

# Työkyvyttömyyseläkkeen ennustaminen koneoppimismenetelmillä

Jarno Varis



**Eläketurvakeskus**  
PENSIONSSKYDDSCENTRALEN

# Eläketurvakeskuksen koneoppimisprojekti

- Yhteistyössä Siili Solutions Oyj:n kanssa (Ilkka Huopaniemi)
- Tarkoituksena hankkia osaamista koneoppimisesta



# Koneoppiminen

- Datan käyttö (ennuste)mallin tekoon. Ennustemallia sovelletaan uuteen, tuntemattomaan dataan
- Samankaltaista kuin monimuuttujamenetelmät tilastotieteessä (luokittimet, regressio, klusterointi, pääkomponenttianalyysi (PCA))



# Koneoppiminen vs. algoritmien perinteinen logiikka

## Perinteinen logiikka

- Asiantuntija määrittelee säännöt, ohjelmoija koodaa
- If () then

## Koneoppiminen

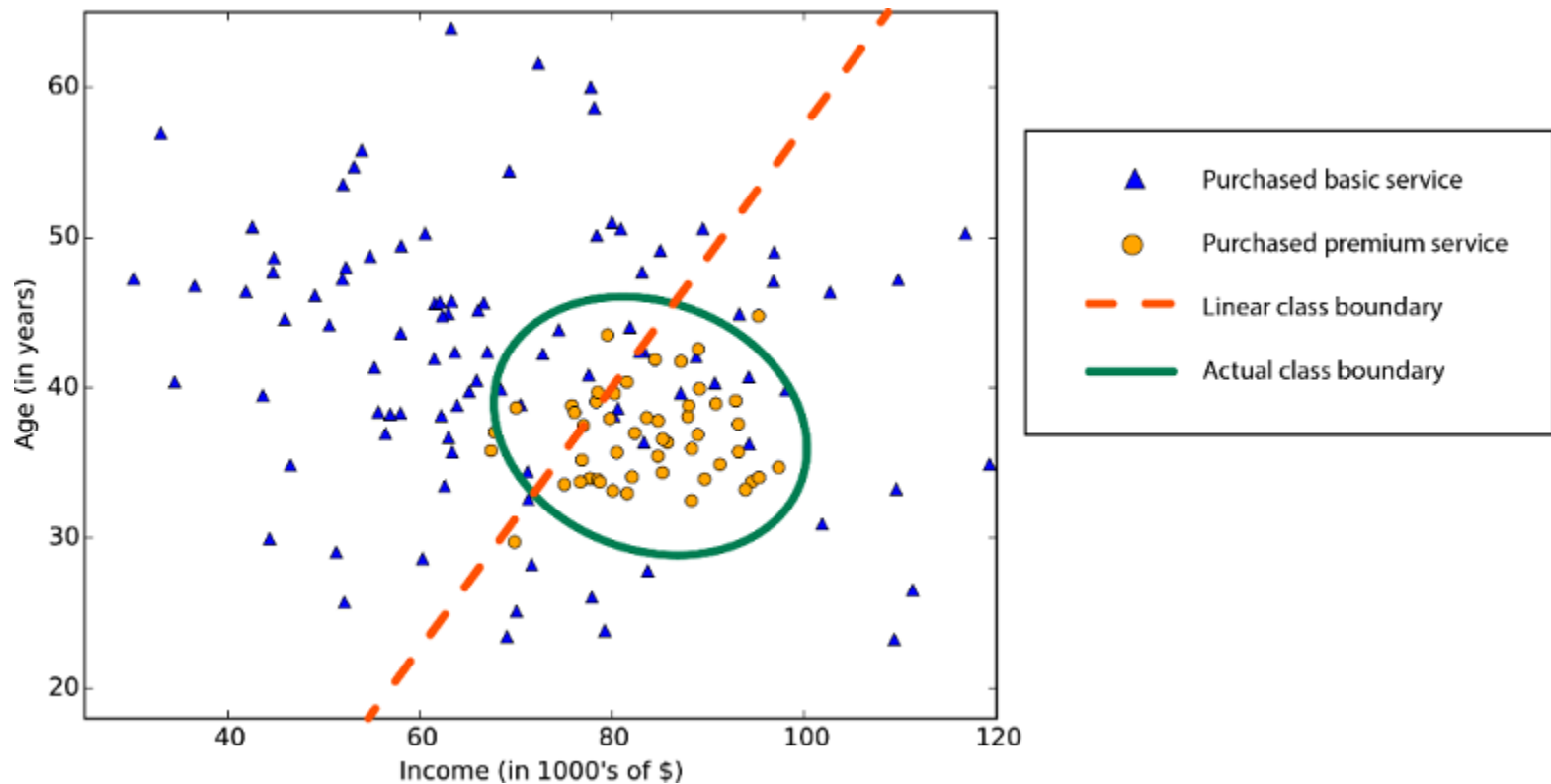
Asiantuntija lokeroi (labeloi) datan  
(mielellään paljon)

Kaikki potentiaalisesti järkevät muuttujat mukaan

Kone oppii mallin ja tarvittavat säännöt datasta automaattisesti



# Lähde: Microsoft, Google kuvahaku



# Milloin koneoppiminen voi olla paras lähestymistapa?

- Jos säännöt vaikea määritellä  
(erota kissat koirista kuva-aineistossa)
- Sääntöjä ei vielä tunneta (koneoppiminen löytää ne)
- Lokeroitua (labeloitua) esimerkkidataa saatavilla paljon
- Onko helpompaa antaa koneelle esimerkkidataa kuin määritellä säännöt kattavasti?



# TK-ennustustehtävä

- Ennustetaan työkyvyttömyyseläkkeen (TK):n alkamista 2 vuoden päästä
- Esim. jos TK alkoi 2013, käytetään kaikkea dataa 2011 ja siitä taaksepäin
- Jaetaan data opetusdataan, jolla luokitin opetetaan ja testidataan joka jää ulkopuolelle
- Ennustustarkkuus määritetään kokeilemalla opetettua luokitinta testidataan



# Määritettävät tulokset

- Vertaillaan eri luokittimia (tarkkuuksia)
- Määritetään parhaiten ennustavat muuttujat
- Verrataan väärin menneitä ennusteita ICD-luokitukseen (=erilaiset sairausdiagnoosit) ja hylättyihin TK-päätöksiin





# Data

- 240 000 Työkyvyttömyyseläkeläistä
- 240 000 Verrokkia (ei työkyvyttömyyseläkettä)
- Data 2 vuotta ennen TK-päätöstä ja siitä taaksepäin
- Ansiot ja saadut etuudet vuosittain 10-v ajalta, sukupuoli, ikä, yksityisen sektorin työnantajan toimiala, ammatti, koulutus, kuntoutusraha, sosioekonominen tilanne, ansiolaki, siviilisääty, asuinkunta
- 2005-2017

## Harhojen poisto

- Verrokeilla aloitusvuosi arvotaan satunnaisesti
- Poistettu yli 61v, alle 18v, kuolleet ennen 2005, eläkkeellä, puuttuvaa dataa, sose-eläkeläiset, vanha tk ennen 2005, ulkomailla asuvat, asuinkunta puuttuu
- Tiedot löydyttävä ansaintarekisteristä



## Datan laatu

- 240 000 TK-tapausta riittävä ja erittäin hyvä määrä. Verrokkeja tarvittaessa jopa noin 4-5 milj.
- Missing-at-random puuttuvia arvoja ei käytännössä ollenkaan, tämä poikkeuksellisen hyvä tilanne
- Puuttuvia blokkeja jonkin verran (<10%), tietyiltä henkilöiltä puuttuu aina samat muuttujat. Näille pitää (ja pitäisi olla mahdollista) ymmärtää selitys ja päättää sitten niiden poistamisesta jottei tulokseen tule harhaa
- Datan määrän ja laadun suhteen ei estettä parantaa tarkkuutta nykyisestä vielä merkittävästi



## Myös muuta aineiston käsittelyä

- Liukulukujen normalisointi
- Kategoristen muuttujien muunnos dummy-muuttujiksi
- Puuttuvien arvojen käsittelyä yms.



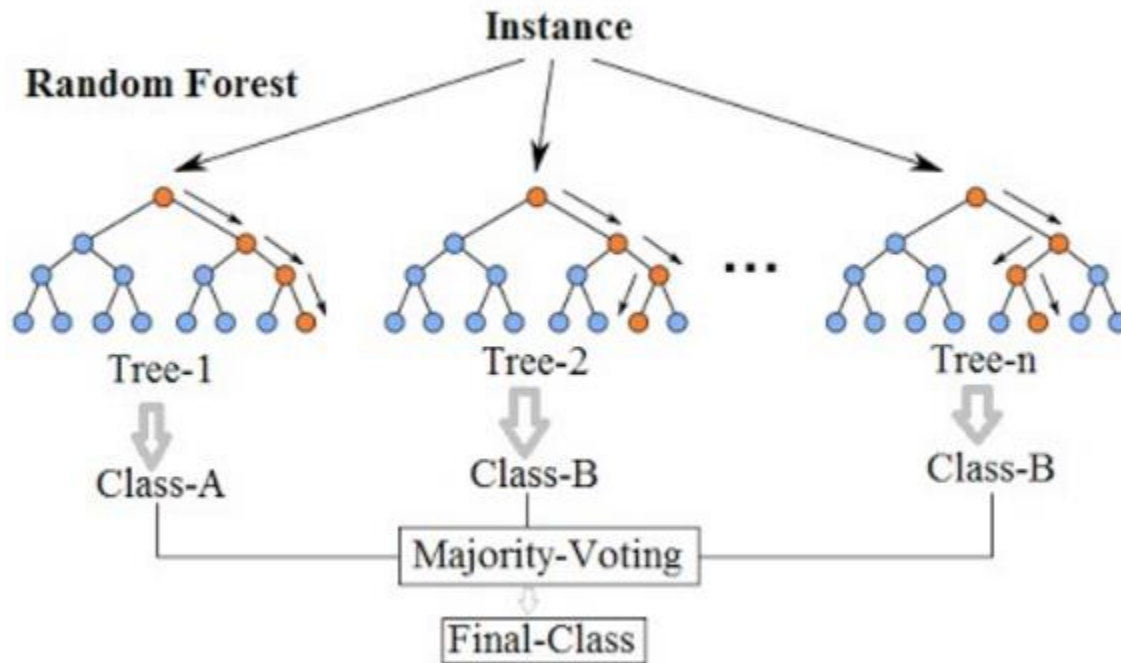
## Tulokset: ennustustarkkuus

- 50% on baseline, eli satunnaisesti arvaamalla saavutettava tarkkuus (240 000 + 240 000)
- Random forest 77.8%
- XGBoost (edistynyt päätöspuu) 77.3%
- Lineaarinen malli: (Logistic regression, L1 regularisointi) 77.0%
- Päätöspuu 69%
- Neuroverkko 78%



Lähde: [medium.com](https://medium.com/@williamkoehrsen), [William Koehrsen](#)

## Random Forest Simplified



# Ennustavat muuttujat

Random forest: suunta ei saatavilla

1. ika (0.119959)
2. etuusPaivat120 (0.067701) 120 Sairauspäiväraha
3. etuusLaji120 (0.046804) 120 Sairauspäiväraha
4. ansiot0 (0.035628)
5. ansiot1 (0.030332)
6. ansiot2 (0.027132)
7. ansiot3 (0.025374)
8. ansiot4 (0.024366)
9. ansiot10 (0.023391)
10. ansiot5 (0.023096)
11. ansiot6 (0.022786)
12. ansiot8 (0.022607)
13. ansiot9 (0.022399)
14. ansiot7 (0.022279)
15. etuusPaivat151 (0.020729) 151 Työmarkkinatuki
16. ansiotPv0 (0.017700)
17. ansiotPv10 (0.015096)
18. ansiotPv9 (0.013473)
19. ansiotPv1 (0.011985)
20. ansiotPv8 (0.011831)



# TK:ta ennustavat muuttujat

## Lineaarinen malli (logistic)

1. ika (0.963152)
2. etuusPaivat120 (0.522155) 120 Sairauspäiväraha
3. etuusLaji120 (0.452786) 120 Sairauspäiväraha
4. sose8 (0.182078) Muut (työttömät ja varusmiehet)
5. onOllutKura (0.171518)
6. etuusLaji151 (0.116641) 151 Työmarkkinatuki
7. etuusLaji130 (0.107410) 130 Kuntoutusrahalain mukainen kuntoutusraha
8. koulutus2 (0.101541) Ei koulutusta
9. sose5 (0.098094) Työntekijät
10. ansiot10 (0.090435)
11. koulutus3 (0.090281) Keskiaste
12. etuusLaji121 (0.087191) 121 Osittainen sairauspäiväraha
13. etuusPaivat301 (0.083538) 301 Liikennevakuutuksen ansiomenetykskorvaus
14. tolK (0.065357) K Rahoitus- ja vakuutustoiminta
15. etuusPaivat300 (0.061091) 300 Tapaturmavakuutuksen ansiomenetykskorvaus (työajan tapaturma)
16. sose6 (0.058441) Opiskelijat
17. tolR (0.055375) R Taiteet, viihde ja virkistys
18. tolC (0.055273) C Teollisuus
19. ltunnus1\_2 (0.042594) 2 = TYELB, kuukausitilitteiset TYEL työsuhteet/sopimustyönantaja ja LEL-tilitykset
20. etuusPaivat121 (0.041700) 121 Osittainen sairauspäiväraha



# Ei-tk:ta ennustavat muuttujat

## Lineaarinen malli (logistic)

1. ansiot0 (-0.204212)
2. SIVIILISAATY\_0.0 (-0.159515) Ei tietoa
3. etuusLaji210 (-0.128780) 210 Ansioon suhteutettu päiväraha
4. ansiotPv0 (-0.098233)
5. etuusLaji102 (-0.097883) 102 Isyysraha
6. etuusPaivat210 (-0.090645) 210 Ansioon suhteutettu päiväraha
7. SIVIILISAATY\_2.0 (-0.086502) Avioliitossa
8. tol64 (-0.085399) 64 Rahoituspalvelut (pl. vakuutus- ja eläkevakuutustoiminta)
9. ltunnus1\_0 (-0.076018) nan
10. tol55 (-0.075723) 55 Majoitus
11. tolG (-0.072032) G Tukku- ja vähittäiskauppa, moottoriajoneuvojen ja moottoripyörien korjaus
12. tol56 (-0.067687) 56 Ravitsemistoiminta
13. tol10 (-0.066769) 10 Elintarvikkeiden valmistus
14. etuusLaji110 (-0.066501) 110 Kotihoidon tuki
15. tolQ (-0.065643) Q Terveys- ja sosiaalipalvelut
16. tol26 (-0.065602) 26 Tietokoneiden sekä elektronisten ja optisten tuotteiden valmistus
17. tol92 (-0.065378) 92 Rahapeli- ja vedonlyöntipalvelut
18. ltunnus0\_Y (-0.063458) Y = YEL
19. etuusLaji100 (-0.061323) 100 Äitiysraha
20. koulutus7 (-0.060575) Ylempi korkeakouluaste



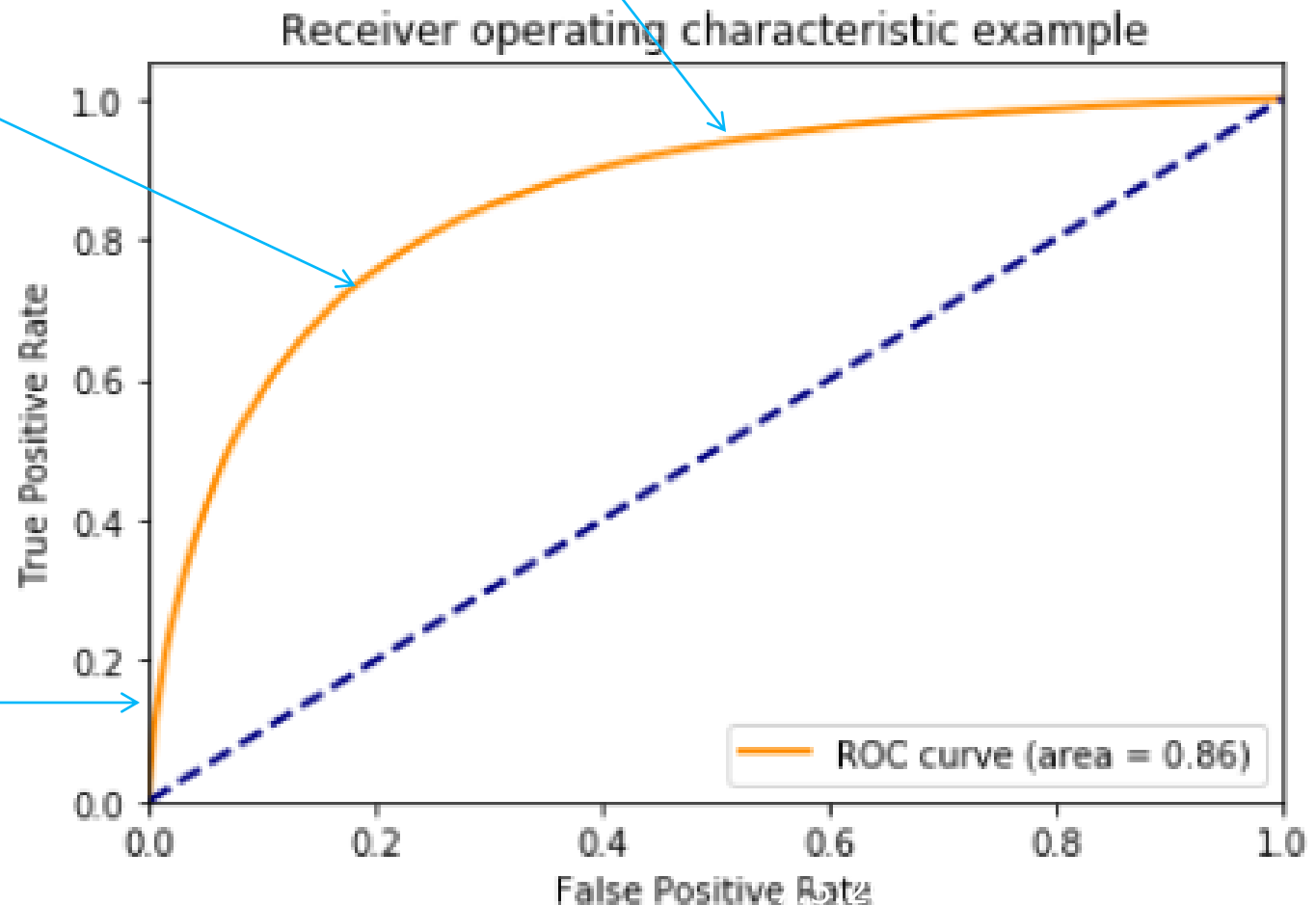


# Ovatko false positivet haitallisempia kuin false negativet?

**Maksimi: löytääksemme 78%  
TK:ista teemme 18% virheellisiä  
TK-ennusteita**

Löytääksemme 90% TK:ista myös  
60%  
terveistä tehdään virheellinen TK-  
ennuste

15%  
suurimmassa  
riskissä  
olevat henkilöt  
vain <1% vääriä  
hälytyksiä



## Kirje kotiin suurimmassa riskissä oleville?

- 15% suurimmassa riskissä olevat, 1% väärää hälytyksiä
- Todellisuudessa kuitenkin 4 milj ei-TK tapausta 100 000 TK-tapausta
- 15% 100 000 = 15 000 oikeaa hälytystä
- 1% 4 000 000 = 40 000 väärää hälytystä
  
- Mallien tarkkuutta kannattaa vielä parantaa käytännön sovelluksia varten



# ICD10 diagnoosit

## Ennustamisen jälkiprosessointina tutkittu ennustustarkkuutta per diagnoosi Diagnoosidataa ei käytetty ennustamiseen

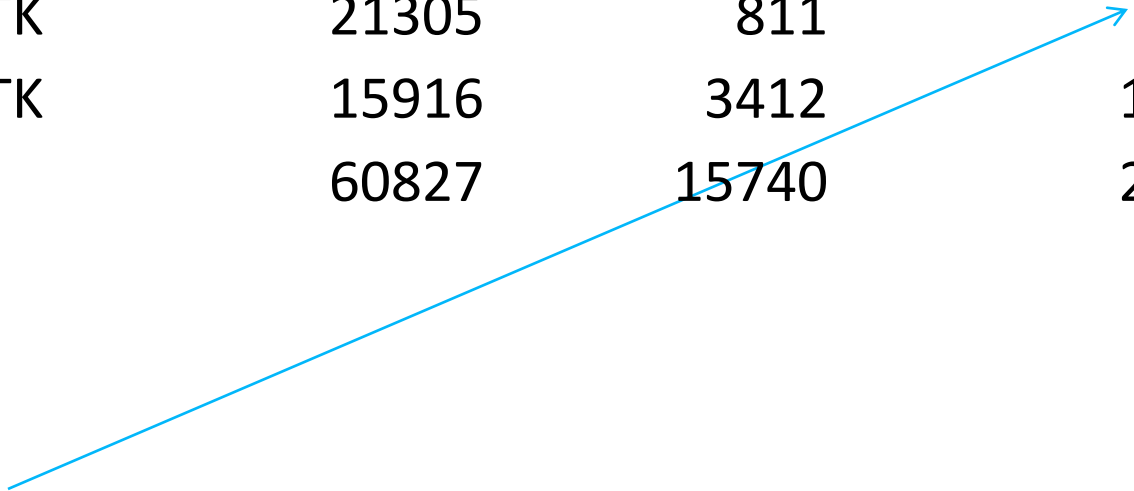
acc=0.73% ['AB', 'Tartunta- ja loistauteja'] n=173  
acc=0.73% ['CD', 'Kasvaimet'] n=6288  
acc=0.71% ['D', 'Veren ja verta muodostavien elinten sairaudet'] n=279  
acc=0.84% ['E', 'Umpierityssairaudet, ravitsemussairaudet ja aineenvaihduntasairaudet'] n=1679  
acc=0.73% ['F', 'Mielenterveyden ja käyttäytymisen häiriöt'] n=27933  
acc=0.76% ['G', 'Hermoston sairaudet'] n=7103  
acc=0.74% ['H', 'Silmän ja sen apuelinten sairaudet'] n=694  
acc=0.80% ['HH', 'Korvan ja kartiolisäkkeen sairaudet'] n=574  
acc=0.82% ['I', 'Verenkiertoelinten sairaudet'] n=7010  
acc=0.92% ['J', 'Hengityselinten sairaudet'] n=1792  
acc=0.80% ['K', 'Ruuansulatuselinten sairaudet'] n=846  
acc=0.86% ['L', 'Ihon ja ihonalaiskudoksen sairaudet'] n=437  
acc=0.89% ['M', 'Tuki- ja liikuntaelinten sekä sidekudoksen sairaudet'] n=31245  
acc=0.70% ['N', 'Virts- ja sukupuolielinten sairaudet'] n=339  
acc=0.67% ['O', 'Raskaus, synnytys ja lapsivuoteus'] n=6  
acc=1.00% ['P', 'Eräät perinataaliaikana alkaneet tilat'] n=4  
acc=0.65% ['Q', 'Synnynnäiset epämuodostumat, epämuotoisuudet ja kromosomipoikkeavuudet'] n=398  
acc=0.79% ['R', 'Muualla luokittelemattomat oireet, sairaudenmerkit sekä poikkeavat kliiniset ja laboratoriolöydökset'] n=428  
acc=0.77% ['ST', 'Vammat, myrkytykset ja eräät muut ulkoisten syiden seuraukset'] n=5453  
acc=0.76% ['Z', 'Tekijöitä jotka vaikuttavat terveydentilaan ja yhteydenottoihin terveyspalvelujen tuottajiin'] n=477

**Tuki- ja liikuntaelinsairaudet ja hengityselinsairaudet voidaan ennustaa parhaiten**



## Hylättyjä päätöksiä takana vs. ennusteet

	ei-hylättyjä	hylättyjä>0	#:lla hylättyjä
Ei TK, ennustettiin Ei TK	70339	304	0.430
Ei TK, ennustettiin TK	21305	811	3.667
TK, ennustettiin Ei TK	15916	3412	17.653
TK, ennustettiin TK	60827	15740	20.557



Tekoäly ennustanut TK:n, henkilö saanut hylättyjä päätöksiä.  
Tapauksia suhteessa vähän.

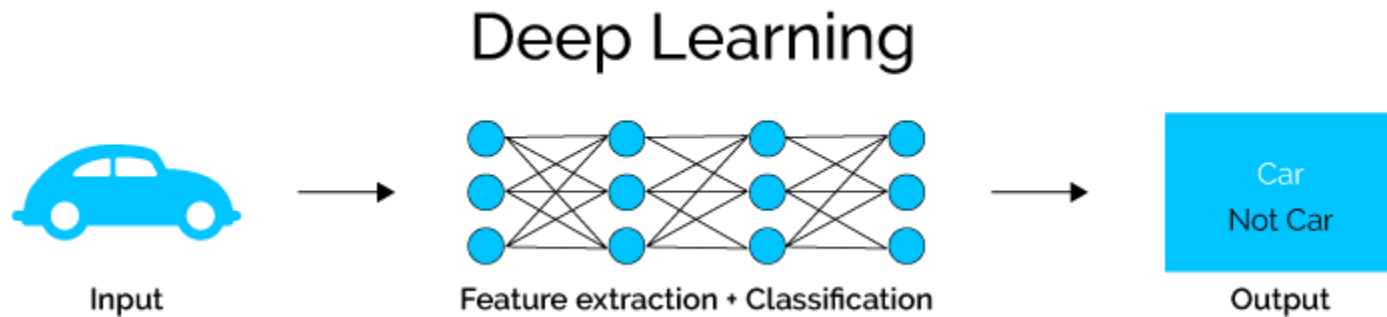
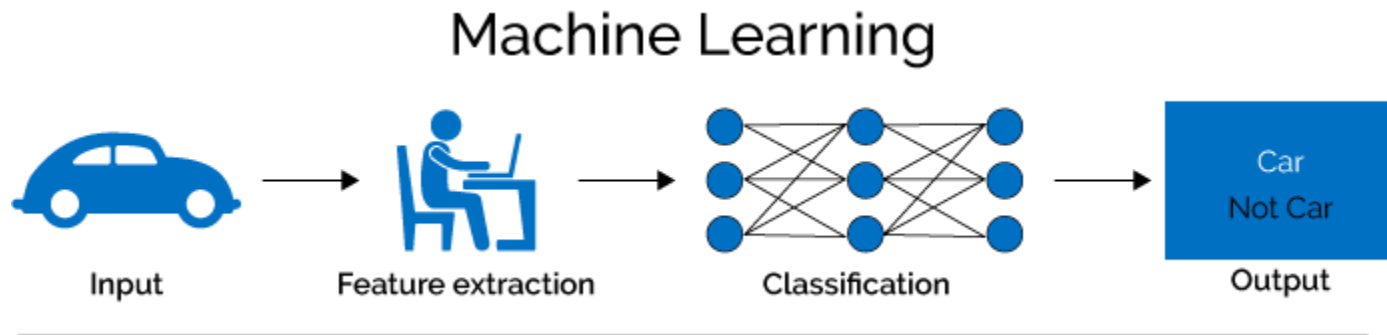


# Deep learning

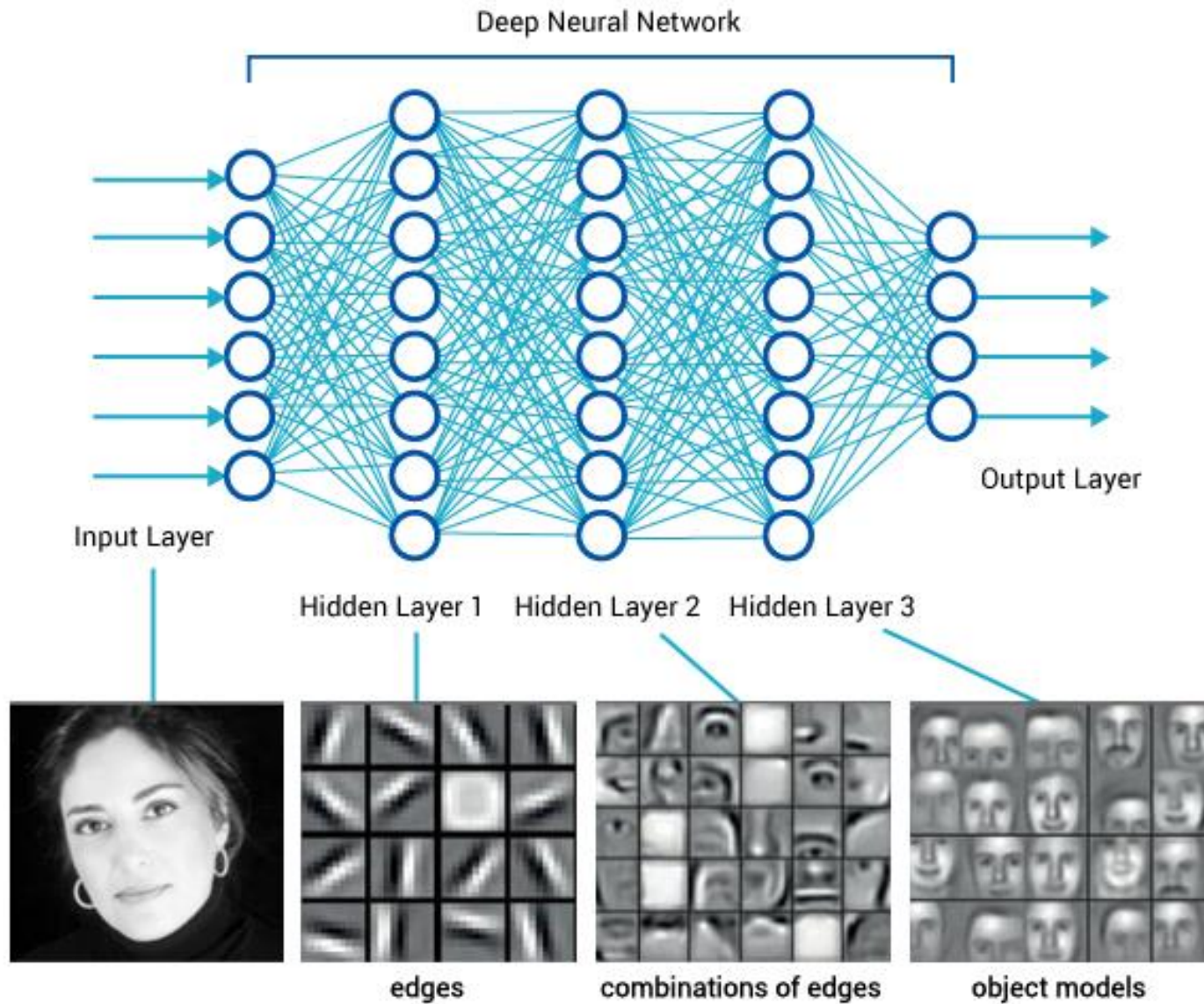
- Kokeiltiin yksinkertaisia 1-5 piilotetun tason neural networkeja
  - Luokittelutarkkuus sama, n. 78%
  - Ei vielä saavutettu parannusta, mutta kokeilut alustavia
  - Tilanne voi muuttua myös uuden datan myötä
- 
- Ennustavien muuttujien tulkitseminen ei tunnetusti ole välttämättä helppoa
- 
- Keras-kielellä deep learning erittäin nopeaa ja helppoa



# Deep learning automates feature extraction



# Deep learning layers



# Riskit

- Datassa edelleen biaksia.
- Esim. jos henkilö saanut TK:n, sosioekonomista asemaa kuvaava luokka "eläkeläinen" saattaa paljastaa henkilön eläkeläisstatuksen koneelle. Luokitin havaitsee tämän
- Kävimme aineistoa tarkasti läpi vastaavien ongelmakohtien vuoksi





# Johtopäätökset

- Datassa on ennustusvoimaa ja saadaan ennustettua TK 2 vuoden päähän 78% tarkkuudella
- Potentiaalia ennustustarkkuuden parantamiseen on paljonkin (terveystiedot, mallien parantaminen)
- Parhaiten TK:ta ennustavat ikä, saadun sairauspäivärahan määrä, ansiotaso



## Lopullinen validointi olisi:

- Longitudinal trial. Ennustetaan 2016 datan perusteella ketkä jäävät 2018 TK:lle. 2017 perusteella ketkä jäävät 2019.
- Onnistumisen edellytys, että datan keruussa tai lainsäädännössä ei isoja muutoksia aiempaan nähden



# Tulevaisuuden suunnat

1. Mallin ja ennustustarkkuuden parantaminen
2. Luokittimien ja muiden mallien käyttö muihin ongelmiin
3. Ennustavan mallin hyödyntäminen käytännössä



# 1. Ennustemallin tarkkuuden parantaminen

Kulmakivet:

- 1.1 Enemmän ja parempaa dataa
- 1.2 Parempi malli



## 1.1 Parempaa dataa

- Lisää muuttujia. Esimerkiksi tarkemmat sairausdiagnoositiedot
- Näytteiden määrä ja datan laatu jo nykyisellään erittäin hyviä



## 1.2. Ennustemallin parantaminen

- Muuttujien esikäsittelyn parantaminen
- Luokittimien parametrien hinkkaaminen
- Aikasarjamallit iso kokonaisuus: muuttujat ja outcomet
  - Muuttujat: Esim. ansioiden kehitys vuosien mittaan: ensin tasaiset, sitten yllättävä pudotus nolnaan ja samalla sairauspäivärahojen kasvu 365pv/vuodessa. Tällainen piirre voisi ennustaa TK:ta todella hyvin.
  - Outcome: Ennustetaan samalla mallilla todennäköisyyttä 1v, 2v, 3v, 4v päähän
- Neuroverkkoja kokeiltu vasta pikaisesti: erityisesti terveystietojen kanssa tulos voi parantua yllättäen ja löytyä kiinnostavia muuttujien yhdistelmiä



## 2. Koneoppimisen hyödyntäminen muihin ETK:n-tutkimuskohteisiin

- Samoja luokittimia voi hyödyntää muissakin tutkimuskohteissa ennustamaan jotain outcomea
- Klusterointi: keskenään samankaltaisten aliryhmien etsiminen ja hyödyntäminen



## 3. Ennusteiden soveltaminen

- Asiantuntijatyön apuväline
  - Aggregaattiennusteet?
  - Tutkimuksen apuväline?
  - Uusien, ennestään tuntemattomien ilmiöiden mallintaminen?





# Loppu, kiitos!

- Hyvä oppimateriaali:
  - Python Machine Learning, Sebastian Raschka, Vahid Mirjalili

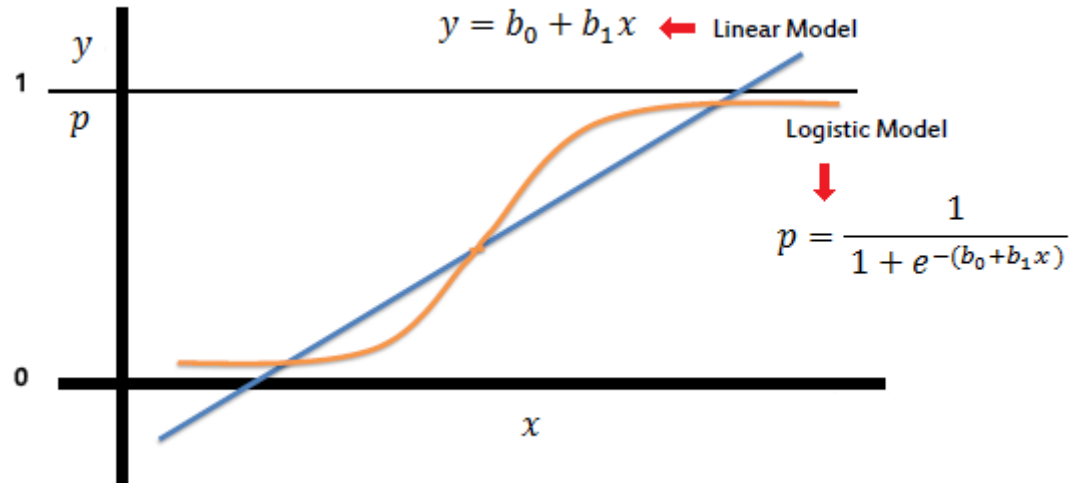


## Liite 2.2. Työkyvyttömyyseläkkeen tasoon vaikuttavat merkittävimmät laskentasäännöt ja muutokset vuodesta 1994 alkaen

Vuosi	Ansaittu työeläke	Tulevan ajan eläkeosa	Työkyvyttömyyseläkkeen tason ylläpitäminen	Vanhuuseläke	Muuta
1994	<ul style="list-style-type: none"> <li>Eläkepalkka 2 v 4:stä</li> <li>Automaattinen katkaisu</li> <li>1,5 % karttuma 23–59 v</li> <li>2,5 % karttuma 60v+</li> </ul>	<ul style="list-style-type: none"> <li>Oikeus, jos asunut 5 v Suomessa ja</li> <li>Jos 360 pv ts:n päättymisestä</li> <li>Tulevan ajan ansio viimeisestä 6 kk jatkuneesta ts:sta</li> <li>Karttumis-% 1,5</li> </ul>	<ul style="list-style-type: none"> <li>Puoliväli-indeksi työeläkkeissä</li> <li>Kansaneläkeindeksi Kelassa</li> </ul>	<ul style="list-style-type: none"> <li>Alin vanhuuseläkeikä 65 v</li> <li>Varhennettu vanhuuseläkeikä 60</li> <li>Lykätty eläke 65+</li> </ul>	<ul style="list-style-type: none"> <li>Vakuuttamisvelvollisuus 14–65 v</li> </ul>
1996	<ul style="list-style-type: none"> <li>Eläkepalkan 10 v keskiansio asteittain voimaan</li> <li>PTEL vähennetään työansioista</li> </ul>	<ul style="list-style-type: none"> <li>10 v aikana 1 v Suomen ansioita ed. lisäksi</li> <li>Karttumisprosentit 1,5/1,2/0,8</li> </ul>	<ul style="list-style-type: none"> <li>Puoliväli-indeksi alle 65 ja alkava työeläke</li> <li>Eläkeikäisen indeksi 65+ työeläkkeisiin</li> </ul>		<ul style="list-style-type: none"> <li>Pohjaosan poisto Kelassa alkoi</li> </ul>
2005	<ul style="list-style-type: none"> <li>Työeläke ansioiden perusteella 18–68 v</li> <li>Karttumisprosentit 1,5/1,9/4,5</li> <li>PTEL (iän mukaan) vähennetään ansiosta</li> </ul>	<ul style="list-style-type: none"> <li>Karttumis-% 1,3 iässä 50–63 v</li> <li>Edellytys 17 455,15 euroa 10 edellisen kalenterivuoden aikana (v. 2017 taso)</li> </ul>	<ul style="list-style-type: none"> <li>Alkava työeläke tarkistetaan palkkakertoimella</li> <li>Työeläkeindeksi kaikkiin työeläkkeisiin</li> <li>Kertakorotus alle 56 v</li> </ul>	<ul style="list-style-type: none"> <li>Alin vanhuuseläkeikä työeläkkeissä 63 (Kela 65), karttuu 68 asti</li> <li>Lykätty työeläke 68+ (Kela 65+)</li> <li>Varhennettu vanhuuseläke 62 v iästä</li> <li>Työeläkettä eläkeaikaisesta työstä</li> </ul>	<ul style="list-style-type: none"> <li>Vakuuttamisvelvollisuus 18–68 v</li> <li>Työttömyyseläke ja YVE lakkaa</li> <li>Vähentämätön vanhuuseläke 62-v., jos työtön</li> </ul>
2010	<ul style="list-style-type: none"> <li>Elinaikakerrointa sovellettiin työeläkkeissä 1. kerran</li> </ul>	<ul style="list-style-type: none"> <li>VEKL mukaan tulevan ajan ansioon</li> <li>Karttumis-% 1,5</li> </ul>	<ul style="list-style-type: none"> <li>Kertakorotusta korotettiin heti, kun sitä alettiin soveltaa, muuten kuten yllä</li> </ul>		<ul style="list-style-type: none"> <li>TyEdL voimaan</li> </ul>
2011					<ul style="list-style-type: none"> <li>Takuueläke voimaan</li> </ul>
2013					<ul style="list-style-type: none"> <li>VAVE poistuu</li> <li>Vähentämätön vanhuuseläke 62-v. poistuu 1957 jälkeen syntyneiltä (Kelassa vielä 64-v. ennen 1962 syntyneillä)</li> </ul>
2017	<ul style="list-style-type: none"> <li>Karttumis-% 1,5</li> <li>4,5 % muuttuu lykkäyskorotukseksi</li> <li>Ei PTEL-vähennystä työansioista</li> <li>Elinaikakerroin lievenee 2027–</li> </ul>	<ul style="list-style-type: none"> <li>Oikeus, jos 17 455,15 e 10 ed. vuoden aikana (v. 2017 taso)</li> <li>Pääteikä nousee</li> <li>Karttumis-% 1,5</li> <li>Elinaikakerroin tulevaan aikaan 2027–</li> </ul>	<ul style="list-style-type: none"> <li>Työeläkkeen indekseistä ei vähennetä enää työntekijän työeläkemaksun muutosta</li> </ul>	<ul style="list-style-type: none"> <li>Alin eläkeikä sidotaan elinaikaan</li> <li>Lykätty työeläke alimmasta vanhuuseläkeikästä</li> <li>Kelassa eläkeikä kuten työeläkkeessä 2030 alkaen</li> </ul>	<ul style="list-style-type: none"> <li>Vakuuttamisvelvollisuus 17 v. yläraja työntekijöillä (18 v. yläraja yrittäjillä)</li> <li>Työuraeläke 2018–</li> <li>OVE korvaa osa-aikaeläkkeen</li> </ul>



# Logistinen regressio



Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

[Lähde: http://www.saedsayad.com/logistic\\_regression.htm](http://www.saedsayad.com/logistic_regression.htm)



# <https://www.quora.com/What-exactly-is-a-logistic-regression-algorithm-in-machine-learning-What-are-its-applications>

In general the error function in logistic regression is given by :

$$\begin{aligned} J(\theta) &= \frac{1}{m} \sum_{i=1}^m \text{Cost}(h_{\theta}(x^{(i)}), y^{(i)}) \\ &= -\frac{1}{m} \left[ \sum_{i=1}^m y^{(i)} \log h_{\theta}(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_{\theta}(x^{(i)})) \right] \end{aligned}$$

Don't let this function scare you. Here **m = no of elements in training set i.e. 7 for us**(refer training set above), **y is simply 1 or 0** and **h(x) is nothing but the 'Sigmoid function'** we talked about above. Since **sigmoid is a function of theta**(refer what is z above), **hence J is a function of theta.**



## 1.1.11. Logistic regression

As an optimization problem, binary class L2 penalized logistic regression minimizes the following cost function:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^n \log(\exp(-y_i(X_i^T w + c)) + 1).$$

Similarly, L1 regularized logistic regression solves the following optimization problem

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^n \log(\exp(-y_i(X_i^T w + c)) + 1).$$

w=coefficient vector,  
X=design matrix (data, explanatory variables),  
yi=outcome (from training data, variable to be predicted, binary value 0 or 1)



# Projektissa: Sklearn LogisticRegression, solver: liblinear

## 2. Large Linear Classification (Binary and Multi-class)

LIBLINEAR supports two popular binary linear classifiers: LR and linear SVM. Given a set of instance-label pairs  $(\mathbf{x}_i, y_i), i = 1, \dots, l$ ,  $\mathbf{x}_i \in R^n$ ,  $y_i \in \{-1, +1\}$ , both methods solve the following unconstrained optimization problem with different loss functions  $\xi(\mathbf{w}; \mathbf{x}_i, y_i)$ :

$$\min_{\mathbf{w}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^l \xi(\mathbf{w}; \mathbf{x}_i, y_i), \quad (1)$$

where  $C > 0$  is a penalty parameter. For SVM, the two common loss functions are  $\max(1 -$

<https://www.csie.ntu.edu.tw/~cjlin/papers/liblinear.pdf>

