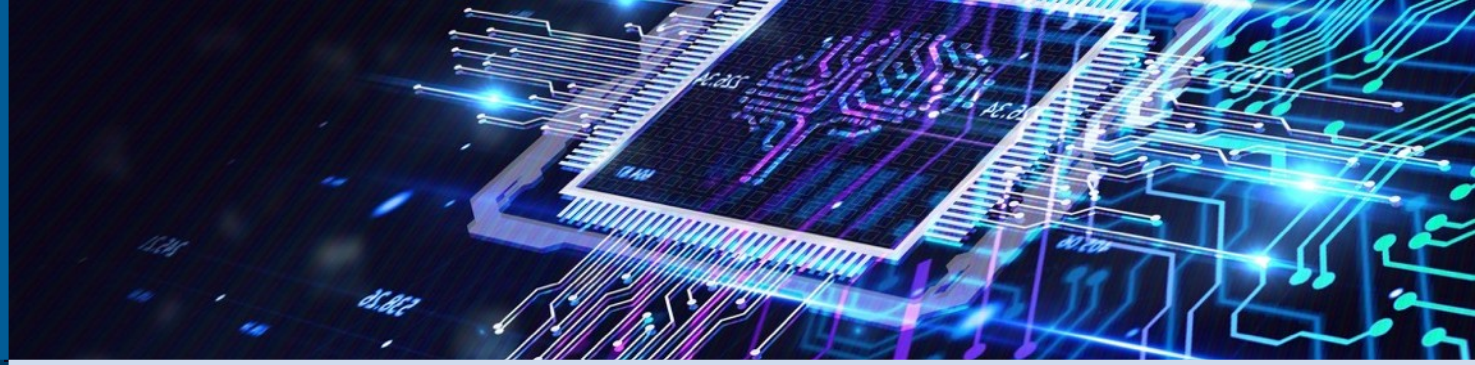
CSC

ICT Solutions for Brilliant Minds



High-performance computing for automatic text annotation

HPD-hankkeen loppuwebinaari, 19.1.2021

Mats Sjöberg, CSC

Automatic text annotation

"Löytyykö ainestoa kädestäennustamisesta tai tarot-korteista, tarttisin pikaista palautetta, kiitos yhteistyöstänne. Tampereen kaupunginkirjastosta löytyy..."

⇒ tarokki, ennustaminen, kädestä ennustaminen, korteista ennustaminen

"Hanna Vehmas Liikuntamatkalla Suomessa Vapaa-ajan valintoja jälkimodernissa yhteiskunnassa Esitetään Jyväskylän yliopiston liikunta- ja terveystieteiden..."

⇒ liikuntamatkailu, liikuntasosiologia, vapaa-aika, vapaa-ajantoiminnat, matkailu, urheilu, liikuntaharrastus, matkailijat, motiivit, luontosuhde

"Nopeat ja näppärät Tämä kauniisti toteutettu keittokirja sisältää veden kielelle nostattavan valikoiman nopeita ja helppoja reseptejä eri puolilta maailmaa..."

⇒ ruokaohjeet

"tarkoitettu vakavuuteen laivan matemaattisia dynaamista Jerzy vakavuutta ISBN sen kallistumista takia koskevia. laivasuunnittelijan käyttäytymistä ..."

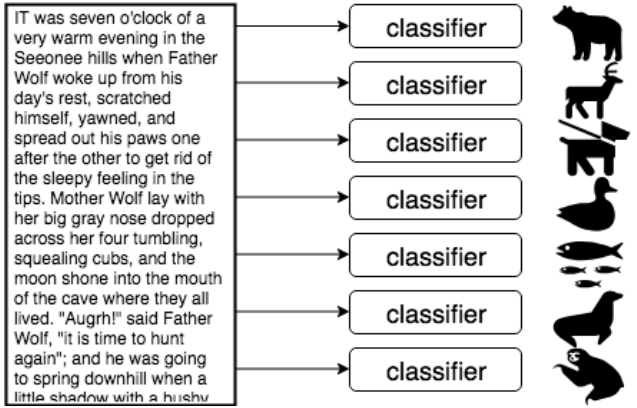
⇒ laivat, hydrodynamiikka, hydrostatiikka

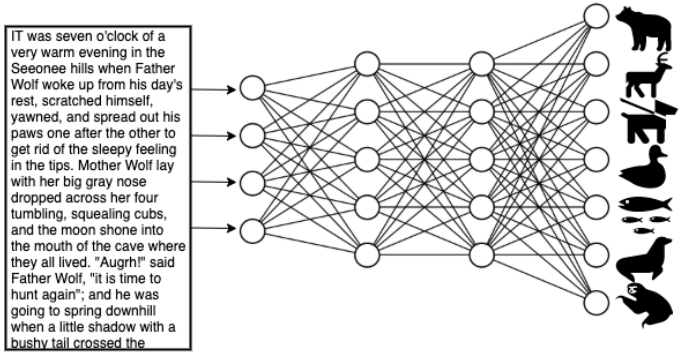
Extreme multi-label text classification (XMTC)

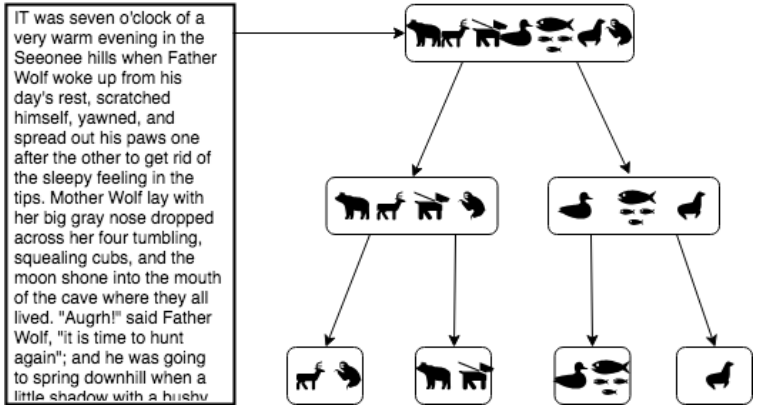
- The number of training examples, the dimensionality of data, and especially **the number of labels** are large
- Issues:
  - sparsity
  - label correlations
  - scalability
  - computational cost
- E.g. YSO contains more than 30,000 concepts

Algorithms for XMTC

- One-vs-all: PDSparse, DiSMEC
- Target embedding: SLEEC, AnnexML
- Tree-based ensembles: FastXML, PFastreXML, Parabel, Bonsai, AttentionXML, CraftML, …
- Neural networks: fastText, XML-CNN, Bow-CNN, X-BERT, …

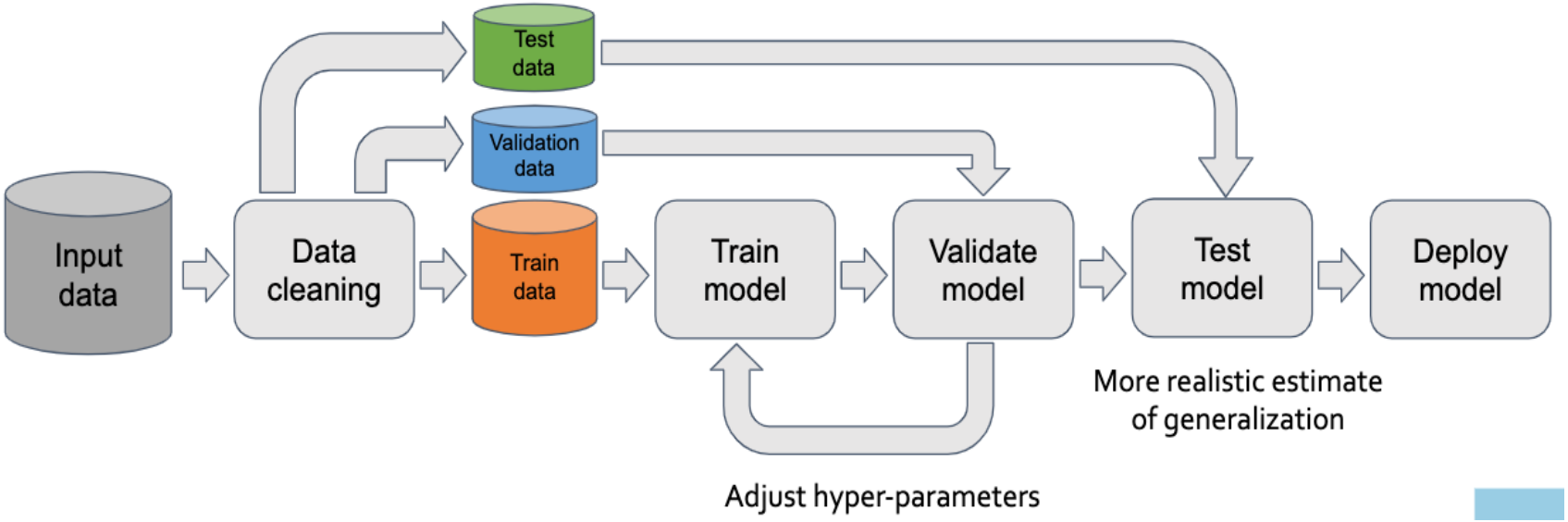






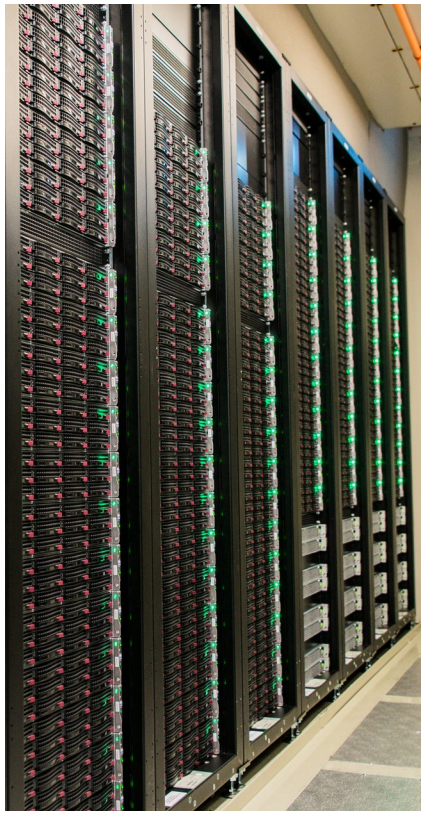
Hyper-parameter optimization

- Most of the algorithms have hyper-parameters that need to be tuned to get optimal results for different datasets
- Requires *lots of computational resources*

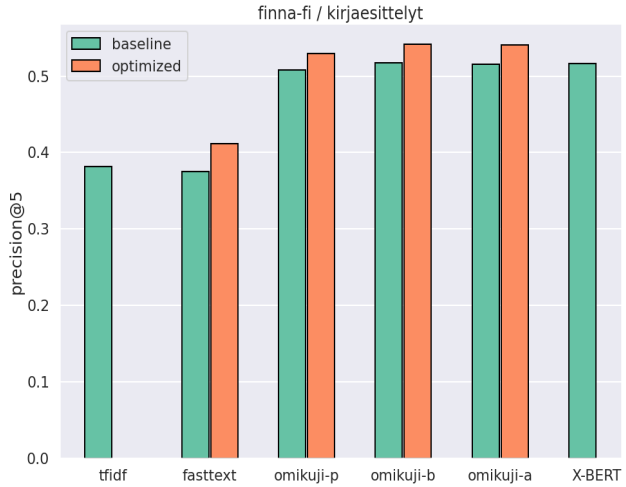


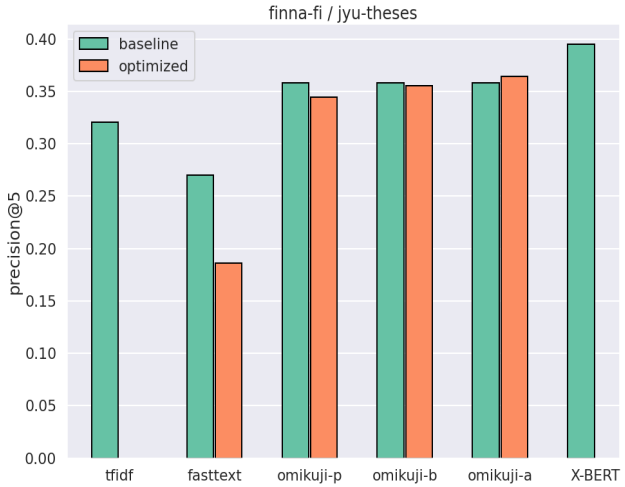
High-performance computing at CSC

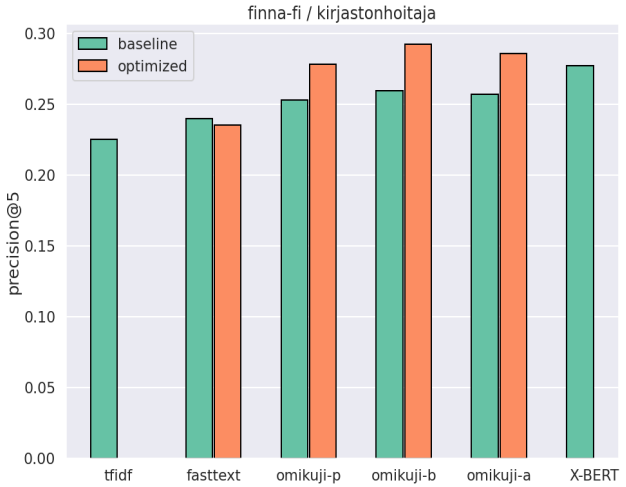
- **Puhti supercomputer**
  - Cluster of 682 computers with Intel CPUs
  - Peak performance of 1,8 petaflops
  - **Puhti AI**: 80 GPU-accelerated computers, each with 4 NVIDIA V100 GPUs
- **Mahti supercomputer**
  - Cluster of 1404 computers with AMD CPUs
  - Peak performance of 7,5 petaflops



Some results: Finnish (YSO)

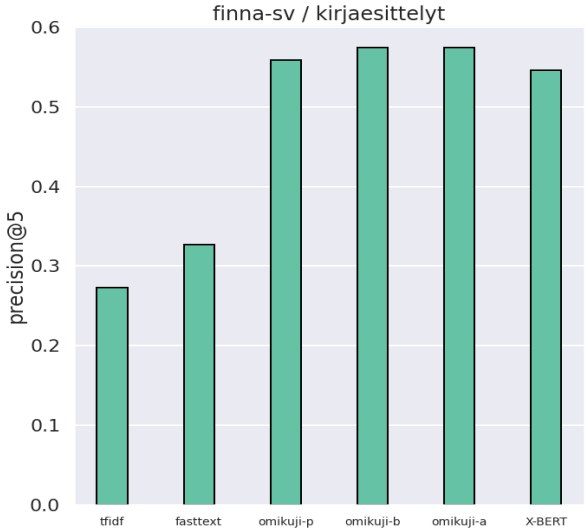


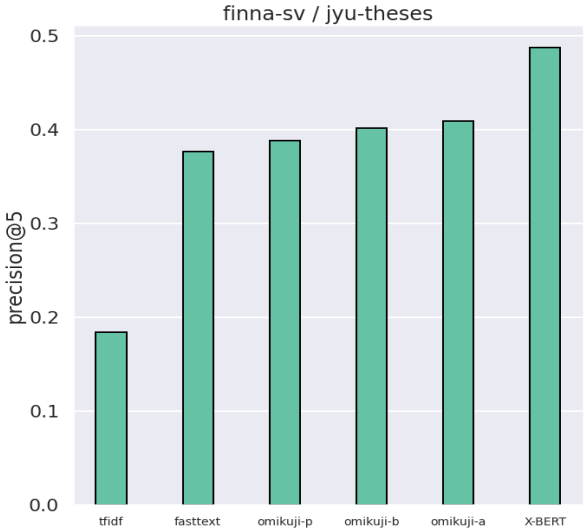


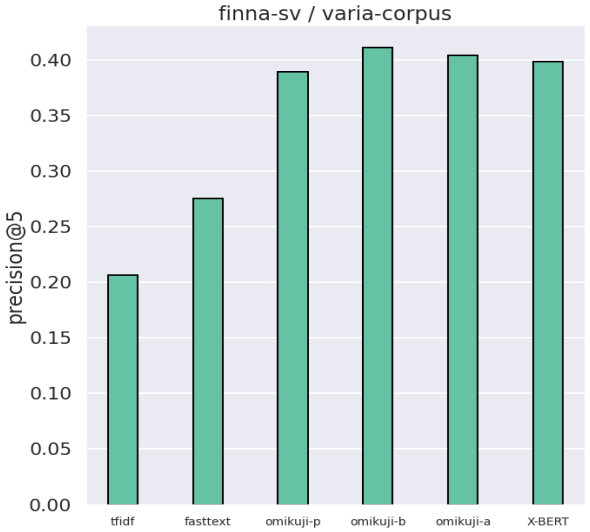


- Optimized = parameters which achieved best results *overall* in all datasets
- Tree-based algorithm (Omikuji) best, neural network-based (X-BERT) promising

Some results: Swedish (YSO)

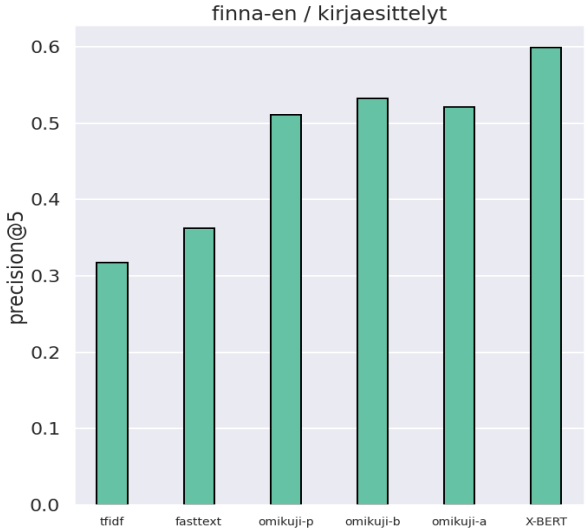


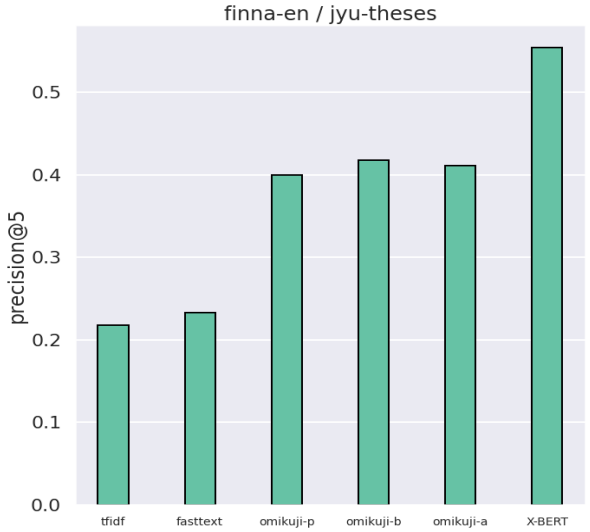


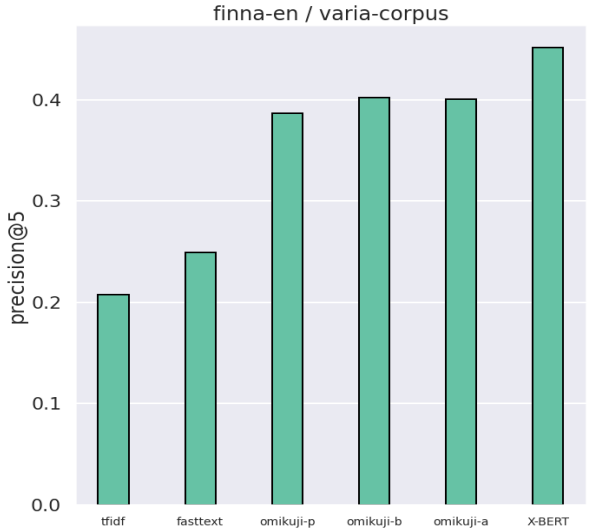


- No hyper-parameter optimisation performed for Swedish and English

Some results: English (YSO)



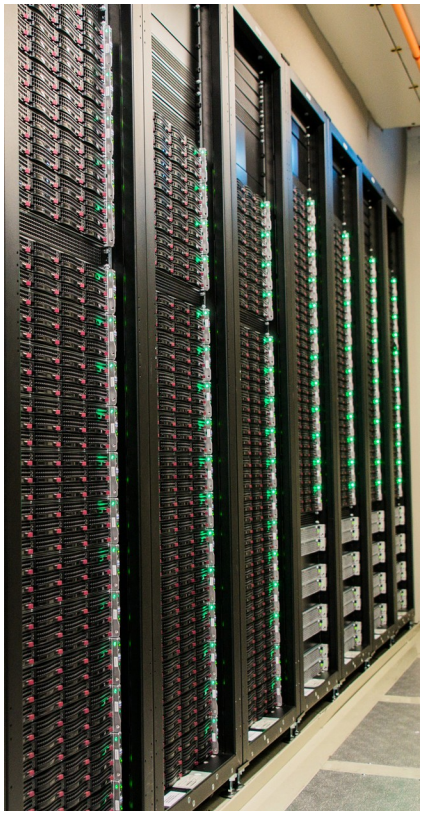




- Neural network-based X-BERT especially good for English (not surprising!)

Conclusions

- Annotation with library classification (YSO, YKL) is extreme multi-label classification (XMTC) problem
  - Lots of algorithms, lots of parameters
  - No problem: CSC has lots of computers!
- Better parameters, new promising algorithms found
  - Some have already been integrated into Annif



Mats Sjöberg

mats.sjoberg@csc.fi

http://staff.csc.fi/msjoberg/

facebook.com/CSCfi

twitter.com/CSCfi

linkedin.com/company/csc---it-center-for-science

github.com/CSCfi