

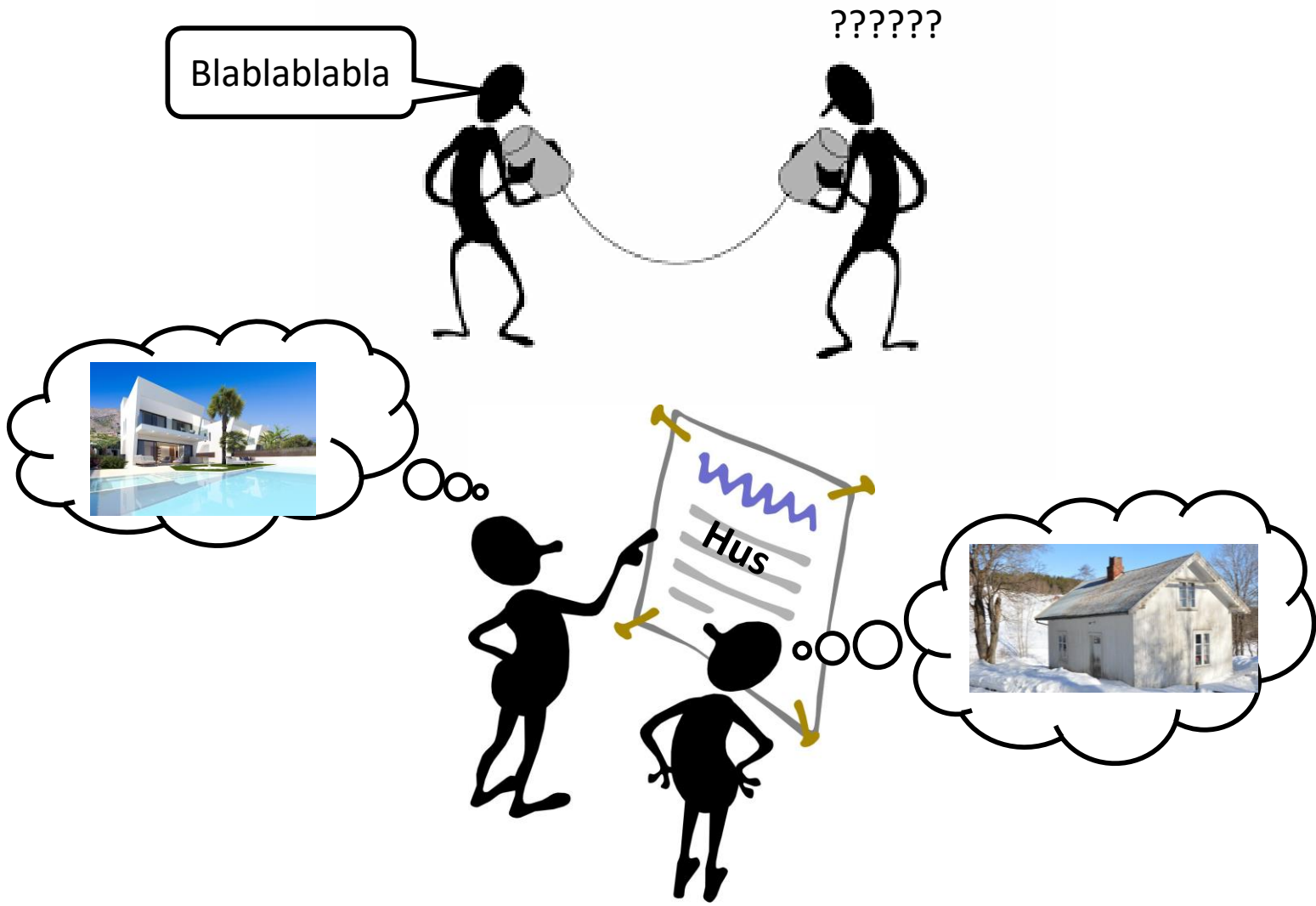
# Semantisk teknologi

– hva er grunnlaget for at digitale systemer kan forstå meningen med ordene – og vi kan utvikle maskinlæring?

*Bernt A. Bremdal, UiT*

*1. 12.2020*

# Å forstå og bli forstått



# Språket vårt

- SYNTAKS: Handler om **språklige termer og ord** og om **strukturelle relasjoner** som definerer lovlige sekvenser.
- SEMANTIKK: Handler om **mening** – det vil si **relasjoner mellom språklige termer og de faktiske ting** de refererer til i den virkelige verden
- PRAGMATIKK handler om **relasjoner mellom bruker i brukernes miljø og de språklige formene** som brukeren anvender for å referere til i den virkelige verden



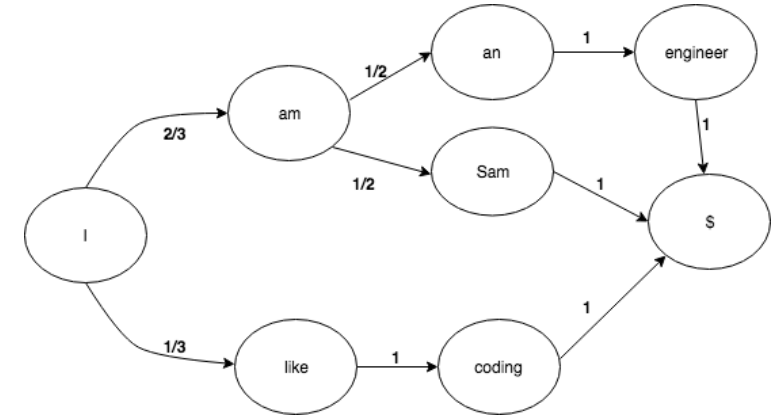
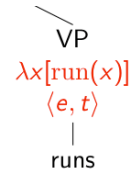
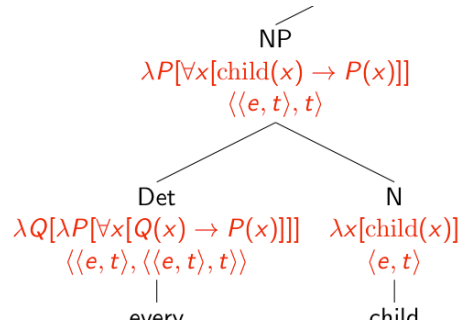
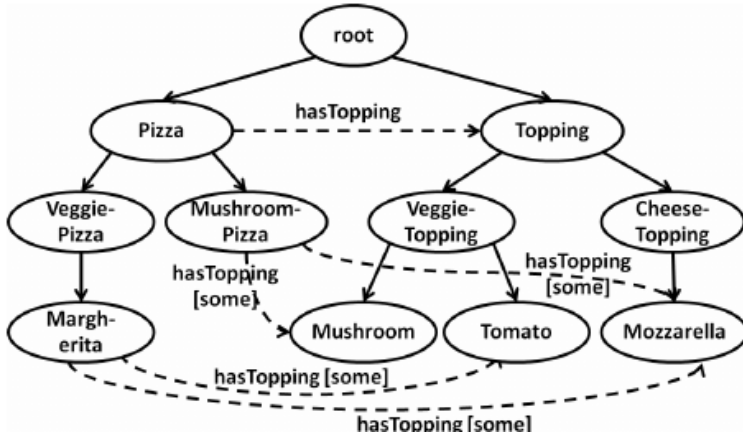


Bisken

Bestevenn

Rufsen

Kjøter

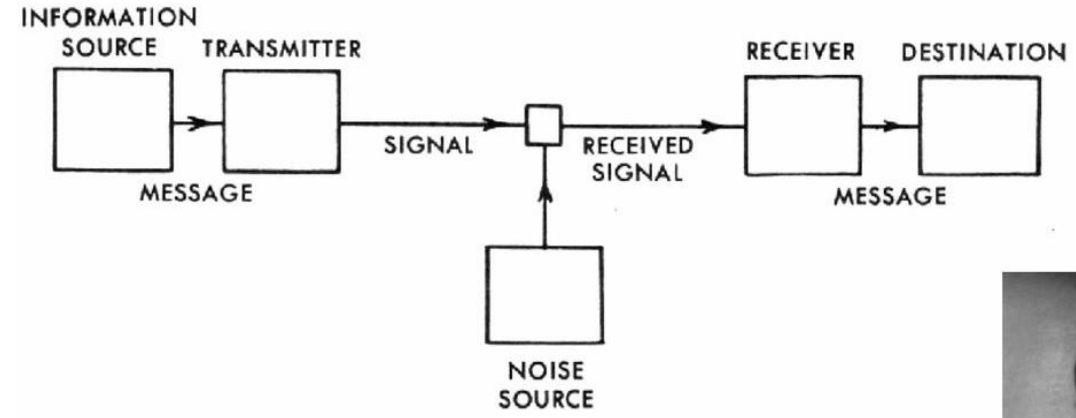


# Forskning rundt naturlig språk

Strukturell analyse, ontologier, statistikk, ontologier, statistikk, ontologier, statistikk.....

# Information theory

The Mathematical Theory of Communication



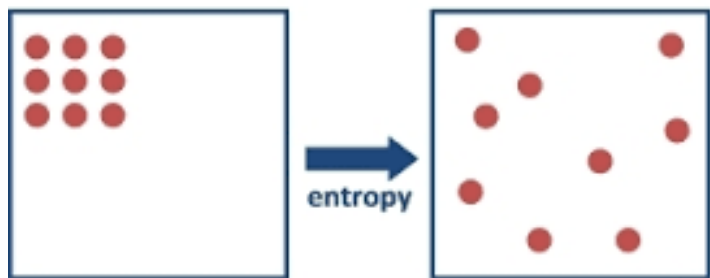
Schematic diagram of a general communication system.

→ Shannon information can only be lost, never gained

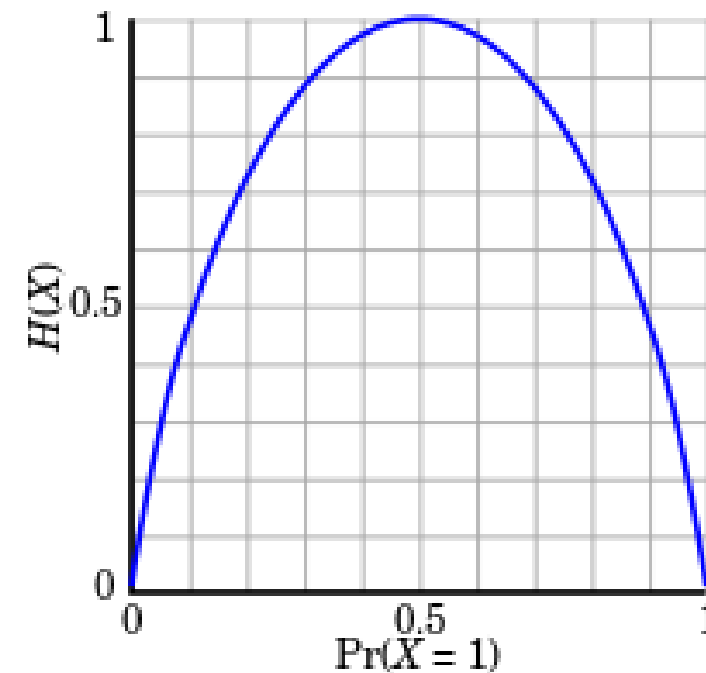
Shannon (1948). *Bell Systems Technical Journal* 27: 379.



Claude Shannon  
1916–2001



$$H = - \sum_{i=1}^n p_i \log p_i$$

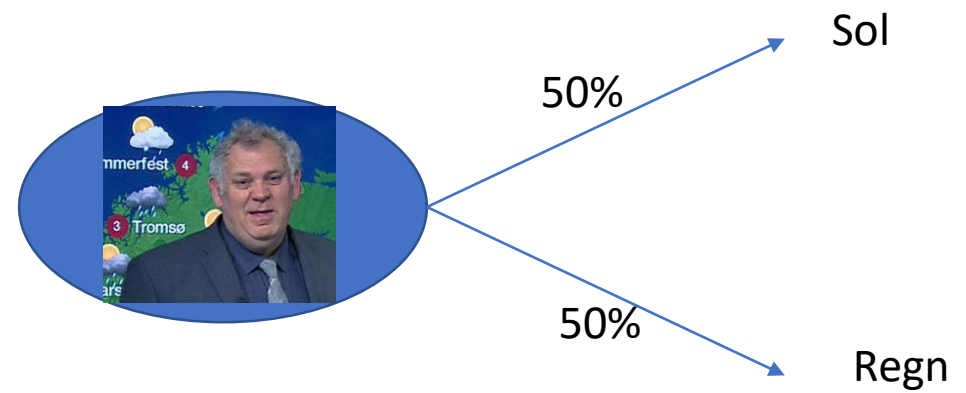
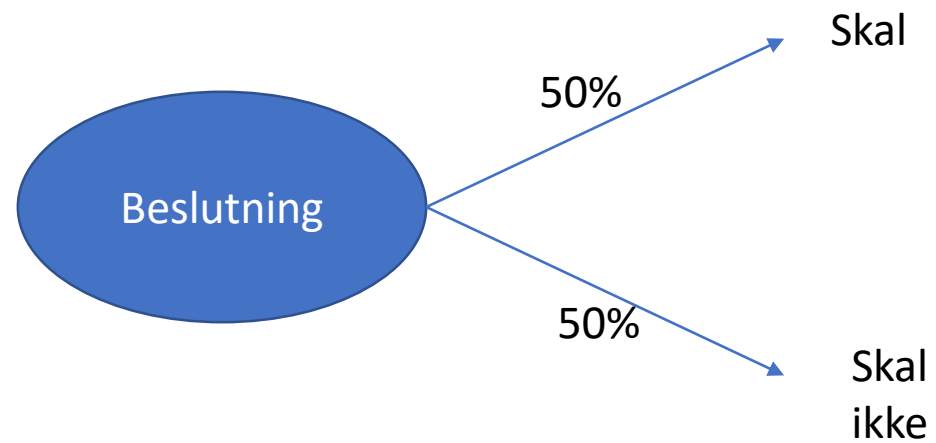



## Entropi viktig i:

- Økonomisk beslutningsteori
- Signalbehandling
- De fleste typer maskinlæring



# Maksimal entropy

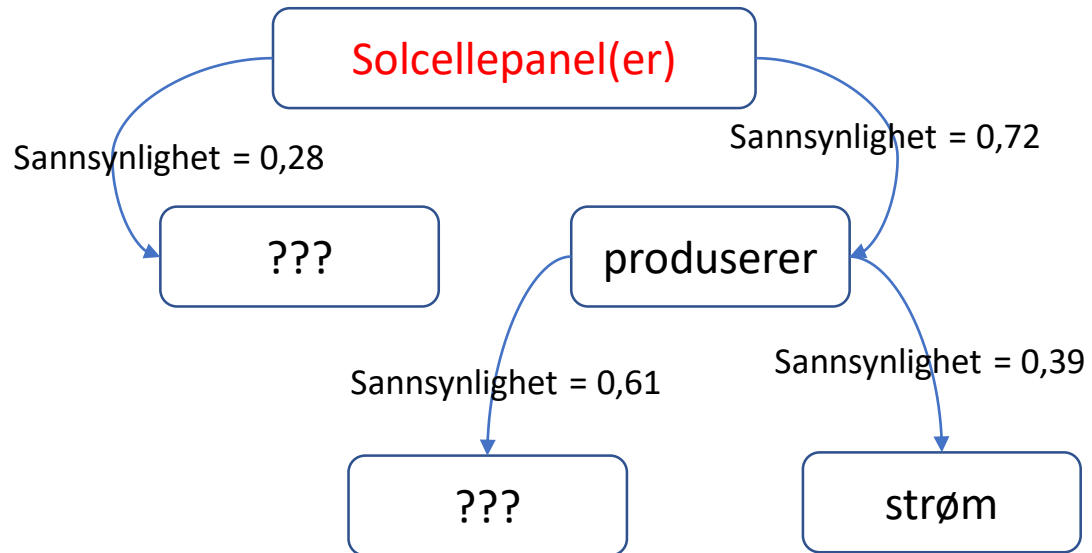




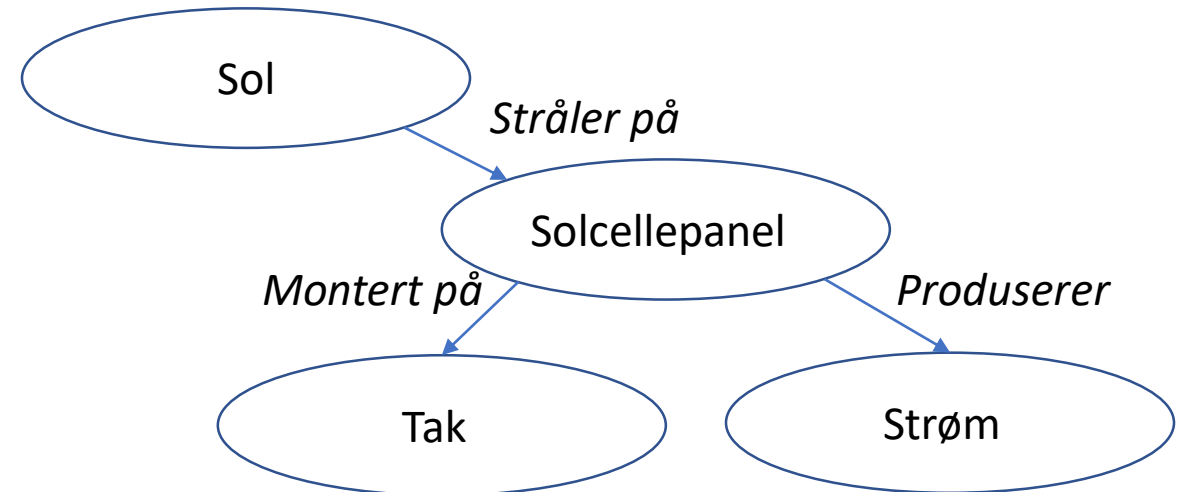
# Gjett neste ord

- Jeg er for.....
- Isen er veldig .....
- Donald .....
- Diego .....
- Solcellepaneler produserer .....
- Du må ha pepper og ..... på kjøttakene
- Pierre kommer fra Frankrike. Han snakker.....
- Tom er fotballspiller. På landslaget har han reddet mange skudd. Han er.....

# Solcellepaneler produserer .....



Statistisk tilnærming



Ontologisk tilnærming

Maskinlæringsteknikker

# Entropi og redundans

Fra Google

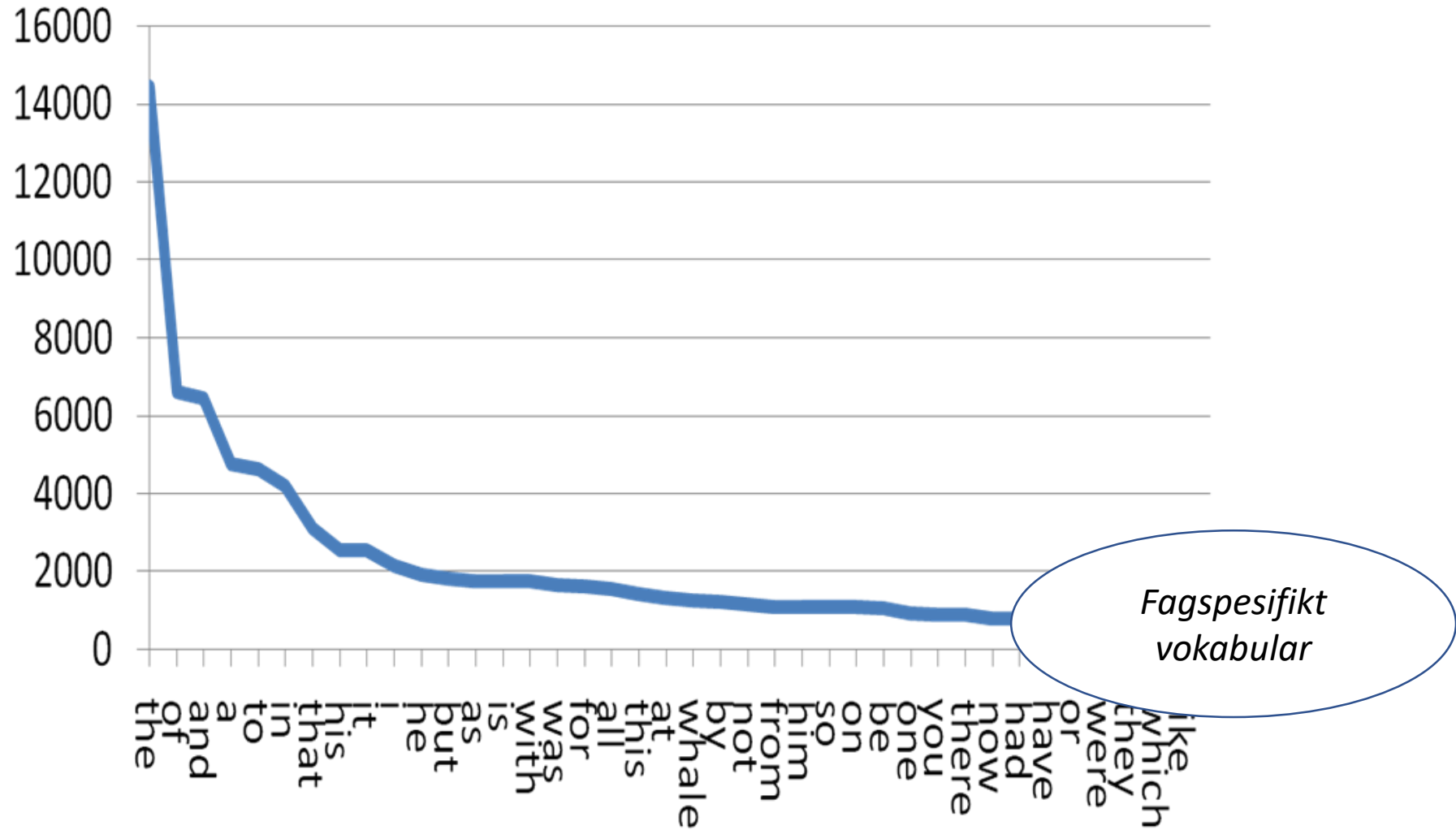
Søkeord	Antall treff
the	Omtrent 25 270 000 000
med	Omtrent 2 310 000 000
aus	Omtrent 2 620 000 000
Gahr-Støre	Omtrent 897 000
støtfanger	Omtrent 700 000
skolisse	Omtrent 244 000
reduksjonsventil	Omtrent 83 700

Grad av entropi bestemmer hvor vanskelig et kryssord kan være på et bestemt språk

The image shows a crossword puzzle grid. On the left side, there is a photograph of a Christmas tree decorated with lights and ornaments. The crossword grid is overlaid on the right side of the image. The grid contains several words and letters. A vertical column of cells is highlighted in yellow. This column contains the letters K, A, and N from top to bottom. Other words visible in the grid include 'KORSE', 'NORSK FUL', 'FOLK', 'KJØPE', 'JULENE DO-SITE', 'THORIA', 'LAKNET', 'TRØVKE', 'DE BLER FOR LITVA', 'DPRNE', 'TRESKET', 'BENNER-KALE', 'DEKKE', 'SPONE', 'DAVE-NISSE', 'LURENE DOLLAR', 'BLISE', 'LOTTAS', 'LOGGE', 'ALARM', 'SØKTEVE', 'K', 'SØTTOKK', 'VØR', 'PERE-NARE', 'LITTE', 'FISY', 'A', 'N', 'COMPACT DISC', 'REMAN-PLI', 'MORFIST', 'LANT', 'FUSSE'.

Fagspesifikke ord er ofte lange og har en lav ordrangering

# Zipf lov



# Kontekst

Kontekstinformasjon



```
graph TD; A(Kontekstinformasjon) --> B(fotballspiller); A --> C(landslaget); A --> D(reddet); A --> E(skudd);
```

Tom er **fotballspiller**. På **landslaget**  
har han **reddet** mange **skudd**. Han  
er.....

# Kommunikasjon innenfor og utenfor en kontekst

Language lean – concept strong



To geologer med samme bakgrunn og samme språk og samme arbeidsoppgaver har allerede en etablert kontekst. Lett å kommunisere.

Language rich – concept weak



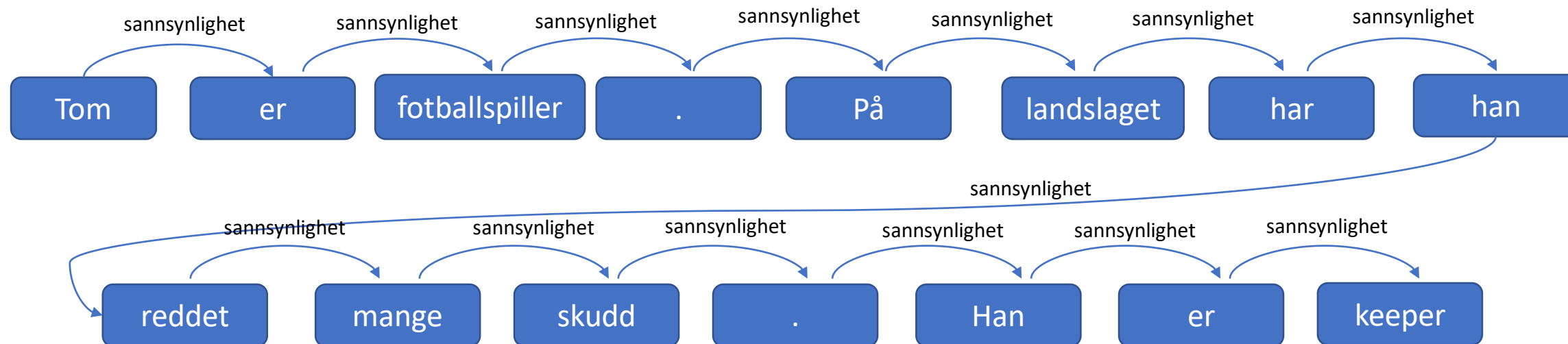
Journalisten må konstruere en kontekst for de uinnvidde. Mange ord må støtte hverandre, mange forklaringer, vanskelig kommunikasjon.

*Datamaskinen er uinnvidd*

# Takk til Intel, AMD, NVIDIA.....



Tom er fotballspiller. På landslaget har han reddet mange skudd. Han er.....



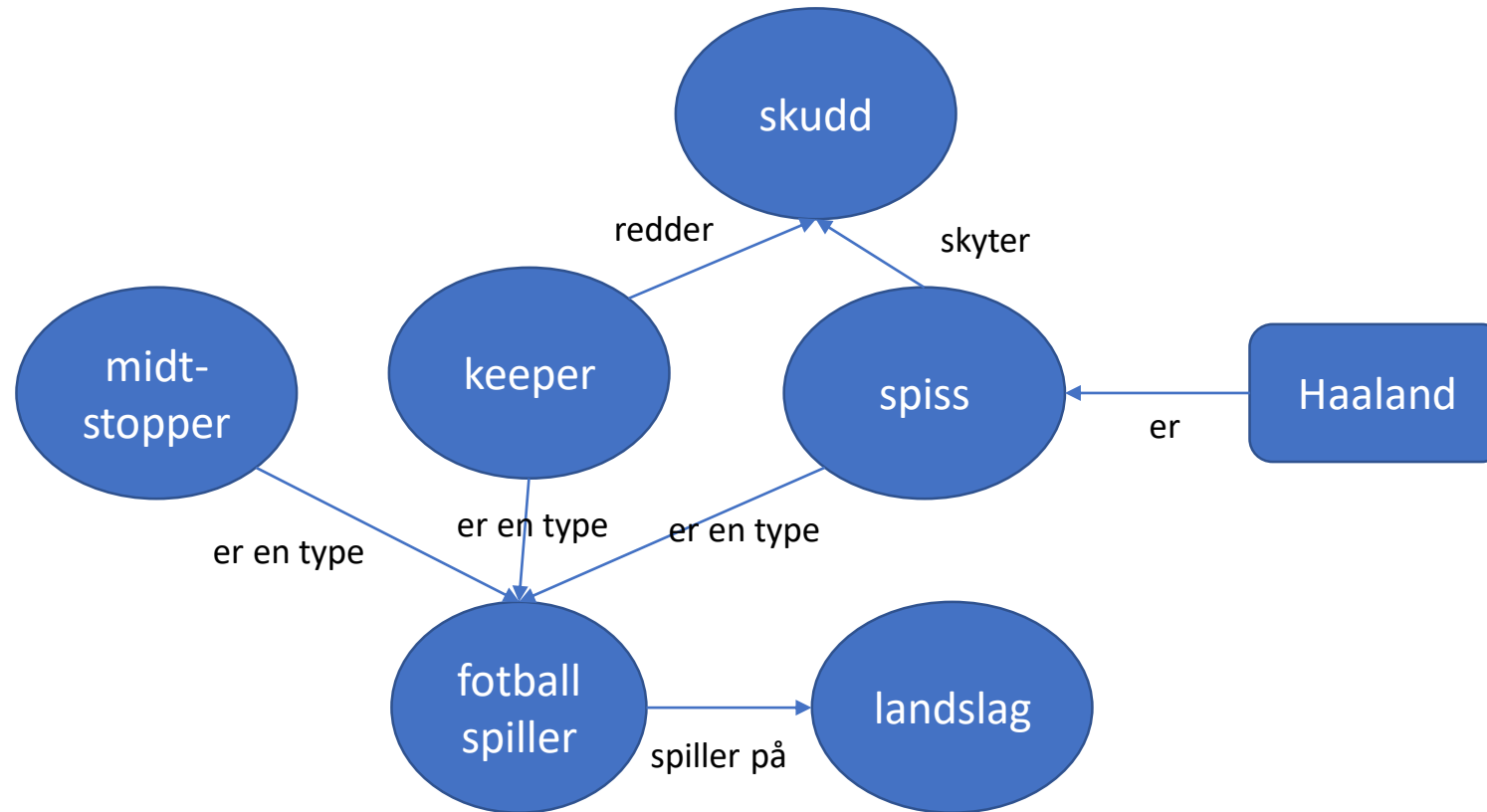
Det er mulig å beregne sannsynligheten for ulike ordrekkefølger med verktøy som Bayesiske nett, rekurrente nevrale nett og mye data

Beregningsmetoden er helt identisk med «pila» i sjakkmesterskap som vises på NRK





# En liten fotball ontologi



Bruk kunnskap når du har det – ontologisk engineering  
Mange ontologiske biblioteker for ulike fagområder finnes  
Neo4J er en veldig potent grafdatabase som kan holde og lære om nye begreper

# Ontologi: Fra media til juss

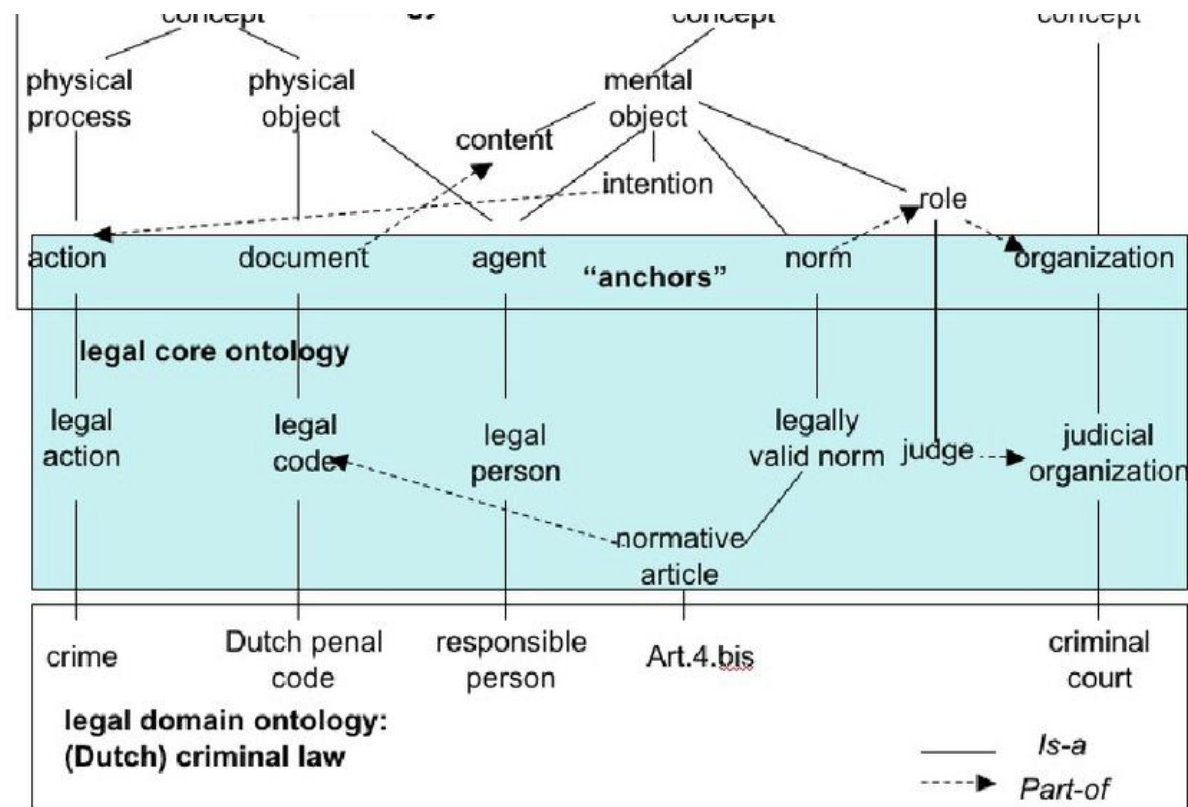
## Dublin Core. Meta-data nøkkel for mediabransjen

Name	Metadata subtag	Definition
Title	Title	A name given to the resource
Creator	Creator	An entity primarily responsible for making the content of the resource
Subject and keywords	Subject	The topic of the content of the resource
Description	Description	An account of the content of the resource
Publisher	Publisher	An entity responsible for making the resource available
Contributor	Contributor	An entity responsible for making contributions to the content of the resource
Date	Date	The date that the resource was published or some other important date associated with the resource
Resource type	Type	The nature or genre of the content of the resource
Format	Format	The physical or digital manifestation of the resource
Resource identifier	Identifier	An unambiguous reference to the resource within a given context: this is the object identifier
Source	Source	A reference to a resource from which the present resource is derived
Language	Language	A language of the intellectual content of the resource
Relation	Relation	A reference to a related resource
Coverage	Coverage	The extent or scope of the content of the resource
Rights management	Rights	Information about rights held in and over the resource

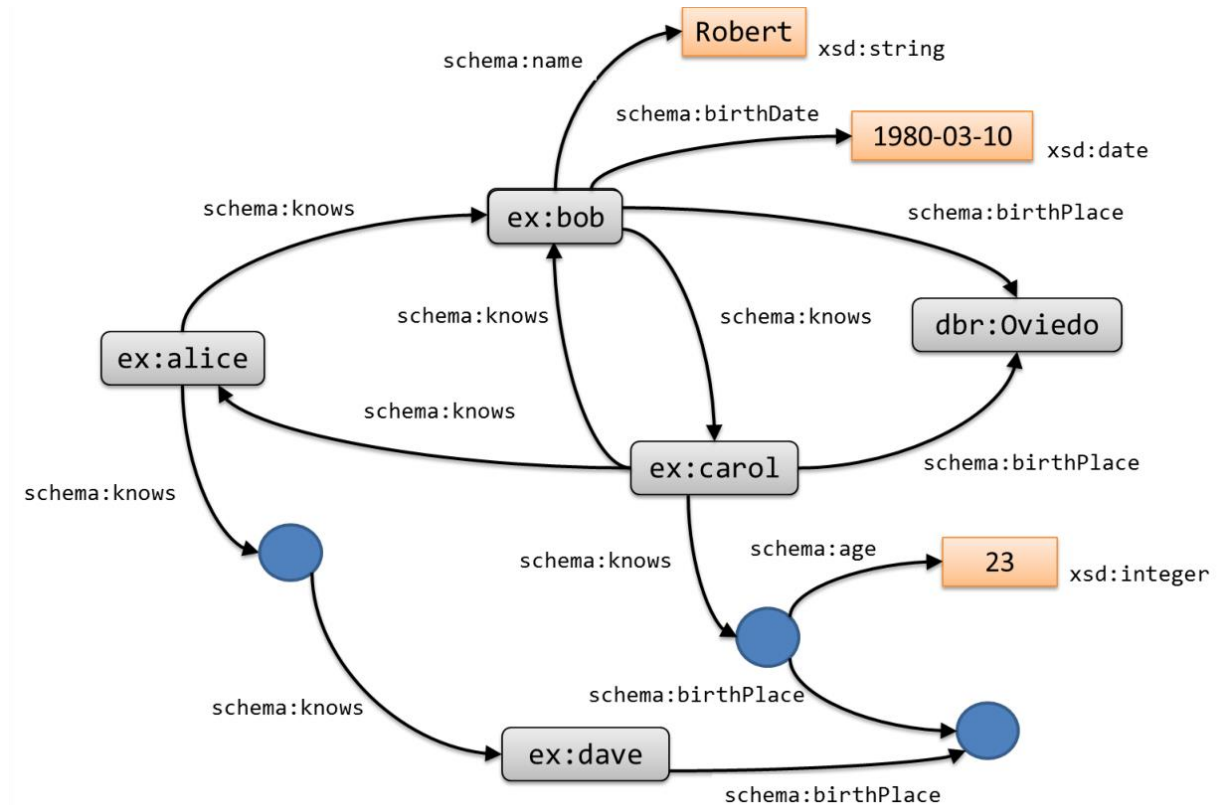
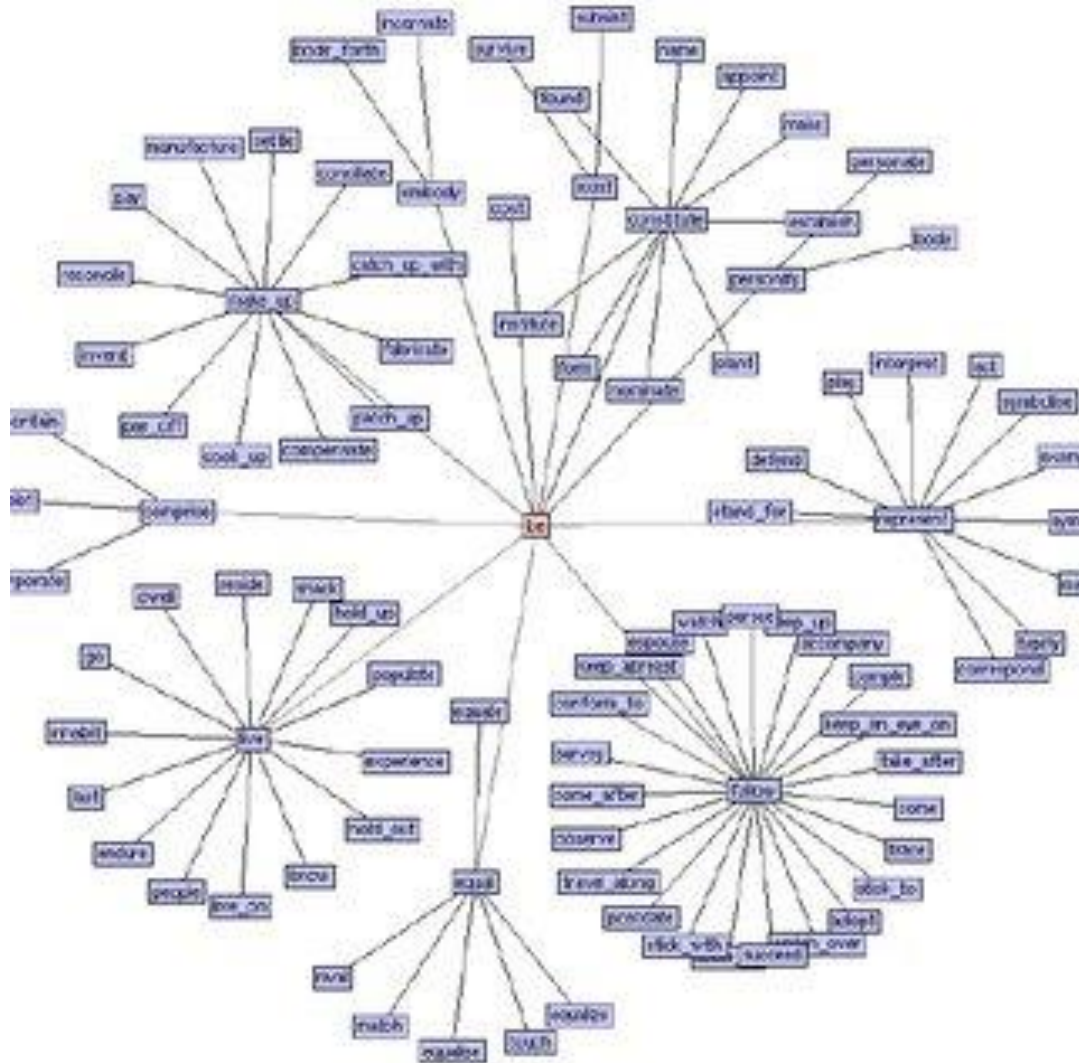
### Greenstone archive documents

In a single document, the Greenstone archive format imposes a limited amount of structure. Documents are divided into paragraphs and are split hierarchically into sections and subsections; these may be nested to any depth. Each document has an associated Open Document Architecture (ODA) ID—these are extended to identify sections and subsections by appending section and subsection numbers, separated by periods. For example, subsection 3 of section 2 of document HASHa7 is referred to as HASHa7.2.3.

## En ontologi for juss



# RDF, OWL, Protégé, OrdNett og flerfoldige ontologiske biblioteker klar for gjenbruk



# “Hvermansens” ontologi = folksonomy

## Folksonomy: Definition

- ❖ Folksonomy is the result of personal free tagging of pages and objects for one's own retrieval
- ❖ The tagging is *usually* done in a social environment (shared and open to others)
- ❖ The act of tagging is done by the person consuming the information

InfoCloud Solutions, Inc. - 2009



Meget nyttig for å kategorisere meldinger og bilder og dermed etablere et maskinlæringsgrunnlag  
Brukeren får hjelp til å søke etter relevant materiale og gruppere disse.

# Når bør fagspesifikke ontologier benyttes?

- Når krav til faglig presisjon er viktig
  - Relativt få, men viktige begreper
  - Viktig at flere peker
- Gir økt standardiseringer i dokumenter og spesifikasjoner
- Betydelig gjenbruk øker verdien av ontologien
- Kan harmonisere innhold i skriftlig materiale produsert av ulike parter og fagfolk
- Stor betydning der opphavet er farget av ulike faglige sjargonger

# Statistisk baserte systemer

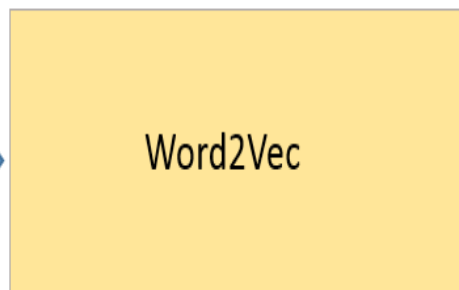
- Regressive metoder som underbygger det som populært kalles «dyp maskinlæring» har hatt en enorm fremgang de siste tiåret
  - Store nok maskiner – krever mye datakraft
  - Nok språkdata
- Talegjenkjenning
- Sentimentanalyse
- Språkoversettelse
- Språkproduksjon
- Omskriving
- I øyeblikket monopoliseres utviklingen av de store plattformaktørene
  - Google, Microsoft, Facebook, Amazon, Alibaba
  - Europa har ingen slike plattformer

# LSTM og regressjonsteknikker med dyp læring

## English Wikipedia Corpus

The Annual Reminder continued through July 4, 1969. This final Annual Reminder took place less than a week after the June 28 Stonewall riots, in which the patrons of the Stonewall Inn, a gay bar in Greenwich Village, fought against police who raided the bar. Rodwell received several telephone calls threatening him and the other New York participants, but he was able to arrange for police protection for the chartered bus all the way to Philadelphia. About 45 people participated, including the deputy mayor of Philadelphia and his wife. The dress code was still in effect at the Reminder, but two women from the New York contingent broke from the single-file picket line and held hands. When Kameny tried to break them apart, Rodwell furiously denounced him to onlooking members of the press.

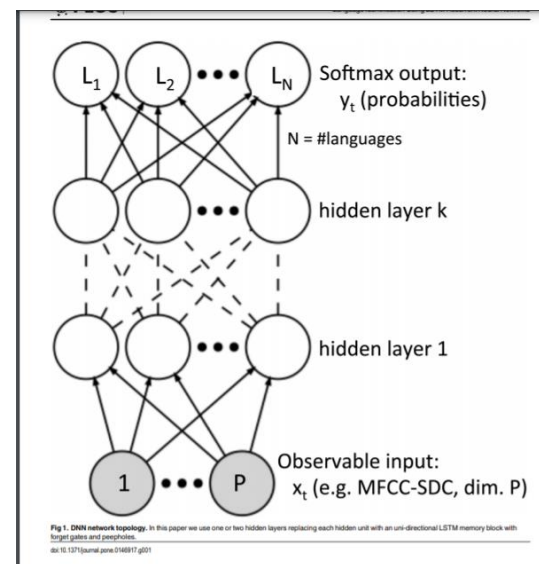
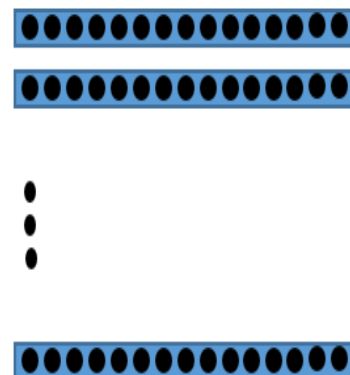
Following the 1969 Annual Reminder, there was a sense, particularly among the younger and more radical participants, that the time for silent picketing had passed. Dissent and dissatisfaction had begun to take new and more emphatic forms in society.<sup>11</sup>The conference passed a resolution drafted by Rodwell, his partner Fred Sargeant, Brody and Linda Rhodes to move the demonstration from July 4 in Philadelphia to the last weekend in June in New York City, as well as proposing to "other organizations throughout the country... suggesting that they hold parallel demonstrations on that day" to commemorate the Stonewall riot. ....



aardvark  
apple  
.  
.  
.  
zoo

## Embedding Matrix

D-dimensional vector



# Oversettelse



[Home](#) > [Tools](#) > **Babel Fish Translation**

## Babel Fish Translation [Help](#)

**Translate a block of text**

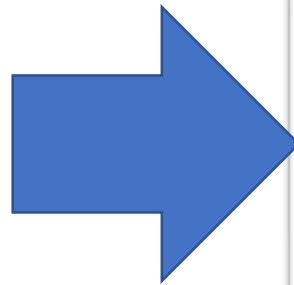
Select from and to languages

---

**Translate a Web page**

Select from and to languages

Add [Babel Fish Translation](#) to your site.  
Tip: You can now translate framed pages.



Norsk ↕ Engelsk

<p>Det er veldig alvorlig og nok et eksempel på at pandemien har rammet noen elever ekstra hardt. Jeg frykter at det betyr større forskjeller mellom elevene og økt risiko for frafall, skriver Thorkildsen i en e-post til VG.</p> <p>Hun setter pris på at det helvses</p>	<p>It is very serious and another example that the pandemic has hit some students extra hard. I fear that this means greater differences between the students and an increased risk of dropping out, Thorkildsen writes in an e-mail to VG.</p> <p>She appreciates that more and more consequences of school operations</p>
--	---

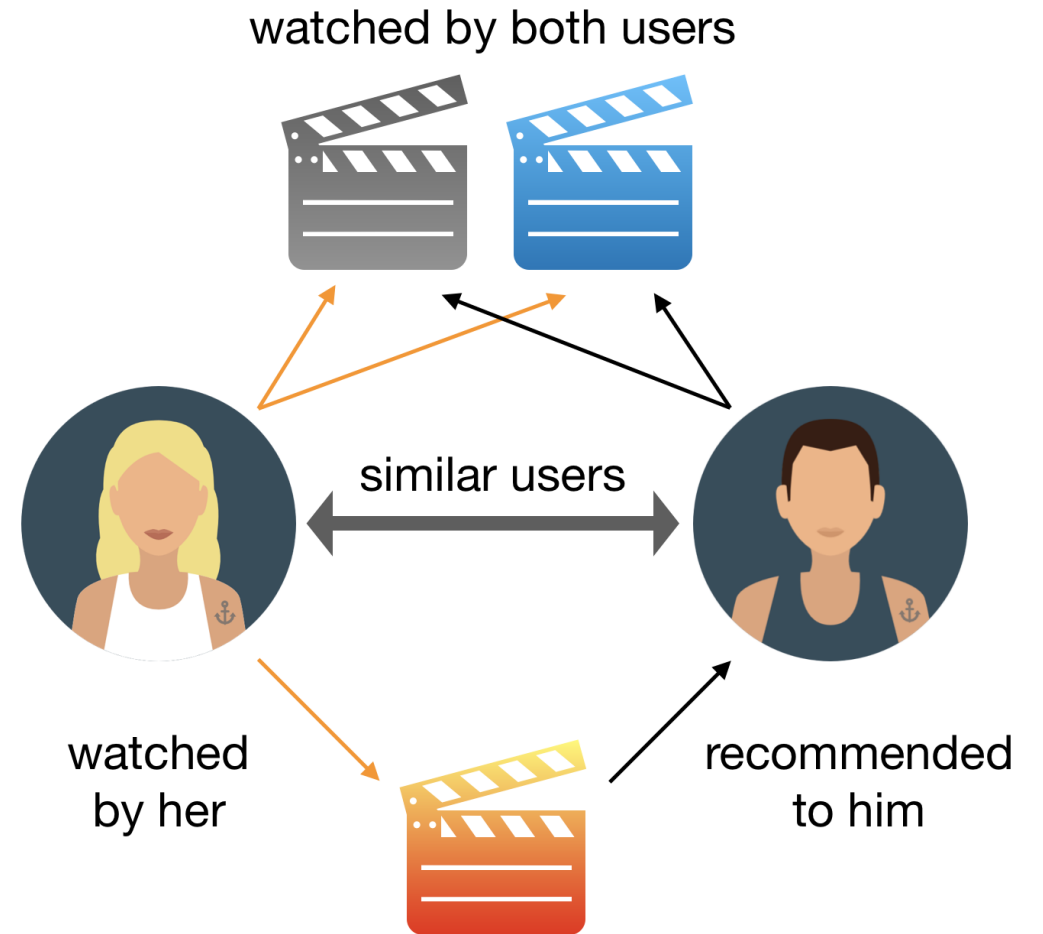


# Sentimentanalyse



# Sentimentanalyse bruker ord og fingerspråk

- Analyser kommentarer på Facebook, Twitter eller diskusjonsforum (flere hundre millioner mennesker)
- Etabler en kontrollgruppe av f.eks. 10.000 personer
  - Gjør en inngående analyse av deres behov, helse, legning, tro, politiske holdninger og hvordan de oppfører seg på sosiale media
  - Lag en profil på hver – grupper de som er like
- Overvåk sosiale media og gjøre en sentimentanalyse av enkeltindivider utenfor kontrollgruppen
  - sammenlign med medlemmer av kontrollgruppen
- Predikter behov, helse, legning, tro, politisk holdning
- Selg resultatet til de som betaler best





# Hvor går vi?

Fremtiden kom kanskje i 2020

# OpenAI



- **Generative Pre-trained Transformer 3 (GPT-3)**
- Lansert mai i år, beta versjon tilgjengelig i juli
- Autoregressiv språkmodell med “dyp læring”
- GPT-3's fullversjon bruker 175 milliarder maskinlærings parameter
- Overgår Microsofts Turing NLP med 10 ganger i antall parameter
- Microsoft har tatt kontroll over kildekoden
- Men en åpen API er tilgjengelig for systemutviklere

# Open AI



Q: What is your favorite animal?

A: My favorite animal is a dog.

Q: Why?

A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?

A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?

A: A giraffe has two eyes.

Q: How many legs does a frog have?

A: A frog has four legs.

Q: Are there any animals with three legs?

A: No, there are no animals with three legs.

Q: Why don't animals have three legs?

A: Animals don't have three legs because they would fall over.

Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm

Subtitle: Joaquin Phoenix pledged to not change for each awards event

Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what.

Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

# Omskriving av juridisk språk

**GPT3**  
*Examples*  
</>



Michael @michaeltefula · 21 Jul

Just taught GPT-3 how to turn legalese into simple plain English. All I gave it were 2 examples 🤖 Might build a term sheet and investment document interpreter out of this 🤖

## 2 Training Examples

**Legal clause:** The Company and the Founders will provide the Investors with customary non-competition, non-solicitation and confidentiality undertakings.

**Plain English:** The startup and its founders will provide the usual assurances and guarantees on facts about the business. The founders will also agree not to work for competitors, poach employees or customers when they leave the startup, and respect confidentiality.

**Legal clause:** In the event of an initial public offering of the Company's shares on a US stock exchange the Investors shall be entitled to registration rights customary in transactions of this type (including two demand rights and unlimited shelf and piggy-back rights), with the expenses paid by the Company.

**Plain English:** If the Company does an IPO in the USA, investors have the usual rights to include their shares in the public offering and the costs of doing this will be covered by the Company.

## AI Translations (AI Output in Blue)

**Legal clause:** Upon liquidation of the Company, the Series A Shareholders shall be entitled to all other shareholders an amount in respect of each Series A Share equal to the Liquidation Preference, plus all accrued but unpaid dividends. If the Company has assets remaining after the distribution of that amount, the Series A Shareholders will participate with the holders of Ordinary Shares pro rata to the number of converted shares.

**Plain English:** If the startup is wound up, the Series A investors will be paid first and they will also share any leftover assets with ordinary shareholders.

**Legal Clause:** Sale of all or substantially all of the assets of the Company involving a change in control (each, a "Corporate Transaction") will be treated as a liquidation and the proceeds of sale will be distributed as set out in paragraph 8.1. If Series A Shares have received any Special Dividend it shall be set off against the Liquidation Preference.

**Plain English:** If the company is sold, or a new owner takes control, the proceeds from the sale will be distributed to the Series A investors first.

**Legal Clause:** If holders of at least 51% of the Series A Shares and Ordinary Shares, there will be drag along rights so that all remaining shareholders are required to sell on the same terms, provided that the dragged shareholder provides to the purchaser any representations or warranties except as to title to the shares.

**Plain English:** If a majority of investors agree to sell their shares, any other shareholders will have to accept the same price and terms as the others.

## 2 Training Examples

**Legal clause:** The Company and the Founders will provide the Investors with customary representations and warranties examples of which are set out in Appendix 4 and the Founders will provide the Investors with customary non-competition, non-solicitation and confidentiality undertakings.

**Plain English:** The startup and its founders will provide the usual assurances and guarantees on facts about the business. The founders will also agree not to work for competitors, poach employees or customers when they leave the startup, and respect confidentiality.

**Legal clause:** In the event of an initial public offering of the Company's shares on a US stock exchange the Investors shall be entitled to registration rights customary in transactions of this type (including two demand rights and unlimited shelf and piggy-back rights), with the expenses paid by the Company.

**Plain English:** If the Company does an IPO in the USA, investors have the usual rights to include their shares in the public offering and the costs of doing this will be covered by the Company.

«The Jetsons» verden kommer nærmere

*Takk for oppmerksomheten*



UIT  
NORGES  
ARKTISKE  
UNIVERSITET

