Koneoppiminen ja tekoälyn opettaminen: tiedon laadun merkitys

Miika Malin

Oulun yliopisto / BISG

23.5.2023

Esityksen rakenne

1. Tekoälyn ja koneoppimisen määritelmä
2. Koneoppimisen eri osa-alueet
3. Tiedon laadun ominaisuudet, laadun arviointi ja parantaminen
4. Esimerkkejä huonon tiedon käytöstä koneoppimisessa
5. Lopputentti / Yhteenveto

Mitä on tekoäly?

Tietokonejärjestelmien kyky suorittaa tehtäviä, jotka yleensä vaativat ihmisen älykkyyttä

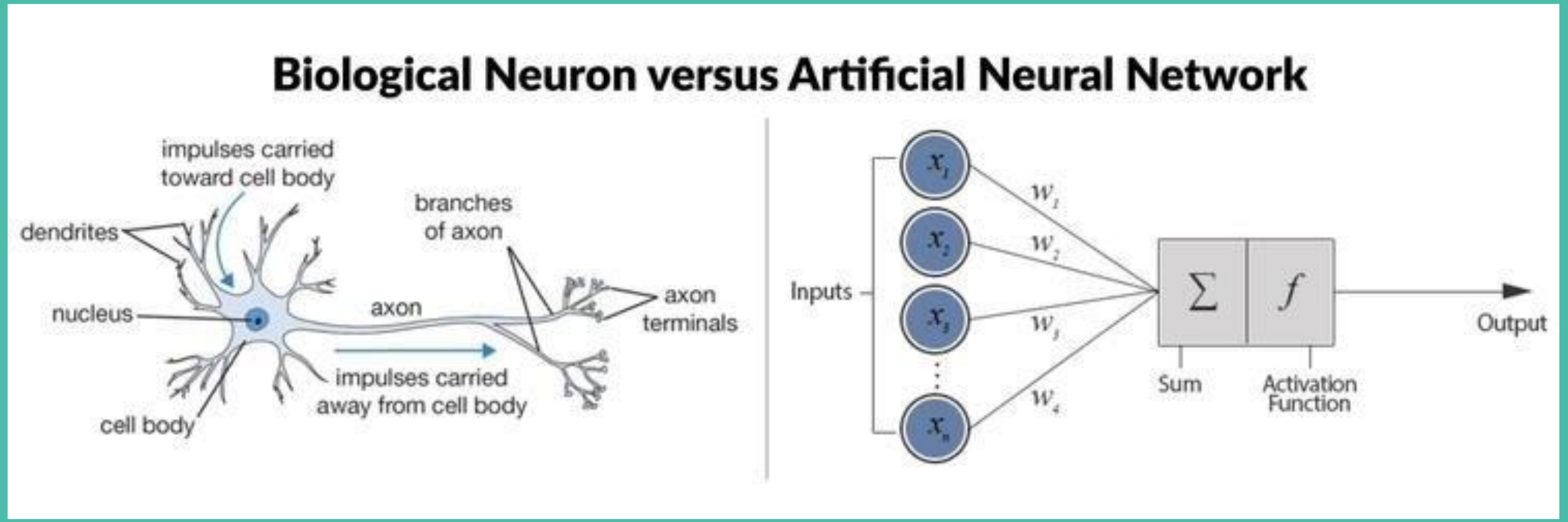
Koneoppiminen

- Tekoälyn osa-alue, missä tietokoneelle luodaan kyky oppia tiedosta
  - Ei sääntöpohjainen järjestelmä
  - Yleisin tekoälyn muoto
  - Opitaan askel kerrallaan toistamalla
  - Oppiminen on vain yksi osa älykkyyttä

- Jaetaan yleensä kolmeen kategoriaan:
  - Ohjattu oppiminen
  - Vahvistusoppiminen
  - Ohjaamaton oppiminen

Ohjattu oppiminen

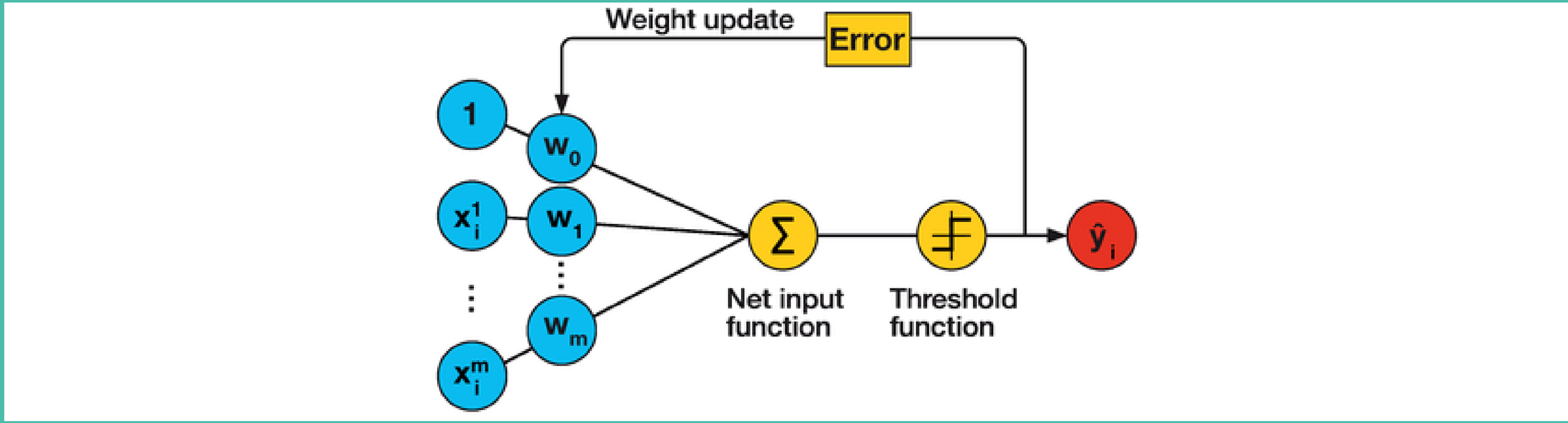
- Datassa on valmiina syöte ja haluttu ulostulo
- Oppiminen tapahtuu säätämällä mallia kohti haluttua tulosta



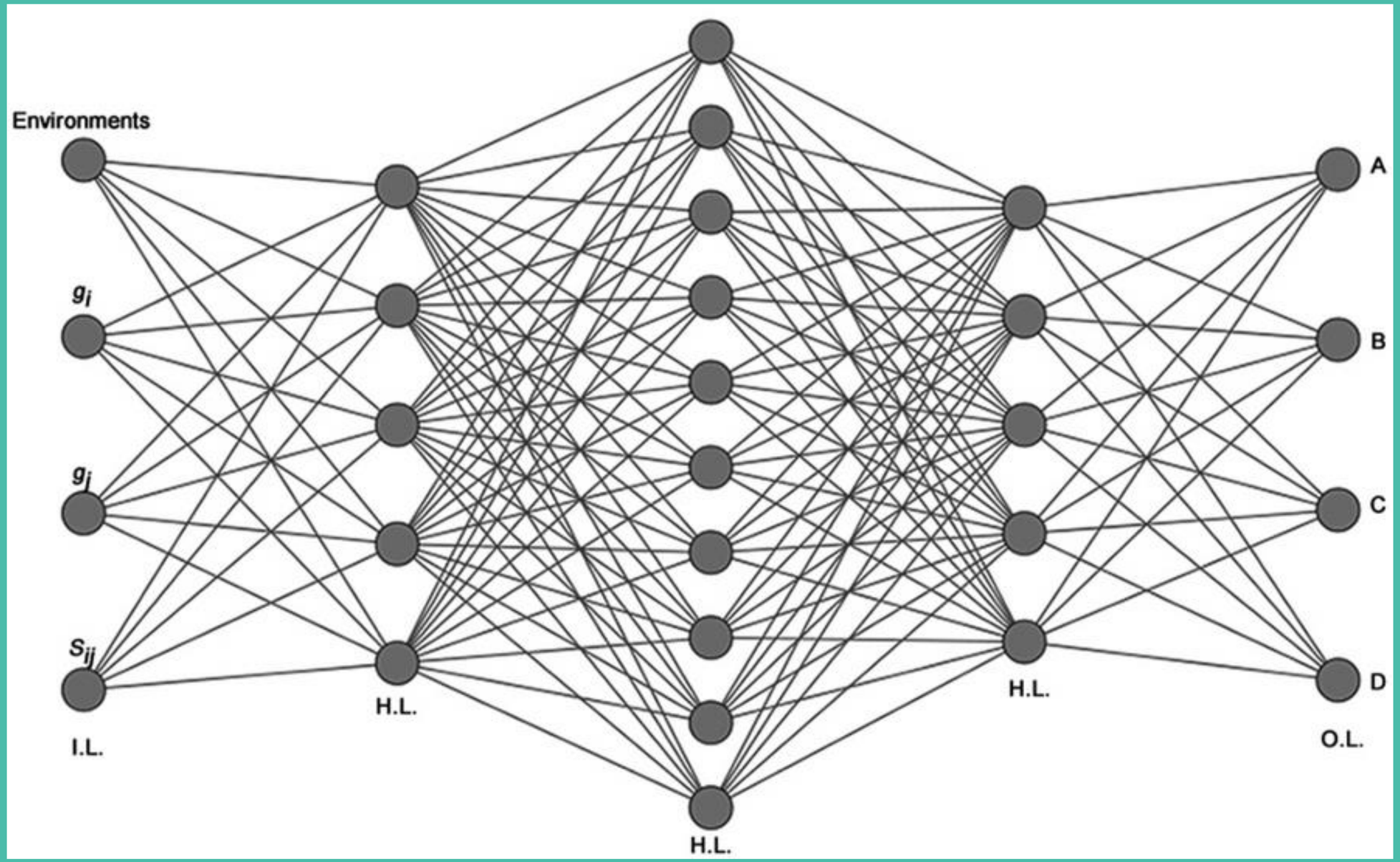
Mwandau, Brian & Nyanchama, Matunda. (2018). Investigating Keystroke Dynamics as a Two-Factor Biometric Security.

Painoja säädellään virheen mukaan

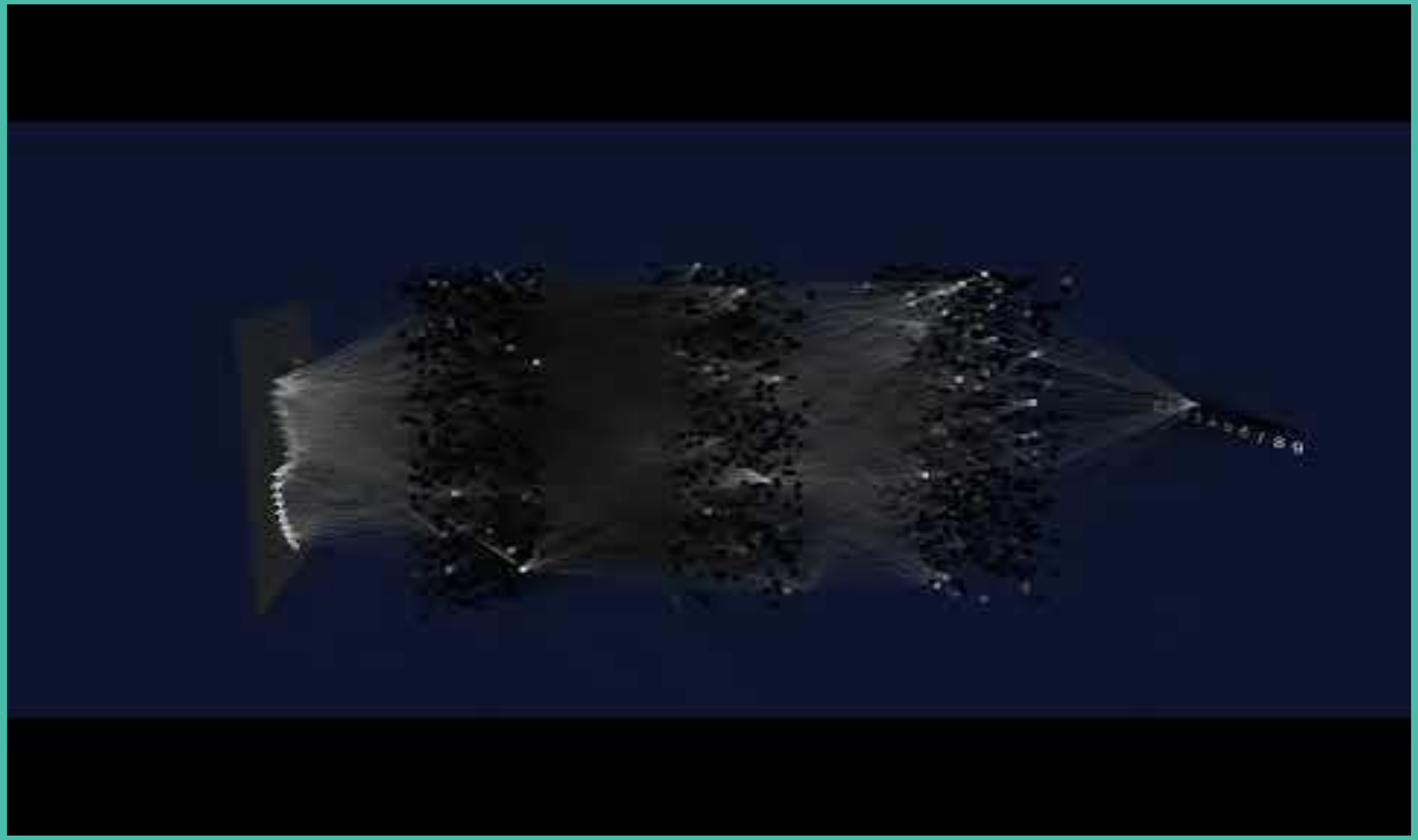
- Opitaan iteratiivisesti virheistä



Ceneda, Nicolo. (2020). Quantile Regression of High-Frequency Data Tail Dynamics via a Recurrent Neural Network. 10.13140/RG.2.2.17219.02086.



*Inocente, G., Garbuglio, D.D., & Ruas, P.M. (2022) Multilayer perceptron applied to genotypes classification in diallel studies. Scientia Agricola, 79.*https://doi.org/10.1590/1678-992X-2020-0365

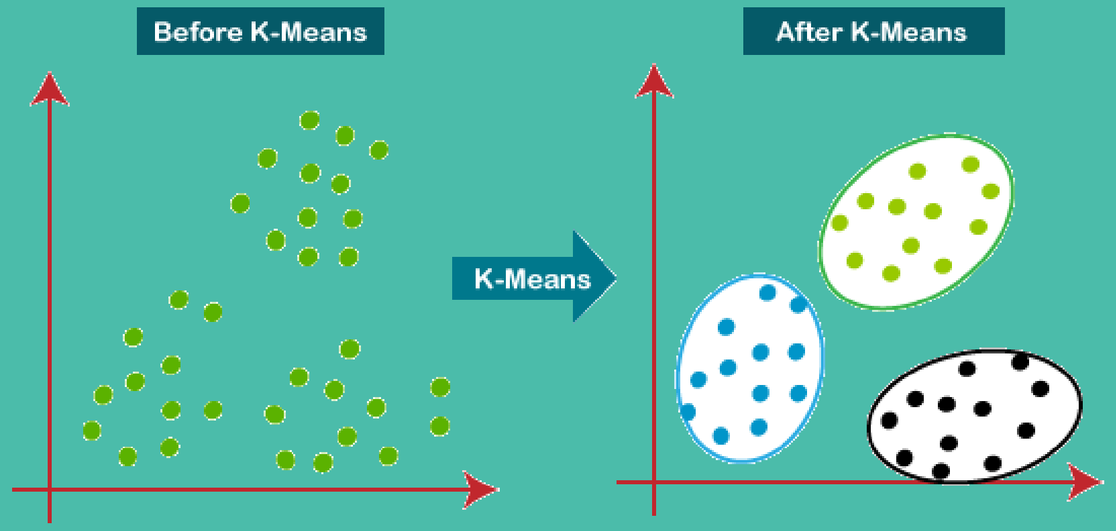


Vahvistusoppiminen

- Agentti toimii ympäristössä jota se pystyy mittaamaan ja saa palautetetta toimintojen perusteella
  - Positiivinen palaute, jos agentti toimii järkevästi
  - Negatiivinen palaute, jos agentti tekee huonon valinnan

Ohjaamaton oppiminen

- Ohjaamaton oppiminen: Datassa ei ole valmiina haluttuja ulostuloja
- Tavoitteena on ymmärtää datassa esiintyviä piileviä rakenteita, kuten samankaltaisuudet, ryhmät tai säännönmukaisuudet
  - Esim. Klusterointi K-means menetelmällä



Tiedon laadun ominaisuudet, laadun arviointi ja parantaminen

Tärkeitä tiedon laadun ominaisuuksia koneoppimiseen

Relevanssi

- Tiedon täytyy olla relevanttia ongelman ratkaisemiseksi

Täydellisyys

- Ei (liikaa) puuttuvaa tietoa

Laatu

- Kohina, epätarkkuus, vääristymät heikentävät mallin tarkkuutta

Monimuotoisuus

- Tiedon pitää kattaa mahdollisimman paljon erilaisia tapauksia ja tilanteita

Hyvät metatiedot!

- Ilman tietoa tiedosta on koneoppimismallien kehittäminen mahdotonta

Tiedon laadun arviointi koneoppimisessa

Tarkastelu ja visuaalinen analyysi

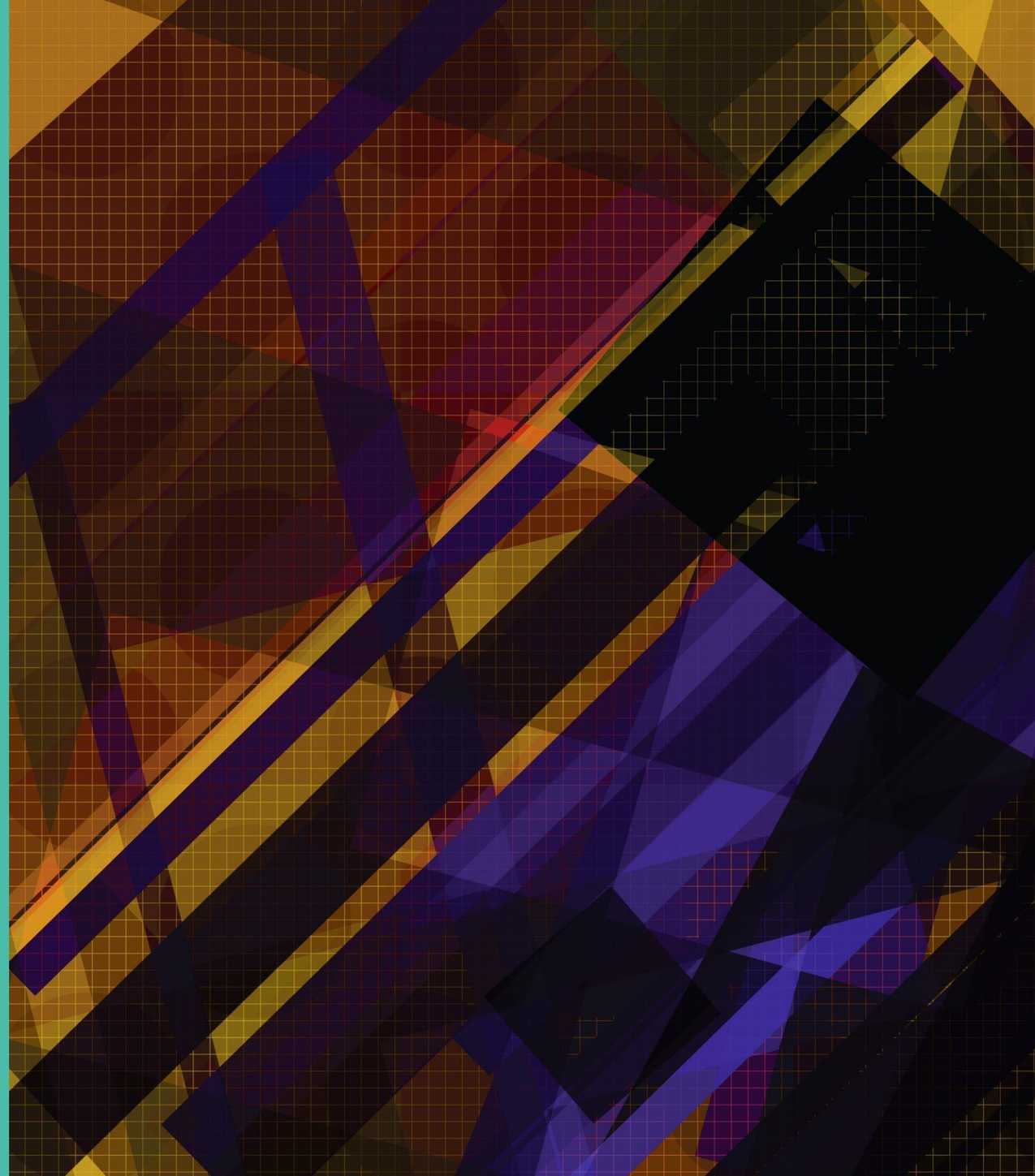
Poikkeavien arvojen tunnistaminen

Tilastolliset mittarit

Tiedon laadun parantaminen esikäsittelyllä

- Datan puhdistus virheistä
  - Poistetaan tai korjataan virheelliset arvot
- Normalisointi
  - Muutetaan kaikki mallin eri syötteet samalle skaalalle
- Luokkien tasapainotus
  - Undersampling, oversampling
- Puuttuvien arvojen imputointi
  - Korvataan puuttuvat arvot esim. toisesta lähteestä

Esimerkkejä huonon tiedon käytöstä koneoppimisessa



Zillow

- Amerikkalainen yritys joka toimii kiinteistövälityksessä
- Automatisoi pitkälle kiinteistöjen hankinnan luottaen omaan AI ratkaisuun
- Malli oli opetettu avoimella ja käyttäjien antamalla datalla ilman riittävää valvontaa

Nov. 2, 2021 Updated Nov. 19, 2021, 8:41 a.m. ET

Daily Business Briefing

Zillow, facing big losses, quits flipping houses and will lay off a quarter of its staff.

The real estate website had been relying on its algorithm that estimates home values to buy and resell homes. That part of its business lost about $420 million in three months.

- Malli teki kolmessa kuukaudessa 420 miljoonan dollarin tappiot
  - Neljäsosa työntekijöistä piti irtisanoa

Amazonin rekrytointiprosessi

- Amazon kehitti koneoppimiseen perustuvan työkalun, jota kutsuttiin nimellä "Automated Talent Acquisition System" (ATAS).
  - Tarkoitus oli auttaa suodattamaan hakijoita CV:n perusteella
- Opetusdata kerättiin 10 vuoden aikana tulleista hakemuksista
- Hakemuksia oli huomattavasti enemmän miehiltä kuin naisilta
  - Malli oppi suosimaan miehiä

Pallo vai kalju pää?

- Skotlantilainen jalkapallojoukkue Inverness Caledonian Thistle FC alkoi käyttää otteluissaan AI kameraa
  - Kamera seuraa automaattisesti palloa pelissä, ja lähettää kuvan streamiin

- Ottelussa olikin kaljupäinen tuomari, ja kamera seurasi tuomaria eikä palloa

Take home messages

Lopputentti:

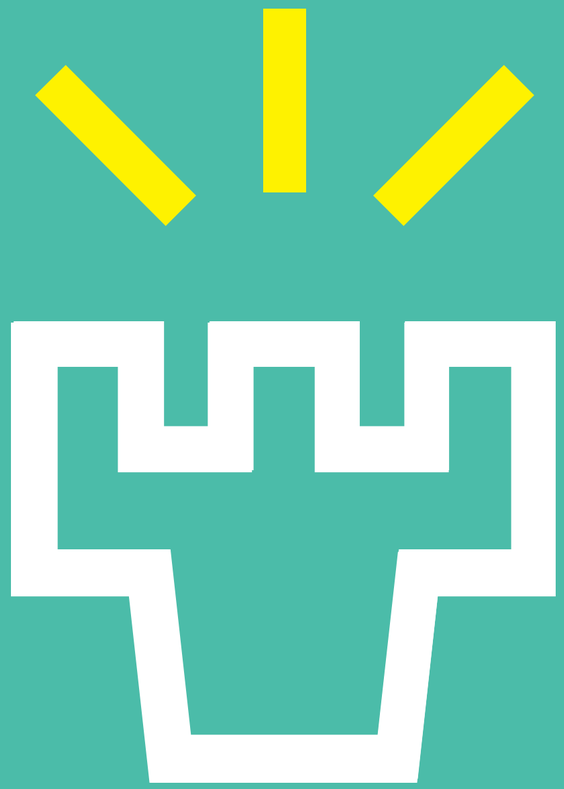
https://pollev.com/miikamalin036

Yhteenveto

– Tekoäly ≠ Koneoppiminen

- Koneoppiminen jaettu kolmeen kategoriaan: Ohjattu oppiminen, vahvistusoppiminen ja ohjaamaton oppiminen

- **Tiedon määrä ei korvaa laatua**



OULUN
YLIOPISTO