Model analytics for feature models: case studies for S.P.L.O.T. repository

16.10.2018, AMMoRe @ MODELS'18

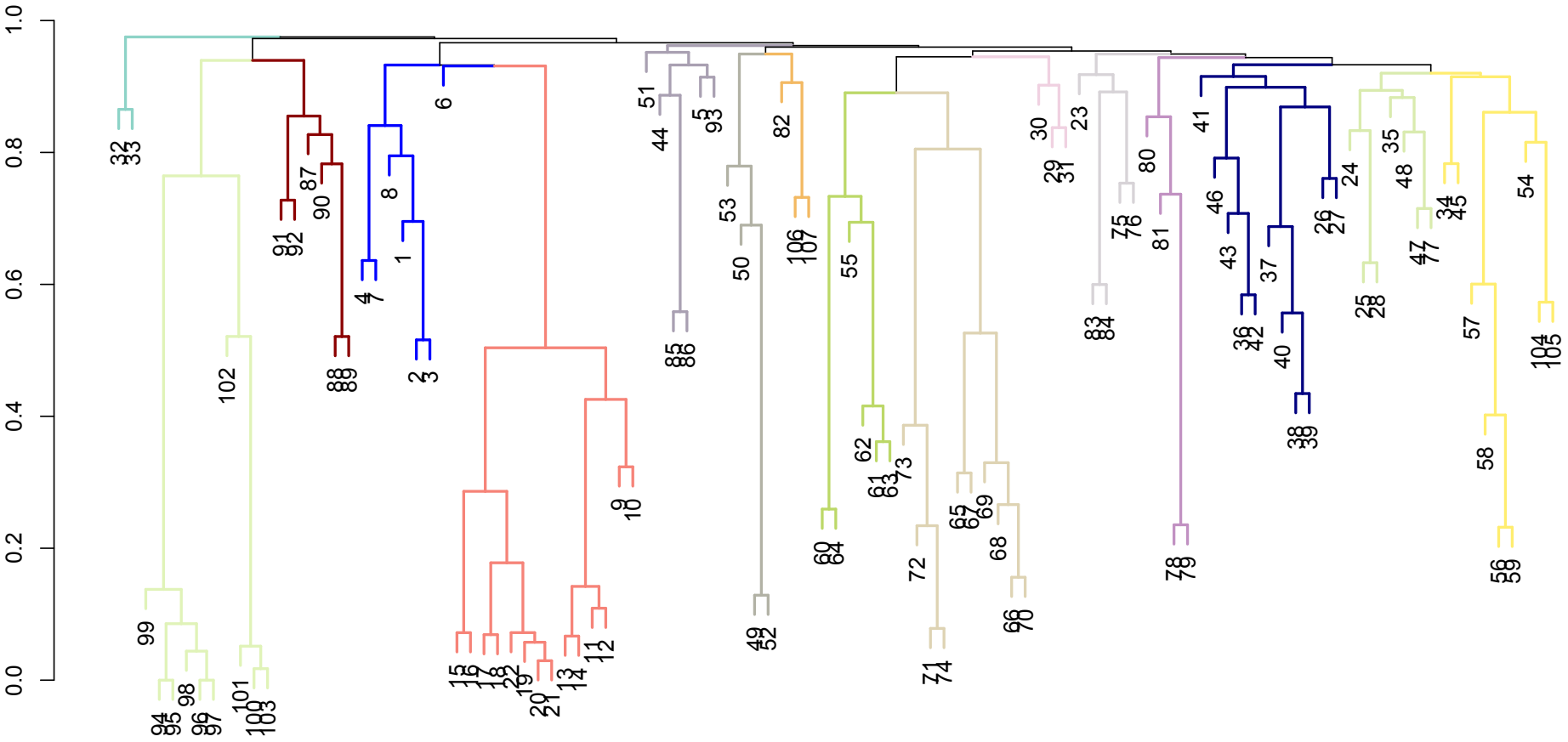
Önder Babur

w/ Loek Cleophas, Mark van den Brand

TU/e Technische Universiteit Eindhoven University of Technology

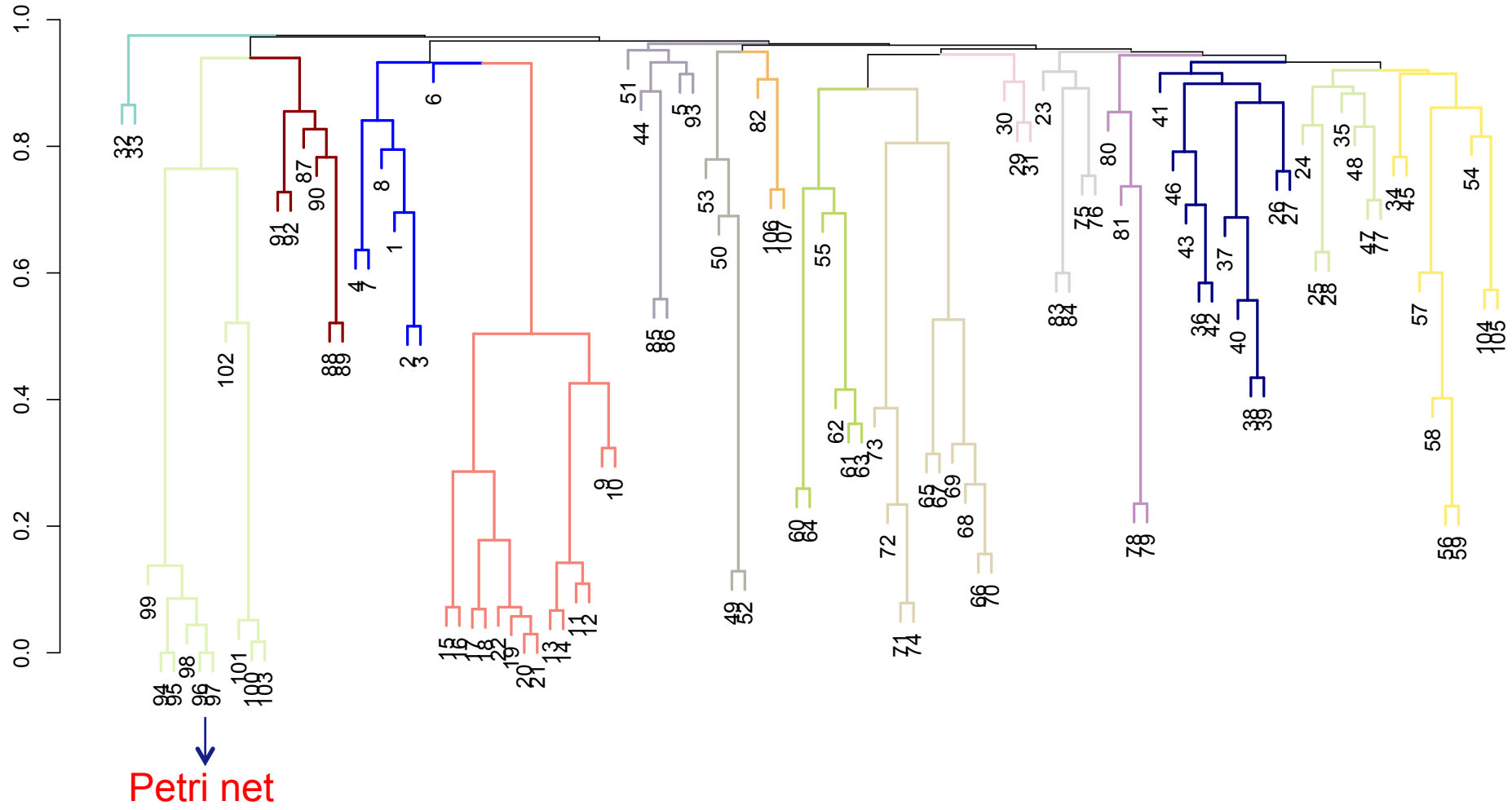
Where innovation starts

Example*: ATL Metamodel Zoo



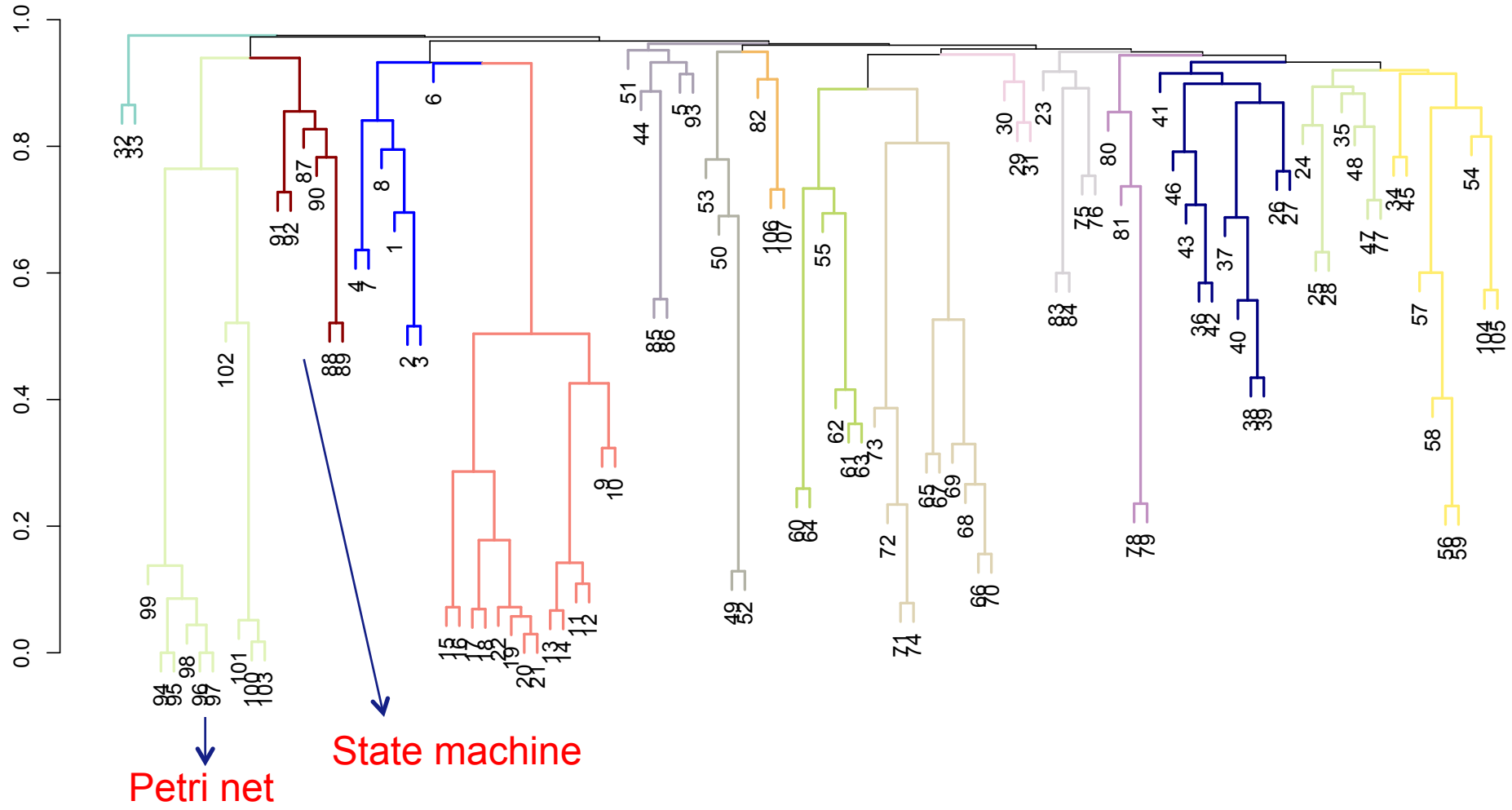
**Babur et al, Hierarchical Clustering of Metamodels for Comparative Analysis and Visualization, ECMFA '16*

Example*: ATL Metamodel Zoo



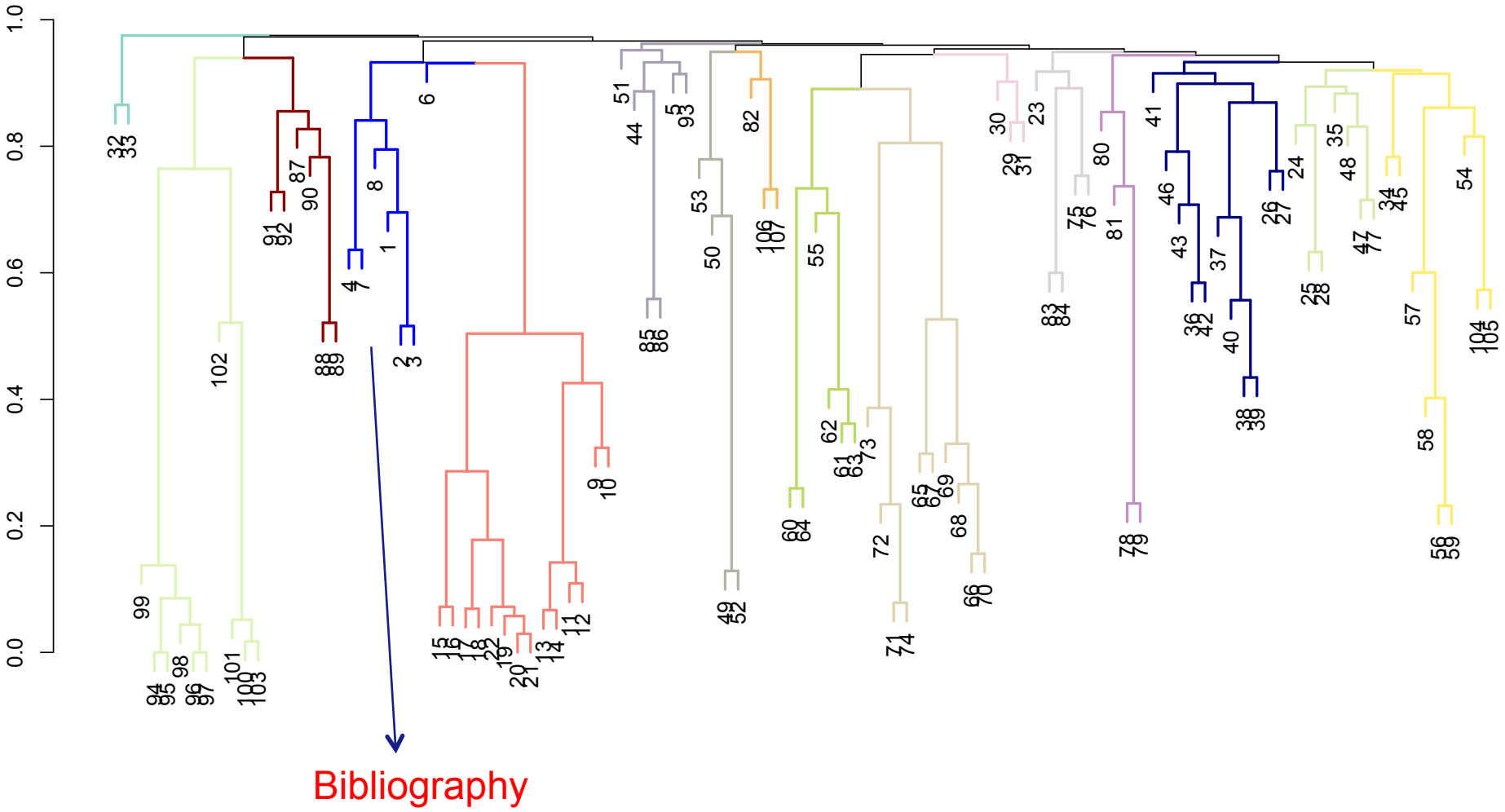
**Babur et al, Hierarchical Clustering of Metamodels for Comparative Analysis and Visualization, ECMFA '16*

Example*: ATL Metamodel Zoo



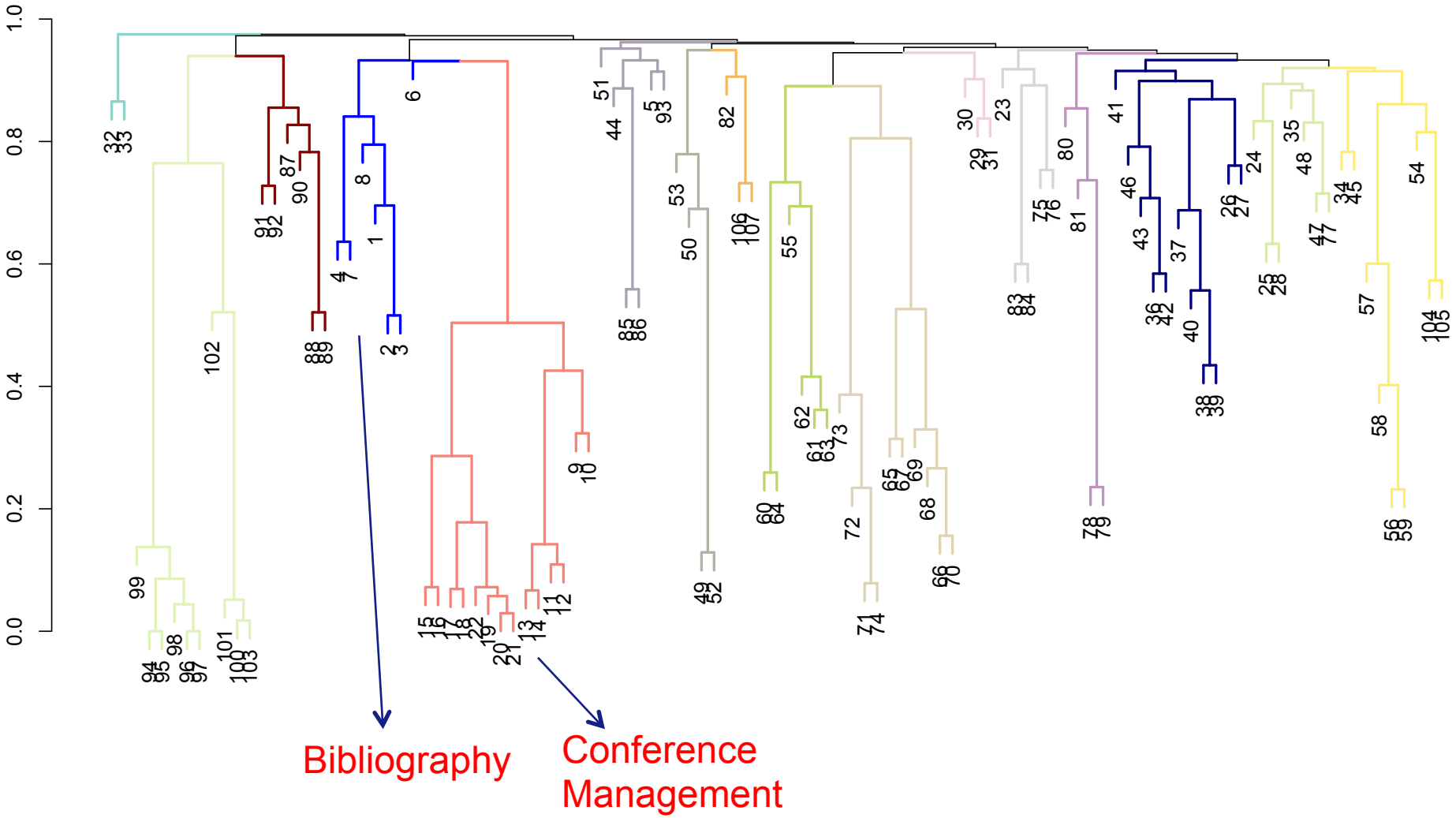
**Babur et al, Hierarchical Clustering of Metamodels for Comparative Analysis and Visualization, ECMFA '16*

Example*: ATL Metamodel Zoo



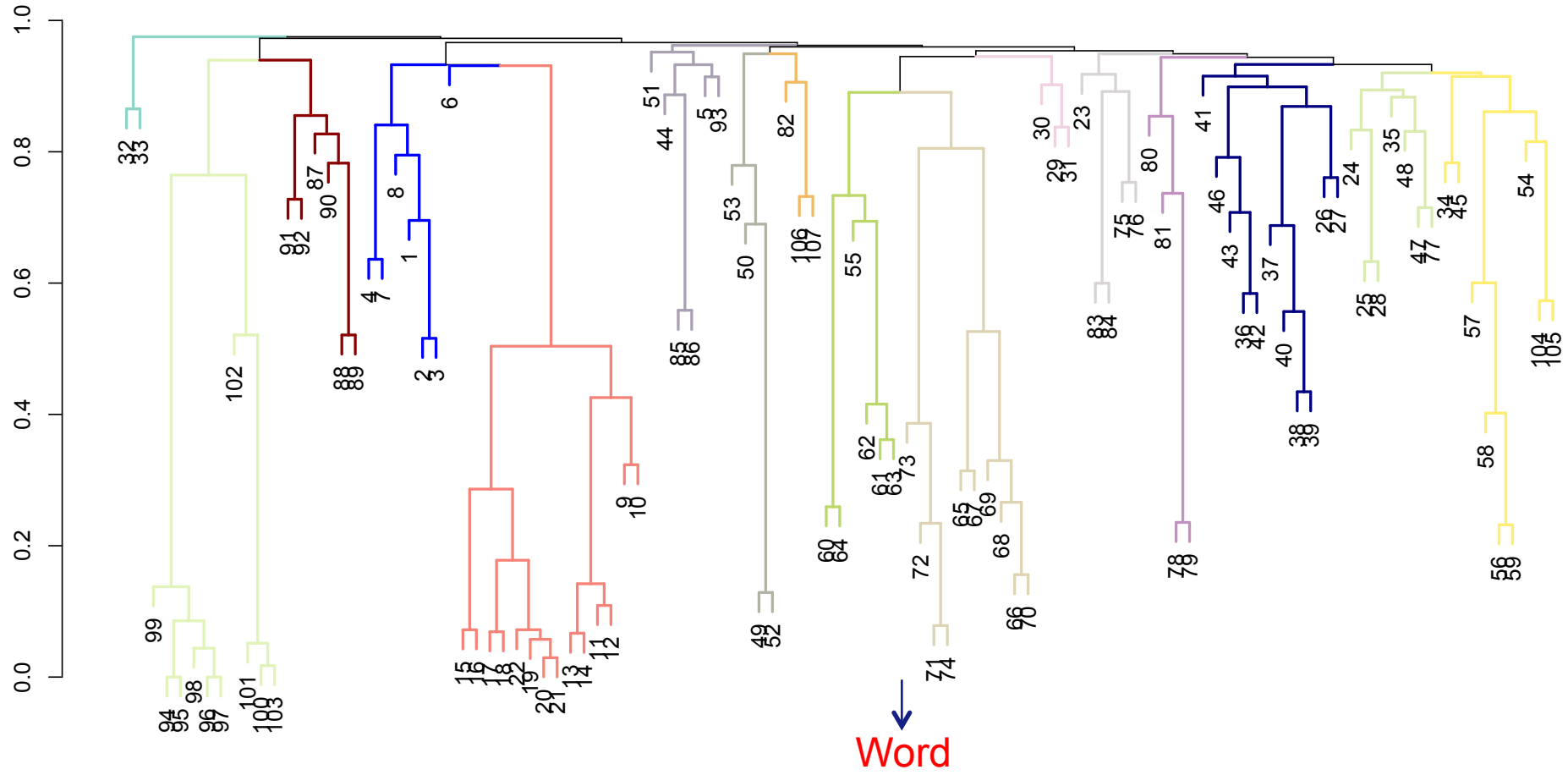
*Babur et al, Hierarchical Clustering of Metamodels for Comparative Analysis and Visualization, ECMFA '16

Example*: ATL Metamodel Zoo



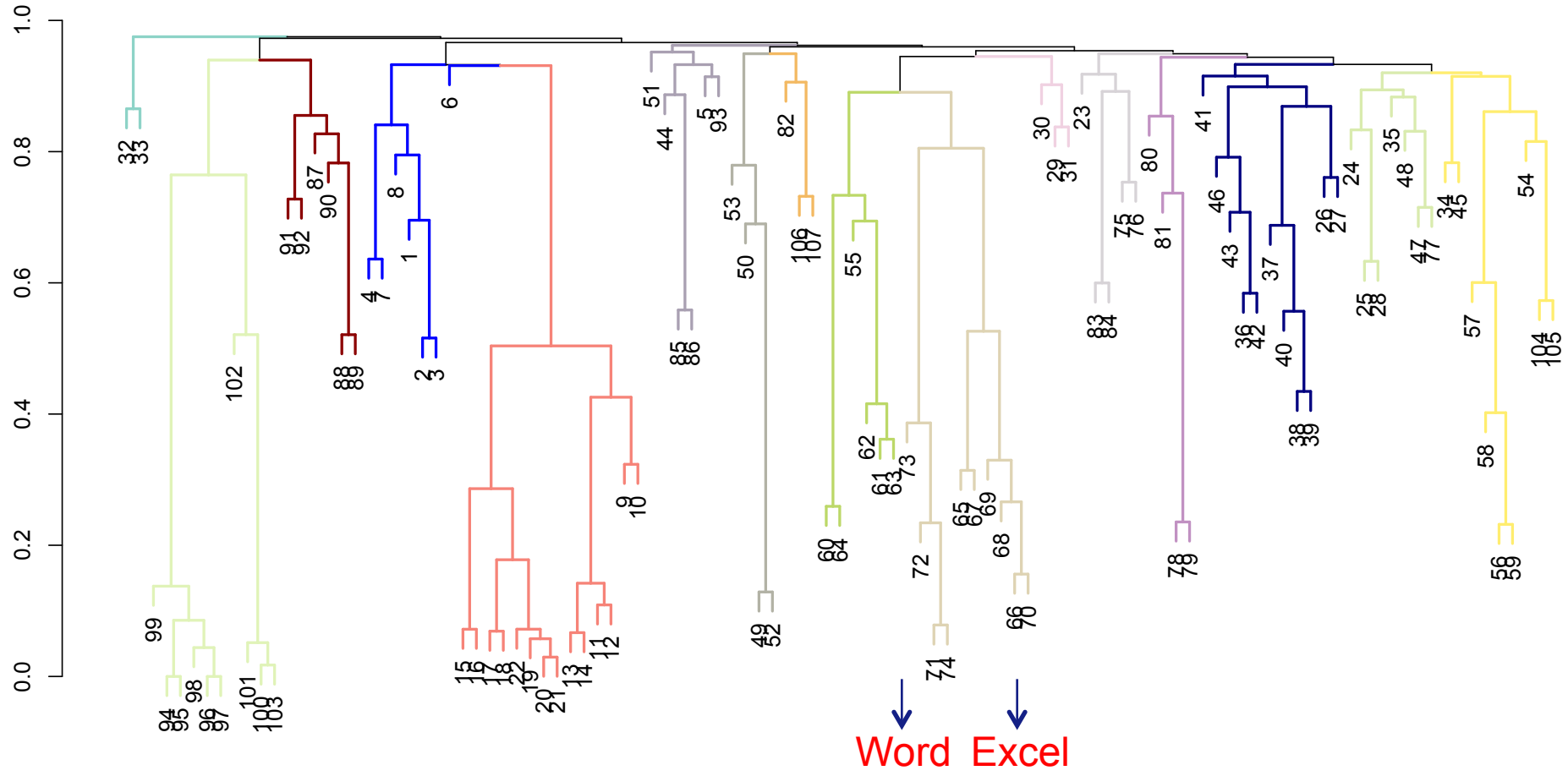
Babur et al, Hierarchical Clustering of Metamodels for Comparative Analysis and Visualization, ECMFA '16

Example*: ATL Metamodel Zoo



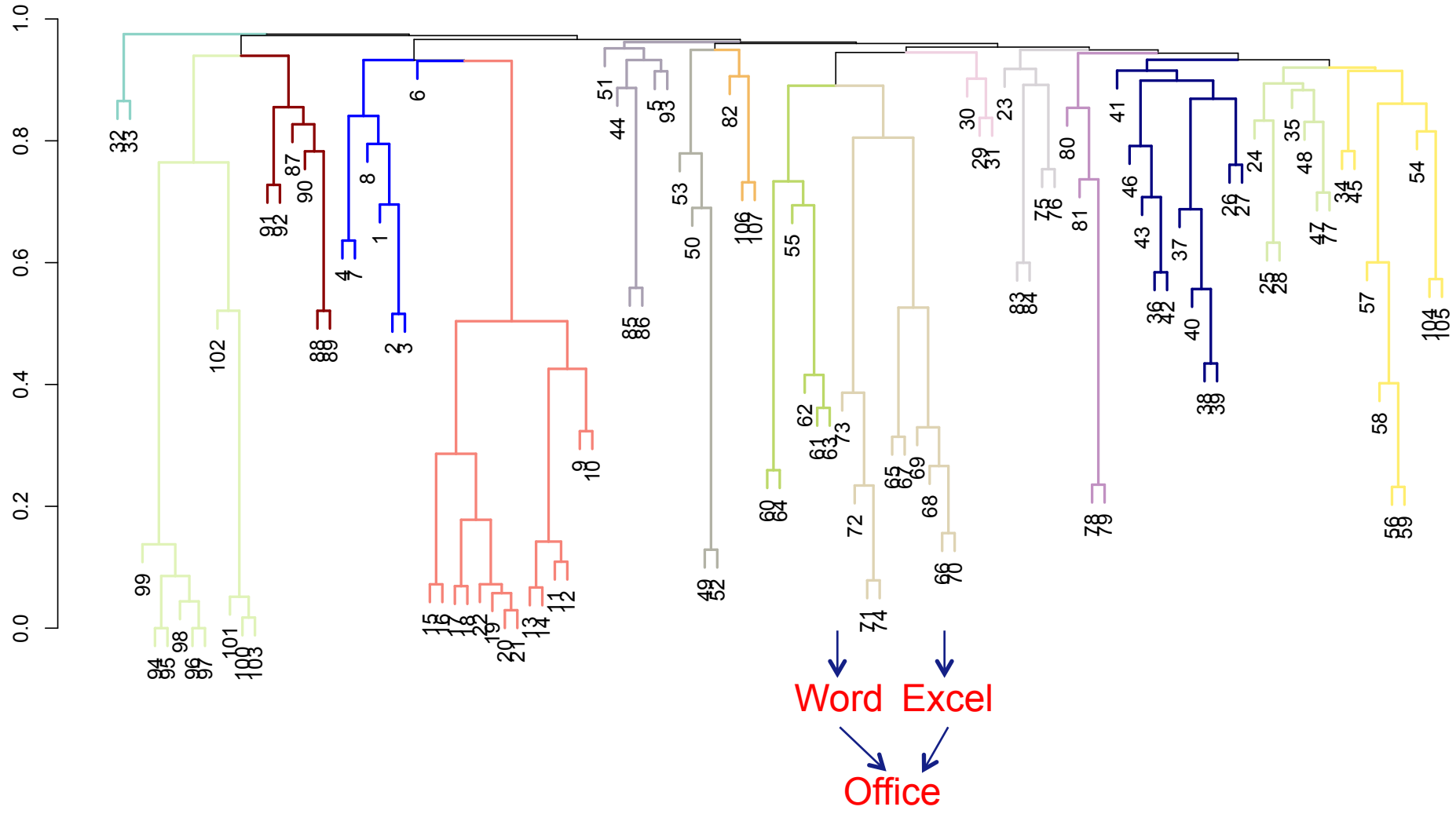
**Babur et al, Hierarchical Clustering of Metamodels for Comparative Analysis and Visualization, ECMFA '16*

Example*: ATL Metamodel Zoo



**Babur et al, Hierarchical Clustering of Metamodels for Comparative Analysis and Visualization, ECMFA '16*

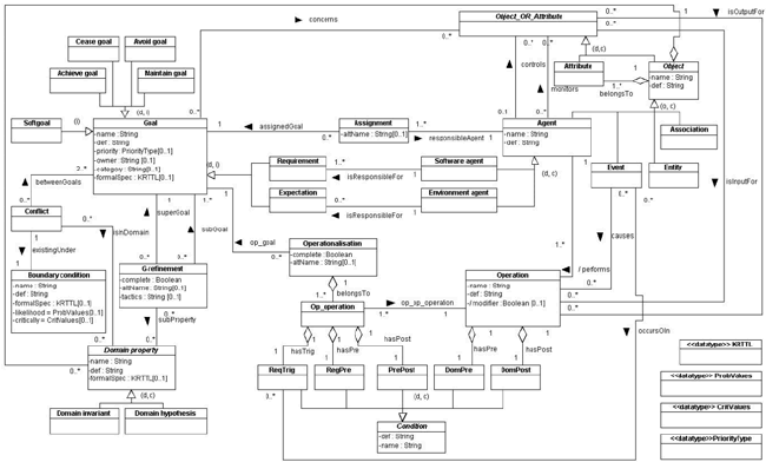
Example*: ATL Metamodel Zoo



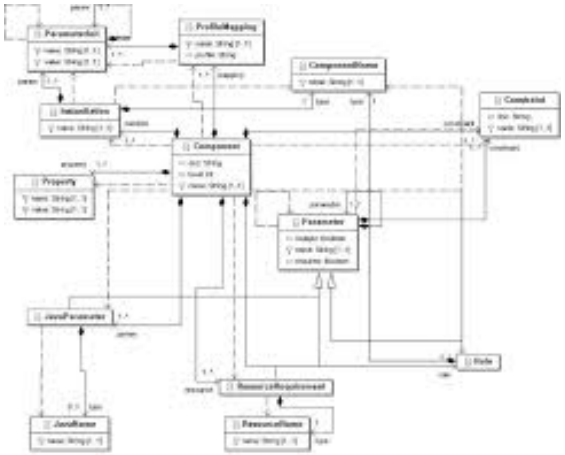
**Babur et al, Hierarchical Clustering of Metamodels for Comparative Analysis and Visualization, ECMFA '16*

Model Comparison

- **Model comparison: a common operation in MDE (and SPL, ...)**
- **Typically ‘deep’ and ‘pairwise’**



graph comparison,
fixpoint calc.,
logical inference
…



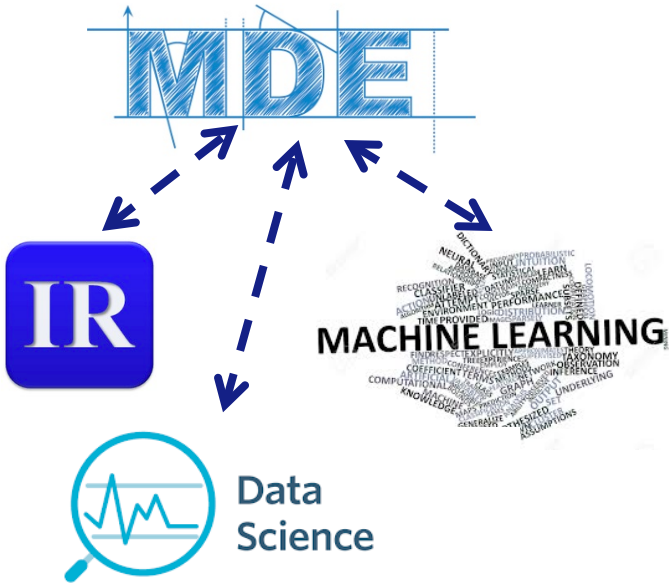
Model Comparison

- Model comparison: a common operation in MDE (and SPL, ...)
- Typically ‘deep’ and ‘pairwise’
- What if
  - Many models (tens? hundreds? millions? ☺)
    - ATL Zoo: ~300 metamodels
    - SPLOT: >1000 feature models, growing
    - Github Ecore crawl: 68k
    - Lindholmen UML dataset: 93k
    - Industry: ~100 metamodels, 55k models

**Babur et al, Models, More Models and then A Lot More, GRAND @ STAF’17*

SAMOS Framework for Model Analytics

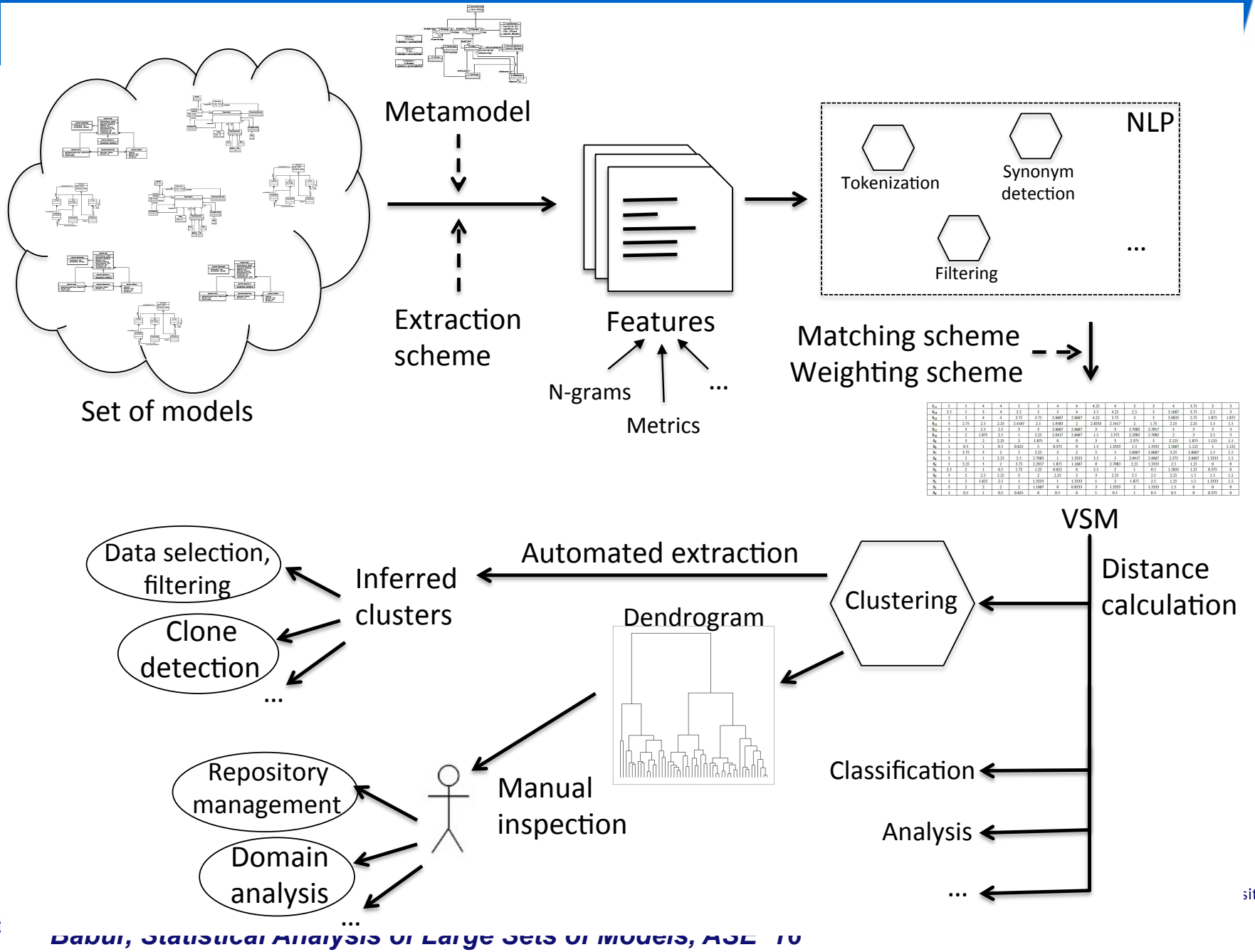
- **Statistical Analysis of ModelS**



- **SAMOS is capable of:**
  - **Feature extraction (fragmentation)**
    - **n-grams, subtrees, metrics, …**
  - **Feature comparison**
    - **natural language processing**
    - **elaborate weight and comparison schemes**
  - **Vector Space Model computation**
  - **Distance measures, statistical analyses in R**
  - **Distributed analysis with Apache Spark**

Babur, Statistical Analysis of Models, ASE '16

SAMOS Framework for Model Analytics



Model Analytics with SAMOS

- **Proposal**
  - **'N-way' model comparison & analysis**
  - **Scalable, fast but approximate**

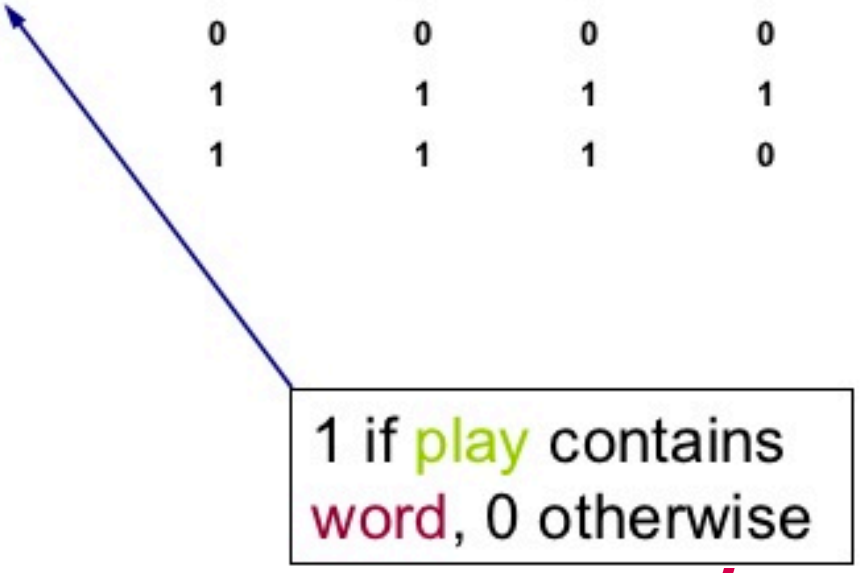
- **Main Components**
  - **[1] Cut the model into fragments**
    -> Information Retrieval
  - **[2] Statistically analyse them**
    -> Machine Learning

Babur, Statistical Analysis of Large Sets of Models, ASE '16

Vector Space Model

- **Vocabulary vs Documents**
  - **occurence / frequency table**

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0



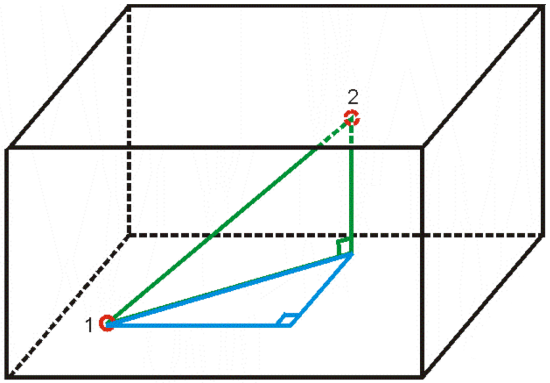
Vector Space Model

- **Vocabulary vs Documents**
  - **occurence / frequency table**

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 if play contains word, 0 otherwise

- **Similarity = distance calculation**



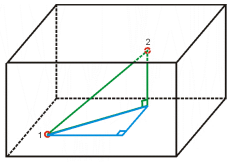
Vector Space Model

- **Vocabulary vs Documents**
  - **occurence / frequency table**

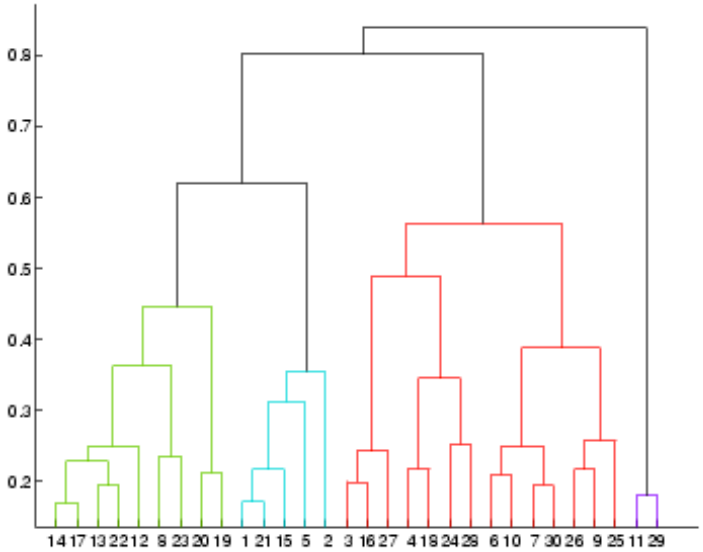
	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 if play contains word, 0 otherwise

- **Similarity = distance calculation**



- **Analysis = clustering**



Vector Space Model

- **Vocabulary vs Models?**
  - **Model element names, types, attributes**

Vector Space Model

- **Vocabulary vs Models?**
  - **Model element names, types, attributes**

<EClass, StateMachine>

<EReference, transitions>

…

Vector Space Model

- **Vocabulary vs Models?**
  - **Model element names, types, attributes**
  - **n-grams of those (=structure)**

<EClass, StateMachine> - contains - <EReference, transitions>

<EClass, InitialState> - supertype - <EClass, State>

…

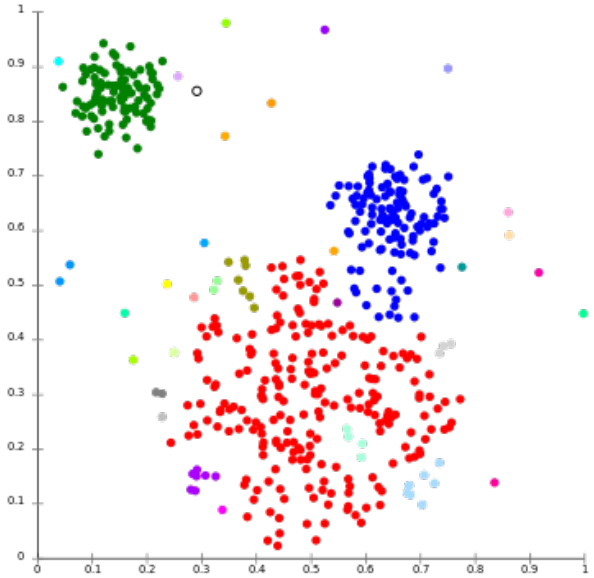
Babur, Cleophas, Using n-grams for the Automated Clustering of Structural Models, SOFSEM'17

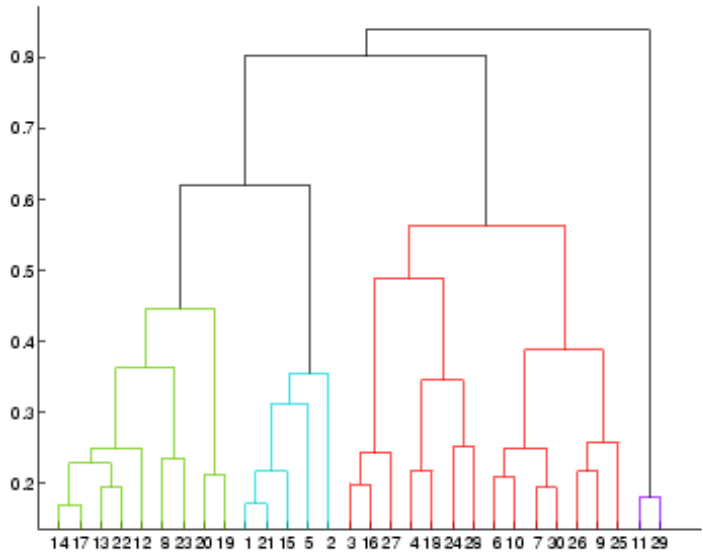
Vector Space Model

- Vocabulary vs Models?
  - Model element names, types, attributes
  - n-grams of those (=structure)
- Comparison, weighting schemes
  - Notably Natural Language Processing
- Many more parameters…

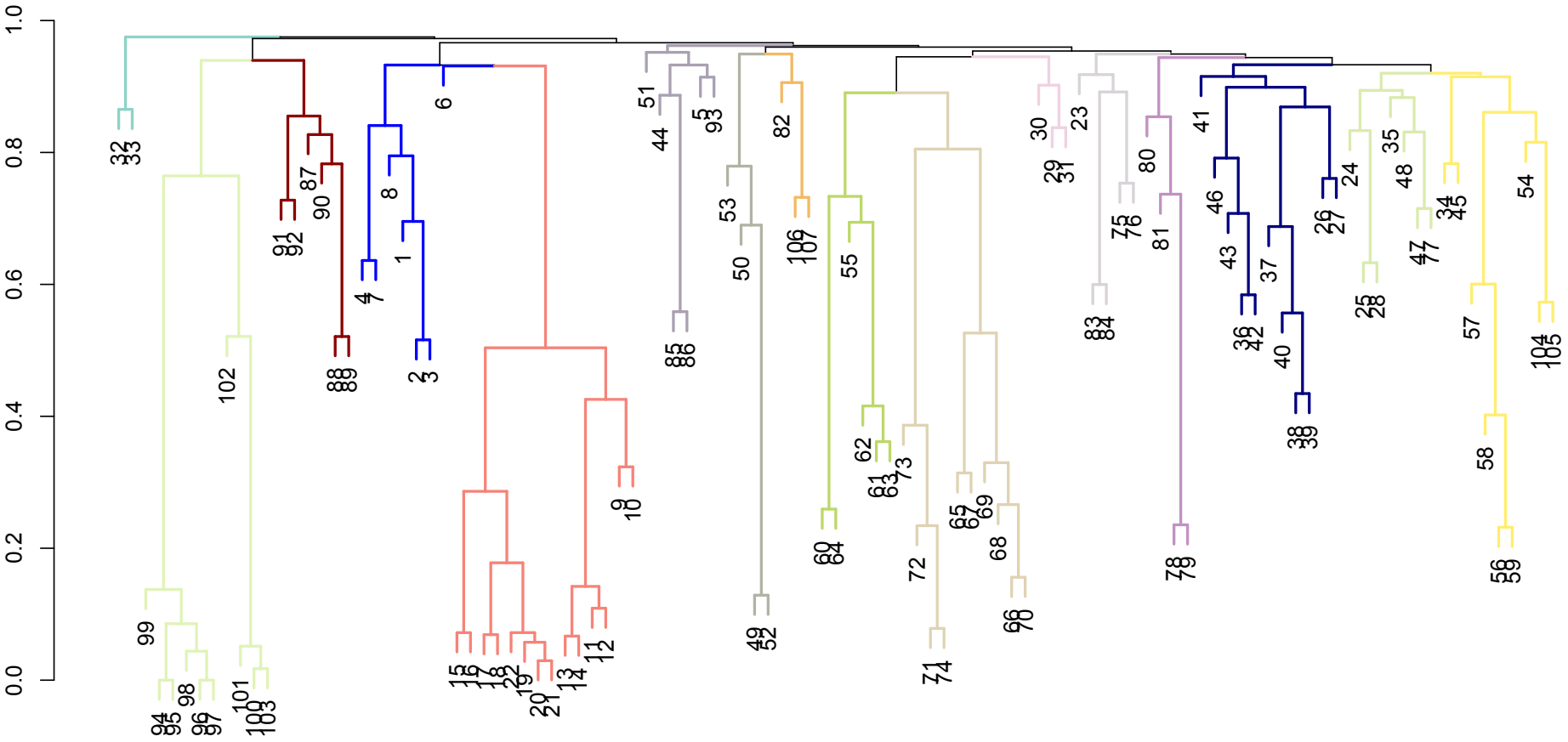
Analysis of Models

- **Clustering (Statistics)**
  - **Unsupervised machine learning**
  - **Grouping similar objects**





Example*: ATL Metamodel Zoo



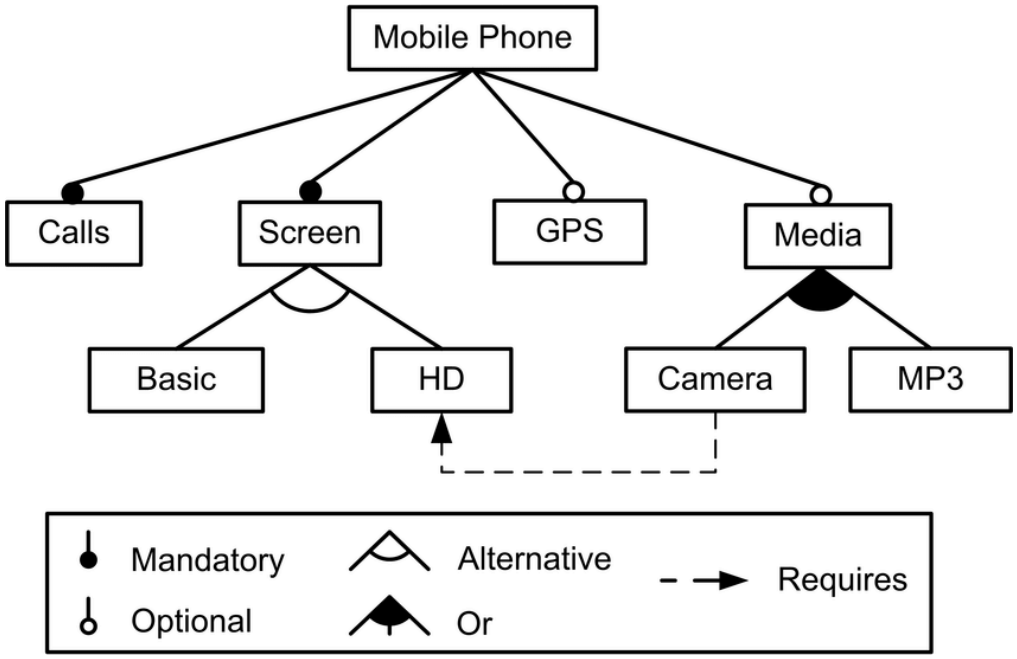
**Babur et al, Hierarchical Clustering of Metamodels for Comparative Analysis and Visualization, ECMFA '16*

Case for SPLOT Repository

- **Feature Model repository**
  - **Many models (>1000), growing**
- **Issues hindering management, search and reuse**
  - **lack of metadata and domain info**
  - **lots of duplicates/clones**
- **Proposal: model analytics using SAMOS for**
  - **domain analysis and repo. overview**
  - **elimination/marking of duplicates/clones**

Case for SPLOT Repository

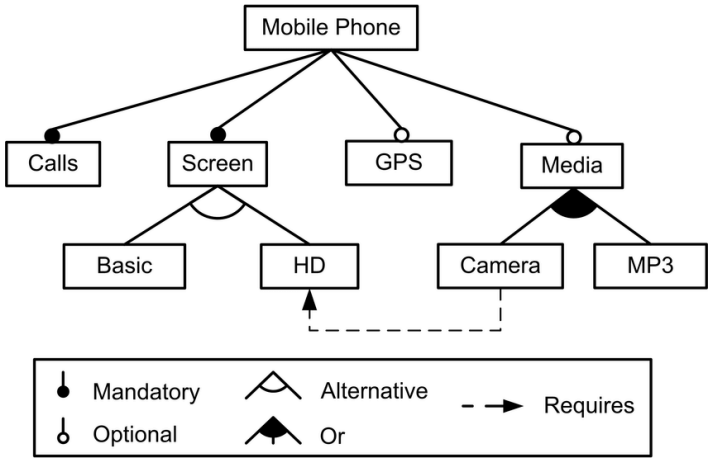
- **Feature Model repository**



Case for SPLOT Repository

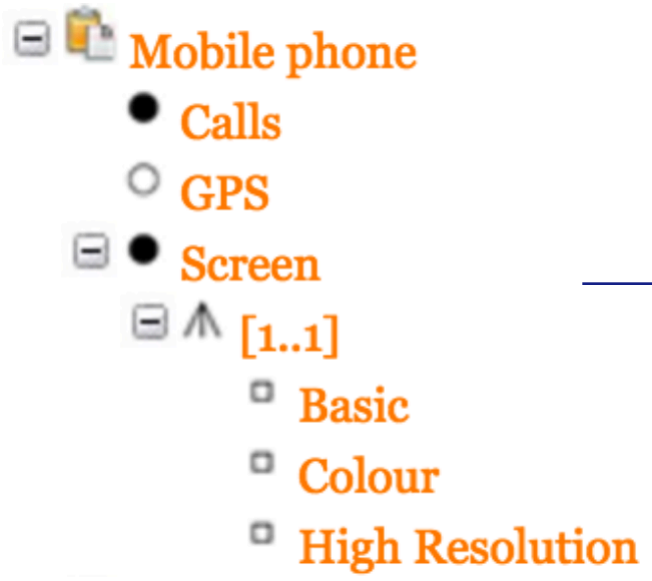
- **Feature Model repository**

- **Note the SPL/MDE literature**
  - **mostly focuses on**
    - **config. semantics**
    - **inference**
  - **while ignoring**
    - **ontological (tree) structure**
    - **(partial) similarity**
    - **natural language processing aspects**



Case for SPLOT Repository

- **(IR-) feature extraction from feature models**



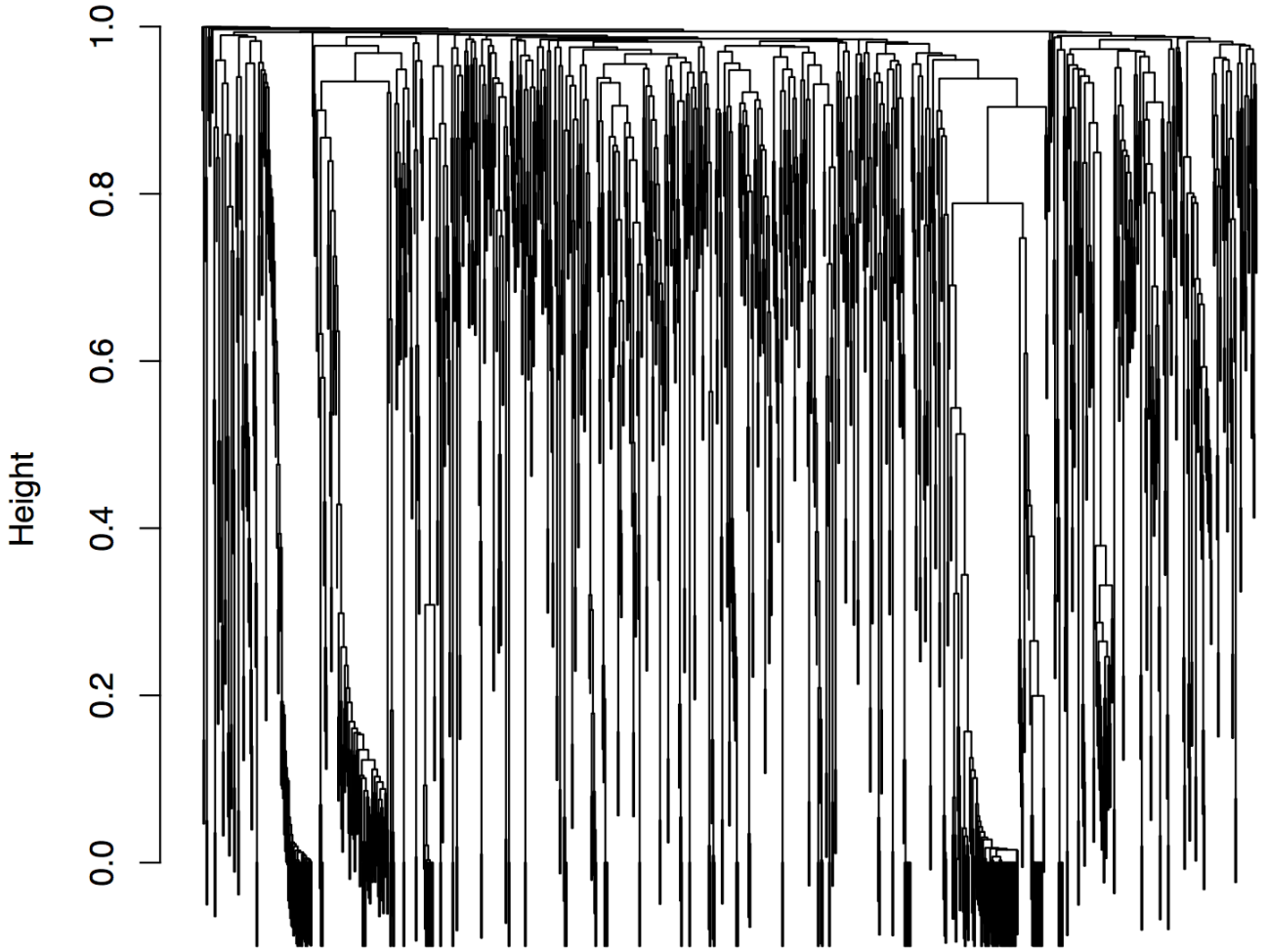
type	IR-features
unigram	(Root-Mobile phone) (Mandatory-Calls) (Optional-GPS) (Mandatory-Screen) (Grouped-Basic) …
bigram	(Root-Mobile phone)-(child[1..1])-(Solitaire-Calls) (Root-Mobile phone)-(child[0..1]-(Solitaire-GPS) (Solitaire-Screen)-(child[1..1])-(Grouped-Basic) (Solitaire-Screen)-(child[1..1])-(Grouped-Colour) …
constr	(~GPS,~Basic) (High resolution,~Basic) …

Cutting Feature Models Open - 1

- Goal: Domain Analysis
- What feature to extract: names
  - enough for domain analysis!! (e.g. bank or car FM)
- How to compare features: NLP
  - tokenization, lemmatization, stemming
  - stopword removal, typos, semantic relatedness
- Analysis: Cosine distance and hierarhical clustering

Repository Overview

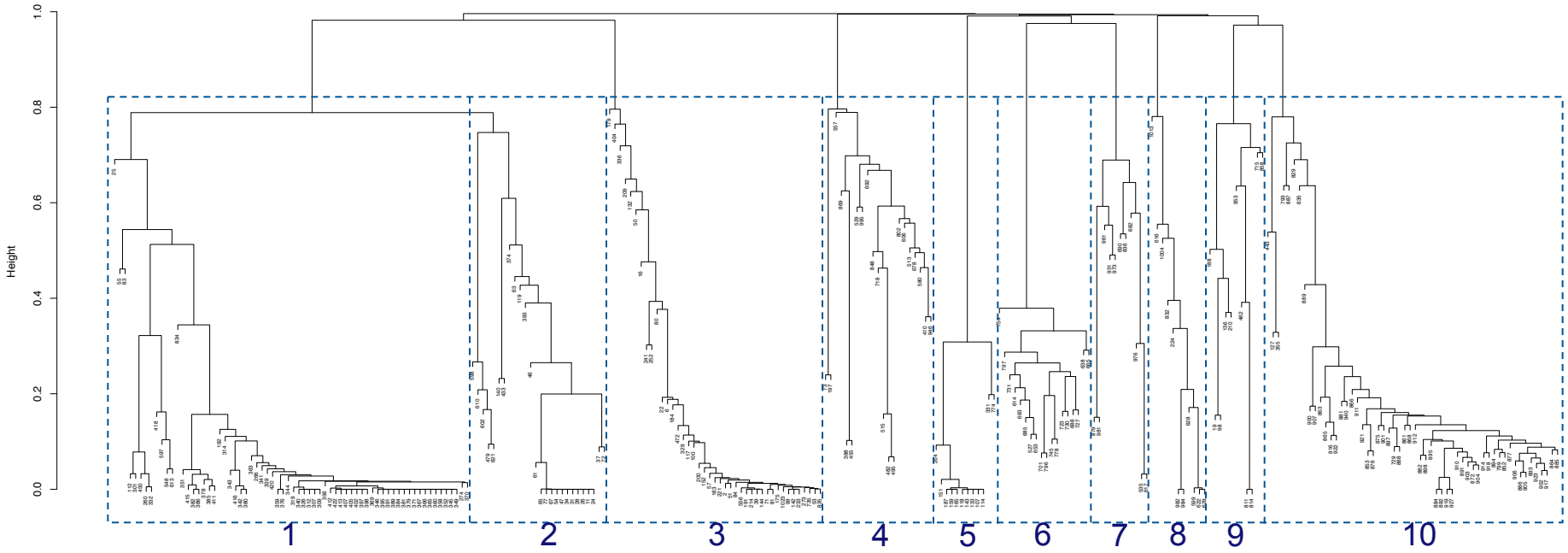
- **Diverse repository, not all clear and thick branching**
  - **Too many domains?**



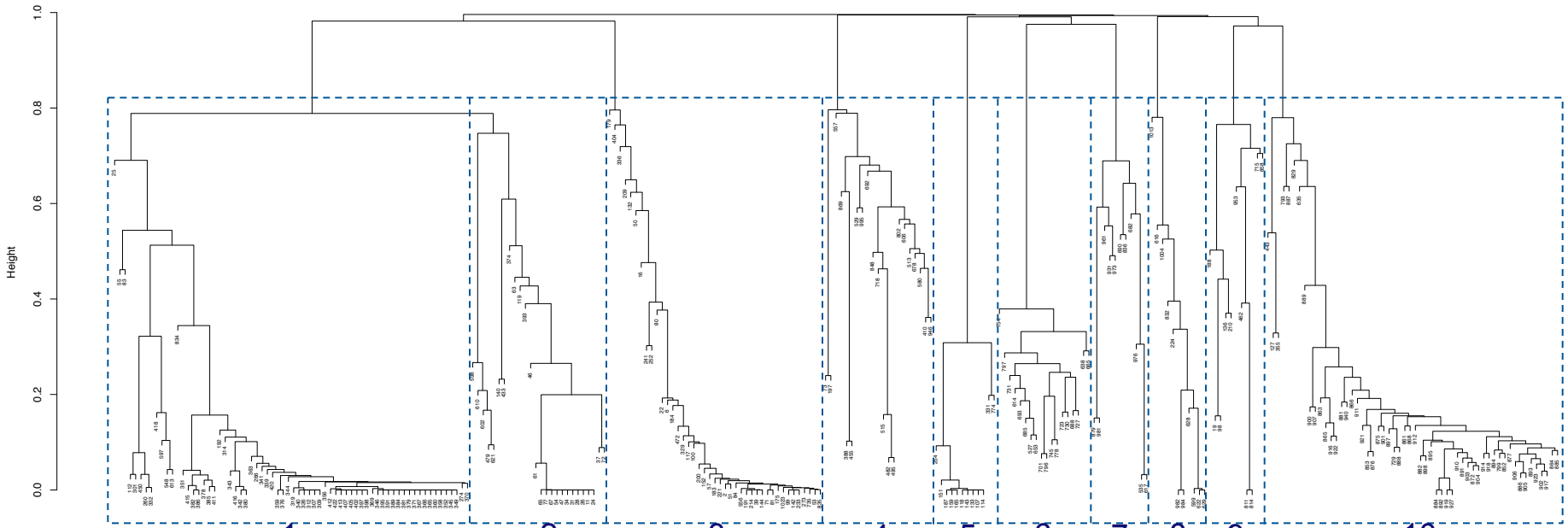
Solution: Repository Subset

- Get a subset of the feature models
  - highly similar (cosine distance < 0.8)
  - in large clusters (size > 20)
- Others are considered “outliers”, excluded from our preliminary analysis
- Result: 275 models in the subset
  - Clustering result manually inspected and labelled

Results: Domain Analysis



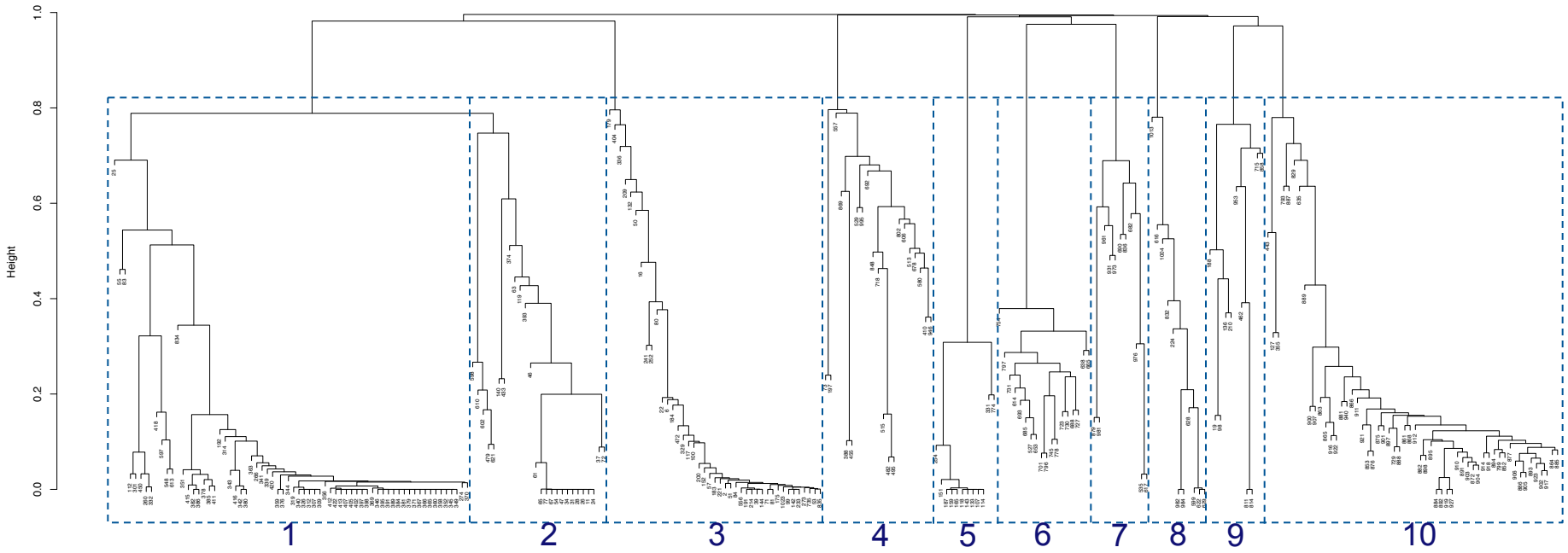
Results: Domain Analysis



mobile media

mobile phone

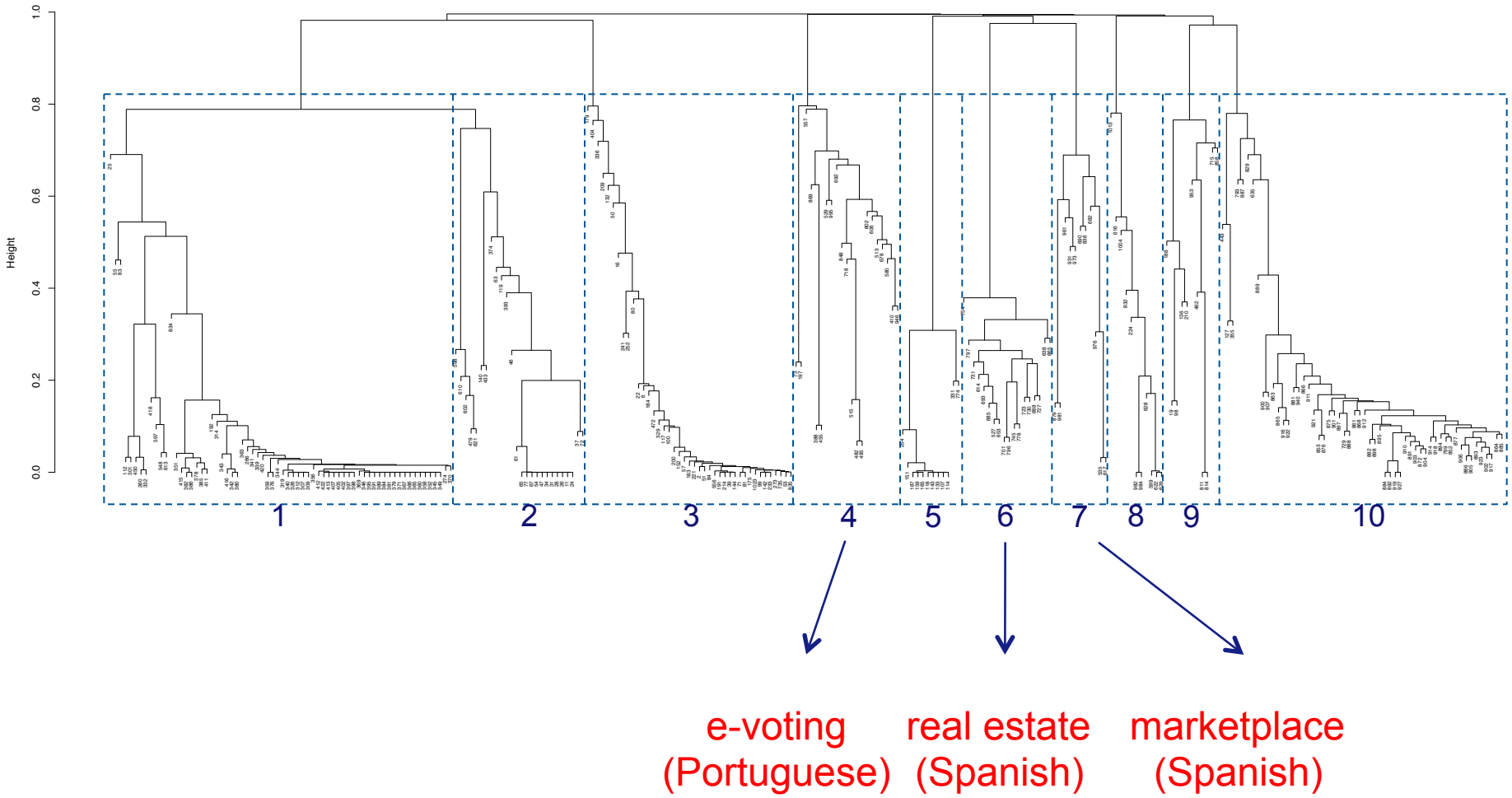
Results: Domain Analysis



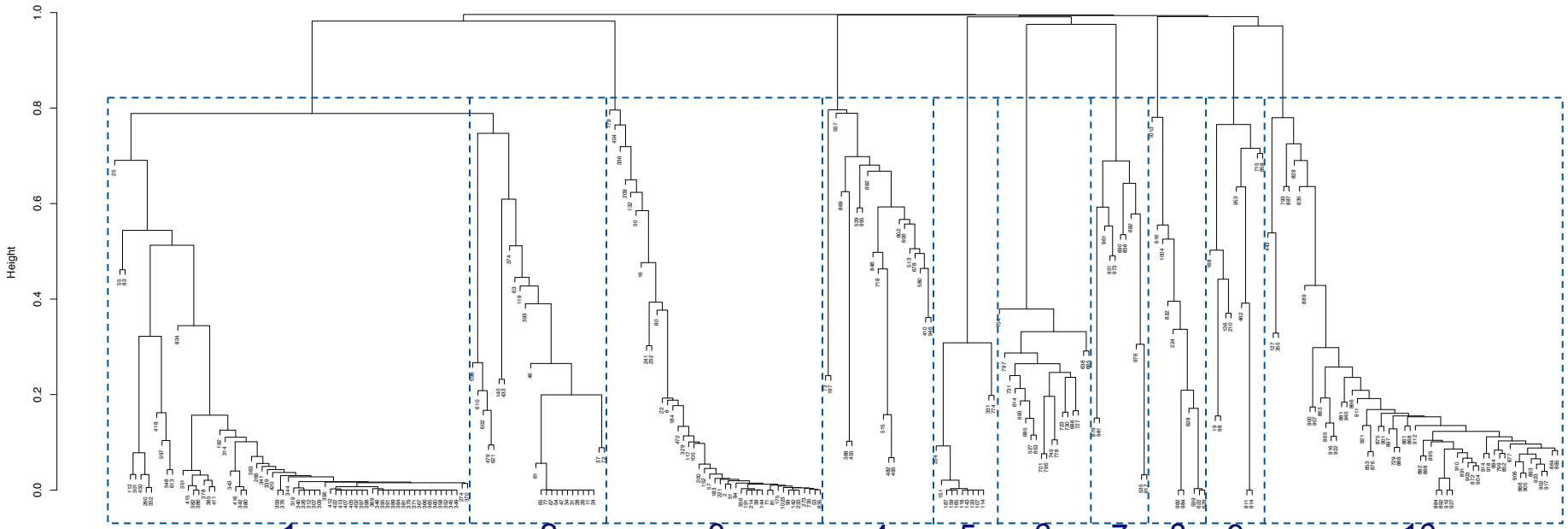
Feature-1
Feature-2
…

F1
F2
F_1
…

Results: Domain Analysis



Results: Domain Analysis



e-shop/commerce …

Domain Analysis - Discussion

- **Not an easy task to fully automate!**
  - **many domains, few models per domain**
  - **domain-specific terminology (hard for NLP)**
  - **multi-lingual**

- **Cannot get all the domains with 100% accuracy, still can help with**
  - **repository exploration (end-user side)**
  - **curation (management side)**

Cutting Feature Models Open - 2

- Goal: Clone detection
- What feature to extract: bigrams, sets
  - names, types
  - structural relations, cardinalities
  - constraints
- How to compare features: n-gram & set comparison
- Analysis: Bray-Curtis distance, density-based clustering

Results: Clone Detection

- **Clone classification:**
  - **Type A: identical models (except cosmetic NLP)**
  - **Type B: highly similar models (<10% difference)**
  - **Type C: considerably similar models (<30% difference)**

clone type	#clusters	#pairs	#models involved
A	22	64	59
B	60	1472	208
C	90	3320	382

Observations w.r.t. Type A clones

- **SAMOS was 100% accurate (in validation subset)**

- **Diff's observed:**
  - **date of creation of the model**
  - **feature model name, metadata,**
  - **constraint names (i.e. not the content),**
  - **order of elements in the feature tree and CNF formulas,**
  - **consistently changed feature id's (which lead to e.g. completely different looking constraint formulas)**
  - **cosmetic changes in feature names (e.g. upper/lower and snake/camel casing),**
  - **…**

Observations w.r.t. Type B clones

- **SAMOS was mostly accurate**
  - **Some manually labelled as Type C (discussion later)**

- **Diff's observed:**
  - **cardinalities,**
  - **textual changes in feature names (e.g. typos, additional tokens),**
  - **addition or removal of features and constraints,**
  - **moving features to elsewhere in the feature tree,**
  - **…**

Observations w.r.t. Type C clones

- SAMOS was generally accurate ☺
- Diff's observed:
  - higher percentage of addition, removal, or changes in feature trees and constraints
- Problematic:
  - similar names: Feature-1 and Feature-2
  - bigram simplification for grouped features
    - need more accurate representation (tree?)
  - weighting might be needed
    - e.g. higher hierarchy = larger weight?

Clone Detection - Discussion

- **Clone detection with SAMOS - work in progress**

- **High accuracy (although not 100%), can help identify (implicit)**
  - **duplicates**
  - **versions (considering time and/or ownership)**
  - **variants**

Discussion and Future Work

- Overall promising approach, qualitatively evaluated
- Room for improvement
  - Quantitative evaluation w/ labelled datasets
  - More accurate representation: e.g. grouped features
  - Configuration semantics (approximately?) captured
  - Weighting, fine tuning of SAMOS for feature models
  - Better (and possibly multi-lingual) NLP

Thanks!

https://modelanalytics.wordpress.com

Contact: O.Babur@tue.nl

Önder Babur

Eindhoven University of Technology