

Andmetöötluse olulised aspektid

Dage Särg, andmeteadlane
STACC OÜ

Teemad

- Andmete märgendamine
- Andmete eelarvamuslikkus
- Aegandmed
- Andmete perioodilisus

Andmete märgendamine

STACC

Andmete märgendamine

- Toorandmete (pildid, tekst, video, ...) tõlgendamine ja neile ühe või enama tähendusliku sildi lisamine
- Võimaldab masinõppemudelil andmetest järeldusi tegema õppida

Masinõppe liigid

Juhendatud õppimine (*supervised learning*)

- Andmestikus on kirjas oodatav väljund
 - Õpime ennustama: nt rämpsposti filtreerimine, tekstiliikide tuvastus, keeletuvastus, autorituvastus, meelsusanalüüs

Juhendamata õppimine (*unsupervised learning*)

- Andmestikus pole kirjas oodatavat väljundit
 - Otsime andmestest struktuuri

Tekstide liigitamine

- Sarnaste dokumentide tuvastamine
- Klassifitseerimine
 - Eeldefineeritud liigid, nt žanride, autorite vms kaupa
- Klasterdamine
 - Tüüpiliselt pole teada võimalike gruppide arv või kuidas neid nimetada



Tekstide liigitamine

- Sarnaste dokumentide tuvastamine
- Klassifitseerimine
 - Eeldefineeritud liigid, nt žanride, autorite vms kaupa
- Klasterdamine
 - Tüüpiliselt pole teada võimalike gruppide arv või kuidas neid nimetada



Andmete märgendamine - tekst

3	positiivne	Isiklikult kohtasin natukegi Kukekese moodi politseinikku viimati kaheksa aasta eest Lätis. Eranditult kõik viimase kümnendi kokkupuuted politseiametnikega on kinnitanud: vaatamata raskustele on Eesti riik suutnud korrakaitsjateks värvata inimesi, kes on arukad, kohusetundlikud, lugupidamist sisendavas füüsilises vormis ja hea väljendusoskusega.
4	vastuoluline	Olen näinud ka, kuidas patrull korrarikkujat taltsutab, ning suur osa sellest seisnes enesekindlas olekus ning vastuvaidlemist välistavalt, kuid rahulikult antud korraldustes. Aprillirahutuste ajal veendusime, et Eesti politsei suudab käituda ründajatega väga karmilt. Vaevalt et mitmeks tunniks näoli porisele asfaldile pandud märatsejat lohutas, et kohus hiljem teda riigijuhtide lubatud viisil aastateks vangi ei pannud.
5	negatiivne	Kummaline on nüüd äkki lugeda politsei ja siseministeeriumi ametnike kurtmist kohtu poolt politseinike ründajatele (lihtsalt öeldes – peksjatele) määratud karistuste üle. Tsiteerin ERRile intervjuu andnud politseinikku: «See ei ole normaalne, et sellised jõmikad tulevad, tümitavad ja nii ongi noh. Mingi tagajärg peab saabuma neile, [kes politseiametniku vastu kätt tõstavad].»
6	negatiivne	Küsimus pole mitte leebetes karistustes, vaid selles, et purjus ja relvadeta kraaklejad läksid täisvarustuses patrullpolitseinikele kallale ja said neist hõlpsalt jagu. Politseiametnikule on riik andnud õiguse vajadusel vägivalda kasutada. Praegu kuuleme aga hoopis politseinike kaastunnet äratavaid kirjeldusi, kuidas neid rusikate ja jalgadega peksti, kuidas kumminui ära võeti, kuidas kägistati ja taga aeti.

Andmete märgendamine - tekst

Selveri	B-ORG	
esindaja	O	
Mari	B-PER	
Maasikas	I-PER	
kommentaare	O	
ei		O
jaganud	O	
.		O

Selveri esindaja Mari Maasikas kommentaare ei jaganud.

Andmete märgendamine - tekst

"<s>"

"<Tean>"

"tead" Ln V main indic pres ps1 sg ps af @FMV #1->0

"<omas>"

"oma" Ls A pos sg el @AN> #2->3

"<käest>"

"käsi" Lst S com sg el @ADV L #3->1

"<,>"

"," Z Com CLB #4->4

"<tuttav>"

"tuttav" L0 A pos sg nom @AN> #5->6

"<tunne>"

"tunne" L0 S com sg nom @SUBJ #6->1

"<. >"

"," Z Fst #7->7

"</s>"

Andmete märgendamine - tekst

Süntaks: lauseliikmed

<i>Väike</i>	<i>laps</i>	<i>sõi</i>	<i>matsutades</i>	
<i>kala</i>				
@AN>	@SUBJ	@FMV	@ADVL	@OBJ
täiend	alus	öeldis	määrus	
sihitis				

Andmete märgendamine - tekst

Süntaks: lauseliikmed

Väike *laps* *sõi* *matsutades*
kala

@AN> @SUBJ @FMV @ADVL @OBJ

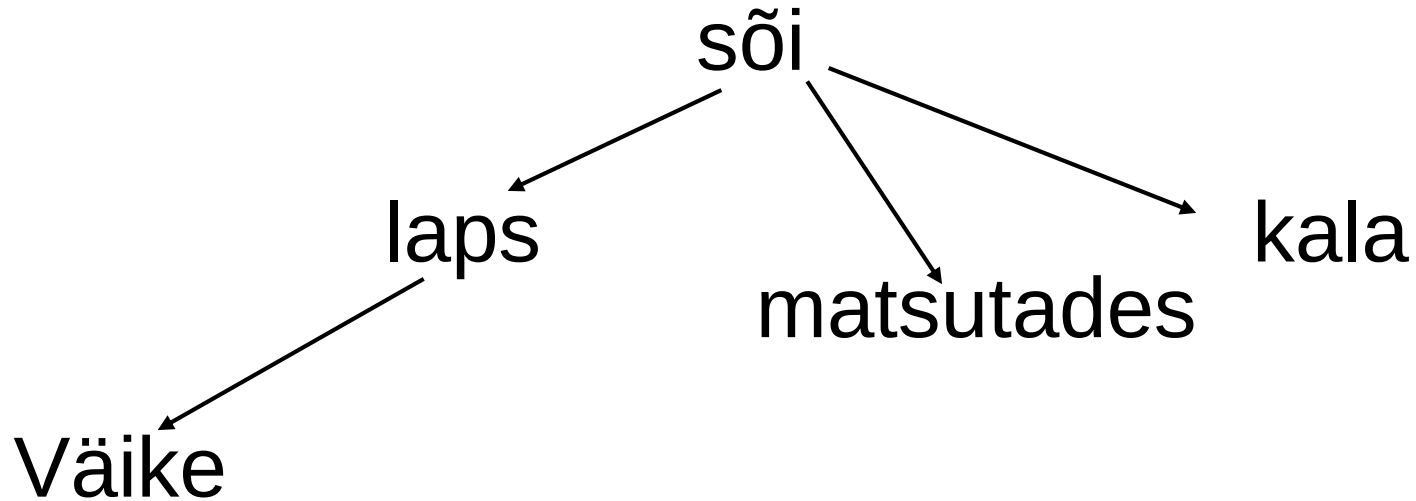


öeldis

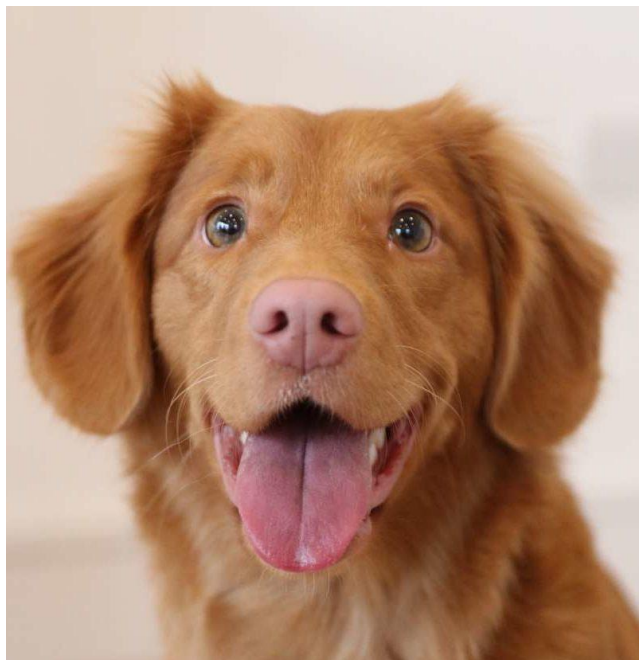


Andmete märgendamine - tekst

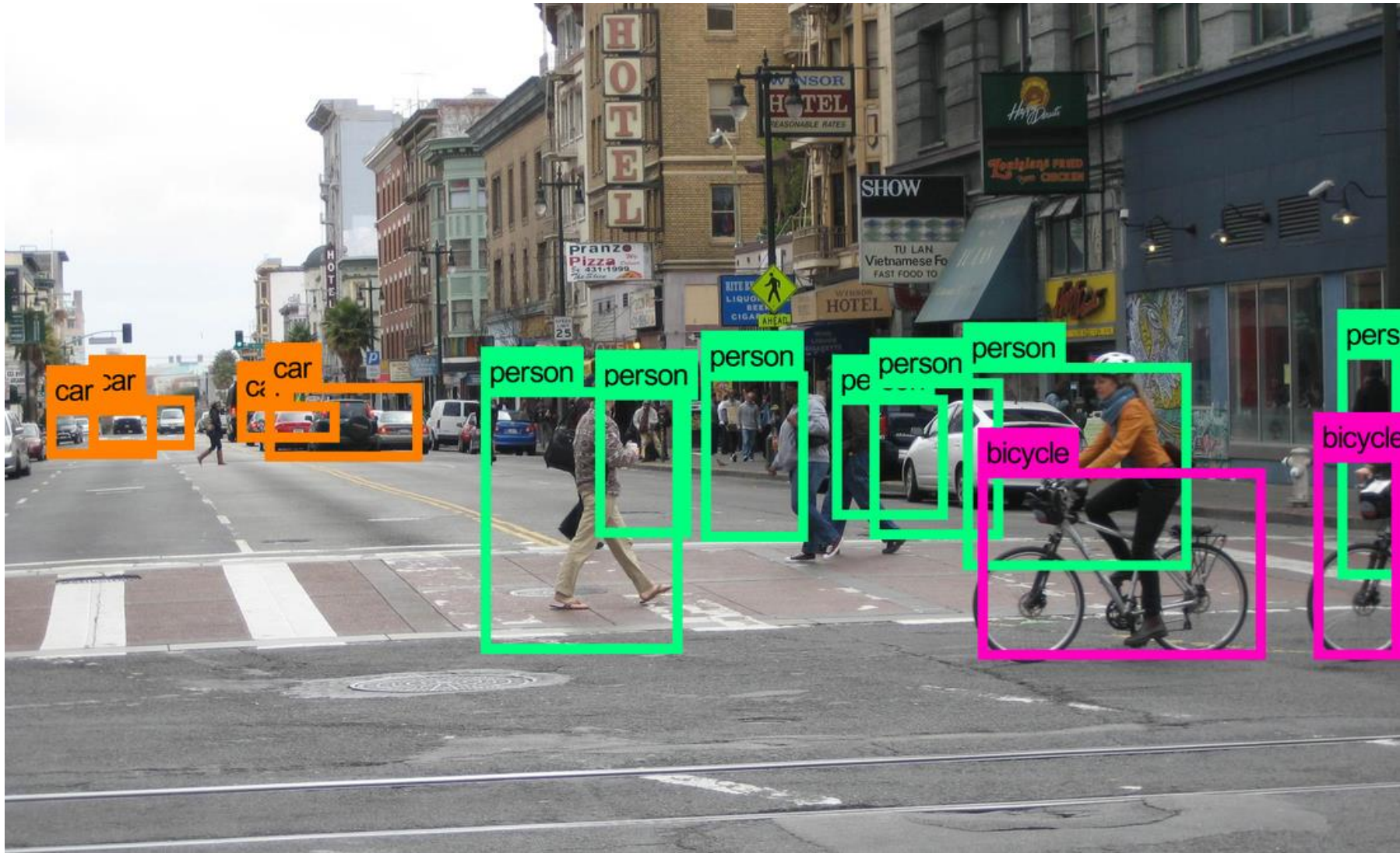
Süntaks: sõnadevahelised seosed lauses



Andmete märgendamine - pildid



Andmete märgendamine - pildid



Andmete märgendamine - olulised aspektid

- Mida märgendada?
- Kuidas märgendada?
- Mis toorandmeid kasutada?
- Kes märgendab?
- Kui palju märgendada?

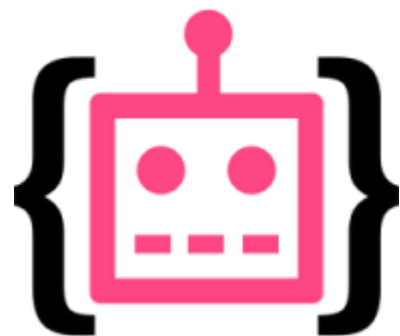
Märgendusformaate

```
persons.csv - Notepad
File Edit Format View Help
Family Name,Given Name,VIAF ID
Ackersdijck,Willem Cornelis,17959345
Adelung,Friedrich von,22963658
Afzelius,Arvid August,49972119
Amerling,Karel,13331054
Anton,Karl Gottlob von,183632821
Arwidsson,Adolf Ivar,8184878
Asbjørnsen,Peter Christen,116587918
Attems,Heinrich,37665468
Atterbom,Per Daniel Amadeus,46819248
Balabin,Viktor Petrovich,44473845
Banks,Joseph,46830189
Beck,Friedrich,44338671
Becker,Reinhold von,42101066
Bernhart,Johann Baptist,69674335
Bertram,Johann,32890043
Bilderdijk,Willem,14882166
Boisserée,Sulpiz,7483155
Bopp,Franz,61614118
Borovský,Karel Havlíček,100277614
Bosković,Jovan,161354270
Buslaev,Fyodor,10074560
Cenowa,Florian Stanislaw,44466031
Chomiakov,Aleksei,66492873
```

```
<?xml version="1.0"?>
- <birds>
  - <owl id="1201">
    <species>Bubo bubo</species>
    <name>Eagle Owl</name>
    <region>Eurasia</region>
  </owl>
  - <owl id="1202">
    <species>Strix occidentalis</species>
    <name>Spotted Owl</name>
    <region>North America</region>
  </owl>
</birds>
```

```
"employees":[
  {"firstName":"John", "lastName":"Doe"},
  {"firstName":"Anna", "lastName":"Smith"},
  {"firstName":"Peter", "lastName":"Jones"}
]
```

Märgendustarkvarasid



Label Studio

Andmete märgendamine - kuidas?

Märgendid (sildid) peavad olema:

- informatiivsed
- eristavad / eristatavad
- sõltumatud
- tõesed

Ground Truth / maapealne tõde

- Masinõppealgoritmi tulemuse võrdlemine sellega, kuidas on päriselt
- Mõiste pärit meteoroloogiast - kohapeal tehtud vaatluste teel saadud info
- Eeldus: märgendajate märgendus on tõene (?)

Andmete märgendamine - mõõdikud

- Täpsus - kuivõrd ühtib märgendajate märgendus tõega
- Kvaliteet - kuivõrd ühtselt / järjekindlalt on andmed märgendatud

Olulisi aspekte märgendamisel

- Domeeni ja konteksti tundmine
- Agiilsus
- Töömahu mõistmine (nt 1h videot - 800 h märgendamist)
- Tasustamine - tunni või mahu põhiselt
- Andmete turvalisus

Rohkem märgendamise kohta

<https://www.kdnuggets.com/2021/12/data-labeling-ml-overview-and-tools.html>



Aegandmed

Aegandmed

- Aegandmed on ajalisse järjekorda pandud andmepunktide kogum
- Tüüpiliselt on andmepunktideks mõõtmised samast allikast mingi ajavahemiku järel
- Kasutatakse muutuste jälgimiseks ajas, mustrite tuvastamiseks ning tuleviku ennustamiseks

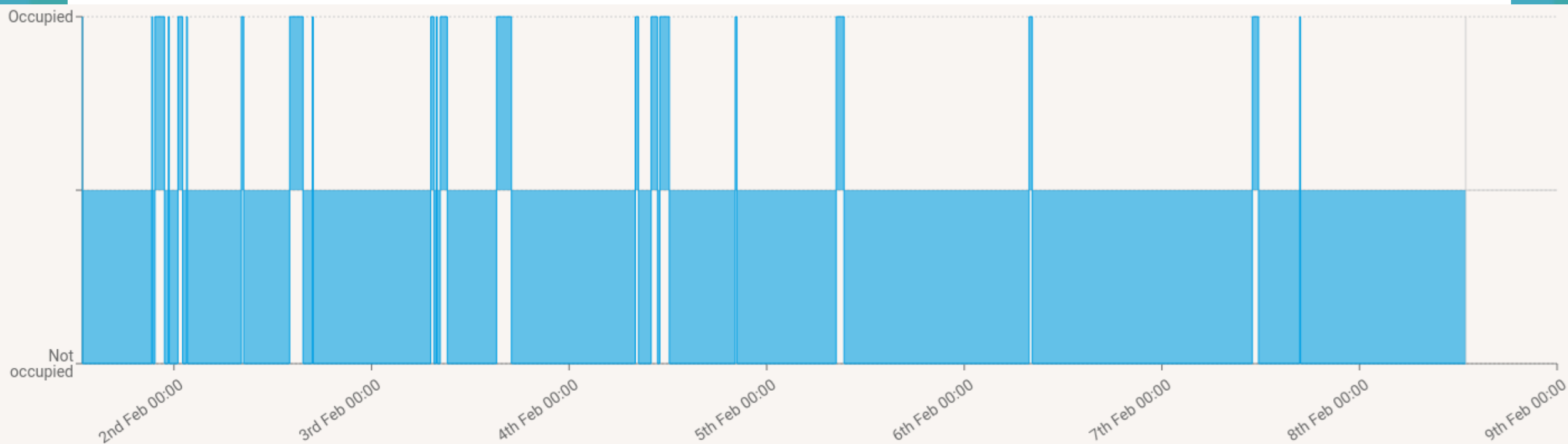
Aegandmed

- Aegandmed on ajalisse järjekorda pandud andmepunktide kogum
- Tüüpiliselt on andmepunktideks mõõtmised samast allikast mingi ajavahemiku järel
- Kasutatakse muutuste jälgimiseks ajas, mustrite tuvastamiseks ning tuleviku ennustamiseks



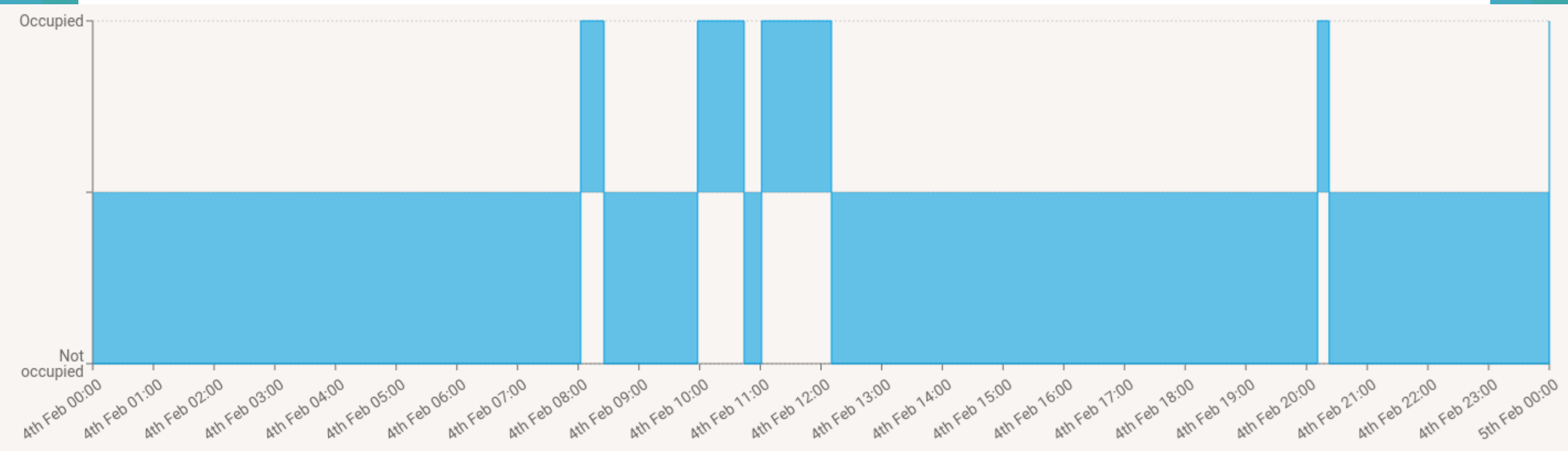
Aegandmed - näide

Kassi maja hõivatus



Aegandmed - näide

Kassi maja hõivatus



Aegridade analüüs - milleks?

- Aitab mõista protsesse ja sündmusi, mille käigus uuritavad andmed genereeritud on
- Sobitades andmetele mudeli, saab selle abil ennustada, monitoorida või tagasisidestada vastavaid protsesse ja sündmusi

Aegandmete järjestatus

- Järjestatuse tõttu unikaalne omadus - seriaalkorrelatsioon
- Seriaalkorrelatsioon näitab vaatlustevahelise korrelatsiooni sõltuvust mõõtmispunktide vahelisest kaugusest
 - $\Rightarrow$ andmepunkti väärtus ühel ajahetkel on statistilises sõltuvuses andmepunkti väärtusest mõnel teisel ajahetkel

Aegandmete tüübid

- Aegrida - kogum mõõtmisi ühe subjekti kohta erinevatel ajahetkedel. Nt maksimaalne temperatuur Tallinnas iga aasta 1. jaanuaril
- Mitmemõõtmelised andmed - kogum mõõtmisi mitme subjekti kohta ühel ajahetkel. Nt maksimaalne temperatuur, õhuniiskus ja tuule kiirus Tallinnas 1.01.2022
- Paneelandmed - mitmemõõtmelised aegread, kombinatsioon eelnevast kahest tüübist. Kogum mõõtmisi mitme subjekti kohta mitmel ajahetkel

Aegandmete analüüsimeetodeid

Aegridade analüüsil kasutatakse minevikus esinenud väärtuste infot ja neilt tuvastatud mustreid, et ennustada, mis juhtub tulevikus

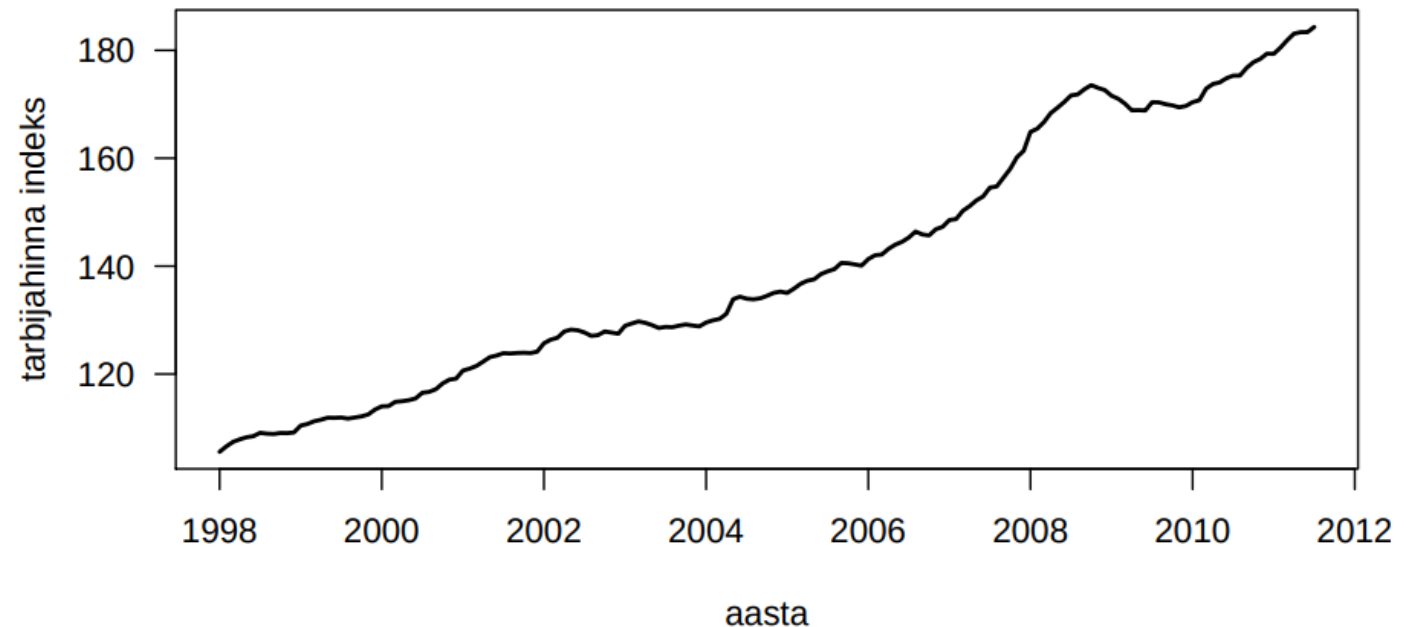
- Trendi tuvastamine
- Sesonsete mustrite analüüs

Trend

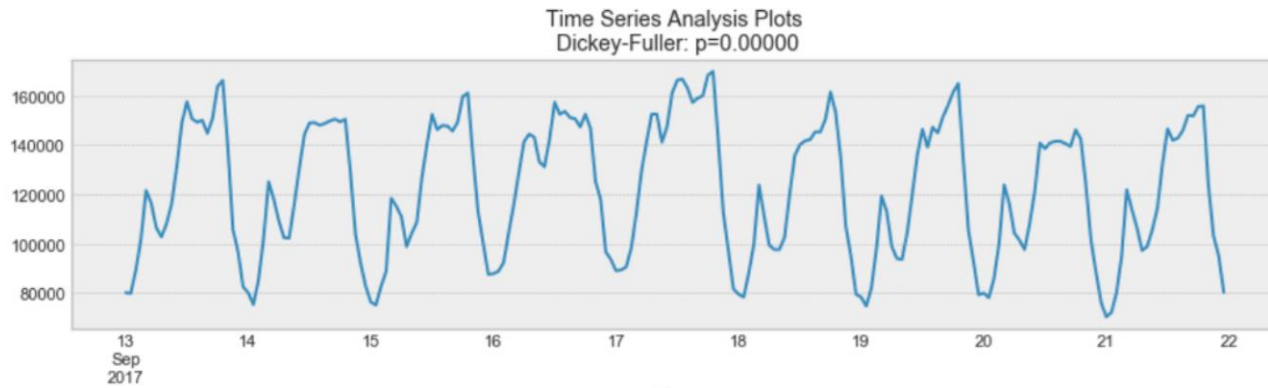
- Süstemaatiline muutus aegreas - pikaajaline suundumus
- Suund ja tõusunurk võivad olla püsivad või muutuda

Trend

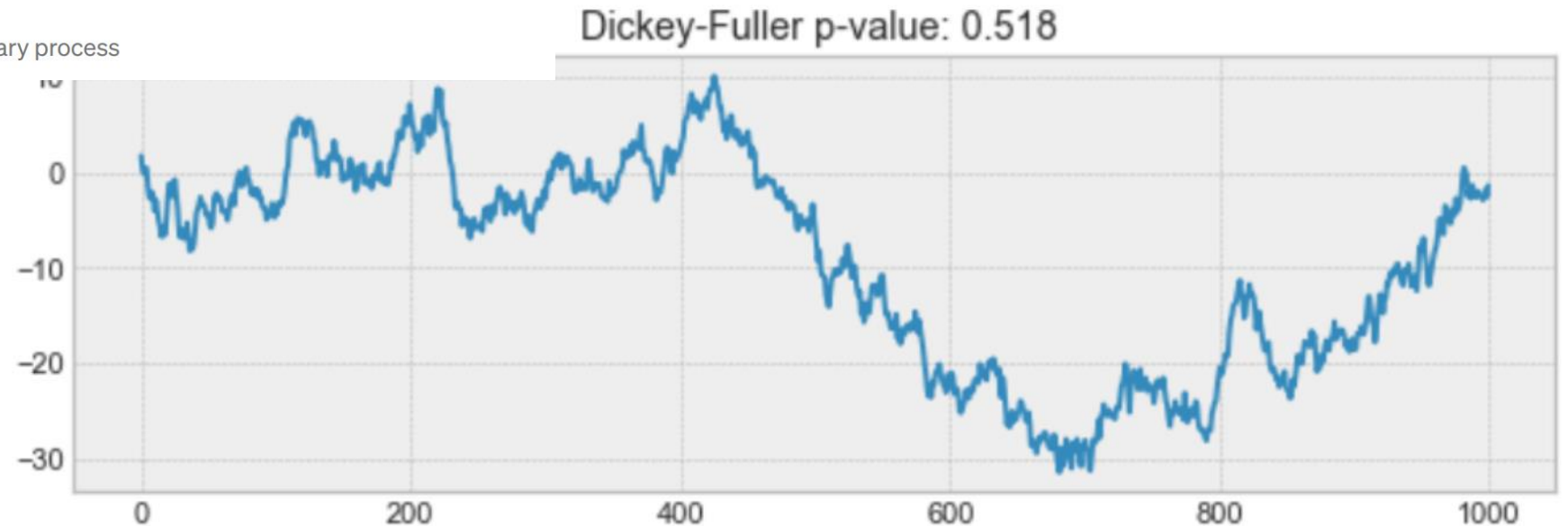
- Süstemaatiline muutus aegreas - pikaajaline suundumus
- Suund ja tõusunurk võivad olla püsivad või muutuda



Stationary vs



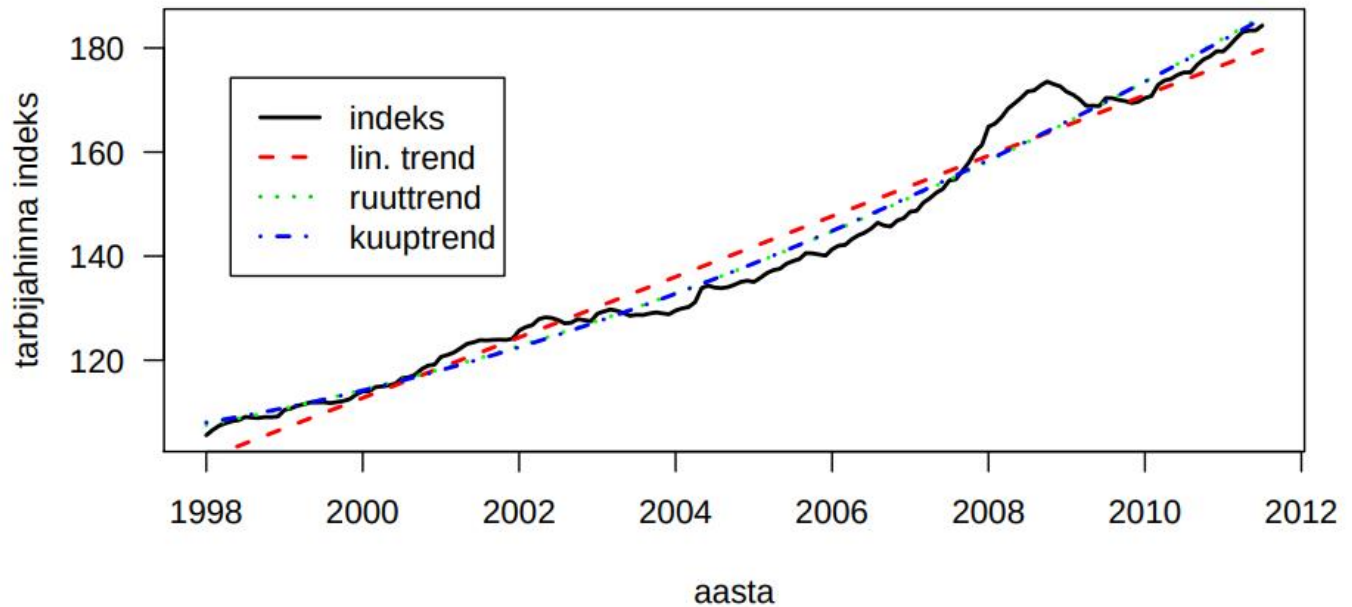
Example of a stationary process



Example of a non-stationary process

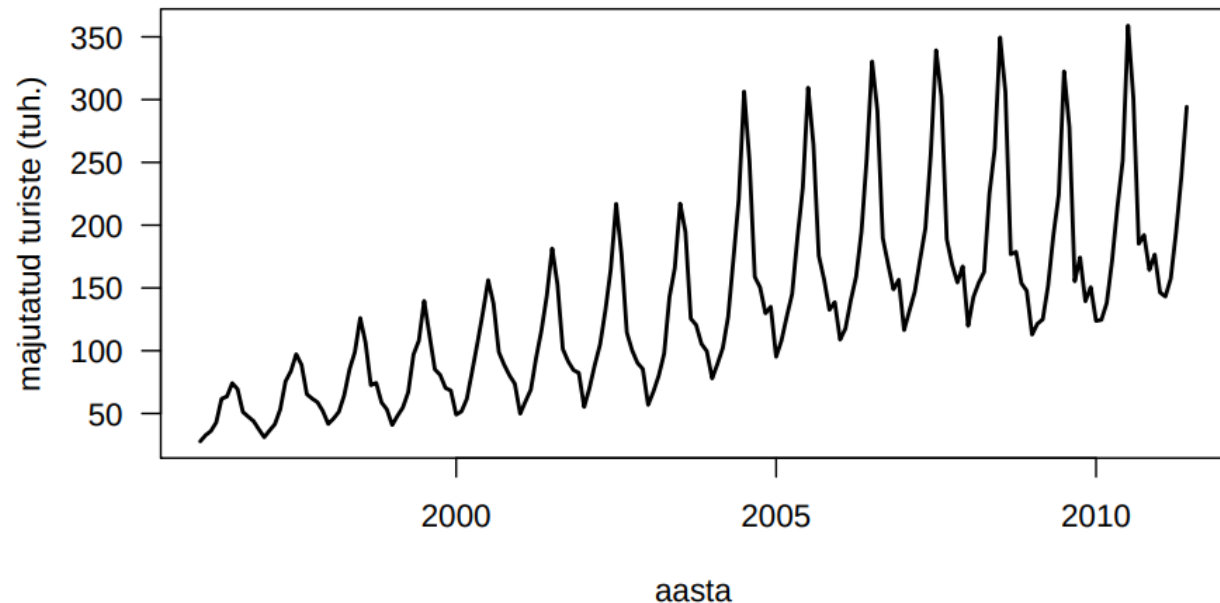
Trend

- Süstemaatiline muutus aegreas - pikaajaline suundumus
- Suund ja tõusunurk võivad olla püsivad või muutuda



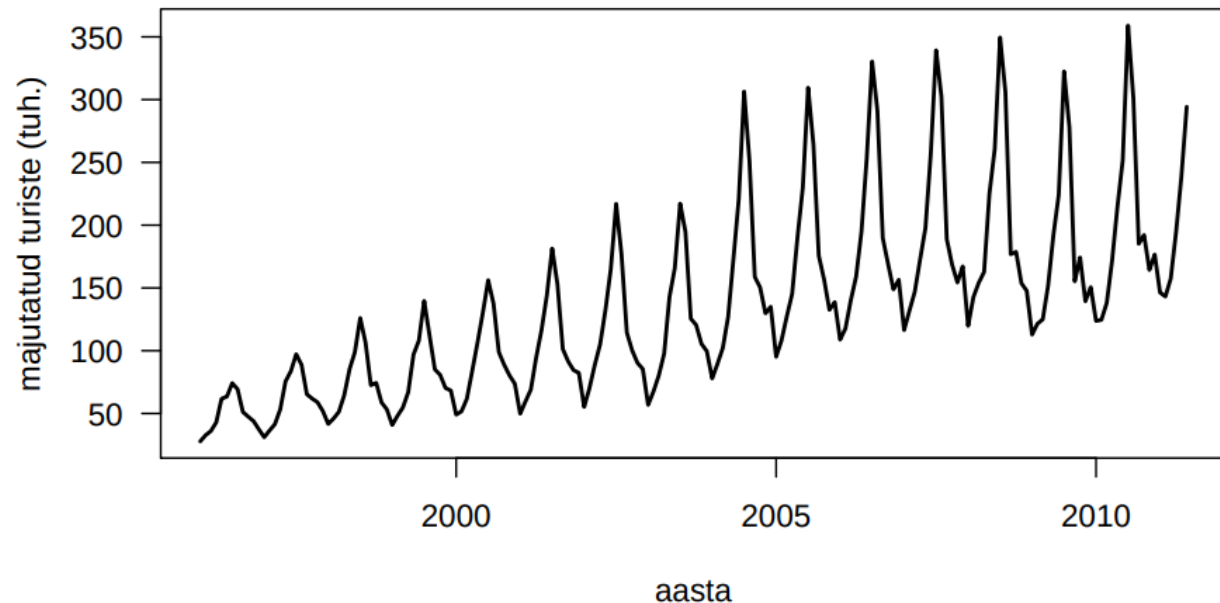
Trend

- Süstemaatiline muutus aegreas - pikaajaline suundumus
- Suund ja tõusunurk võivad olla püsivad või muutuda



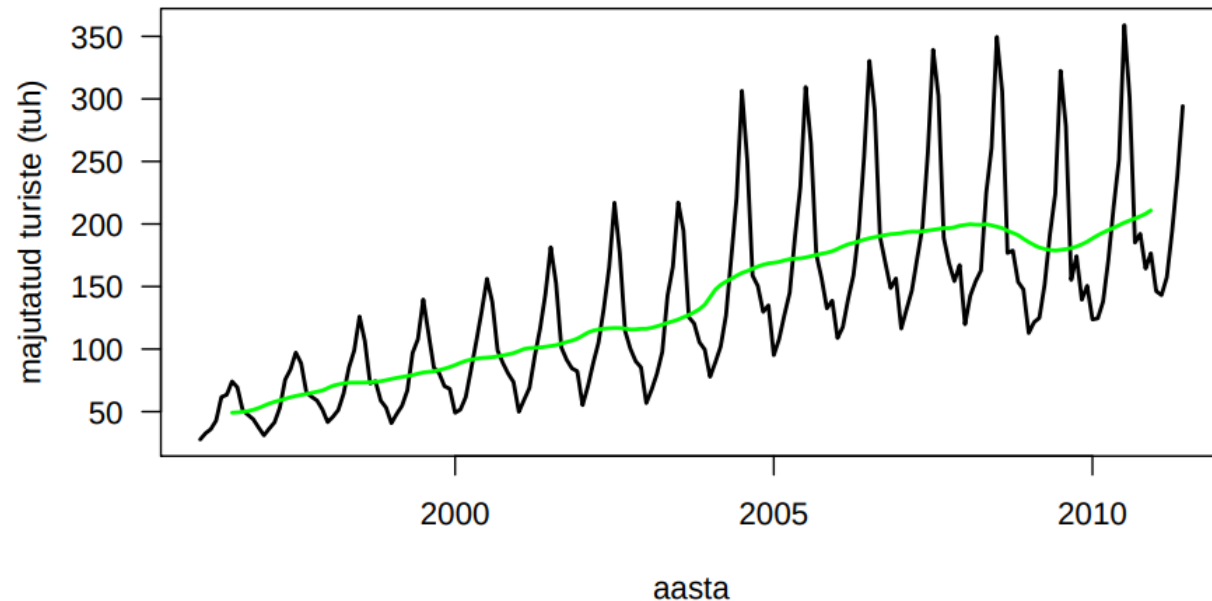
Perioodilisus

- Regulaarsed ja ennustatavad mustrid aegreas väiksema kui aastase intervalliga
- Mugav tuvastada graafikult



Perioodilisus

- Regulaarsed ja ennustatavad mustrid aegreas väiksema kui aastase intervalliga
- Mugav tuvastada graafikult



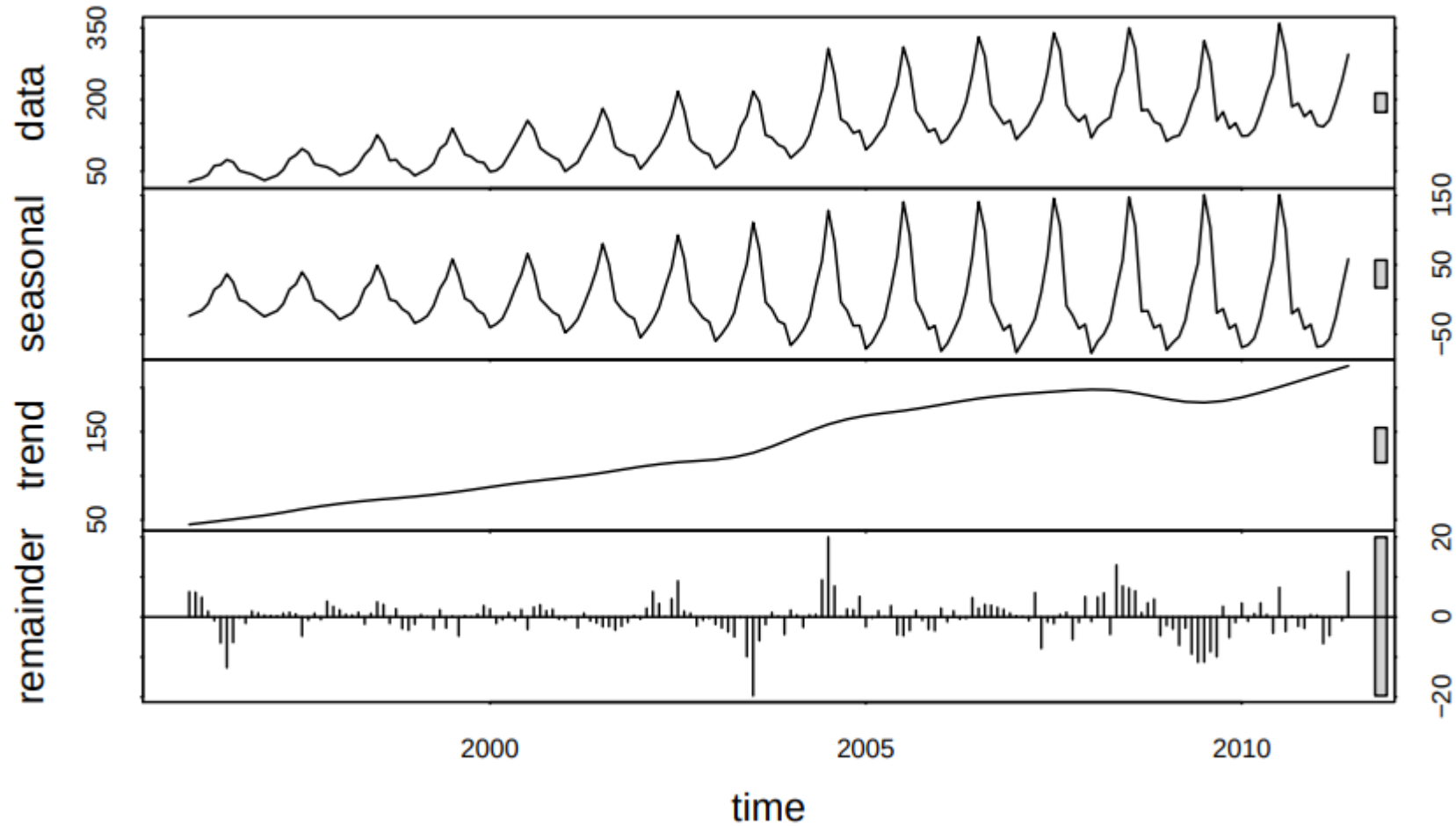
Perioodilisuse tuvastamine

- Tausta- ja kontekstiteadmus
- Graafikud
- Statistiline analüüs ja testid

Perioodilisuse arvestamine

- Mudelid, millele on perioodilisusega arvestamine juba sisse ehitatud, nt SARIMA perekonna mudelid
- Perioodilisuse eemaldamine andmete silumise teel. Ennustuste tegemiseks vaja taastada
- Kasutada juba silutud algandmeid, kui on saadaval (nt USA tööhõiveandmed)

Aegridade analüüs



Aegridade modelleerimine

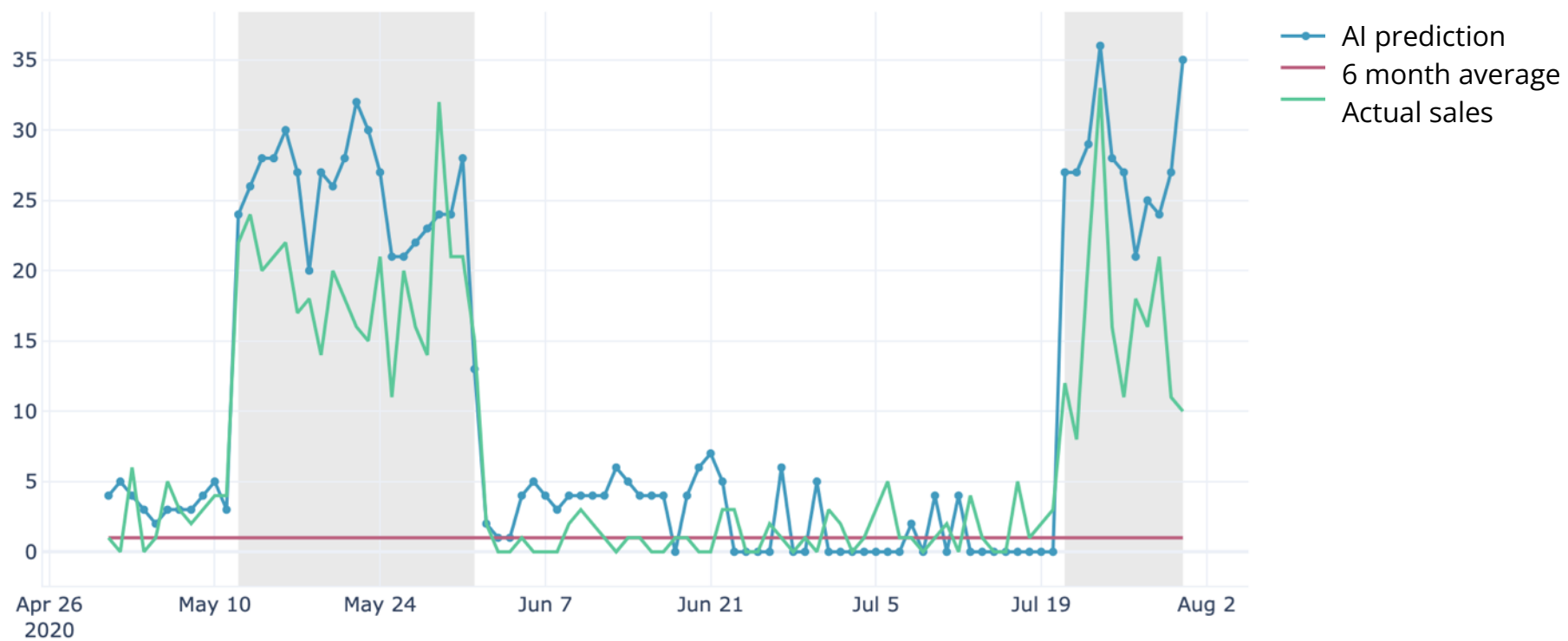
Enimkasutatavad mudelid

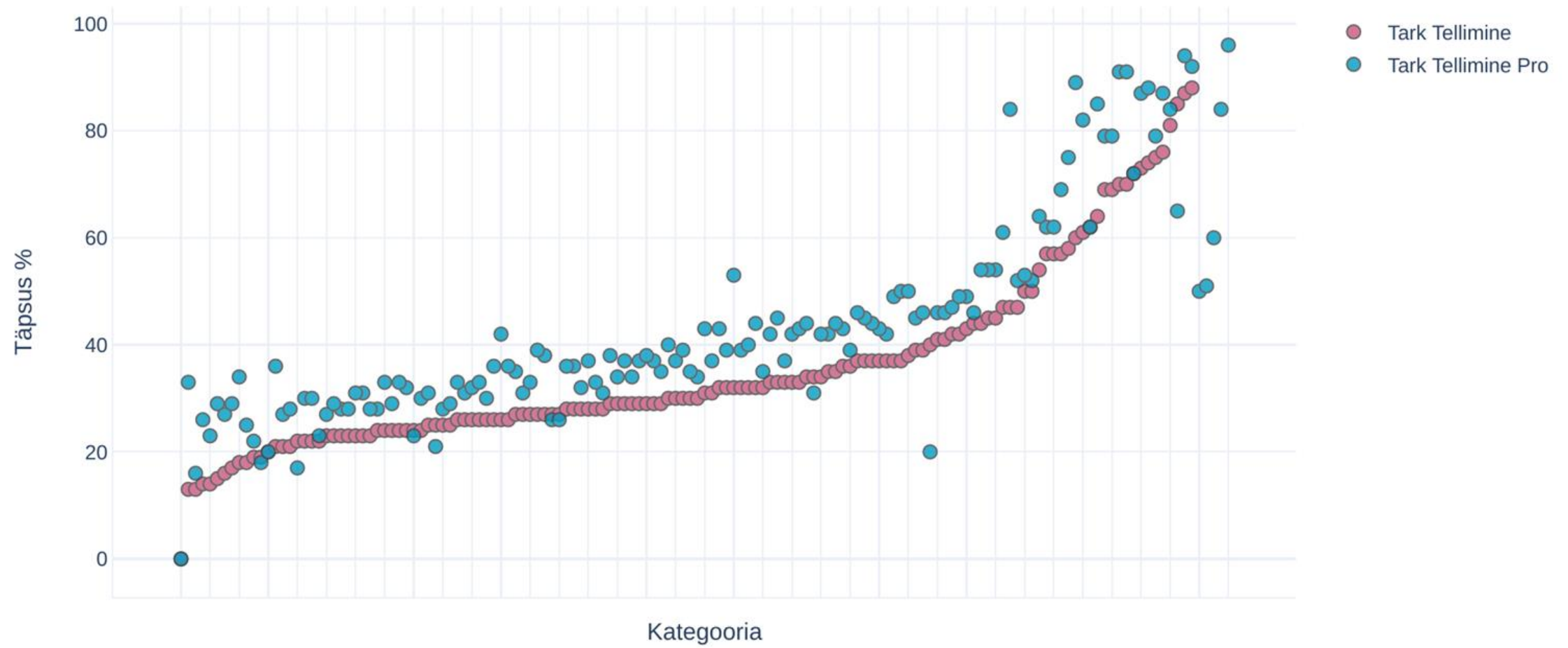
- libiseva keskmise põhised mudelid
- ARIMA->SARIMA->SARIMAX mudelid
- tehisnärvivõrkudel põhinevad mudelid

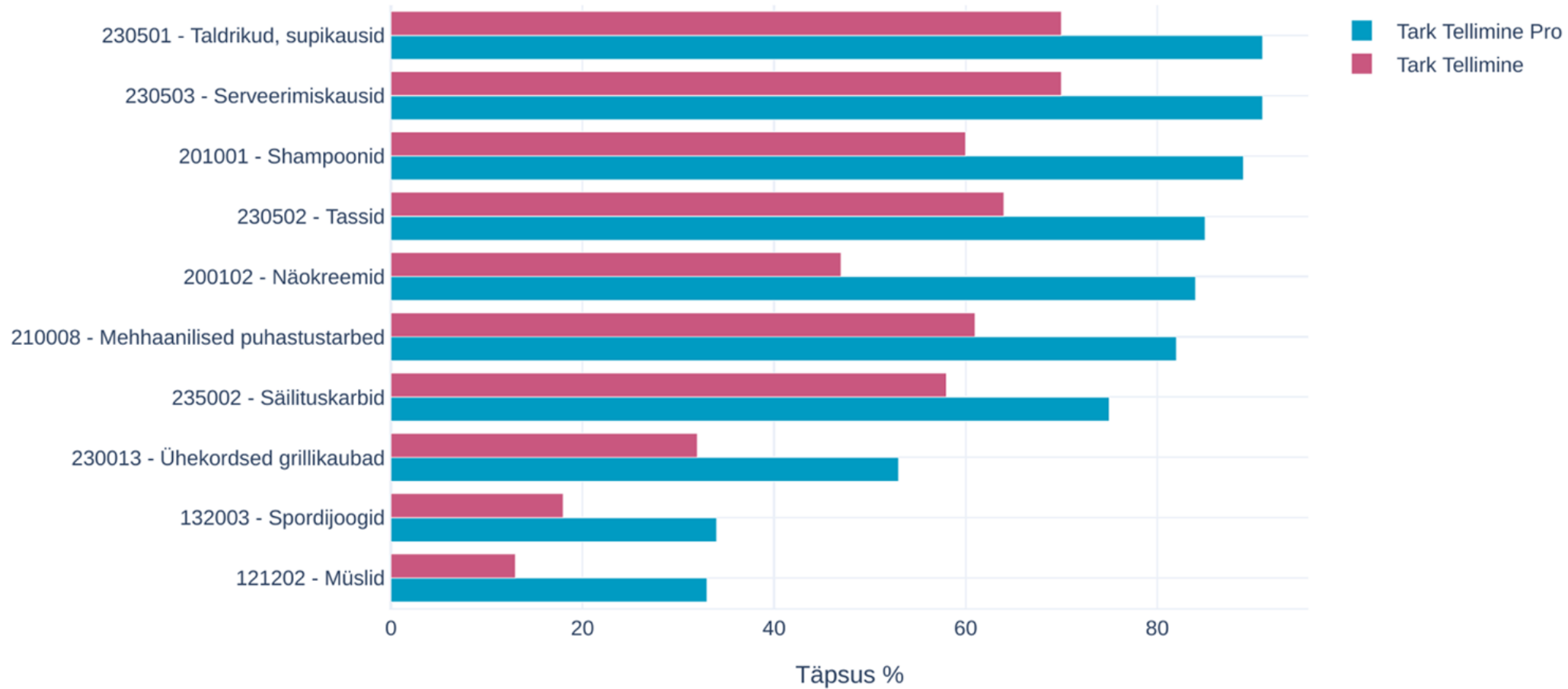
Aegread - kasutuslugu: nõudluse ennustamine

STACC

Kohv Paulig Classic filter







Andmete eelarvamuslikkus

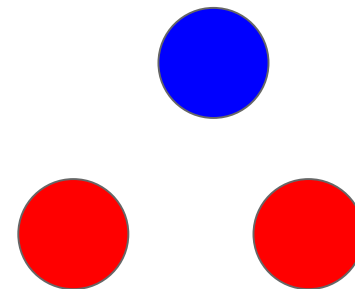
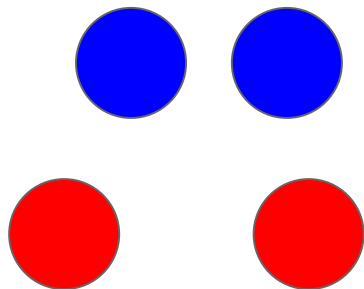
STACC

Masinõppe nõudmised andmetele

- Andmete **hulk**
 - Mida rohkem, seda parem. Mida keerulisem mudel, seda enam andmeid vajab
- Andmete **jaotus**
 - Treeningandmed kirjeldama reaalsust
- Andmete **kvaliteet**
 - Vigaste andmetega saame vigase mudeli
- Tunnuste **relevantsus** sõltuva muutuja suhtes
 - Teose žanrit ei saa ennustada ainult autori abil

Andmete hulk & jaotus

Väikese treeningandmestiku korral võime leida lihtsa kirjeldava mudeli



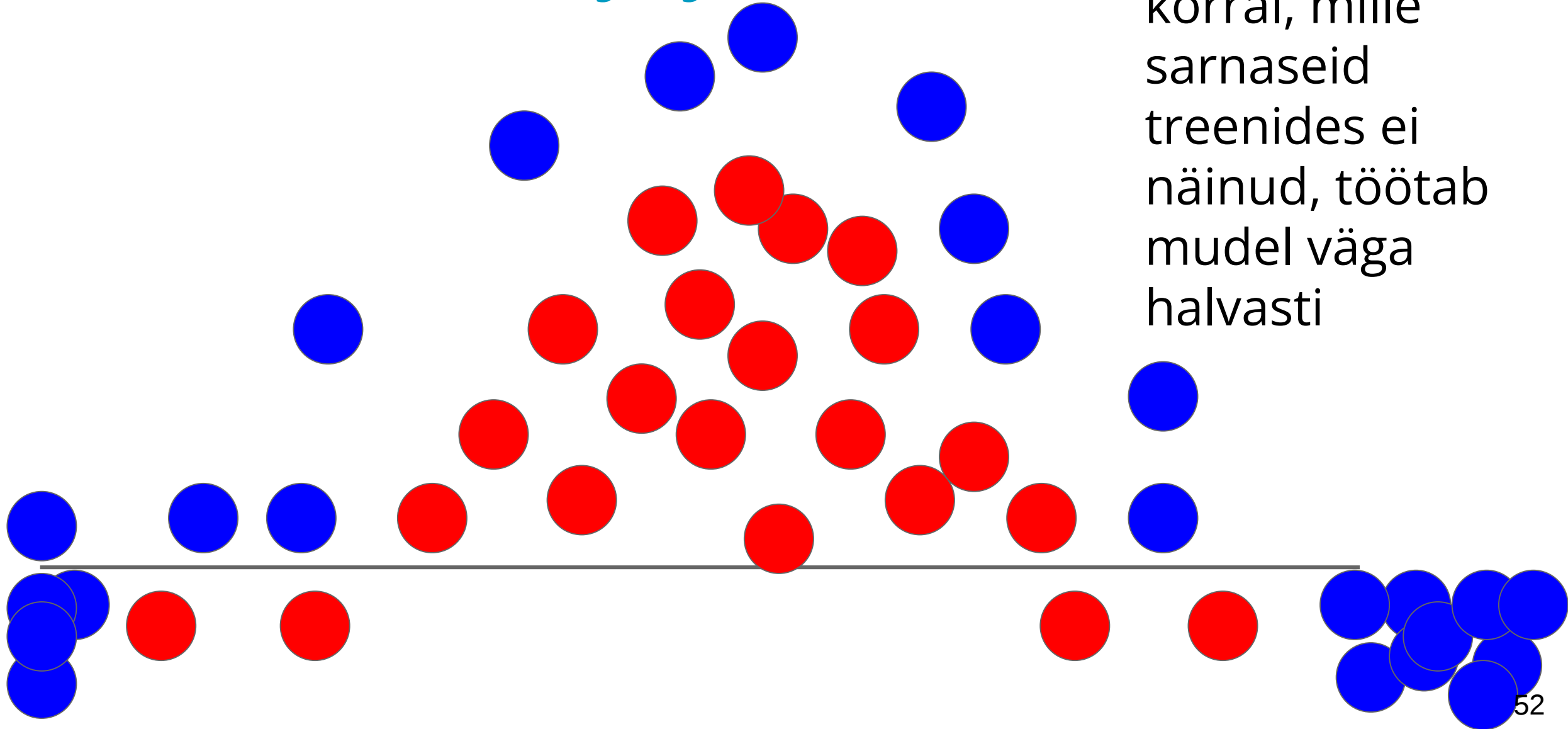
Andmete hulk ja jaotus

Väikese treeningandmestiku korral võime leida lihtsa kirjeldava mudeli



Andmete hulk ja jaotus

Uute andmete korral, mille sarnaseid treenides ei näinud, töötab mudel väga halvasti



Andmete kvaliteet

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

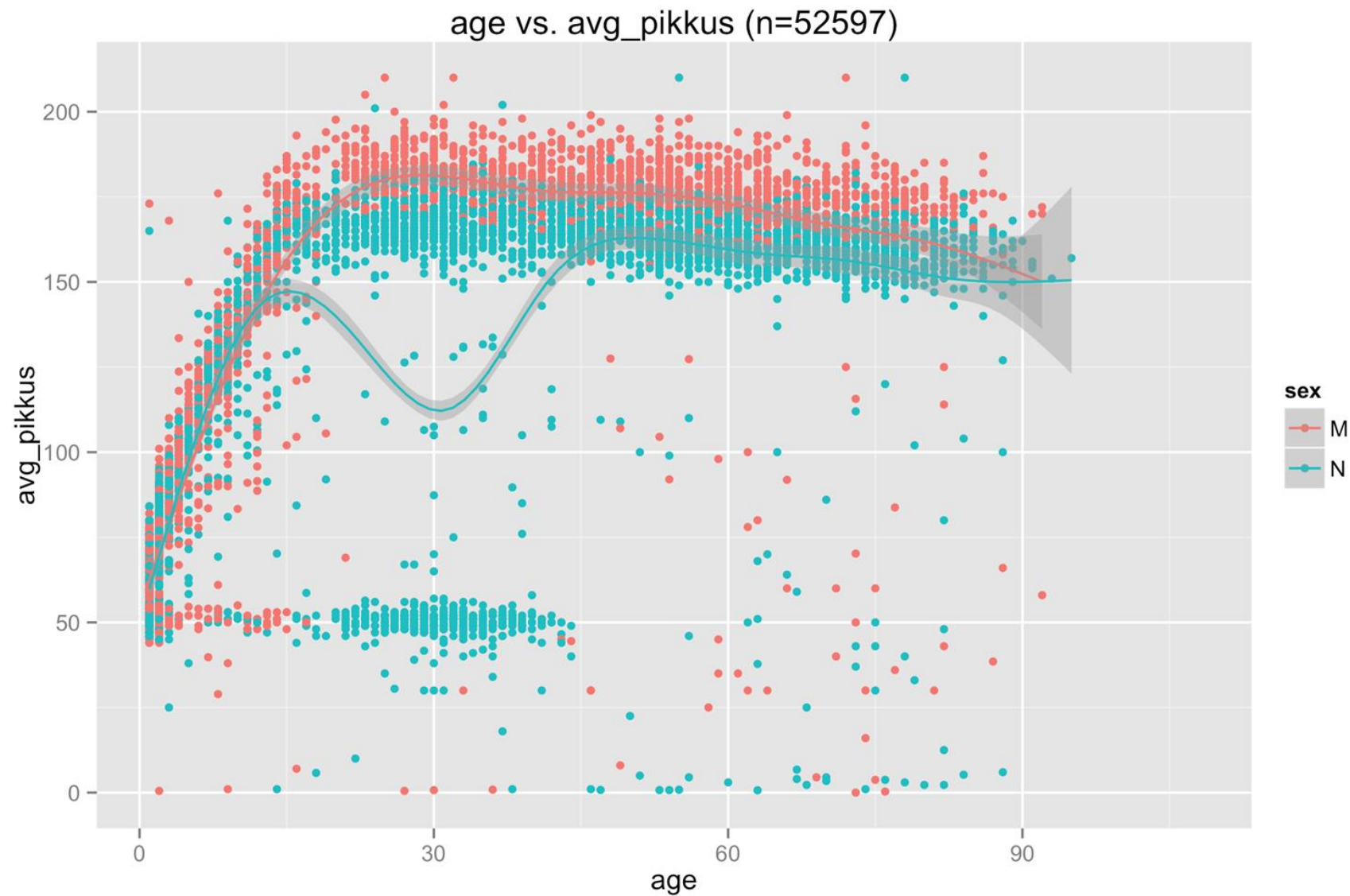
Stsenaarium: Mingil põhjusel salvestati süsteemi “katkised” andmed.

Tulem: Tehisintellekt treenib end öösel rumalaks!

Lahendus:

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

Probleemid andmetes



Andmete kallutatus - põhjuseid

- Aktiivsus eri gruppida lõikes - kes loovad / märgendavad andmeid?
- Kasutaja tagasisidel põhinev õppimine - mudeli enda toimimine mõjutab edasisi treeningandmeid
- Muudatused süsteemis - uued väljad / tegevusvõimalused
- Puudulikud andmed - peamiselt inimtegur
- Ühiskondlikud tegurid

Andmed on külmalt matemaatilised



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*

Andmete loojate eelarvamused



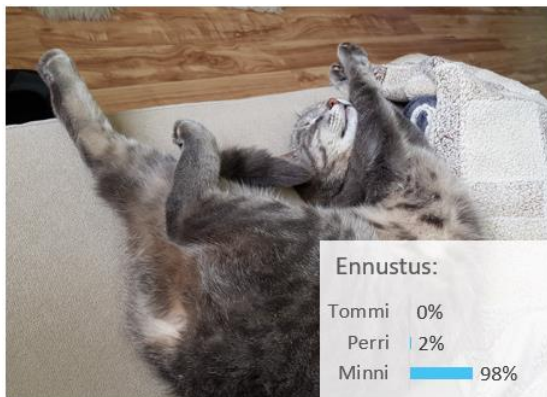
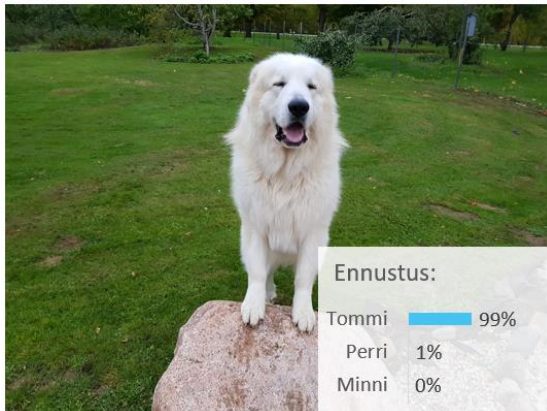
Stsenaarium: 2016. aasta 23. märtsil avas Microsoft Twitteri konto vestlusrobotile Tay; Tay suutis õppida nendelt, kellega ta suhtleb;

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

Lahendus:

- Eetika ja moraalireeglite kontroll tehisintellekti tulemile.

Seninägemata andmed



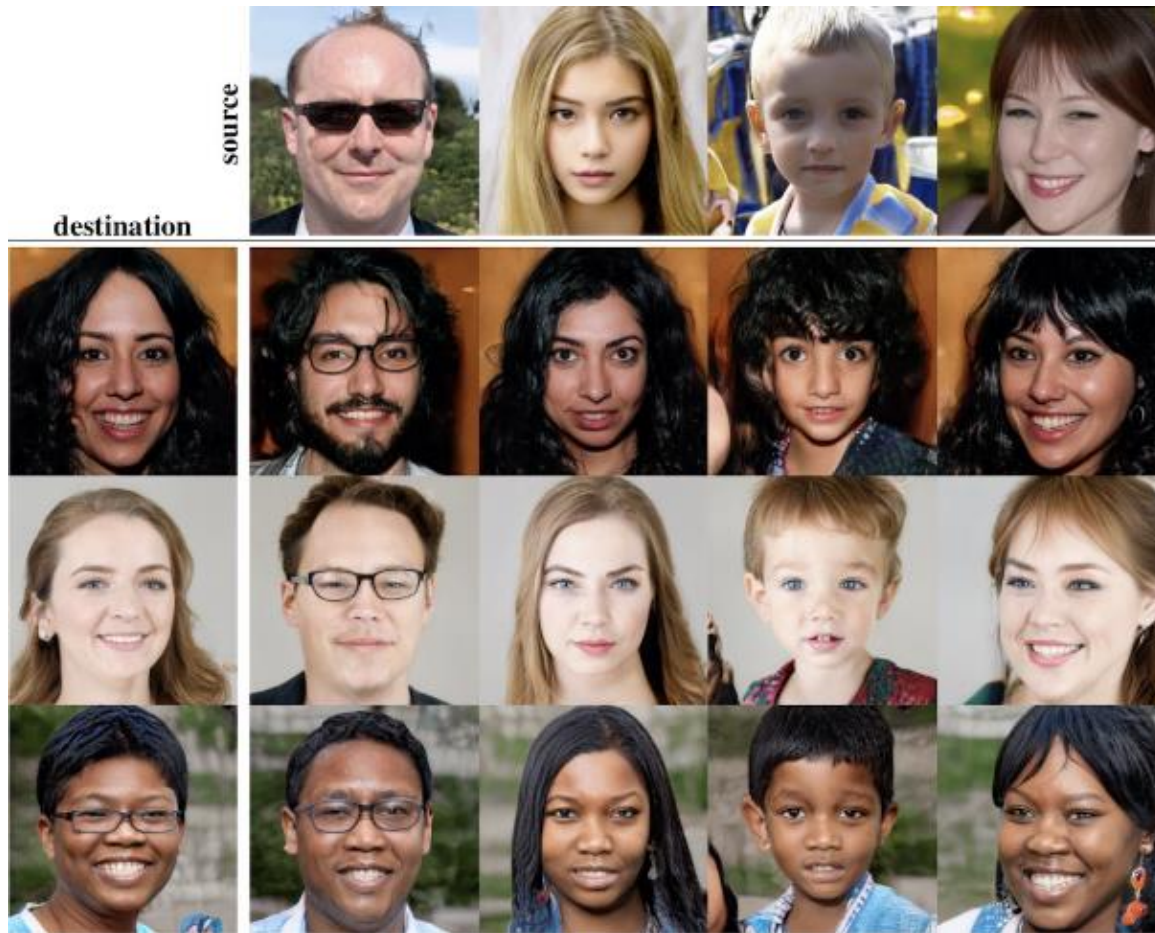
Stsenaarium: Tehisintellekt annab väga täpseid vastuseid olukordades, mille põhjal ta on treenitud ja hakkame süsteemi liigselt usaldama (*overconfidence*)!

Tulem: Uudse andmekillu nägemisel annab süsteem täiesti vale vastuse

Lahendus:

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

Valeandmed



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!?

Tehisintellekti eetika



- Vastutuse küsimus
- Ohutus ja “trammiprobleem”
- AI kasutamise eetika
- AI kui oht inimkonnale
- Privaatsus ja andmete turvalisus
 - Pseudonüümimine
 - Krüpteerimine
 - Anonüümimine

Katsetame ise ka

<https://tinyurl.com/ai-kratt2>

Kokkuvõtteks

- Tehisintellekti ja masinõppe aluseks on andmed
- Prügi sisse, prügi välja
- Paljude ülesannete jaoks on vaja andmed märgendada
- Edukas märgendamine eeldab pühendumust ja konteksti tundmist
- Masinõpe pole maagia – see on (keerukas) matemaatika
- Alusta sammhaaval: teostatavuse hinnang, piloot...

Mõtteid, küsimusi, kommentaare?

Mõtteid, küsimusi, kommentaare?

STACC OÜ

stacc.ee

Anna Laaneväli anna.laanevali@stacc.ee

Dage Särg dage.sarg@stacc.ee

Koolitust korraldab Majandus- ja Kommunikatsiooniministeerium.
Projekti rahastatakse Euroopa Liidu struktuuritoetuse
toetuskeemist "Infoühiskonna teadlikkuse tõstmine", mida
rahastab Euroopa Regionaalarengu Fond.

Üritust aitab läbi viia Conference Expert / Reisiekspert



RIIGI INFOSÜSTEEMI AMET



Euroopa Liit
Euroopa
Regionaalarengu Fond



Eesti
tuleviku heaks

TARK



e-RIIK

STACC