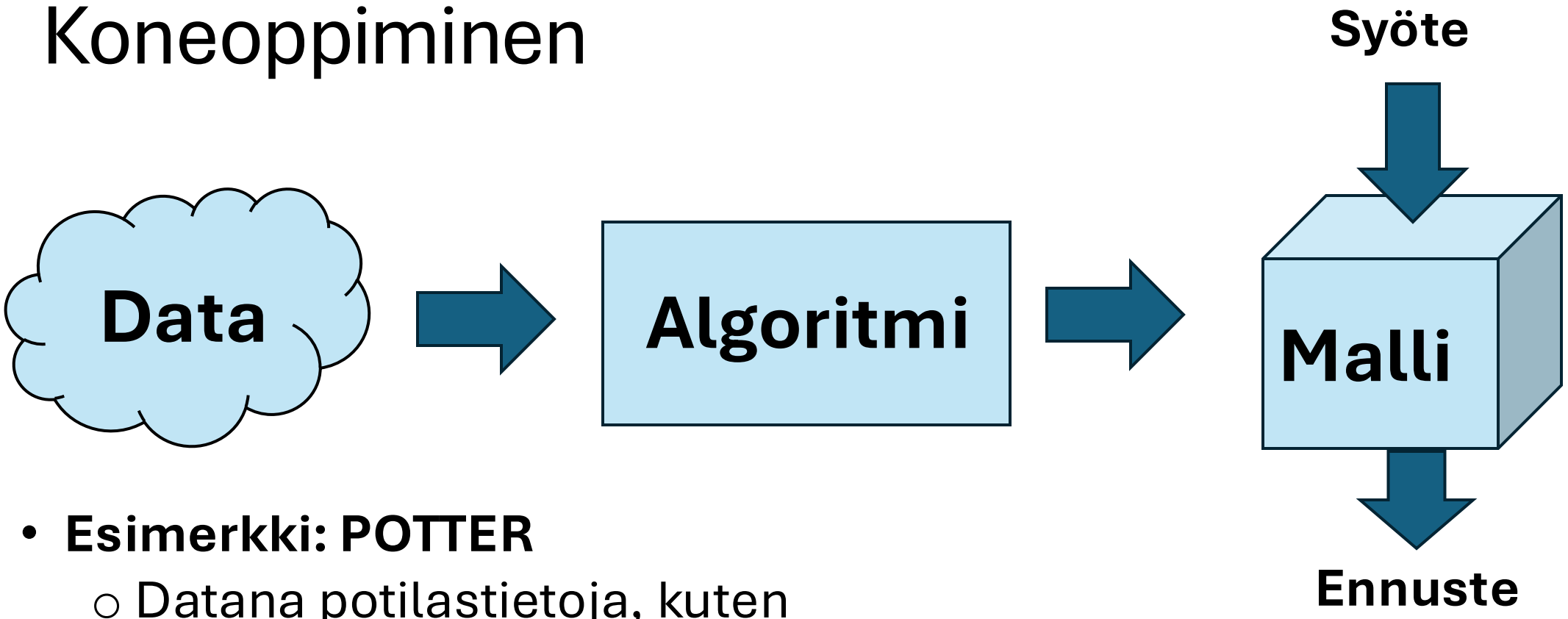
Moderni selitettävä tekoäly

Reijo Jaakkola

Tampere University

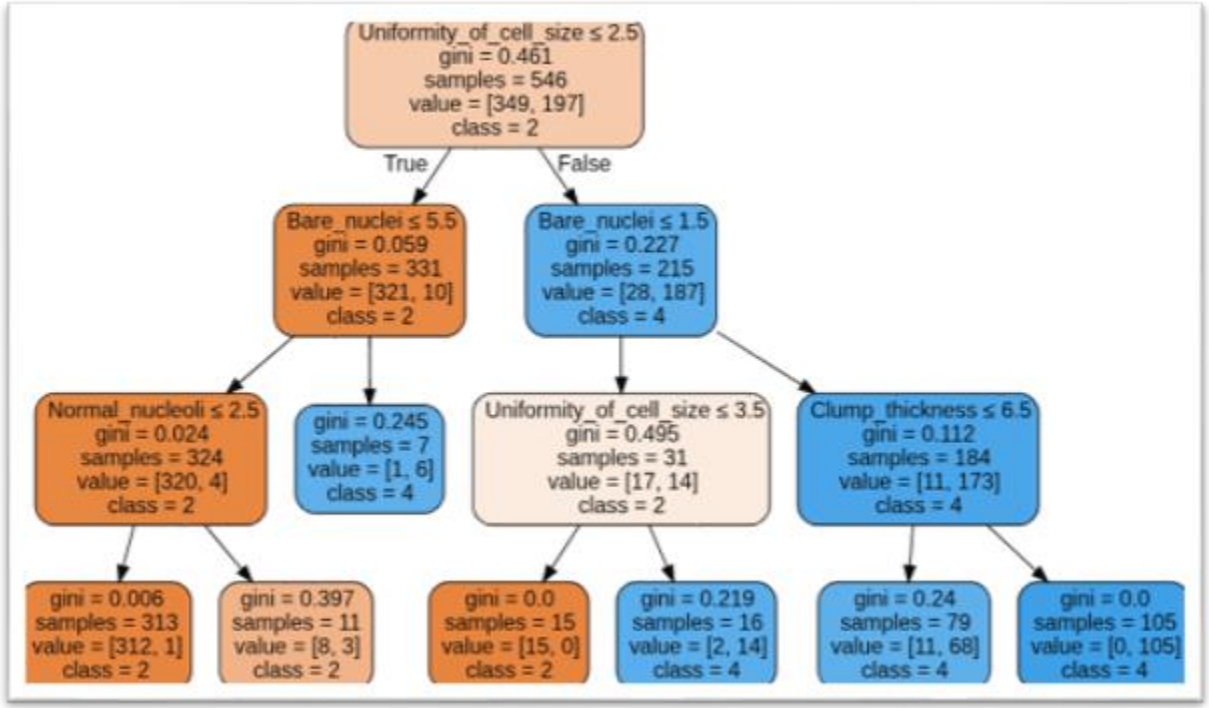
Koneoppiminen



- **Esimerkki: POTTER**
  - Datana potilastietoja, kuten
    - ikä,
    - laboratoriokokeiden tuloksia ja
    - olemassa olevat sairaudet.
  - Ennustetaan riskiä hätäleikkauksen jälkeiselle komplikaatiolle.

Selitettävä vs. ei-selitettävä koneoppiminen

- **Black-box mallit**
  - Ennusteiden selittäminen vaikeata/mahdotonta.
  - **Esim**. Neuroverkot.
- **White-box mallit**
  - Ennusteet voidaan selittää.
  - **Esim**. Päätöspuut, lineaarinen regressio.
- Teollisuudessa käytetään usein black-box malleja niiden tarkkuuden vuoksi.

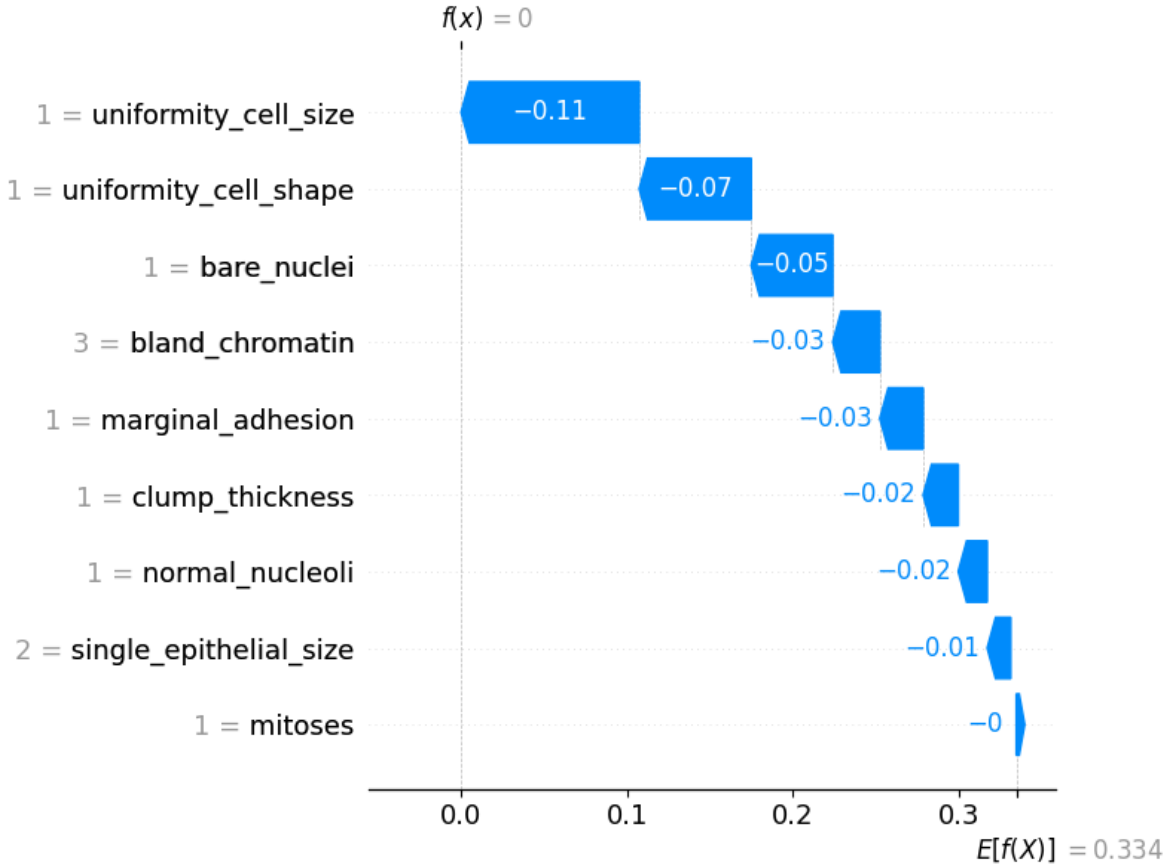


Black-box mallien haaste

- **Esimerkki:** Tutkimus vuodelta 1997 jossa tarkasteltiin potilaita, joilla oli keuhkokuumeen oireita.
- Malli ennusti, onko potilas korkean riskin vai ei. Sen päätökset perustuivat esimerkiksi ikään ja ennestään olemassa oleviin sairauksiin, kuten astmaan.
- Malli oli **tarkka**, mutta osoittautui, että sen mielestä astma pienensi potilaan riskiä.
- Tutkimusdatan astmapotilaat saivat nopeampaa ja intensiivisempää hoitoa.

Black-box mallien selittäminen

- Posthoc-selittäminen suosittua.
- Yksi käytetyimmistä työkaluista on **SHAP**, joka antaa painoja eri ominaisuuksille.
- **Haasteita:**
  - Hankala selittää mitä painot itseasiassa tarkoittavat.
  - Painot voivat olla harhaanjohtavia.



Modernit white-box mallit

> **Cynthia Rudin**: *"It is a myth that there is necessarily a trade-off between accuracy and interpretability."*

- White-box mallien suurin vahvuus on se, että näemme, mitä virheitä ne tekevät, jolloin voimme korjata ne.
- White-box-mallien oppimisalgoritmit ovat viime vuosina kehittyneet huomattavasti:
  - Optimal Decision Trees
  - CORELS
  - Explainable Boosting Machines

White-box mallit Tampereen yliopistossa

- Kehitämme matemaattiseen logiikkaan perustuvia white-box malleja.
- Monissa tapauksissa yhtä tarkkoja kuin tarkimmat black-box mallit.
- Oppiminen on hyvin nopeata ja laskennallisesti kevyttä.

```
IF
   uniformity_of_cell_shape ⩾ 2.50 AND uniformity_of_cell_size ⩾ 2.50
   OR bland_chromatin ⩾ 3.50 AND uniformity_of_cell_shape < 4.50 AND
   uniformity_of_cell_size ⩾ 2.50
   OR bland_chromatin ⩾ 3.50 AND uniformity_of_cell_shape ⩾ 2.50 AND
   uniformity_of_cell_shape < 4.50 AND uniformity_of_cell_size < 4.50
THEN class = 4

ELSE class = 2
```