Mapping the Internet

Modelling Entity Interactions in Complex Heterogeneous Networks

Ing. Šimon Mandlík, Supervisor: doc. Ing. Tomáš Pevný, Ph.D

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Motivation

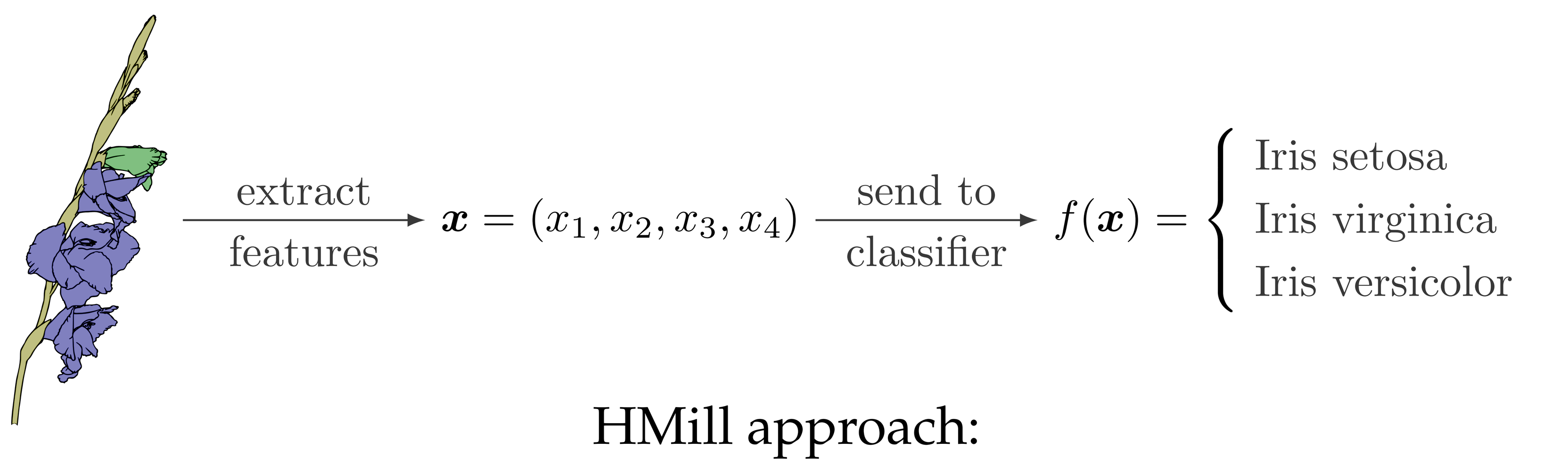
In many domains, application of standard machine learning methods on modern data sources is still hindered due to:

- Unrealistic **assumption about independence** and identical distribution of input data
- Unknown set of informative **features** to represent samples
- **Heterogeneous** or **hierarchical** nature of the data
- **Missing data** on various levels of abstraction
- Insufficient **scalability**
- Unsatisfactory **explainability** and **interpretability**

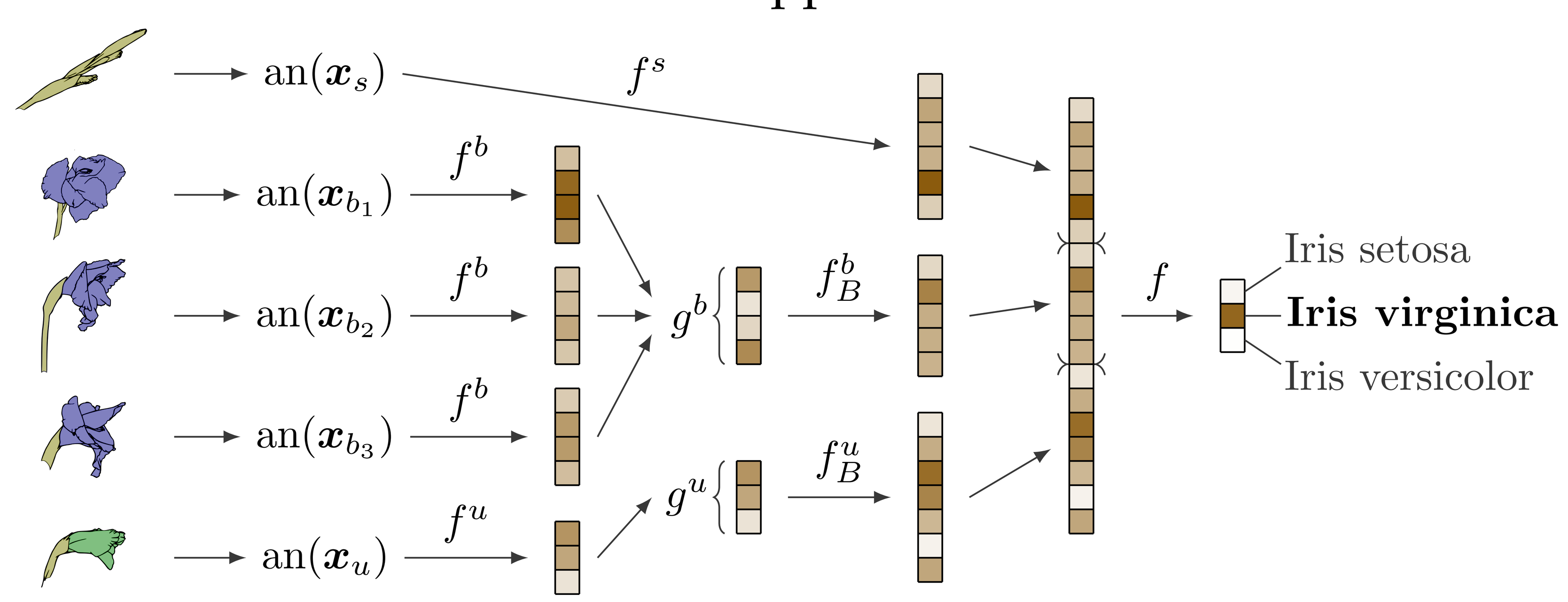
HMill framework

- Hierarchical Multi-instance Learning Library
- general-purpose, unified **framework** for sample representation and model definition
- high modelling **flexibility** and overall **versatility**

Classical machine-learning approach:



HMill approach:

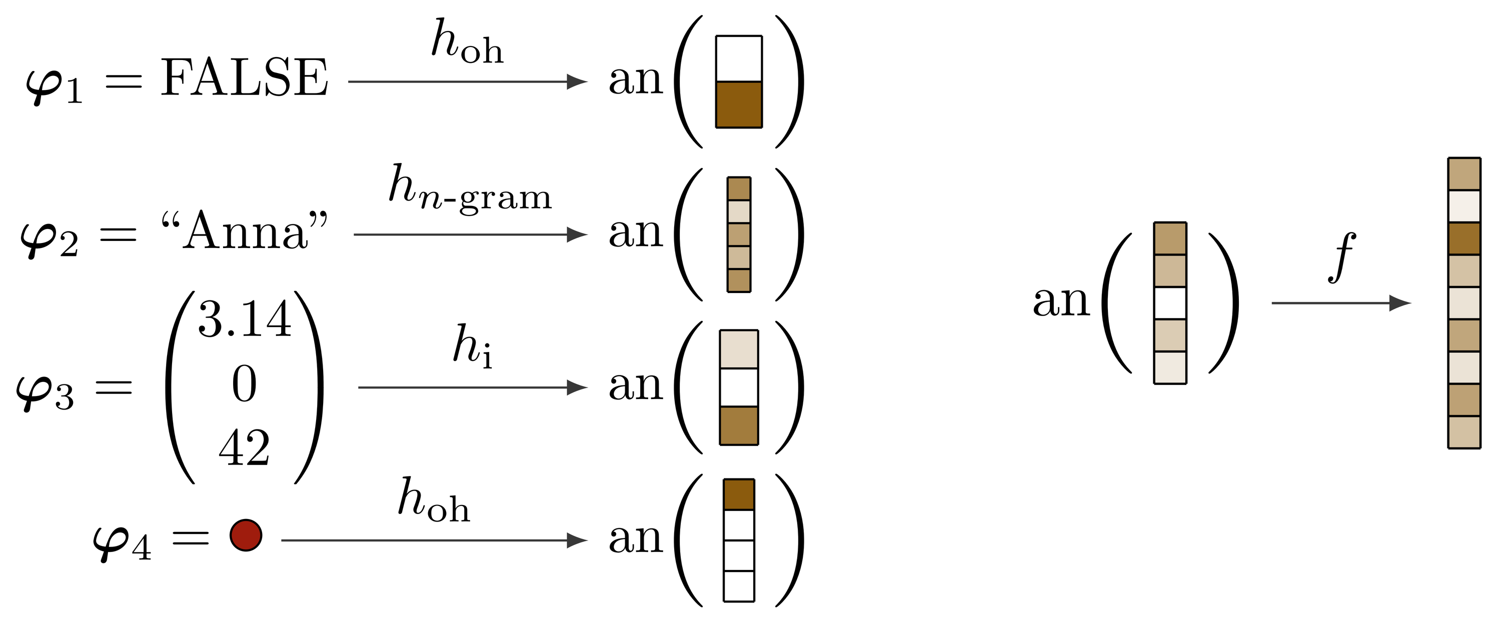


HMill components

- tree-based sample and model representations
- each layer handles a different level of abstraction

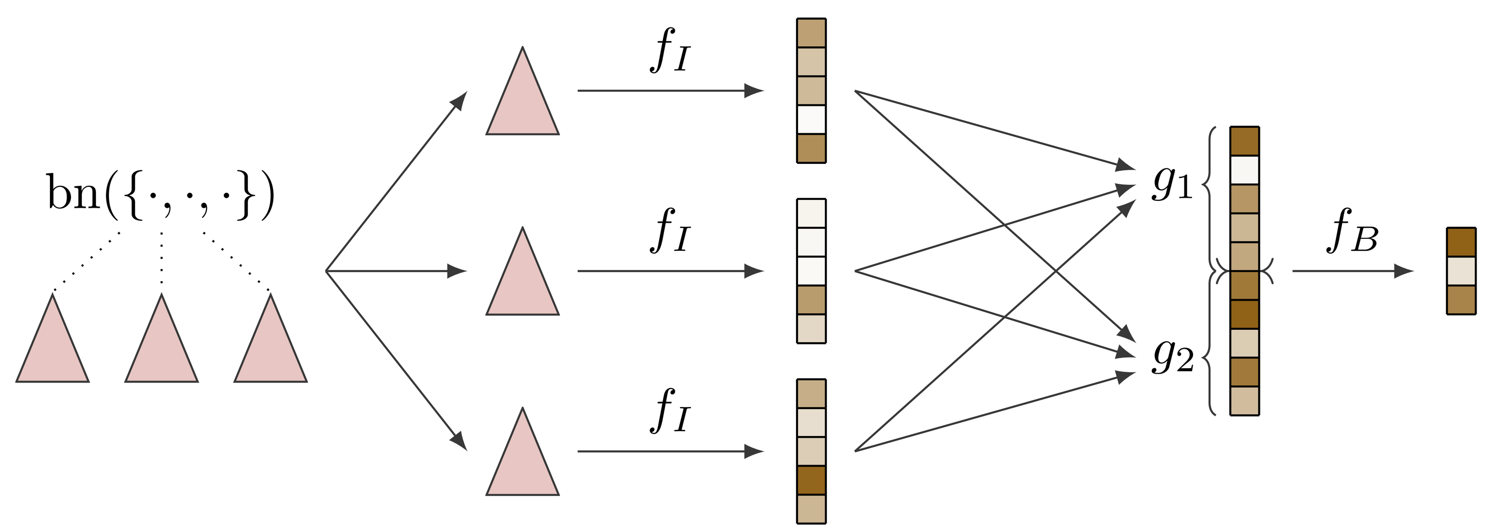
Array nodes

For modelling lowest-level raw observations



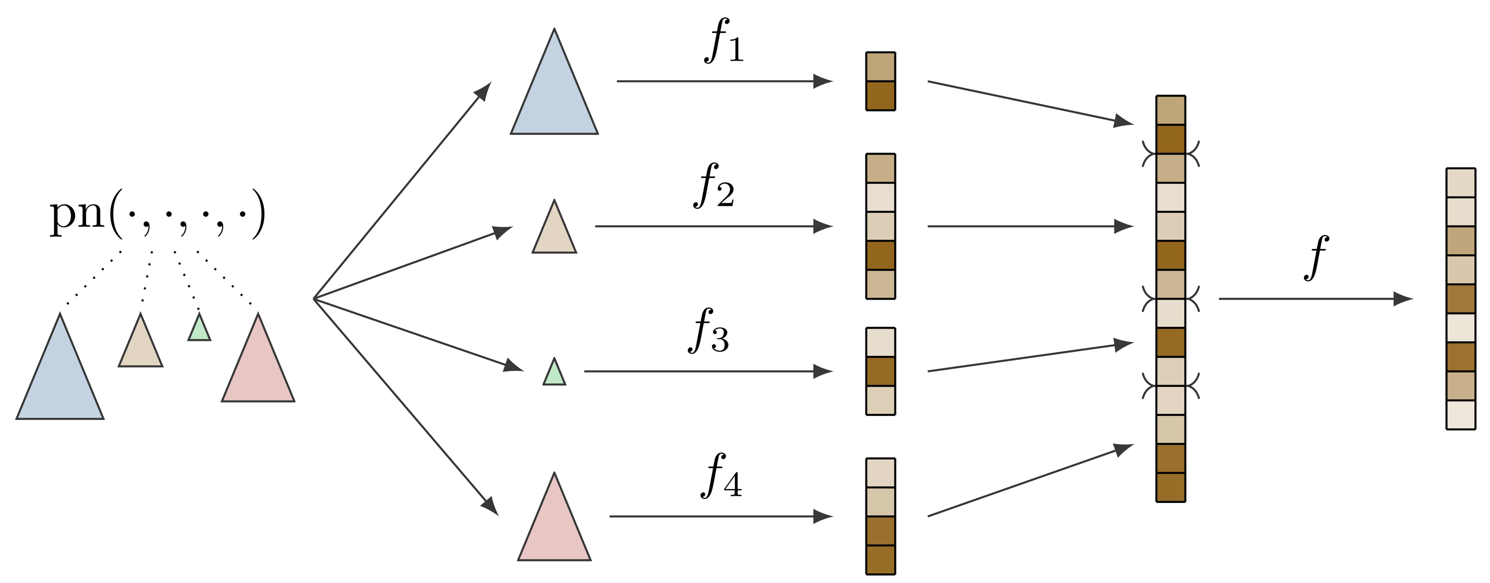
Bag nodes

For modelling compact sets of probability measures



Product nodes

For modelling Cartesian products



HMill traits

- theoretically **justified** (extension of the UA theorem)
- efficient **batching** and **gradient computation**
- elegant dealing **missing data**
- convenient **sampling** techniques for large inputs

Real-world use cases

- framework tested on **three completely different tasks**
- **cybersecurity domain** — very relevant and difficult for ML
- baseline models achieved **comparable or better performance** than specialized methods on all three tasks

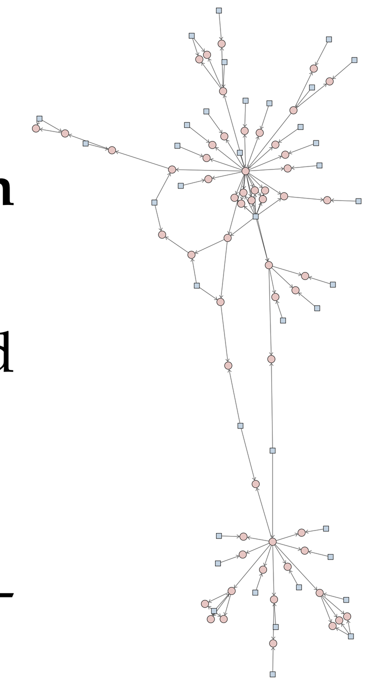
Use case: Classifying IoT device over network

- classifying **the type of IoT device**
- based on measurements obtainable by network scanning
- structured, hierarchical and heterogeneous data
- some items are missing
- input: **JSON/XML documents**
- Avast data
- HMill performs better on the provided dataset than a specialized method


```
{
  "mac": "80:5e:c0:41:ad:39",
  "ip": "192.168.0.147",
  "services": [
    {
      "port": 80,
      "protocol": "tcp"
    },
    {
      "port": 5353,
      "protocol": "udp"
    }
  ],
  "device_class": "IP_PHONE",
}
```

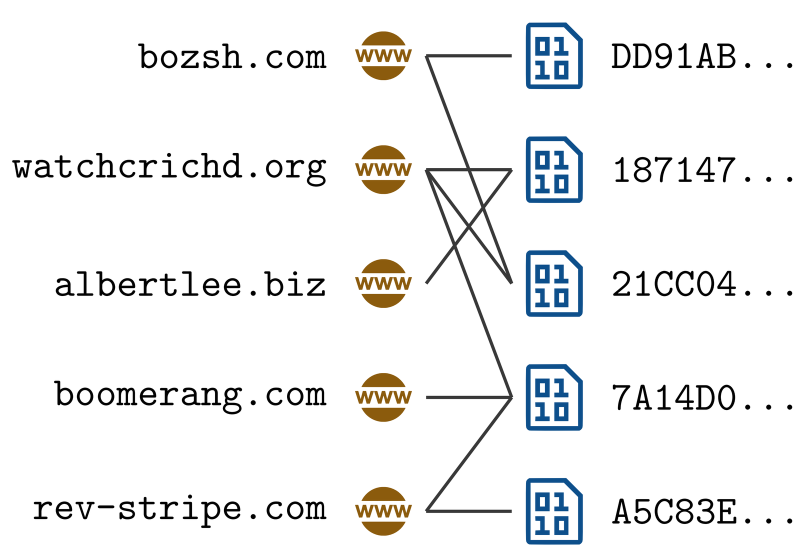

Use case: Detecting malware with behavioral graphs

- detecting **malicious binary files**
- based on the behavior in Windows OS
- input: snapshot of the OS represented **as a graph**
- nodes represent files and processes
- edges represent interaction between files and processes
- data obtained from Avast
- HMill more accurate than methods ignoring relations



Use case: Harmful domain detection from relations

- detecting **harmful domains**
- input: **binary relations**
- example: domain D in relation with binary B, because B connected to D
- Cisco cooperation
- HMill performs comparably to state-of-the-art



Conclusion

- HMill offers **high versatility** with **no performance compromises**
- excels at automated, **Auto ML style approach to learning** from real-world data
- out-of-the-box availability and little to no preprocessing needed enable **application to many problems**
- implementation available at https://github.com/pevnak/Mill.jl