MACHINE LEARNING

Vad och Varför

Thomas Hellström
Department of Computing Science
Umeå Universitet

Medicinteknikdagarna

2018-11-10

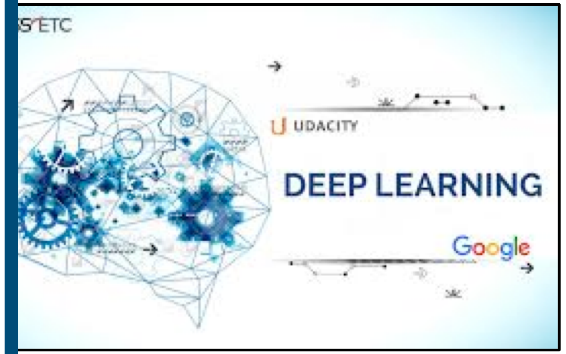
UMEÅ UNIVERSITY

Intelligent robotics @ Umeå University

Professor Thomas Hellström
Institutionen för datavetenskap
Umeå Universitet

Machine Learning är HETT!

- Google, Amazon, Yahoo, Facebook, Microsoft, IBM öser in resurser i ML
- Google har skapat en ny forskningsgrupp i Zurich, specifikt för ML
- Machine Learning (CS 229) är den mest populära kursen på Stanford University
- …



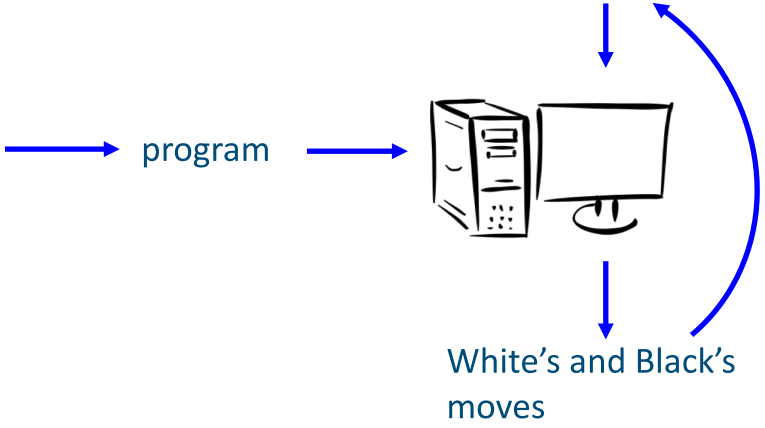
Jag ska försöka hinna med det här

- Vad är ML?
  - Historik, exempel
  - Definition
- Varför är ML så populärt?
  - Varför just nu?
  - Begränsningar, utmaningar

Vad är ML?

ML är inte nått nytt!

- De första artiklarna från 1950-talet !
- Arthur Samuel’s *Dam*-program (1959)



- Programmet lärde sig av att spela dam mot sig själv
- Slutligen spelade datorn bättre än Samuel själv
- Samuel myntade termen ”machine learning”

38 år senare

- IBM's Deep Blue slår stormästaren Gary Kasparov i schack (1997)

Tusentals gamla mästarmatcher

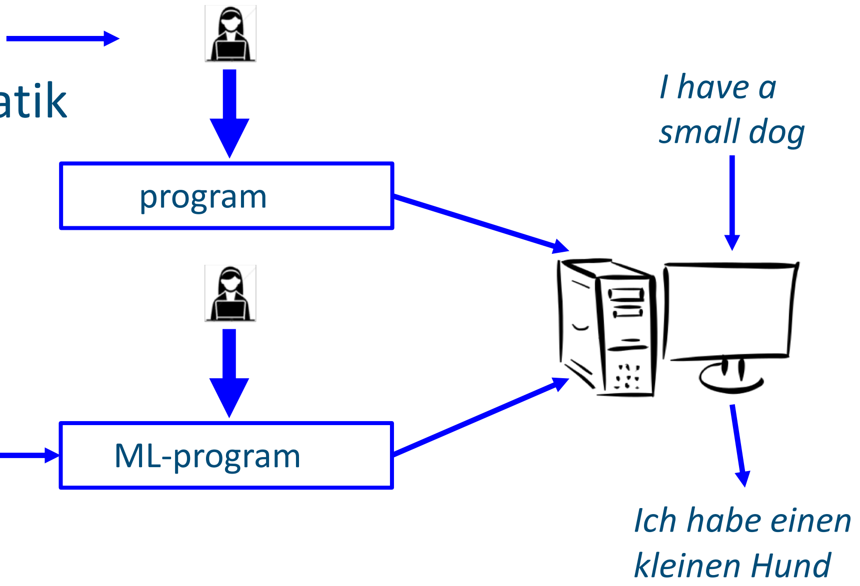
program

- Deep Blue lärde sig en evalueringsfunktion och använde rå datorkraft för att välja nästa drag

ML för automatisk översättning av naturligt språk

- Utan ML:

1700+ sidor med engelsk grammatik & en tysk grammatik



- Med ML (2015):

10^9 sidor av översatt text

- Använde Deep Learning
- Mycket bättre än existerande system som INTE använde ML

ML för att generera bildtexter

- Lär sig från en databas med bilder/bildtext (>1M bilder)

A group of young people playing Frisbee

A person riding a motorcycle on a dirt road

A refrigerators filled with lots of food and drinks

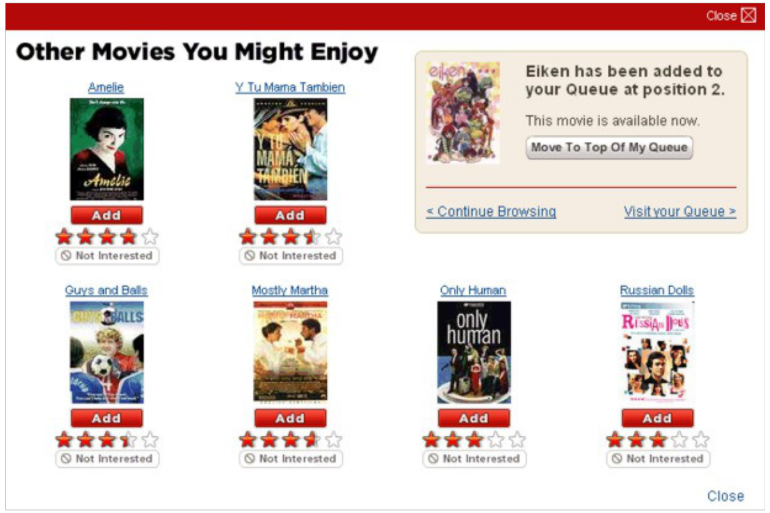
- Mycket bättre än state-of-the-art
- Utfördes av fyra Googleforskare:

Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, *Show and Tell: A Neural Image Caption Generator* , Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15).

Exampel: Analys av människor i videos

Exampel: Rekommendationssystem

- Vilken film ska rekommenderas till en Netflixkund?

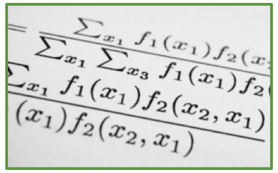


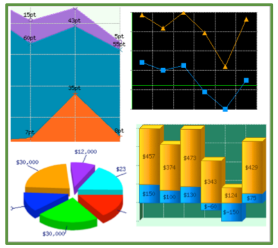
- Vem ska få specifika produkterbjudanden?

Så vad är Machine Learning?

- Använd data för att skapa ett program som kan
  - *Utföra en uppgift*
    - Spela schack
    - Översätta
    - Generera bildtexter
    - Analysera röntgenbilder
    - Detektera fotgängare i självkörande bilar
  - *Extrahera information från data*
    - Vilken film ska rekommenderas till en Netflixkund?
    - Vem ska få specifika produkterbjudanden?

Relaterade “buzz words”



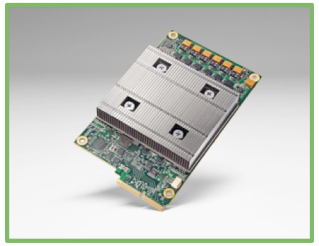


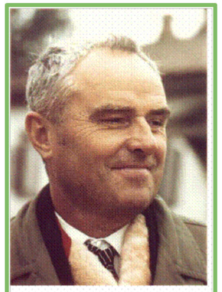
- Data Mining
  - Används ofta synonymt med ML
  - Alternativ definition: *"The non trivial extraction of implicit, previously unknown, and potentially useful information from data”*
- Knowledge Discovery in Databases (KDD)
- Information Extraction
- Data Science
- Business Intelligence, Business Analytics, Predictive Analytics, Big Data
- Alla inkluderar grundläggande ML, men också annat
  - preprocessing, cloud management, data base management, data visualization, reporting, …

Varför?

Varför just nu?

- Big Data!
  - Enorma mängder potentiellt värdefull data
  - Används i ML för inlärning
- En explosion av datorkraft
  - TPU (Tensor Processing Unit) är specifikt anpassad till TensorFlow
  - Beräkningar i molnet
- Ny (?) utveckling av algoritmer och teori
  - Deep learning (1965)
- Uppenbara begränsningar med andra ansatser




О.Г. Івахненко (1967 р.)

Begränsningar med en modeldriven ansats

- “The Unreasonable Effectiveness of Mathematics in the Natural Sciences”[1]:
  - Fysik kan ofta beskrivas elegant med enkla matematiska formler som $f = ma$ eller $e = mc^2$
- Vetenskap som inkluderar människor är ofta resistent mot elegant matematik
  - En (ofullständig) Engelsk grammatik är på över 1700 sidor
  - Traditionella system för röstigenkänning och språkförståelse fungerade dåligt och verkade inte kunna bli så mycket bättre
- “The Unreasonable Effectiveness of Data“
  - State-of-the-art inom röstigenkänning, maskinell översättning, och bildanalys är datadrivna

1. E. Wigner, “The Unreasonable Effectiveness of Mathematics in the Natural Sciences,” *Comm. Pure and Applied Mathematics*, vol. 13, no. 1, 1960, pp. 1–14.

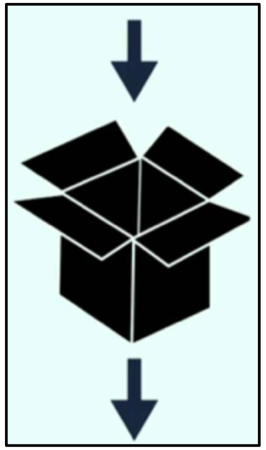
Begränsningar med en datadrivna ansats (ML)

- "Låt data tala för sig själv"
- Lätt att dra felaktiga slutsatser
  - "Data snooping", Overfitting, Overtraining
  - "If you torture data long enough it will confess to anything"

Förväxling av statistisk korrelation med orsakssamband

Begränsningar med en datadriven ansats (ML)

De flesta ML-tekniker leder till “black boxes”



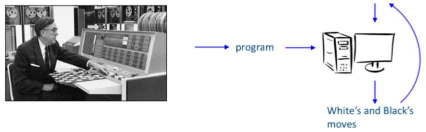
- Det sätt på vilket en given input genererar en output är inte känd
- Detta gör det svårt att
  - Nå insikt i de studerade processerna
  - Förklara prediktioner och beslut
  - Debugga program
  - Validera systemen
- Beslutssystem kan lära sig att diskriminera p.s.s. de människor från vilka de lär sig
- En vanlig strategi är att använda ML bara som ett rådgivande system, parallellt med (ett system som är gjort av) en mänsklig expert.

När är ML användbart?

- När man inte vet hur man skriver programmet
  - Sånt som människor kan, men som är svåra att formalisera
    - Schack, bildanalys, översättning, bedömning av löneansökningar, detektering av spam, medicinska diagnoser eller behandlingar.
  - Sånt som är svårt för en människa att utföra
    - Svåra förutsägelser tex.: solaktivitet, underhåll av maskiner, annonskampanjer, kreditkortsbedrägerier.
- När det tar för lång tid att skriva programmet
  - Automatisk detektering av föremål i bilder
- När processen förändras med tiden
  - Programmet MÅSTE bero på data
- När vi inte ens vet vad programmet ska göra
  - ”Frequent pattern mining”, t.ex. analys av inköpskvitton för att finna köpmönster

ML är inte nått nytt!

- De första artiklarna från 1950-talet !
- Arthur Samuel’s checkers-program (1959)



- Programmet lärde sig av att spela checkers mot sig själv
- Slutligen spelade datorn bättre än Samuel själv

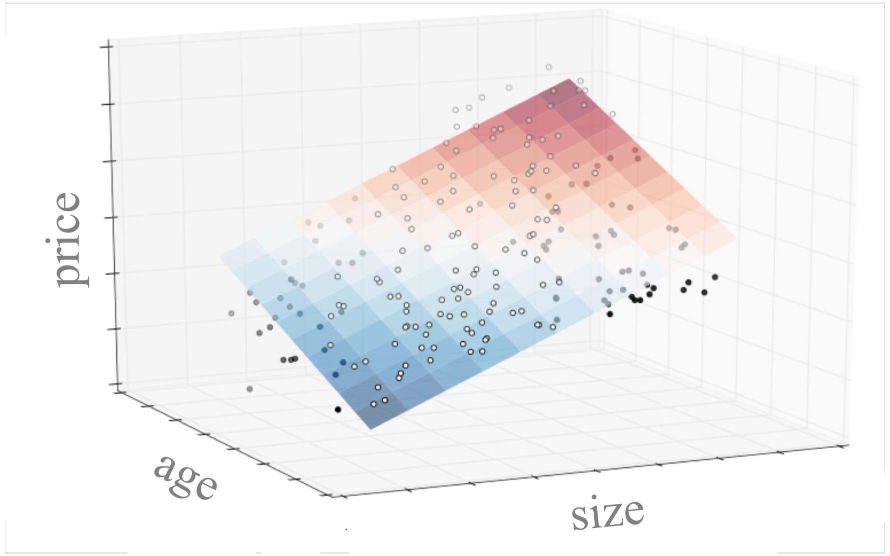
How?

General categories of ML problems

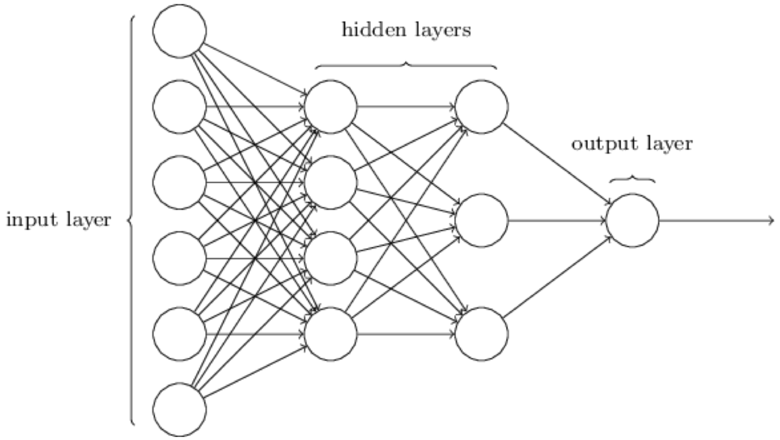
- Regression
- Classification
- Clustering

Regression – prediction of continuous variables

- How can the price of a house be predicted from parameters like size, location, no. of rooms, age, view, size of yard, … ?
- Find a real-valued function $f\,(rooms, age, view, size, \ldots)$
- Use data to find parameter values that makes data fit a model like $price = p_0 + size * p_1 + age * p_2$

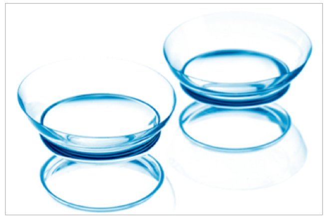


Usually non-linear models - Artificial **neural networks**:

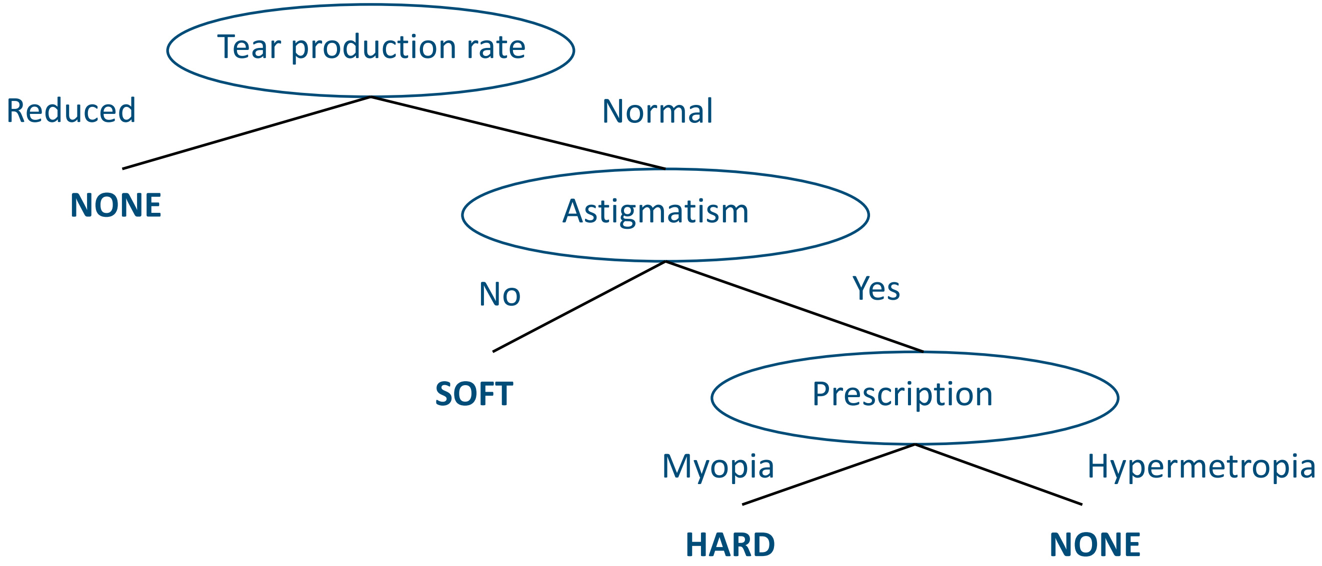


Classification – prediction of categorical variables

- What type of contact lenses (soft, hard, none) should a customer be recommended, based on age, prescription, astigmatism, tear production rate?

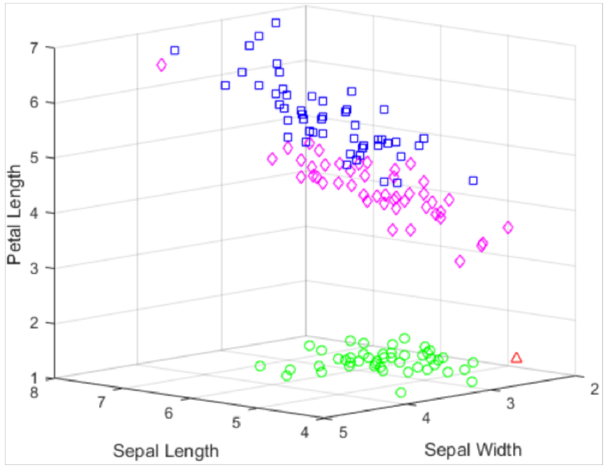


- Find a discrete function *C*(*age, prescription*, …) that returns the type of lens to recommend.
- Based on collected data, a **decision tree** is computed:



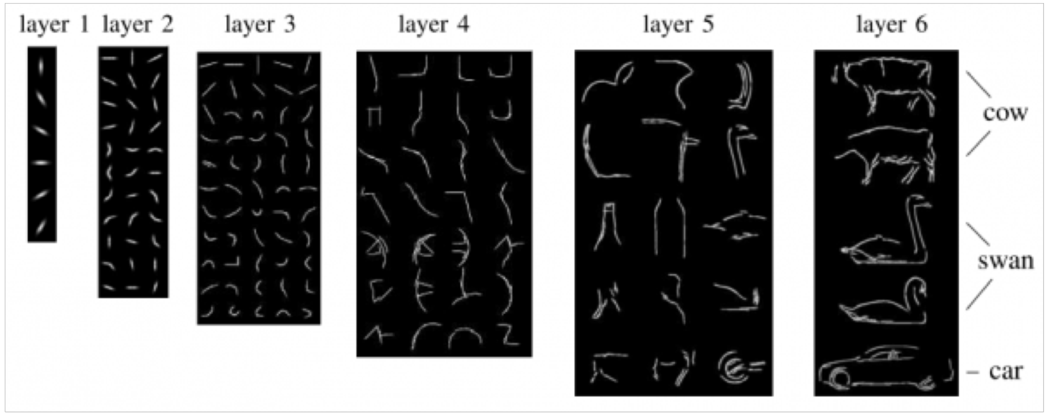
Clustering – grouping observations

- Can iris flowers be categorized based on *Petal Length, Sepal Length, Sepal Width*?
- Data for a lot of iris flowers is collected
- Running a clustering algorithm reveals 3 distinct clusters
- These clusters correspond to the iris species *setosa, versicolor, virginica*
- Clustering can be used for *Market segmentation*
  - Identification of homogeneous groups of customers that may respond differently to promotions or advertising.
- Typical algorithms: **C-means clustering** algorithm



What about this Deep Learning?

- Not really new either (2006 or actually already 1965)
- Usually multi-layer (> 2) artificial neural networks
- Hierarchical: higher level (layer) concepts are learned from the lower level (layer) ones



- In this way, traditional feature extraction is not necessary
- Need a LOT of data and takes a LONG time to train
  - Finding optimal architecture (number of layers/units, the learning rate, initial weights etc.) requires new methods
  - Highly parallelizable (GPU, and dedicated hardware)
- Seriously black-box