Esimene aasta praktiseeriva andmeteadlasena tarkvaraettevõttes

Peeter Piksarv

piksarv@gmail.com datasci.ee/meetup 7. sept 2017, Tartu

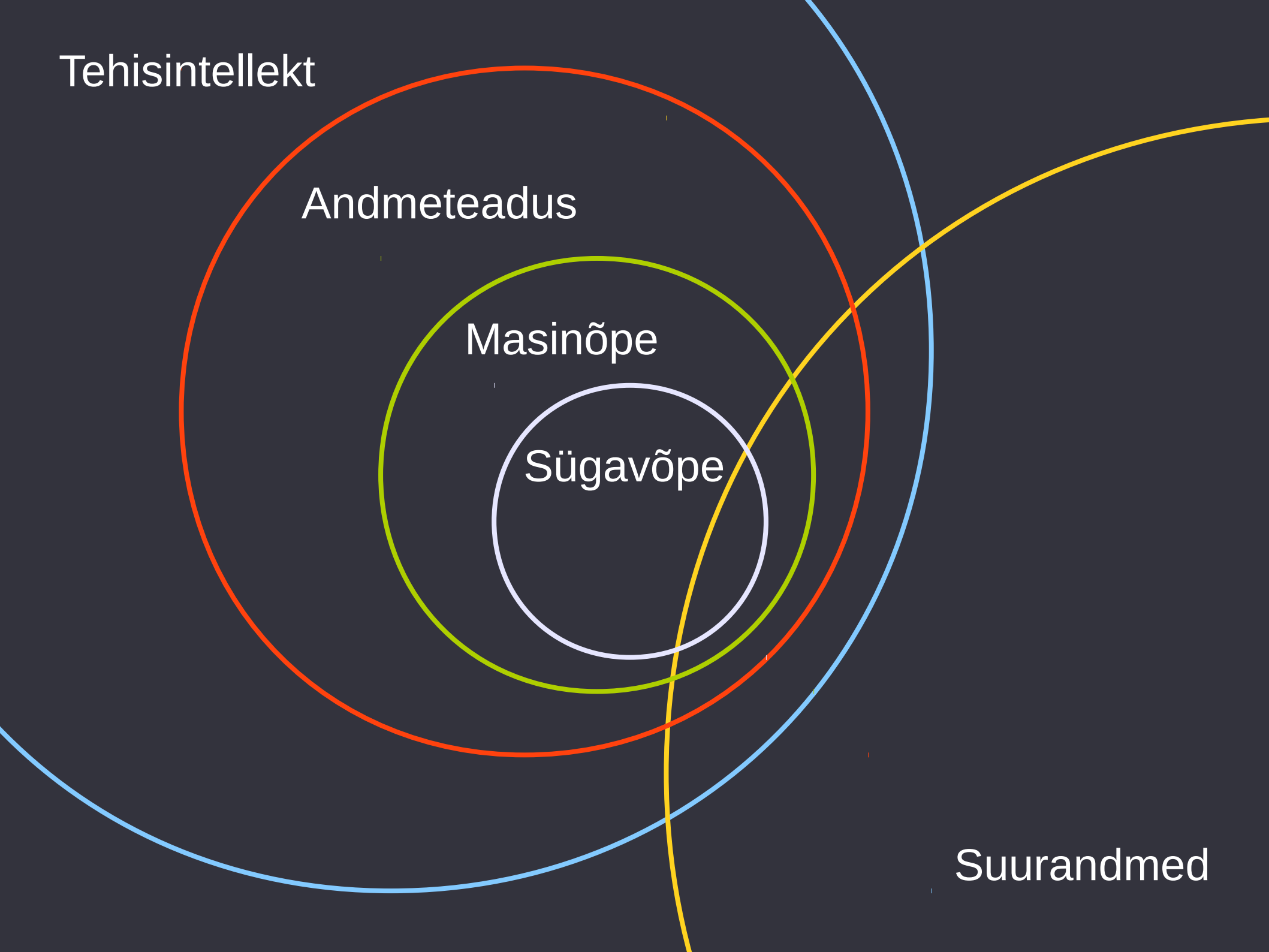
- Andmeteadus?!
  - Kes on kes andmeteaduses?
- Teadus -> andmeteadus
  - Mis kasu on doktorikraadist?
- Praktiline töö andmeteadlasena
- Akadeemia vs erasektor

- 2013 PhD füüsikas
- 2014-2016 järeldoktorantuur
- 2016- andmeteadlane

TARTU ÜLIKOOL
1632
UNIVERSITAS TARTUENSIS

University of
St Andrews

MOONCASCADE


Tehisintellekt
Andmeteadus
Masinõpe
Sügavõpe
Suurandmed

Kes tegelevad andmeteadusega?

- Ülikoolid
  - Andmeteaduse valdkond, arvutiteaduse osana
  - Tööriist (genoomika, osakeste füüsika, astronoomia)
- Ettevõtted
  - Tootearendus
  - Ärianalüütika
- Tarkvara/andmeteaduse konsultatsioonifirmad

Vt ka datasci.ee/ressursid

Andmeteadlane

Analüütik

Andmeinsener

Andmeteaduse (projekti)juht

Andmeteadlase ülesanded sõltuvad ettevõttest

Statistik

Andmehaldur

Ärianalüütik

Andmearhitekt

Miks mulle andmeteadus?

- Esmane huvi ettekandest S2DS programmi kogemusest (2015 sügis)
- Andmeteadus kui tõusev trend
- Tehisintellekt kui intrigeeriv probleem
- Eestis väljaspool ülikoole PhD-ga valikuvõimalused mõneti piiratud
- Pragmaatilised põhjused

Peamiseks õppimismaterjaliks avatud online-kursused

- Statistical Learning
  Trevor Hastie, Rob Tibshirani / Stanford Online
- Machine Learning
  Andrew Ng / Coursera
- Mining Massive Datasets
  Jure Leskovec, Anand Rajaraman, Jeff Ullman / Stanford Online
- Kaggle
- Datascience kokkutulekud
- Konverentsi ettekanted
  (PyData, PyCon, SciPy, GOTO, etc)
- DataTau, Data Science Weekly, Data Elixir, KDNuggets..

PhD-st ülekantavad oskused

- Õppimisoskus
- Probleemilahendusoskus
- Tehniline lugemisoskus
- Teaduslik meetod
- Füüsikaliste mõõtmiste taust
- Andmete visualiseerimine

Lisaks masinõppe meetoditele vajalik omandada efektiivne programmeerimisoskus

- Milline on hea kood?
- Lähtekoodi haldamine (git)
- Tarkvara testimine
- Virtualiseerimine, konteinerid
- SQL

Hea ülevaate kasutatavatest tehnoloogiatest saab töökuulutusi sirvides

Aasta jooksul osaline kuues eri projektis

- 4 seotud loomuliku keele töötlemisega
  - sh 3 vestlusrobotit
- Teadus- ja arenduskoostöö TÜ-ga mobiilside kõnelogi põhjal asukohapõhiste teenuste jaoks
- Garage48 Mooncascade ülesanne
- Pikim 6 kuud, lühim paar nädalat
- Aega ka enesetäienduseks

```
>>> from sklearn import *
```

Teadusest ~~erasektorisse~~ Mooncascade'i

- Aja jälgimine
- Igapäevane PPP (progress/plaanid/probleemid)
- Klient
- 1-2-nädalased arendustsüklid
- Programmerimise koostöövahendid (Git, JIRA)
- Tarkvaraarenduse tööriistad (Gitlab CI, Docker)
- Minimaalne elujõuline toode (MVP)

Suur osa andmeteadust on õpitav ainult läbi kogemuse

- Kuidas saada toimiv mudel, kui aega on ainult esmase versiooni väljatöötamiseks?
- Kuidas panna väljatöötatud mudel toote või teenuse sisse?
- Kuidas treenida mudeleid siis, kui andmeid ei ole?

- Igapäevane ülesannete ülevaatamine ja planeerimine annab hea struktuuri päevaks
- Olemas eraldi müügi- ja tugimeeskond
- Iseenesest programmeerimine on hea kiire tagasisidega protsess, kus tulemus ruttu näha
- Lühemad projektid ja väiksemateks osadeks jaotatud ülesanded annavad saavutamise tunde
- On olnud võimalus väga palju õppida ja katsetada

Aitäh!