STATISTICS DENMARK

Classification of the Purchasing Power Parities Nomenclature

22/5 2025

AIML4OS WP10 cluster 3 w. Austria
Thank you ---- and ----!

Outline

STATISTICS DENMARK

1. Case: What Official Statistics process are we talking about?

2. Highlights from the project history
   - What has been done?
   - Results

3. Lessons and Perspectives
   - Suggestions (are welcome!!!)
   - Questions

4. What is hierarchy and is it useful?

Raise hands or make a noise/comment

STATISTICS DENMARK

- Who of you have heard or know a little about
  - COICOP?
  - Hierarchical classification?

STATISTICS DENMARK

Case

Case: Purchasing Power Parities (PPP)

STATISTICS DENMARK

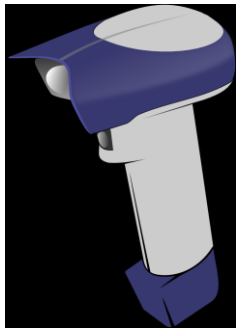
- To ensure that Eurostat enables international comparisons
  a) in real values
  b) of price levels

- The business process of Official Statistics:

  Eurostat ← PPP-calculations ← PPP-classification ← data…

Case: Classify scanner data (1)

- Major retailers provide to Statistics Denmark the ”scanner-data”



Placeholder – data removed

Case: Classify scanner data (2)

STATISTICS DENMARK

- Eurostat's triannual .XML
  - Defines a hierarchical product taxonomy
  - PPP = COICOP + 3 levels

A.01.1.7.5.01.bb Potato crisps, BL (UID: 7428)
3

Brand:	---
Reference Quantity:	**200 g**
Minimum quantity:	150
Maximum quantity:	350
Exclude:	multipack; fried in olive oil

SPECIFY: **Label |**

A.01.1.7.5.01.ba Potato crisps, WKB (UID: 2101)
2

Brand:	**Well Known**
Reference Quantity:	**200 g**
Included in ICP global list:	yes
Minimum quantity:	130
Maximum quantity:	300
Type:	salted or flavoured
Exclude:	PRINGLES; multipack; exotic flavours; fried in olive oil

SPECIFY: **Brand |**

Case: Classify scanner data (3)

STATISTICS DENMARK

- Classification currently done manually
  - ”The boys” look up information online to discriminate classes
  - Provides a dataset of scanner-data with the classes i.e. supervised

Case: Classify scanner data (4)

STATISTICS DENMARK

- Analytical situation
  - ~400 theoretical classes at level 1.2.3.4.5.6.7
  - ~15.000 skewed/unbalanced observations of the classes
  - *Is it possible to gain leverage over the classification problem by use of the class hierarchy?*

STATISTICS DENMARK

Project History

Highlights from the Project History

STATISTICS DENMARK

- 2010-20: Statistics DK increasingly systematizes scanner data as retailers returns more of it

- 2022: --------- develops the project now discussed in AIML4OS
  - Experimenting with text prep. & bag of words; then different embeddings
  - PyTorch NN BERT finetuning gives rise to ~70% classification accuracy (at L2), but…
  - Accuracy varies substantially by the frequency of the class observations

- 2024, second half: An intern reviews --------- work

- 2025, spring: ------ are given a go at the project
  - Understand and get some of Peter's code to run
  - FastText gives rise to ~90% classification accuracy, but…

STATISTICS DENMARK

Lessons and Perspectives

Lessons We Believe We Learned

STATISTICS DENMARK

- Perhaps more about project and code organization than ML/AI
  - Describe the business process of Official Statistics
  - Make realistic assumptions about what data will be available → modelling
  - Make it clear to readers of the code what is a preliminary experiment, what is a final solution, etc.

Perspectives (1)

STATISTICS DENMARK

- Will _____ wreck havoc with learned classifications?
  - Items not yet observed
    - Low accuracy demands manual labor
  - Triannual PPP.XML updates
    - make accurate predictions at t_0 incorrect at t_1
- Model differences in the .XML?

Perspectives (2)

STATISTICS DENMARK

- Estimate distance/similarity, not classification probability?
  - Few-shot learning or Siamese NN
  - .XML as ”support set”
  - Scanner data as ”query set”
  - Positive and negative pairs and perhaps even an anchor
    - Negative pairs can be selected using information about the hierarchical structure

➔ *Is it possible to gain leverage over the ~~classification~~ comparison problem by use of the class hierarchy definitions?*

Perspectives (3)

STATISTICS DENMARK

- Predict longer product descriptions from shorter ones?

STATISTICS DENMARK

Suggestions and Questions

STATISTICS DENMARK

Hiercarchical classes

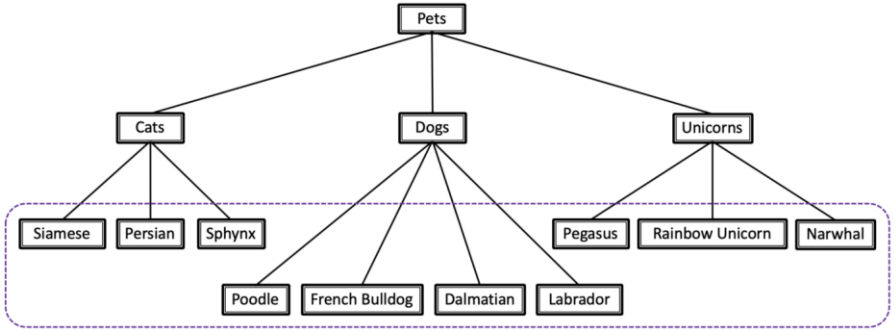
STATISTICS DENMARK

Screenshot of some results has been removed

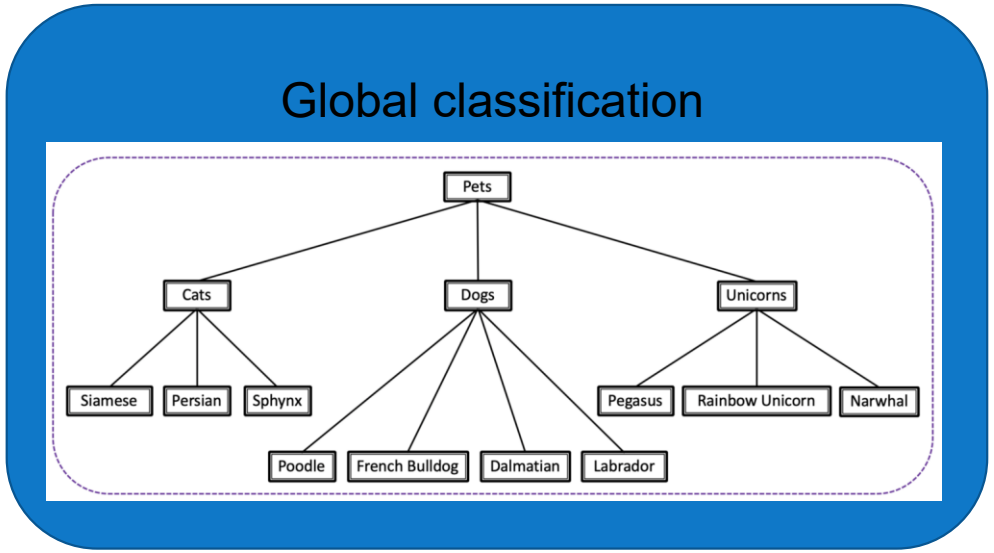
What is Hierarchy?

STATISTICS DENMARK

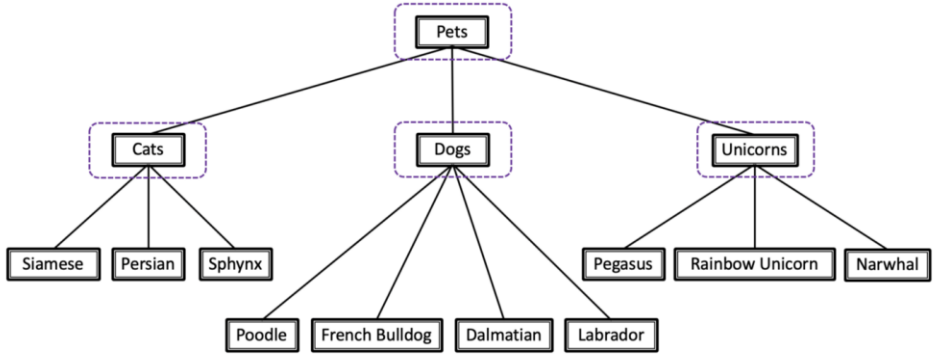
Flat classification



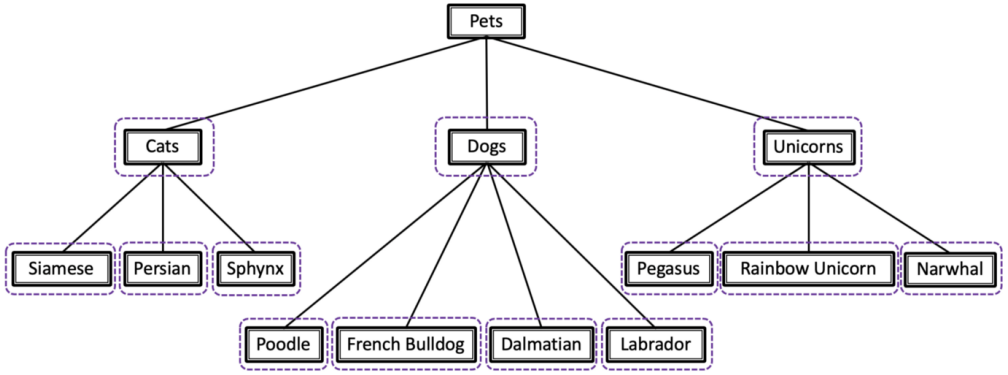
Global classification



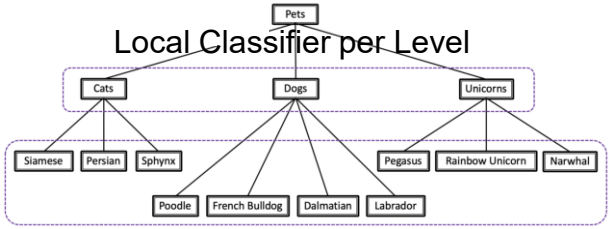
Local Classifier per Parrent Node



Local Classifier per Node



Local Classifier per Level



Needn’t Hierarchy?

STATISTICS DENMARK

- Experience with data at DE-stasis and DST suggests not