AUTONOMOUS SYSTEMS INITIATIVE

UNIVERSITY OF ALBERTA
UNIVERSITY OF CALGARY
University of Lethbridge
RDC
SAIT
Alberta

AUTONOMOUS SYSTEMS INITIATIVE

Path Based Hierarchical Clustering on Knowledge Graphs

Marcin Pietrasik and Marek Reformat

UNIVERSITY OF ALBERTA · UNIVERSITY OF CALGARY · University of Lethbridge · RDC · SAIT · Alberta

Overview

- Knowledge graphs - introduction
- Motivation for our work
- Brief overview of proposed method

What is a knowledge graph?

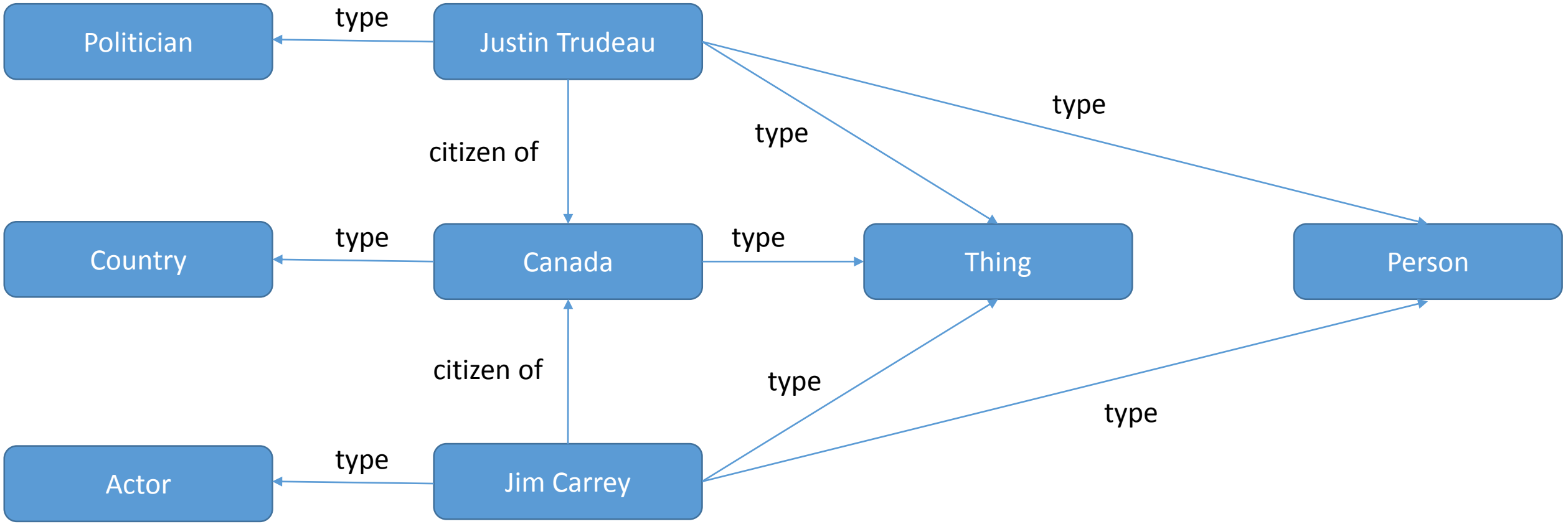
AUTONOMOUS SYSTEMS INITIATIVE

- A medium for storing information as a graph
- Information in knowledge graphs is expressed as **triples**
- Triples link a subject to an object via a predicate

<subject, predicate, object>

<Justin Trudeau, citizen of, Canada>

Toy knowledge graph



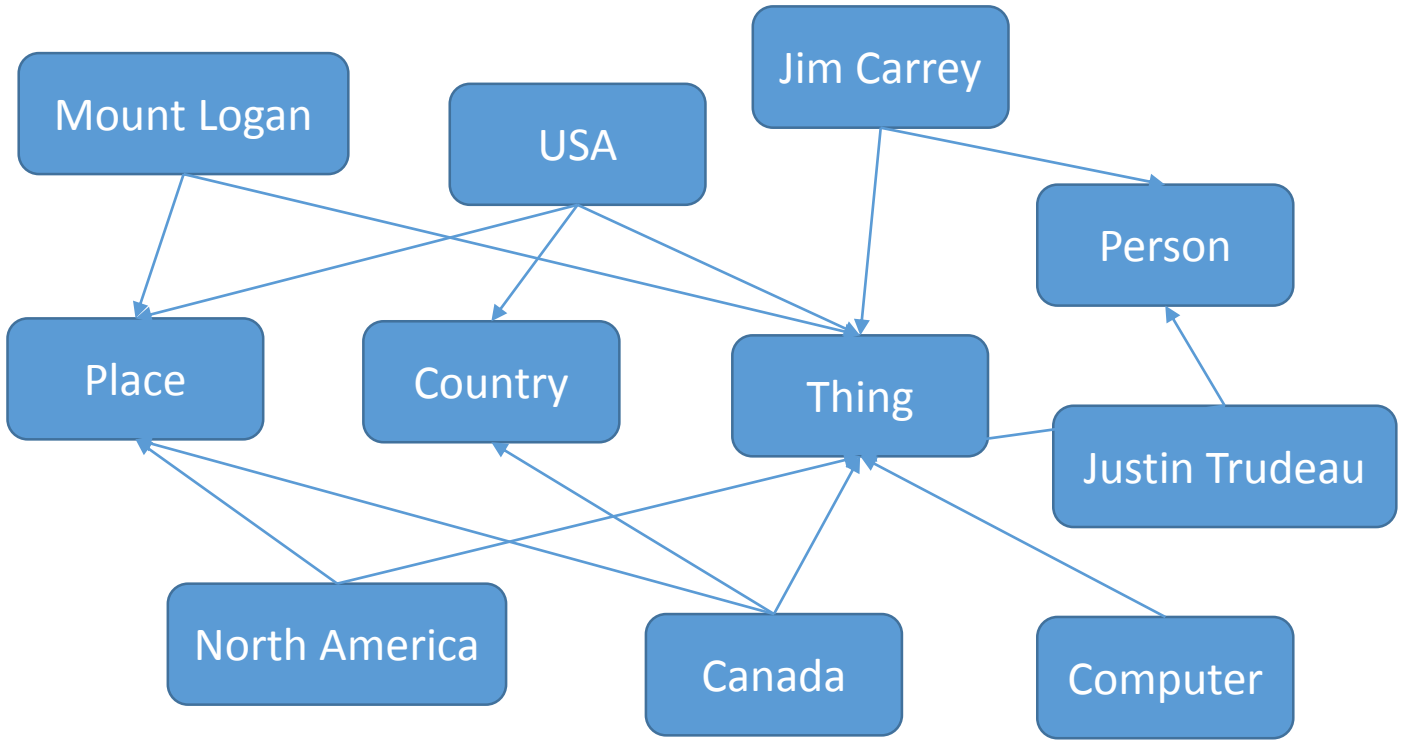
Motivation

- Problem: knowledge graphs are
  - Flat
  - No hierarchical relations between entities
- Objective:
  - find a way to induce a hierarchical clustering of knowledge graph subjects and provide a label for each cluster

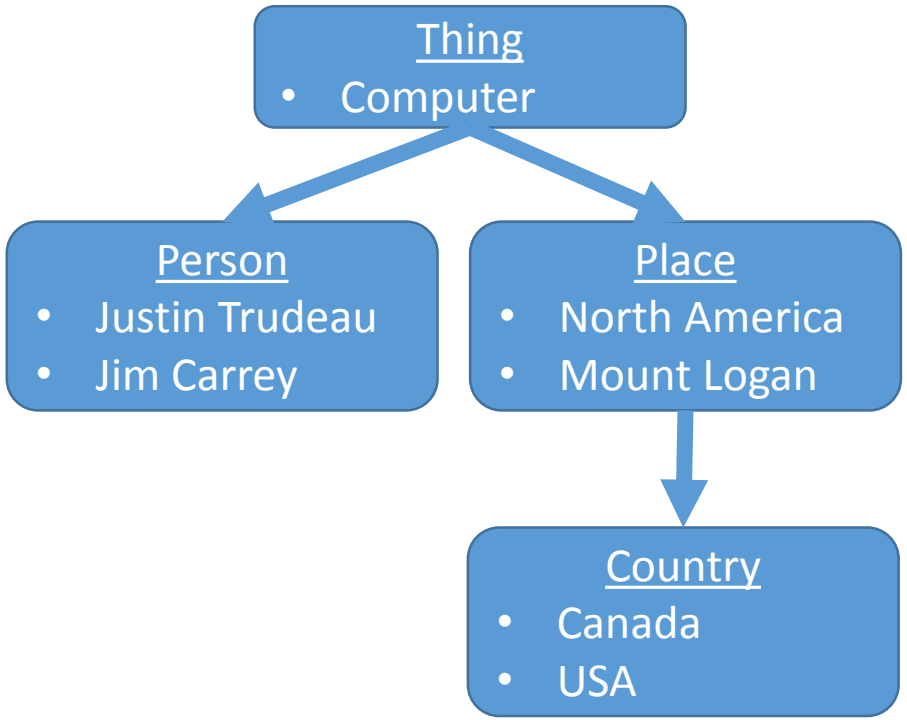
Structure in a knowledge graph

AUTONOMOUS SYSTEMS INITIATIVE

Flat



Hierarchical



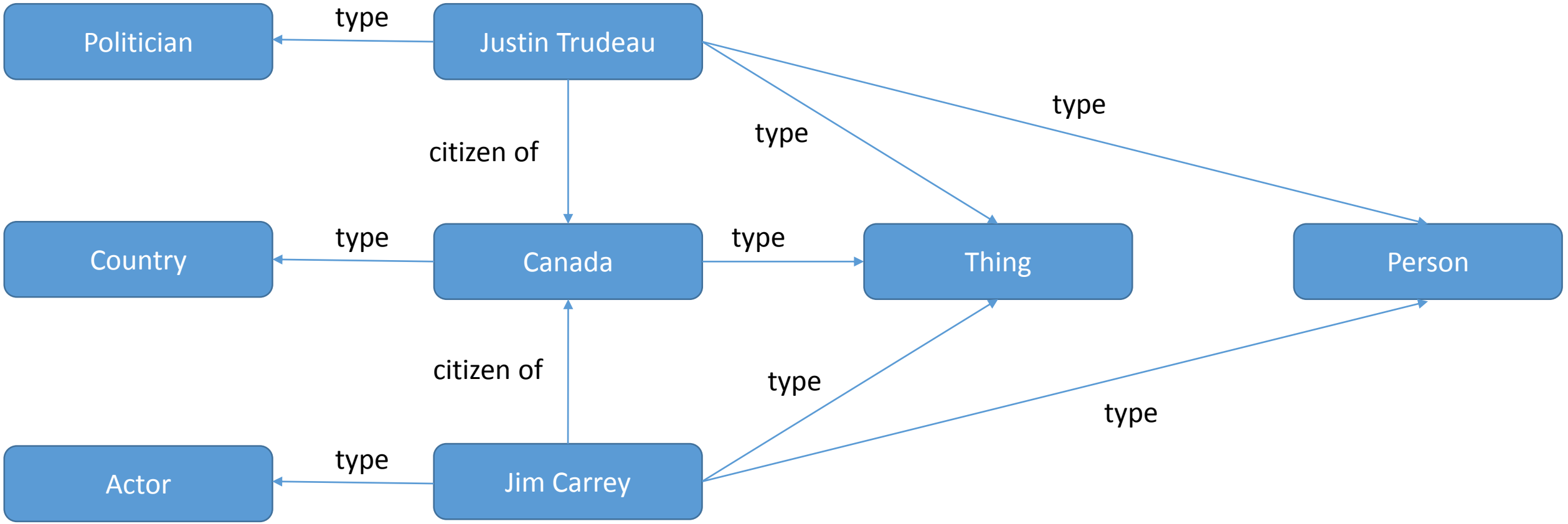
Approach

- Three steps:
  1. Induce a hierarchy of classes
  2. Assign subjects to the induced hierarchy
  3. Prune the hierarchy of empty clusters

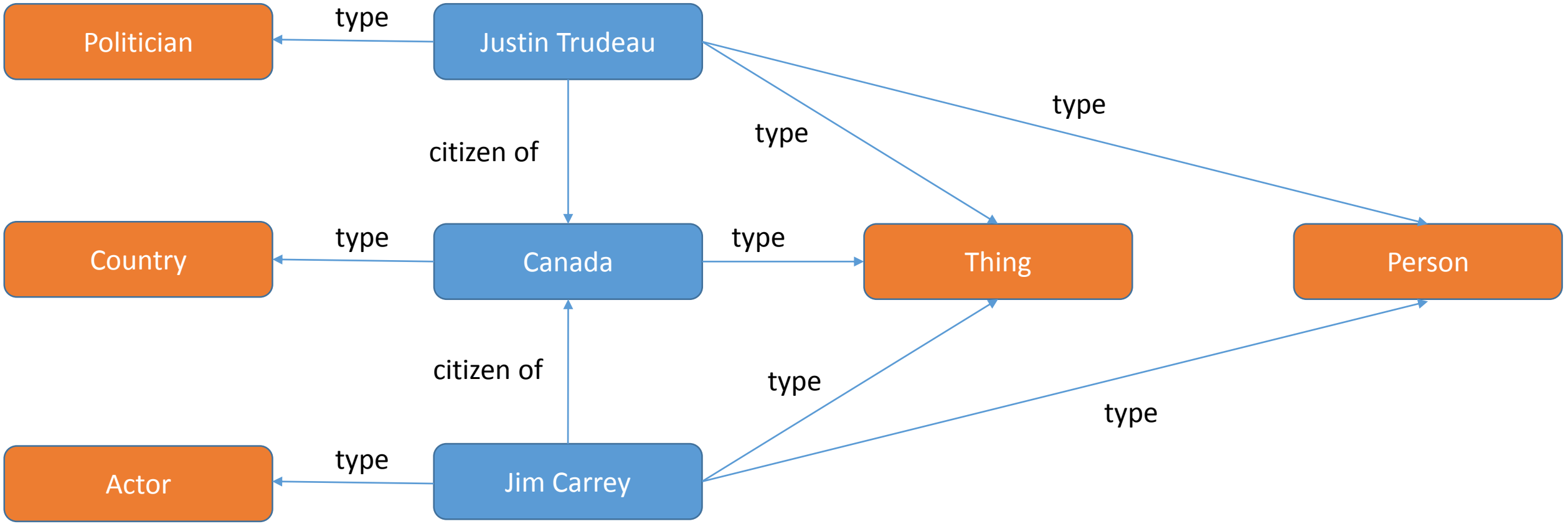
Step 1

- We induce a class hierarchy using the *smict* method proposed in our earlier work
  - Pietrasik, Marcin, and Marek Reformat. "A Simple Method for Inducing Class Taxonomies in Knowledge Graphs." In *European Semantic Web Conference*, pp. 53-68. Springer, Cham, 2020. 2020.
- Hierarchy built on class frequencies and co-occurrences

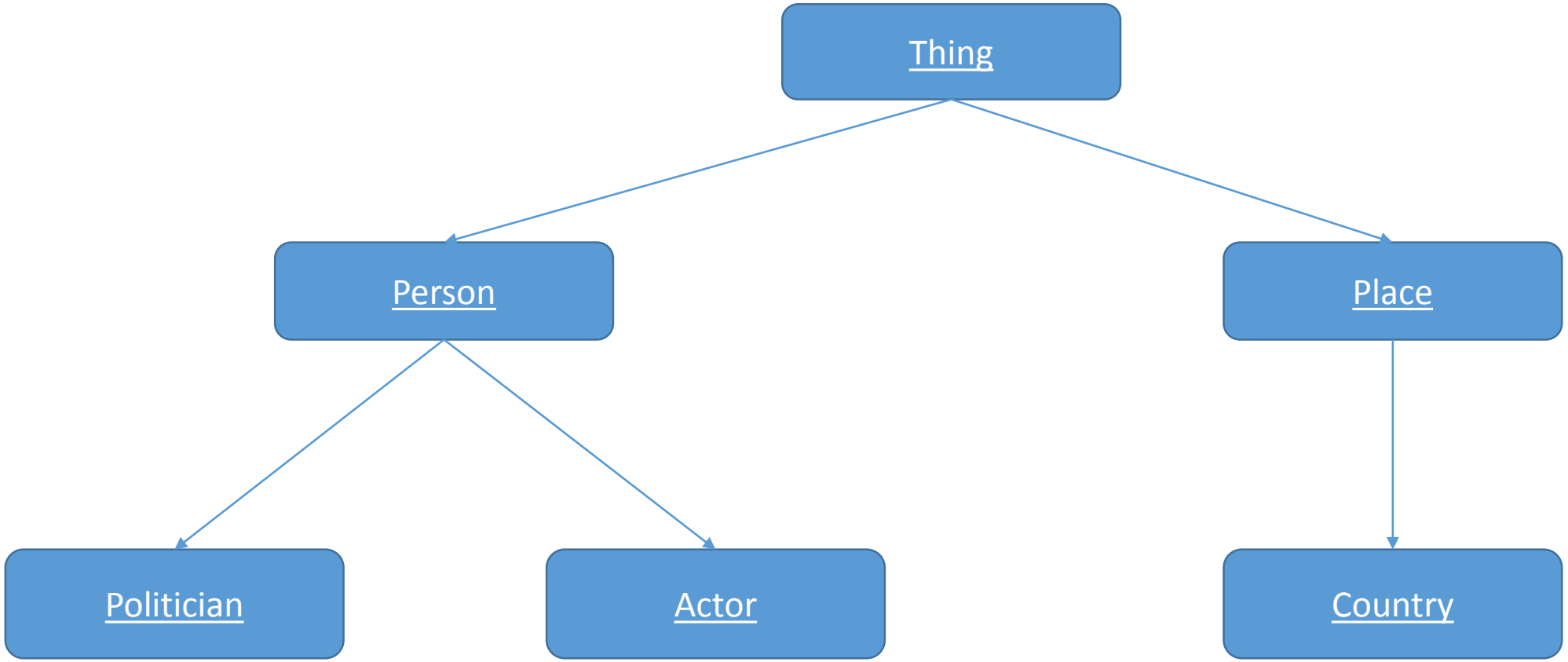
Step 1 Visualization



Step 1 Visualization



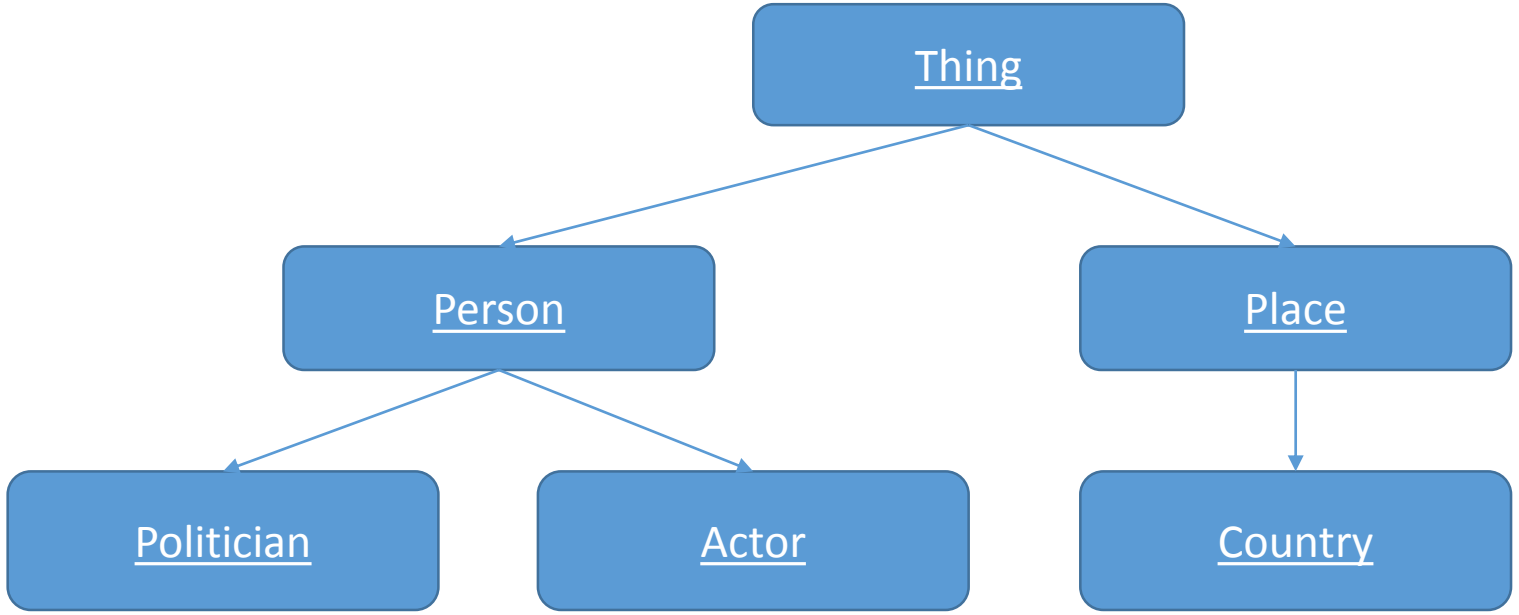
Step 1 Visualization



Step 2

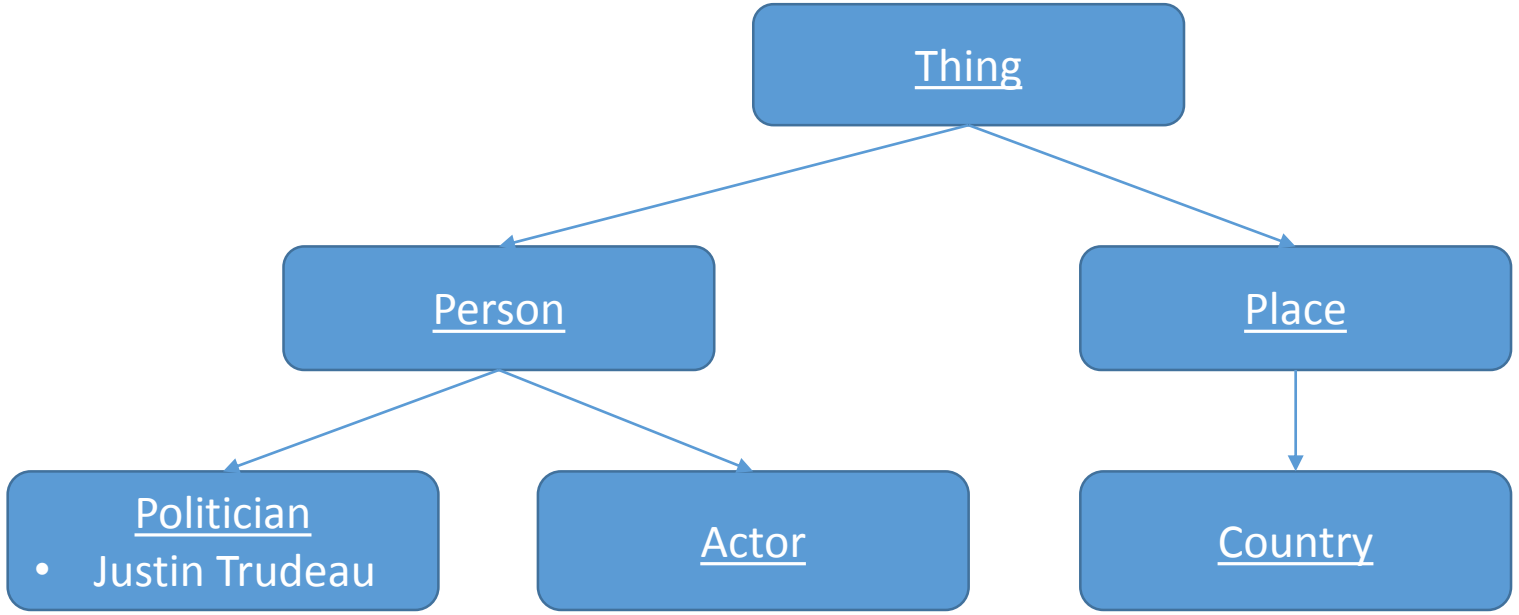
- Subjects are added to the hierarchy
- This process generates hierarchical clusters from the knowledge graph

Step 2 Visualization



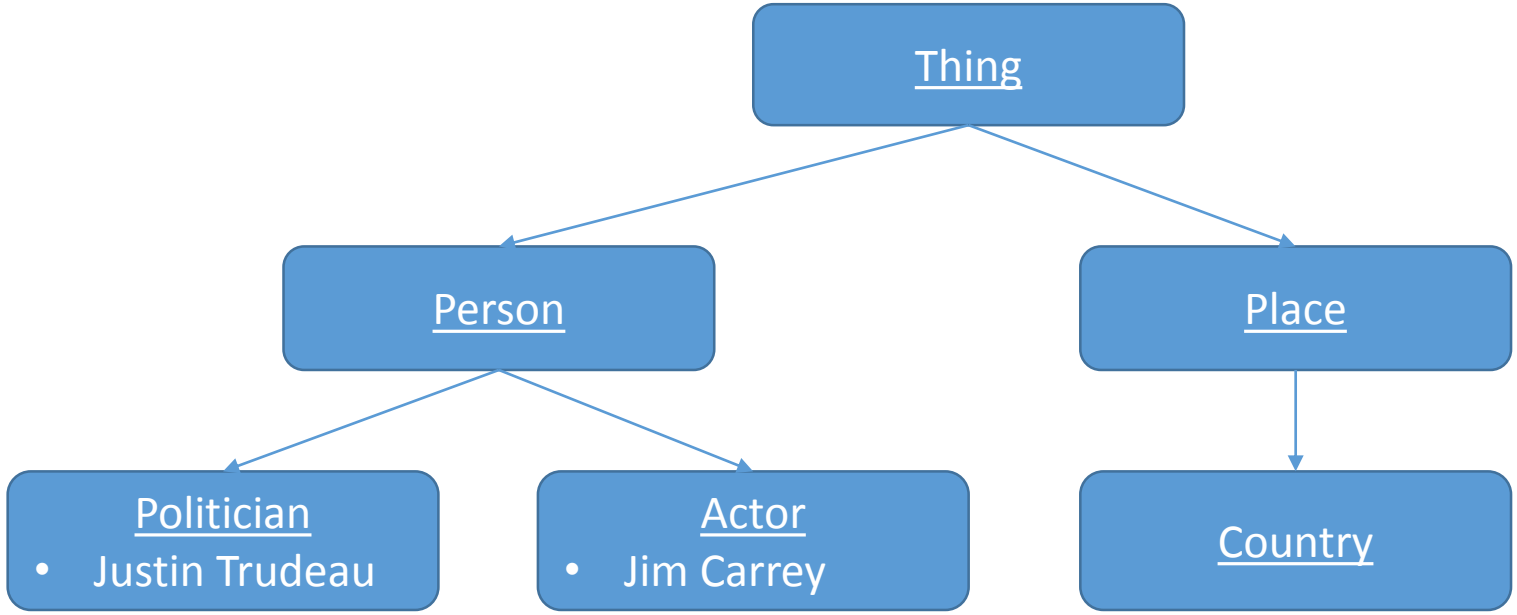
Justin Trudeau
Jim Carrey
Canada
Donald Trump
Wayne Gretzky
North America

Step 2 Visualization



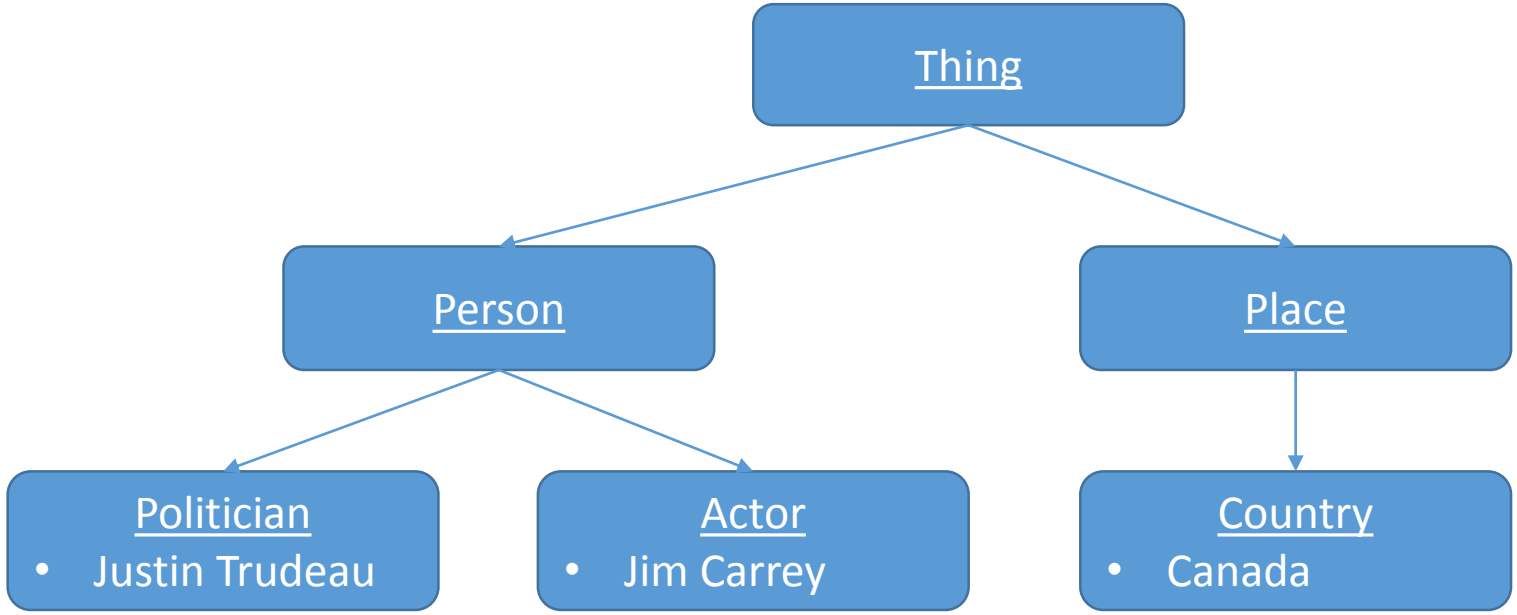
Jim Carrey
Canada
Donald Trump
Wayne Gretzky
North America

Step 2 Visualization



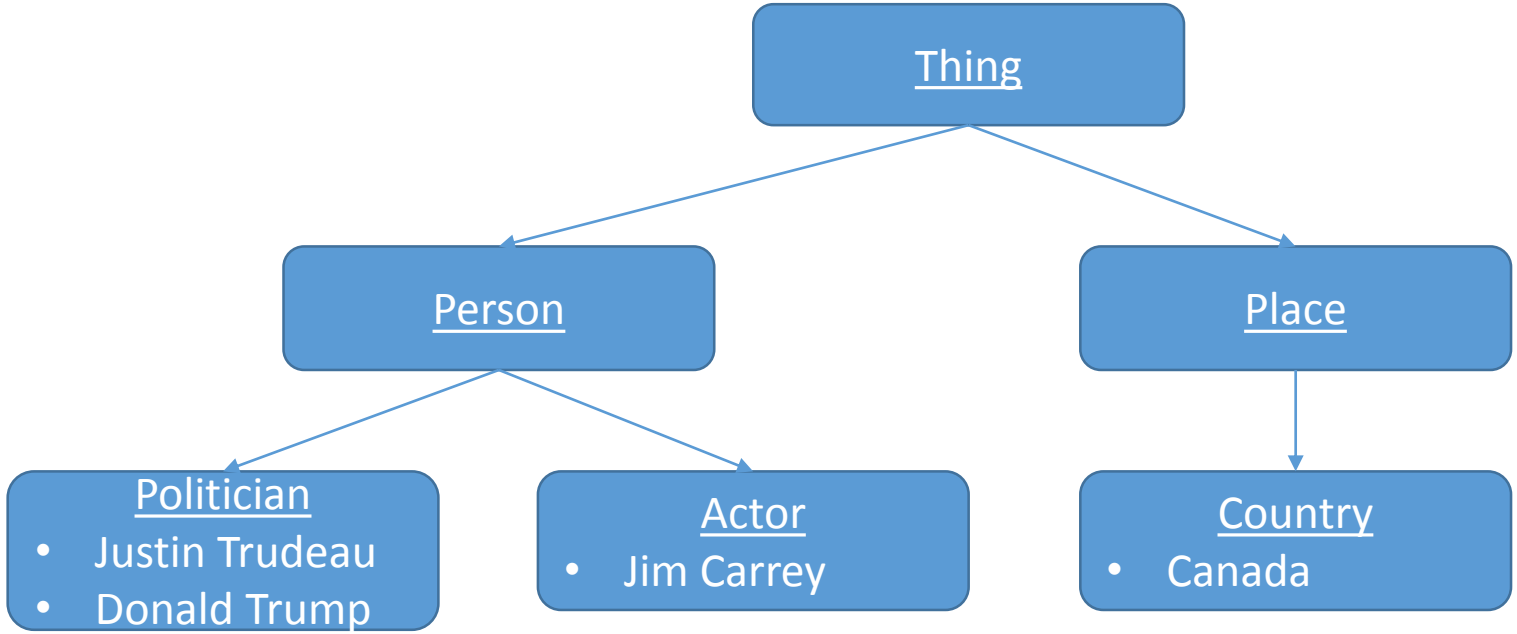
Canada
Donald Trump
Wayne Gretzky
North America

Step 2 Visualization



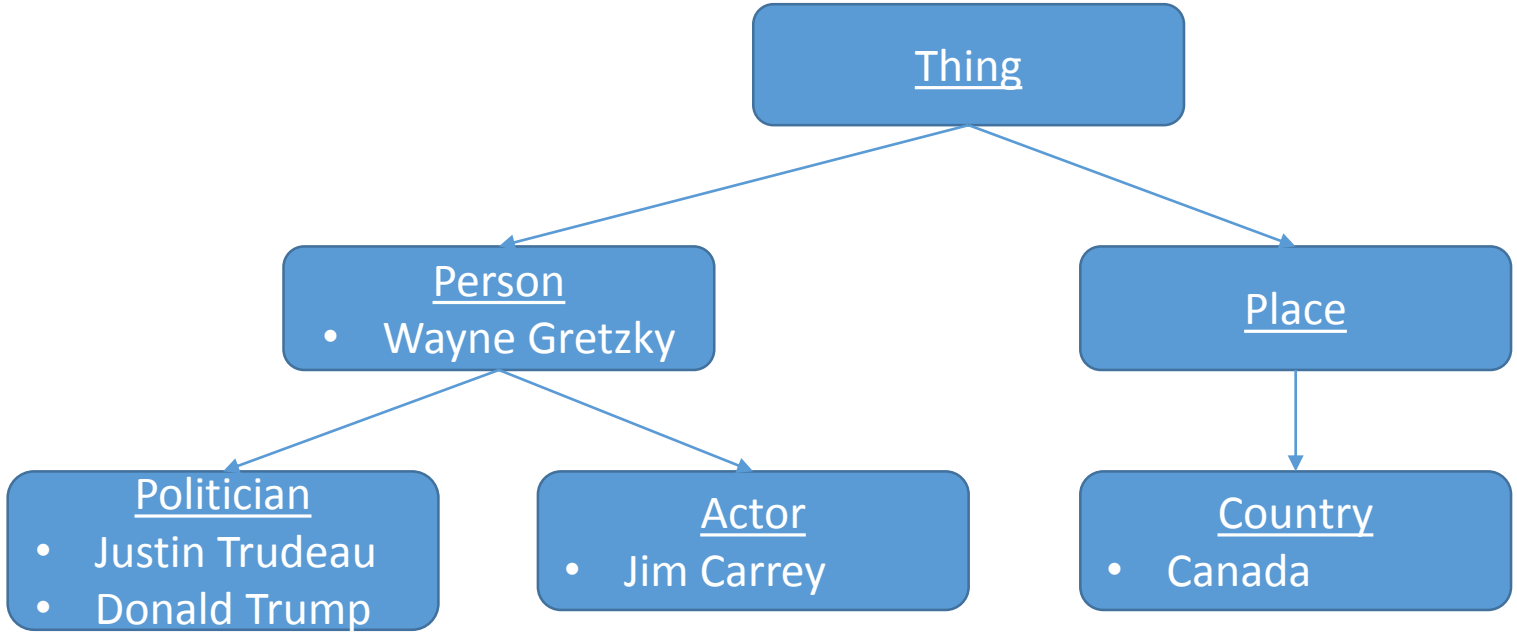
Donald Trump
Wayne Gretzky
North America

Step 2 Visualization



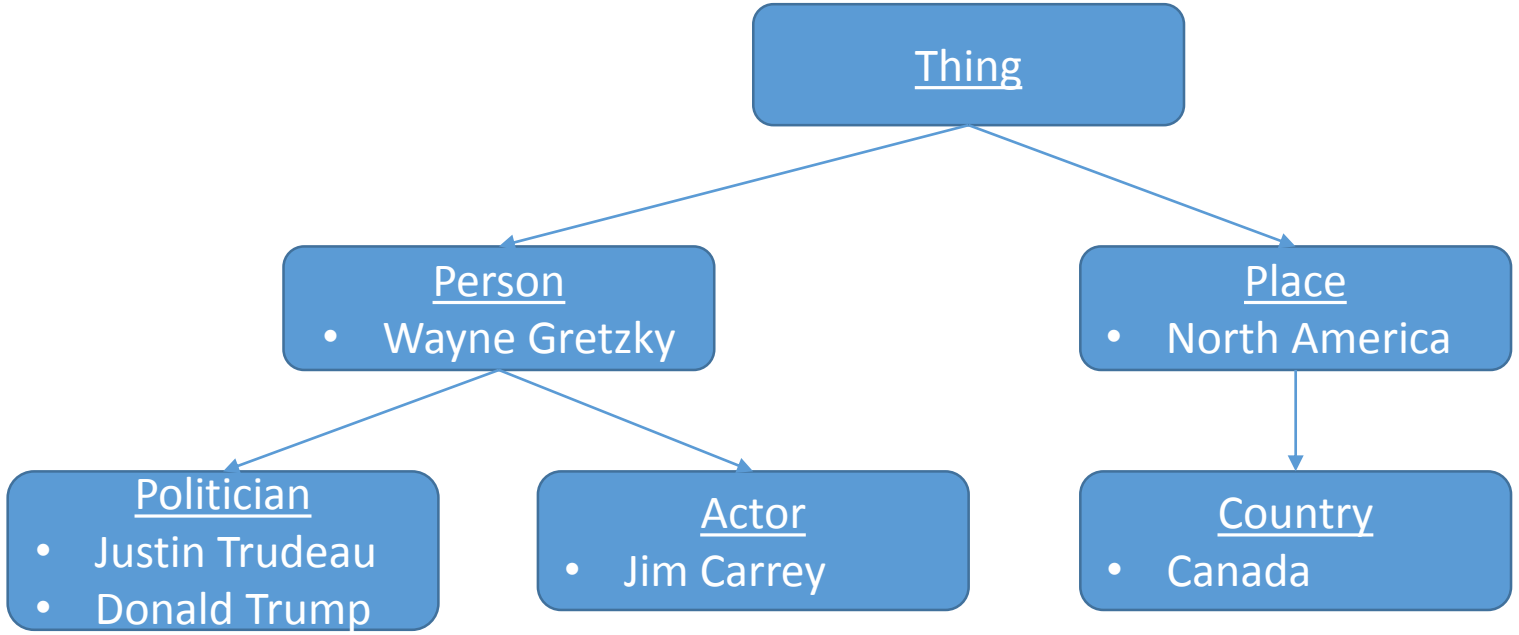
Wayne Gretzky
North America

Step 2 Visualization



North America

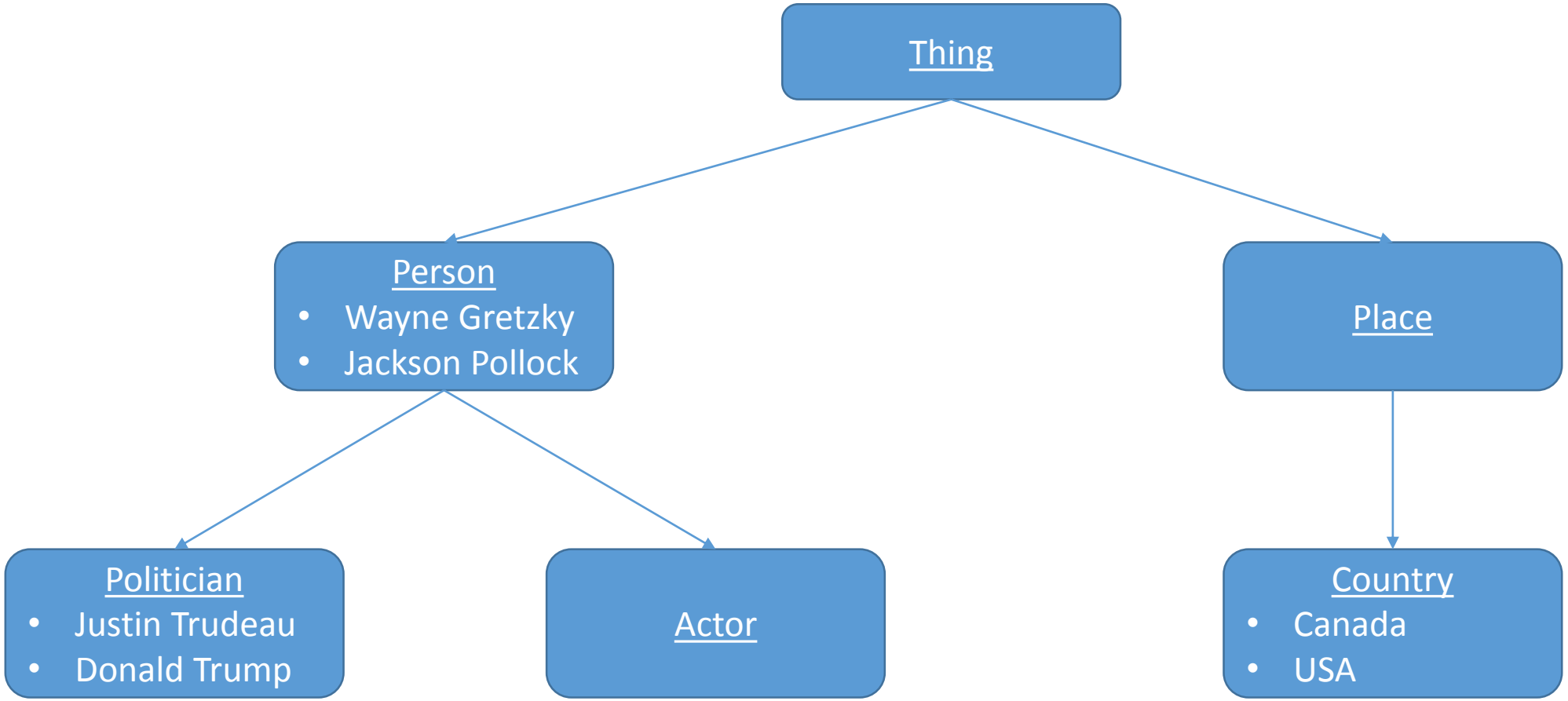
Step 2 Visualization



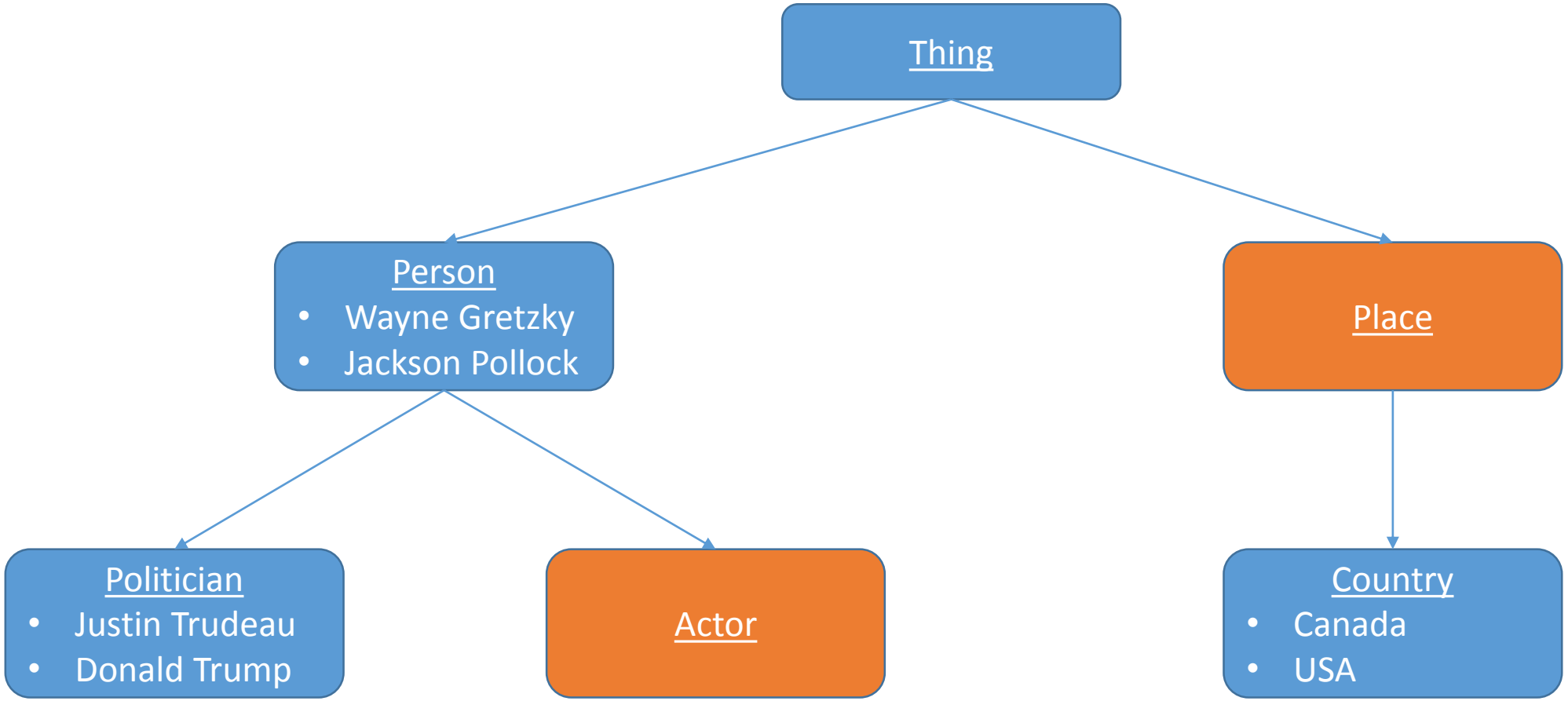
Step 3

- Prune the hierarchy by removing empty clusters and reattach orphaned clusters to next ancestor in line
- Root cluster is never removed

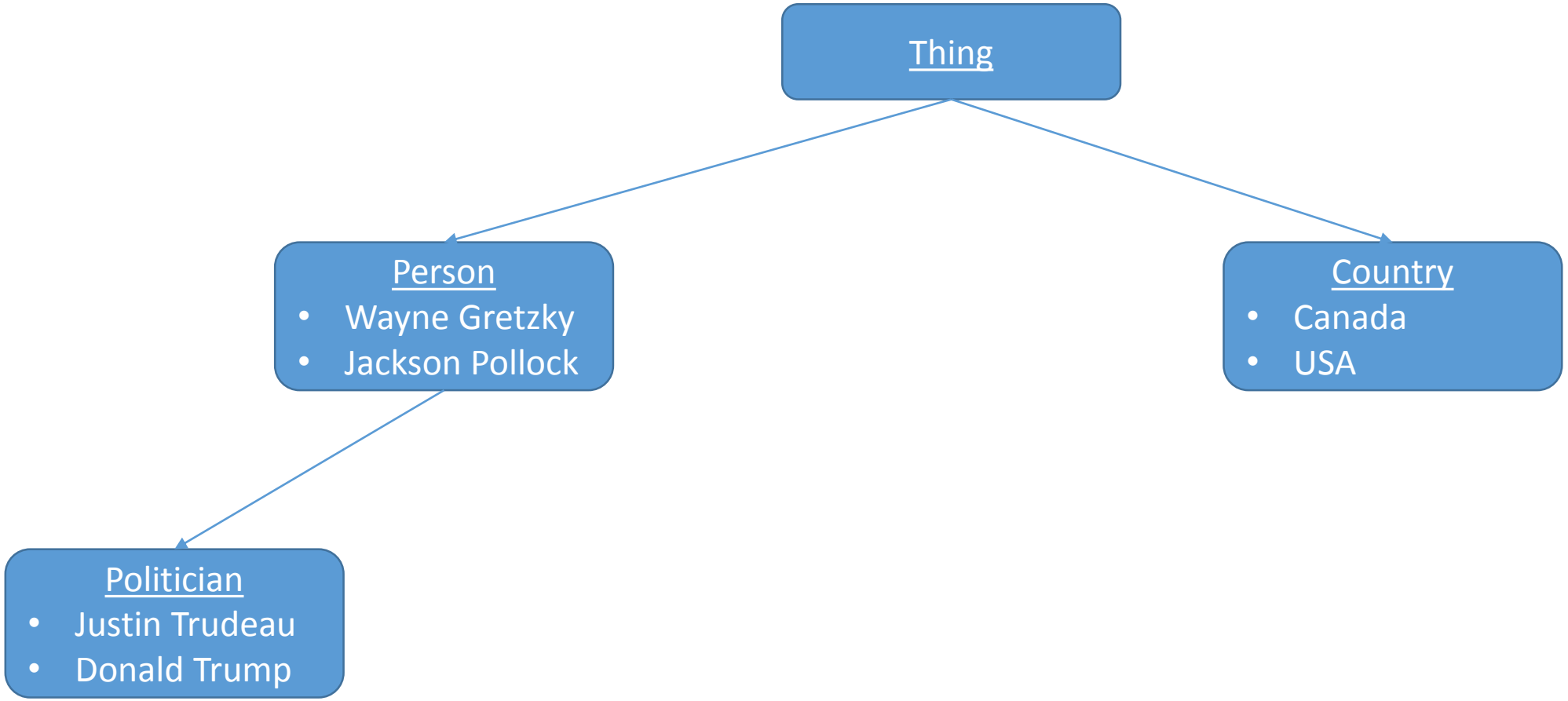
Step 3 Visualization



Step 3 Visualization



Step 3 Visualization



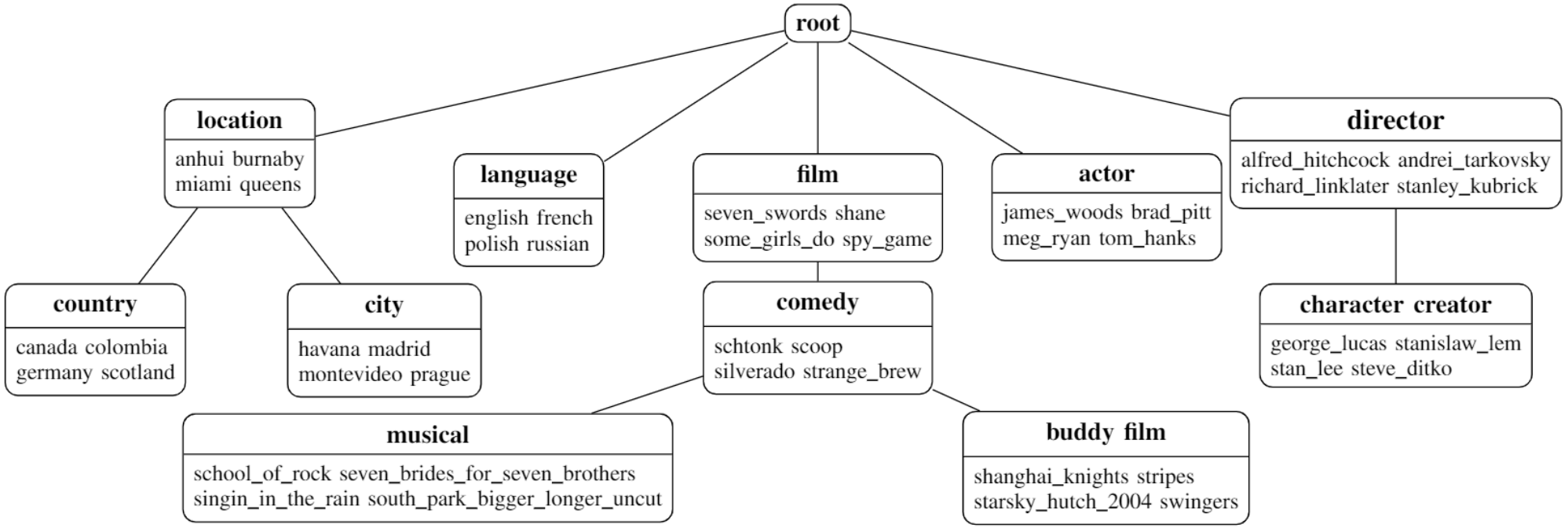
Evaluation

- We apply our method on three datasets:
  - IIMB (movies)
  - DBpedia (Wikipedia facts)
  - WordNet (English language)
- We use three metrics to measure performance: Hie-F1, Sub-F1, and Tag-F1.

Results

	Hie-F1	Sub-F1	Tag-F1
IIMB	0.4444	0.8905	0.7843
DBpedia	0.8627	0.9659	0.9603
WordNet	0.6579	0.9212	0.8998

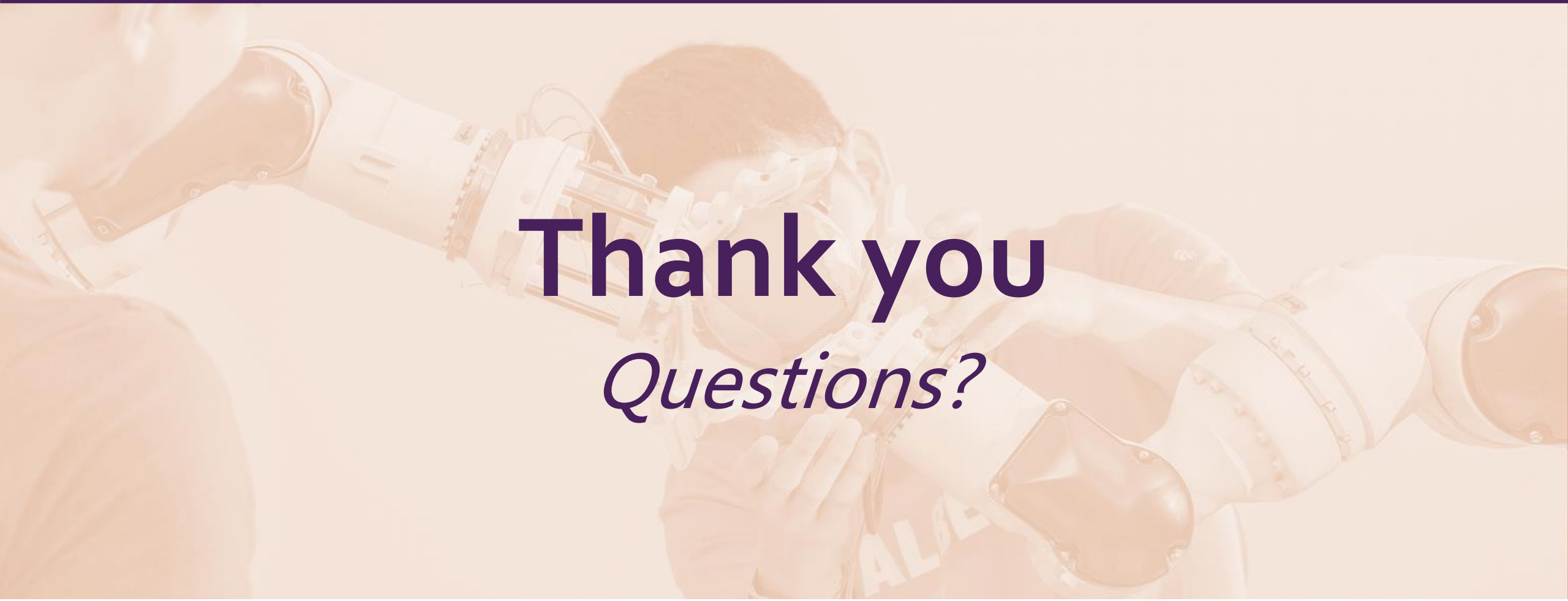
IIMB Results Excerpt



Summary

- Method for performing hierarchical clustering on knowledge graph subjects
- Source code and experimental results available at www.github.com/mpietrasik/smich
- Feel free to email me about specifics at pietrasi@ualberta.ca

AUTONOMOUS SYSTEMS INITIATIVE



Thank you

Questions?