STK1100 våren 2023

Sannsynlighetsregning og statistisk modellering

Generell introduksjon

Omhandler delvis stoffet i avsnitt 1.1 i læreboka
(annet stoff fra kapittel 1 tas ved behov)

Matematisk institutt
Universitetet i Oslo

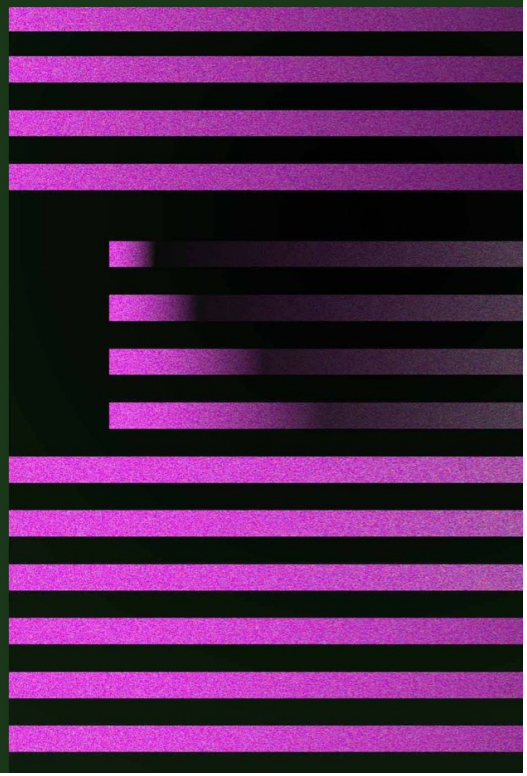
OpenAI

API

ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

TRY CHATGPT ↗

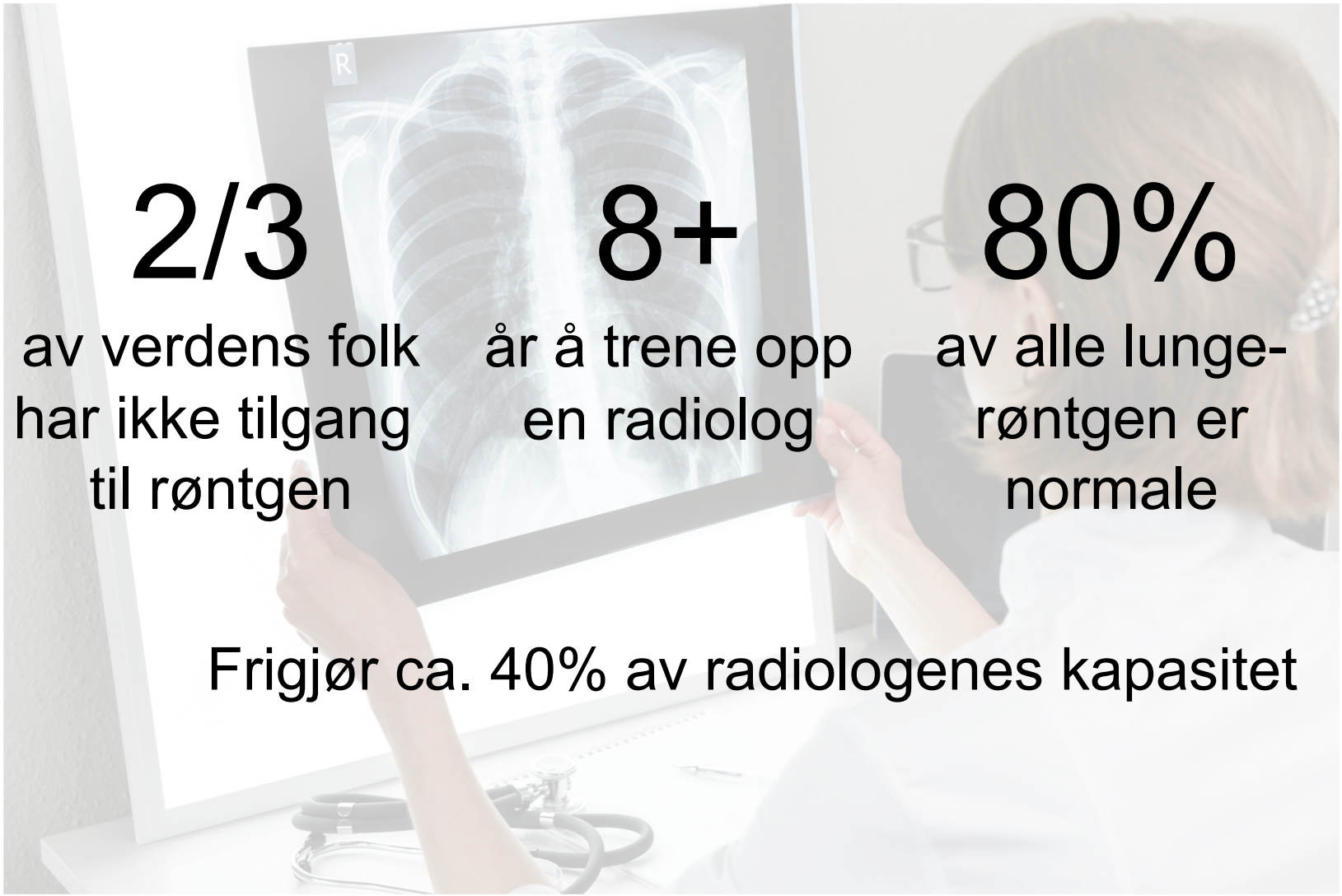


Illustrasjonsfoto: Colourbox.com

2022: Verdens første CE-sertifiserte hel-automatiske billed-diagnostiseringsverktøy

Identifiserer normale lunge-røntgen og produserer rapport

Kilde: https://oxipit.ai/news/first-autonomous-ai-medical-imaging-application/



2/3
av verdens folk har ikke tilgang til røntgen

8+
år å trene opp en radiolog

80%
av alle lunge-røntgen er normale

Frigjør ca. 40% av radiologenes kapasitet

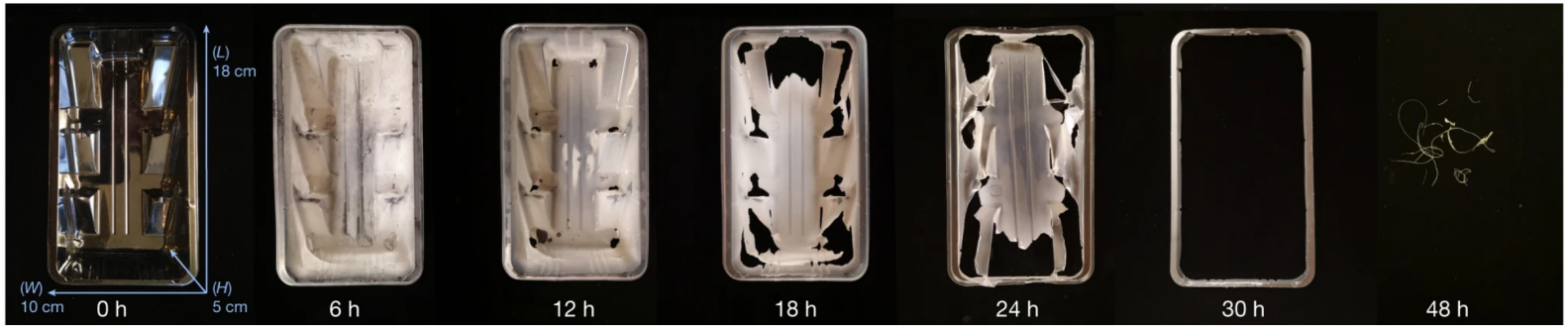
Illustrasjonsfoto: Colourbox.com

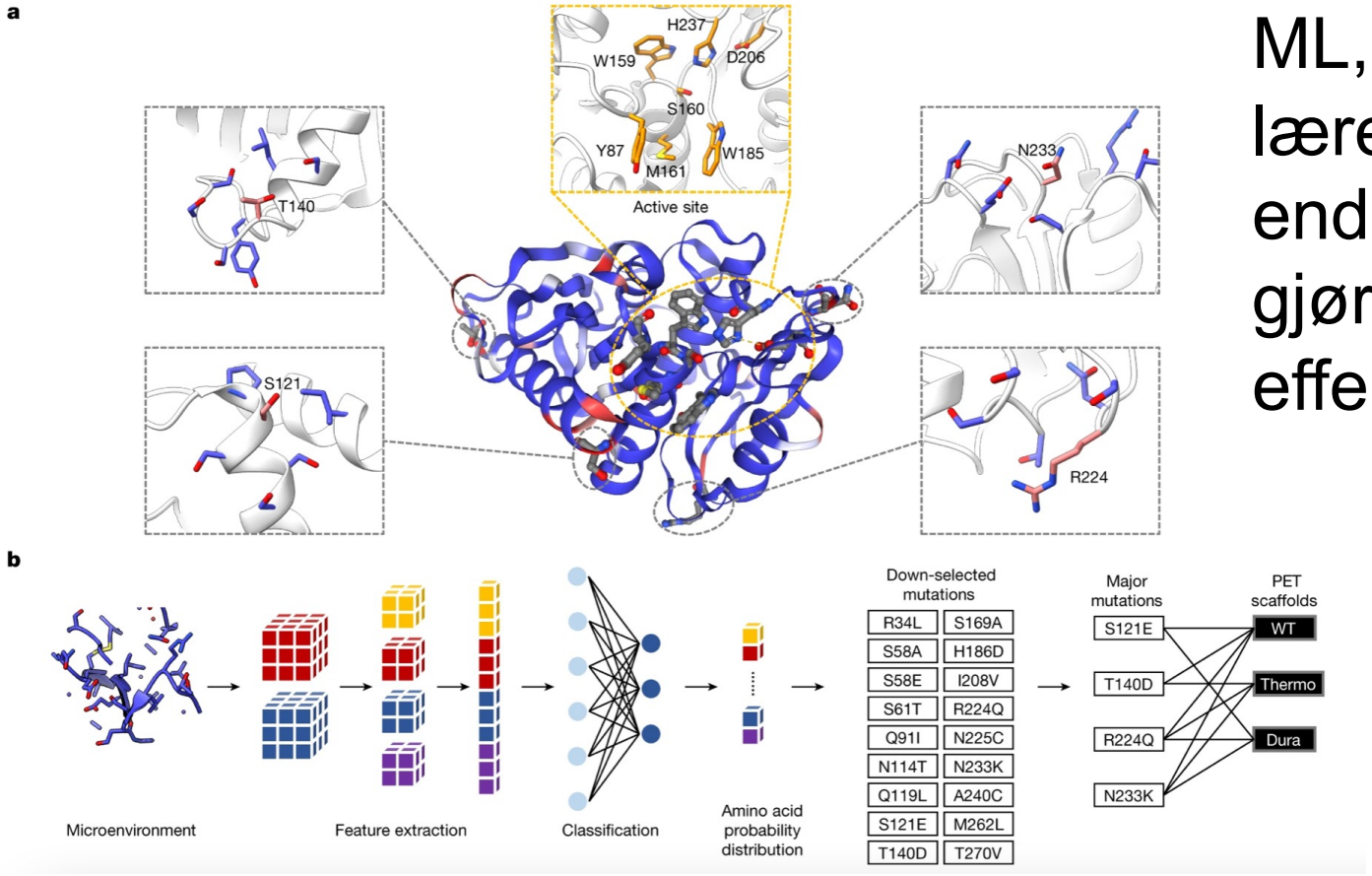
Machine learning-aided engineering of hydrolases for PET depolymerization

Hongyuan Lu[1], Daniel J. Diaz[2], Natalie J. Czarnecki[1], Congzhi Zhu[1], Wantae Kim[1], Raghav Shroff[3,4], Daniel J. Acosta[3], Bradley R. Alexander[3], Hannah O. Cole[1,3], Yan Zhang[3], Nathaniel A. Lynd[1], Andrew D. Ellington[3] & Hal S. Alper[1]✉

nature

April 2022



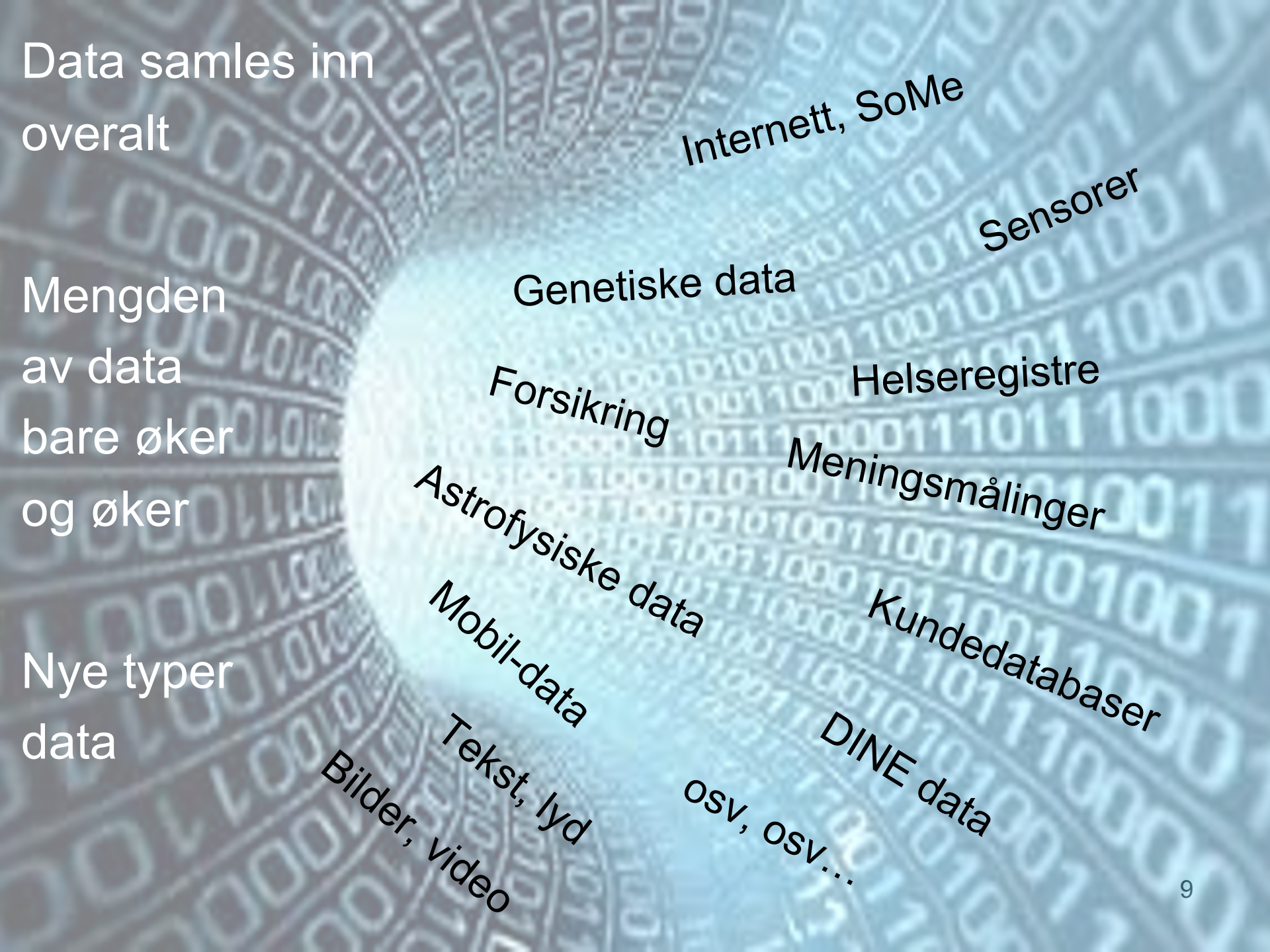


ML, nevrale nett: lærer seg hvilke endringer som gjør enzymet mer effektivt og stabilt

Data samles inn overalt

Mengden av data bare øker og øker

Nye typer data



Data må behandles og fortolkes for å gi ny kunnskap og kunne danne grunnlag for beslutninger.

Da må vi ta hensyn til at data som oftest er påvirket av individuelle variasjoner, måleusikkerhet og andre faktorer av tilfeldig natur.

Vi må ta hensyn til disse tilfeldige variasjonene når dataene skal behandles og fortolkes.

Vi beskriver da dataene med en statistisk modell som tar hensyn til de tilfeldige variasjonene og bruker en statistisk metode for å få informasjon ut av dataene.

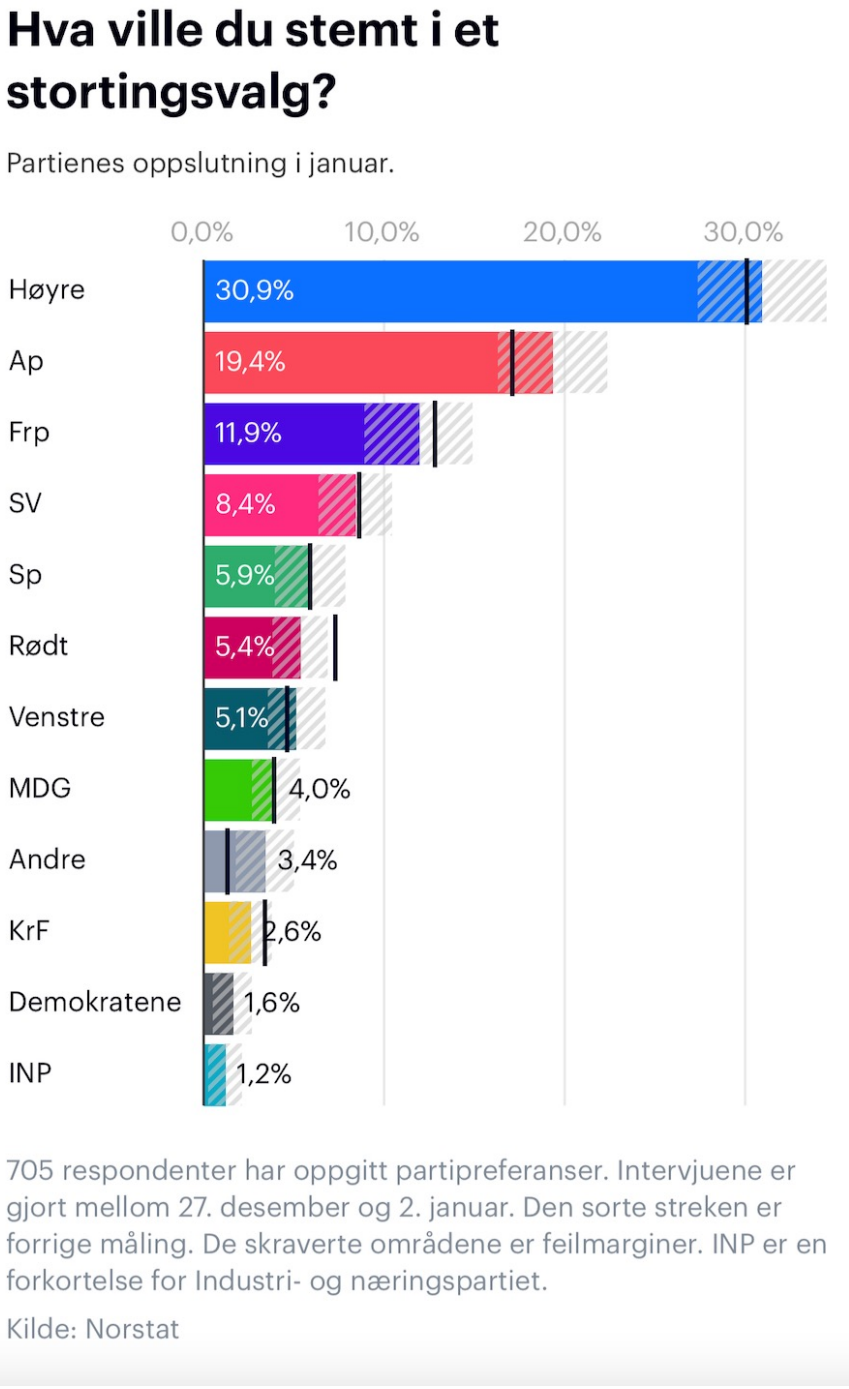
Sannsynlighetsregningen danner grunnlaget for statistiske modeller og metoder.

Vi vil se kort på et veldig enkelt eksempel.

Publisert 03.01.23, 955 intervjuet, 705 har oppgitt partipreferanse

Av et utvalg på 705 som ville ha stemt hvis det var valg, ville 137 ha stemt Arbeiderpartiet (Ap).

Hva er Ap's oppslutning i populasjonen av alle stemme-berettigede?



Vi får en (noe forenklet) statistisk modell for meningsmålingen ved å se på den som et binomisk forsøk (jf. Matematikk R1):

- Hver av de n = 705 personene i utvalget vil enten ha stemt Ap, eller så ville de ikke det
- Sannsynligheten for at en person ville ha stemt Ap er p = «andelen i populasjonen som ville ha stemt Ap»
- Personene i utvalget ville ha stemt Ap eller ikke uavhengig av hverandre

Et estimat på Ap's oppslutning er $\hat{p} = \frac{137}{705} = 0.194$

Ved å bruke kunnskap om den tilfeldige variasjonen i et binomisk forsøk, kan vi regne ut at en feilmargin for dette estimatet er $\pm$ 3.0 prosentpoeng

STK1100 er et basalt kurs som legger hovedvekten på grunnleggende sannsynlighetsregning og statistisk modellering.

Vi vil også se litt på hvordan sannsynlighetsregningen danner grunnlaget for statistiske metoder.

STK1110: Statistiske metoder og dataanalyse

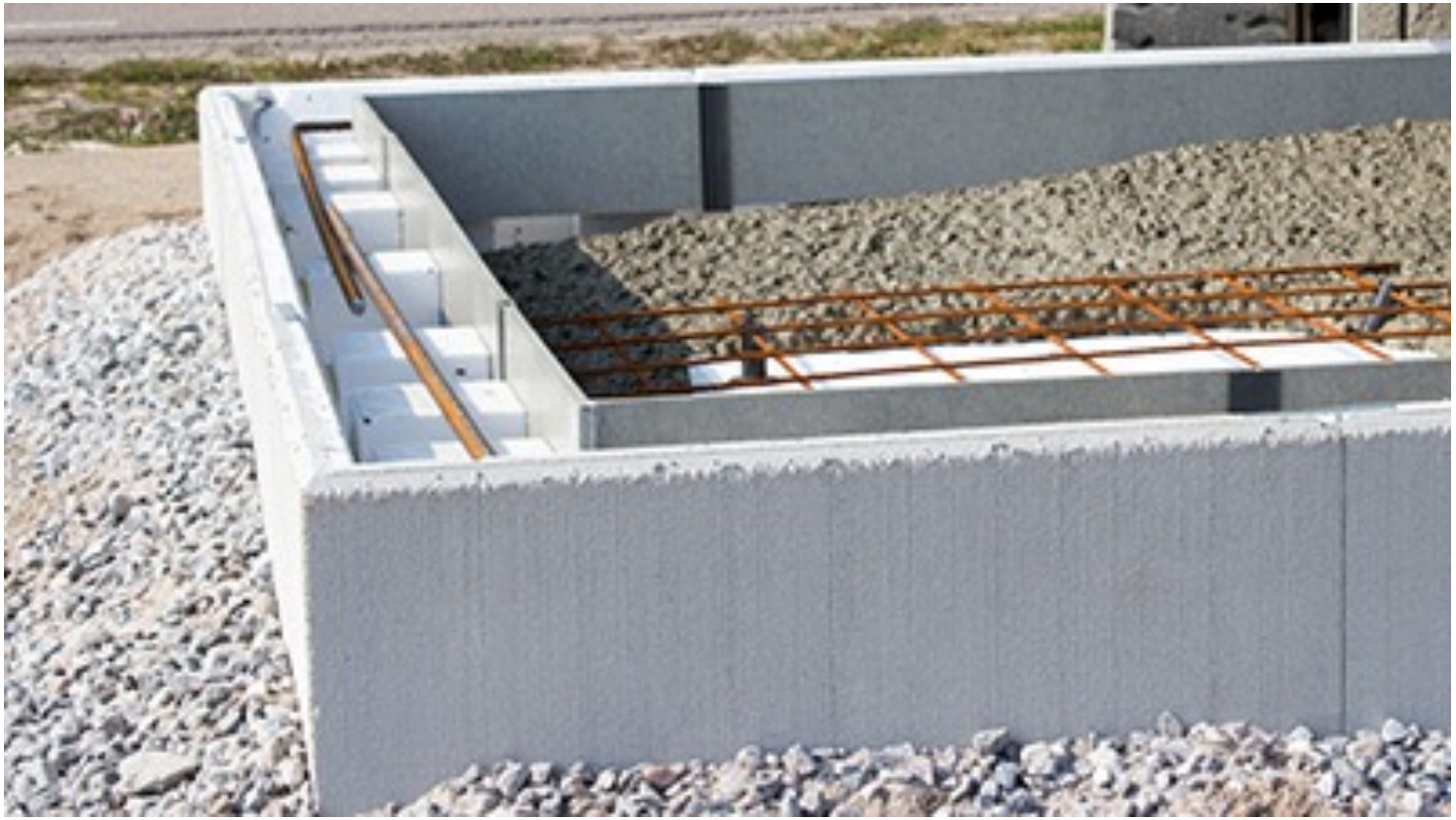
STK2100: Maskinlæring og statistiske metoder for prediksjon & klassifikasjon

Og deretter en rekke påfølgende kurs!

Hovedtrekk i pensum i STK1100
(fra læreboka til Devore, Berk & Carlton):

- Kapittel 1: Beskrivende statistikk (tas ved behov)
- Kapittel 2: Grunnleggende sannsynlighetsregning (mye repetisjon fra Matematikk R1)
- Kapitlene 3-6: Diskrete og kontinuerlige stokastiske variabler. Sannsynlighetsfordelinger, forventning, varians, store talls lov, sentralgrensesetningen, m.m.
- Kapitlene 7 og 8: Litt om statistiske metoder (estimering og konfidensintervall)
- Kapittel 12: Mer om statistiske metoder (regresjon)

STK1100 = grunnmur!



Vitenskaps-slottet – “Castle of science”– DALL-E