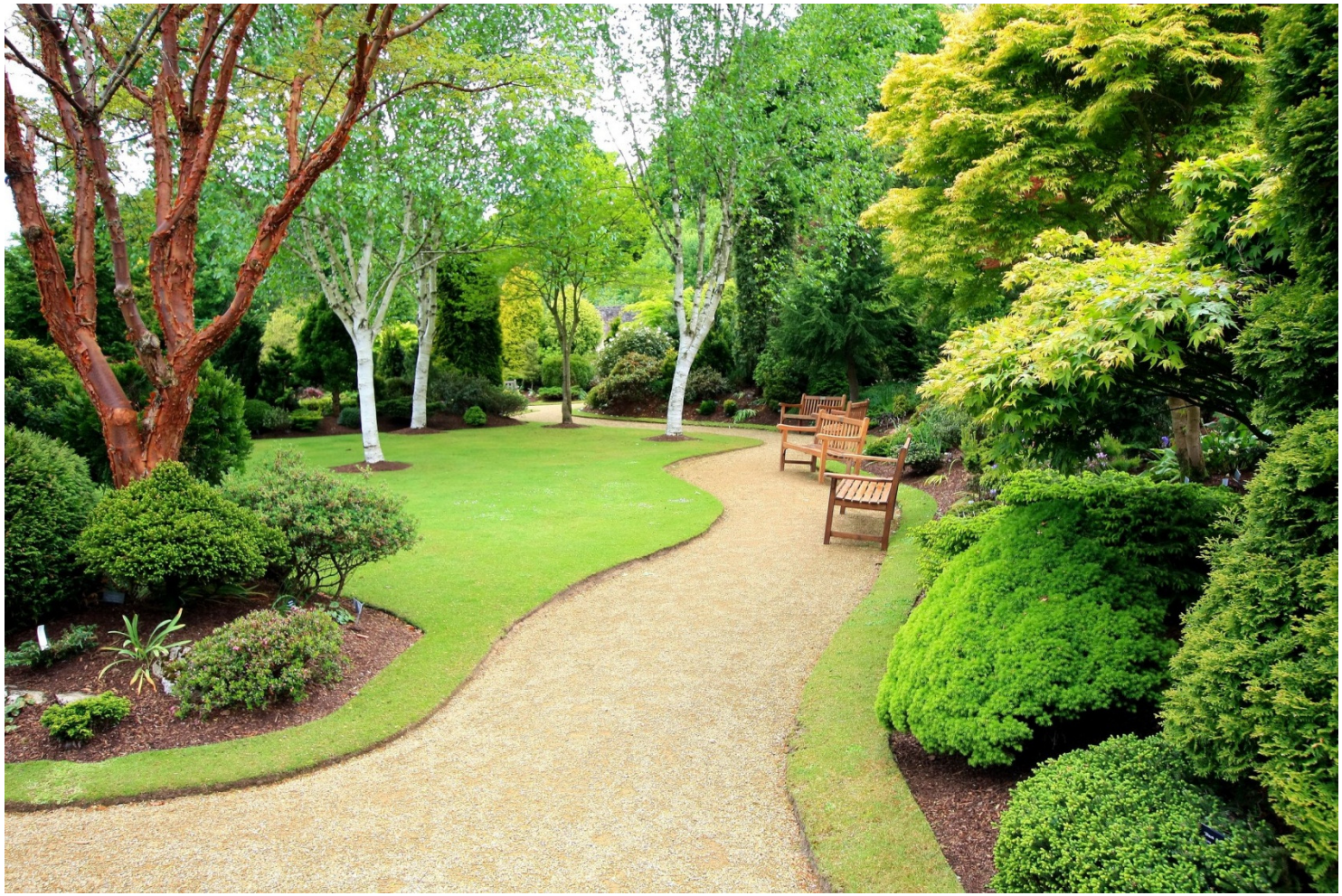
Data Landscaping to Support Coordination at Scale

Peter Gates

Janssen Research & Development LLC.

Note: The views expressed herein are solely those of the author, and should not be construed as representing the views of the author's employer.

What We Want



What We Have

The Problem

- Pharmaceutical R&D rising costs
- Cost inflation largely attributed to clinical trials

Informatics Opportunity

- Drive down cost associated with operations
  - Primary use of data
- Improve ability to translate historical and preclinical data into an increase in probability of clinical success
  - Secondary use of data
- Assemble public and internally authored data in support of licensing, mergers and acquisitions.

Coordination as Communication

1. Internal thoughts
   a. Qualia (perceptions, sensations, reactions, moods)
   b. Ideas
2. Direct and interactive
   a. An interactive face to face meeting
   b. Q&A session between an "authority" and a group of interested individuals
3. Direct and non-interactive
   a. Speech
   b. Lecture/presentation
4. Indirect (technology mediated) and interactive
   a. Phone conversation (synchronous)
   b. Instant messenger (synchronous)
   c. E-mail exchange (asynchronous)
5. **Indirect (technology/artifact mediated) and non-interactive**
   a. Communication at scale
      i. Across time
      ii. Across geography
      iii. Across large audiences

The Structure of Language

- Grammar
  - Rules for assembling instances of lexical categories (nouns, verbs, adjectives, …) into communication artifacts
- Vocabulary
  - Sets of words assigned to lexical categories.

Grammatical vs. Lexical Meaning

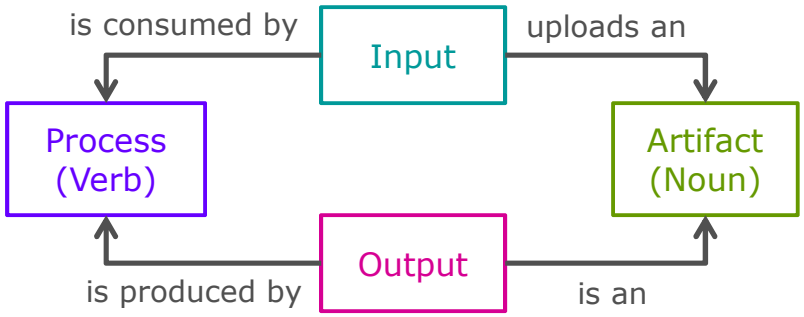
Colorless green ideas sleep furiously (Chomsky)

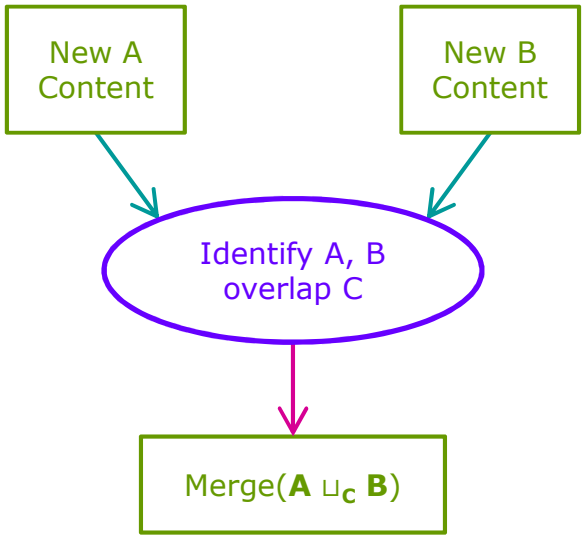
It can only be the **thought** of **verdure** to come, which prompts us in the autumn to buy these **dormant** **white** lumps of vegetable matter covered by a brown papery skin, and lovingly to plant them and care for them. It is a marvel to me that under this cover they are **laboring** unseen at such a rate within to give us the sudden awesome beauty of spring flowering bulbs. While winter reigns the earth **reposes** but these **colorless** **green** **ideas** **sleep** **furiously**. (CM Street)

colorless	→	white
green	→	verdure
ideas	→	thought
sleep	→	dormant, reposes
furiously	→	laboring

Structured Authoring and Databases

- Schema as grammar
  - A network of interdependent types
- Type signature as vocabulary
  - A set of valid values available for each type of the schema.





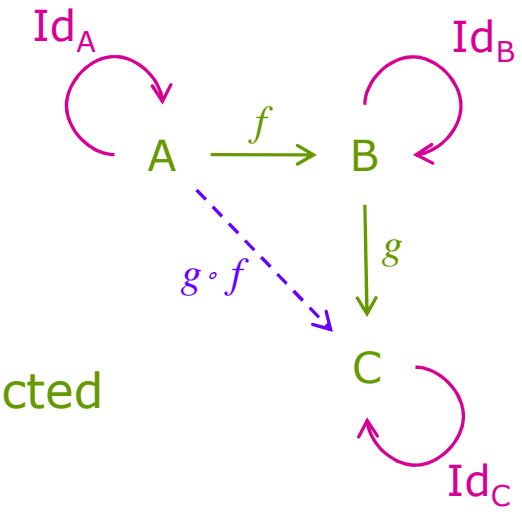
From Local to Global

- We currently have evolving strategies for building local, domain specific, databases.
- Top down attempts to define a generic database that standardizes database construction are unsatisfactory e.g. Basic Formal Ontology (BFO).
- We propose a bottom up approach that looks for overlaps between databases.

What is Data Landscaping?

1. Build technology partnerships.
2. Survey the data landscape.
3. Implement domain specific information authoring tools.
4. **Identify overlaps between domain specific content.**
5. **Integrate across domain specific content and their versions.**
6. Translating data into conclusions; knowledge discovery, feature detection, parameter estimation, annotation, …
7. Connect data authors and consumers through collaborative workflows.

Category Theory: A Theory of Structure (Schema)

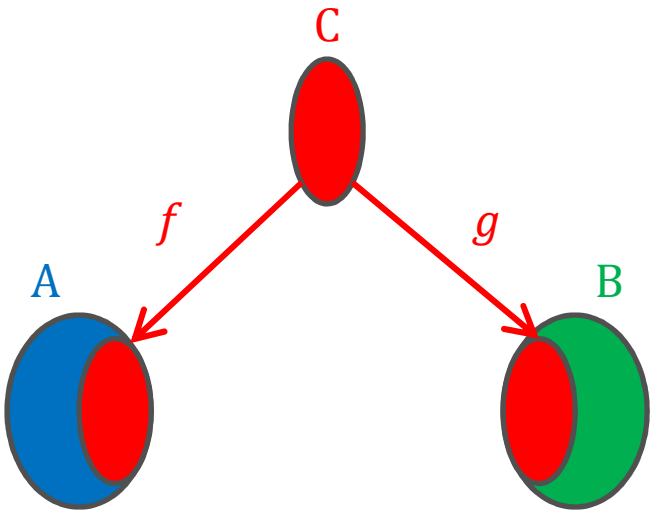


- A Category is:
  - Objects connected by arrows (a directed graph)
  - Every object has an identity arrow
  - Arrows can be composed head to tail to define paths.
  - Paths can be declared to be equivalent.

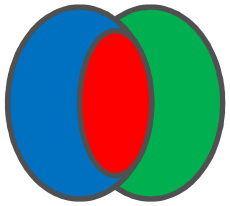
Gluing: A Categorical Construction

- Gluing captures the idea that one can assemble a global view from overlapping local views.
- There are a variety of different related constructions that arise from consideration of equivalent paths that all relate to the idea of gluing.
- In what follows we will focus on one such construction known as a pushout that abstracts the notion of the union of two sets.

Gluing A and B with C

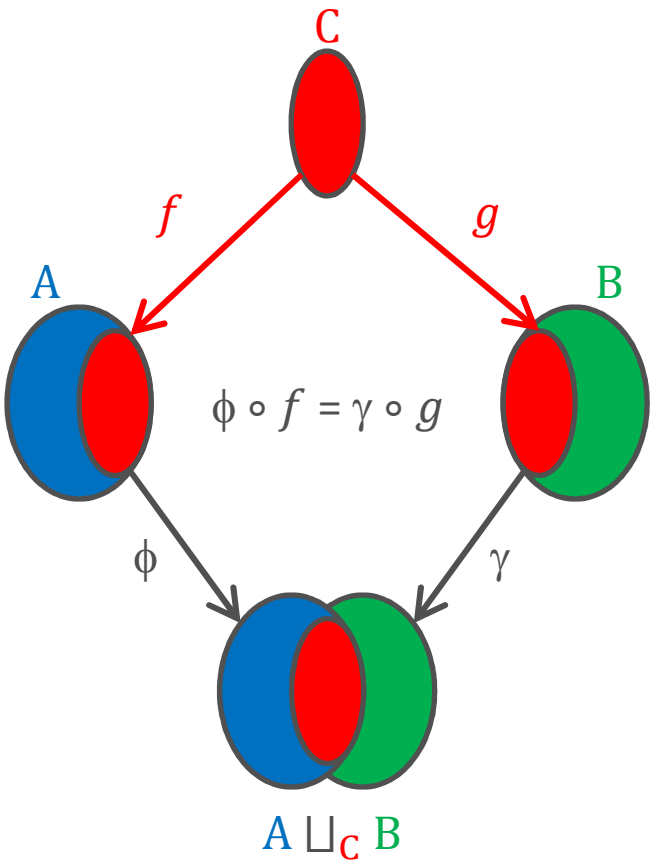


The Union of A and B with Intersection C

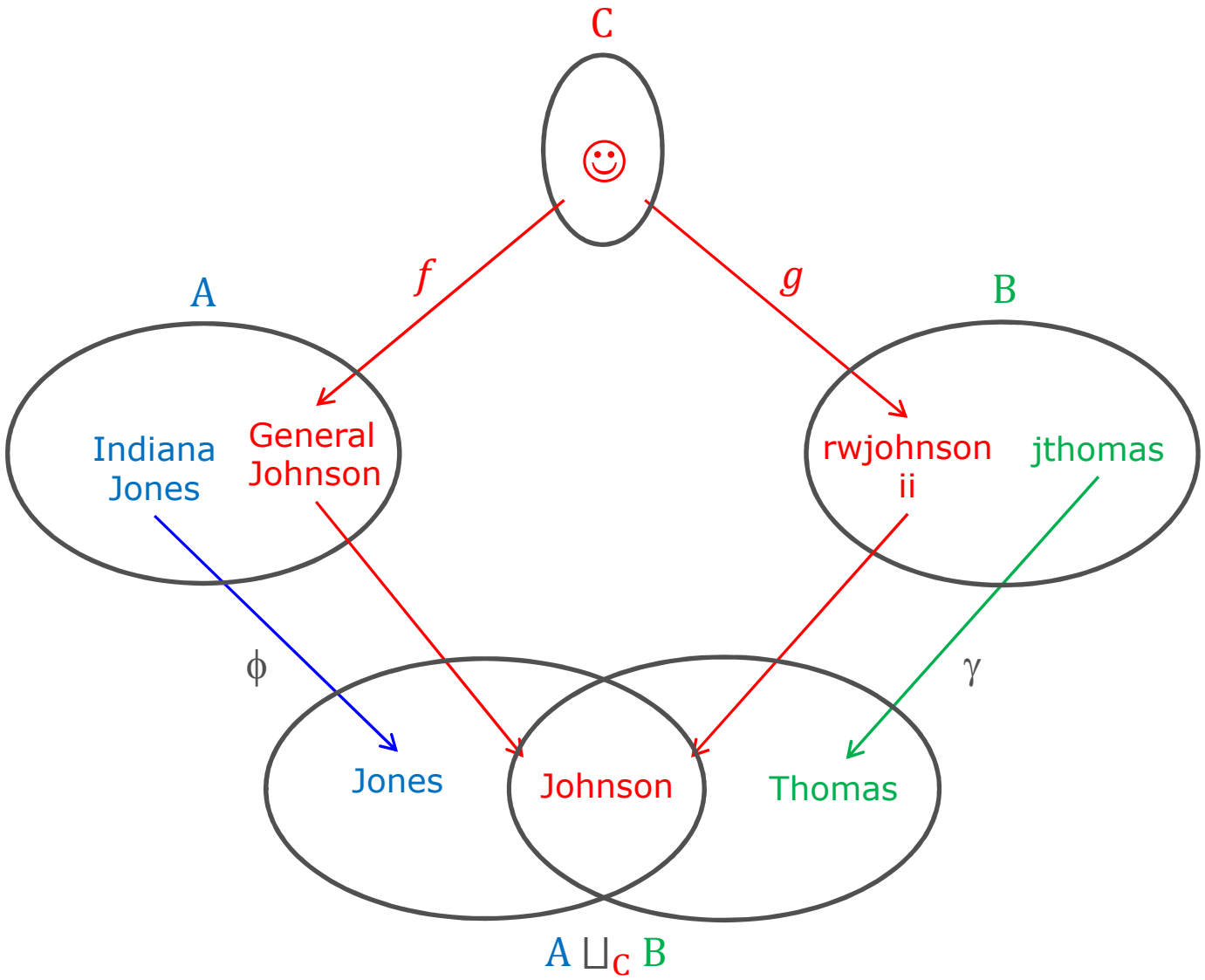


$A \sqcup_C B$

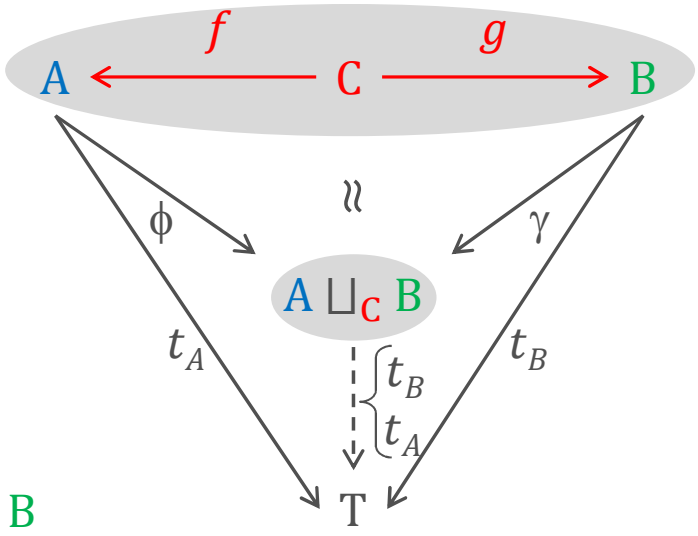
Overlaps and Gluing

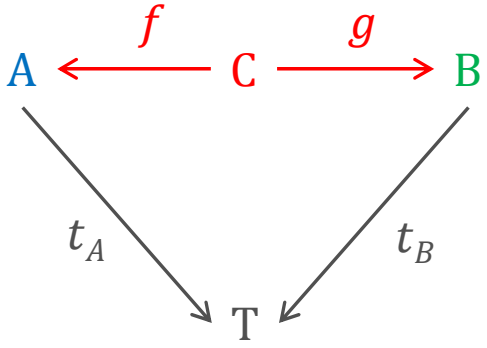


Set Union as Gluing

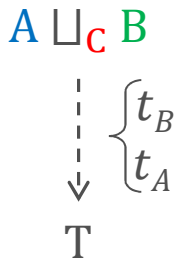


Gluing is a Universal Construction





$(t_A, t_B) \longleftrightarrow \begin{cases} t_B \\ t_A \end{cases}$

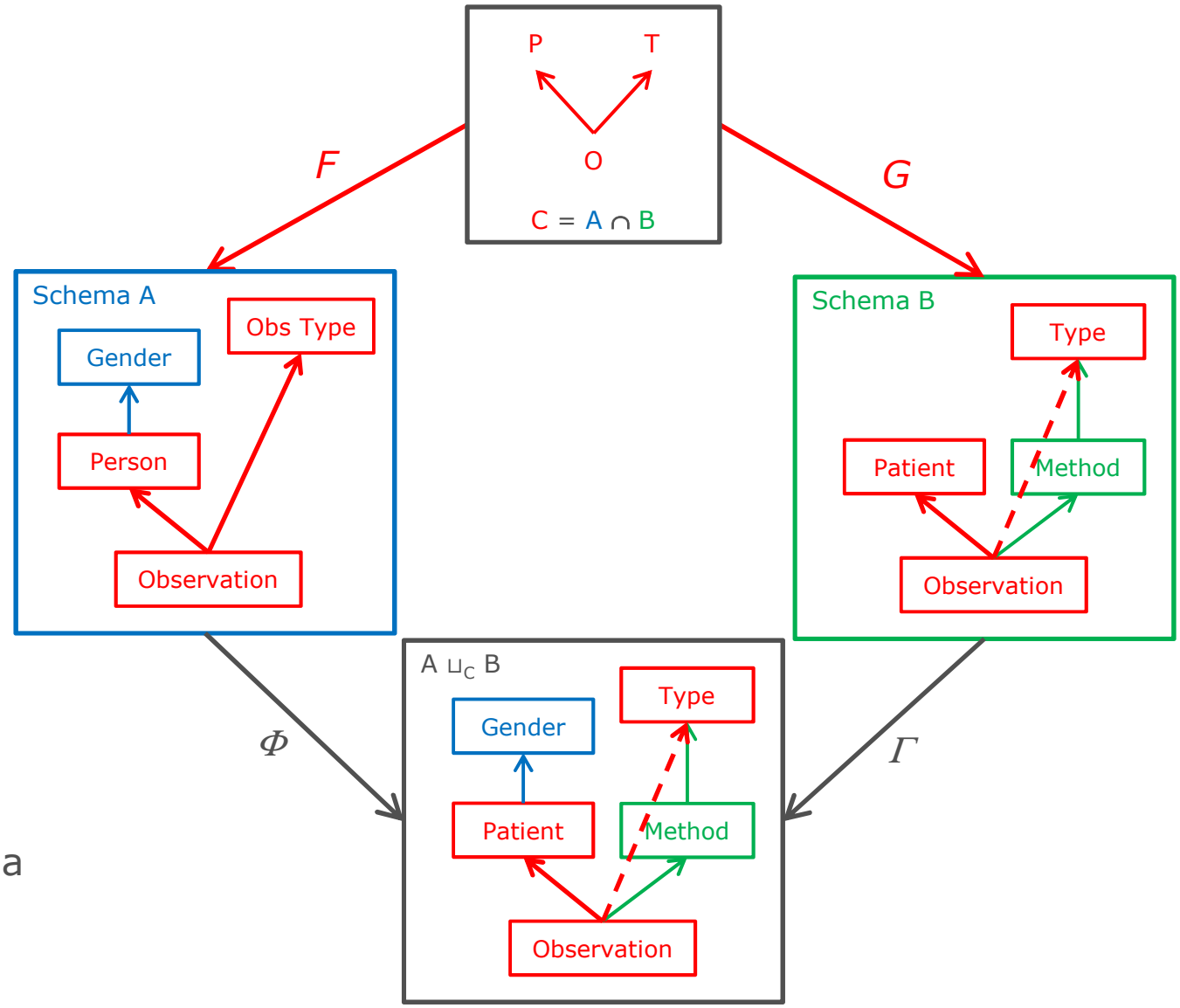


We Can Glue Schemas/Grammar!

Overlap/Glue

Source DBs

Union
Integrated Schema



We Can Glue Vocabulary Too!

- Once we have glued two or more schemas the universal construction tells us how to glue the database states.
- Where sources disagree on the overlap we need to have a model for “trusting”

How Do We Identify Overlaps?

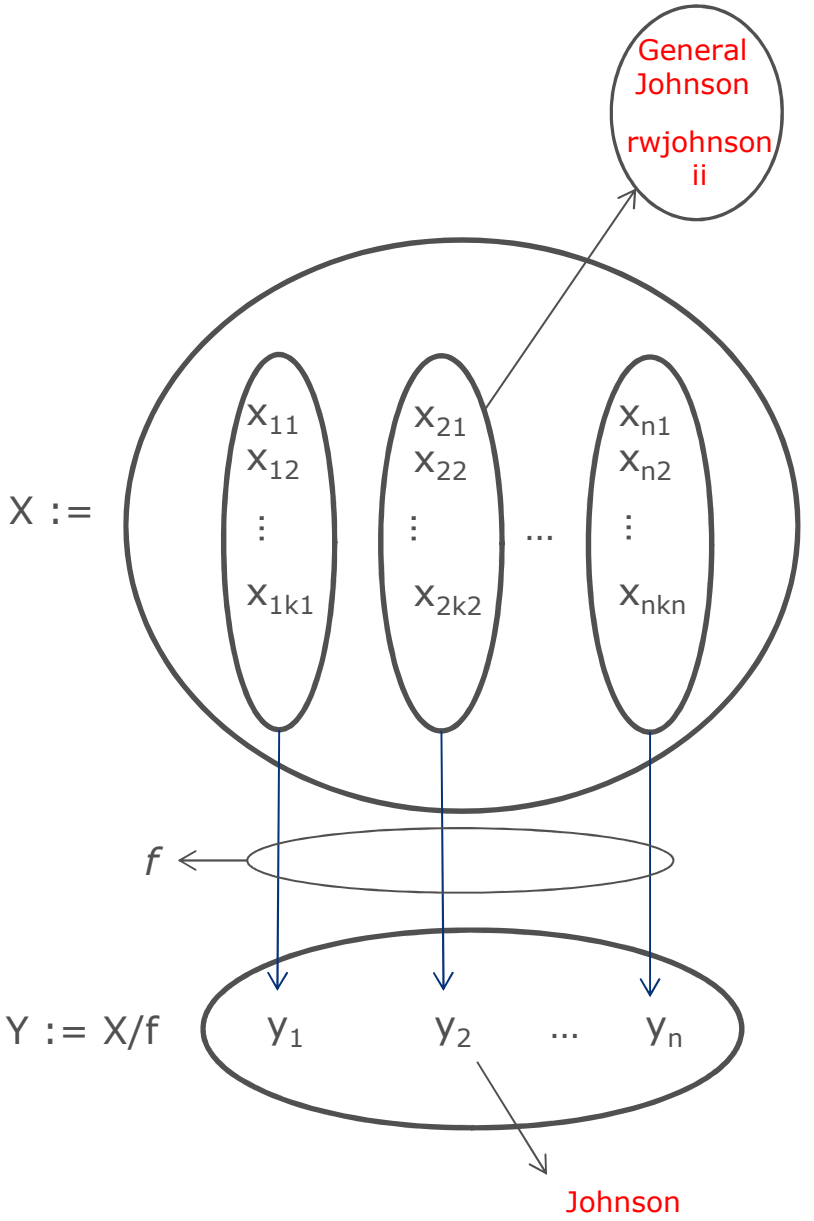
It Is All About Equality

- Given a pair of subjects, are they the same or different?
- For our discussion we will adopt a "simple" model:

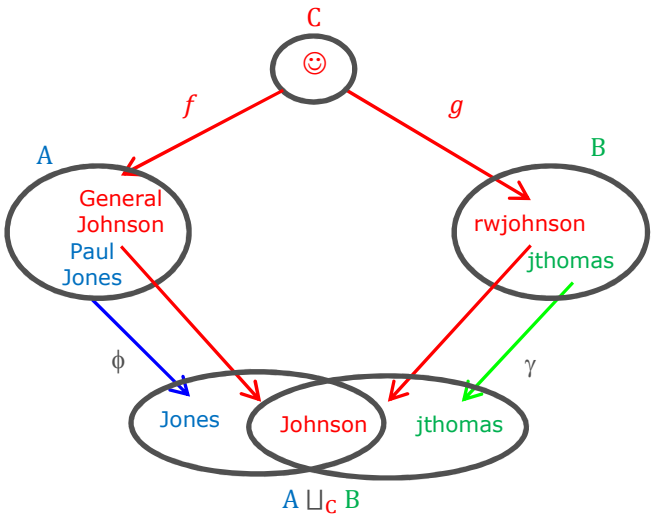
> We partition subjects into groups that are equivalent.
>
> Note: Subjects could be lexical categories/entities, properties or vocabulary instances associated with a lexical category.

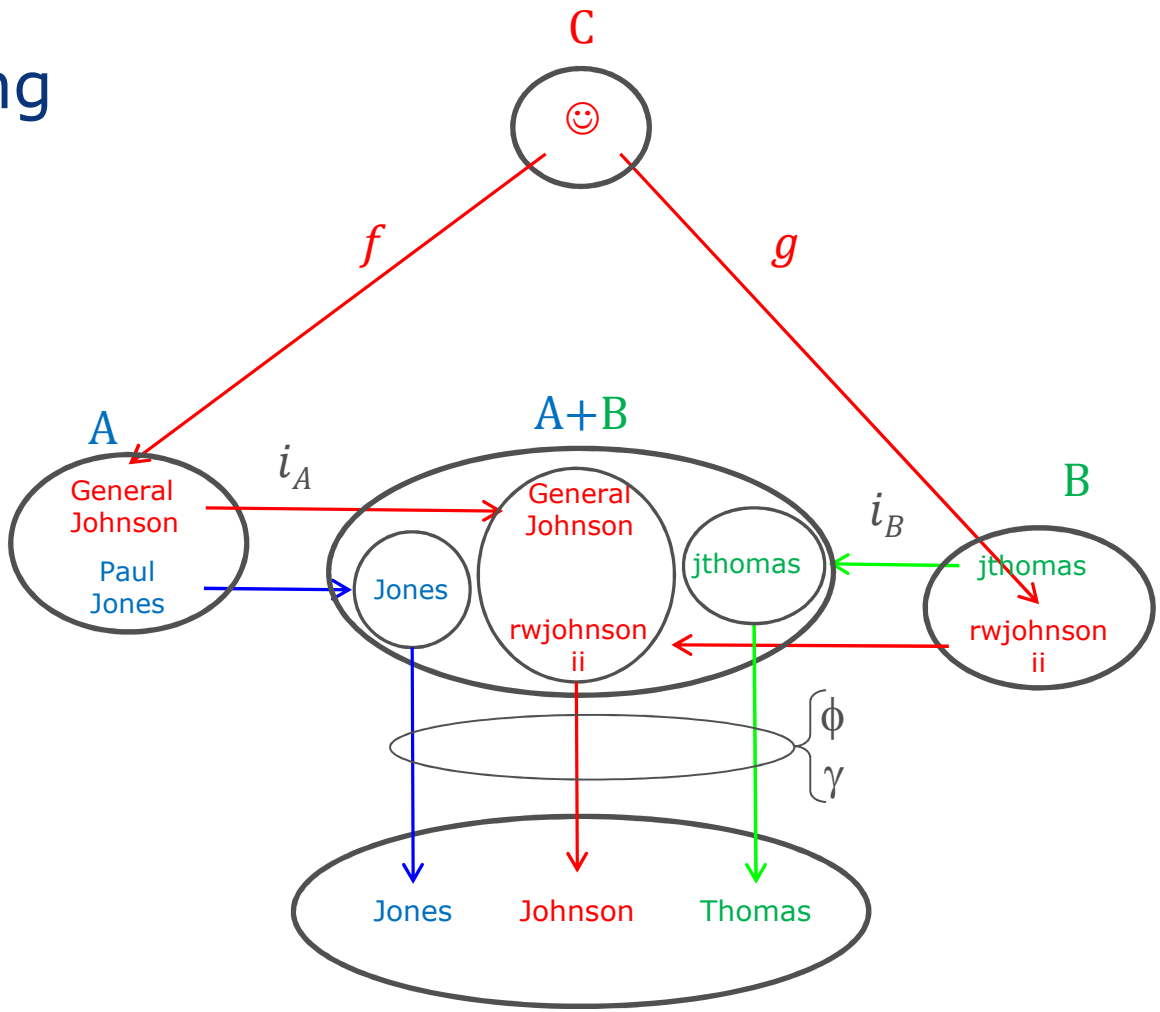
Equivalence Classes

- Each y can be considered a canonical label for a subject.
- The stalk over a y_i contains germs x_{ij} that can be considered synonyms for the equivalence class.
- This defines a functional dependency from the set X of synonyms to the set Y of canonicals.



Equivalence as Gluing





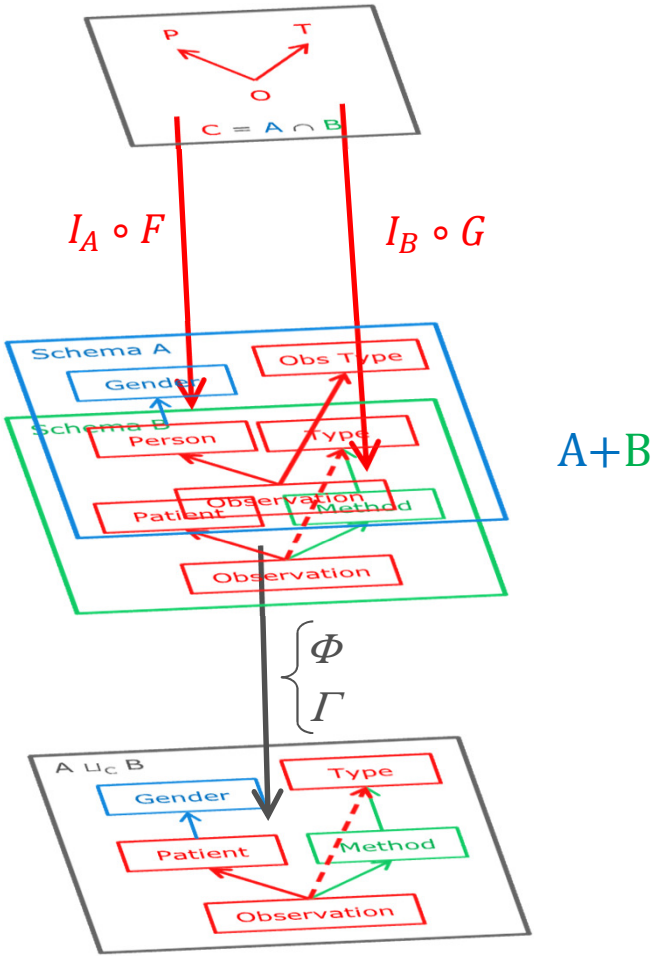
$$A \sqcup_C B \approx Coeq(i_A \circ f, i_B \circ g) := (A + B)/(i_A \circ f \sim i_B \circ g)$$

We Can Build a Quotient Database!

Overlap/Glue

Source DBs

Union
Integrated Schema
Quotient DB



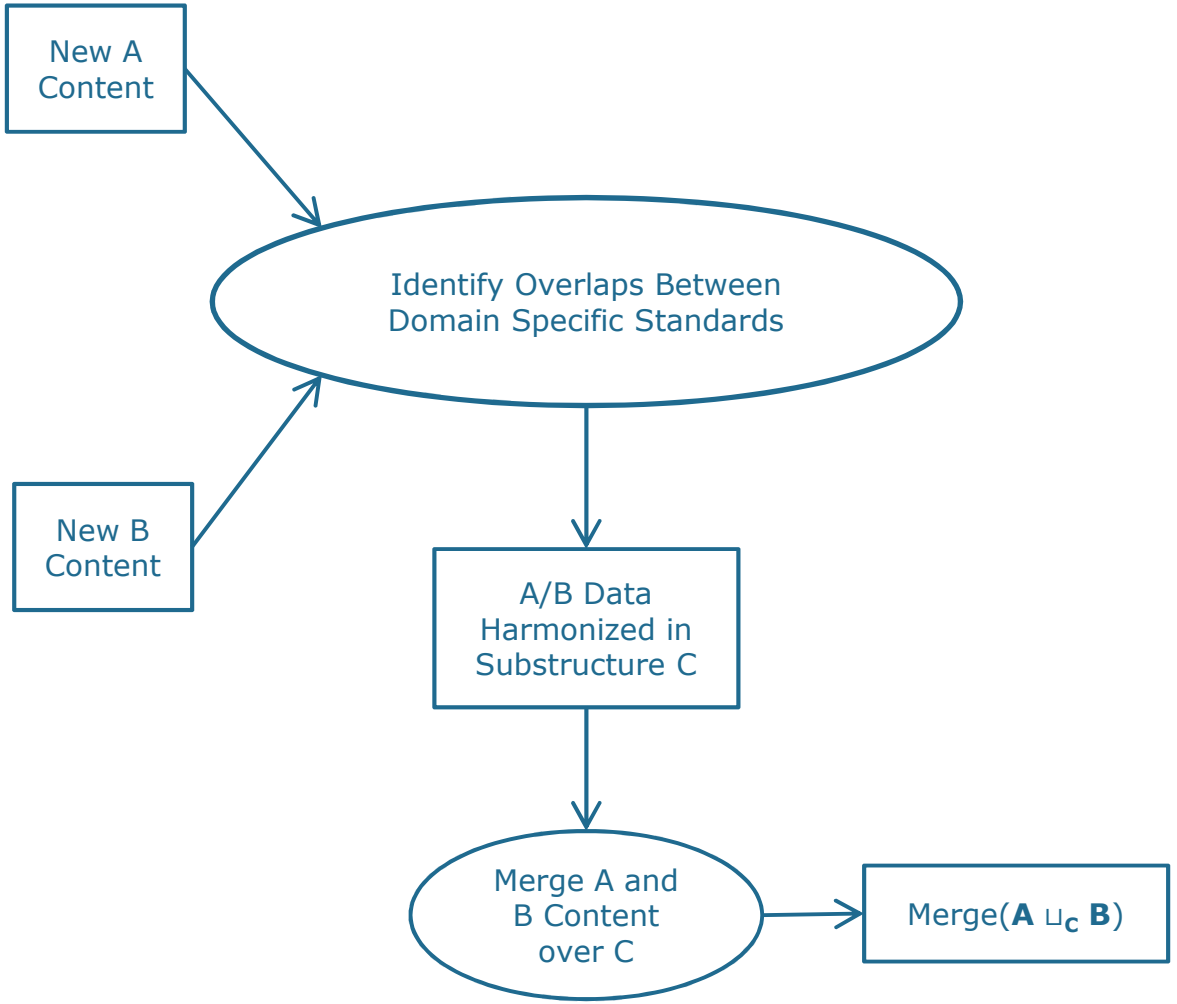
Care and Feeding

- The construction and maintenance of these classification schemes is a database application!
- Initialize the database with existing standards.
- As new unidentified subjects arrive attempt to match to existing synonyms.

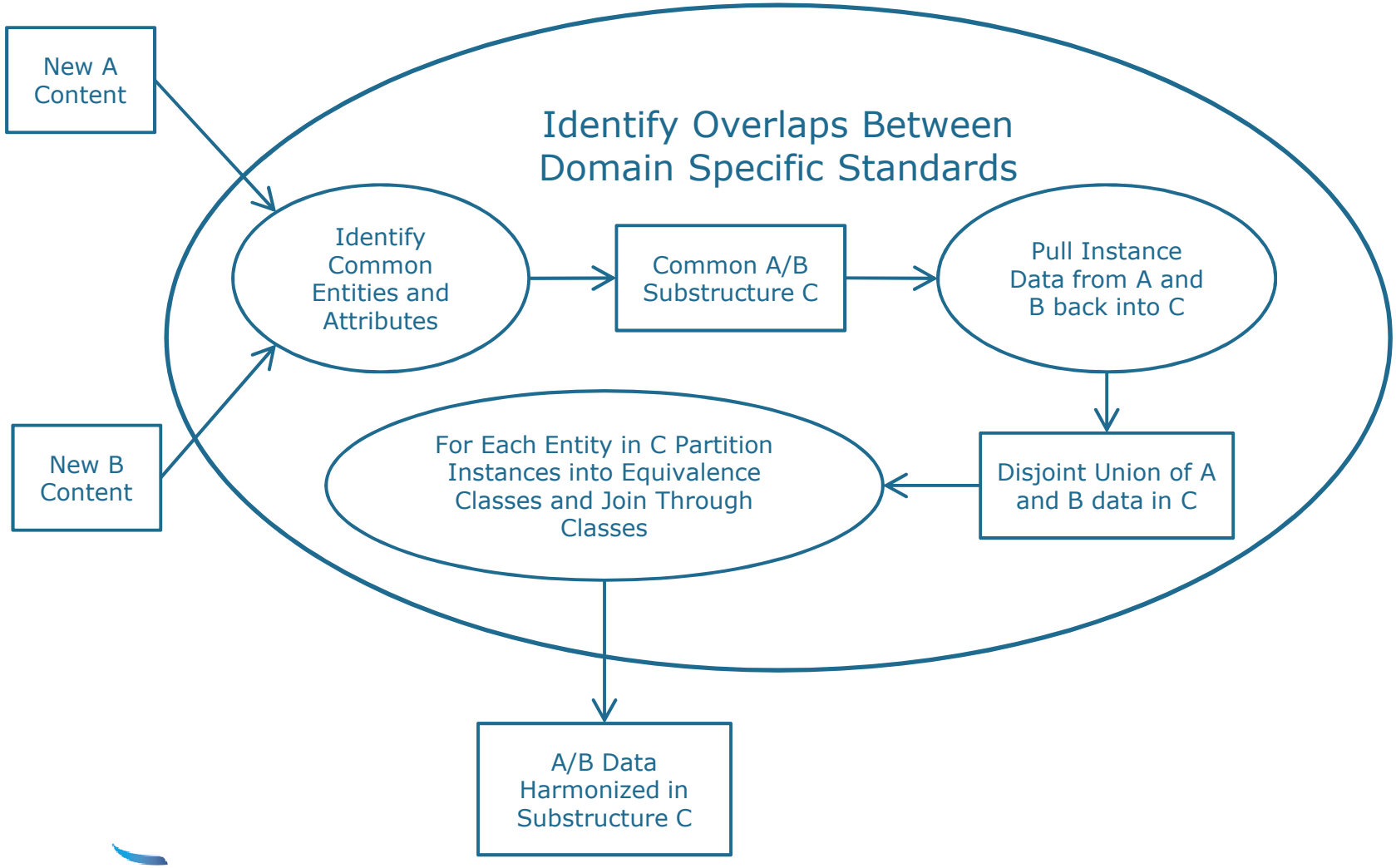
Matching

- Generate match candidates using exact, similarity or fuzzy matching.
- Present a subset of candidate matches to subject mater experts.
- Use SME input to train automated classification algorithms.

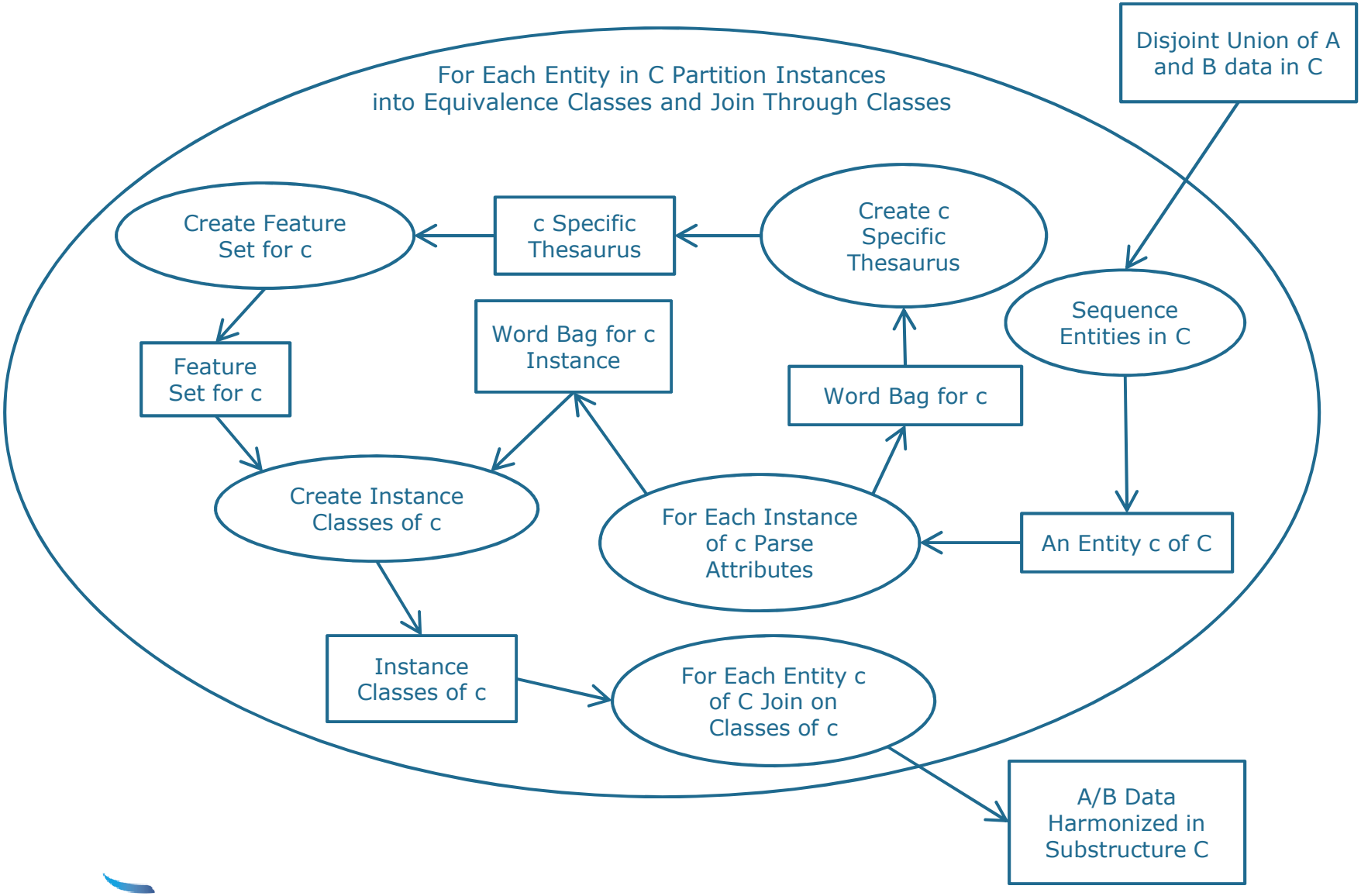
Data Landscaping Data Flow: 4 & 5



Identify Overlaps



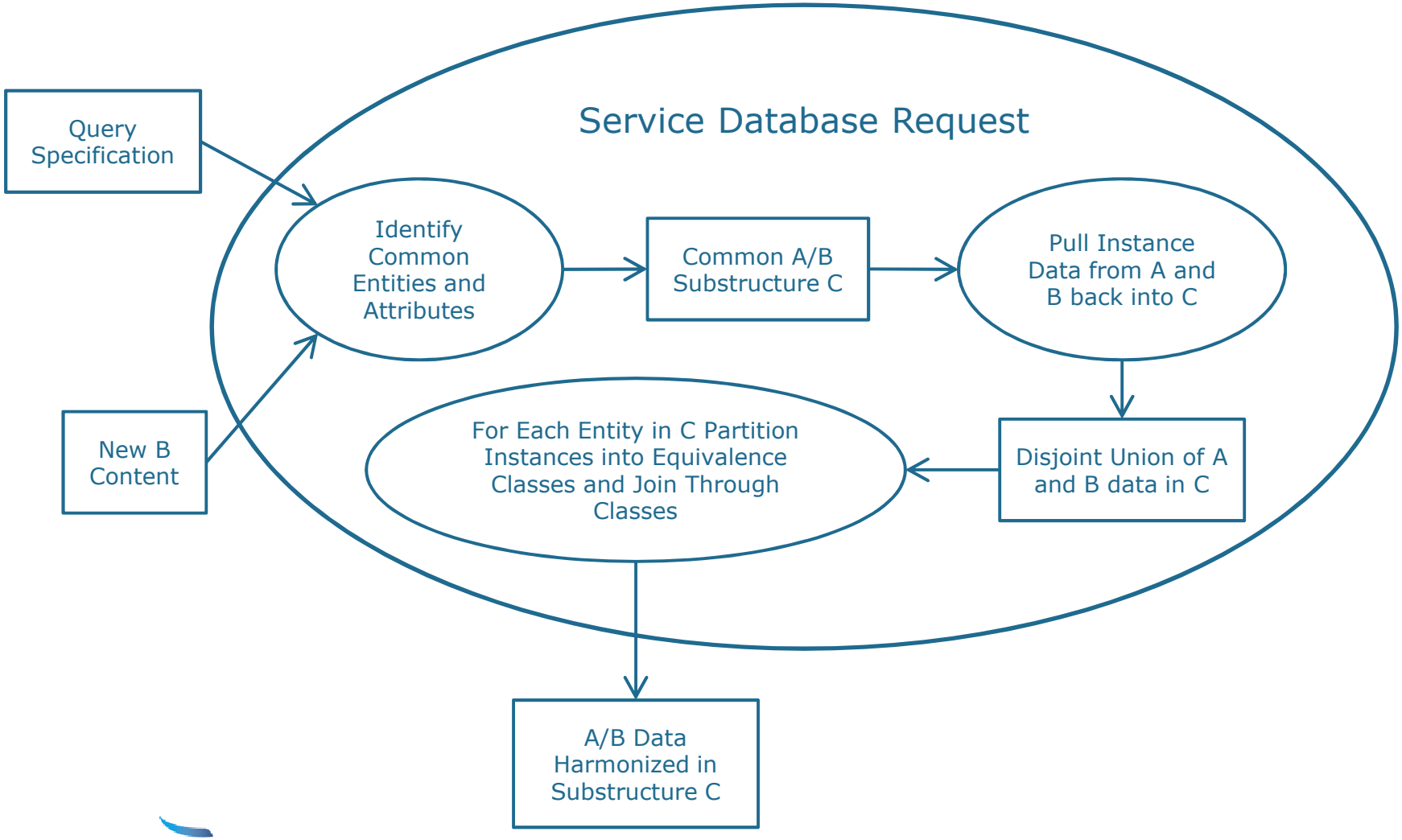
Matching


For Each Entity in C Partition Instances
into Equivalence Classes and Join Through Classes
Disjoint Union of A and B data in C
Create Feature Set for c
c Specific Thesaurus
Create c Specific Thesaurus
Sequence Entities in C
Feature Set for c
Word Bag for c Instance
Word Bag for c
Create Instance Classes of c
For Each Instance of c Parse Attributes
An Entity c of C
Instance Classes of c
For Each Entity c of C Join on Classes of c
A/B Data Harmonized in Substructure C

What is Data Landscaping Revisited

1. Build technology partnerships.
2. Survey the data landscape.
3. Manage recipes
    1. Implement domain specific information authoring tools.
    2. Connect data authors and consumers through collaborative workflows.
4. **Manage database fibers**
    1. **Identify overlaps between domain specific standards.**
    2. **Integrate domain specific standards and their versions.**

Database Management Systems (DBMSs)



Acknowledgements

David Spivak
MIT Math Department

Gluing Can Also Merge Elements