**Diagram

Description automatically generated**

Text

Description automatically generated

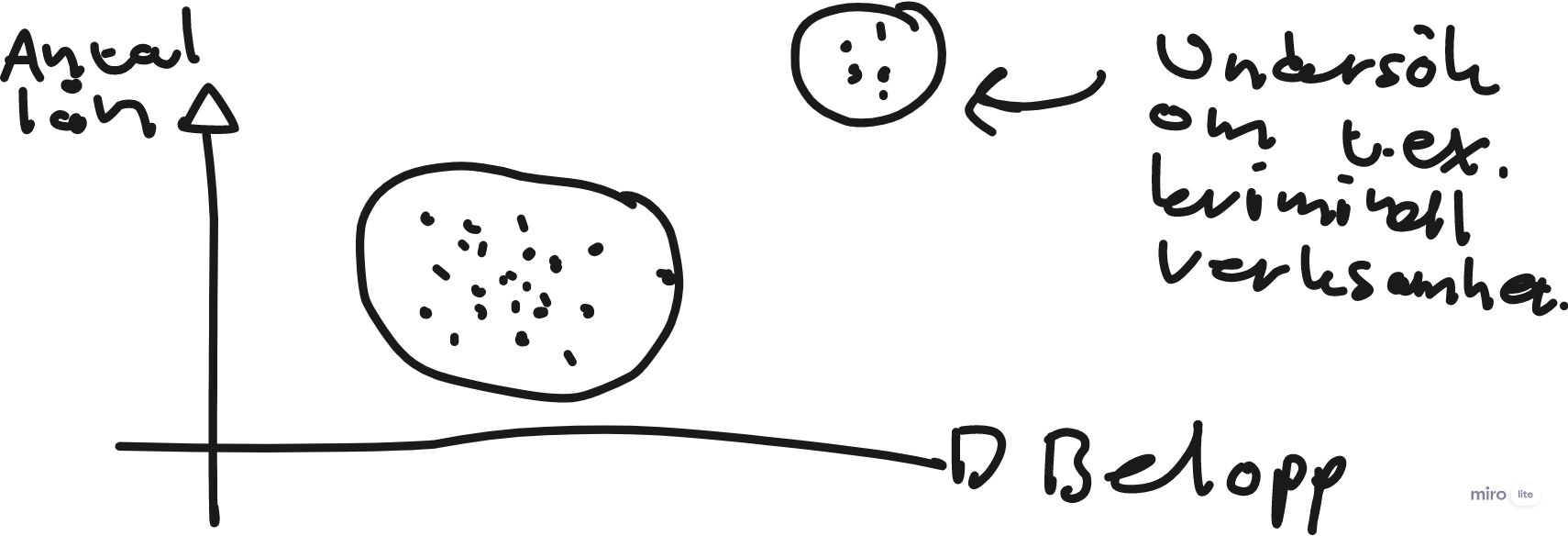
* Stor del av datan i verkligheten har inga "labels". Därför kan unsupervised learning techniques (såsom klustring) användas i många sammanhang.
* Exempel: I ditt familjealbum har du bilder på dig själv, din mamma, din pappa. Stoppar du in bilderna i datorn så har du data utan labels (y), då är det ett "unsupervised learning" dataset.   
  Om vi för varje bild hade skrivit vem bilden är på (dig själv, din mamma eller din pappa) så har vi ett "supervised learning" dataset.
* Klustring (att man sätter datan i olika "kluster" / grupper) används för "unlabeld" data set och har flera tillämpningsområden. Några exmepel:   
    
  1. Kundsegmentering: Gruppera kunder på t.ex. ålder, lön, inkomst, köpbeteende).   
  - Troligtvis så vill man hantera kunder som är unga med låg inkomst annorlunda än förmögna pensionärer som har en hög inkomst.

A picture containing graphical user interface

Description automatically generated

2. "Anomalie Detection"

 - Observationer som inte passar in "så bra" till något kluster/grupp kanske är en anomalie. T.ex. vissa bankkunder tar extremt många och stora lån under en kort tid. Detta kan t.ex. bero terroristorganisationer, krimininell verksamhet eller dylikt. Dessa transaktioner kanske man vill kolla lite extra på.



3. Semi-supervised learning.

 - Skapar kluster och kollar på några observationer i varje kluster (ofta kollar man på centroiden) och sätt manuellt labels på dem observationerna du kollar på. Därefter så sätter du samma labels på samtliga observationer i klustret.

"Woho" :D Nu kan du göra supervised learning för du har manuellt satt labels på ditt data set.

Graphical user interface

Description automatically generated

**K-means (s.238)**

* Se bild på s.239.
* Med hjälp av K-means så kan vi träna en modell med t.ex. 5 kluster (vi väljer själva antalet kluster).
* K-means algoritmen försöker hitta varje klusters centrum (centroid observation). Därefter så "klassificierar man" alla punkter till det kluster vars centrum är närmast. Chart

  Description automatically generated with low confidence
* "Food for though": Uppenbart att k = 5 (d.v.s. 5 kluster) är lämpligt i bilden ovan. Men, om vi har t.ex. ett 10 dimensionellt dataset (x1, x2, ..., x10), hur väljer vi då antalet kluster när vi inte kan visualisera datan rakt av?
* Vi pratar om detta strax.
* Vi kollar på koden på (s.239-240)
* attributet centers\_: Koordinater för varje klusters centrum.
* transform metoden mäter avståndet till varje klusters centrum för varje observation.

**Hur fungerar K-means algoritmen? (Se bild s.242)**

1. Vi initerar modellen genom att slumpmässigt placera ut centroider.
2. Klassificiera alla data punkter (varje observatio "tillhör" det kluster vars centroid är närmast).

Itera steg 1 och 2 tills centroiderna slutar röra på sig.

Scatter chart, qr code

Description automatically generated

* Algoritmen kommer garanterat konvera i ett ändligt (d.v.s. ej oändligt) antal iterationer.
* **Problem:** Vi kan konvergera till en suboptimal lösning.
* **Lösning till problemet:** Vi gör flera slumpmässiga initialiseringar och väljer den bästa lösningen bland dessa olika initialiseringar.   
  - I sklearn så kan antal initialiseringar styras med hjälp av hyperparametern n\_init (se s.243)
* **Hur vet vi vilken lösning som är bäst?** Svar: Vi använder en perfomance metric som heter **inertia** (för linjär regression var t.ex. vår Performance Metric "Mean Squared Error" [MSE] ).

A picture containing text

Description automatically generated

* K means algoritmen kör ovanstående iterativa process flera gånger (antalet gånger styrs av n\_init) och väljer den modell som har lägst inertia.
* Nu har vi en överblick och god förståelse för hur K-means modellen fungerar.
* I sklearn så görs vissa saker för att optimera prestandan men vi bryr oss inte om detaljerna eftersom vi nu vet hur modellen principellt sett fungerar.

**Hur skall vi välja antalet kluster?**

* Tidigare (se s.239) så såg vi att 5 kluster var rimligt. Men hur gör vi generellt sett när vi har fler dimensioner?
* Vi hade kanske kunnat välja den modell med lägst inertia? Problemet är, att ju fler kluster du har ju lägre blir inertia.
* Potentiell lösning (se bild s.246): välj k (antal kluster) där intertia slutar avta så mycket (vid "armbågen"). Se bild nedan, där hade vi valt k = 4.

Chart, line chart

Description automatically generated

* Ovanstående metodik med inertia är en relativt "grov" metod.
* En mer precis metod är att använda "**silhouette score**" som är medelvärdet av varje datapunkts "**silhouette coefficient**".

Text

Description automatically generated

Chart, line chart

Description automatically generated

A picture containing text, blackboard, night sky

Description automatically generated

* En annan visualisering som innehåller mycket information kan vi se på (s.248) och kallas "Silhouette diagram".
* Chart

  Description automatically generatedVarje "subplot" har en "kniv" per kluster.
* Knivens höjd visar antalet datapunkter varje kluster har.
* Knivens bredd representerar de sorterade "silhouette coefficients" för datapunkterna i klustret som kniven representerar.
* Den röda sträckade linjen visar "mean silhouette coefficient" d.v.s. "silhouette score".
* Generellt sett så vill man att så många kvinar som möjligt skall passera den röda sträckade linjen och att klustren har ungefär liknande antal observation är också bra.
* I bilden ser vi att k = 4 eller k = 5 är rimliga alternativ. Men vi väljer k = 5 för alla knivar har passerat den röda sträckade linjen och dessutom så har varje kluster ungefär lika många observationer.

**Begräninsningar med K-Means (s.248)**

+ K-means är generellt sett en snabb och skalbar (skalbar betyder att den kan hantera stora dataset utan att den tar extremt lång tid) algoritm.

* Man bör skala datan innan man kör k-means algoritmen.

- Man måste köra algoritmen flera gånger för att undvika att få en suboptimal lösning.

- Man måste välja k (antal kluster) och detta kan vara svårt.

- K-means kan ha "dåligt beteende" när klustrena har olika storlek eller "icke sfäriska former" (sfär = multidimensionell generalisering av en cirkel) eller olika denisteter (d.v.s. hur tätt packade observationerna är)

A picture containing text, night sky

Description automatically generated

* Exempel s.251.   
  Där använder man klustring för "preprocessing". Variabler ersätts med avståndet till olika kluster. Läs i boken för ett exempel kopplat till MNIST.
* Exempel s.253.   
  Klustring kan användas för semi-supervised learning där klustrenas centroider (vi kallar dem "representative images") används för att göra någon form (finns olika sätt) av manuell labeling. Läs exemplet som använder MNIST datan.
* Vid intresse så kan ni läsa om DBSCAN på s.255.
* Andra klustringsalgoritmer kan du läsa om på s.258 vid intresse.
* Gaussian Mixtures ingår inte i kursen.