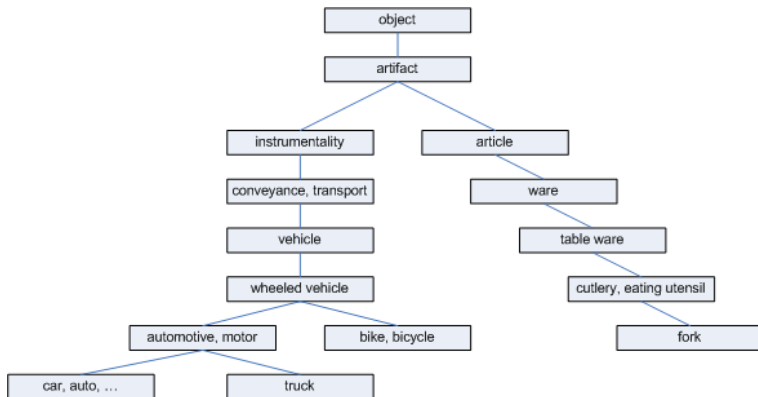
The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard. The mental abilities of a four-year-old that we take for granted — recognizing a face, lifting a pencil, walking across a room, answering a question — in fact solve some of the hardest engineering problems ever conceived. . . As the new generation of intelligent devices appears, it will be the stock analysts and petrochemical engineers and parole board members who are in danger of being replaced by machines. The gardeners, receptionists, and cooks are secure in their jobs for decades to come.

Steven Pinker, *The language instinct*

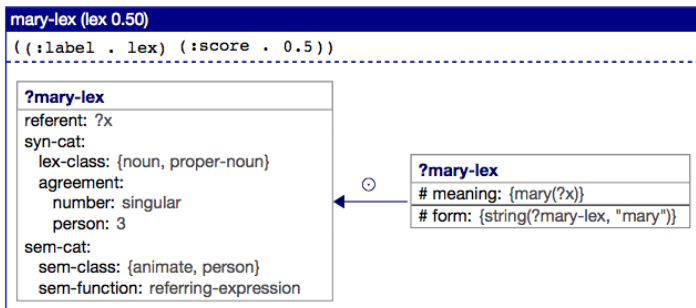
- En robot som bretter et håndkle (videoen er 50 ganger normal hastighet): <http://www.youtube.com/watch?v=gy5g33S0Gzo>

- ▶ Vi **menesker** tolker språklige uttrykk basert på delt **bakgrunnskunnskap** og gjensidige **forventninger** i en gitt **kontekst**.
- ▶ **Språkforståelse** handler mye om **entydiggjøring**.
- ▶ Språkteknologi, og **IN1140**, handler i stor grad om strategier for hvordan maskiner kan takle dette.

→ 2000-tallet: manuelt utformede regler og leksikon



→ 2000-tallet: manuelt utformede regler og leksikon

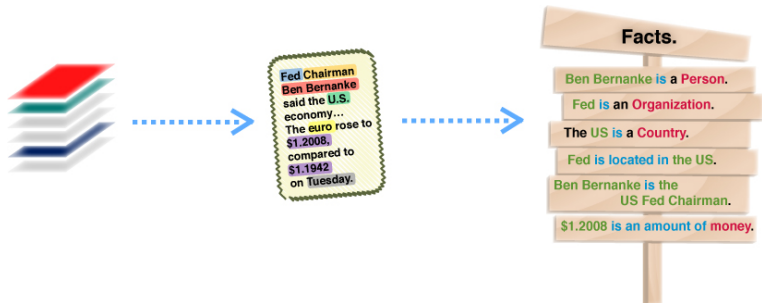


- ▶ 2000-tallet →: empirisk revolusjon
- ▶ **Maskinlæring**
  - ▶ Datamaskiner kan lære fra data: fange opp mønstre og generalisere til nye eksempler

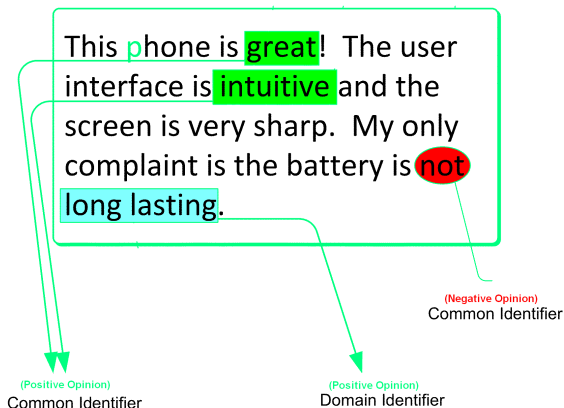




## Informasjonsekstraksjon

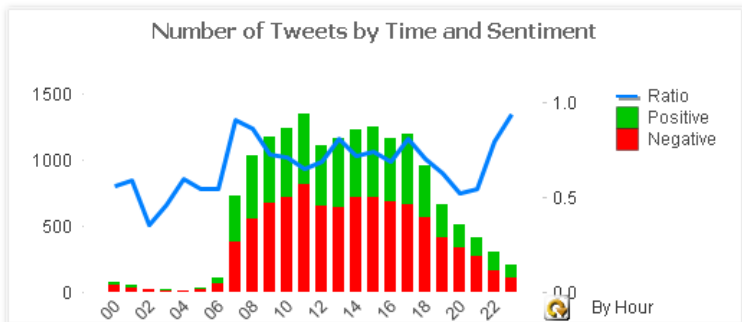


**Sentiment Analyse:** automatisk analyse av subjektivt språk

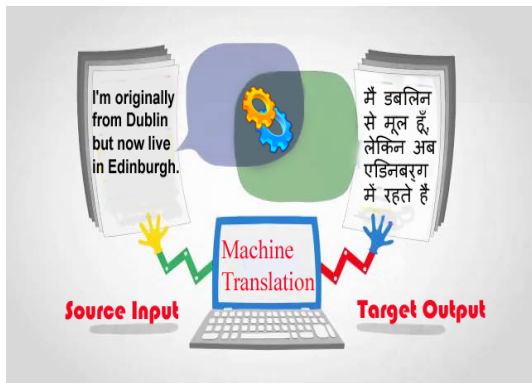




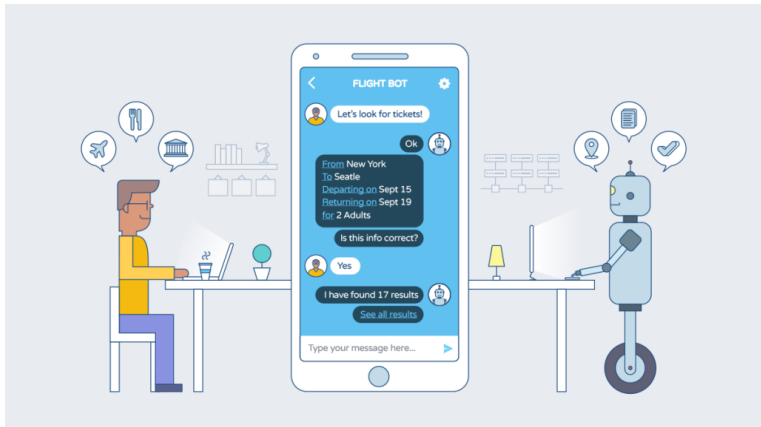
# Medieovervåkning



## Maskinoversettelse



## Dialogsystemer





- ▶ 2 obliger.
- ▶ Oblig 1 har to deler ( $a + b$ ).
- ▶ Oblig 2 har to deler ( $a + b$ ).
- ▶ Dvs. 4 innleveringer tilsammen:  $1a + 1b$ ,  $2a + 2b$ .
- ▶ Begge obligene må bestås for å kunne ta eksamen.
- ▶ Ingen omlevering.

- ▶ 2 obliger.
- ▶ Oblig 1 har to deler ( $a + b$ ).
- ▶ Oblig 2 har to deler ( $a + b$ ).
- ▶ Dvs. 4 innleveringer tilsammen:  $1a + 1b$ ,  $2a + 2b$ .
- ▶ Begge obligene må bestås for å kunne ta eksamen.
- ▶ Ingen omlevering.

## Poengsystemet

- ▶ Man kan oppnå opptil 100 poeng per innlevering
- ▶ For å bestå en oblig kreves minst 100 poeng (av 200 mulige).
- ▶ Eksempel:
  - ▶ 37 poeng på 1a
  - ▶ 68 poeng på 1b
  - ▶ = 105 poeng på oblig 1 (= bestått).

- ▶ **Absolutte frister:**
- ▶ Utsettes *kun* ved egenmelding (opptil 3 dager) eller legeerklæring.
- ▶ **Kopiering/plagiat godtas ikke.** Sett deg inn i reglene.
- ▶ Husk at hvis du distribuerer løsningsforslaget ditt på nett (f.eks via Github), kan du bidra til juks. Styr unna.
- ▶ Benytt deg av gruppeundervisningen, og planlegg tiden din.
- ▶ Tidsregnskap:
  - ▶ Arbeidsinnsats (minimum):  $37,5 / 3 = 12,5$  timer
  - ▶ Etter forelesning+gruppe: 9,5 timer
- ▶ **Konkurranse:** den/de som får flest poeng tilsammen på obligene gjennom semesteret får en premie (overraskelse)!

- ▶ Oppgaver til gruppetimene fokuserer ofte på teori
- ▶ Nytt av året: en del ekstraoppgaver som **Trix**-oppgaver

## List comprehension 1

hest2020 in1140 list\_comprehension python

Gitt følgende liste (a), lag ei ny liste (b) hvor alle verdiene i (a) er ganget med to ved å bruke list comprehension.

`a = [1, 4, 6, 8, 9]`

Hvordan løste du oppgaven?

På egenhånd

Med hjelp

☒ Ikke løst

[Vis løsningsforslag >](#)



- ▶ Skriftlig (digital) hjemmeksamen på fire timer
  - ▶ 25 november kl. 09:00
- ▶ Pensumlitteratur + forelesningsnotater
- ▶ **NB! Ikke en programmeringseksamen.**
- ▶ Fokus på teoretiske konsepter.

- ▶ Emnesiden: timeplan, pensum, lesehenvvisninger, beskjeder etc.
- ▶ Lesehenvvisninger: forbered deg til forelesning
- ▶ Still spørsmål
- ▶ Gruppetimer:
  - ▶ forbered deg
  - ▶ delta aktivt
  - ▶ gjør oppgaver (også de ikke-obligatoriske!)
- ▶ Benytt deg av medstudentene dine

- ▶ Hva lærer jeg i IN1140?
- ▶ Hva er lingvistikk?
- ▶ Språkteknologiske komponenter
- ▶ Metoder