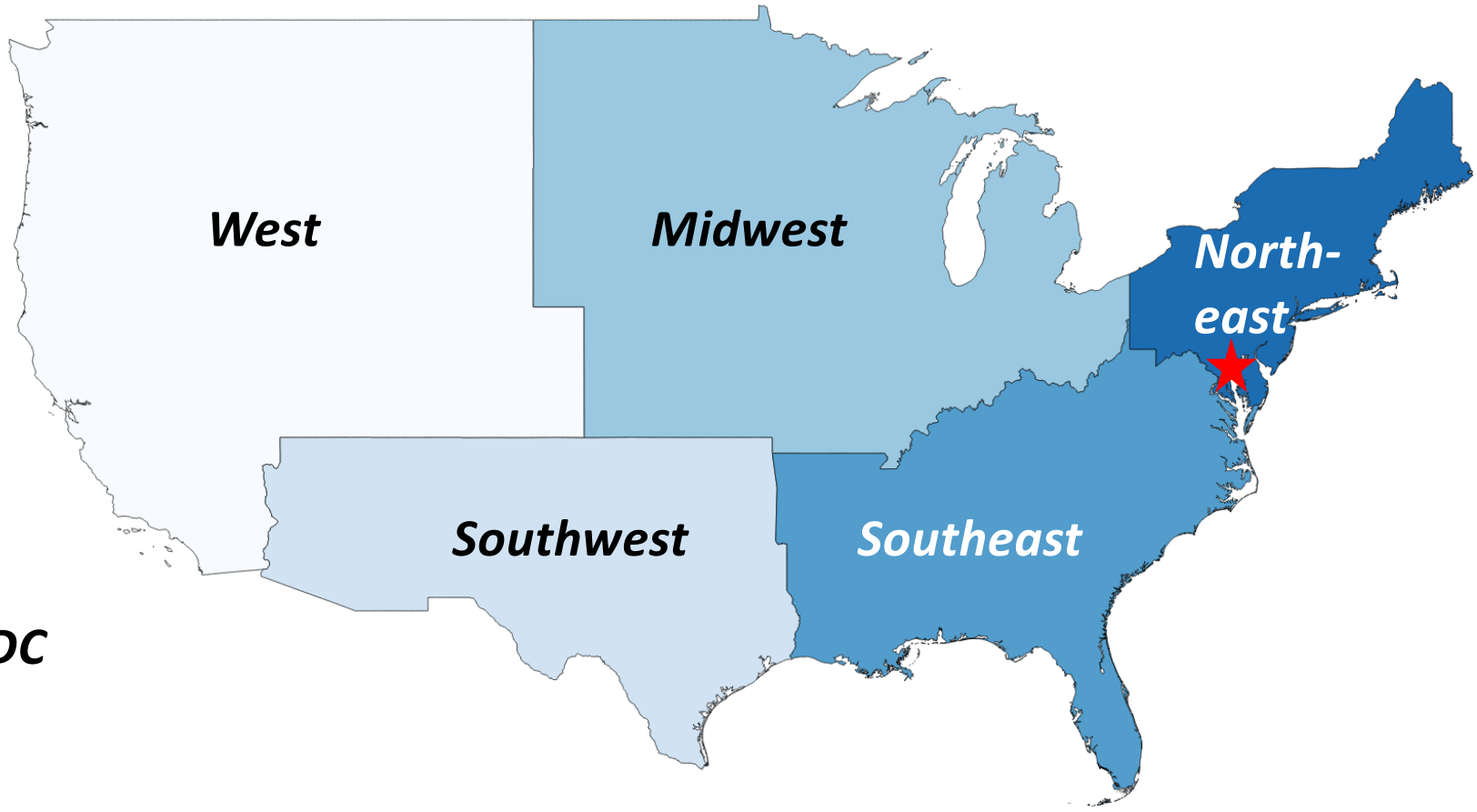
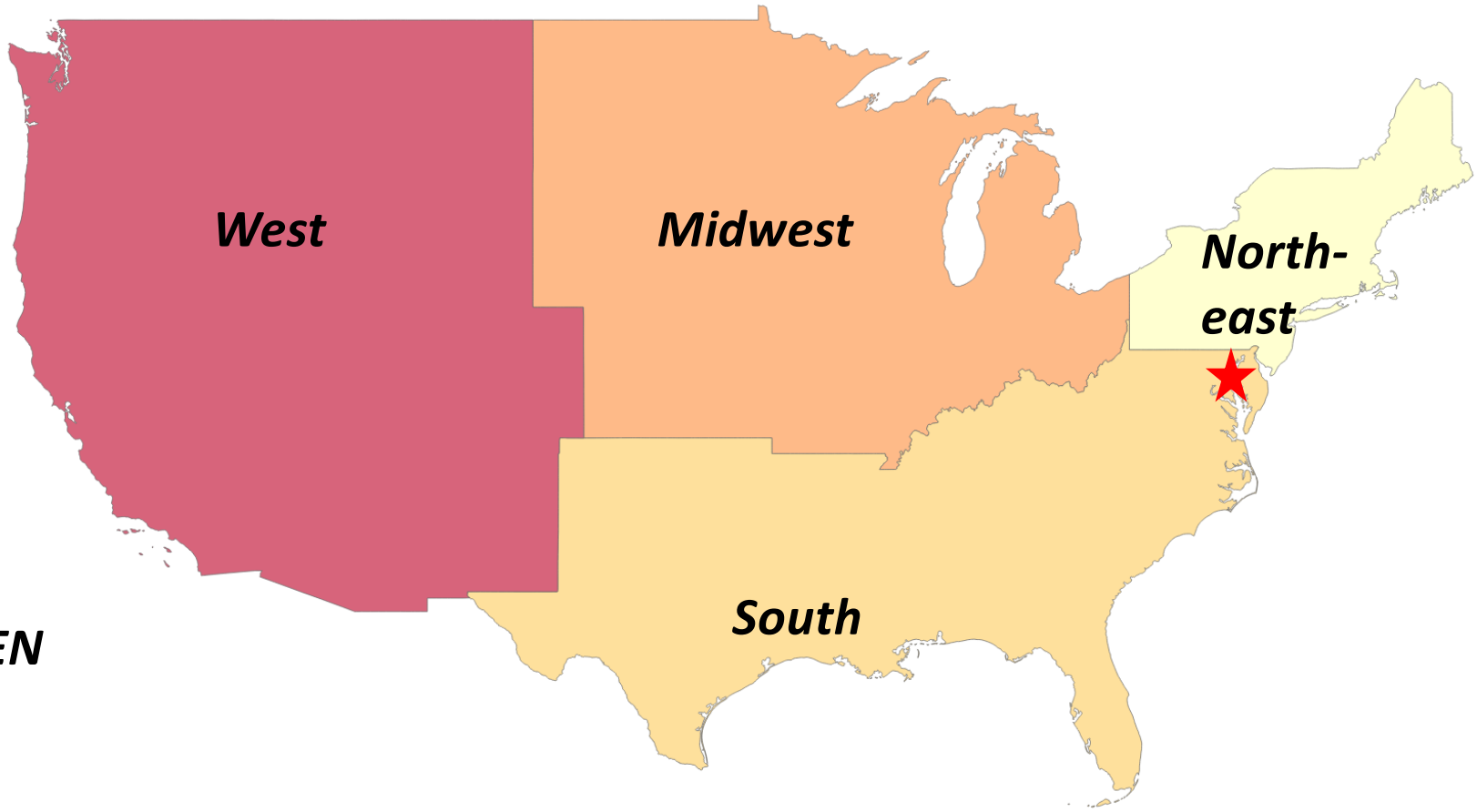
Theory & Practice of Data Cleaning

From Controlled Vocabularies to Taxonomy Alignment

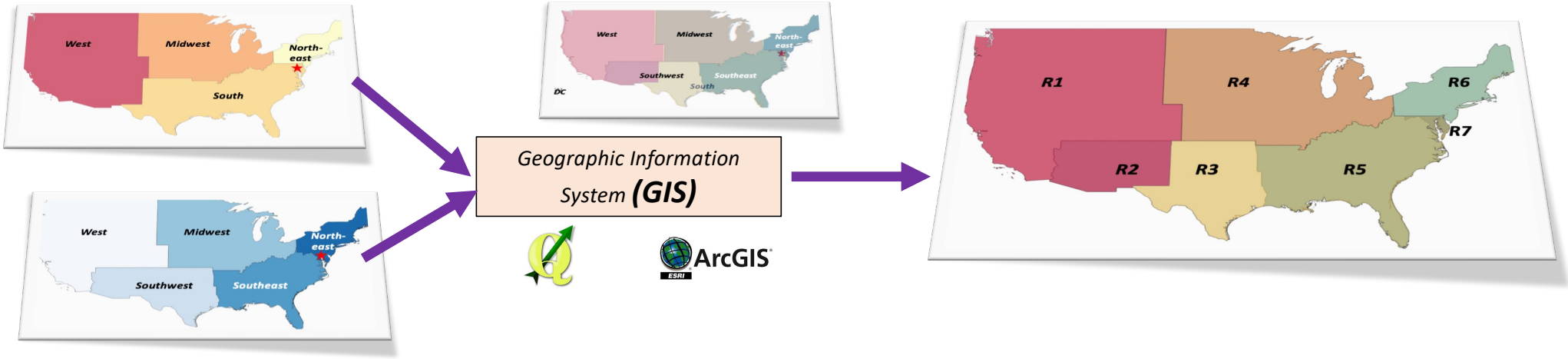
Controlled Vocabularies & Taxonomies

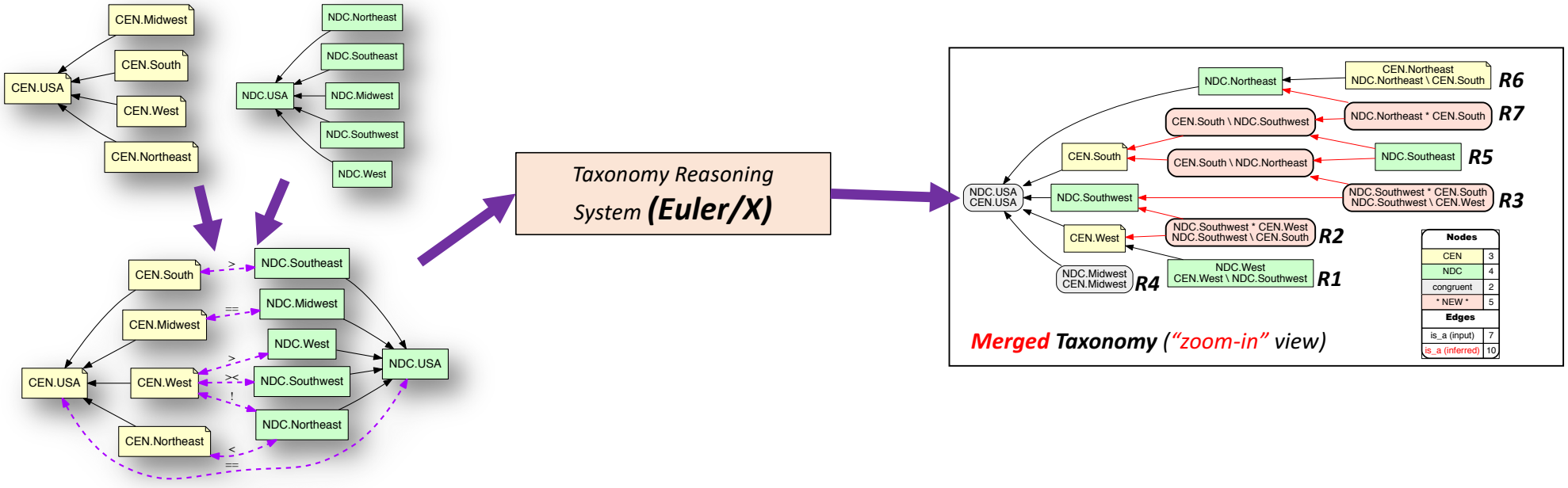
- **Tina**: Hey, any recommended signature dish from where you live?
- **Amy**: Oh, definitely the half-smokes from the Northeast! They are these tasty half-pork and half-beef sausages.
- **Tina**: What a coincidence! We have half-smokes in the South, too! Where do you live in the Northeast? New York? Boston?
- **Amy**: Wrong guesses! Where do you live in the South?
- **Tina and Amy together**: Washington, D.C. [The two of them look at each other, confused.]


West
Midwest
North-
east
Southwest
Southeast
NDC



CEN


West
Midwest
North-east
South
Southwest
Southeast
DC
Geographic Information System (GIS)
ArcGIS
ESRI
R1
R2
R3
R4
R5
R6
R7


CEN.Midwest
CEN.South
CEN.USA
CEN.West
CEN.Northeast
NDC.Northeast
NDC.Southeast
NDC.USA
NDC.Midwest
NDC.Southwest
NDC.West
Taxonomy Reasoning System (Euler/X)
NDC.USA
CEN.USA
NDC.Northeast
CEN.Northeast \ NDC.Northeast \ CEN.South
R6
NDC.Northeast * CEN.South
R7
CEN.South \ NDC.Southwest
CEN.South
CEN.South \ NDC.Northeast
NDC.Southeast
R5
NDC.Southwest
NDC.Southwest * CEN.South
NDC.Southwest \ CEN.West
R3
CEN.West
NDC.Southwest * CEN.West
NDC.Southwest \ CEN.South
R2
NDC.Midwest
CEN.Midwest
R4
NDC.West
CEN.West \ NDC.Southwest
R1
Nodes
CEN 3
NDC 4
congruent 2
* NEW * 5
Edges
is_a (input) 7
is_a (inferred) 10
Merged Taxonomy ("zoom-in" view)

Overview

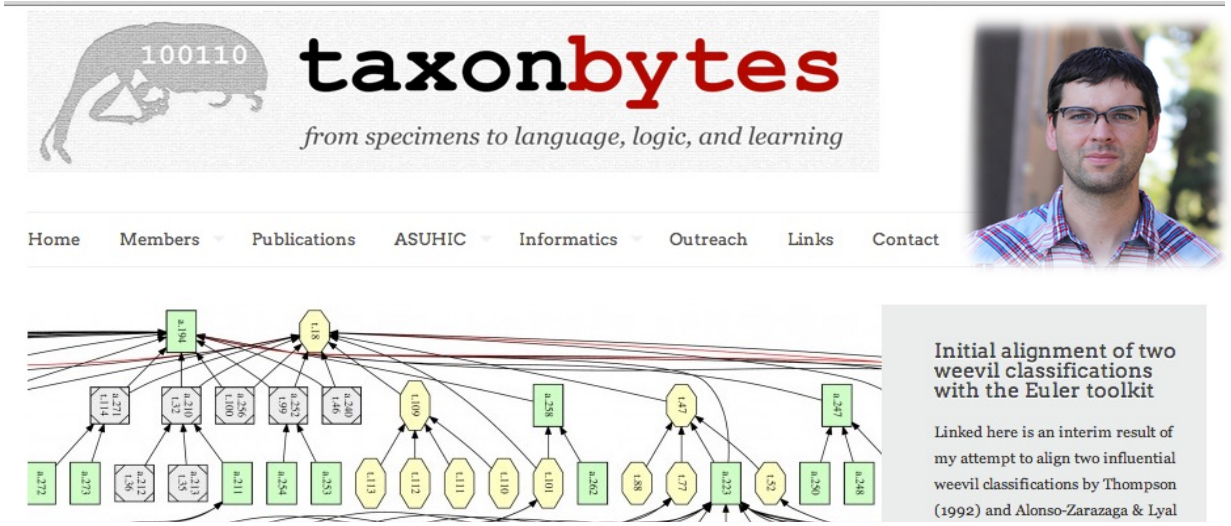
- The **Taxonomy Alignment Problem**
  $T_1 + T_2 + A \Rightarrow T_3$ (*ambiguous .. unique .. **inconsistent***)
- Model-based **Diagnosis** [Reiter'87]
  - Black-box
- Hybrid Approach
  - Black-box & White-box combined

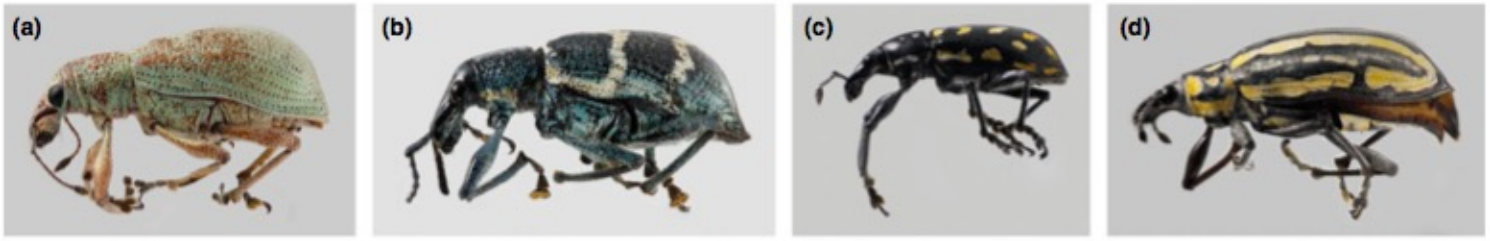
Optional Readings:

Cheng, Yi-Yun, Nico Franz, Jodi Schneider, Shizhuo Yu, Thomas Rodenhausen, and Bertram Ludäscher. "**Agreeing to Disagree: Reconciling Conflicting Taxonomic Views Using a Logic-Based Approach**." *Proceedings of the Association for Information Science and Technology* 54, no. 1 (January 1, 2017): 46–56. https://doi.org/10.1002/pra2.2017.14505401006.

Chen, Mingmin, Shizhuo Yu, Nico Franz, Shawn Bowers, and Bertram Ludäscher. "**A Hybrid Diagnosis Approach Combining Black-Box and White-Box Reasoning**." In *Rules on the Web. From Theory to Applications*, edited by Antonis Bikakis, Paul Fodor, and Dumitru Roman, 127–41. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-09870-8_9.

Meet Prof. Nico Franz: Curator of Insects @ ASU

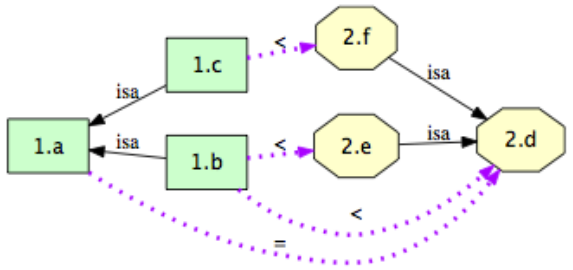




What Nico *et al.* do for a living …

What Nico does for a living (cont'd): The Indoors Part

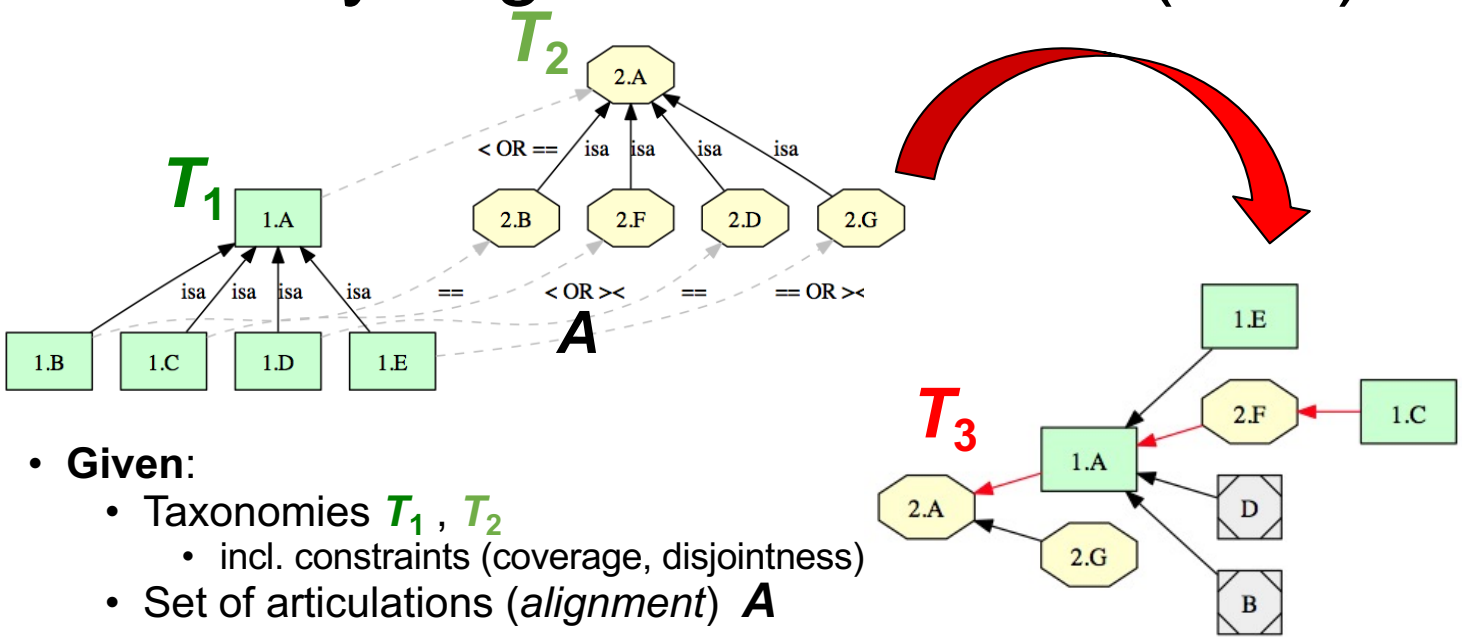
- First: go fun places, find new bugs, study them …
  - "Bugs-R-Us" (see **taxonbytes.org**)
- **Now**: Compare, **align** and **revise** taxonomies, based on careful observation, "character" data, expertise …
- **Formally:**
  - **Input**: T_1 + T_2 (***taxonomies***) + ***A*** (*expert articulations*)



Taxonomy T_1	Taxonomy T_2	Articulations
r_1: b isa a	r_5: e isa d	A_1: a = d
r_2: c isa a	r_6: f isa d	A_2: b $\subsetneq$ e
r_3: a = $b \cup c$	r_7: d = $e \cup f$	A_3: c $\subsetneq$ f
r_4: b $\cap$ c = $\emptyset$	r_8: e $\cap$ f = $\emptyset$	A_4: b $\subsetneq$ d

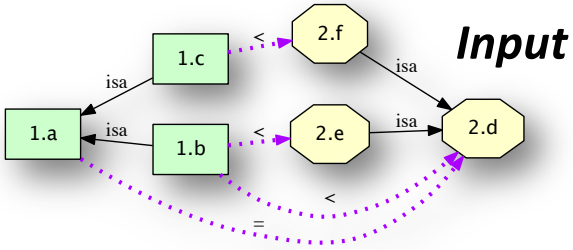
  - **Output**: revised, "merged" taxonomy (-ies) T_3

Taxonomy Alignment Problem (TAP)

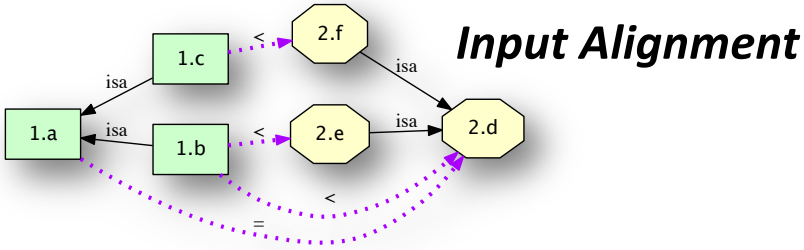


- **Given**:
  - Taxonomies T_1 , T_2
    - incl. constraints (coverage, disjointness)
  - Set of articulations (*alignment*) ***A***
- **Find**:
  - **Combined ("merged") taxonomy T_3 (= T_1 + T_2 + *A*)**
    - **Is it a taxonomy? Or a DAG?**
  - Optional:
    - Final alignment (should be minimal)

TAP: Possible Outcomes



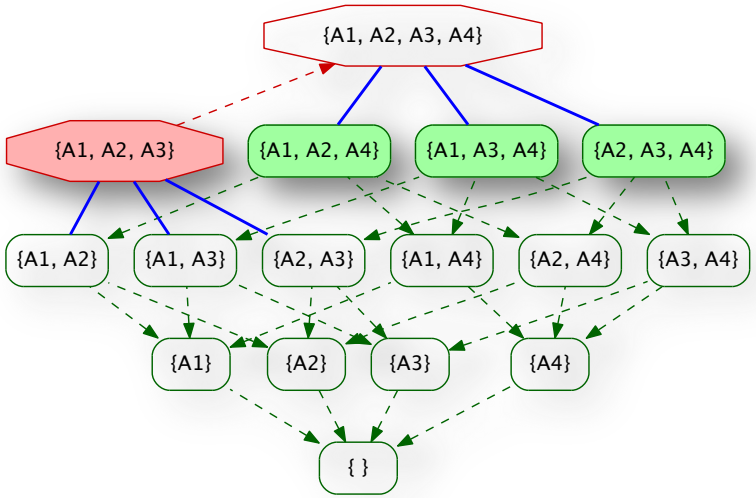
TAP: Possible Outcomes



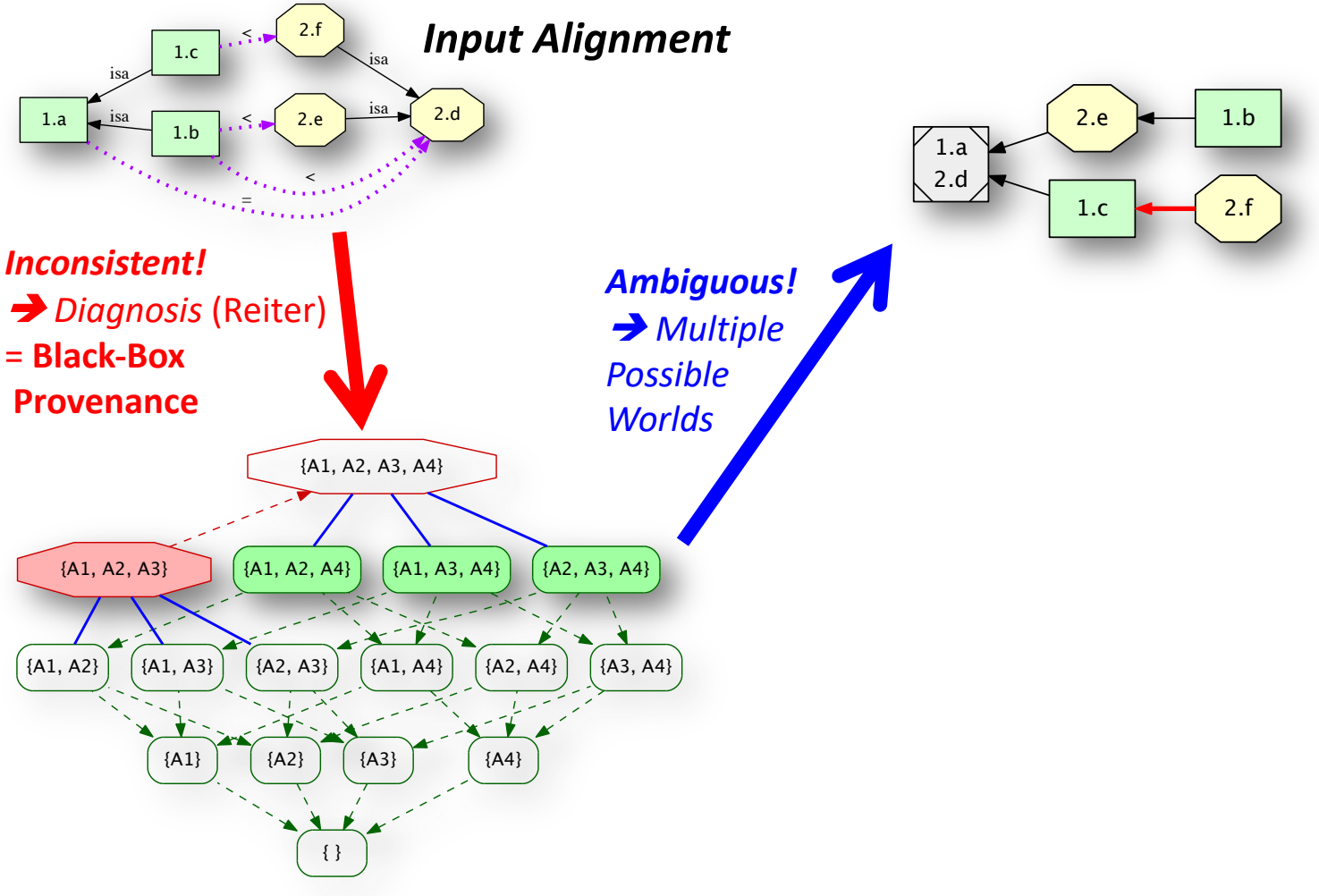
Inconsistent!

➔ *Diagnosis* (Reiter)

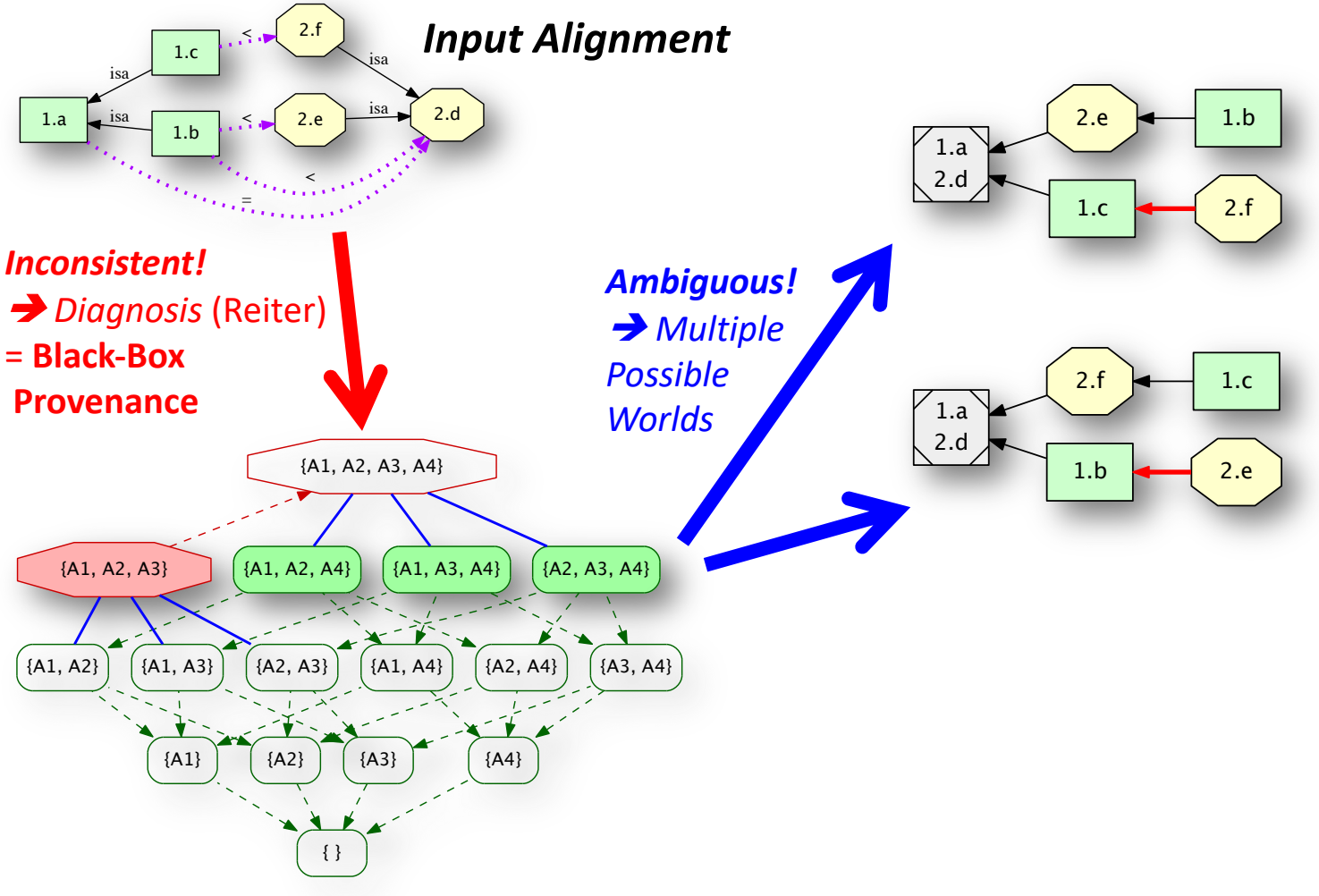
= Black-Box Provenance



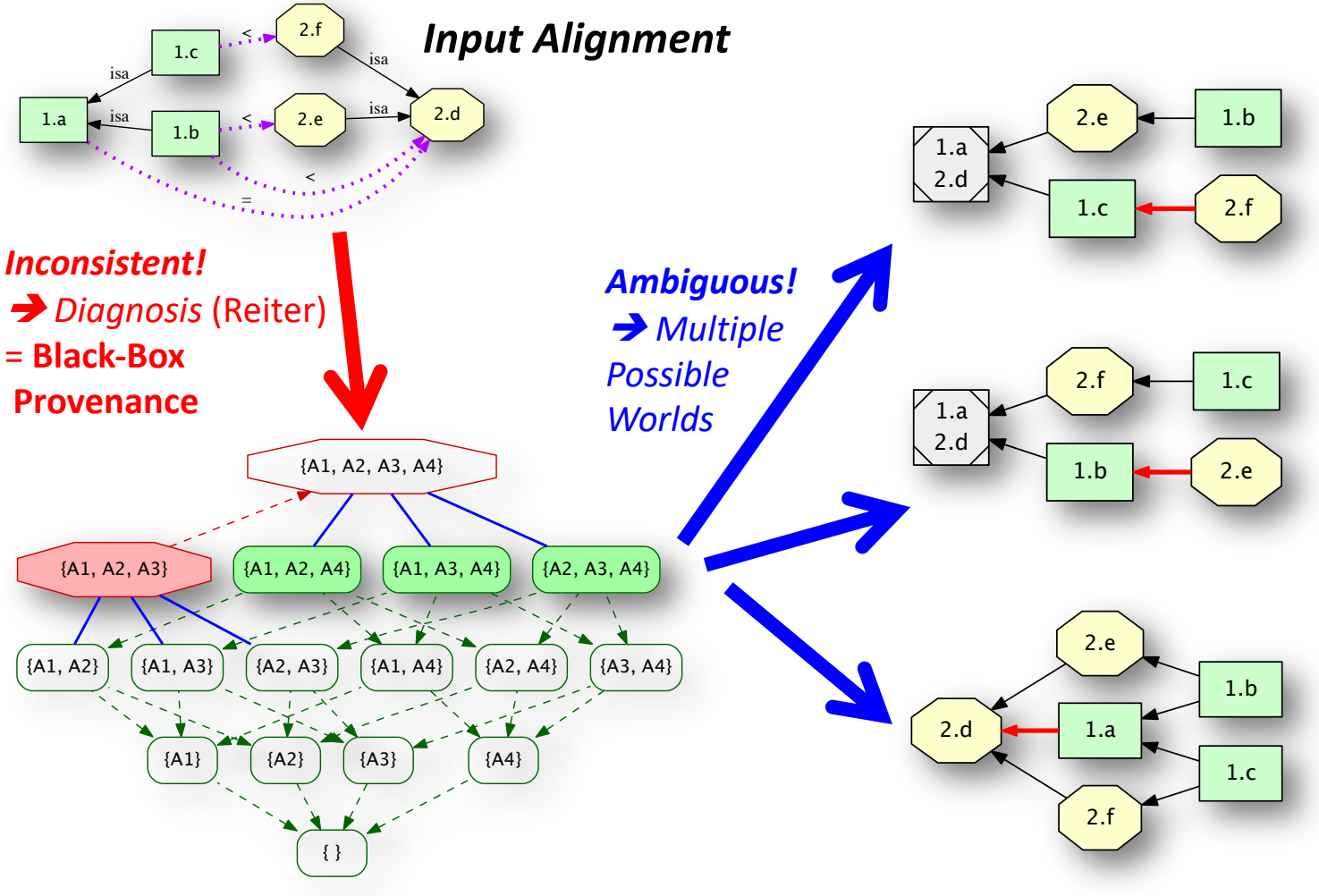
TAP: Possible Outcomes



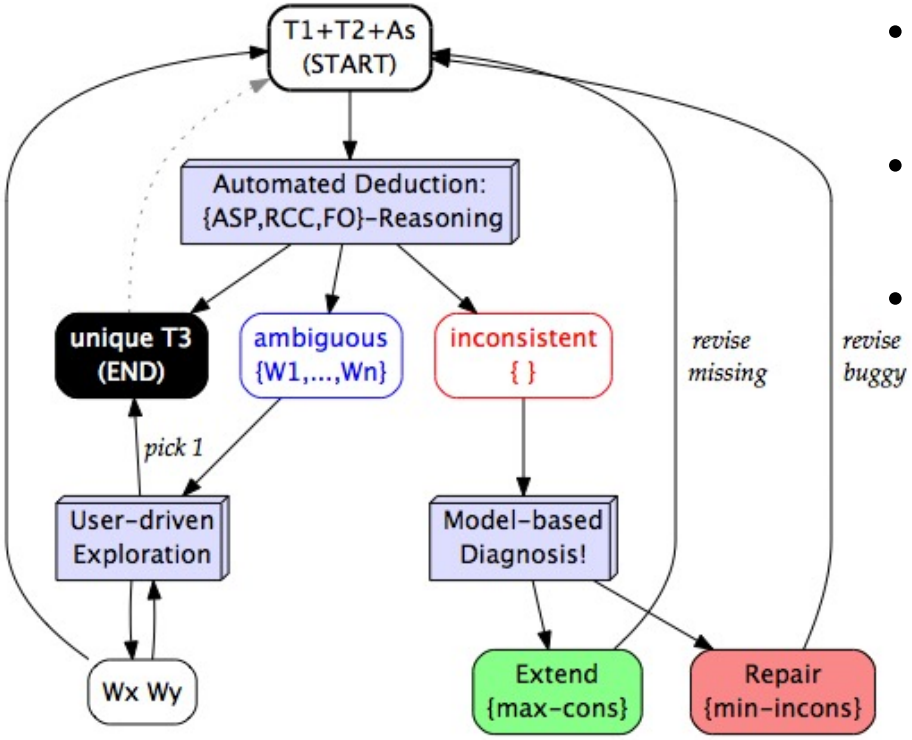
TAP: Possible Outcomes



TAP: Possible Outcomes

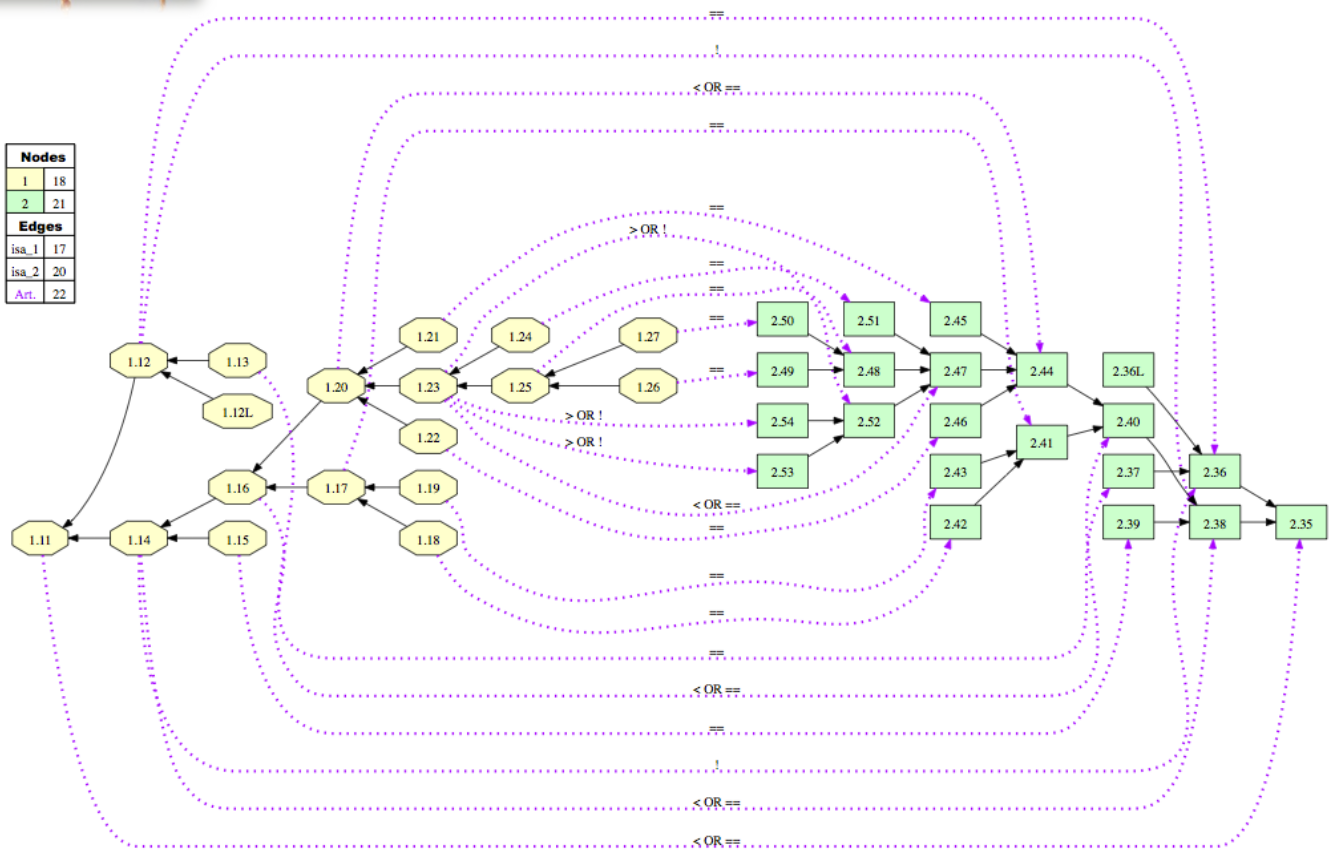


Euler/X Toolkit and Workflow

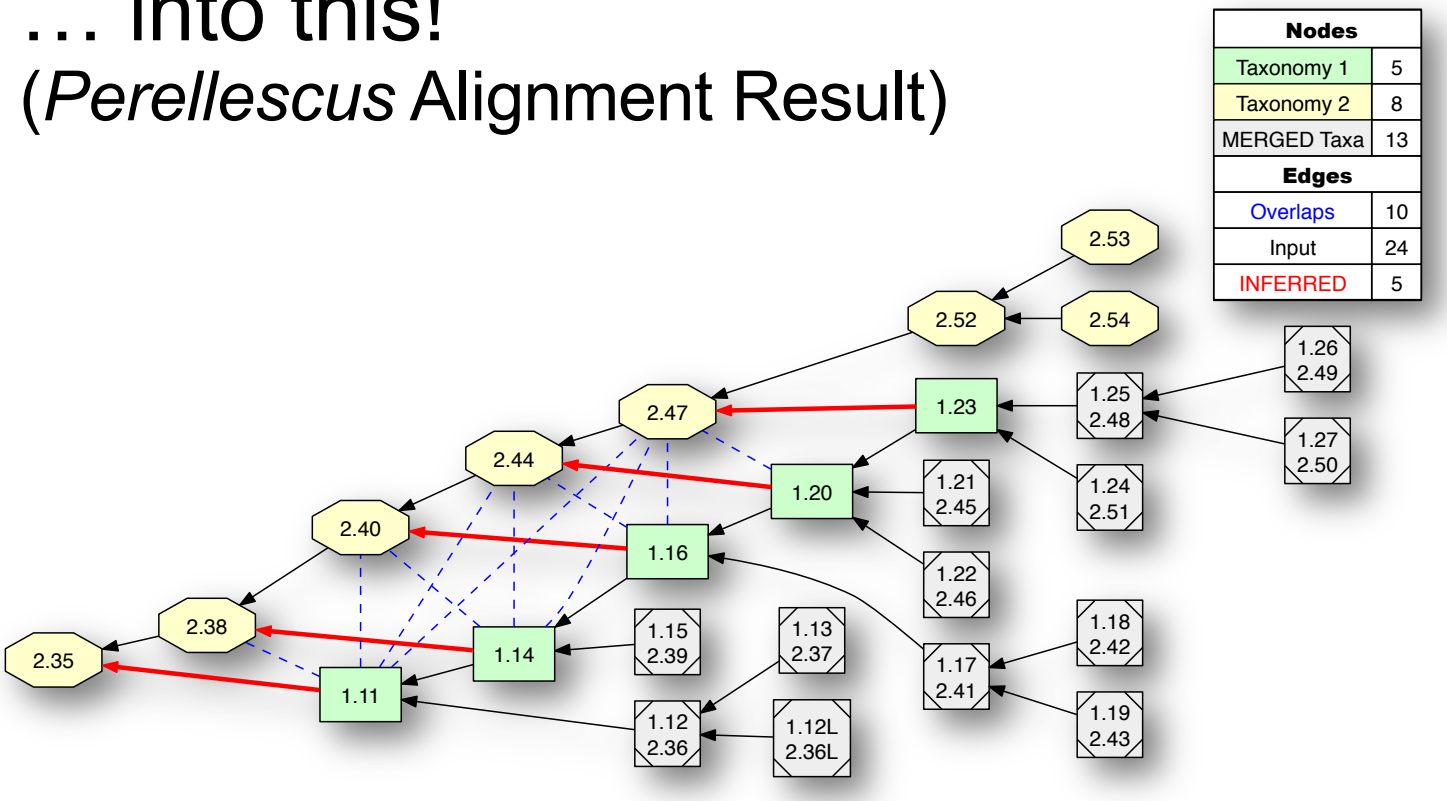


- **FO reasoning** about taxonomies (MFOL)
- Earlier: ***CleanTax***
  - Prover9/Mace4
- Now: ***Euler***
  - **ASP** Reasoners (DLV, Clingo)
  - **Specialized** reasoners (PyRCC)
  - …
  - **X** = ASP, RCC, …

Real-world examples: Turn this …

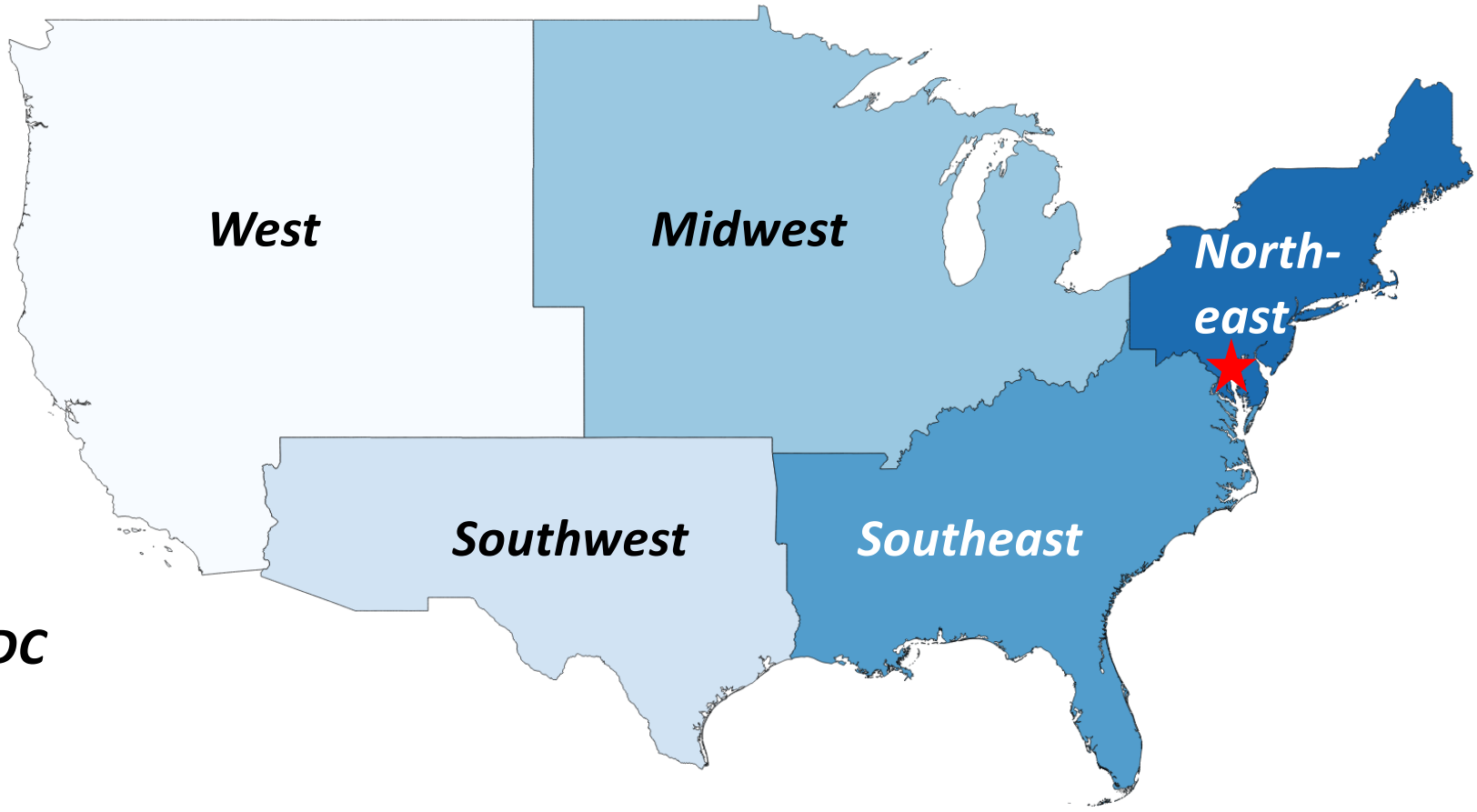


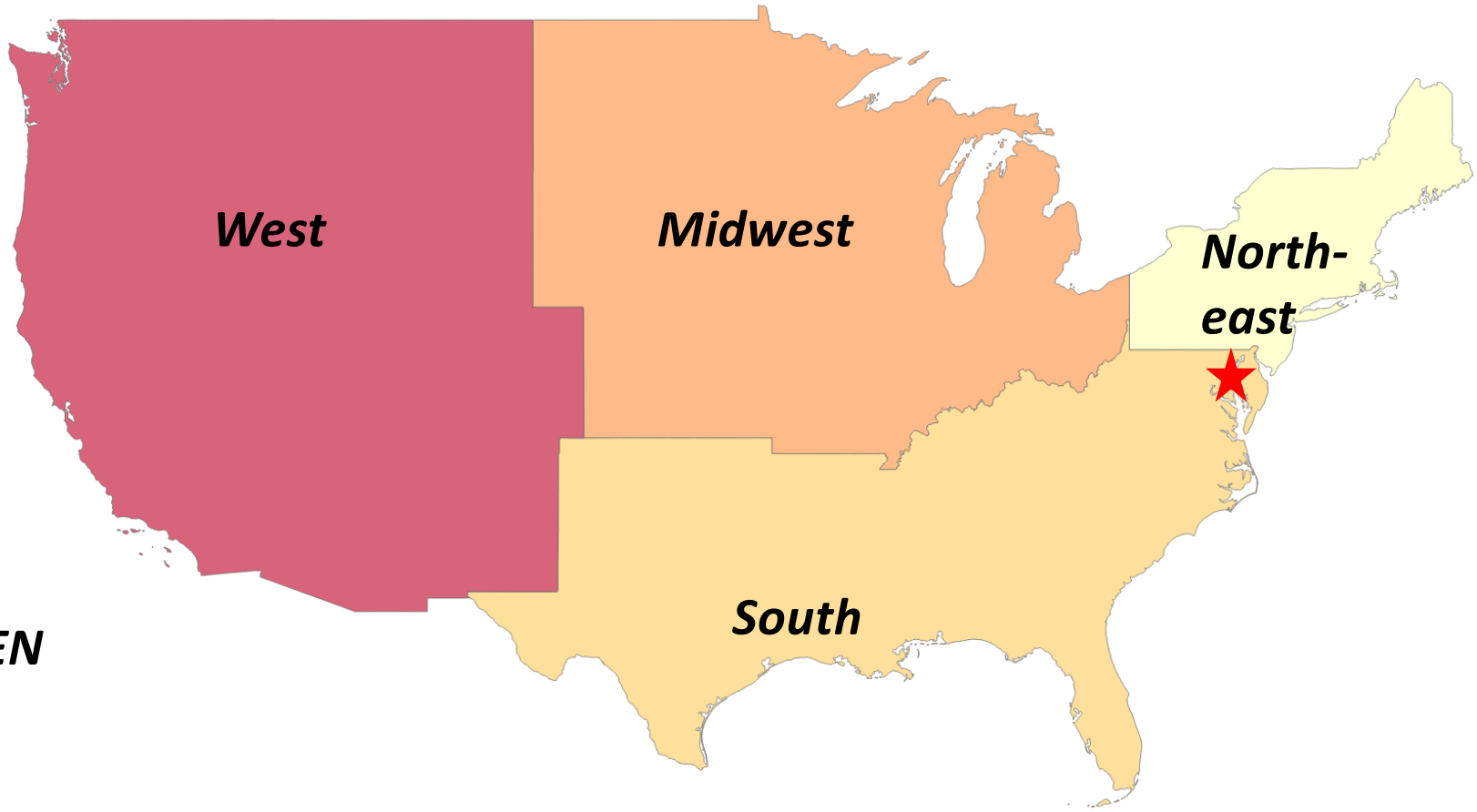
… into this!
(*Perellescus* Alignment Result)



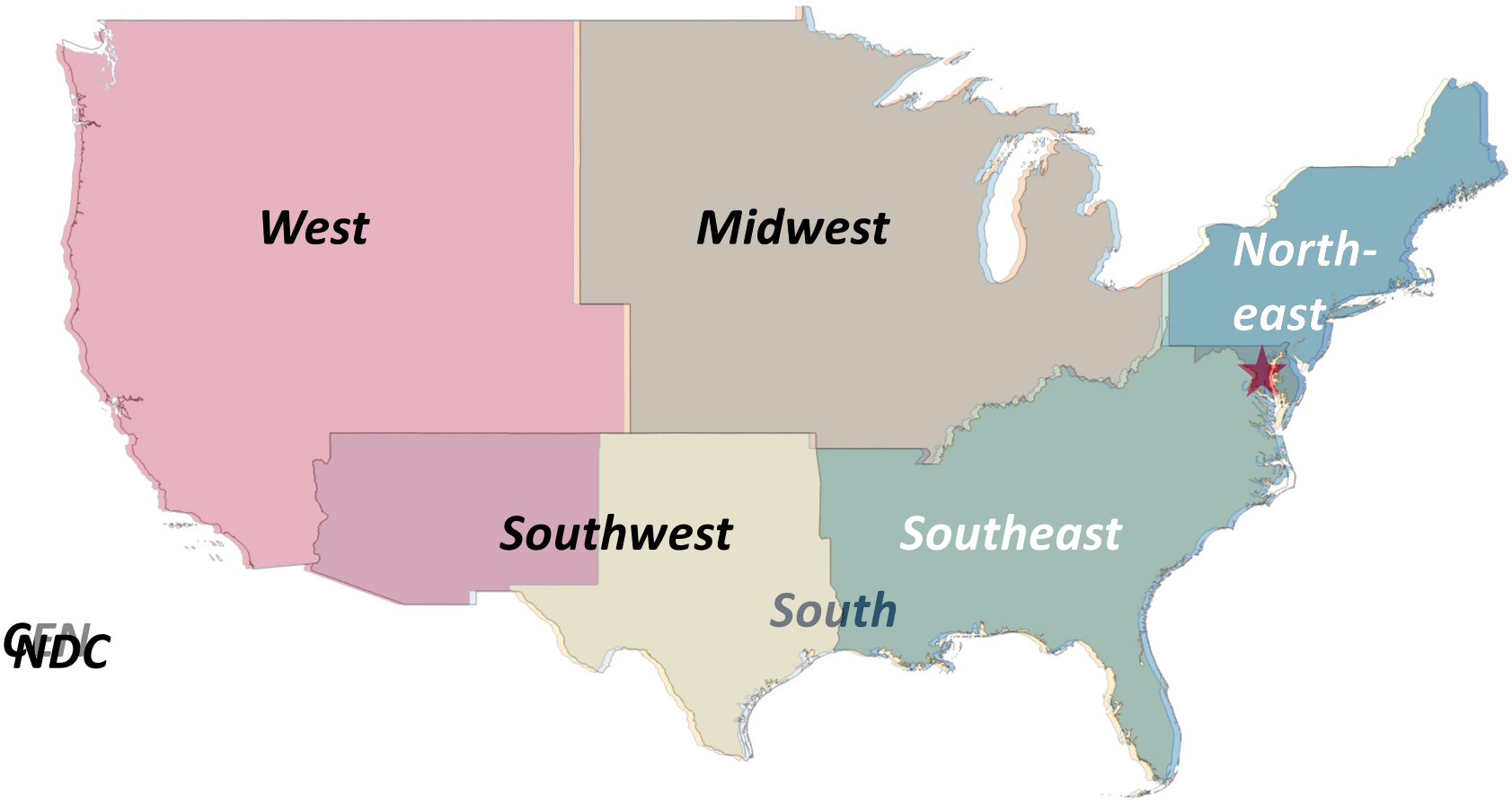
- T_3 := T_1 and T_2 are "merged"
  - Blue dashed: overlaps ➔ resolve via "zoom-in view"

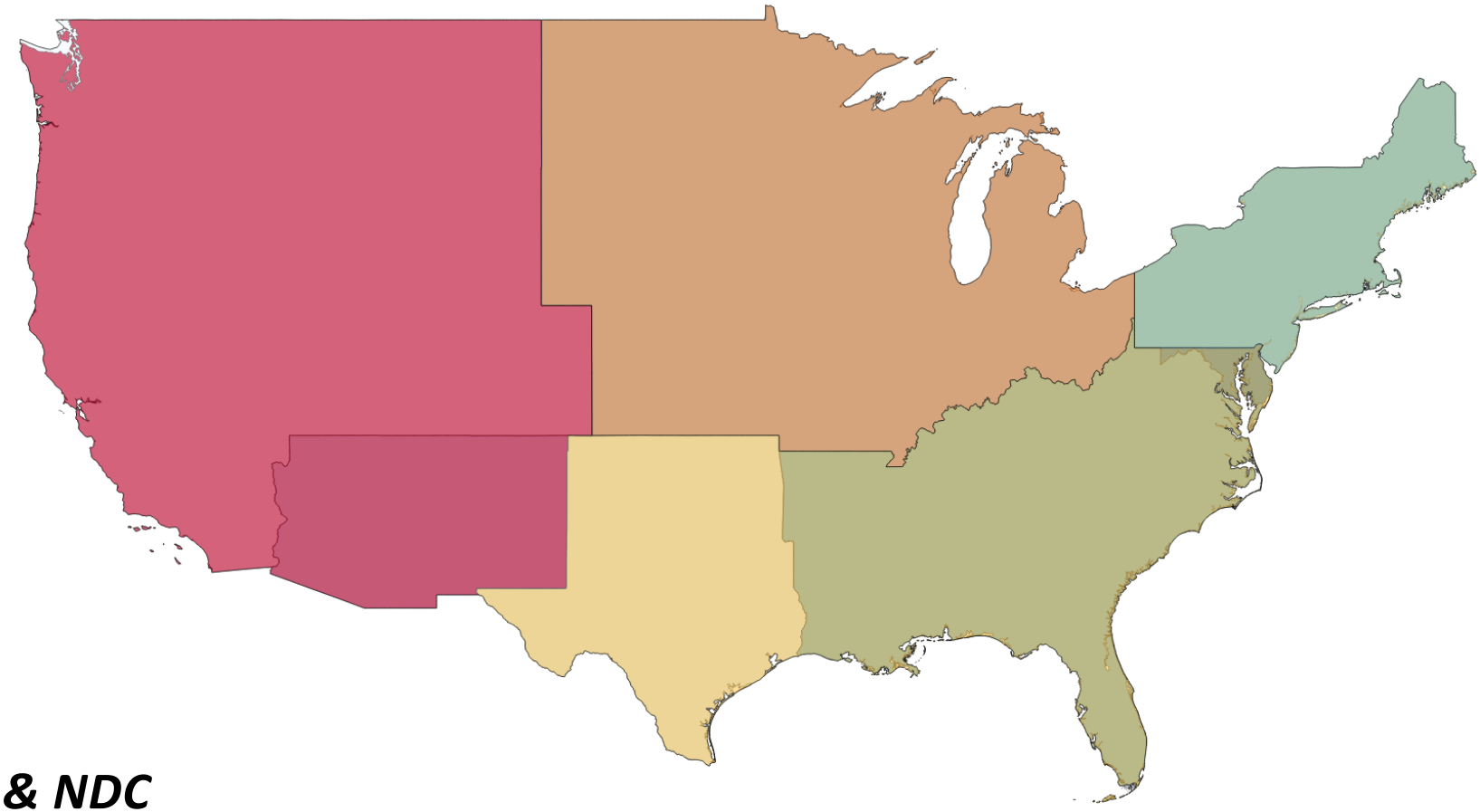
CEN-NDC-Regions
Alignment


West
Midwest
North-
east
Southwest
Southeast
NDC

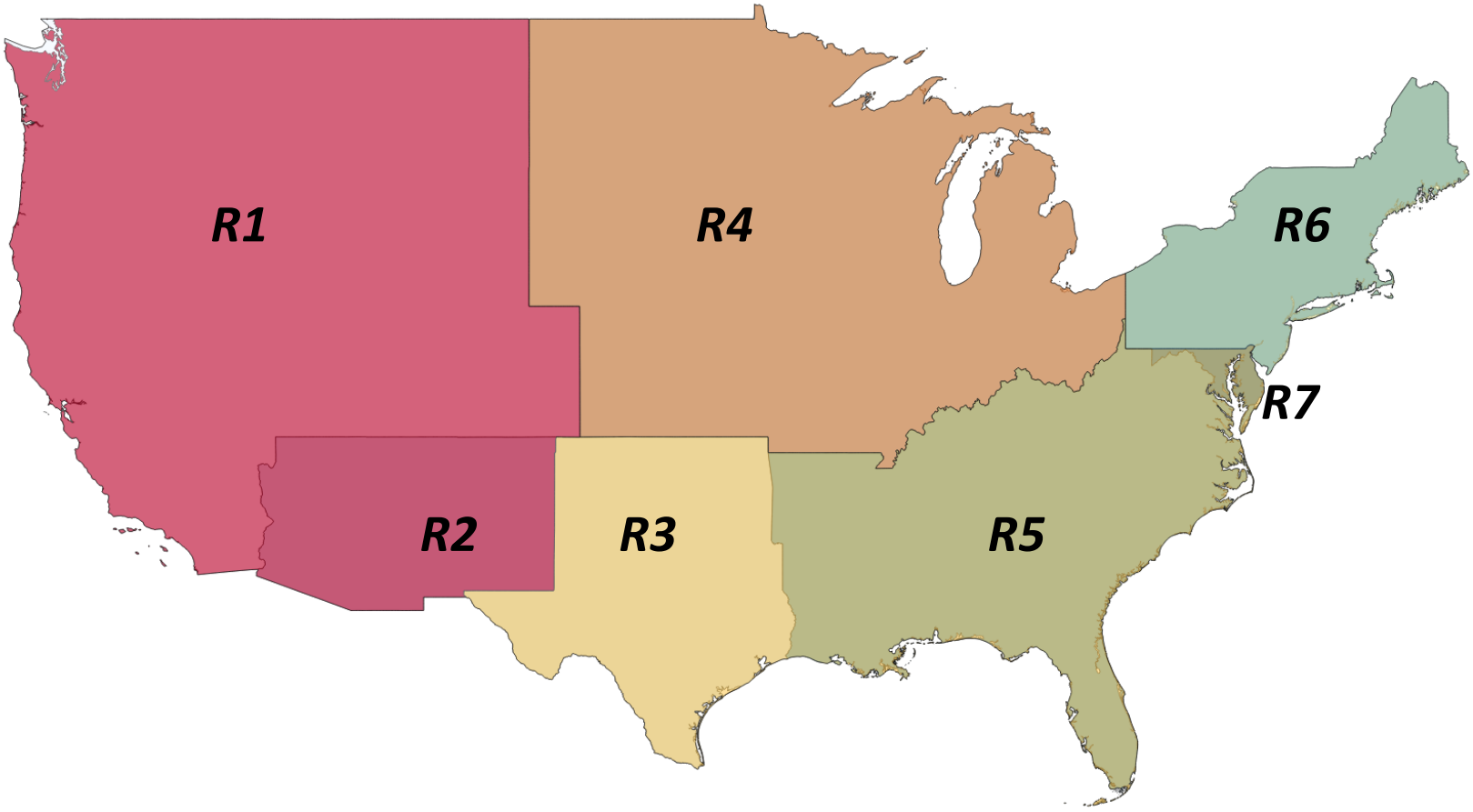


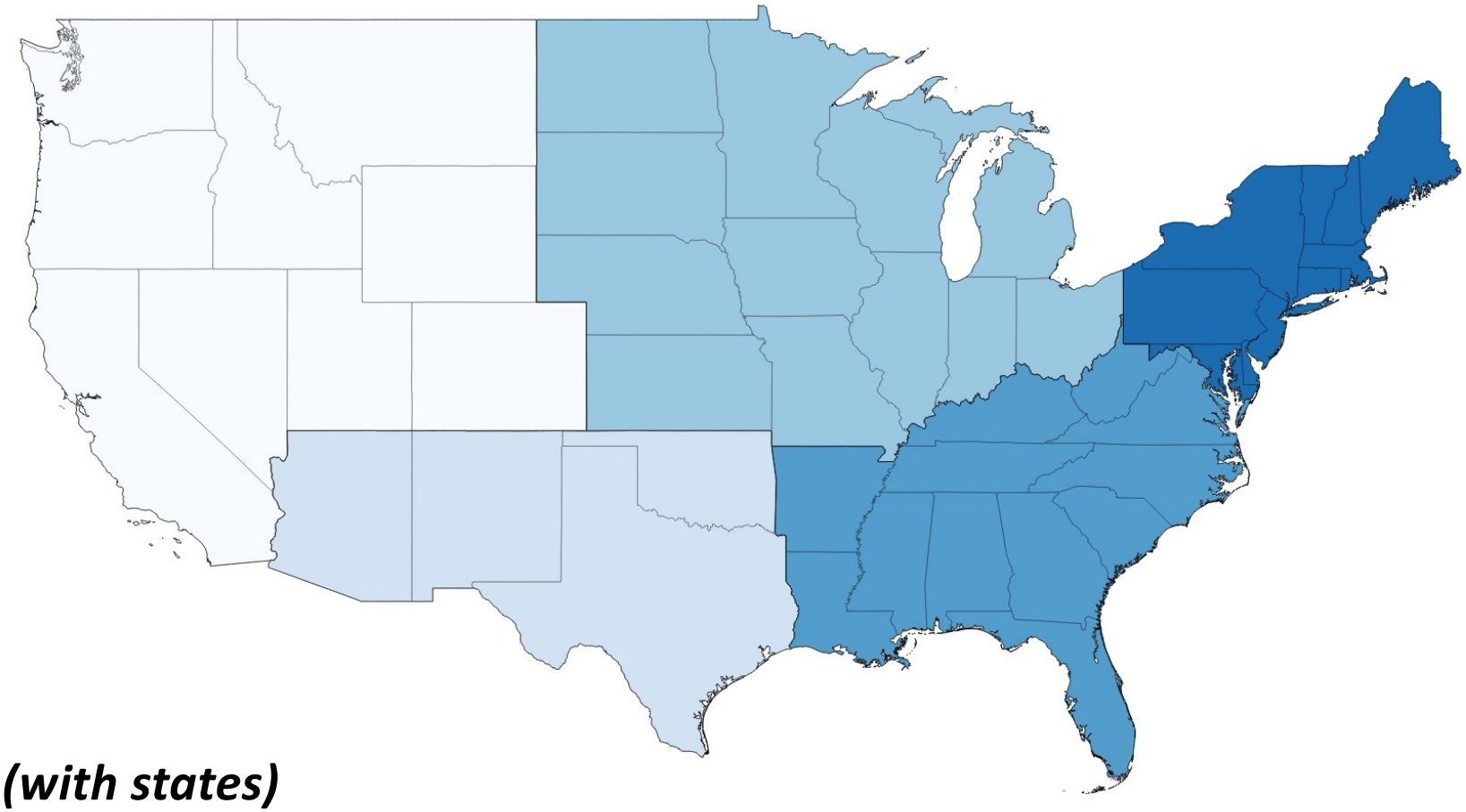
CEN


West
Midwest
North-
east
Southwest
Southeast
South
GFN
NDC

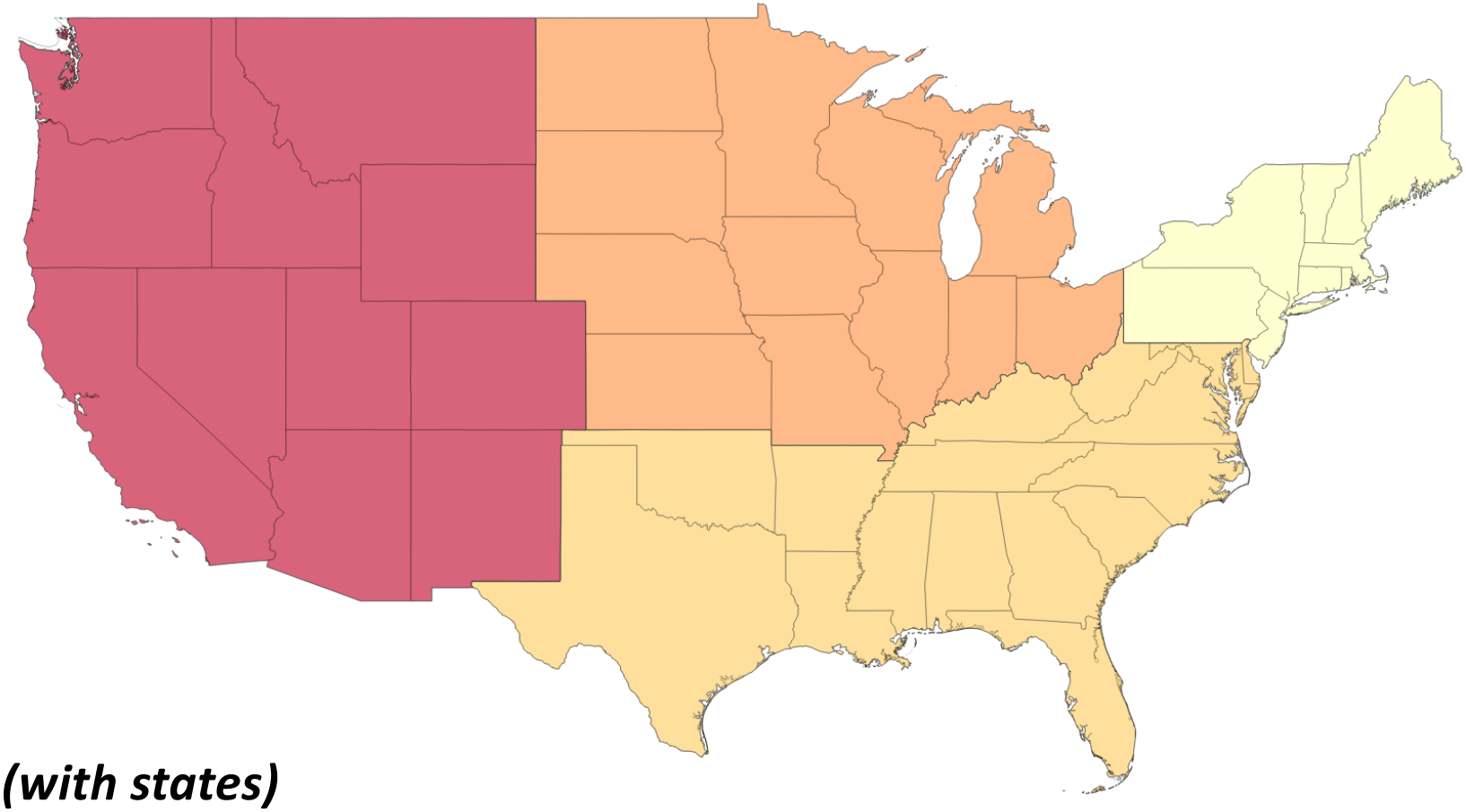


CEN & NDC

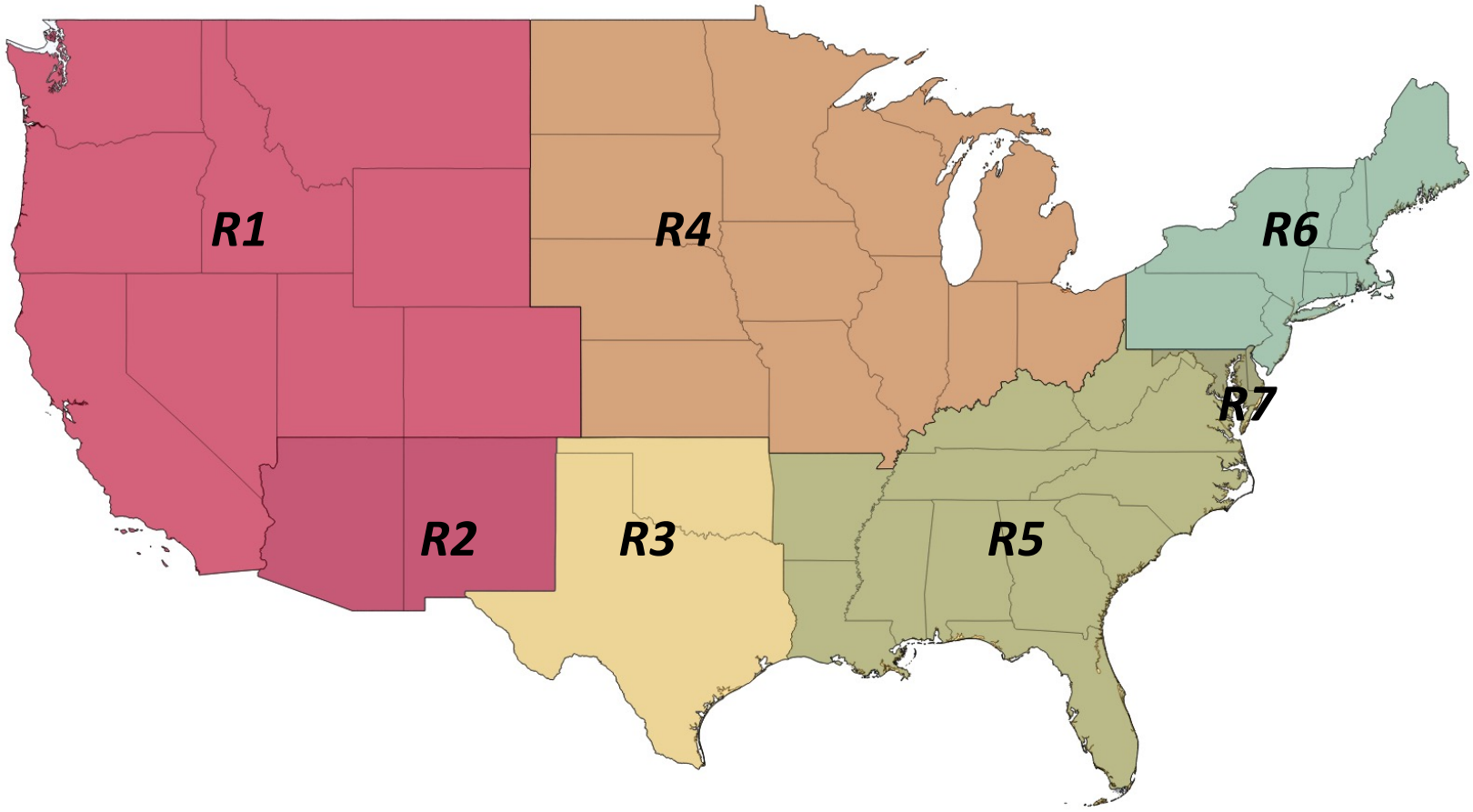

R1
R4
R6
R7
R2
R3
R5

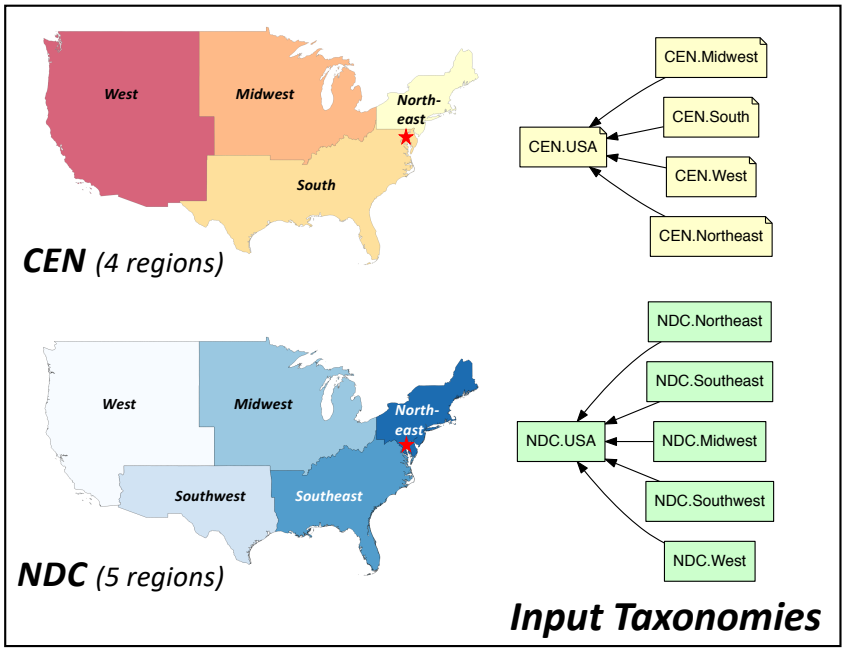


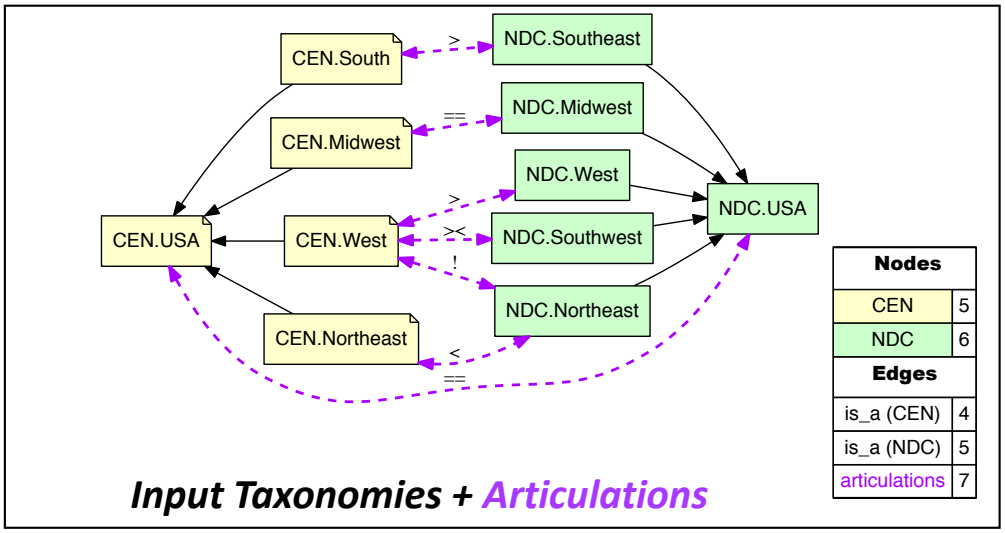
NDC (with states)

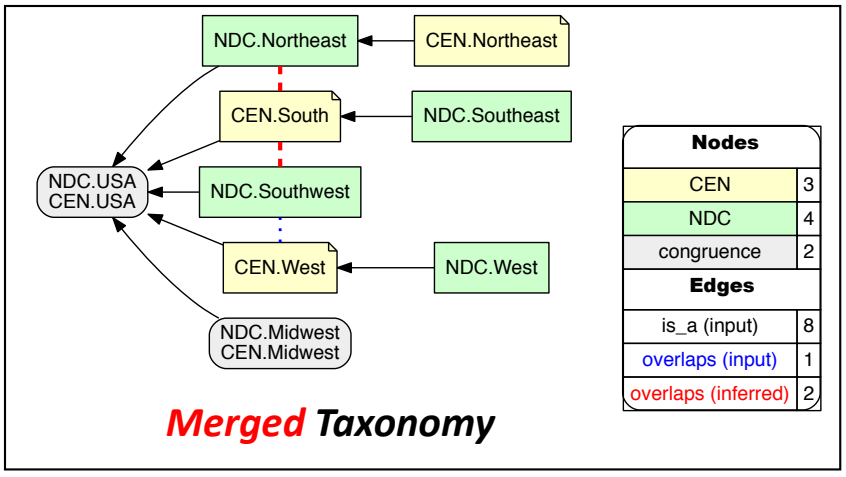


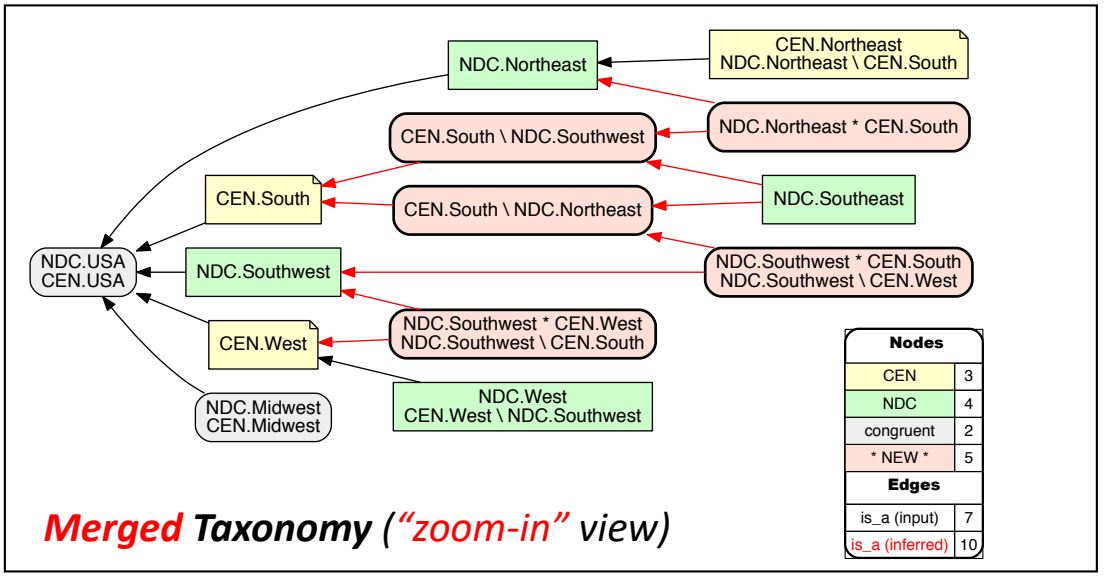
CEN (with states)

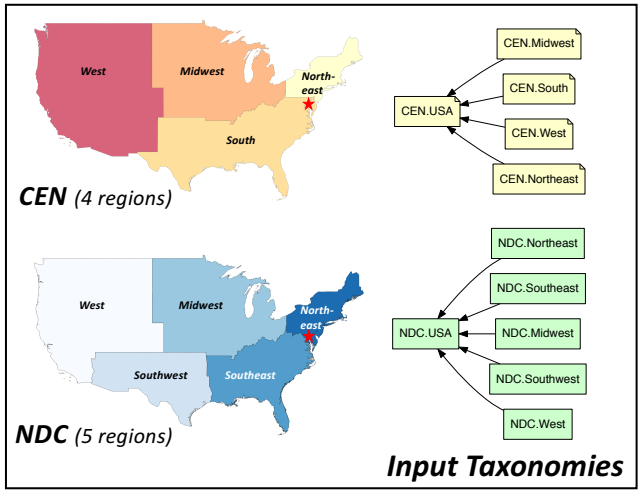

R1
R4
R6
R7
R2
R3
R5

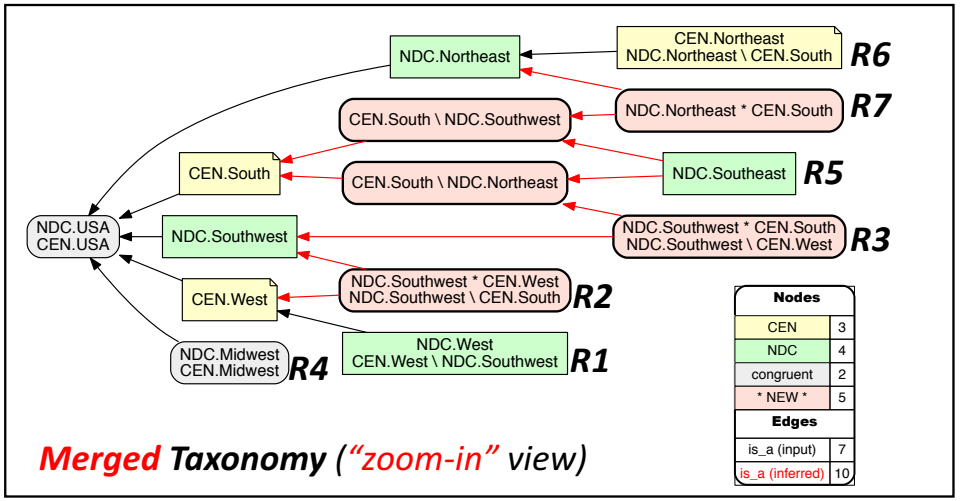

West
Midwest
North-east
South
CEN (4 regions)
CEN.Midwest
CEN.South
CEN.USA
CEN.West
CEN.Northeast
West
Midwest
North-east
Southwest
Southeast
NDC (5 regions)
NDC.Northeast
NDC.Southeast
NDC.USA
NDC.Midwest
NDC.Southwest
NDC.West
Input Taxonomies

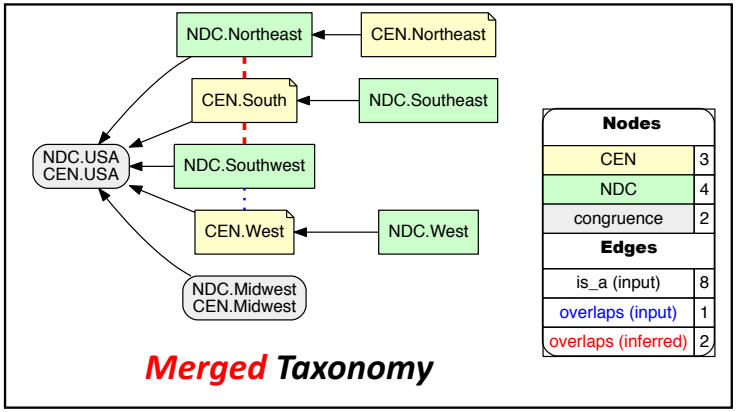

CEN.South
NDC.Southeast
CEN.Midwest
NDC.Midwest
NDC.West
NDC.USA
CEN.USA
CEN.West
NDC.Southwest
NDC.Northeast
CEN.Northeast
Input Taxonomies + Articulations
Nodes
CEN 5
NDC 6
Edges
is_a (CEN) 4
is_a (NDC) 5
articulations 7

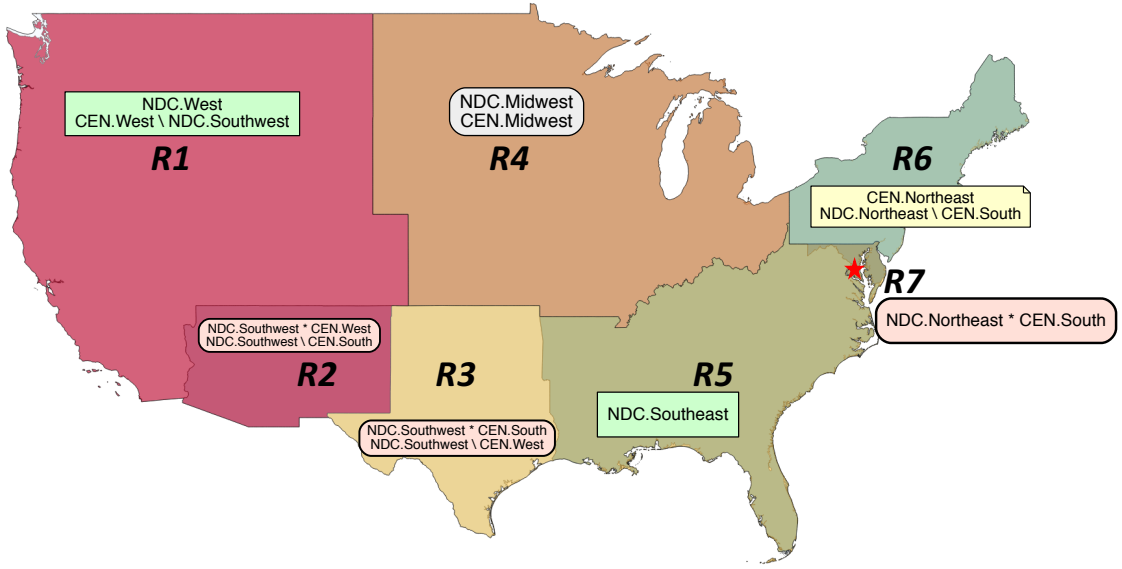

NDC.Northeast
CEN.Northeast
CEN.South
NDC.Southeast
NDC.USA
CEN.USA
NDC.Southwest
CEN.West
NDC.West
NDC.Midwest
CEN.Midwest
Merged Taxonomy
Nodes
CEN 3
NDC 4
congruence 2
Edges
is_a (input) 8
overlaps (input) 1
overlaps (inferred) 2

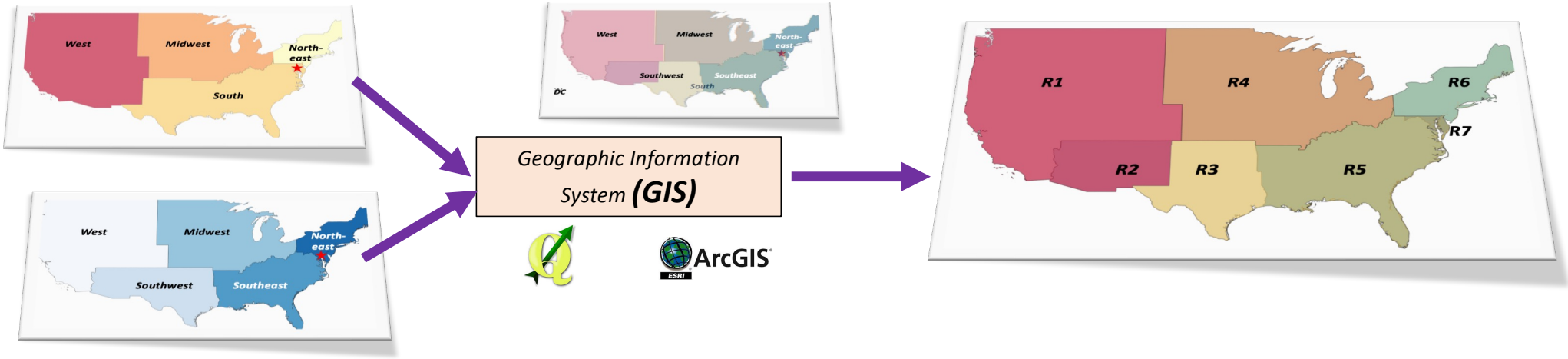

NDC.Northeast
CEN.Northeast
NDC.Northeast \ CEN.South
CEN.South \ NDC.Southwest
NDC.Northeast * CEN.South
CEN.South
CEN.South \ NDC.Northeast
NDC.Southeast
NDC.USA
CEN.USA
NDC.Southwest
NDC.Southwest * CEN.South
NDC.Southwest \ CEN.West
NDC.Southwest * CEN.West
NDC.Southwest \ CEN.South
CEN.West
NDC.Midwest
CEN.Midwest
NDC.West
CEN.West \ NDC.Southwest
Merged Taxonomy ("zoom-in" view)
Nodes
CEN 3
NDC 4
congruent 2
* NEW * 5
Edges
is_a (input) 7
is_a (inferred) 10

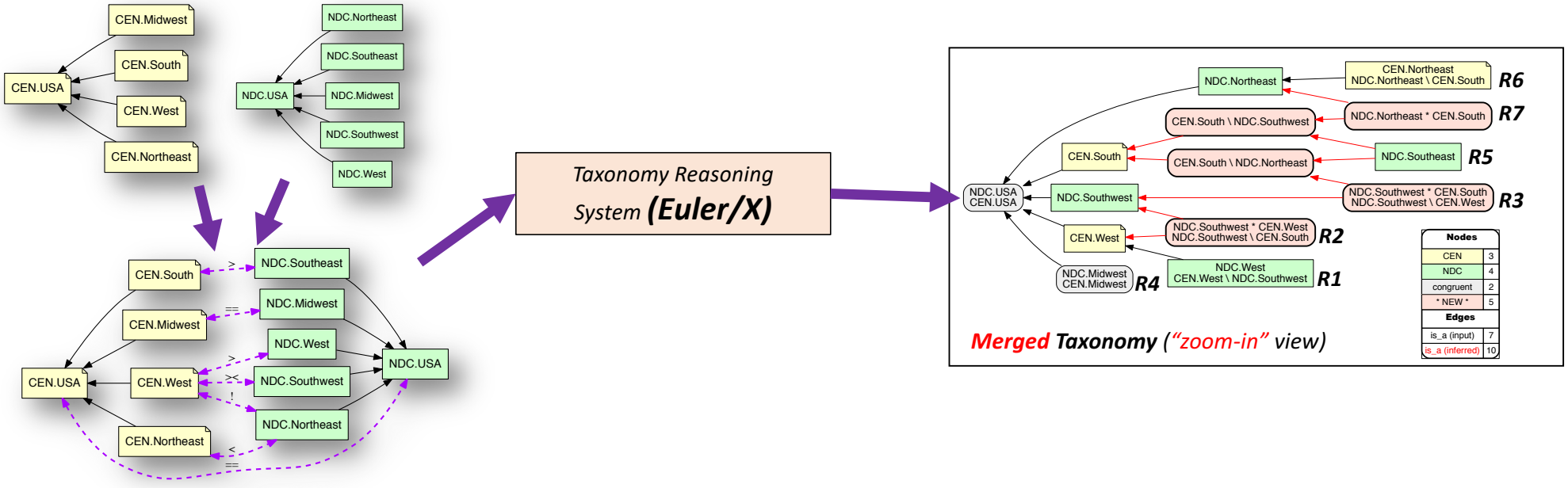

West
Midwest
North-east
South
CEN (4 regions)
CEN.USA
CEN.Midwest
CEN.South
CEN.West
CEN.Northeast
West
Midwest
North-east
Southwest
Southeast
NDC (5 regions)
NDC.USA
NDC.Northeast
NDC.Southeast
NDC.Midwest
NDC.Southwest
NDC.West
Input Taxonomies

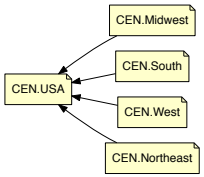

NDC.USA
CEN.USA
NDC.Northeast
CEN.Northeast
NDC.Northeast \ CEN.South
R6
NDC.Northeast * CEN.South
R7
CEN.South
CEN.South \ NDC.Southwest
NDC.Southeast
R5
CEN.South \ NDC.Northeast
NDC.Southwest
NDC.Southwest * CEN.South
NDC.Southwest \ CEN.West
R3
CEN.West
NDC.Southwest * CEN.West
NDC.Southwest \ CEN.South
R2
NDC.Midwest
CEN.Midwest
R4
NDC.West
CEN.West \ NDC.Southwest
R1
Nodes
CEN 3
NDC 4
congruent 2
* NEW * 5
Edges
is_a (input) 7
is_a (inferred) 10
Merged Taxonomy ("zoom-in" view)

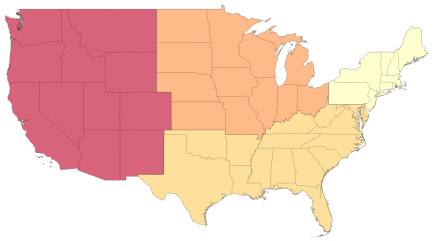

NDC.Northeast
CEN.Northeast
CEN.South
NDC.Southeast
NDC.USA
CEN.USA
NDC.Southwest
CEN.West
NDC.West
NDC.Midwest
CEN.Midwest
Nodes
CEN 3
NDC 4
congruence 2
Edges
is_a (input) 8
overlaps (input) 1
overlaps (inferred) 2
Merged Taxonomy

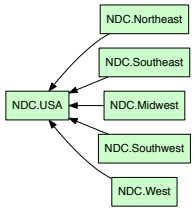

NDC.West
CEN.West \ NDC.Southwest
R1
NDC.Midwest
CEN.Midwest
R4
R6
CEN.Northeast
NDC.Northeast \ CEN.South
R7
NDC.Northeast * CEN.South
NDC.Southwest * CEN.West
NDC.Southwest \ CEN.South
R2
R3
R5
NDC.Southeast
NDC.Southwest * CEN.South
NDC.Southwest \ CEN.West

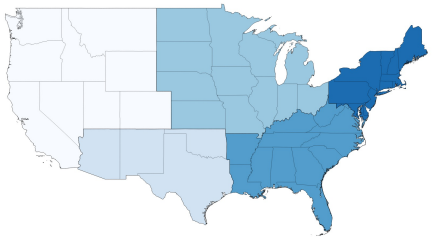

West
Midwest
North-east
South
Southwest
Southeast
DC
Geographic Information System (GIS)
ArcGIS
ESRI
R1
R2
R3
R4
R5
R6
R7

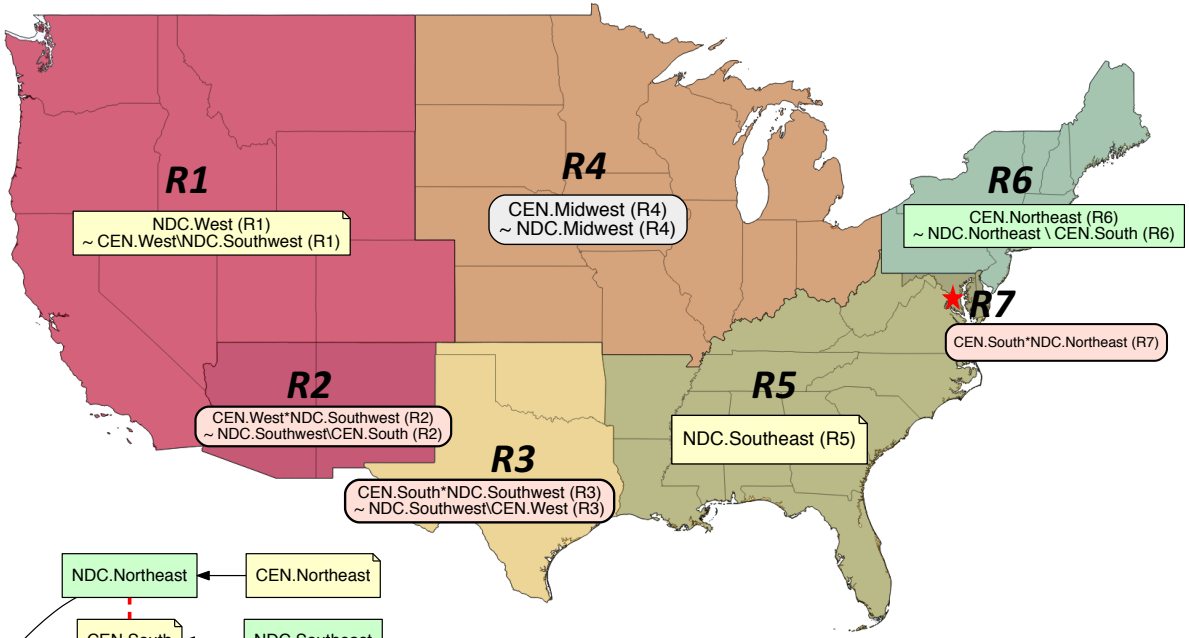

CEN.Midwest
CEN.South
CEN.USA
CEN.West
CEN.Northeast
NDC.Northeast
NDC.Southeast
NDC.USA
NDC.Midwest
NDC.Southwest
NDC.West
Taxonomy Reasoning System (Euler/X)
NDC.USA
CEN.USA
NDC.Northeast
CEN.Northeast
NDC.Northeast \ CEN.South
R6
NDC.Northeast * CEN.South
R7
CEN.South \ NDC.Southwest
CEN.South
CEN.South \ NDC.Northeast
NDC.Southeast
R5
NDC.Southwest
NDC.Southwest * CEN.South
NDC.Southwest \ CEN.West
R3
NDC.Southwest * CEN.West
NDC.Southwest \ CEN.South
R2
CEN.West
NDC.West
CEN.West \ NDC.Southwest
R1
NDC.Midwest
CEN.Midwest
R4
Nodes
CEN 3
NDC 4
congruent 2
* NEW * 5
Edges
is_a (input) 7
is_a (inferred) 10
Merged Taxonomy ("zoom-in" view)



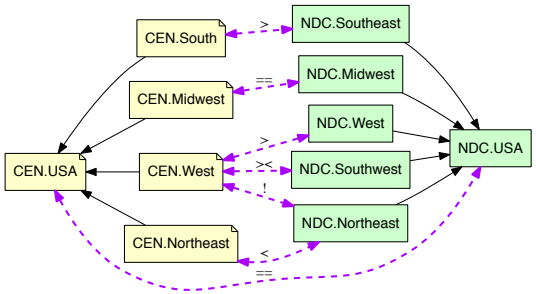


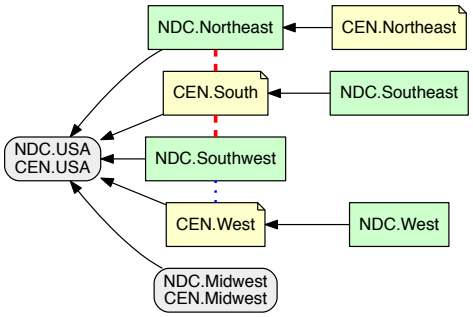






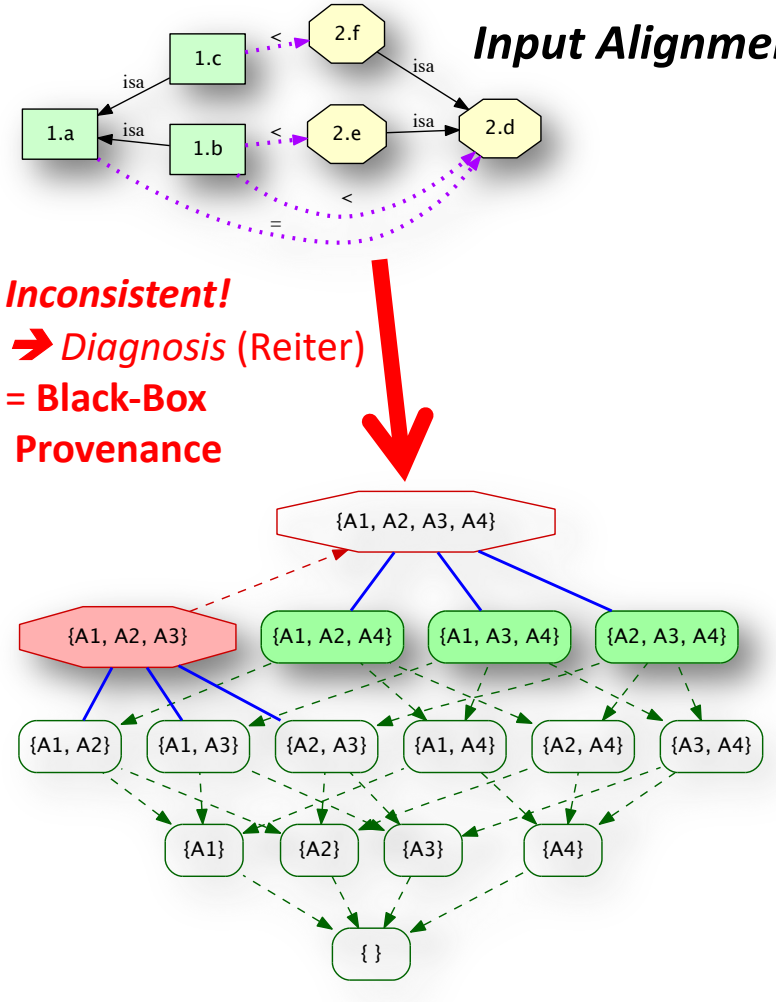
Nodes	
CEN	5
NDC	6
Edges	
is_a (CEN)	4
is_a (NDC)	5
articulations	7





Nodes	
CEN	3
NDC	4
congruence	2
Edges	
is_a (input)	8
overlaps (input)	1
overlaps (inferred)	2

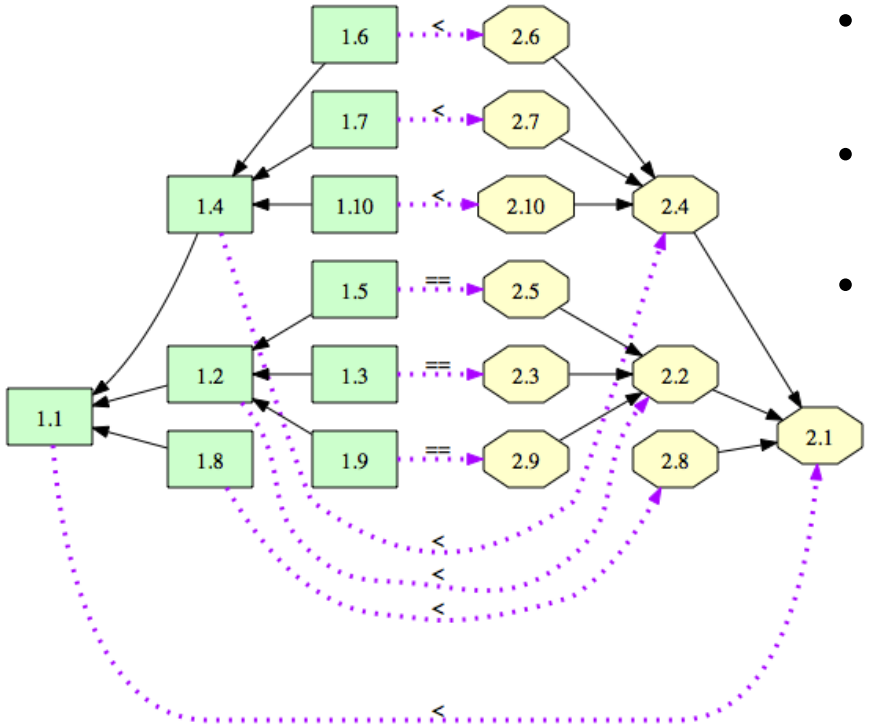
Possible Outcome: Inconsistency!



- Need to **debug** the input articulations ➔ (black-box) diagnosis!
- Focus:
  - How do we **efficiently** compute the diagnostic lattice?
- Also:
  - How to **visualize**..

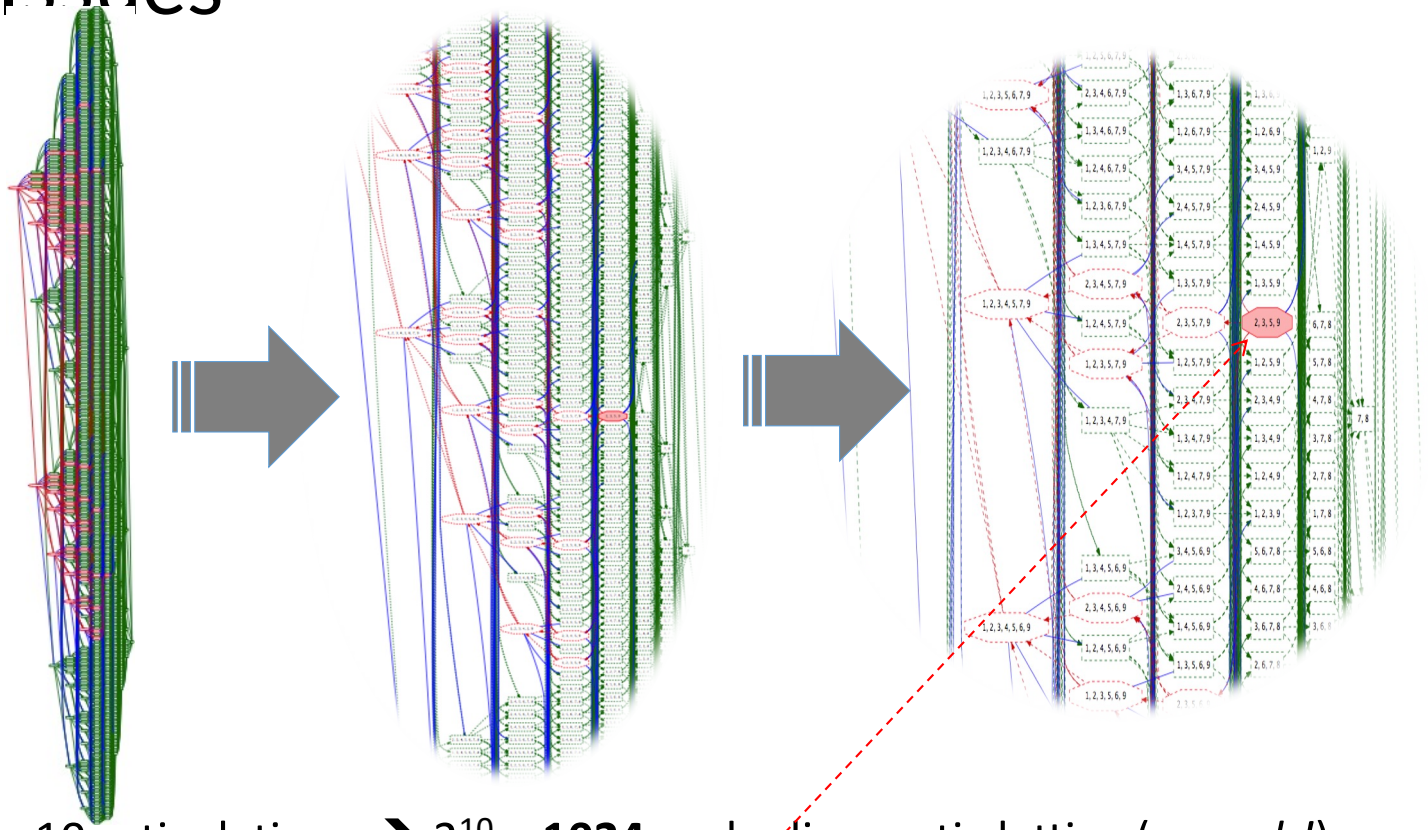
Nodes	
1	10
2	10
Edges	
isa_1	9
isa_2	9
Art.	10

Example Instance
(from synthetic benchmark suite)



- Here: N = **10** taxa in T_1, T_2
- Euler/X finds: **inconsistent**!
- ➔ **diagnostic lattice** of 2^{10} = **1024** nodes
  - ➔ Find minimal inconsistent subset **(MIS)**
  - ➔ maximal consistent subset **(MCS)** ..
  - ➔ **show** to user!

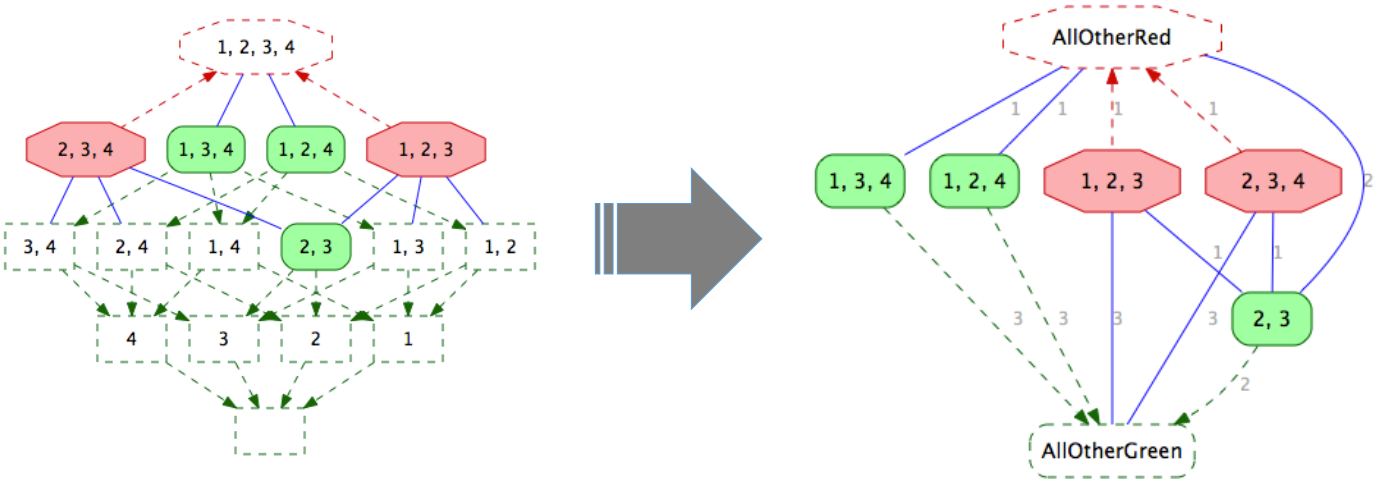
Visualizing Diagnoses: Scalability Issues



N = 10 articulations ➔ 2^{10} = **1024** node diagnostic lattice (… *ouch!*)

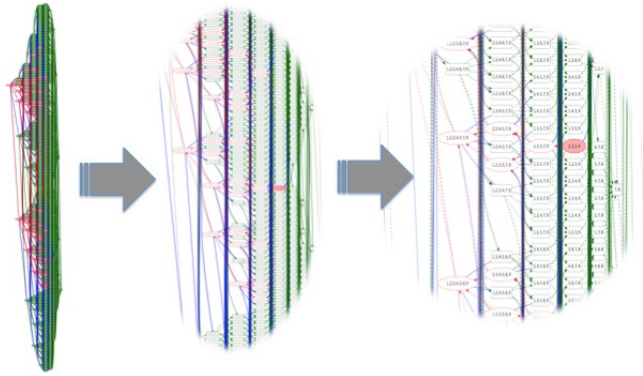
… but only one MIS …

Better Idea: Just show MIS, MCS



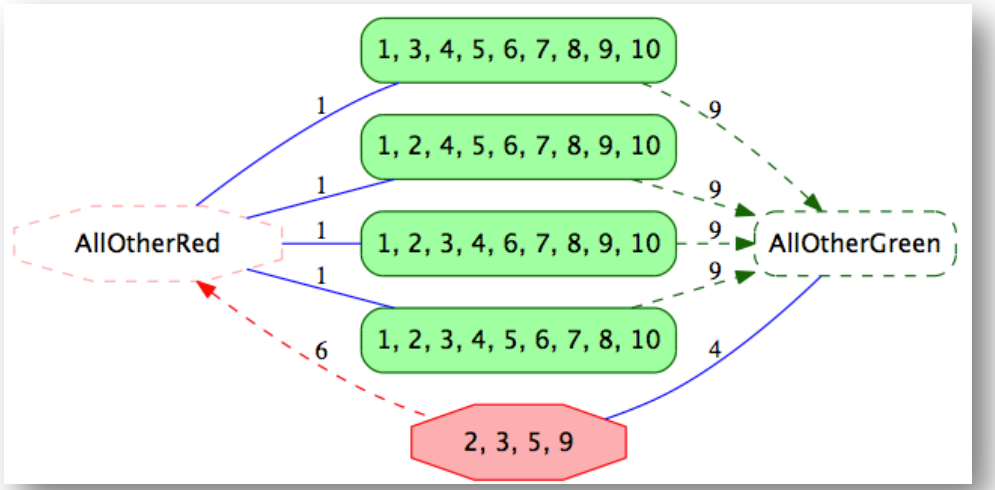
N = 4 articulations ➔ 2^4 = **16** node diagnostic lattice,
but 3 **MCS** and 2 **MIS** are enough!

Visualizing Diagnoses: Focusing on MIS (and MCS)



1024 node lattice

.. but **4 MCS** and **1 MIC** tell it all!



Visualizing Diagnoses

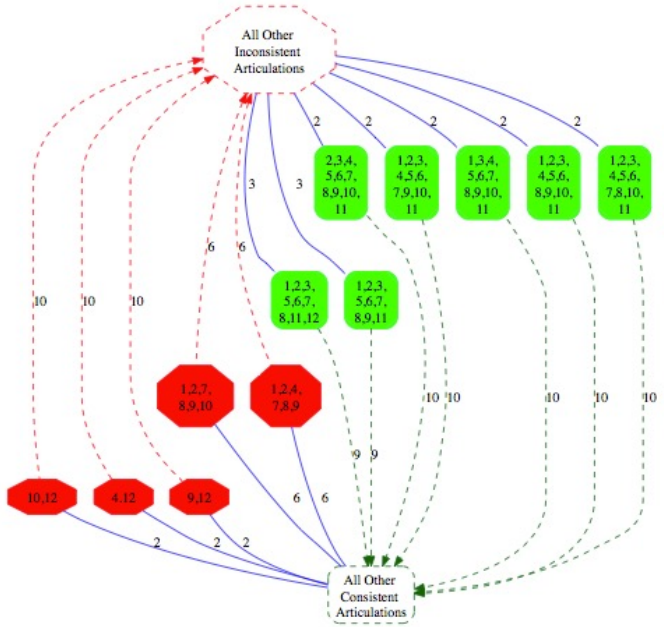
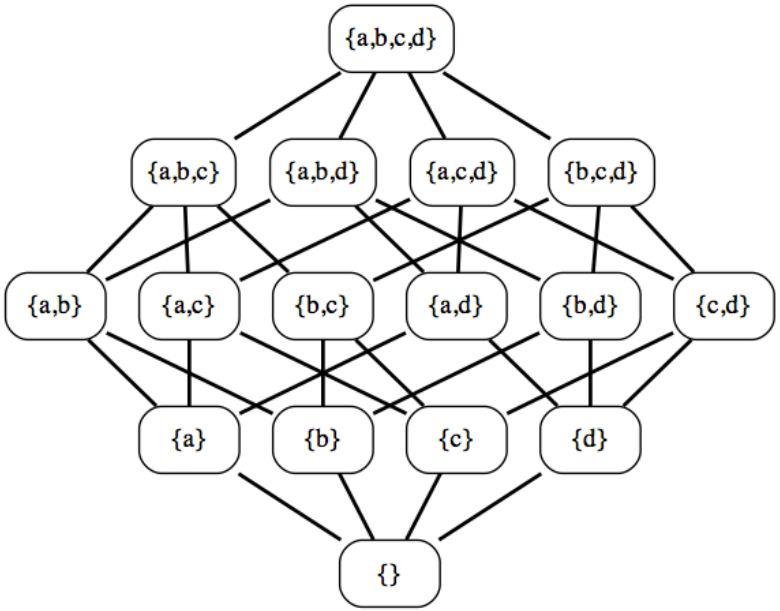


Fig. 7. $\mathcal{MIS}$ (Octagon) and $\mathcal{MCS}$ (Rounded Box) for Example 2. All other non-minimal inconsistent subsets and non-maximal consistent subsets are collapsed as "clouds", the labels of edges show the path length from MIS/MCS to the top/bottom of the lattice.

Example from paper: N=12 ➔ 4096 nodes

.. but **7 MCS** and **5 MIC** tell it all!

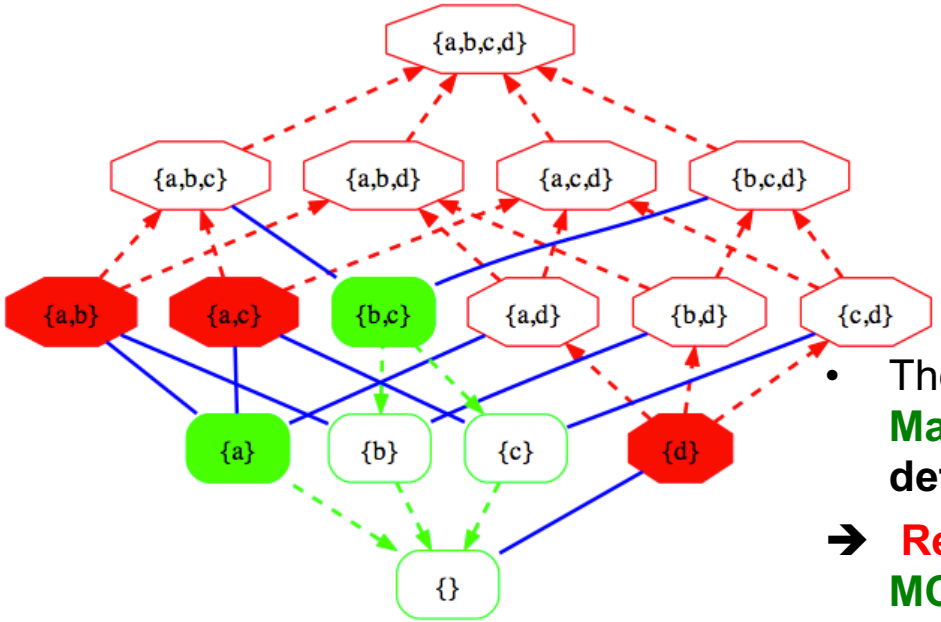
Black-Box Inconsistency Analysis (Diagnostic Lattice)



What happens if you can't have all (here: 4) articulations together?

- Then:
  - Repair: find & revise minimal inconsistent subsets (Min-Incons)
  - Expand: find maximal consistent subsets (Max-Cons) & revise *outs*

Inconsistency Analysis (Diagnostic Lattice)



- The **Min-Incons (MIS)** and **Max-Cons (MCS)** sets **determine all others**

➔ **Repair MIS** and/or **Expand MCS**

- **Black-box Analysis** (Hitting Set algo.) yields a **Diagnosis** (lattice)
  - for *n*=4 articulations, there are 168 possible diagnoses
  - depending on expected "red/green areas" ➔ explore space differently
- |articulations| = *n* ➔

  |possible diagnoses| = |monotonic Boolean functions|

  = Dedekind Number (*n*): 2, 3, 6, 20, **168**, 7581, 7828354, …

Improving Diagnosis

- Reiter's "black-box" (model-based) diagnosis helps debug the articulations
- Limited scalability (inherent complexity)
- But every bit helps:
  - Hitting Set Algorithm ("logarithmic extraction")
- Our idea:
  - Exploit "white-box" reasoning information

  ➔ **RULES** to the rescue

Key Idea: exploit white-box info

- We use Answer Set Programming (ASP) to solve Taxonomy Alignment Problem (TAP)
- Inconsistency = “False” is derived in the head:

  False **:-** ***<denial of integrity constraint>***

- Apply **provenance trick** from databases ☺
  - What articulations contribute to a derivation of “False” ?
  - **Eliminate those that don’t!**

  ➔ an example of **reusing** inferences **across** separate black-box tests!

The Provenance “Trick”

$$H_1(\bar{Y}, r_1 \otimes (P_1 \otimes \ldots \otimes P_n)) :- B_1(\bar{X}_1, P_1), B_2(\bar{X}_2, P_2), \ldots, B_n(\bar{X}_n, P_n).$$

For any constraint rule without head predicate (i.e., false is the head):

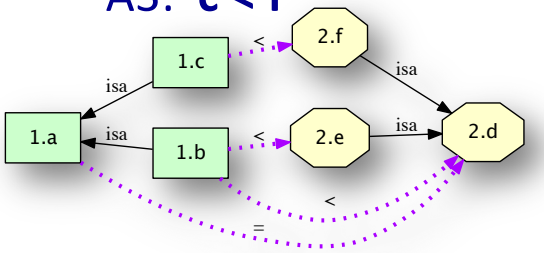
$$r_2 : \mathsf{false} :- B_1(\bar{X}_1), B_2(\bar{X}_2), \ldots, B_n(\bar{X}_n).$$

We rewrite it to a constraint with head predicate NOK where P_i is the provenance of $B_i(\bar{X}_i)$ for $1 \leq i \leq n$ and NOK stands for “Not OK”, i.e. inconsistency:

$$\mathsf{NOK}(r_2 \otimes (P_1 \otimes \ldots \otimes P_n)) :- B_1(\bar{X}_1, P_1), B_2(\bar{X}_2, P_2), \ldots, B_n(\bar{X}_n, P_n).$$

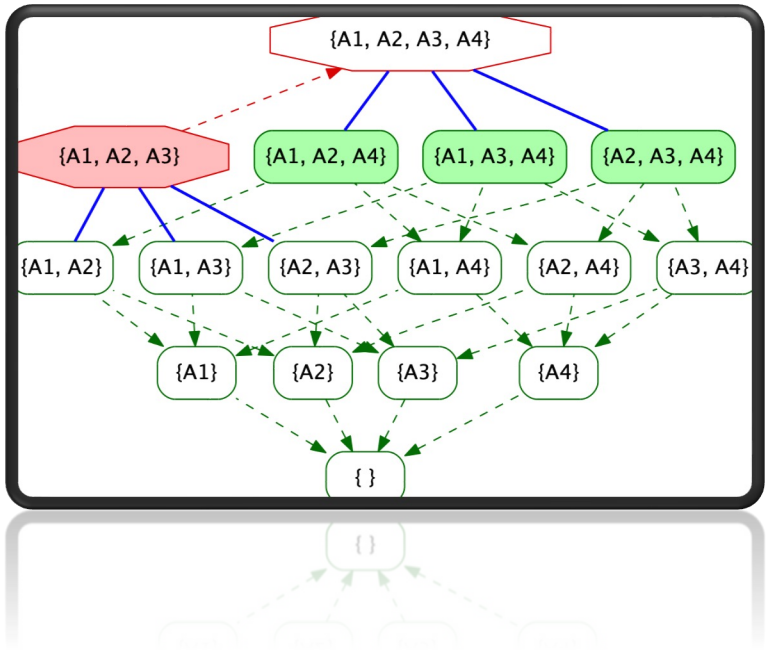
Hybrid Provenance

A3: **c < f**



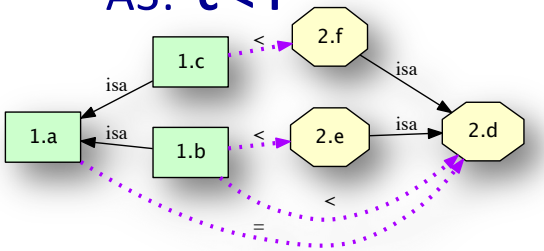
Input Alignment

Black-box Provenance



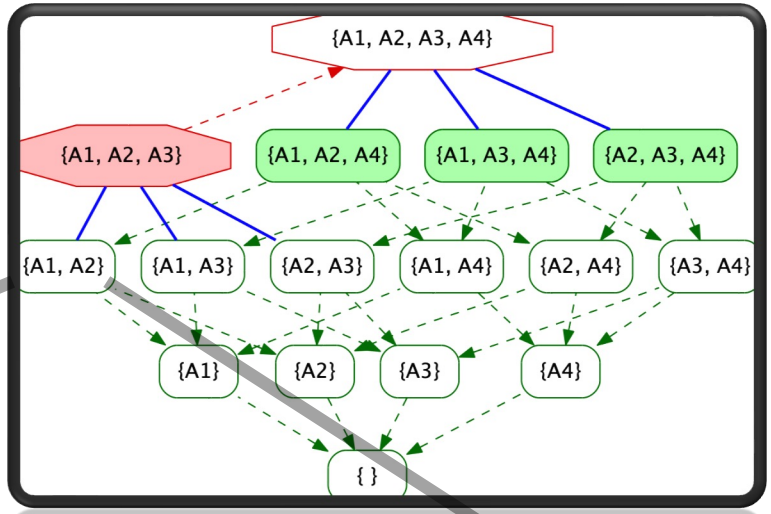
Hybrid Provenance

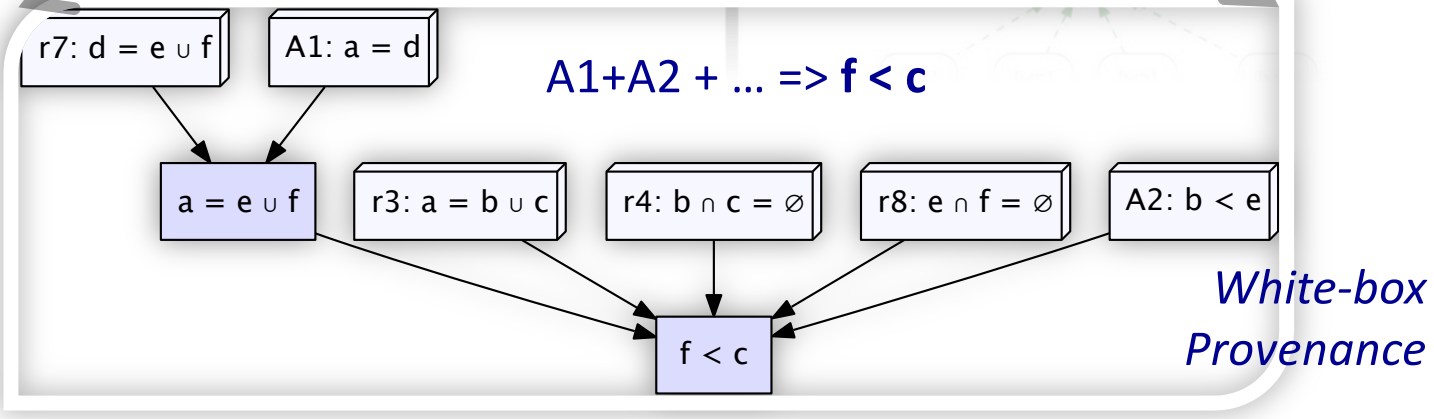
A3: **c < f**



Input Alignment

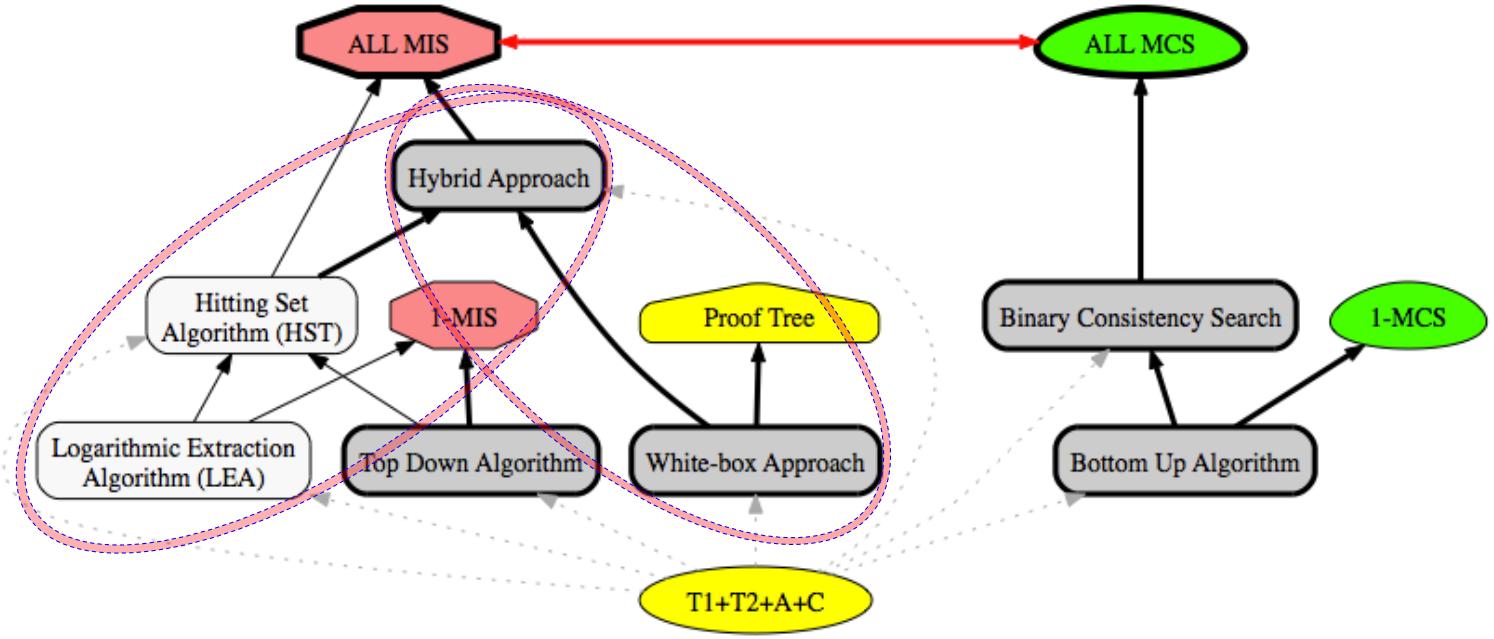
Black-box Provenance





White-box Provenance

The Hybrid Approach



Hybrid Approach

Algorithm 3 White-Box Approach

Input: System description *SD*, a set of constraints *C*
Output: All diagnosis proof trees

ComputeAllProofTrees(SD, C):
1: Encode *SD* and *C* in Datalog rules
2: Rewrite Datalog rules to ones with provenance
3: Run ASP reasoner to get boolean expressions for NOK
4: Construct diagnosis proof trees using the boolean expressions

Algorithm 4 Hybrid Approach

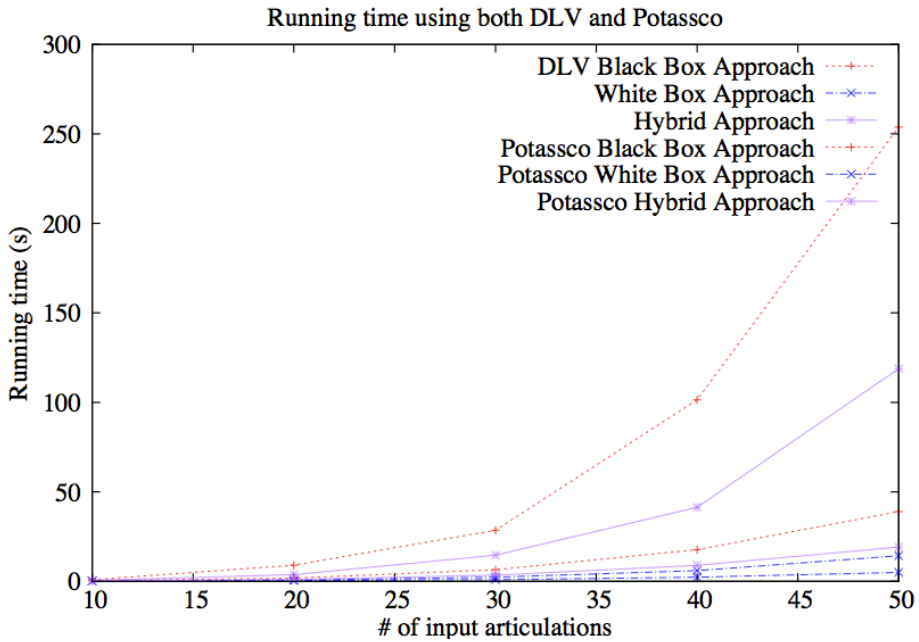
Input: System description *SD*, a set of constraints *C*
Output: All diagnoses ($\mathcal{MIS}$)

ComputeAllMISHybrid(SD, C):
1: $Ts \leftarrow$ *ComputeAllProofTrees(SD, C)*
2: $C' \leftarrow$ set of leaf nodes of the proof trees Ts
3: **return** *ComputeAllMIS(SD,C')*

*What articulations **contribute** to some inconsistency?*

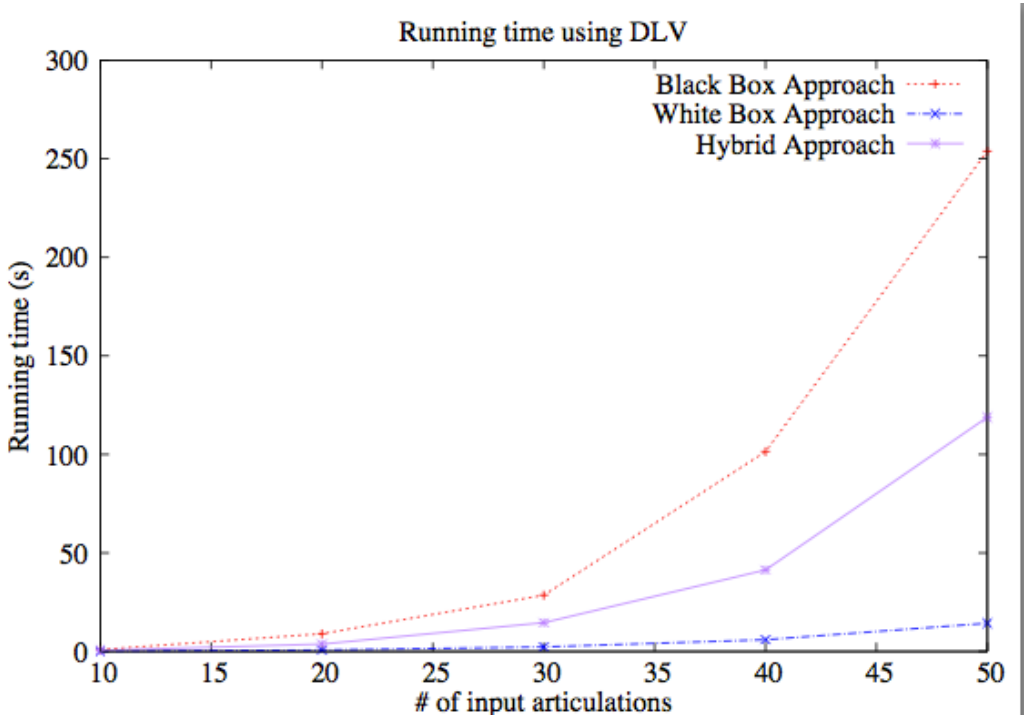
Good old black-box (HST)

Benchmark Results



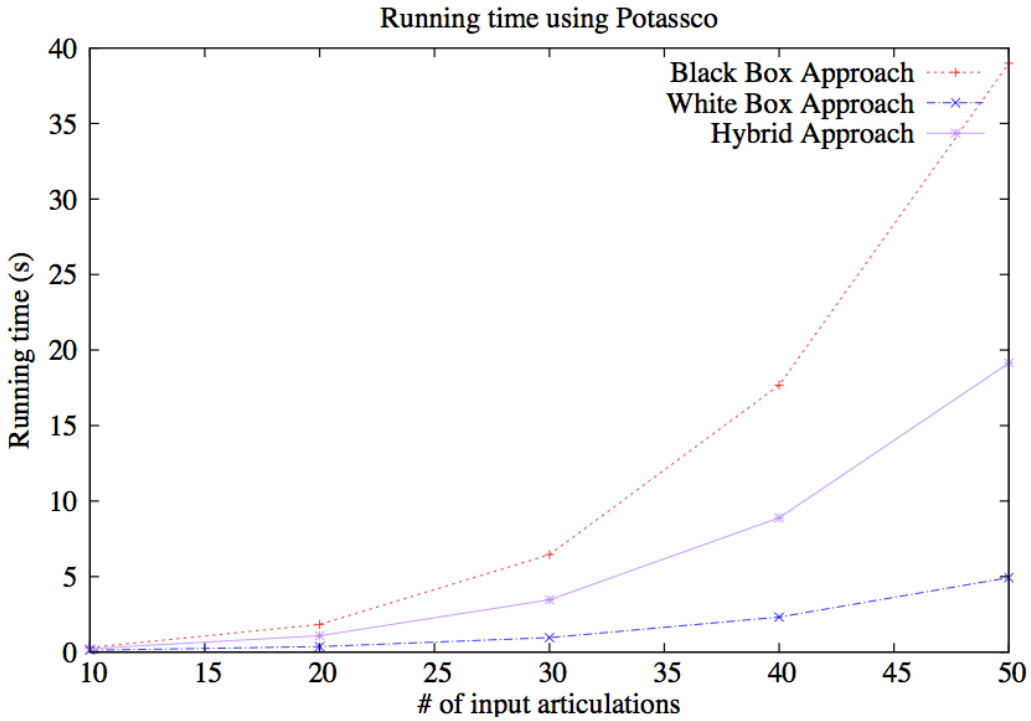
- White-box < Hybrid < **Black-box** (runtimes)
- Note: white-box does **not** give you a diagnosis
- Potassco < DLV

Benchmark DLV



- White-box < Hybrid < **Black-box** (runtimes)
- Potassco < DLV

Benchmark Clingo



- White-box < Hybrid < **Black-box** (runtimes)
- Potassco < DLV

Conclusions

- ASP rules can be used to efficiently solve real-world taxonomy reasoning problems
- Reiter's diagnosis useful to debug inconsistent alignments
- **Adding a "white-box" provenance approach speeds up state-of-the-art HST algorithm by eliminating independent articulations**
- Future work:
  - Further improvements, including **parallelism:**
    - **Trade-off** with **sharing** inferences **across parallel** instances