Rümmler, Arne
Chair of Geoinformatics, TU Dresden

How to evaluate and reduce the complexity of geospatial provenance graphs?

01.09.2020

P R O V E N A N C E

- Provenance describes the origination history of data

PROVENANCE

- Provenance describes the origination history of data
- Workflow graphs are a widely known form of provenance

PROVENANCE

- Provenance describes the origination history of data
- Workflow graphs are a widely known form of provenance
- Provenance data can be standardized (e.g. PROV-O)

PROVENANCE

- Provenance describes the origination history of data
- Workflow graphs are a widely known form of provenance
- Provenance data can be standardized (e.g. PROV-O)
- Standardized provenance graphs can be linked

PROVENANCE

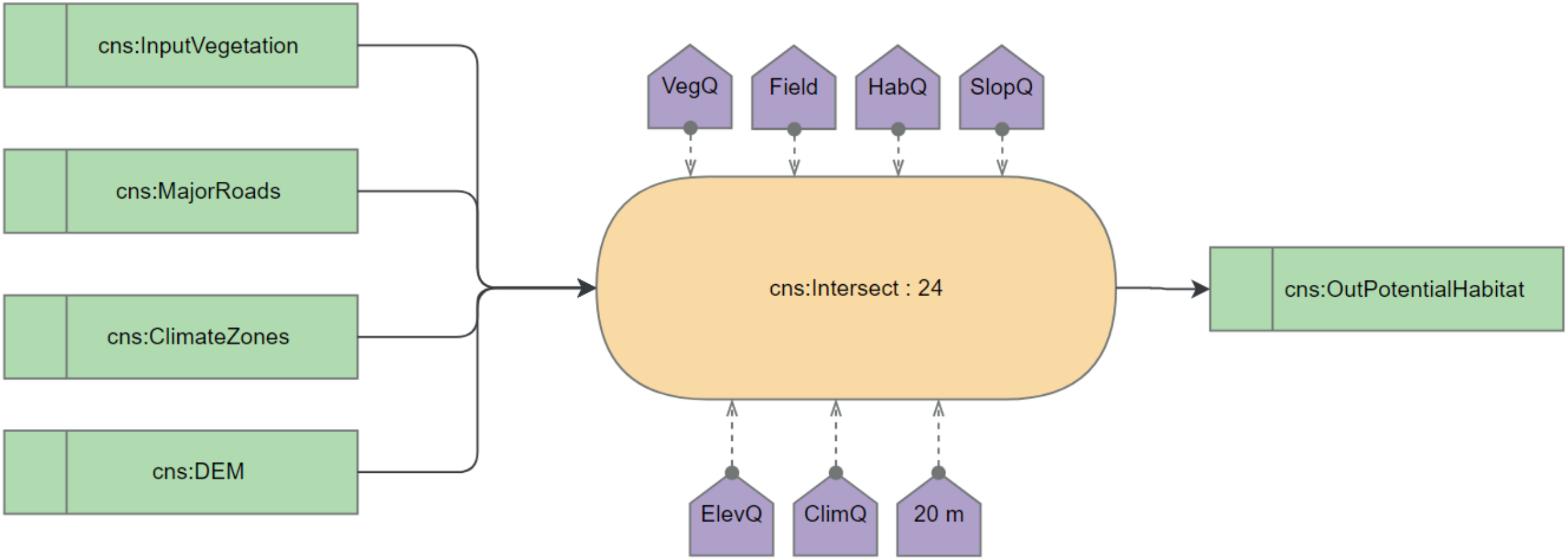
- Provenance describes the origination history of data
- Workflow graphs are a widely known form of provenance
- Provenance data can be standardized (e.g. PROV-O)
- Standardized provenance graphs can be linked
- Provenance graphs can become indefinitely large

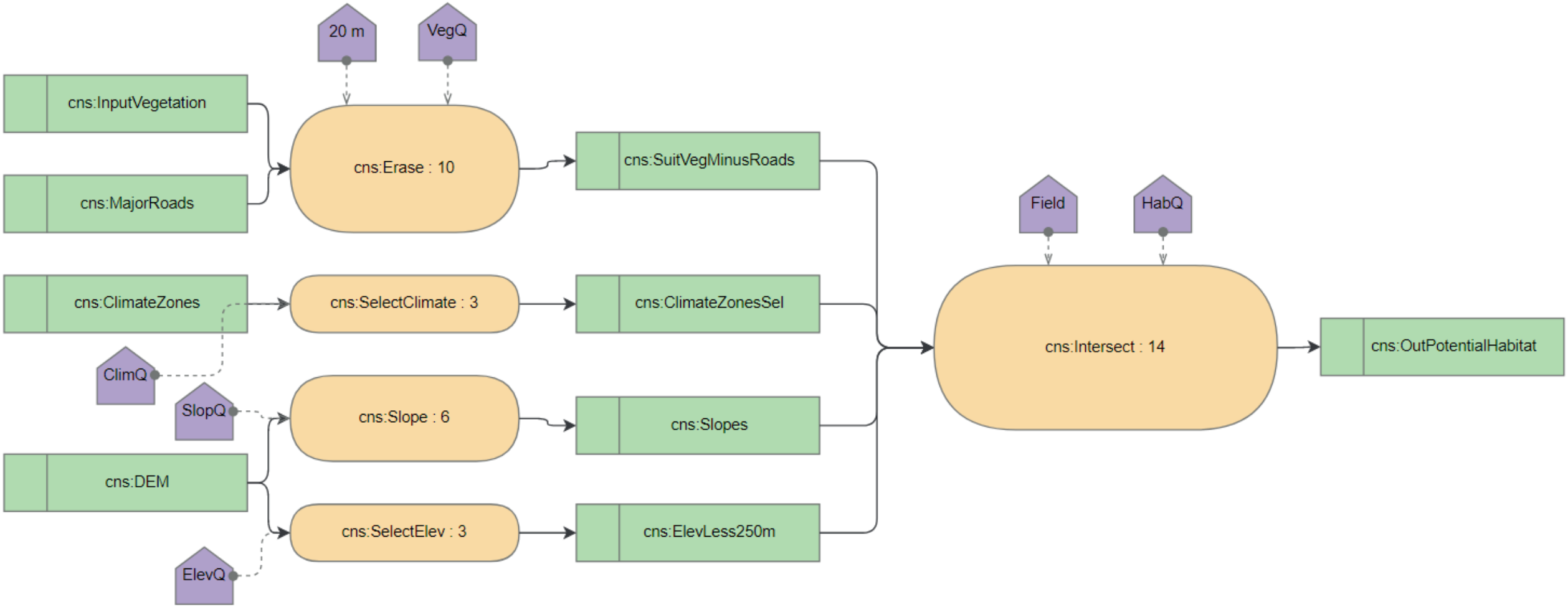
PROVENANCE

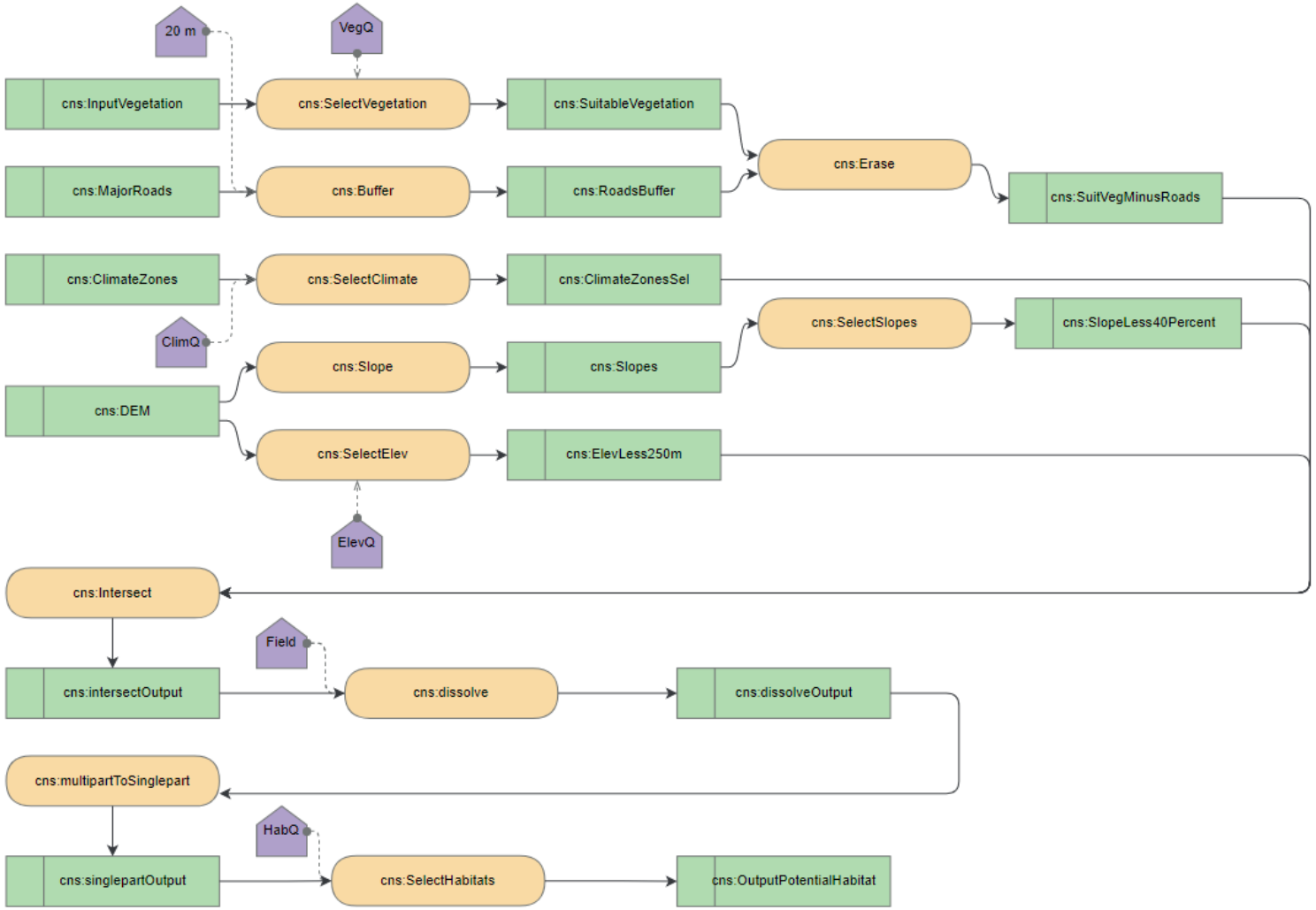
- Provenance describes the origination history of data
- Workflow graphs are a widely known form of provenance
- Provenance data can be standardized (e.g. PROV-O)
- Standardized provenance graphs can be linked
- Provenance graphs can become indefinitely large

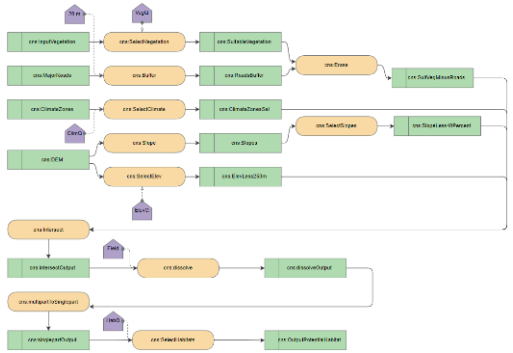
How to make use of such provenance data?

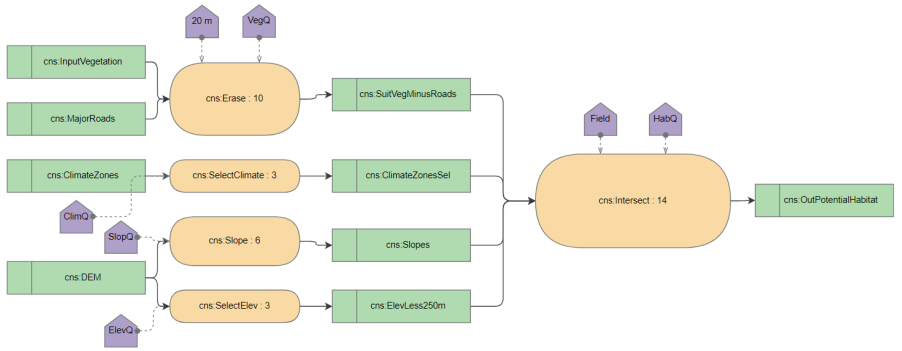
Leverage standardized provenance data to enhance the understandability of the origination history of geospatial data.

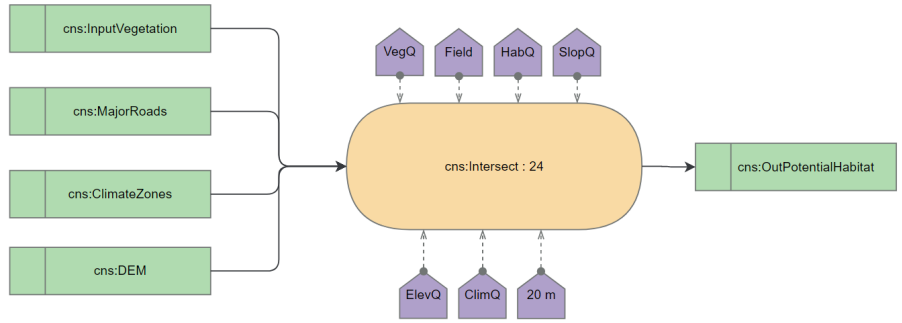

cns:InputVegetation
cns:MajorRoads
cns:ClimateZones
cns:DEM
VegQ
Field
HabQ
SlopQ
cns:Intersect : 24
ElevQ
ClimQ
20 m
cns:OutPotentialHabitat


20 m
VegQ
cns:InputVegetation
cns:MajorRoads
cns:Erase : 10
cns:SuitVegMinusRoads
cns:ClimateZones
cns:SelectClimate : 3
cns:ClimateZonesSel
ClimQ
SlopQ
cns:Slope : 6
cns:Slopes
cns:DEM
cns:SelectElev : 3
cns:ElevLess250m
ElevQ
Field
HabQ
cns:Intersect : 14
cns:OutPotentialHabitat


20 m
VegQ
cns:InputVegetation
cns:SelectVegetation
cns:SuitableVegetation
cns:Erase
cns:MajorRoads
cns:Buffer
cns:RoadsBuffer
cns:SuitVegMinusRoads
cns:ClimateZones
cns:SelectClimate
cns:ClimateZonesSel
cns:SelectSlopes
cns:SlopeLess40Percent
ClimQ
cns:Slope
cns:Slopes
cns:DEM
cns:SelectElev
cns:ElevLess250m
ElevQ
cns:Intersect
Field
cns:intersectOutput
cns:dissolve
cns:dissolveOutput
cns:multipartToSinglepart
HabQ
cns:singlepartOutput
cns:SelectHabitats
cns:OutputPotentialHabitat

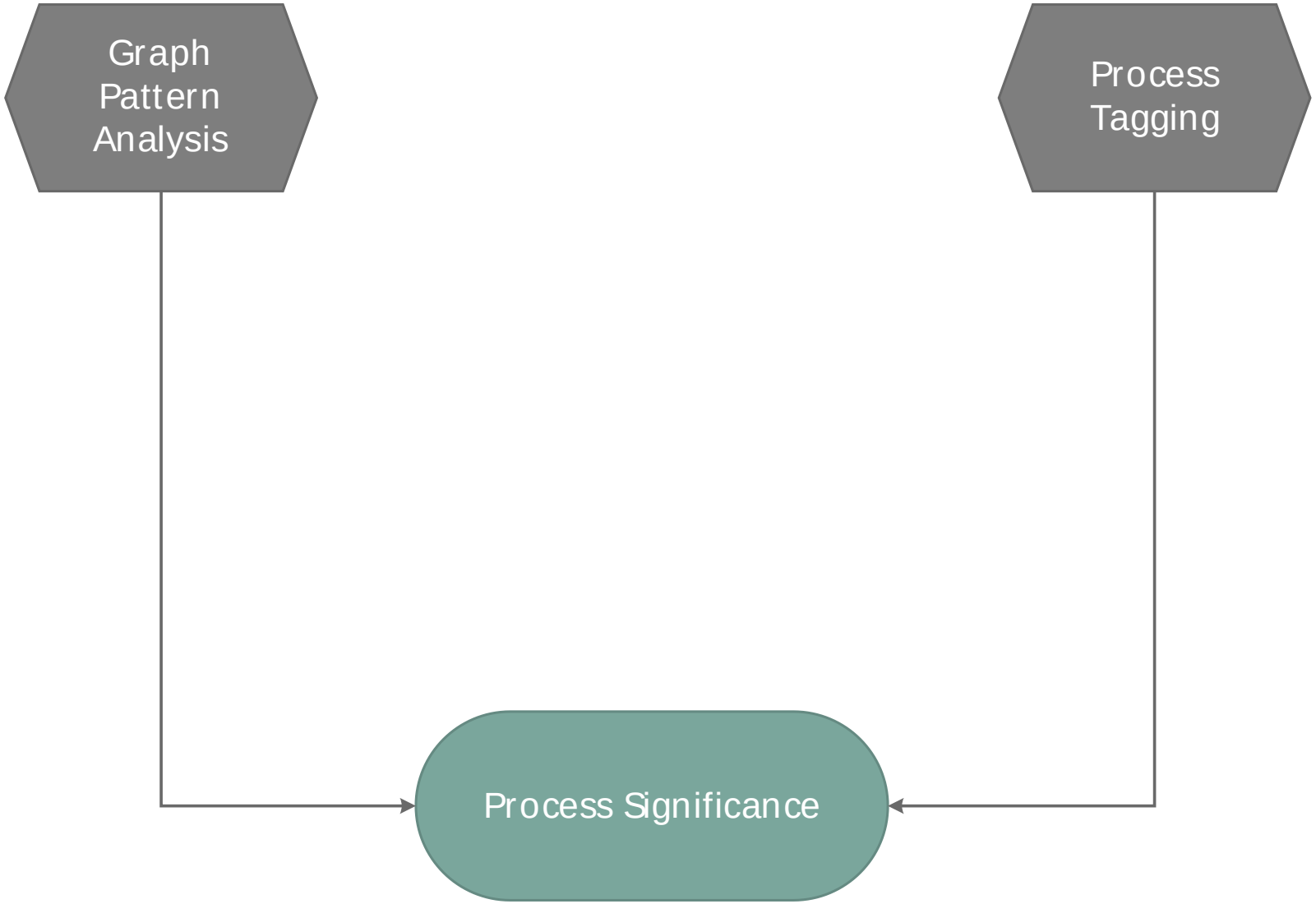


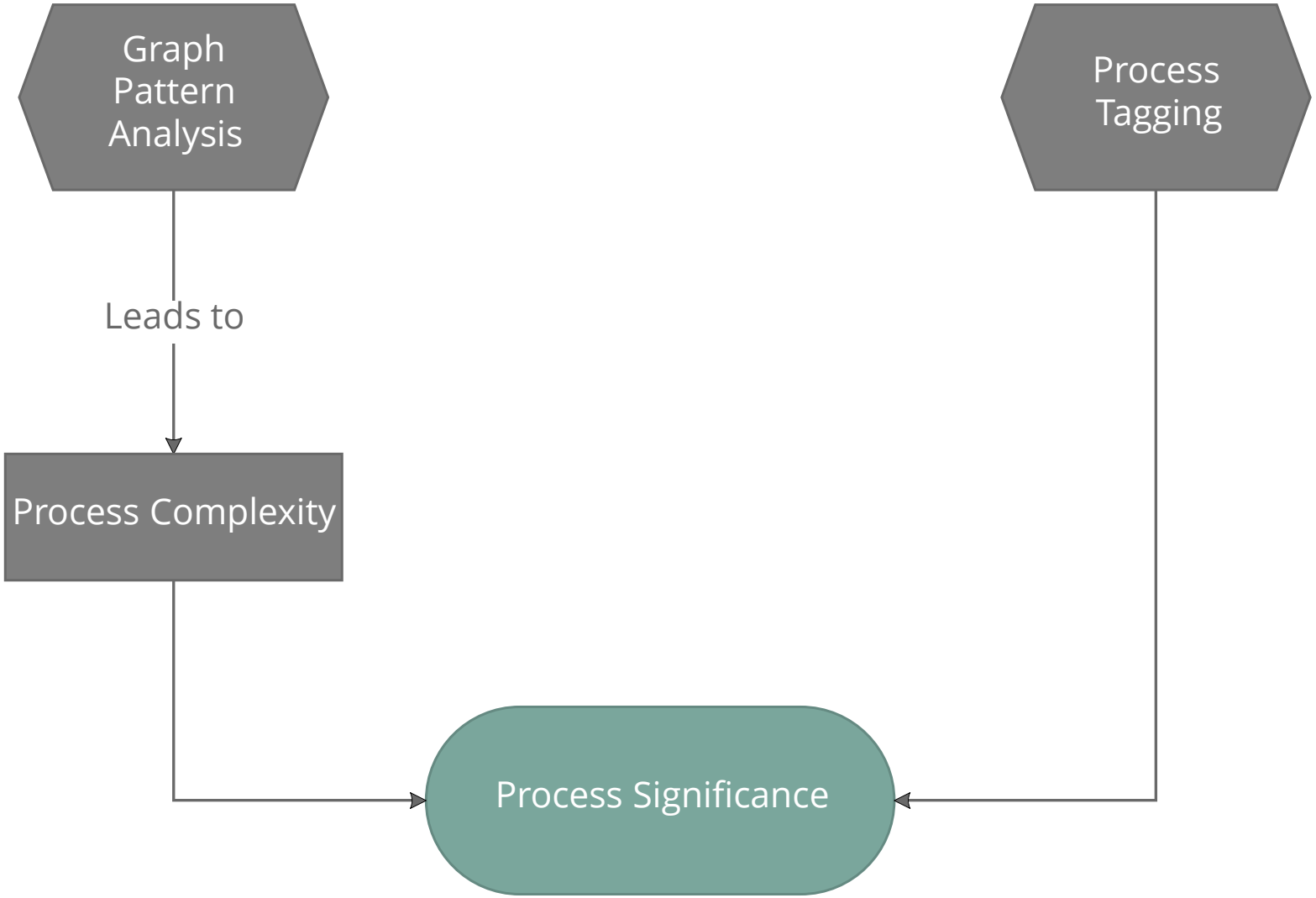


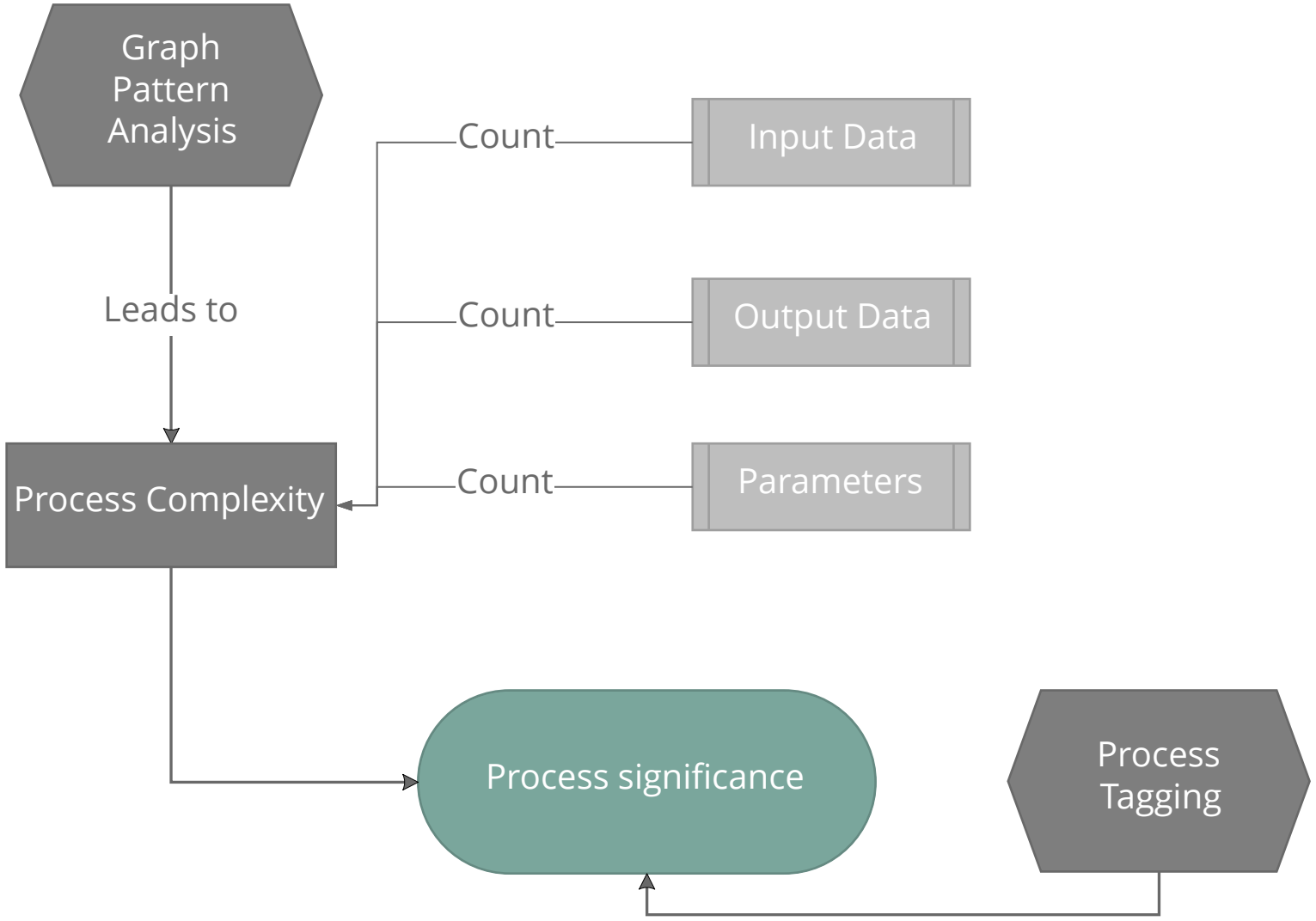


How to get from the left to the right?

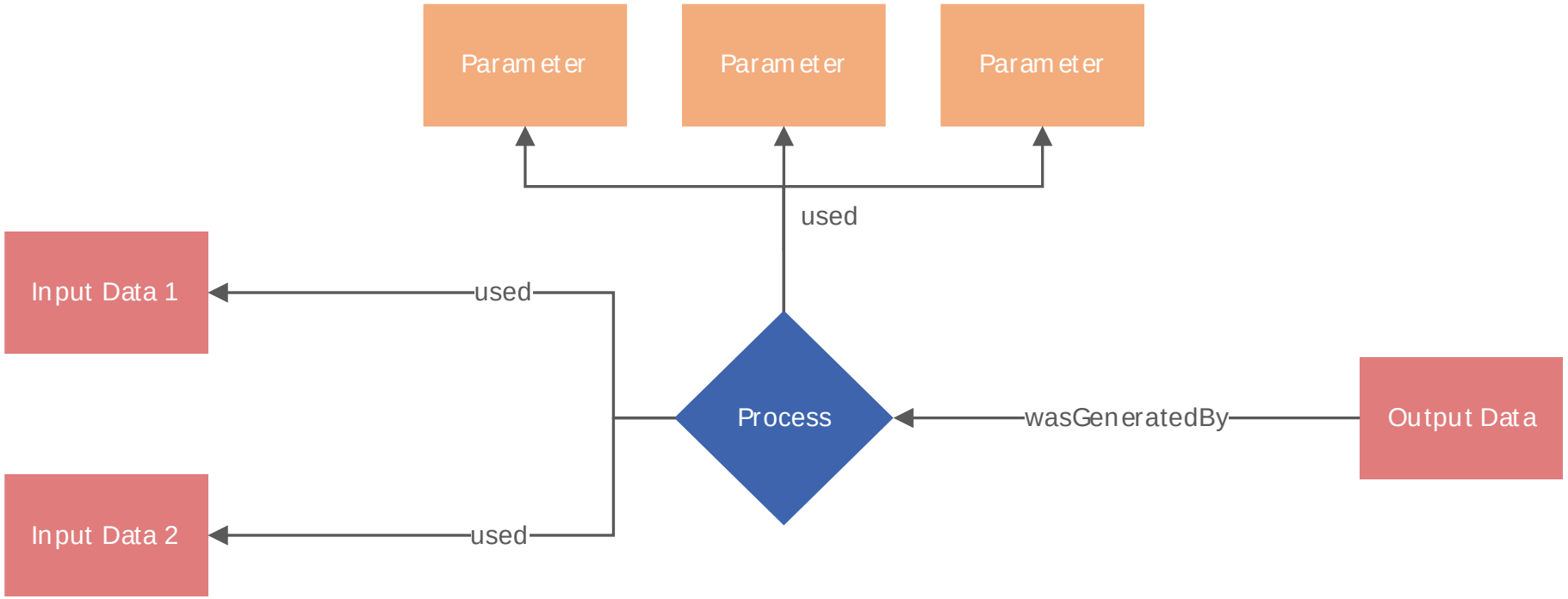
-> **Process-significance based generalization**


Graph Pattern Analysis
Process Tagging
Process Significance

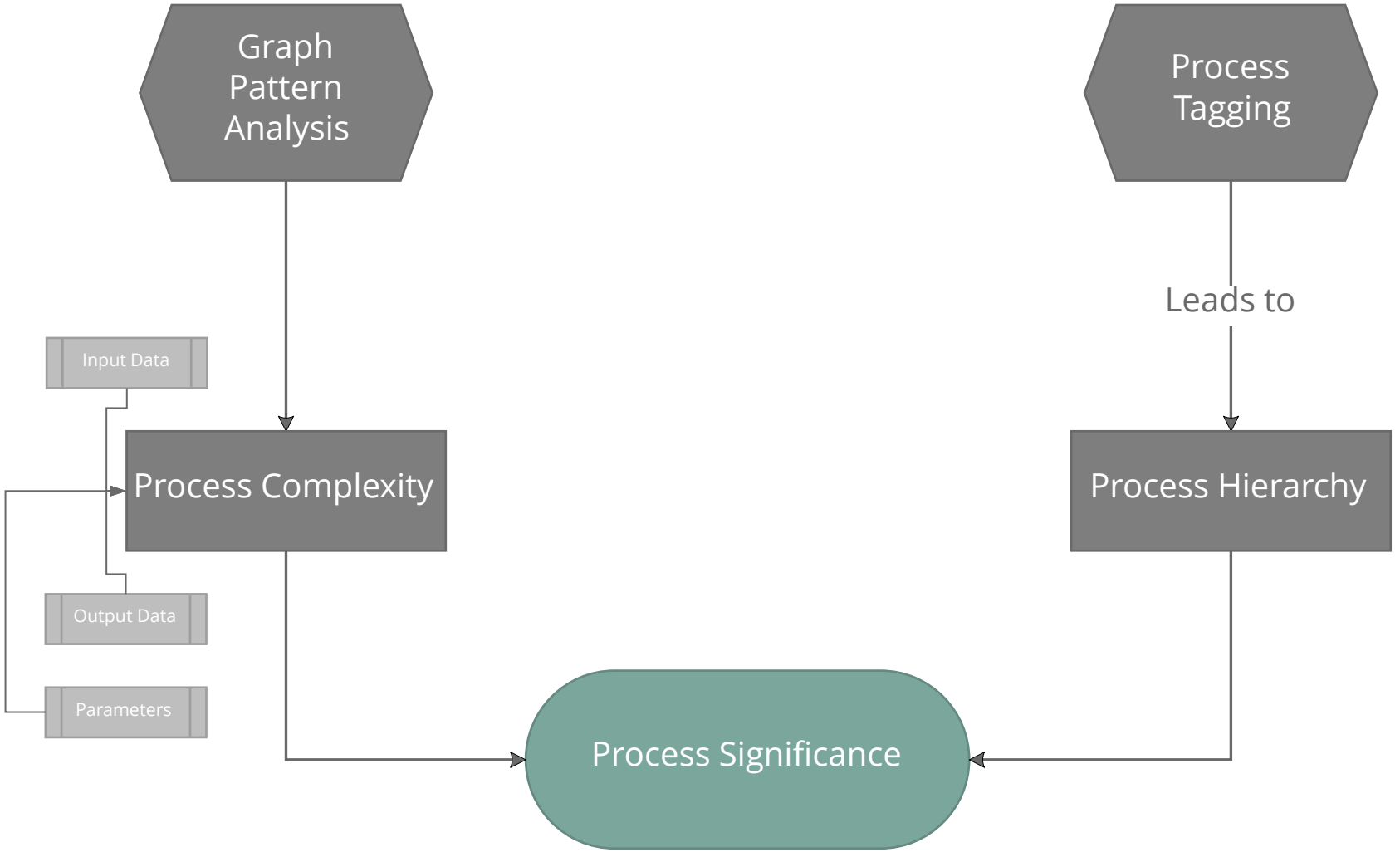

Graph Pattern Analysis
Process Tagging
Leads to
Process Complexity
Process Significance

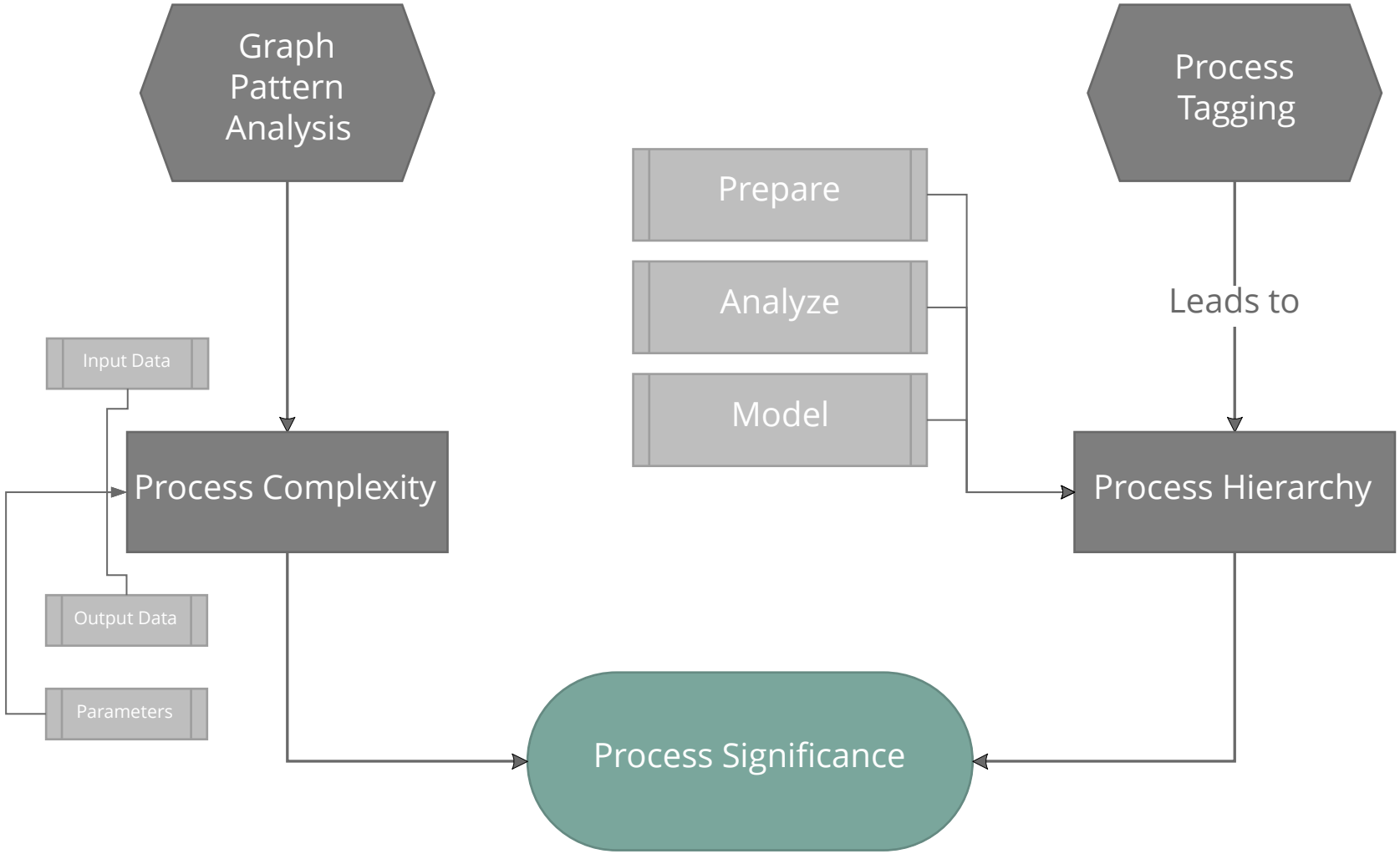

Graph Pattern Analysis
Leads to
Process Complexity
Count
Input Data
Count
Output Data
Count
Parameters
Process significance
Process Tagging

Process complexity analysis by graph pattern



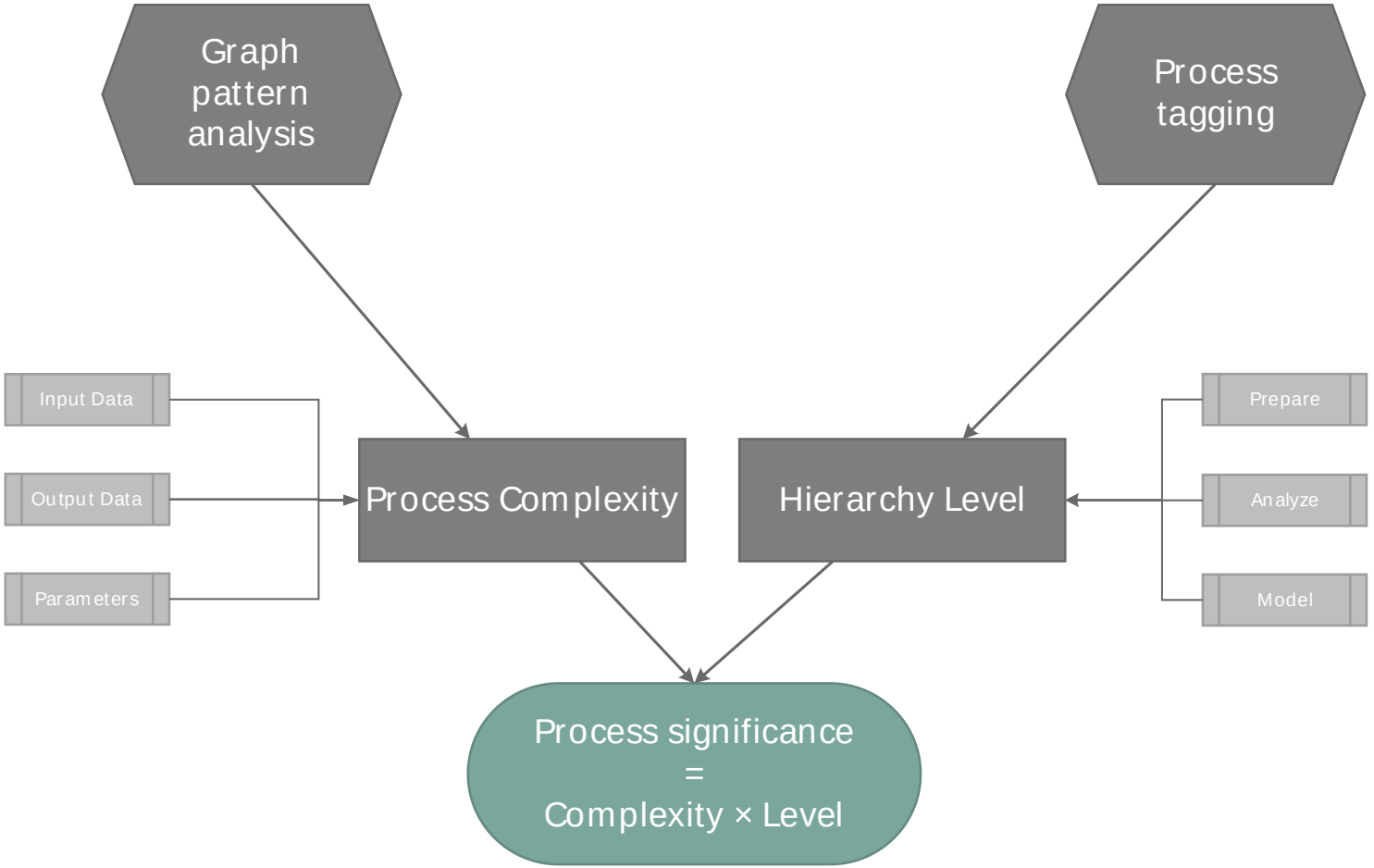
- 2 Inputs, 3 Parameter, 1 Output -> Complexity **C = 2 + 3 + 1 = 6**
- Alternative complexity calculation possible, e.g. sqrt(params)
- Further patterns can be included


Graph Pattern Analysis
Process Tagging
Leads to
Input Data
Process Complexity
Process Hierarchy
Output Data
Parameters
Process Significance

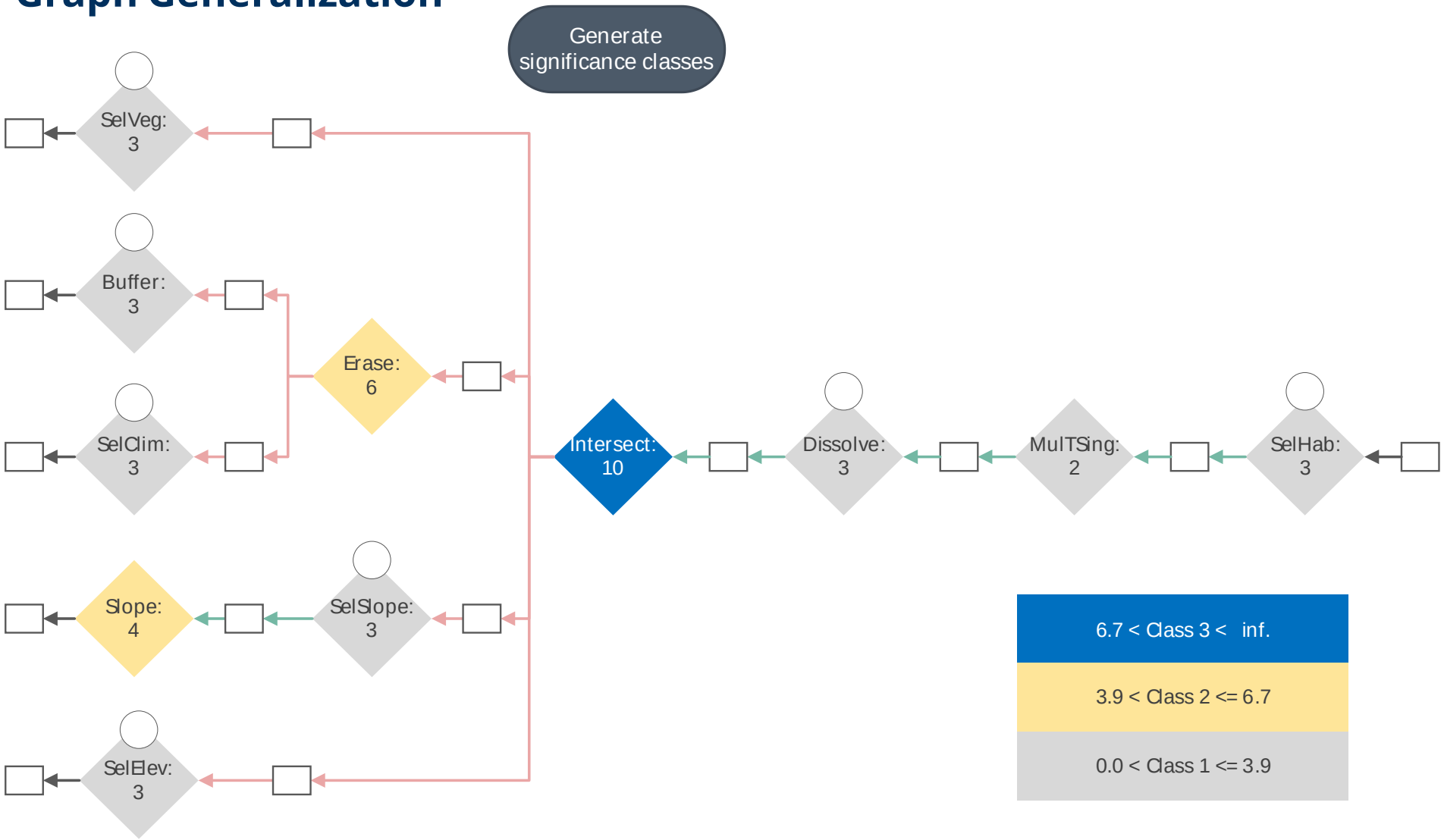

Graph Pattern Analysis
Process Tagging
Prepare
Analyze
Model
Leads to
Input Data
Process Complexity
Process Hierarchy
Output Data
Parameters
Process Significance

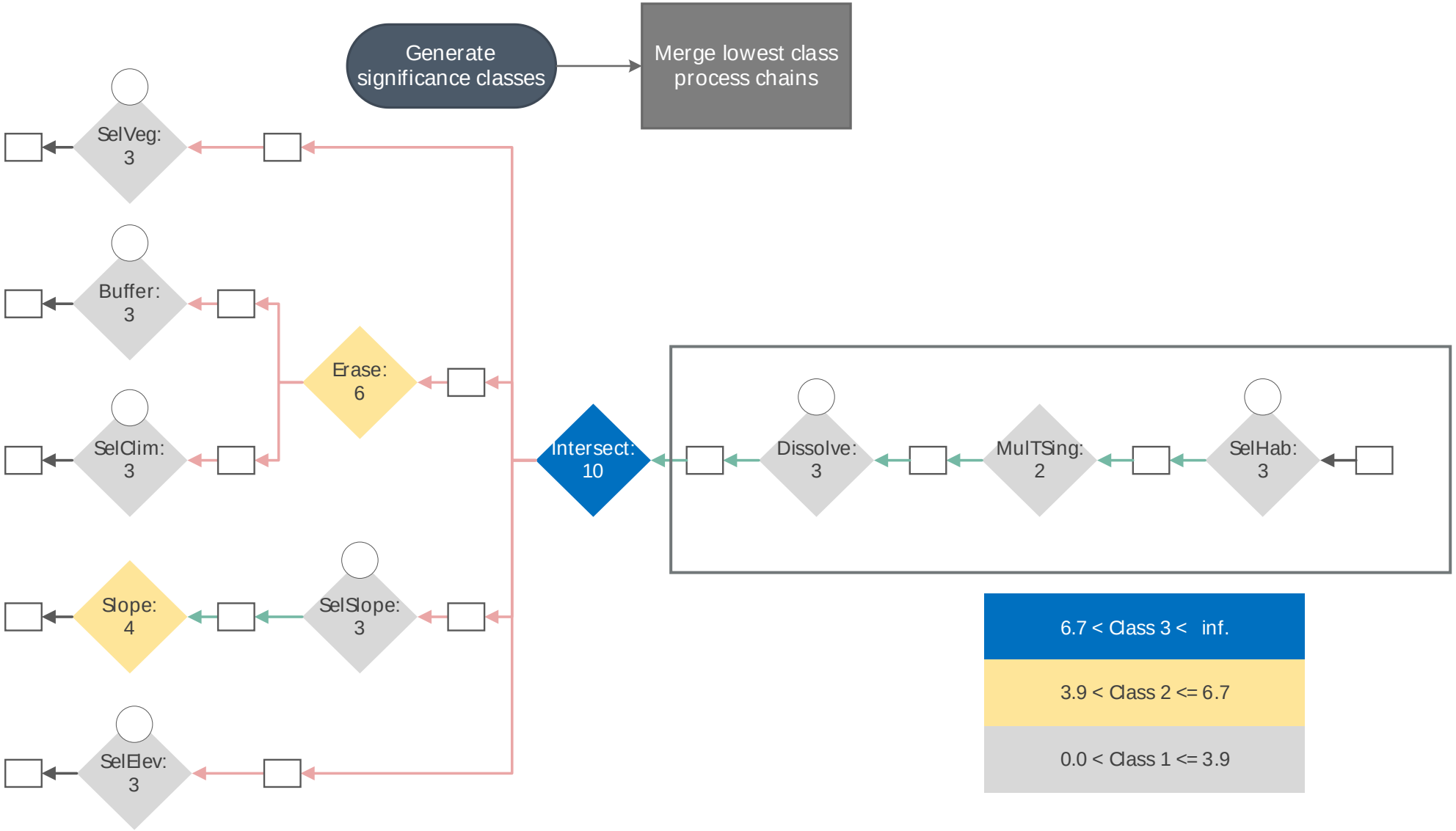
Prepare	Analyze	Model
▪ Selection ▪ Transformation ▪ Reclassification ▪ Geometric Computation ▪ Visualization	▪ Neighborhood Analysis ▪ Map Overlay ▪ Network ▪ Statistics	▪ Model (back-box-modules, neural networks, …)
1	2	3

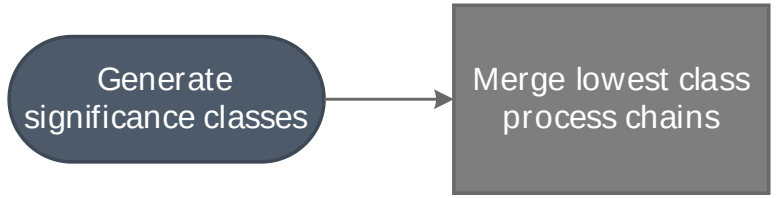
- Subcategories derived from Albrecht (Albrecht, J. Universal analytical GIS operations: a task-oriented systematization of data structure-independent GIS functionality 1998.)
- Value serves a multiplier

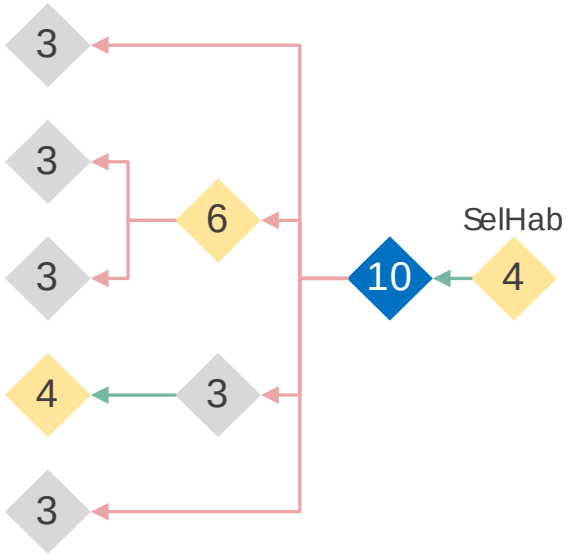

Graph pattern analysis
Process tagging
Input Data
Output Data
Parameters
Process Complexity
Hierarchy Level
Prepare
Analyze
Model
Process significance
=
Complexity × Level

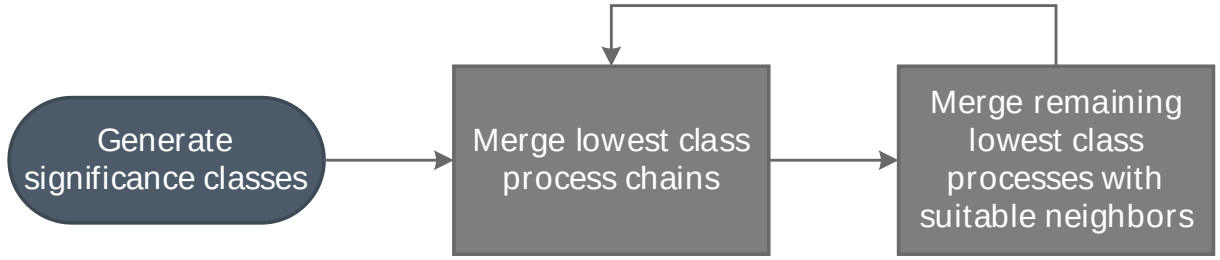
Graph Generalization

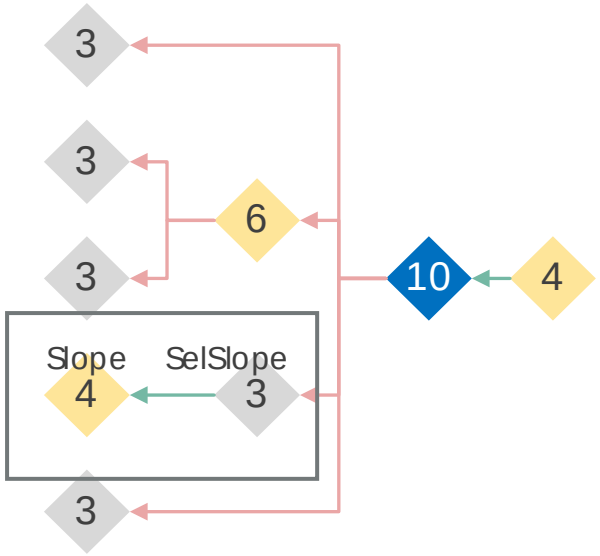


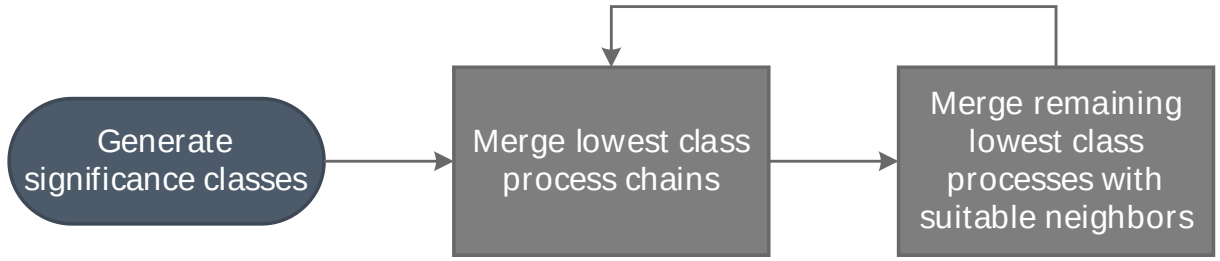

Generate significance classes
Merge lowest class process chains
SelVeg: 3
Buffer: 3
Erase: 6
SelClim: 3
Slope: 4
SelSlope: 3
SelElev: 3
Intersect: 10
Dissolve: 3
MulTSing: 2
SelHab: 3
6.7 < Class 3 < inf.
3.9 < Class 2 <= 6.7
0.0 < Class 1 <= 3.9

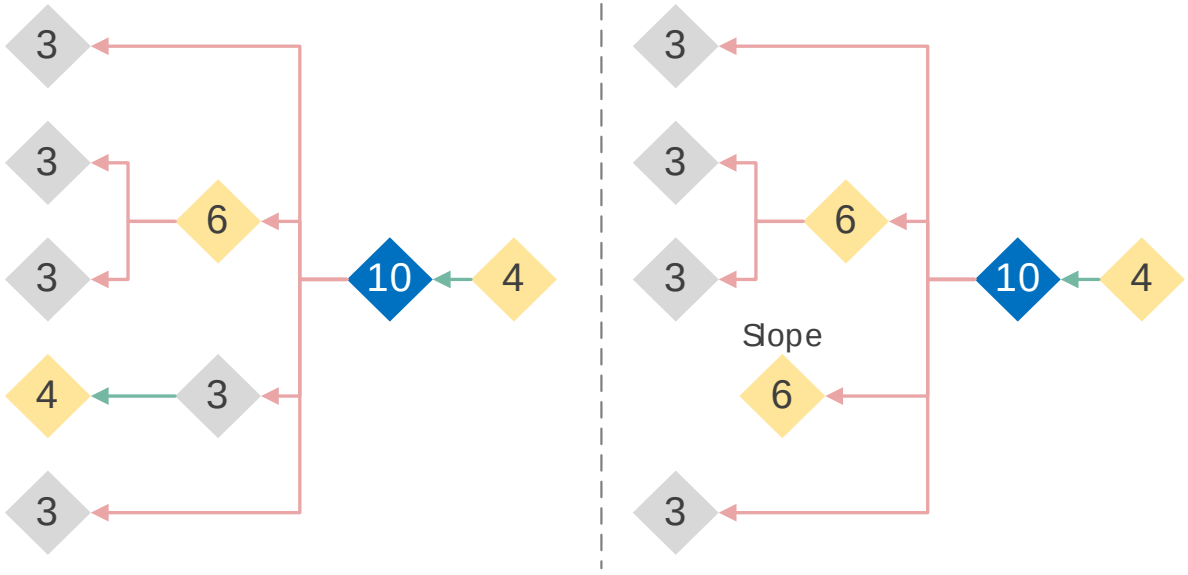

Generate significance classes
Merge lowest class process chains

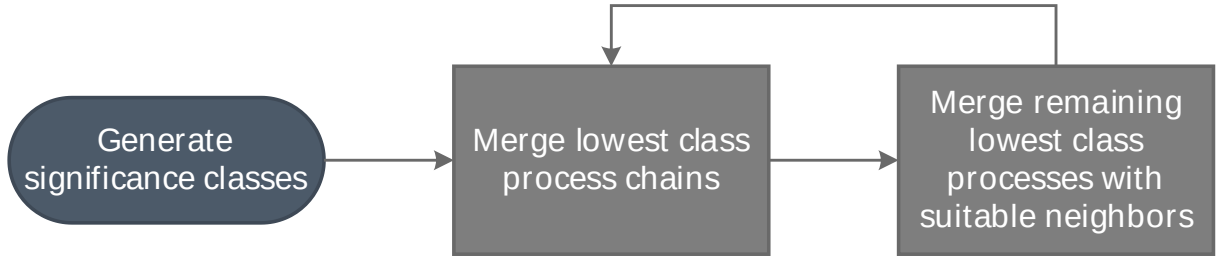

3
3
6
3
10
SelHab
4
4
3
3

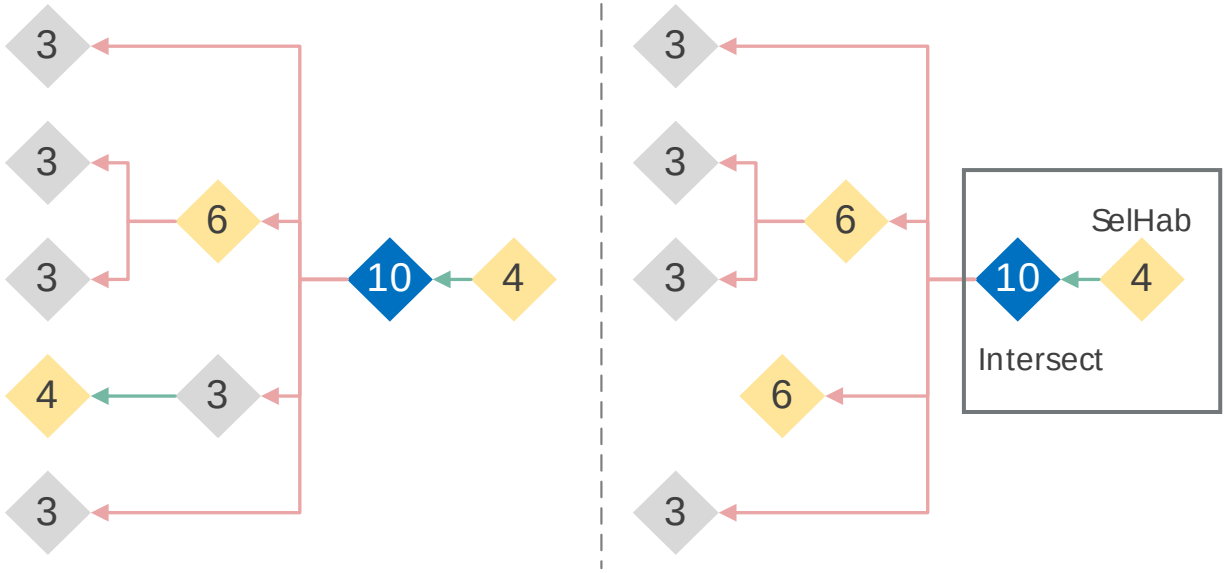

Generate significance classes
Merge lowest class process chains
Merge remaining lowest class processes with suitable neighbors

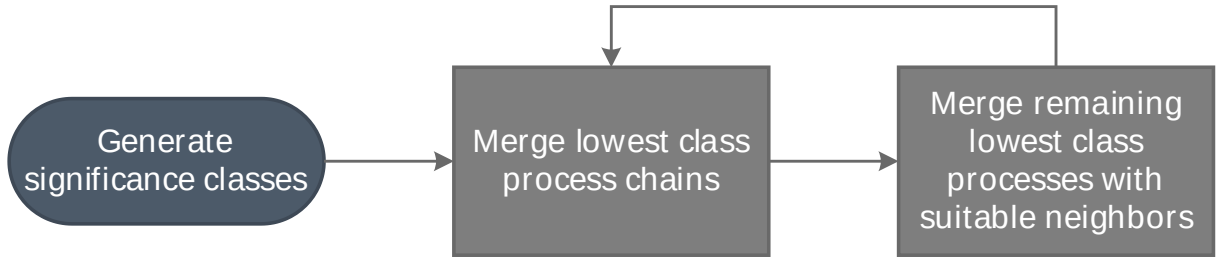

3
3
6
3
10
4
Slope
SelSlope
4
3
3

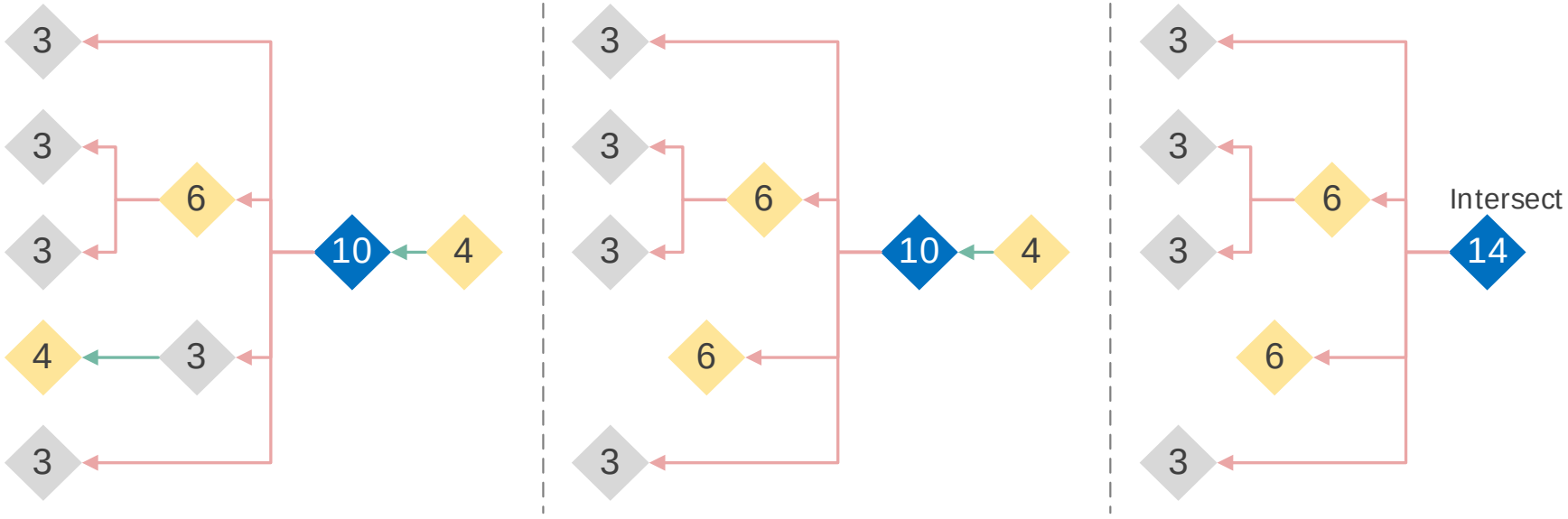

Generate significance classes
Merge lowest class process chains
Merge remaining lowest class processes with suitable neighbors

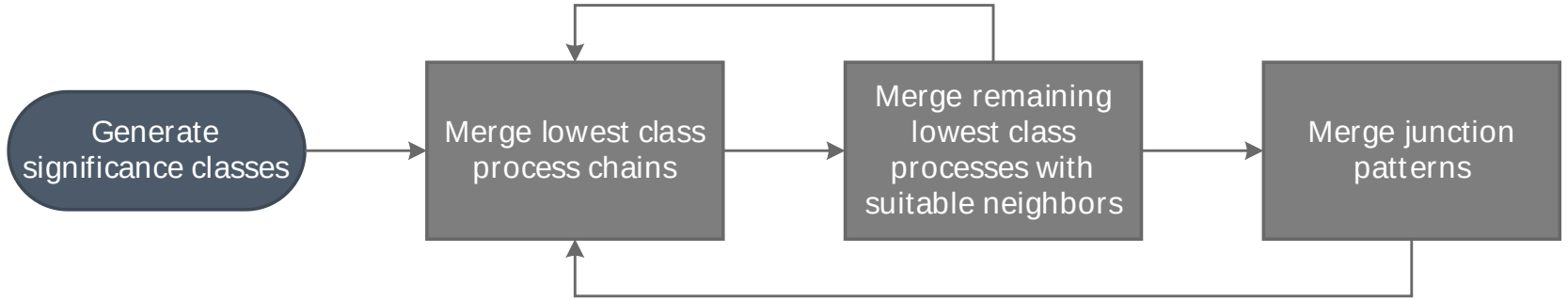

3
3
6
3
10
4
4
3
3
3
3
6
3
10
4
Slope
6
3

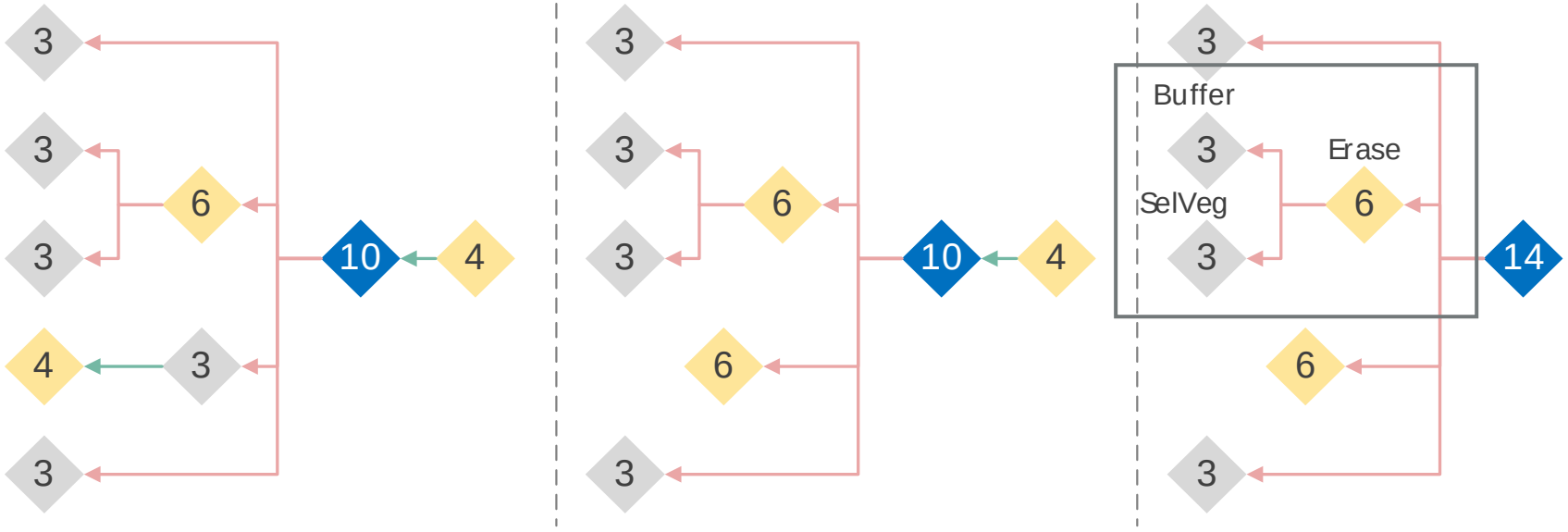

Generate significance classes
Merge lowest class process chains
Merge remaining lowest class processes with suitable neighbors

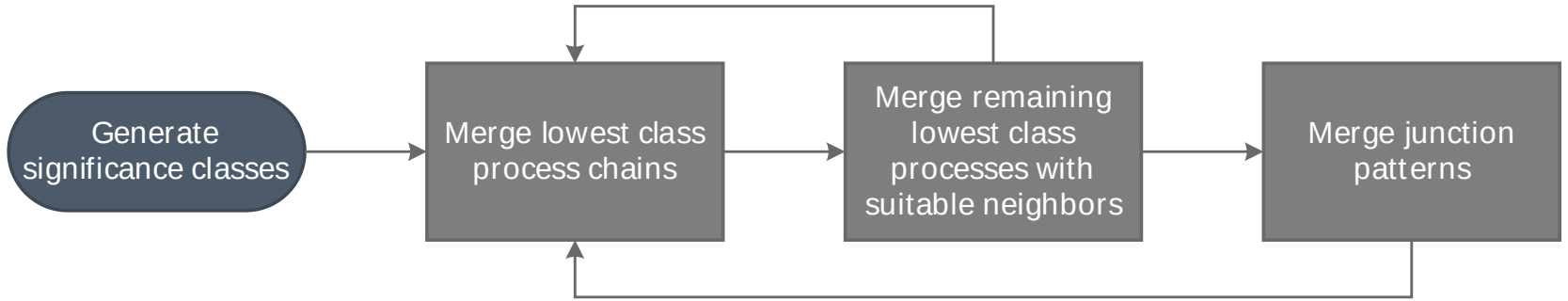

3
3
3
6
10
4
4
3
3
3
3
3
6
10
4
SelHab
Intersect
6
3

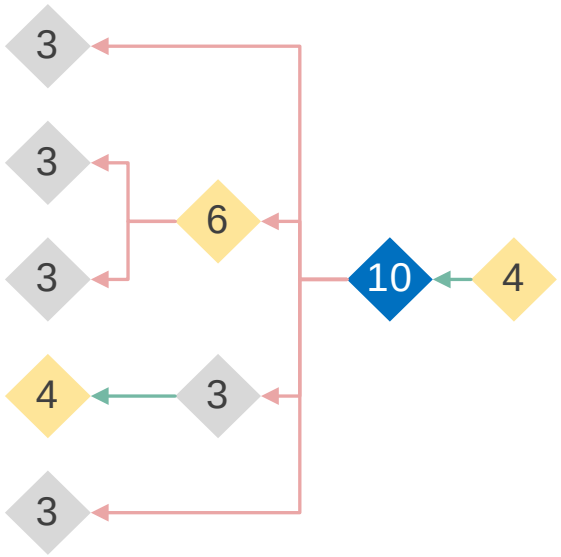

Generate significance classes
Merge lowest class process chains
Merge remaining lowest class processes with suitable neighbors

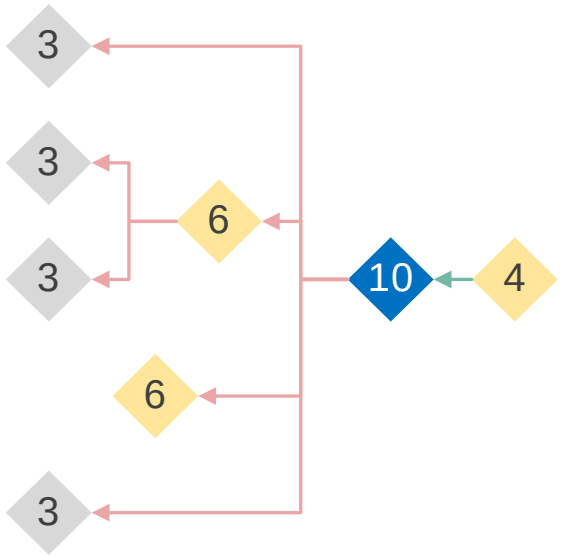

3
3
6
3
10
4
4
3
3
3
3
6
3
10
4
6
3
3
3
6
Intersect
3
14
6
3

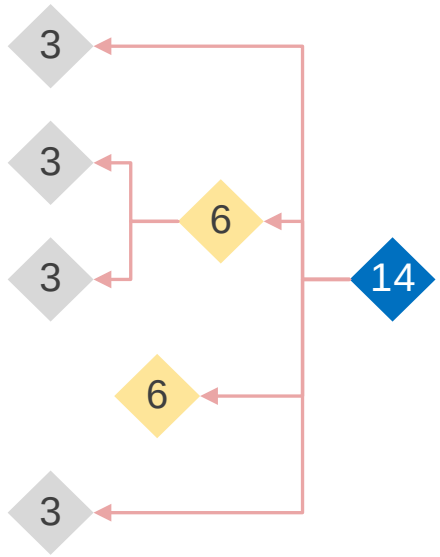

Generate significance classes
Merge lowest class process chains
Merge remaining lowest class processes with suitable neighbors
Merge junction patterns

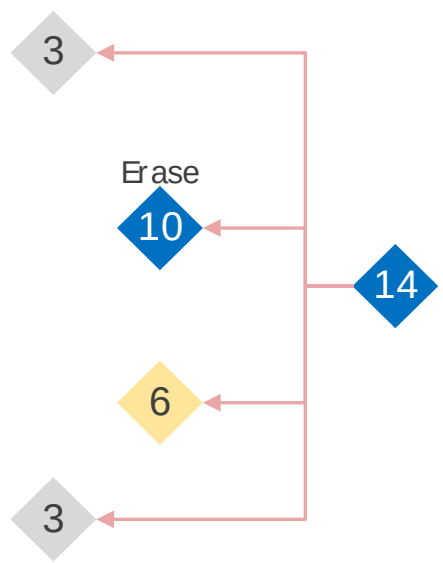

3
3
6
3
10
4
4
3
3
3
3
6
3
10
4
6
3
3
Buffer
3
Erase
SelVeg
6
3
14
6
3

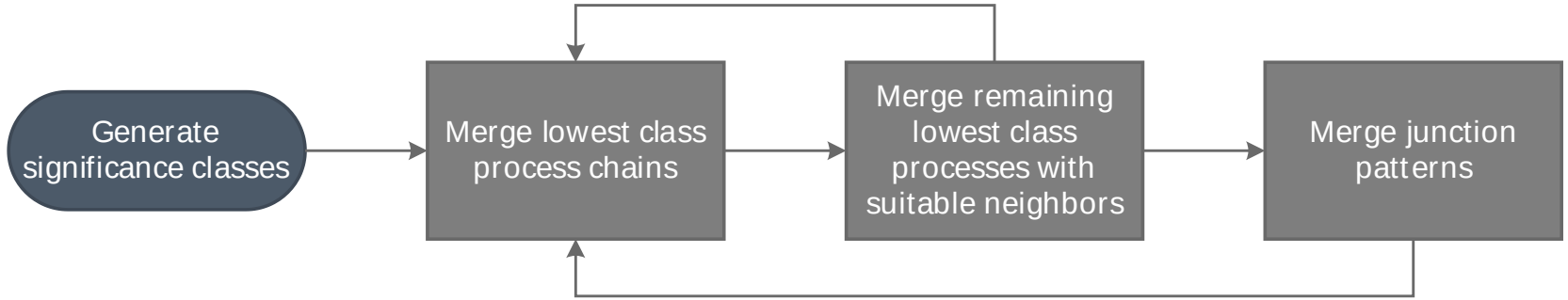

Generate significance classes
Merge lowest class process chains
Merge remaining lowest class processes with suitable neighbors
Merge junction patterns

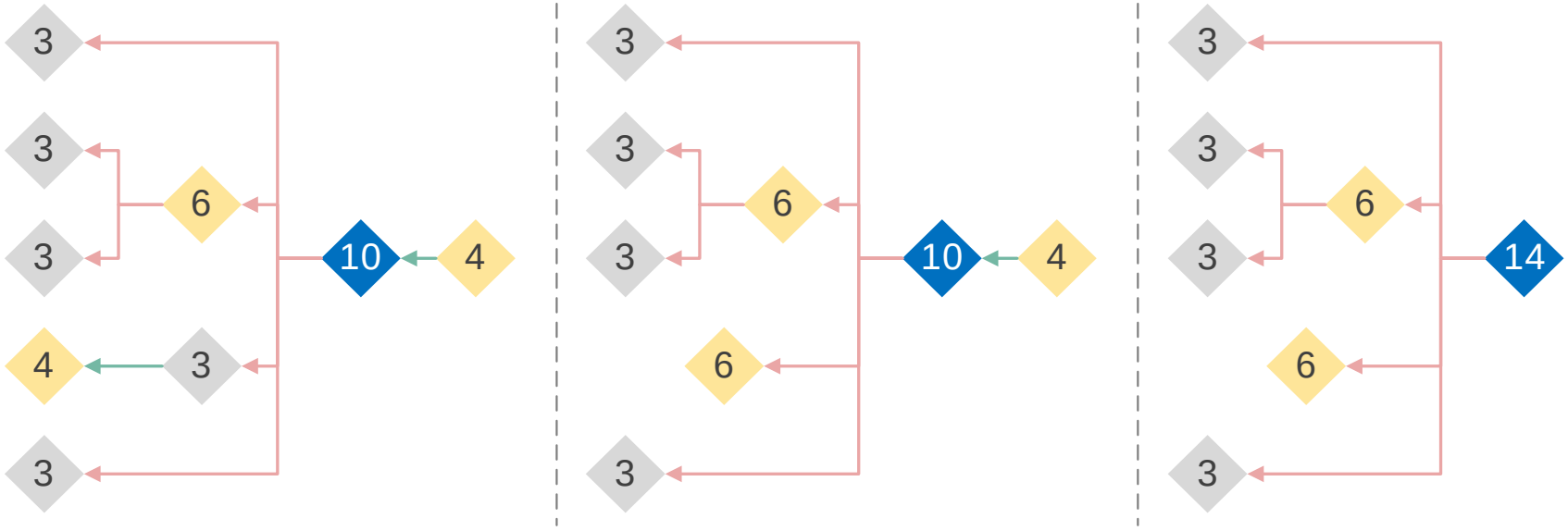

3
3
6
3
10
4
4
3
3

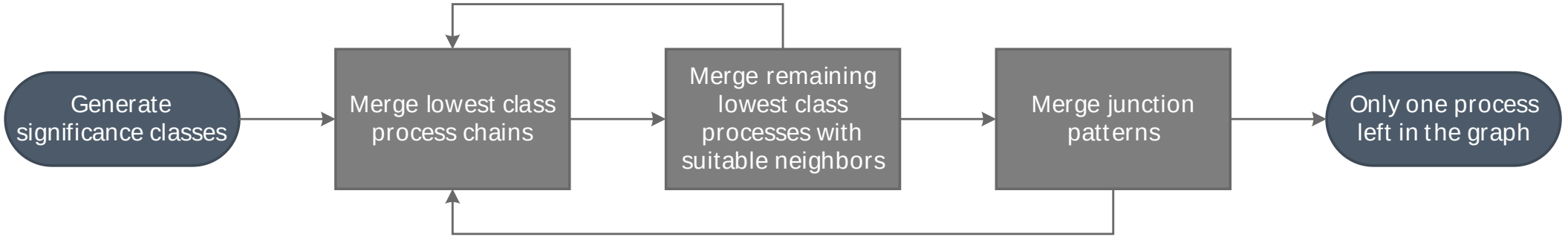

3
3
6
3
10
4
6
3

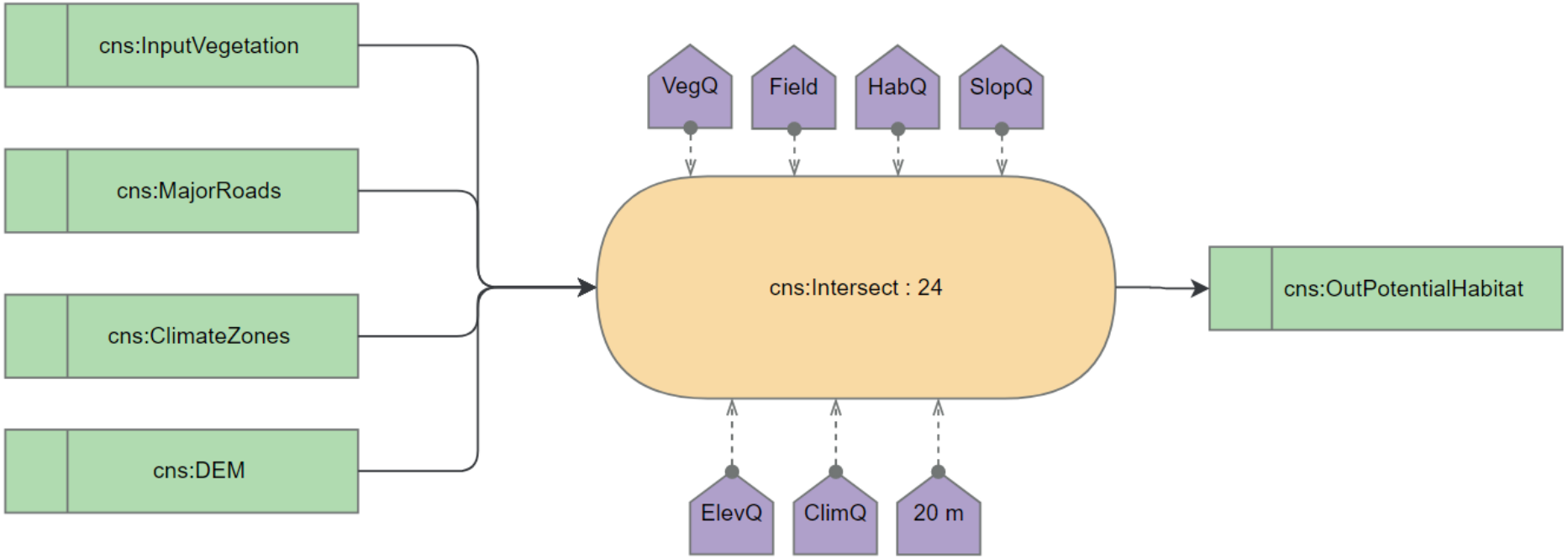

3
3
6
3
14
6
3

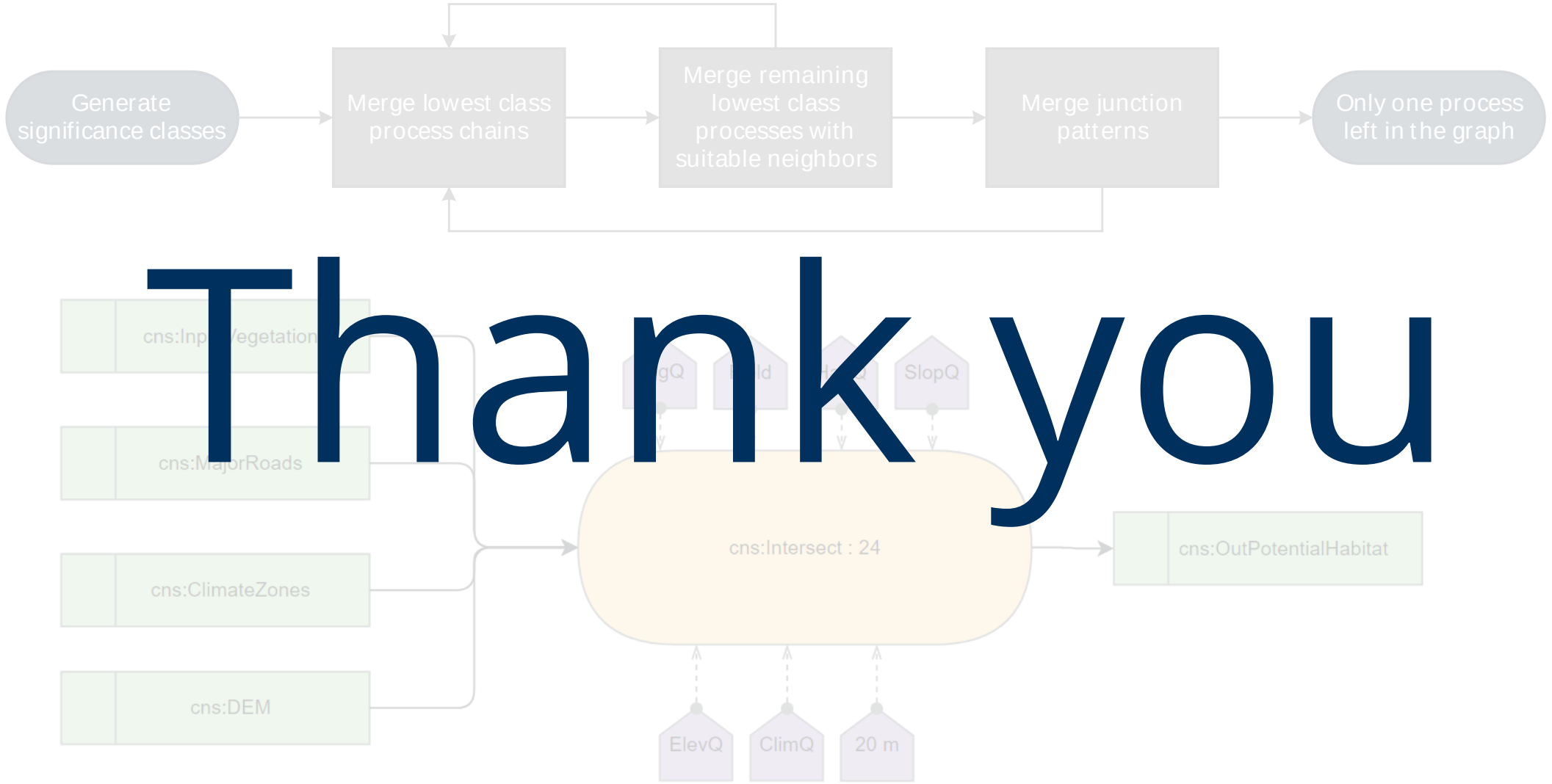

3
Erase
10
14
6
3


Generate significance classes
Merge lowest class process chains
Merge remaining lowest class processes with suitable neighbors
Merge junction patterns

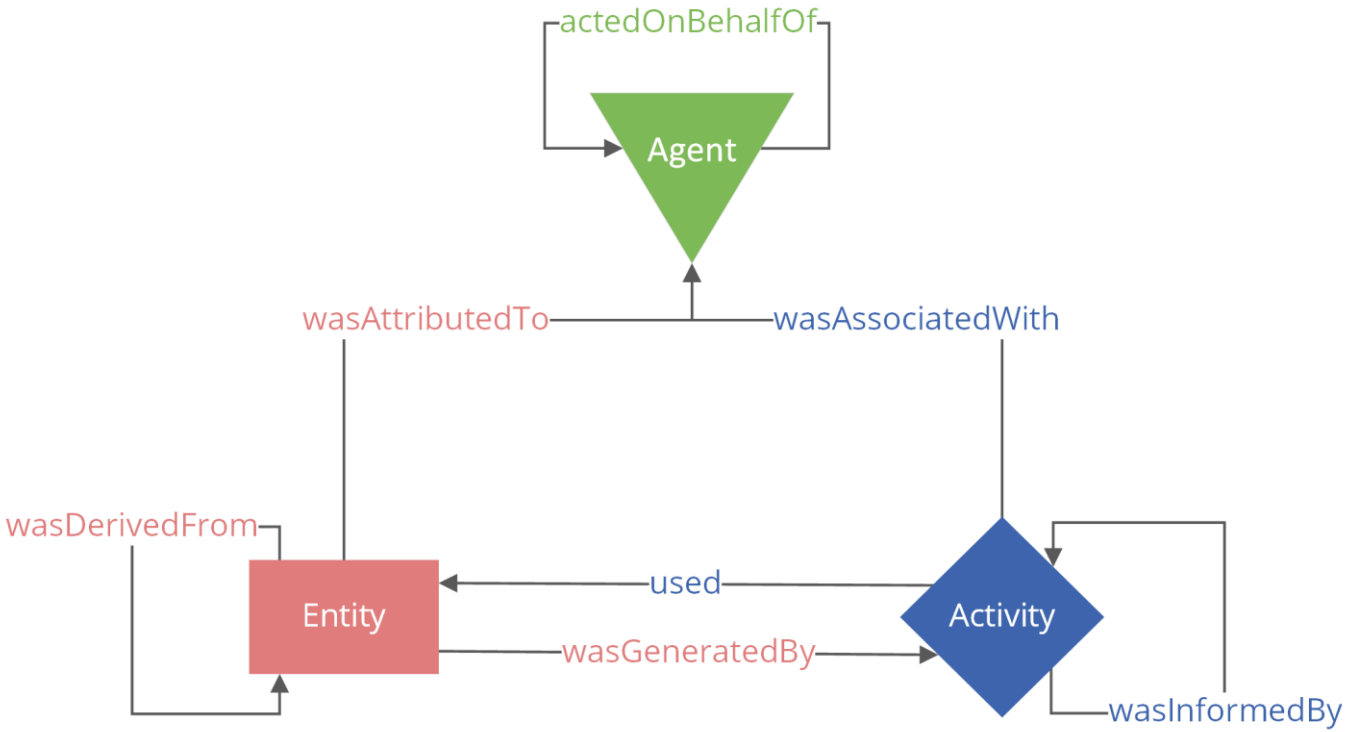

3
3
3
6
10
4
4
3
3
3
3
3
6
10
4
6
3
3
3
6
14
SelClim
3
Erase
10
Intersect
14
Slope
6
SelElev
3

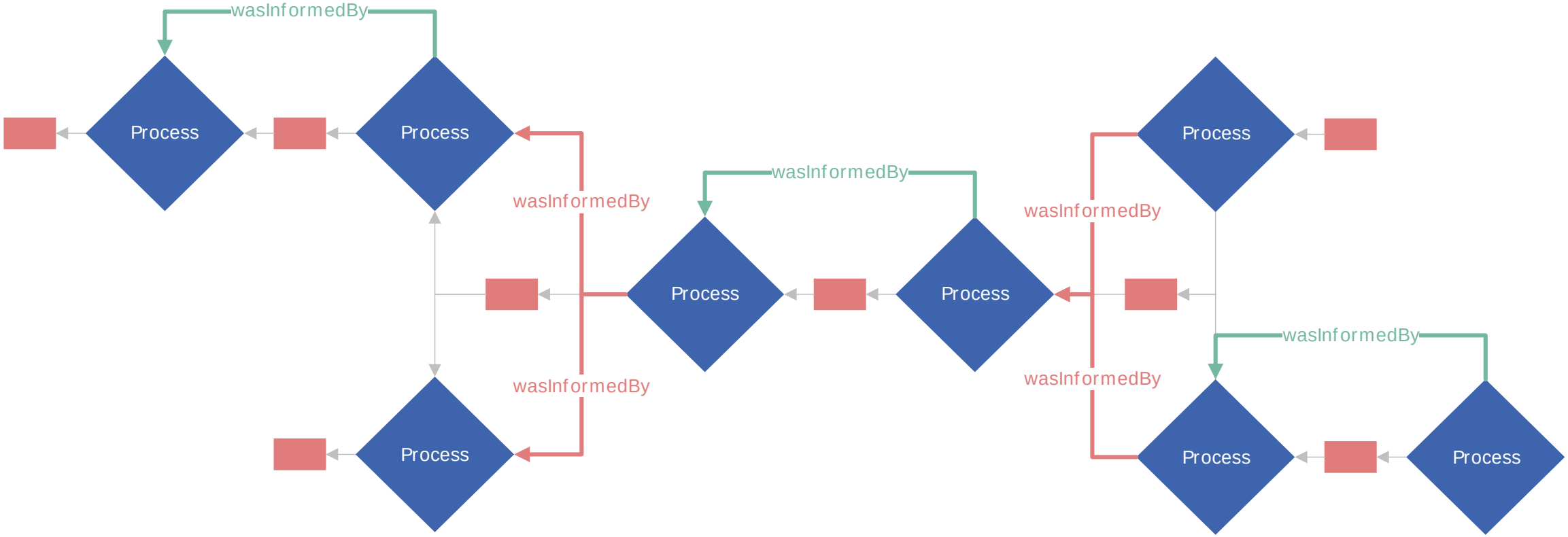

Generate significance classes
Merge lowest class process chains
Merge remaining lowest class processes with suitable neighbors
Merge junction patterns
Only one process left in the graph

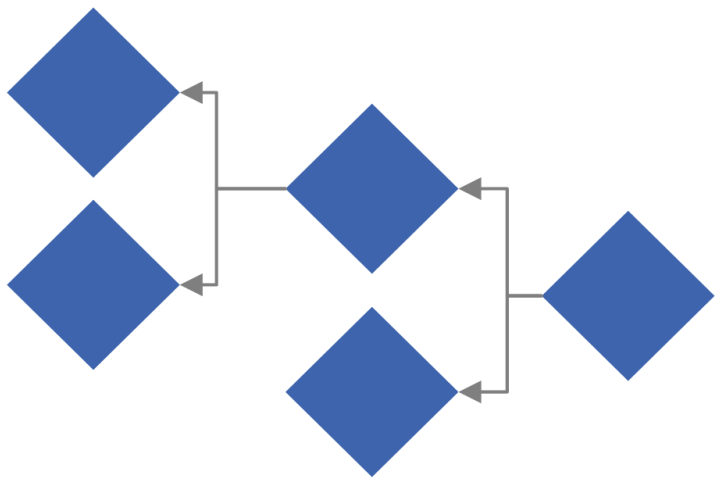

cns:InputVegetation
cns:MajorRoads
cns:ClimateZones
cns:DEM
VegQ
Field
HabQ
SlopQ
cns:Intersect : 24
cns:OutPotentialHabitat
ElevQ
ClimQ
20 m


Generate significance classes
Merge lowest class process chains
Merge remaining lowest class processes with suitable neighbors
Merge junction patterns
Only one process left in the graph
cns:Inp Vegetation
cns:MajorRoads
cns:ClimateZones
cns:DEM
SlopQ
cns:Intersect : 24
cns:OutPotentialHabitat
ElevQ
ClimQ
20 m

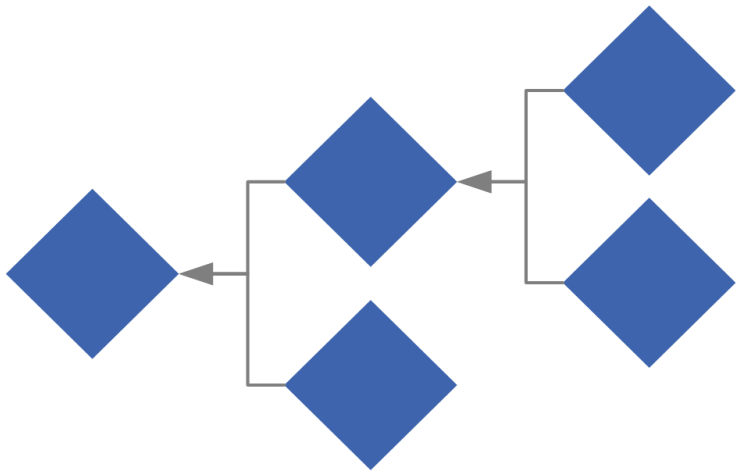
Thank you


actedOnBehalfOf
Agent
wasAttributedTo
wasAssociatedWith
wasDerivedFrom
Entity
used
Activity
wasGeneratedBy
wasInformedBy

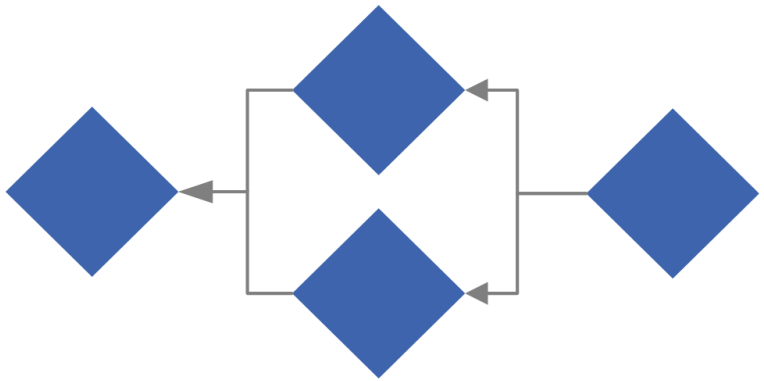

wasInformedBy
Process
Process
wasInformedBy
wasInformedBy
Process
Process
wasInformedBy
Process
wasInformedBy
wasInformedBy
Process
Process
Process



nested



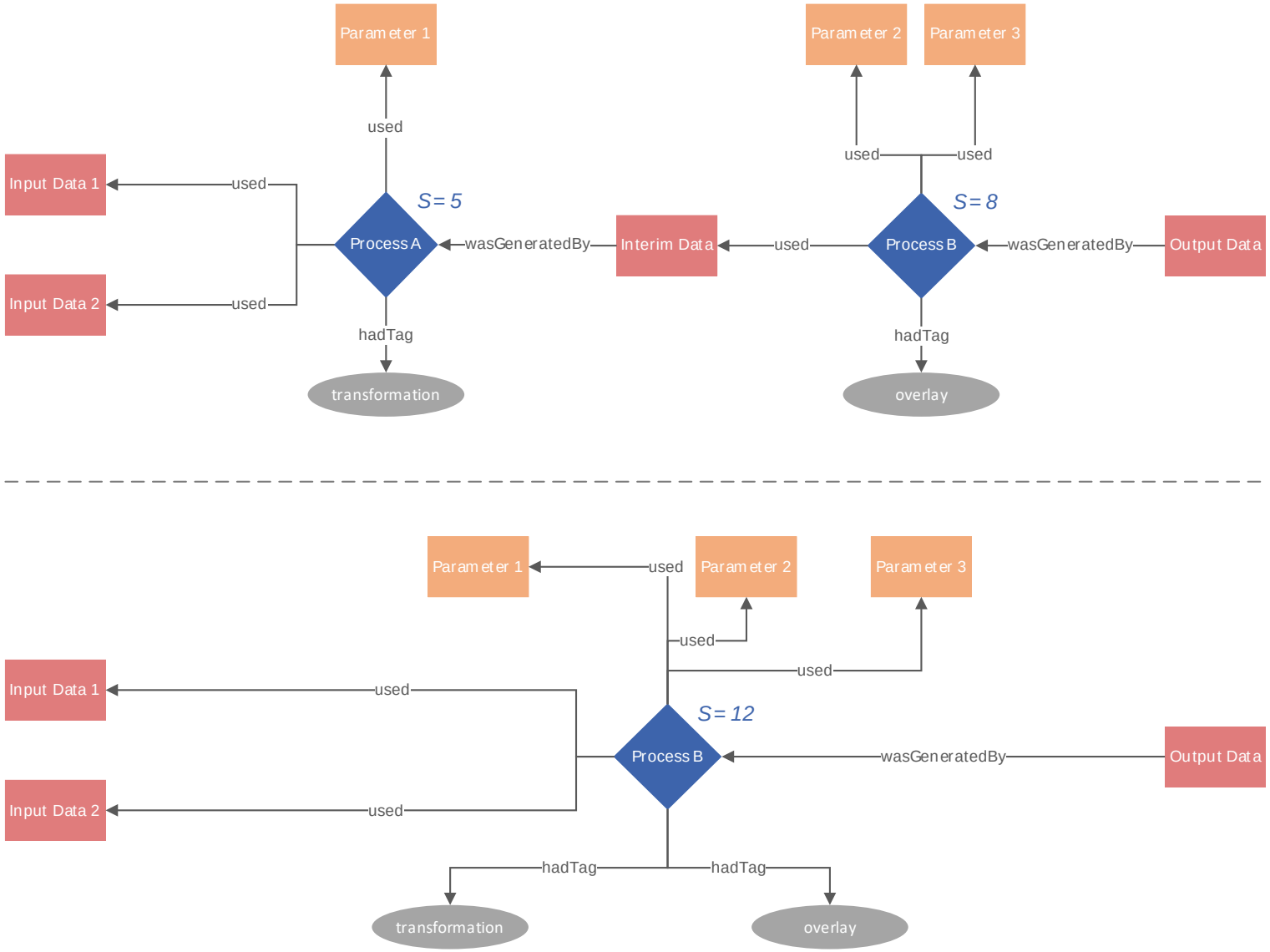
nested

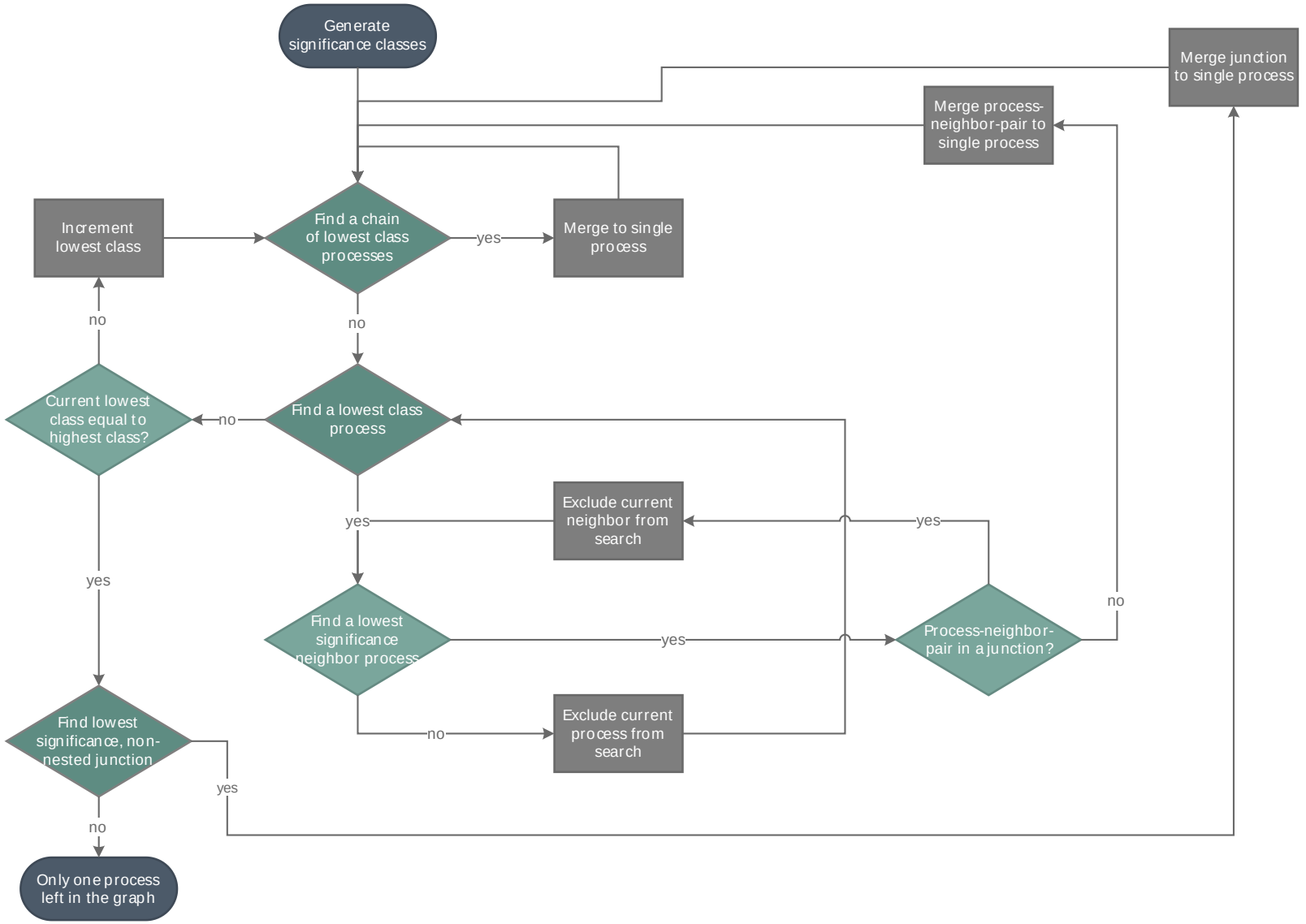


not nested

Merging Rules

- Any interim data between the merged Processes is ignored. Interim data is data that is at the same time input data and output data (only of the Processes that are getting merged).
- The input data of the new Process equals the source data of the merged units. Source data is input data that is no interim data
- The output data of the new Process is equal to the final data of the merged units. Final data is output data that is no interim data.
- The new Process has all parameters of the merged units.
- The new Process has all tags of the merged processes, but without duplicates.
- (The Multiplier of a process with multiple Tags is determined by the highest tag−group (*prepare, analyze, model*) present in the Tags.)
- The name of the process of the new Process is equal to the most significant merged process.


Parameter 1
used
Input Data 1
used
S = 5
Process A
wasGeneratedBy
Interim Data
used
Process B
wasGeneratedBy
Output Data
Input Data 2
used
hadTag
transformation
Parameter 2
Parameter 3
used
used
S = 8
hadTag
overlay
Parameter 1
used
Parameter 2
Parameter 3
used
used
Input Data 1
used
S = 12
Process B
wasGeneratedBy
Output Data
Input Data 2
used
hadTag
hadTag
transformation
overlay


Generate significance classes
Merge junction to single process
Merge process-neighbor-pair to single process
Increment lowest class
Find a chain of lowest class processes
yes
Merge to single process
no
no
Current lowest class equal to highest class?
no
Find a lowest class process
yes
Exclude current neighbor from search
yes
yes
Find a lowest significance neighbor process
yes
Process-neighbor-pair in a junction?
no
Exclude current process from search
no
Find lowest significance, non-nested junction
yes
no
Only one process left in the graph

Group	Tag (original)	Multiplier	Description	Examples (non-comprehensive)
prepare	selection (Selection)	1	Selection operations retrieve data from the database, [...].	Search, Logical/Arithmetic/Algebraic Condition, Suppress, Filter, Data Retrieval
	transformation (Transformation)	1	Transformation operations transform selected data for scale, orientation, projection, and presentation.	Scale Change, Orientation, Projection Change, Georeferencing, Map Registration, Resampling, raster-vector-conversion
	reclassification (Reclassification)	1	Reclassification operations assign new attribute values to a set of objects based on initial attribute values, geometry or topological relations.	Filtering, Merge Attributes, Generalization
	geometric (Geometric Computation)	1	Geometric Computation operations calculate simple geometric values like area, length, perimeter, and distance, but also more complex calculations like buffer zone, and average distance.	Area, Length, Perimeter, Distance, Buffer Zone, Generalization, Centroid, Dissolve Lines, Thiessen, Merge, Surface Calculation, Image Enhancement/Edge Detection
	visualization (Visualization)	1	Visualization covers any data visualization technique. Used in data explorations, process validation, result visualization, etc.	Block Diagram, Scene Generation, Cross Section, Histogram, Choropleth Map, Error Bars

analyze	neighborhood (Neighborhood Analysis)	2	Neighborhood operations use neighborhood information to calculate slope, diversity, profile, and interpolation.	Slope, Diversity, Interpolation, Proximity, Complex Generalization, Connectivity, Adjacency, Nearest Neighbor, Diffusion Model, Line-of-Sight, Pattern Analysis, Viewshed
	overlay (Map Overlay)	2	Map Overlay combines two or more maps; Both spatial and thematic attribute values can be combined.	Line-on-Polygon, Polygon Overlay, Corridors, Line Intersection, Route Allocation, Map Algebra, Optimal Location
	network (Transitive Closure Analysis)	2	Transistive Closure Analysis [...] calculate shortest path, travel time, and maximum flow. This can be done on a line graph or a polygon network.	Shortest Path, Travel Time, Connectivity, Maximum Flow/Flow Analysis, Drainage Network, Watershed Boundaries, Diffusion Model, Catchment
	statistics (Statistics)	2	Statistics operations calculate statistical characteristics like percentage, distribution, frequency, and correlation of objects of different classes over a specified area.	Percentage, Distribution, Frequency, Correlation, Spatial Statistics, Pattern Analysis, Complexity and Variation Measures
model	model	3	Model describes any black-box module. It is also applicatable for scripted processing steps that can not be abstracted by the categories above.	Neural Network,