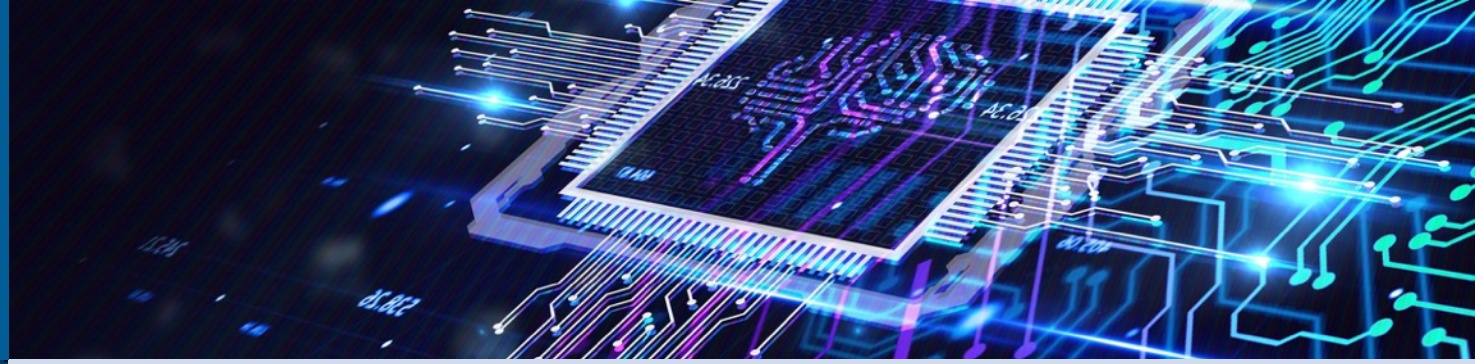
Today's agenda

12:00-12:20	Extreme multi-label text classification (XMTC)
12:20-12:40	Tree-based methods for XMTC
12:40-13:00	Neural networks for XMTC
13:00-	Other matters

CSC

ICT Solutions for Brilliant Minds

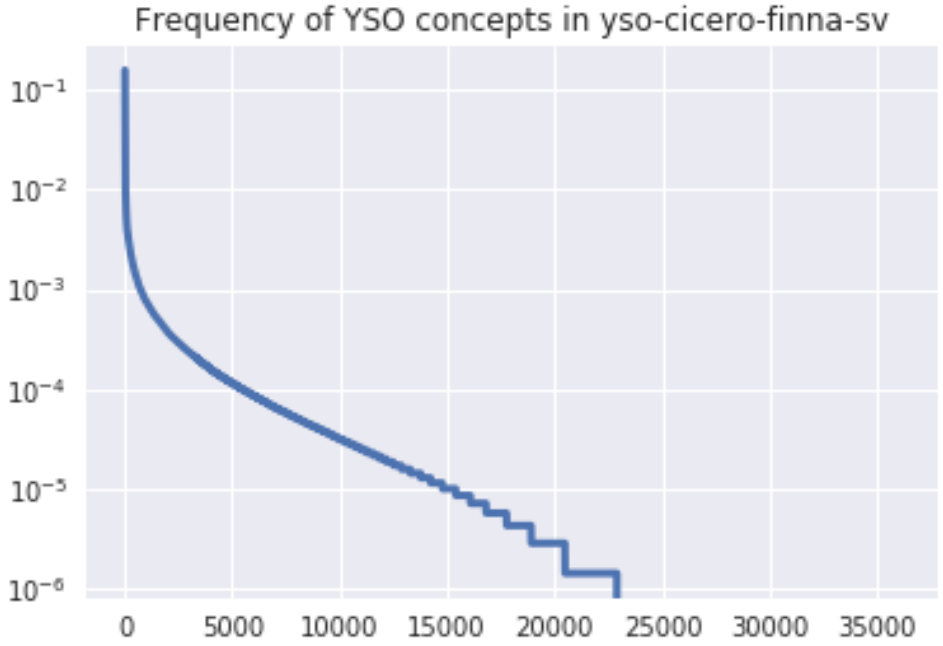


Extreme multi-label text classification (XMTC)

Markus Koskela

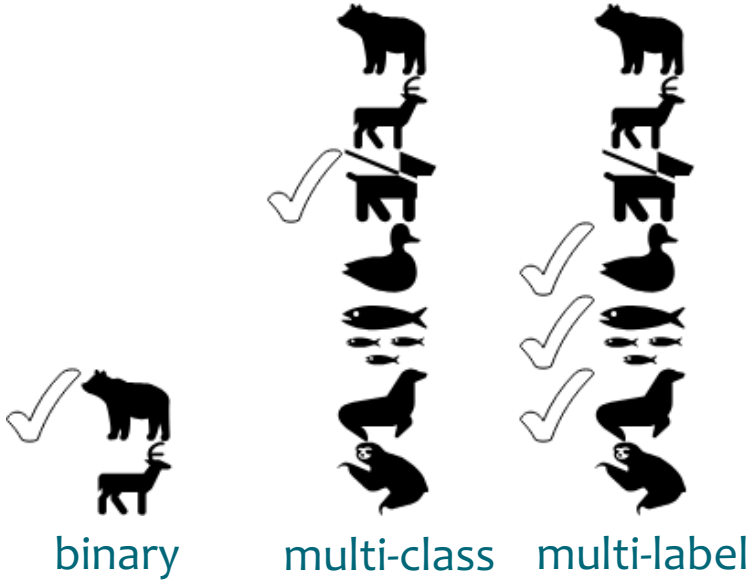
Text classification

- Annotating or classifying documents are hard problems for both humans and machines:
  - subjectivity
  - long tail phenomenon
  - vocabulary updates



Multi-label text classification

- Finding each document the most relevant subset of labels instead of a single correct class



Extreme multi-label text classification

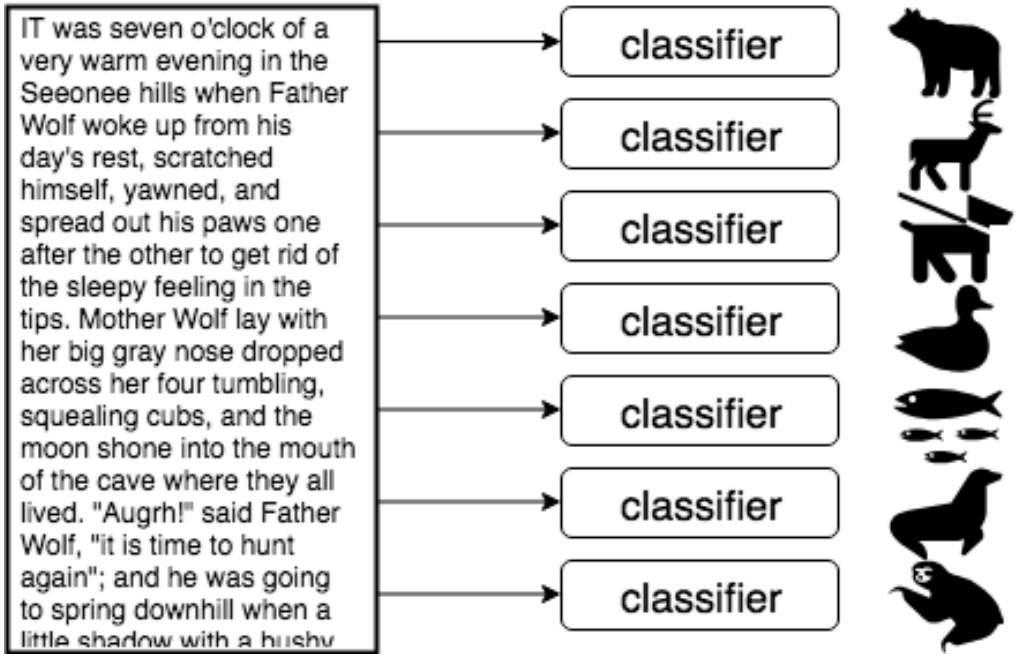
- The number of training examples, the dimensionality of data, and especially **the number of labels** are large
- Issues:
  - sparsity
  - label correlations
  - scalability
  - computational cost
- E.g. YSO contains 31050 concepts

Main approaches for XMTC

One-vs-all

- Learn a separate classifier for each label
- logistic regression, linear SVM
- PDSparse, DiSMEC, …

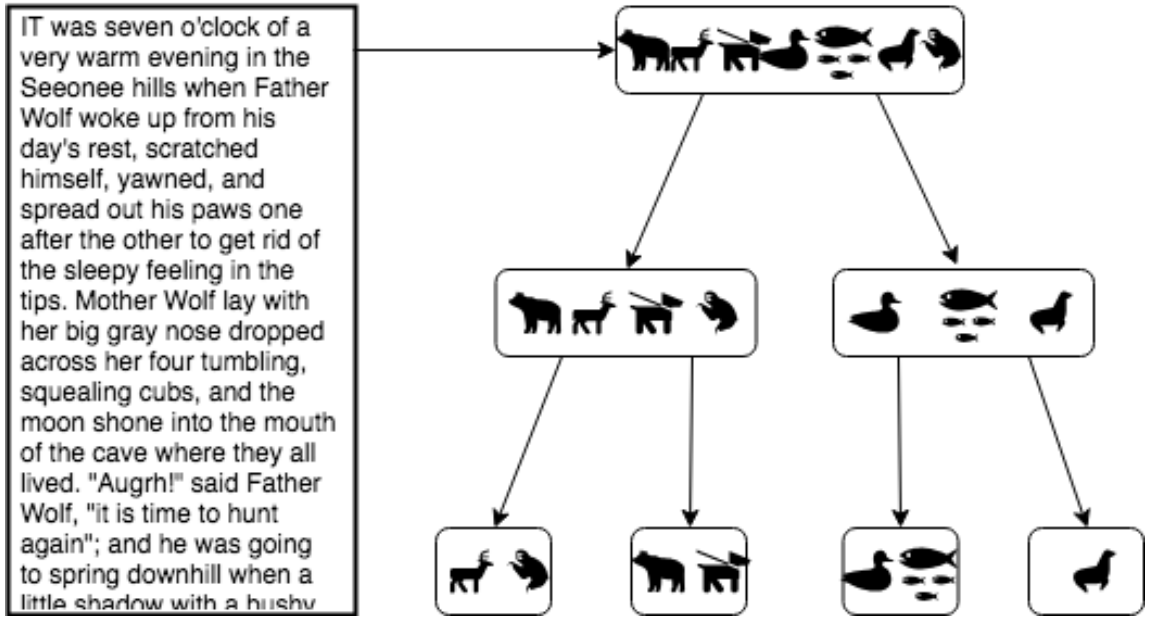
Slow and does not utilize similarities between classes. Still needed as the final step in most approaches.



Tree-based ensembles

- Recursively divide the space of labels or features **based on data**
- FastXML, PFastreXML, Parabel, Bonsai, SwiftXML, CraftML, …

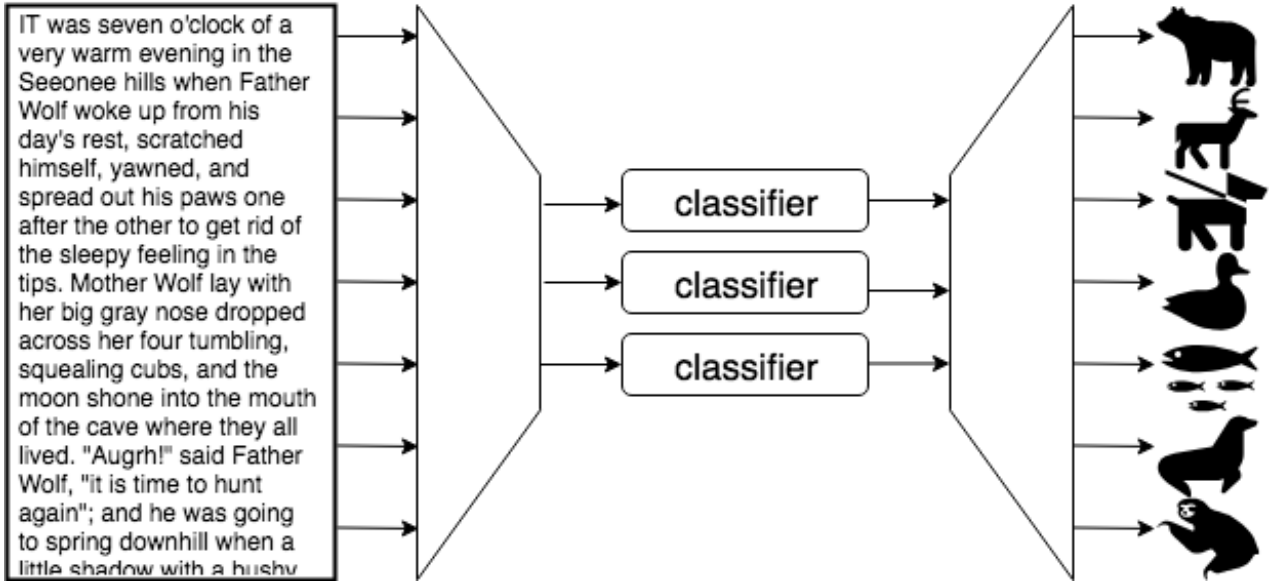
Good results, fast, but models can be large (GBs).



Target embedding

- Exploit label correlations and sparsity to compress the label space
- SLEEC, AnnexML

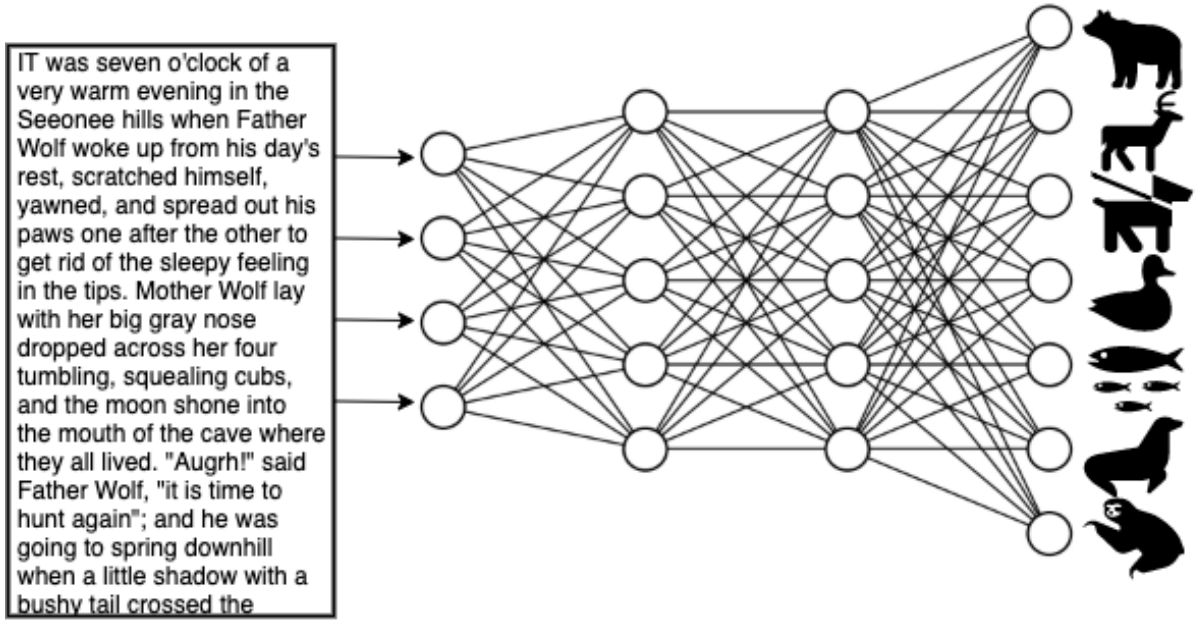
Work, but usually tree-based methods perform better.



Neural networks

- Extract context-sensitive features from raw text using deep learning
- fastText, XML-CNN, Bow-CNN, AttentionXML, X-Transformer, …

Computationally heavy, but some promising results and interesting properties.



High-Performance Digitisation

Project idea

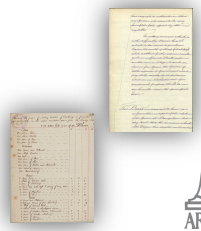
- Use of digital resources is hampered by insufficient search functions and findability due to deficits in data quality and lacking metadata
- Objective is to create an intelligent annotation pipeline for enriching archived material, such as scanned newspapers, journals, books, images, and official documents
- Runs in CSC's environment and uses GPU accelerated machine learning for computer vision and natural language processing
- Result is a service that will be in production use in CSC's cloud computing platform and offered to memory organizations

About the project

- INEA / CEF Telecom project (CEF-TC-2017-3 - Public Open Data)
- From 9/2018 to 8/2020 (2 years), extended to 12/2020
- Total budget of 360 000 euros
- CSC is the only applicant; The National Library of Finland and the National Archives of Finland listed as collaborators
- Website at https://www.csc.fi/en/-/high-performance-digitisation

INEA
The Innovation and Networks Executive Agency

THE NATIONAL LIBRARY OF FINLAND



ARCHIVVM FINLANDIAE PVBLICVM

THE NATIONAL ARCHIVES OF FINLAND

OCR, handwritten text recognition, image layout

High-performance digitisation

- document annotation
- image classification
- topic segmentation
- integration

Open source service for memory organizations hosted in CSC cloud

Markus Koskela

markus.koskela@csc.fi

facebook.com/CSCfi

twitter.com/CSCfi

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

github.com/CSCfi