# Strojno učenje na velikih podatkih

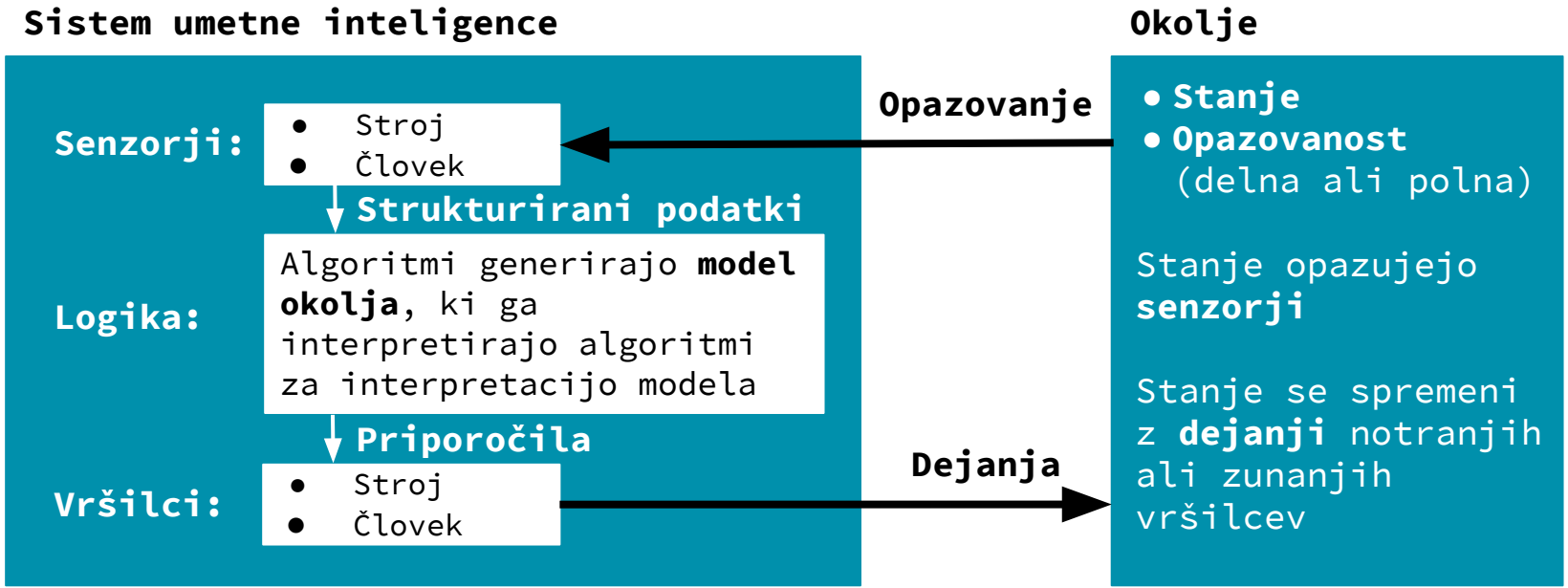
## Machine learning for big data

**Jasna Urbančič**
Laboratorij za umetno inteligenco
20. Junij 2019

Institut
”Jožef Stefan”
Ljubljana, Slovenija

# Strojno učenje

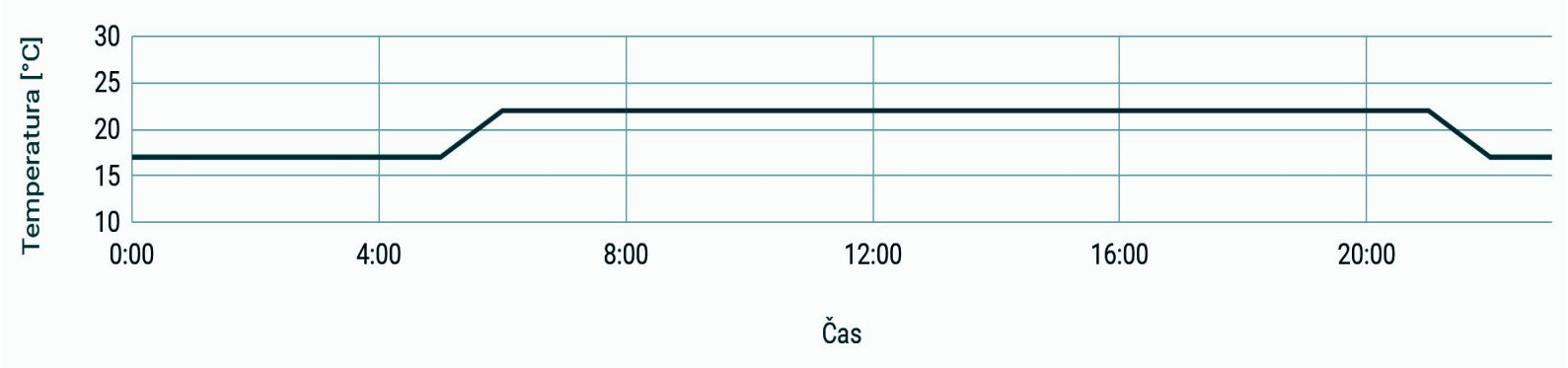
# Umetna inteligenca in strojno učenje




Prilagojeno po priporočilih ekspertne skupine OECD za umetno inteligenco.

# Umetna inteligenca in strojno učenje - ilustrativni primeri

- Termostat: senzor in vršilec je stroj




- Odobritev kredita: vršilec je človek, ki odločitev sprejme (delno) na podlagi priporočila modela

# Strojno učenje

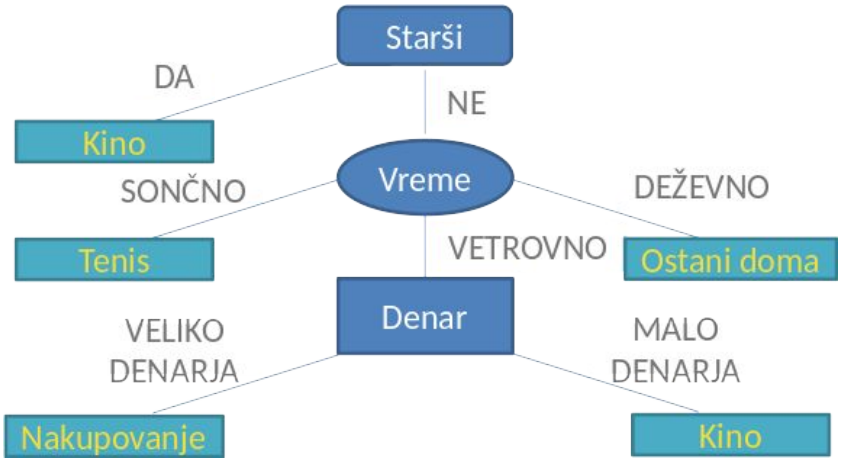
**Podatki**

Značilke

Tarčna spremenljivka

| Teden | Vreme | Starši na obisku | Denar | Odločitev (kategorija) |
|---|---|---|---|---|
| T1 | Sončno | Da | Veliko | Kino |
| T2 | Sončno | Ne | Veliko | Tenis |
| T3 | Vetrovno | Da | Veliko | Kino |
| T4 | Deževno | Da | Malo | Kino |
| T5 | Deževno | Ne | Veliko | Ostal doma |
| T6 | Deževno | Da | Malo | Kino |
| T7 | Vetrovno | Ne | Malo | Kino |
| T8 | Vetrovno | Ne | Veliko | Nakupovanje |
| T9 | Vetrovno | Da | Veliko | Kino |
| T10 | Sončno | Ne | Veliko | Tenis |

**Model**



# Strojno učenje - model

---

Iščemo **funkcijo f(x)** s **parametri a, b, c, ...**, ki opiše **podatke** $(x_i, y_i)$.

**Optimiziramo parametre**, tako da je **napaka** čim manjša.

Definicija napake se razlikuje od aplikacije do aplikacije.

**Model:** Funkcija f in množica optimiziranih parametrov a, b, c...

# Strojno učenje - variacije

**Nadzorovano učenje**

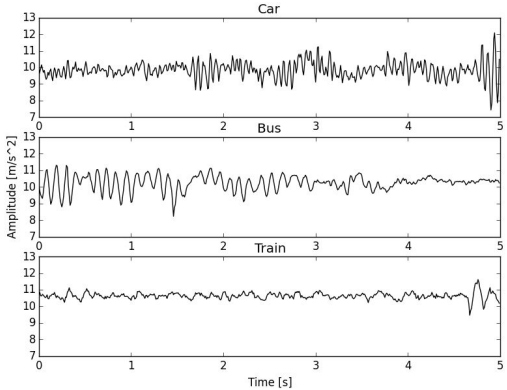
- **Znane** vrednosti tarčnih spremenljivk ($y_i$)
- Primeri: klasifikacija, regresija

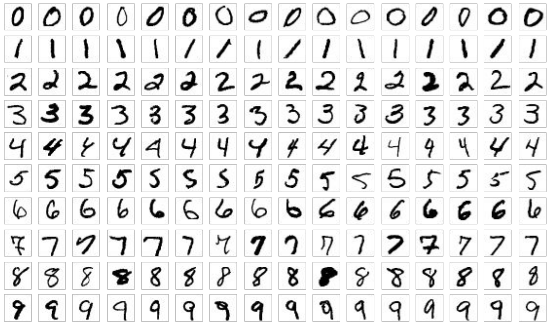
**Nenadzorovano učenje**

- Tarčne spremenljivke in njihove vrednosti **niso znane** - pogosto gre za iskanje podobnosti
- Primeri: razvrščanje v skupine, detekcija anomalij

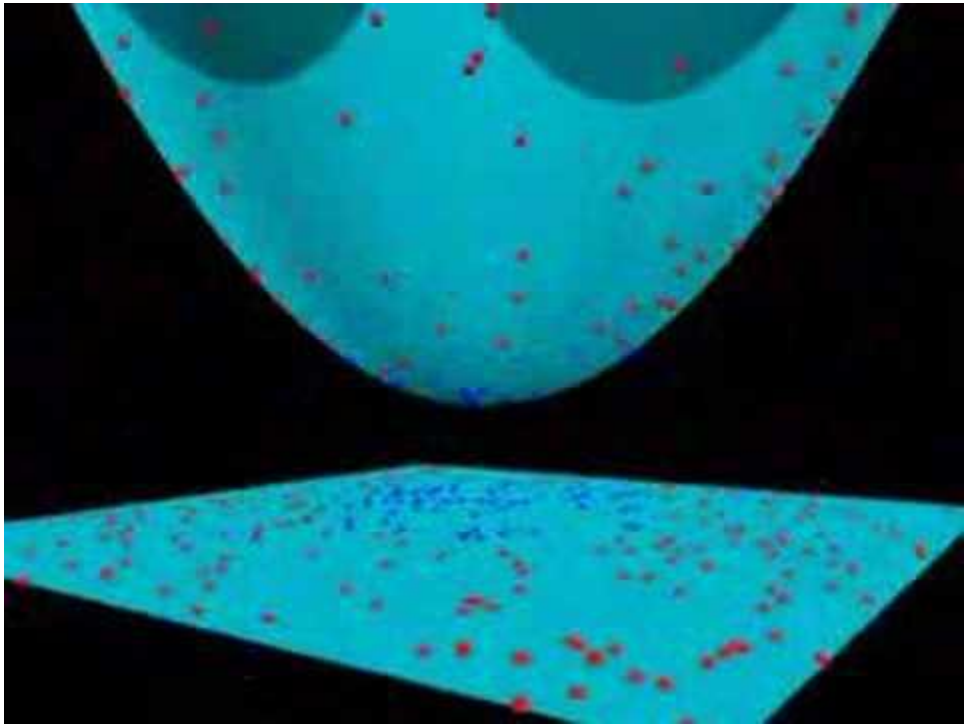
# Strojno učenje - klasifikacija

- **Razvrščanje v razrede**
- Primerjava na podlagi podobnosti ali razdalje
- V praksi:
  - Detekcija prevoznega sredstva
  - Razpoznava ročno zapisanih števk



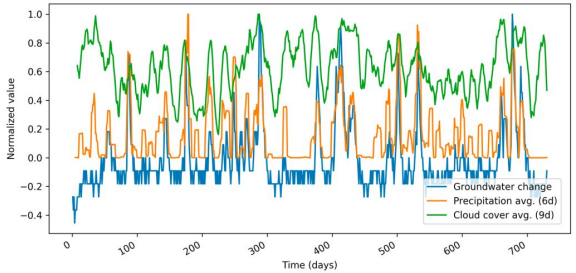
# Strojno učenje - klasifikacija

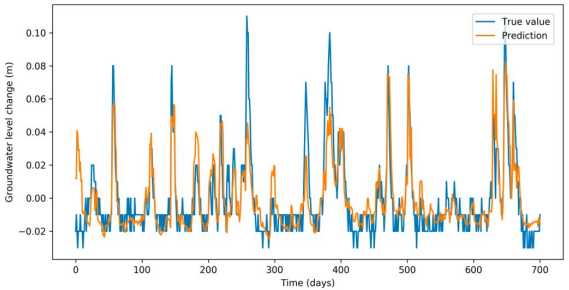


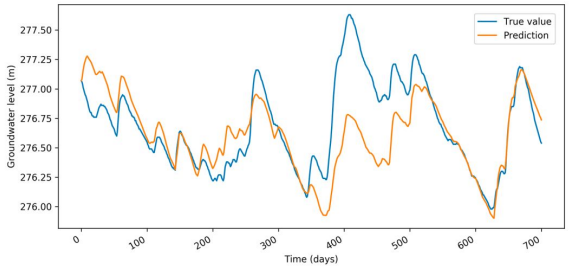
https://www.youtube.com/watch?v=3liCbRZPrZA

# Strojno učenje - regresija

- Uporabno pri **napovedovanju številskih vrednosti**
  - Občasno lahko uporabimo regresijo, nato pa izhod interpretiramo kot kategorije - klasifikacija
- V praksi
  - Napovedovanje porabe energije
  - Ocenjevanje nivoja podtalnice

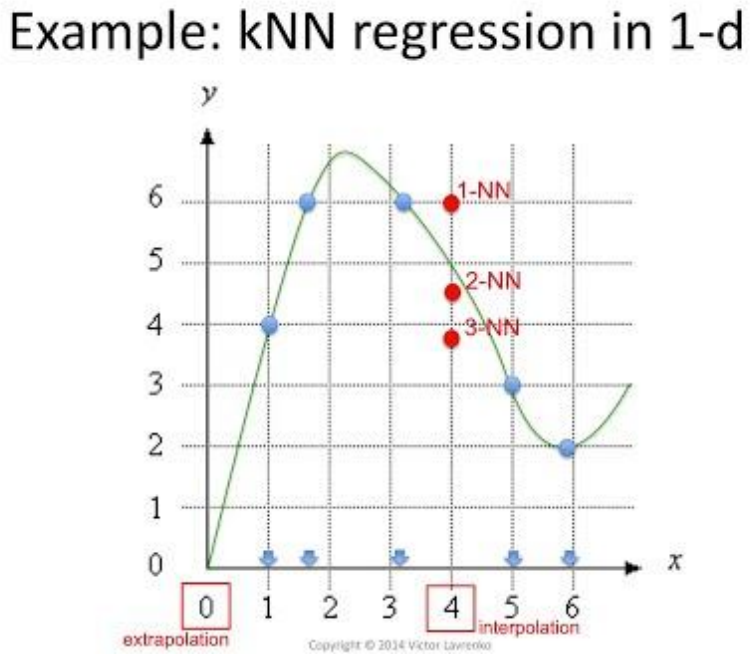



Kenda, K., Čerin, M., Bogataj, M., Senožetnik, M., Klemen, K., Pergar, P., ... & Mladenić, D. (2018). Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer. In *Multidisciplinary Digital Publishing Institute Proceedings* (Vol. 2, No. 11, p. 697).

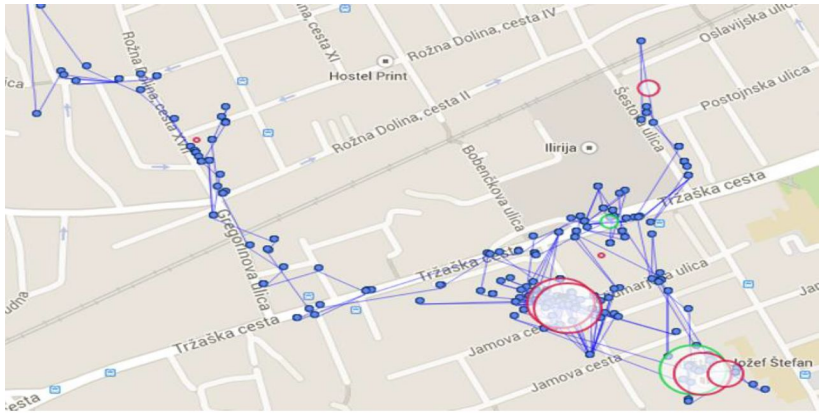
# Strojno učenje - regresija

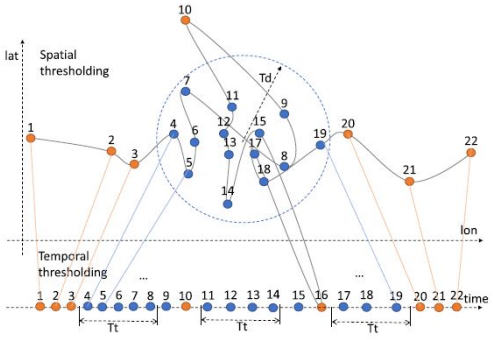



https://www.youtube.com/watch?v=3lp5CmSwrHI

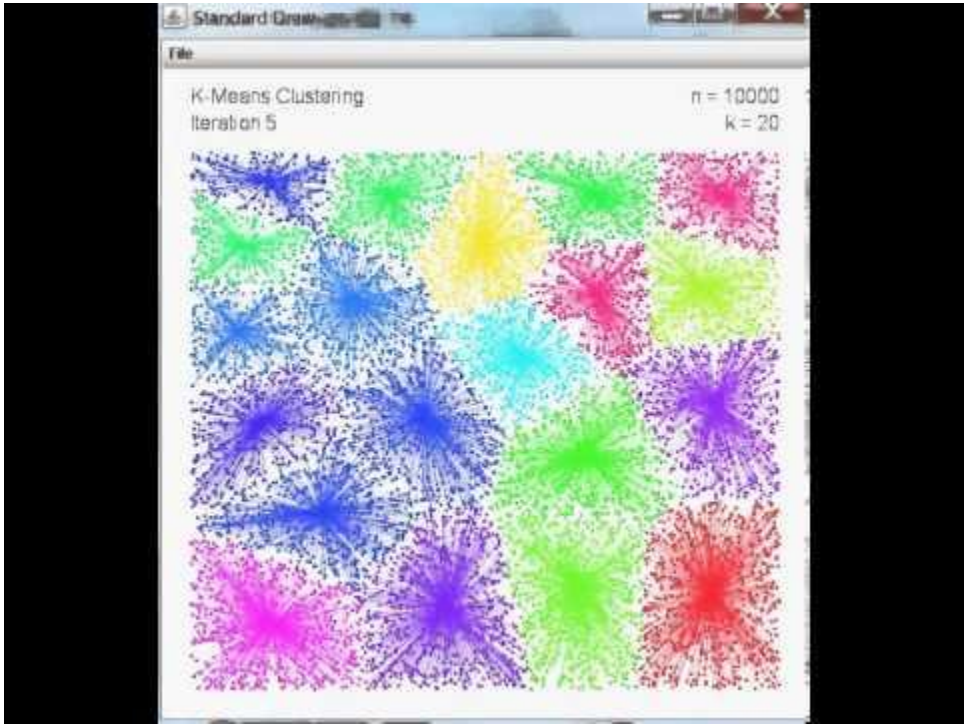
# Strojno učenje - razvrščanje v skupine

- Vnose želimo razporediti v skupine **glede na podobnost** njihovih značilk
- V praksi:
  - Odkrivanje interesnih točk iz sledi GPS



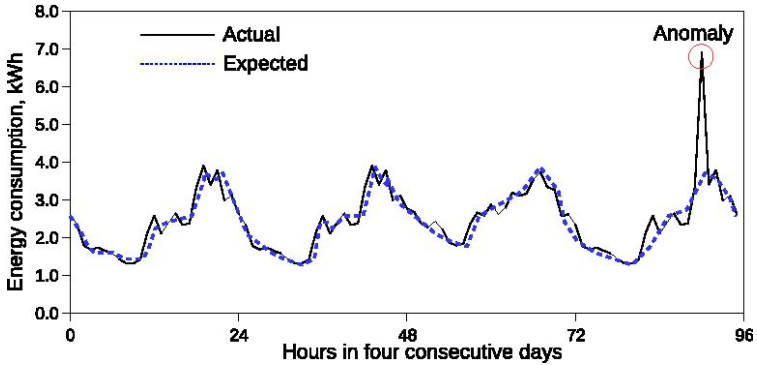
# Strojno učenje - razvrščanje v skupine




https://www.youtube.com/watch?v=BVFG7fd1H30

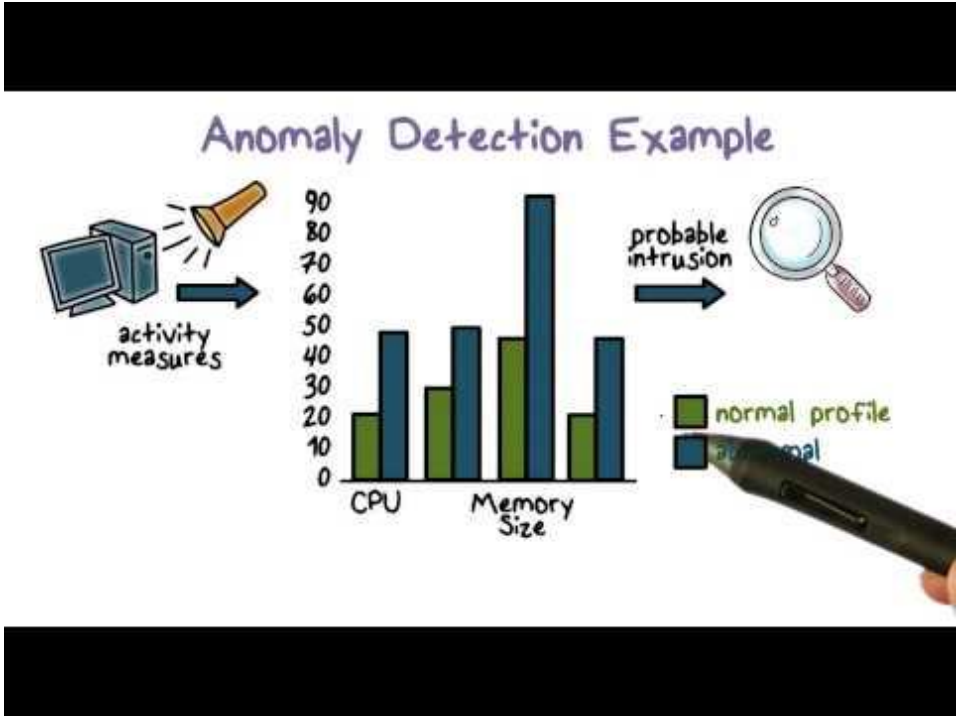
# Strojno učenje - detekcija anomalij

- V podatkih želimo **identificirati redke dogodke**, elemente, pojave, ki se **močno razlikujejo** od ostalih vrednosti
- Anomalije so lahko tudi osamelci, novitete, šum …
- V praksi:
  - Bančne goljufije
  - Spletni incidenti in vdori
  - Strukturni defekti




Vir slike: Liu, Xiufeng and Per Sieverts Nielsen. “Regression-based Online Anomaly Detection for Smart Grid Data.” *CoRR* abs/1606.05781 (2016): n. pag.

# Strojno učenje - detekcija anomalij




https://www.youtube.com/watch?v=stdKi0grdhg

# Primeri strojnega učenja na velikih podatkih

# Primer 1: Modeliranje verjetnosti neplačila

---

**Opis problema:**

- Podjetje nosi tveganje, da stranke ne poplačajo svojih obveznosti.
- Podjetje lahko omeji svoje tveganje:
  - Zavarovanje obveznosti
  - **Preventivni ukrepi**

**Cilj:**

- Modeliranje verjetnosti neplačila posamezne stranke
- Priporočilo višine limita kredita glede na izračunano verjetnost neplačila

Herga, Z., Rupnik, J., Škraba, P., & Fortuna, B. Modeling Probability of Default and Credit Limits. In Conference on Data Mining and Data Warehouses.

# Primer 1: Modeliranje verjetnosti neplačila - podatki

- **Finančni podatki** - javnodostopni podatki o prihodkih in odhodkih podjetja, bilance, letna poročila podjetij

- **Podatki o trgovanju** med podjetjem in njegovimi strankami - mesečni podatki o obsegu trgovanja, obstoječih dolgovih, izpodbijanih terjatvah in zapoznelih plačilih

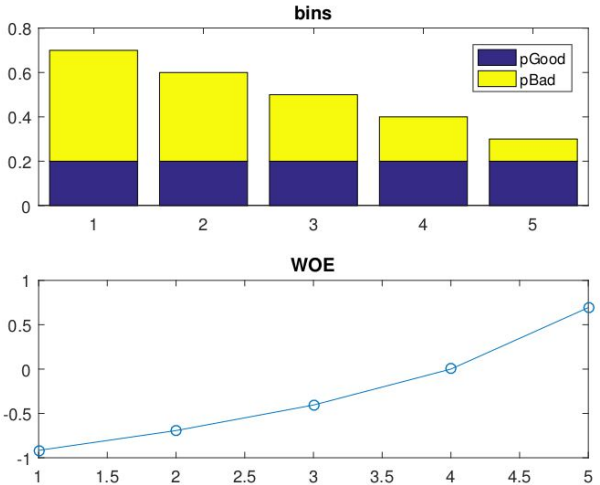
Herga, Z., Rupnik, J., Škraba, P., & Fortuna, B. Modeling Probability of Default and Credit Limits. In Conference on Data Mining and Data Warehouses.

# Primer 1: Modeliranje verjetnosti neplačila - značilke

- Finančni podatki ⟼ finančni indikatorji

- Finančni indikatorji ⟼ značilke
  - Razvrščanje v razrede
  - Transformacija v *weight of evidence (WOE)*

$$WOE = log \frac{P(company=good)}{P(company=bad)}$$



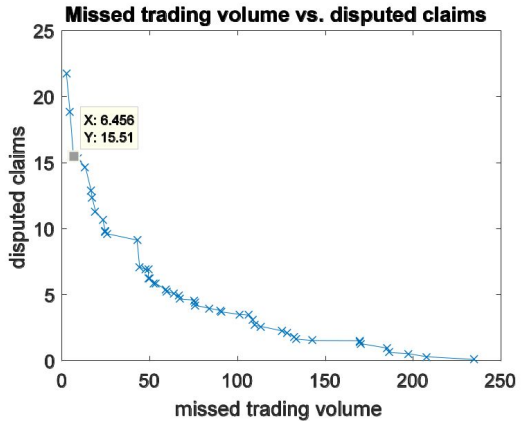

Herga, Z., Rupnik, J., Škraba, P., & Fortuna, B. Modeling Probability of Default and Credit Limits. In Conference on Data Mining and Data Warehouses.

# Primer 1: Modeliranje verjetnosti neplačila - model

---

Verjetnost neplačila za podjetje

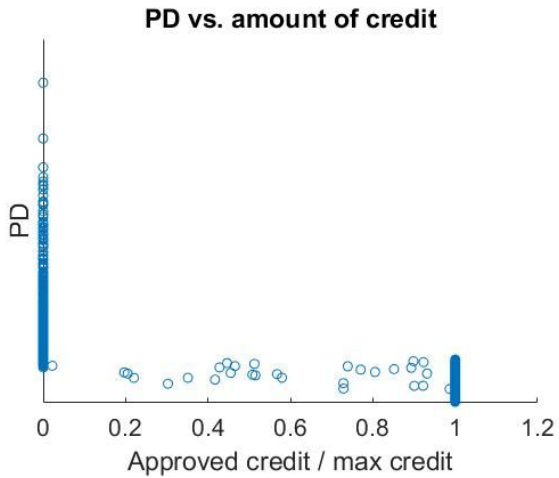
$$PD(x) = \frac{1}{1+exp[-(\beta_0+\beta_1 WOE(x_1)+\beta_2 WOE(x_2)+\cdots+\beta_n WOE(x_n))]}$$




Herga, Z., Rupnik, J., Škraba, P., & Fortuna, B. Modeling Probability of Default and Credit Limits. In Conference on Data Mining and Data Warehouses.

# Primer 1: Modeliranje verjetnosti neplačila - kreditni limit

Optimizacija izgubljenega prometa proti izgubi zaradi neplačil




Herga, Z., Rupnik, J., Škraba, P., & Fortuna, B. Modeling Probability of Default and Credit Limits. In Conference on Data Mining and Data Warehouses.

# Primer 2: Napovedovanje porabe električne energije

---

**Opis problema:**

- Modelirati želimo energetske fenomene (npr. poraba električne energije, cena energije na trgu, proizvodnja električne energije)
- Senzorski podatki prihajajo v sistem z visoko frekvenco (npr. električni tok, električna moč ...)
- Podatki iz napovedi (npr. vremenska napoved) se običajno spreminjajo
- Statične podatke lahko izračunamo vnaprej (npr. dan v tednu, čas v dnevu, dan v letu, lunine mene, prazniki ...)

Kenda, K., & Škrjanc, M. Modelling in Energy Related Scenarios. In Conference on Data Mining and Data Warehouses.

# Primer 2: Napovedovanje porabe električne energije

---

**Cilj:**

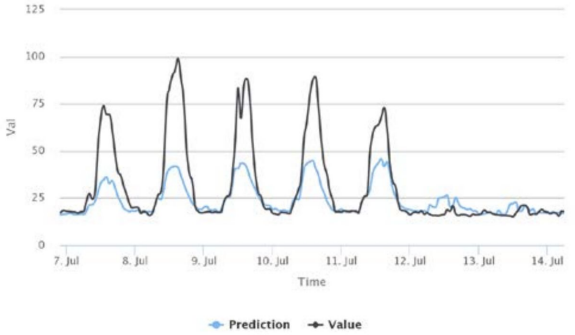
- Čim bolj natančno in zanesljivo želimo modelirati (**napovedovati**) energetske fenomene ob upoštevanju senzorskih in statičnih podatkov ter podatkov iz napovedi

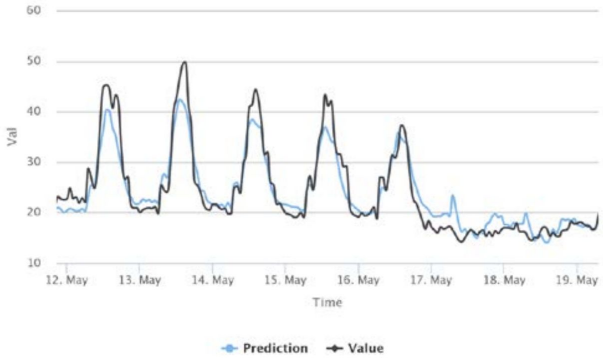
Kenda, K., & Škrjanc, M. Modelling in Energy Related Scenarios. In Conference on Data Mining and Data Warehouses.

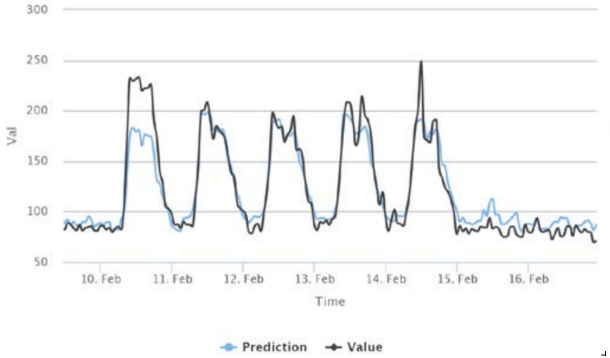
# Primer 2: Napovedovanje porabe električne energije

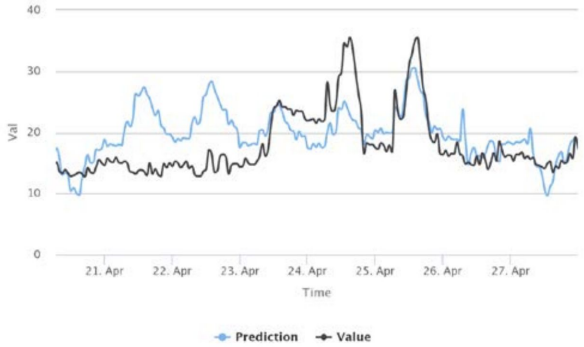
## Poraba javne ustanove

- Dobre značilke: tedenski agregati senzorskih značilk količina sončnega obsevanja, delovni čas, kurilna sezona, etc.
- Kakovost podatkov je včasih vprašljiva: manjkajoče in napačne vrednosti



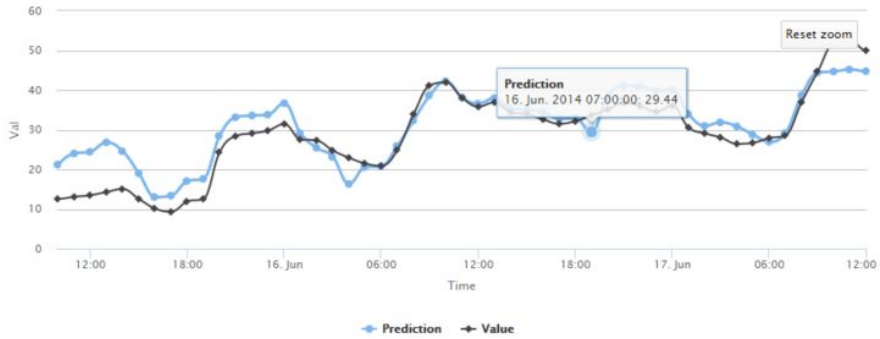



# Primer 2: Napovedovanje porabe električne energije

---

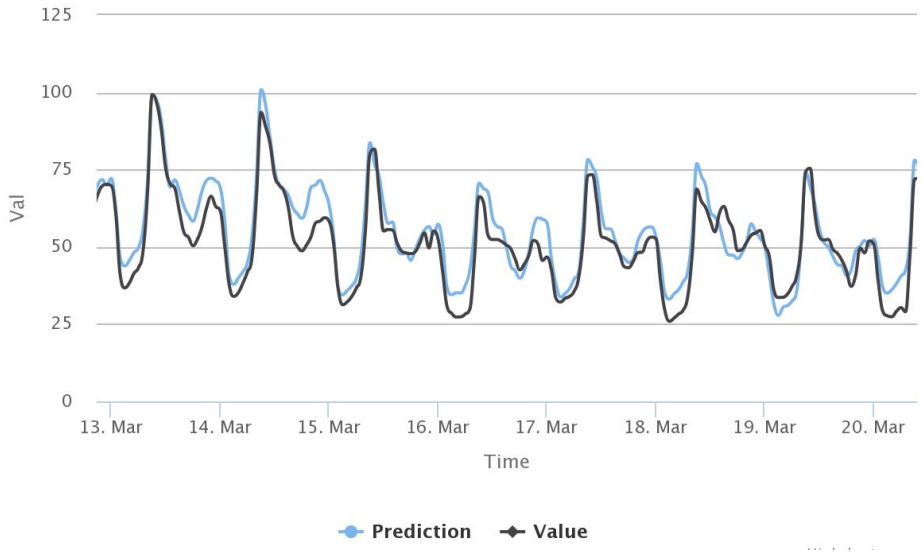
**Cena elektične energije (Energy Spot Market - EPEX)**

- Močno odvisna od proizvodnje elektrike iz alternativnih virov
  - Cena proizvodnje je nizka, a omrežje ni optimizirano za neredne vire, zato se cena elektrike zniža
- Dobre značilke: cena v prejšnih dneh, povprečja v tednu/mesecu, smer vetra



# Primer 2: Napovedovanje porabe električne energije

## Modeliranje proizvodnje v termoelektrarni




Kenda, K., & Škrjanc, M. Modelling in Energy Related Scenarios. In Conference on Data Mining and Data Warehouses.

# Vprašanja in diskusija