Koolitus „Andmeteadus on Popp“

Andmeteadus, Masinõpe ja tehisintellekt

04. november 2021

Kristjan Eljand

Euroopa Liit
Euroopa Sotsiaalfond

Eesti
tuleviku heaks

Päevakava

Teema	Sisu
Andmeteadus, masinõpe ja tehisintellekt	Mis on andmeteadus, masinõpe ja tehisintellekt. Kuidas masinõpe töötab ehk mis on musta kasti sees. Miks me tehisintellekti nii kaua ootama pidime!? Kuna kasutada kirjeldavat analüütikat, statistikat ja masinõpet. *Paus*
Tehisintellekti olemuslikud riskid	Privaatsus ja andmekaitse. Eetilised dilemmad ja tehisintellekti kallutatuse oht. Generatiivsed mudelid ehk võltssisu loomine.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Minust

- 2006 - 2011 – Majandusteadus, korporatiivrahandus, väärtpaberiinvesteeringud.
- 2007-2011 – Töö Civitta ärianalüütikuna.
- 2010 - 2015 – Andmeteadus OÜ ja konkurentsikaart.ee
- 2016 - 2019 – Töö STACCi tegevjuhina (45+ andmeteadurit).
- 2019-… – Töö Eesti Energia Technology Scoutina.

Mis on andmeteadus, masinõpe ja tehisintellekt?

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Jube kasulik idee

Kiirus 6 km/h

Kiirus 12 km/h

Kohalik omavalitsus on pannud videokaamerad kõigi kergliiklusteede äärde ja soovib rakendust, mis ütleks, millega inimesed seal tegelevad.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Actionfy 1.0

Kiirus 6 km/h

Kiirus 12 km/h

```
If (kiirus <= 6) {
    tegevus = kõndimine
} else {
    tegevus = jooksmine
}
```


Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Uued huvilised

Kiirus 6 km/h

Kiirus 25 km/h

Kiirus 12 km/h

Mitte väga kauges tulevikus tekivad teedele rattasõidu huvilised.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Actionfy 2.0

Kiirus 6 km/h

Kiirus 25 km/h

Kiirus 12 km/h

```
If (kiirus <= 6) {
    tegevus = kõndimine
} else if (kiirus<=12) {
    tegevus = jooksmine
} else {
    tegevus = rattasõit
}
```


Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Uued huvilised

Kiirus 6 km/h

Kiirus 25 km/h

Kiirus 0 km/h

Tuleb välja, et tegevus kergliiklusteedel on mitmekesisem, kui võis algselt arvata.

Kiirus 77 km/h

Kiirus 17 km/h

Kiirus 12 km/h

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Actionfy 3.0

Kiirus 6 km/h

Kiirus 25 km/h

Kiirus 0 km/h

if ...

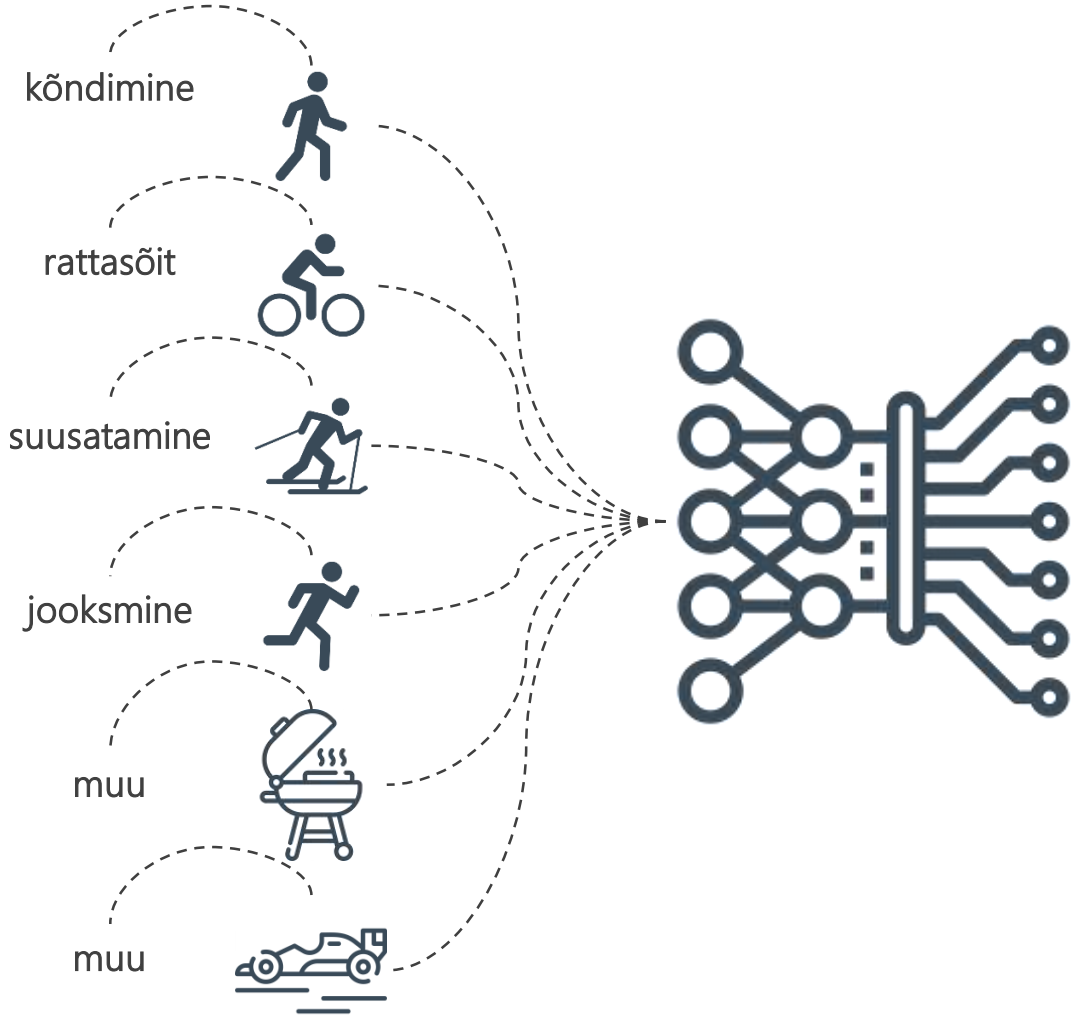
Kiirus 77 km/h

Kiirus 17 km/h

Kiirus 12 km/h

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

ActionAI



Masinõppe puhul **anname** aga ette **näited** tegevustest ja palume mudelil selgeks õppida, kuidas neid tegevusi ära tunda!

Euroopa Liit
Euroopa Sotsiaalfond
Eesti tuleviku heaks

Programmeerimine vs Masinõpe

Programmeerimine

- Selgitame oma soovi **koodi** abil.
- Programmi muutmiseks tuleb see uuesti **programmeerida**.
- Annab **täpse vastuse.**

Masinõpe

- Selgitame oma soovi ~~koodi~~ **näidete** abil.
- Mudeli muutmiseks tuleb see uuesti **treenida.**
- Annab **tõenäosusliku vastuse.**

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Programmeerimine vs Masinõpe

Programmeerimine

- Selgitame oma soovi **koodi** abil.
- Programmi muutmiseks tuleb see uuesti **programmeerida**.
- Annab **täpse vastuse**.

Masinõpe

- Selgitame oma soovi ~~koodi~~ **näidete** abil.
- Mudeli muutmiseks tuleb see uuesti **treenida.**
- Annab **tõenäosusliku vastuse**.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Programmeerimine vs Masinõpe

Programmeerimine

- Selgitame oma soovi **koodi** abil.
- Programmi muutmiseks tuleb see uuesti **programmeerida**.
- Annab **täpse vastuse.**

Masinõpe

- Selgitame oma soovi ~~koodi~~ **näidete** abil.
- Mudeli muutmiseks tuleb see uuesti **treenida.**
- Annab **tõenäosusliku vastuse.**

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Masinõpe on mustrite leidmine näidete baasil!

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Masinõppe liigid – juhendatud õpe

1. Juhendatud õpe (*supervised learning*)
2. Juhendamata õp
3. Stiimulõpe (*rein*

Näide: müügitulu prognoosimine müügitiimi andmete põhjal.

Kampaania	Pühad	Allah %	Müük
0	0	0%	50
1	0	-7%	80
0	0	-5%	60

Numbriline tulemus

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Masinõppe liigid – juhendatud õpe

1. Juhendatud õpe (*supervised learning*)
2. Juhendamata õp
3. Stiimulõpe (*rein*

Näide 2: Katkiste elektriliinide tuvastamine piltide põhjal

Objektide tuvastamine: maa, taevas, elektriliin

Liini seisundi tuvastamine: katki/terve

Euroopa Liit
Euroopa Sotsiaalfond
Eesti tuleviku heaks

Masinõppe liigid – juhendamata õpe

1. Juhendatud õpe (*supervised learning*)

2. Juhendamata õpe (*unsupervised learning*)

3. Stiimulõpe (*reinf*

Anname ette andmed ja palume mudelil iseseisvalt leida mustrid.

Energia	PV	Asukoht
Jah	Jah	Tallinn
Jah	Ei	Tartu
Ei	Jah	Tartu
Jah	Ei	Tallinn
Ei	Jah	Põlva

Sisend: kliendiprofiili andmed.

Eesmärk: Jaota kliendid andmete baasil 5 segmenti!

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Masinõppe liigid - stiimulõpe

1. Juhendatud õpe (*supervised lea*
2. Juhendamata õpe (*unsupervised*
3. Stiimulõpe (*reinforcement learni*

NB: Kui teaksime ette, millised tegevused viivad maksimaalse väärtuseni, saaksime treenida mudeli juhendatud õppega!

Laseme mudelil leida tegevusplaani, mis viib meie poolt soovitava tulemuseni.

Keskkond: Go lauamängu simulaator.
Preemia: Võit/Kaotus

Eesmärk: Õppida selgeks, kuidas Go mängus alati võita.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Masinõppe liigid - stiimulõpe

1. Juhendatud õpe (*supervised lea*
2. Juhendamata õpe (*unsupervised*
3. Stiimulõpe (*reinforcement learni*

Laseme mudelil leida tegevusplaani, mis viib meie poolt soovitava tulemuseni.

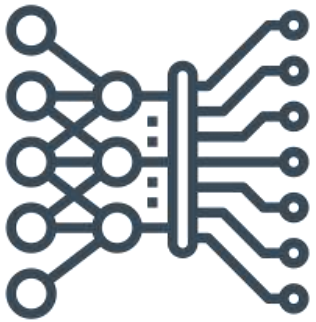
Keskkond: Andmekeskuse HVAC süsteem
Preemia: Jahutusega kaasnev kulu.

Eesmärk: Õppida selgeks tegevusplaan jahutuskulu minimeerimiseks.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Masinõpe ja Tehisintellekt

Masinõpe



=

?

Tehisintellekt

- Tehisintellekt on laiem kui masinõpe, hõlmates ka robootikat.
- Masinõpe näitab protsessi, Tehisintellekt protsessi tulemust.
- Meie koolituse kontekstis: Tehisintellekt = Masinõpe

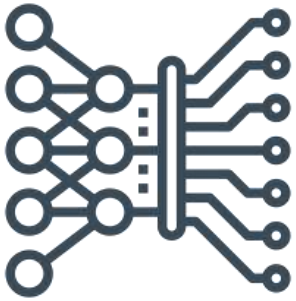
Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Kiire vaade tehisintellekti ajalukku

- 1805 – esimene regressioonmudel (Legrande);
- 1940 – algas tehislike närvivõrkude (Artificial Neural Networks) uurimine;
- 1971 – esimene süvaõppe (Deep learning) mudel (Alexey Ivakhnenko);
- 2005 - … Süvaõppe võidukäik.

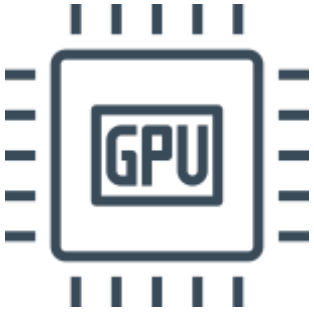
Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Miks me tehisintellekti nii kaua ootama pidime?



Paremad algoritmid

Geoffrey Hinton, Yoshua Bengio, Yann LeCun



Suurem arvutusvõimsus

GPUde kasutamine treenimiseks (Oh ja Jung 2004)

Suuremad andmed

Fei-Fei Li jt.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Ülesanne – millise masinõppeliigiga on tegemist

Kuidas masinõpe töötab, ehk mis on musta kasti sees?

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Kuidas masin õpib: Matemaatika

Väga lihtne matemaatiline valem on kujul:

$$y = m*x$$

y mudeli tulemus (N: müügitulu)

x on mudeli muutuja (N: reklaamikulu)

m on mudeli muudetav parameeter, mis näitab, kui palju muutuja **x** mudeli tulemust mõjutab

Euroopa Liit
Euroopa Sotsiaalfond
Eesti tuleviku heaks

Kuidas masin õpib: Matemaatika

Kahe muutujaga valem on:

$$y = m*x + n*z$$

x ja z on mudeli muutujad

Näiteks:

Müügitulu = m*Reklaamikulu + n*Toote hind

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Õpime õppima: müügitulu prognoosimudel

Reklaamikulu	Toote hind	Müügitulu
20	5	**5**
30	4	**18**
20	3	**11**

Meil on ajaloolised andmed müügitulu, reklaamikulu ja toote hinna kohta.

Eesmärgiks on luua mudel, mis prognoosiks müügitulu.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Õpime õppima: müügitulu prognoosimudel

Meie mudel:

Müügitulu = m*Reklaam + n*Hind

Meie mudelis on kaks muutujat (hind ja reklaamikulu).

Reklaamikulu	Toote hind	Müügitulu
20	5	**5**
30	4	**18**

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Õpime õppima: Alge mudel

Mudeli treeningu alguses antakse parameetritele juhuslikud väärtused näiteks reklaamikulu parameeter = 2 ja hinna parameeter =-2.

Müügitulu = 2*Reklaam -2*Hind

Reklaamikulu	Toote hind	Müügitulu
20	5	**5**
30	4	**18**

Prognoos	
30	**= 2*20 - 2*5**
52	**= 2*30 - 2*4**

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Õpime õppima: Algne mudeli viga

Meie mudel prognoosib müügitulu liiga optimistlikult.

Reklaamikulu	Toote hind	Müügitulu	Prognoos	Viga
20	5	**5**	**30**	**30-5 = 25**
30	4	**18**	**52**	**52-18 = 34**

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Õppimise reegel

1. **Kui viga on 0**, siis on mudel täiuslik;

2. **Kui viga on <0**, siis suurenda positiivse väärtusega muutujate parameetreid ja vähenda negatiivse väärtusega muutujate parameetreid;

3. **Kui viga on >0**, siis vähenda positiivse väärtusega muutujate parameetreid ja suurenda negatiivse väärtusega muutujate parameetreid.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Mudeli treeningprotsess

Kui viga on >0, siis vähenda positiivse väärtusega muutujate parameetreid ja suurenda negatiivse väärtusega muutujate parameetreid.

Müügitulu = 2*Reklaam -2*Hind

Vähendame selle parameetri 2 pealt 1 peale.

Suurendame selle parameetri 2 pealt 3 peale.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Õpime õppima: Mudel pärast treenimist

Õppimise reegli järgi vähendasime mõlemat parameetrit

Müügitulu = 1*Reklaam -3*Hind

Reklaamikulu	Toote hind	Müügitulu
20	5	**5**
30	4	**18**

Meie uus mudel prognoosib täpselt!

Prognoos	
5	**= 1*20 - 3*5**
18	**= 1*30 - 3*4**

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Testime mudelit valideerimisandmestiku peal

Müügitulu = 1*Reklaam -3*Hind

Reklaamikulu	Toote hind	Müügitulu
20	3	**11**

Prognoos	
11	**= 1*20 - 3*3**

Selle näite jätsime treeningandmetest välja.

Ja ka siin prognoosib mudel õigesti!

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Aga testandmed!?

Müügitulu = 1*Reklaam -3*Hind

Reklaamikulu	Toote hind	Müügitulu
20	6	**4**
30	3	**24**

Need on testandmed, mida mudel treeningprotsessi käigus näha ei tohtinud – isegi valideerimiseks mitte!

Prognoos	
2	**= 1*20 - 3*6**
21	**= 1*30 – 3*3**

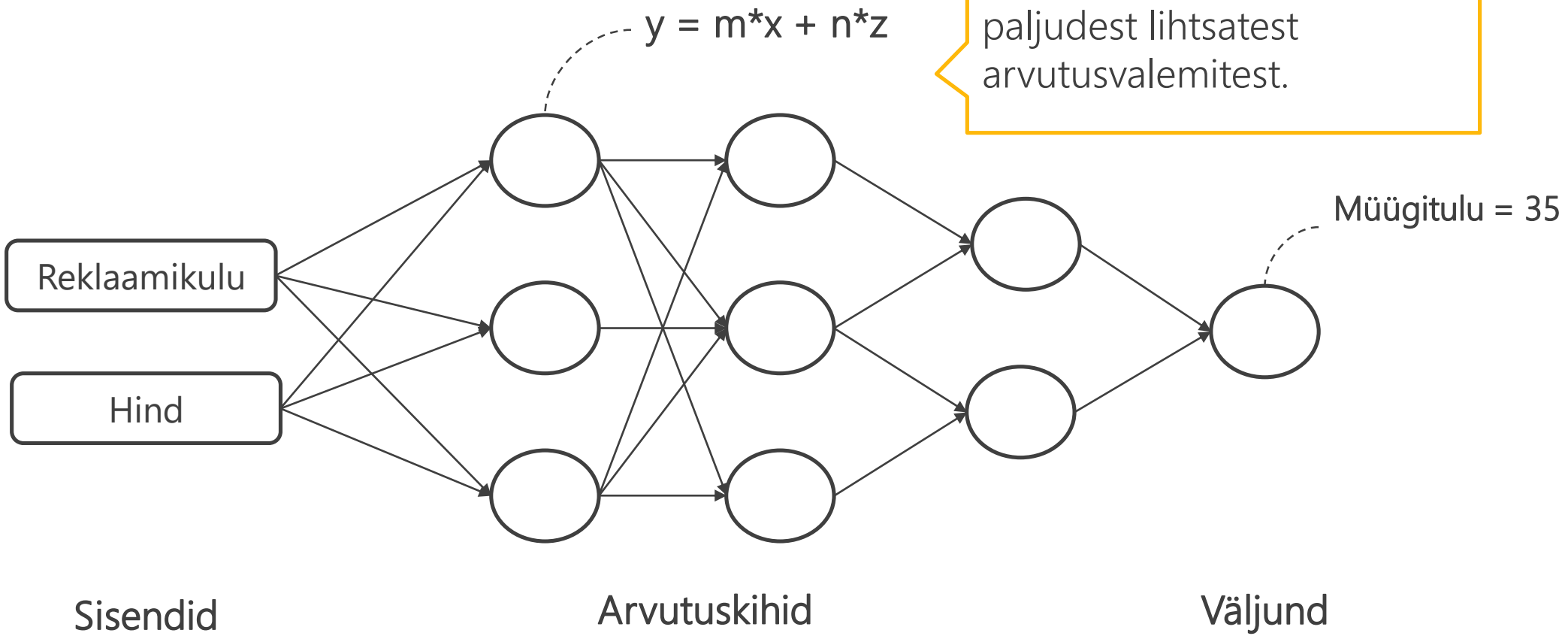
Viga
2-4 = -2
21-24 = -3

Ja alles nüüd saame öelda, et meie mudeli keskmine viga on 2.5€

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Aga tehislik närvivõrk (*artificial neural network*) ja süvaõpe (*deep learning*)!?

Tehislik närvivõrk koosneb paljudest lihtsatest arvutusvalemitest.



Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Tehisintellekt õpib, proovides leida parameetrite väärtusi, millega kaasneb võimalikult väike viga.

Keerukad tehisintellekti mudelid (tehislikud närvivõrgud) koosnevad paljudest lihtsatest arvutusvalemitest.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

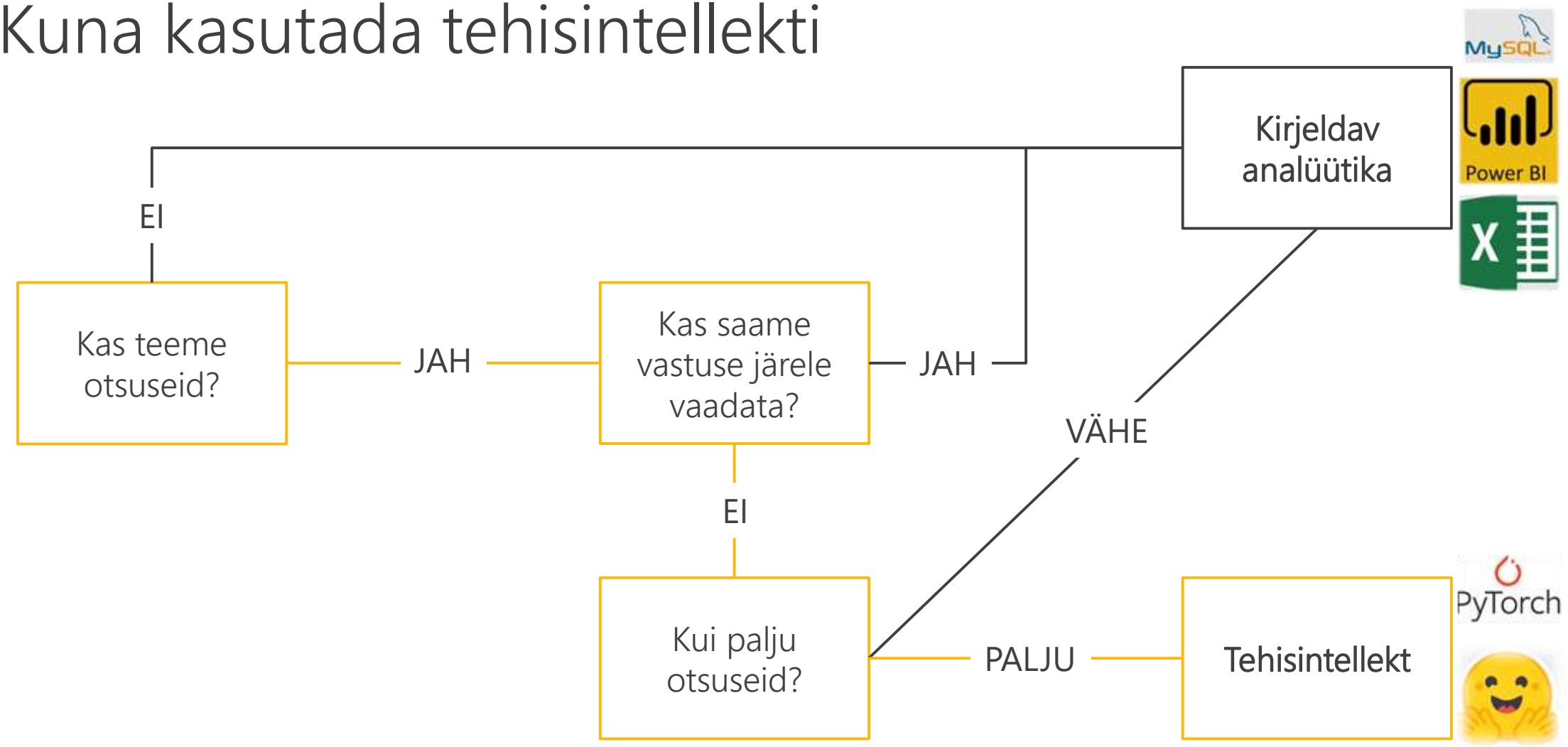
Kuna kasutada kirjeldavat analüütikat, statistikat või masinõpet?

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Kuna kasutada tehisintellekti!?

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Kuna kasutada tehisintellekti



Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Tehisintellekti on mõistlik kasutada, kui meil on vaja teha palju otsuseid ja neid ei saa andmebaasist järele vaadata.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Tehisintellekti olemuslikud riskid?

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Probleem 1: Isikuandmete seaduse potentsiaalne rikkumine

Allikas: VectorStock.com

- **Isikuandmed** – igasugune teave tuvastatud või tuvastatava füüsilise isiku („andmesubjekti“) kohta. Tuvastatav füüsiline isik on isik, **keda saab otseselt või kaudselt tuvastada**.
- Isikuandmete töötlemine on seaduslik, kui isik on andnud nõusoleku töödelda tema andmeid konkreetsel eesmärgil.
- **Nõusolek** – vabatahtlik, konkreetne, teadlik ja ühemõtteline tahteavaldus (toimib *opt-in* mitte *opt-out* põhimõttel!).

NB: kas treenitud masinõppemudel sisaldab isikuandmeid!?

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Probleem 2: “Katkised” andmed

Kuupäev	Kiirus	Lisad	Teostaja	Asend
23.01.2017	6	EI	Mari	Püsti
24.01.2017	10	EI	Mari	Püsti
20.03.2017	12	EI	Mari	Püsti
05.02.2017	25	JAH	Tõnu	Püsti

Kuupäev	Kiirus	Lisad	Teostaja	Asend
23.01.2017		EI	Mari	Püsti
24.01.2017	10		Mari	Püsti
28.01.2017	18	JAH	Mari	Istuv
20.03.2017	12	EI	Mari	
05.02.2017	25	JAH	Tõnu	Püsti

Stsenaarium: Mingil põhjusel salvestati süsteemi “katkised” andmed.

Tulem: Tehisintellekt treenib end öösel rumalaks!

Lahendus:

- Automaatne anomaaliate tuvastamise süsteem andmetele;
- *Sanity check;*
- *Canary deployment;*

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Probleem 3: Tehisintellekt on külmalt matemaatiline

Sulle võib veel meeldida:

Stsenaarium: Inimene ostab poest ainult alkoholi ja tehisintellekti eesmärk on soovitada inimesele "talle sobivaid tooteid".

Tulem: tehisintellekt soovitab inimesele ainult alkoholi -> inimesele ei meeldi, et pood teda alkohoolikuks peab.

Lahendus:

- Lisada tehisintellektile ärireeglite kiht.
- *Sarnane olukord ka rassi ja soo küsimustes.*

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Probleem 4: Inimeste pahatahtlikkus

Allikas: Twitter

Stsenaarium: 2016. aasta 23. märtsil avas Microsoft Twitteri konto vestlusrobotile Tay; Tay suutis õppida nendelt, kellega ta suhtleb;

Tulem: 16 tunni pärast oli Tay rassistlik, vägivalda õhutav tegelane.

Lahendus:

- Eetika ja moraalireeglite kontroll tehisintellekti tulemile.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Probleem 5 (level 2): Inimeste pahatahtlikkus

Allikas: Jonathan Lampel (unsplash.com)

Stsenaarium: Droonile või isesõitvale autole sisestatakse masinõppemudel, mis on treenitud tapma inimesi (N: piltide/video põhjal).

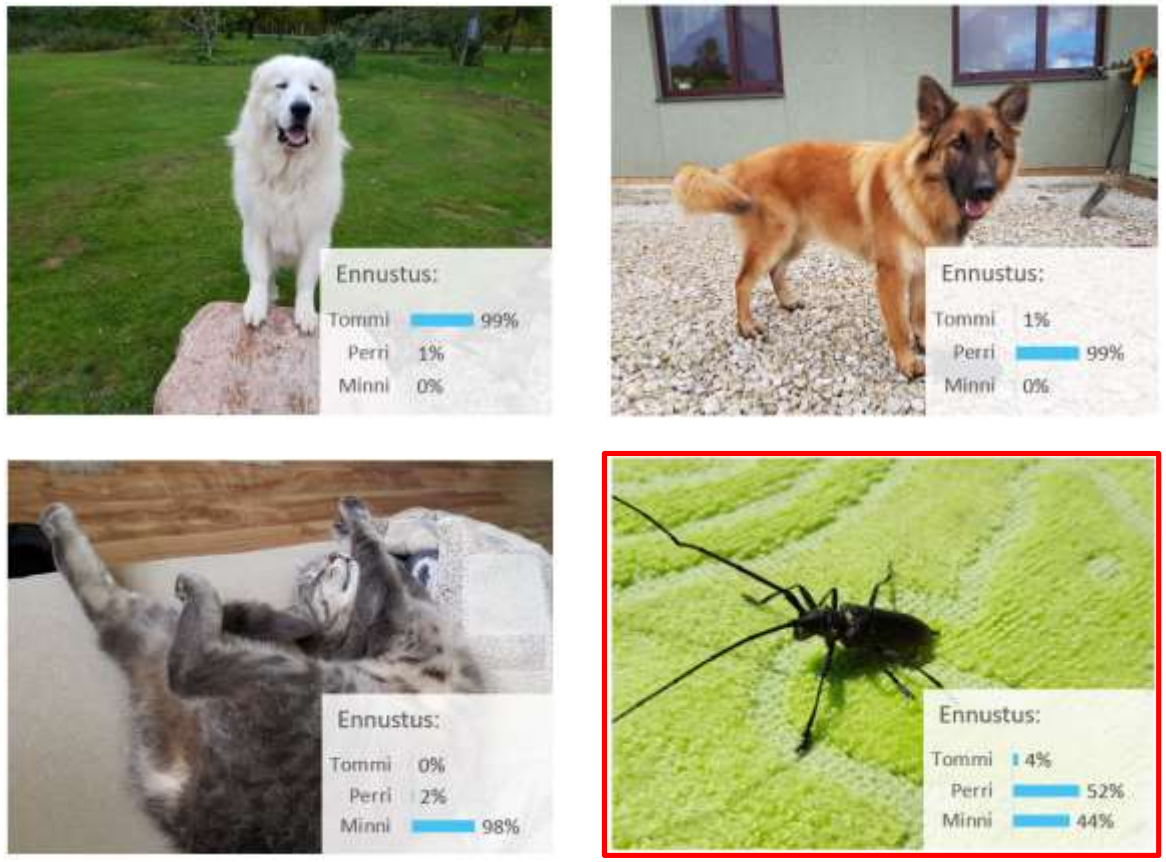
Tulem: ?

Lahendus:

- Aktiivne ja jõuline vastumeetmete arendamine;

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Probleem 6: Liigne usaldus tehisintellekti suhtes



Allikas: Autori koostatud

Stsenaarium: Tehisintellekt annab väga täpseid vastuseid olukordades, mille põhjal ta on treenitud ja hakkame süsteemi liigselt usaldama *(overconfidence)!*

Tulem: Uudse andmekillu nägemisel annab süsteem täiesti vale vastuse

Lahendus:

- Võõraste olukordade tutvustamine treenimisel;
- Säilitada pragmaatiline skeptilisus AI tulemustesse.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Probleem 7: võltssisu ja valeuudised

Allikas: Karras, Aila, Laine, Lehtinen 2017

Stsenaarium: *Generative Adversary Networks (GANs)* suudavad edukalt luua valepilte ja -videosid*!*

Tulem: Saab olema äärmiselt keeruline eristada, kas videol nähtav sündmus on ka tegelikult aset leidnud.

Lahendus:

- Tehnoloogiad valepiltide tuvastamiseks.
- Topeltkontroll õigsuse üle!?

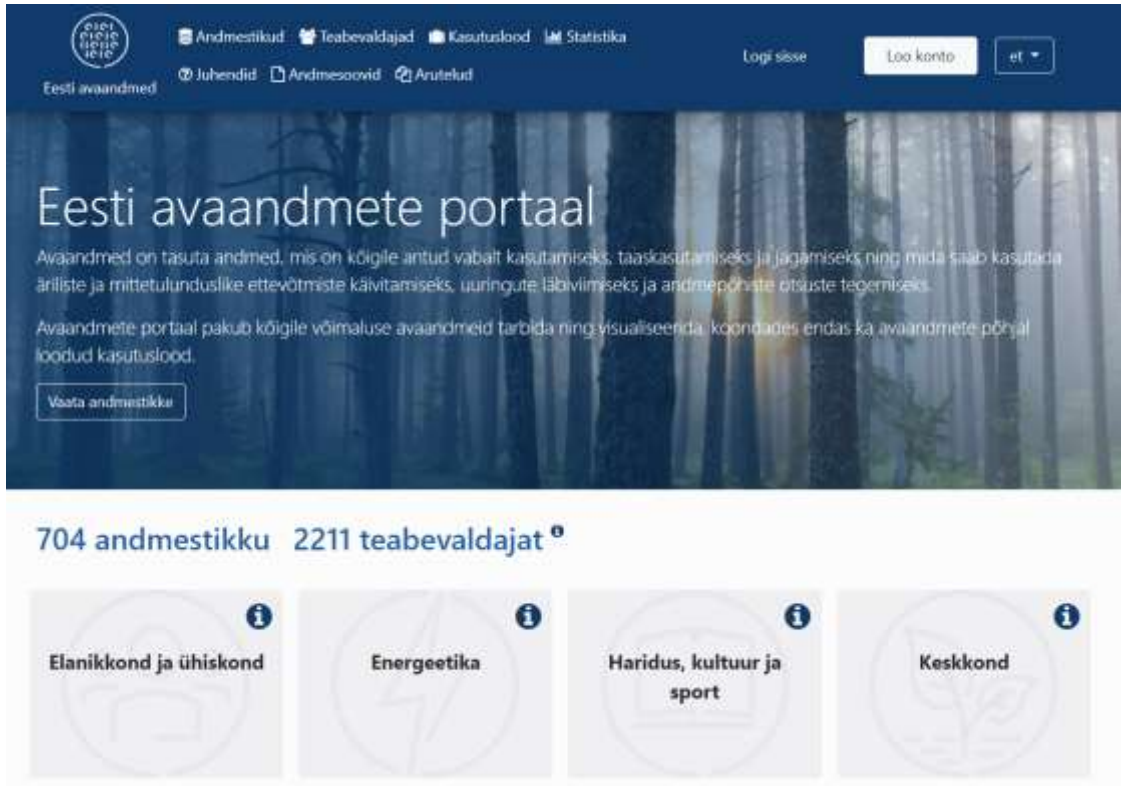
Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Ülesanne – tuletame meelde, mis on mis!

Eesti avaandmete portaali tutvustus?

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Eesti avaandmete portaal



Allikas: avaandmed.eesti.ee

- Avaandmed on tasuta andmed, mis on kõigile antud vabalt kasutamiseks.
- Avaandmete portaal pakub kõigile võimaluse avaandmeid tarbida ning visualiseerida, koondades endas ka avaandmete põhjal loodud kasutuslood.

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Avaandmete portaal – vaatame ringi

Allikas: avaandmed.eesti.ee

- Andmestikke saab otsida nii valdkondade, aasta, piirkonna kui ka otsingusõnade järgi.
- Andmete lisamise ja allalaadimiste kohta saab näha statistikat.
- Avaldatud on kasutuslood ja näidisrakendused.
- Juhendid (N: API kasutamise kohta)

Euroopa Liit
Euroopa Sotsiaalfond
Eesti
tuleviku heaks

Koolitus „Andmeteadus on Popp“

Andmeteadus, Masinõpe ja tehisintellekt

04. november 2021

Kristjan Eljand

Euroopa Liit
Euroopa Sotsiaalfond

Eesti
tuleviku heaks