

Abbreviation Expansion in Lexical Annotation of Schema

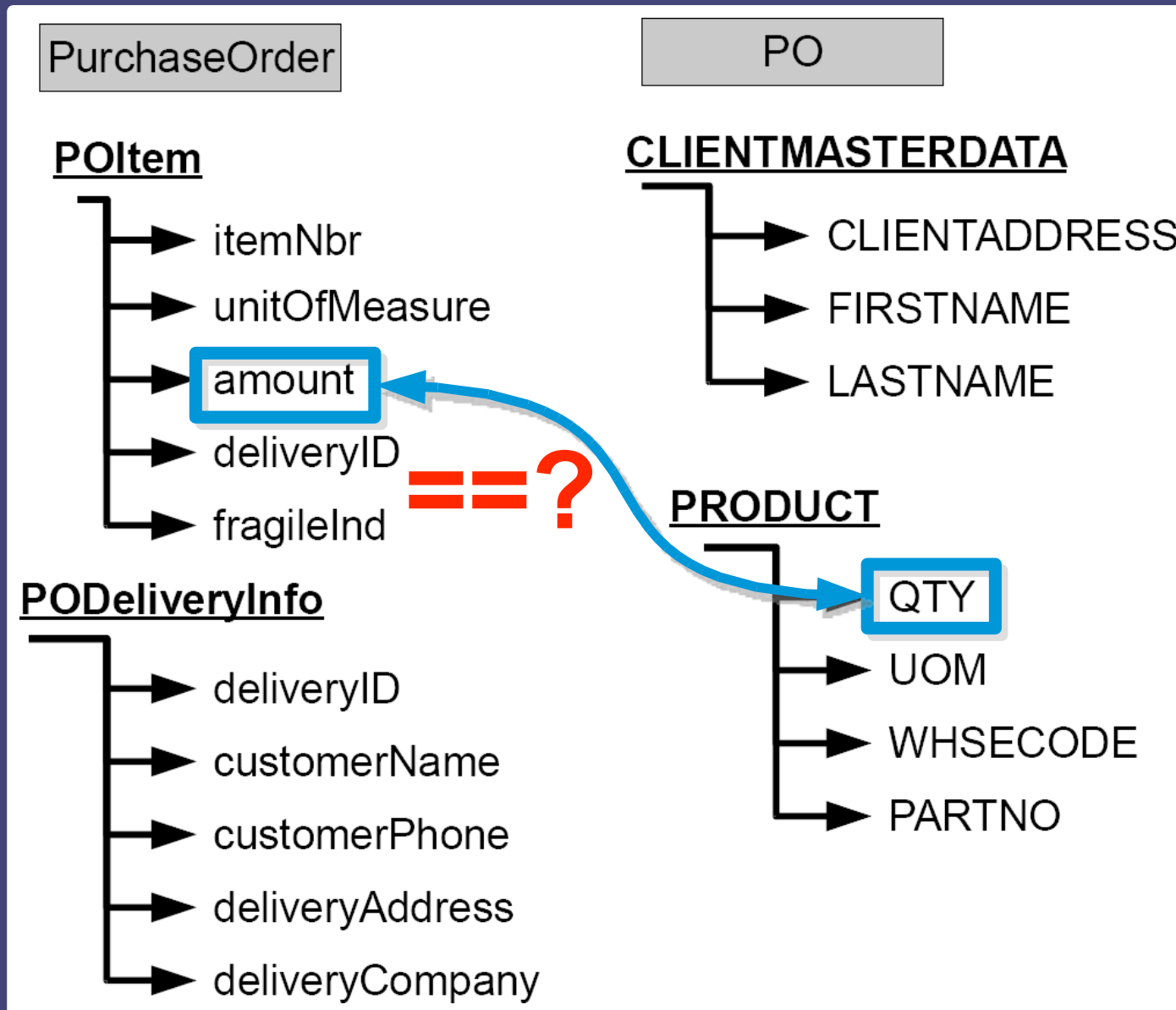
Maciej Gawinecki



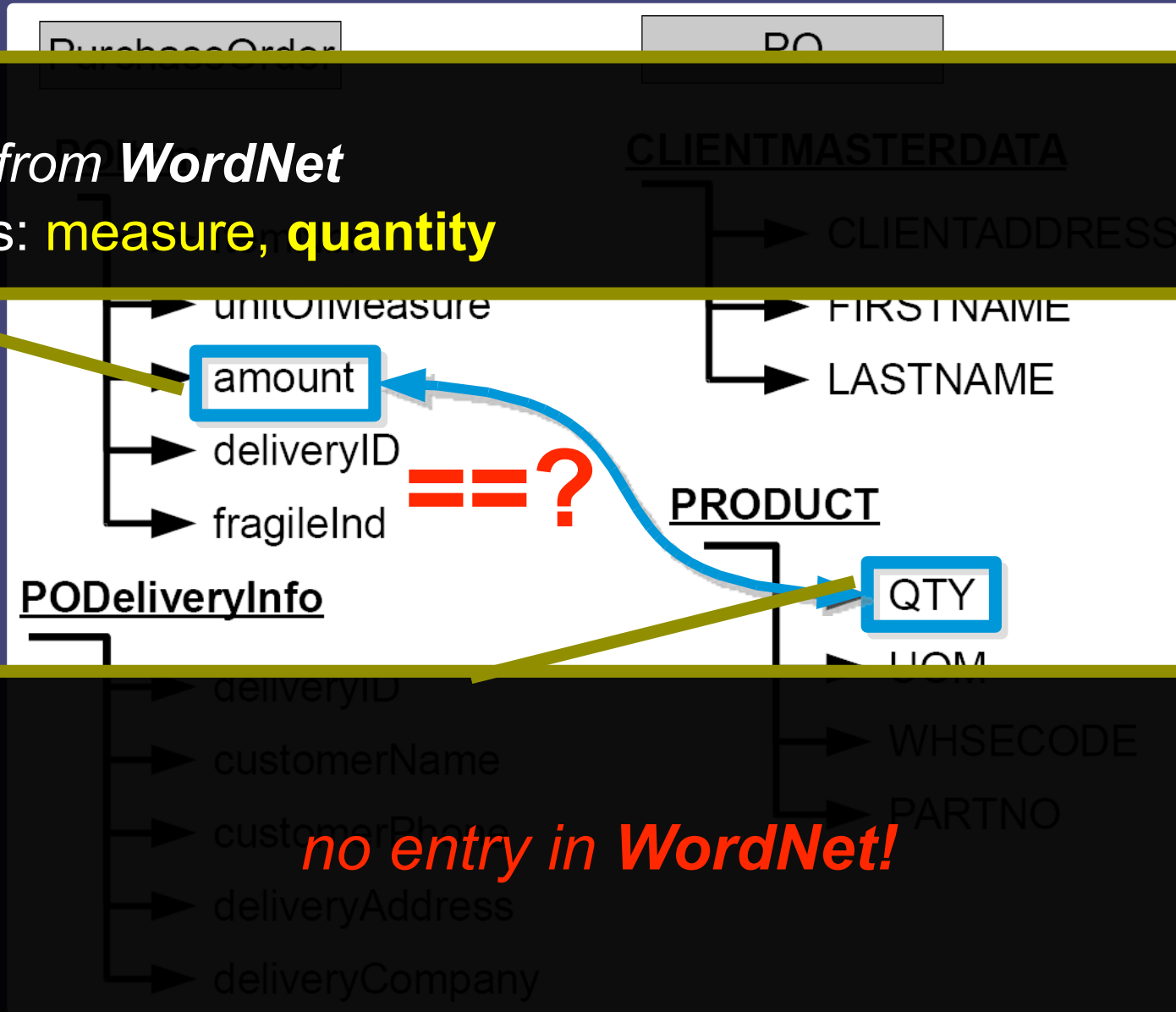
International Doctorate School in
Information and Communication Technologies
Università degli Studi di Modena e Reggio Emilia



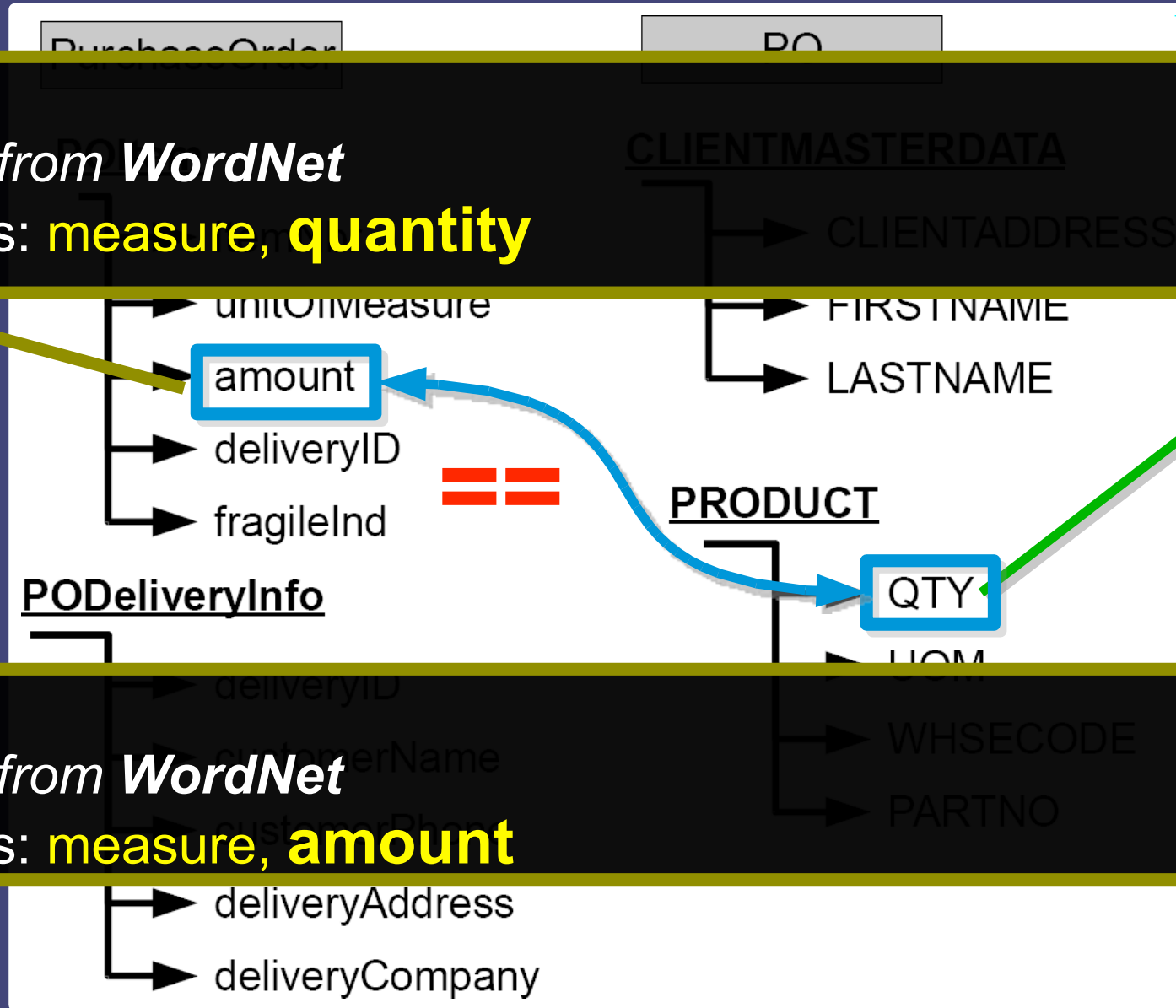
Schemata Integration: Finding the Same Meaning



Schemata Integration: Finding the Same Meaning



Schemata Integration: Finding the Same Meaning



annotation from *WordNet*
synonyms: **measure**, **quantity**

annotation from *WordNet*
synonyms: **measure**, **amount**

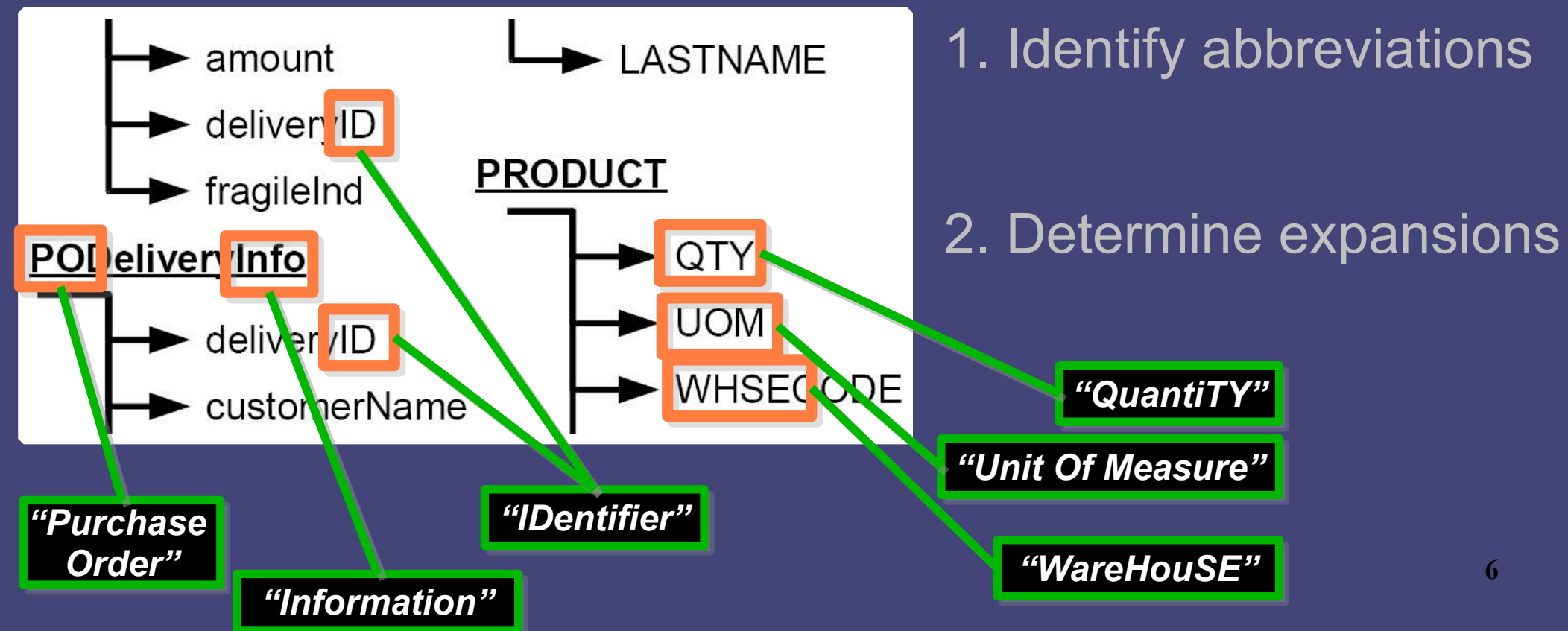
"Quantity"

Why do we care?

- **We need to make abbreviations meaningful**
 - to improve effectiveness of lexical annotation and thus schema mapping discovery
- Most data integration tools ignore the problem
- User-defined dictionary (COMA++, Cupid) is not scalable
 - one abbreviation -- several expansions
 - vocabulary evolves -- dictionary must be updated
 - schema/domain expert still needed
- **We propose effective and scalable solution**

Automatic Abbreviation Expansion

- Given two schemata to integrate, identify character sequences that are abbreviations and determine expansions



Abbreviation Identification

- Determining whether a given word has been used for abbreviation in the given schema label
- Heuristic #1: *non-dictionary words are abbreviations*
 - False negatives: legitimate English words may be used for abbreviations
 - Some words (**standard schema abbreviations**) are always used for abbreviations in schema labels
- Heuristic #2: ***standard schema abbreviations and non-dictionary words are abbreviations!***

Tokenizing Labels

- Reason: **A label can be an abbreviation or it may be multi-word and contain abbreviation(s)**
- Word boundaries
 - punctuation, camel case, e.g. *fragileInd*
 - **no boundaries**, e.g. *WHSECODE*
- Tokenization methods
 - simple
 - **greedy** [Feild 2006]
 - isolating the longest prefixing/suffixing dictionary word/
standard schema abbreviation

Abbreviation Expansion

- The task of finding a relevant expansion for a given identified abbreviation
- There can be more than one expansion candidate for an abbreviation
 - e.g. *PO* can be expanded to:
 - *Purchase Order*
 - *Parents Of*
 - *Post Office*
 - etc.

Types of Abbreviations in Schema

- Standard schema abbreviations
 - describe how a value of an element is *represented*
 - e.g. **Ref** (*Reference*), **Nbr** (*Number*)
- Standard for domain
 - denote *important* and *repeating domain* concepts
 - e.g. **UOM** (*Unit of Measure*)
- Ad hoc abbr. [Ratinov 2004]
 - created to *save space*, from phrases that would not be abbreviated in a normal context
 - e.g. **WHSE** (*Warehouse*), **bk** (*book*)

Where can I find Expansions?

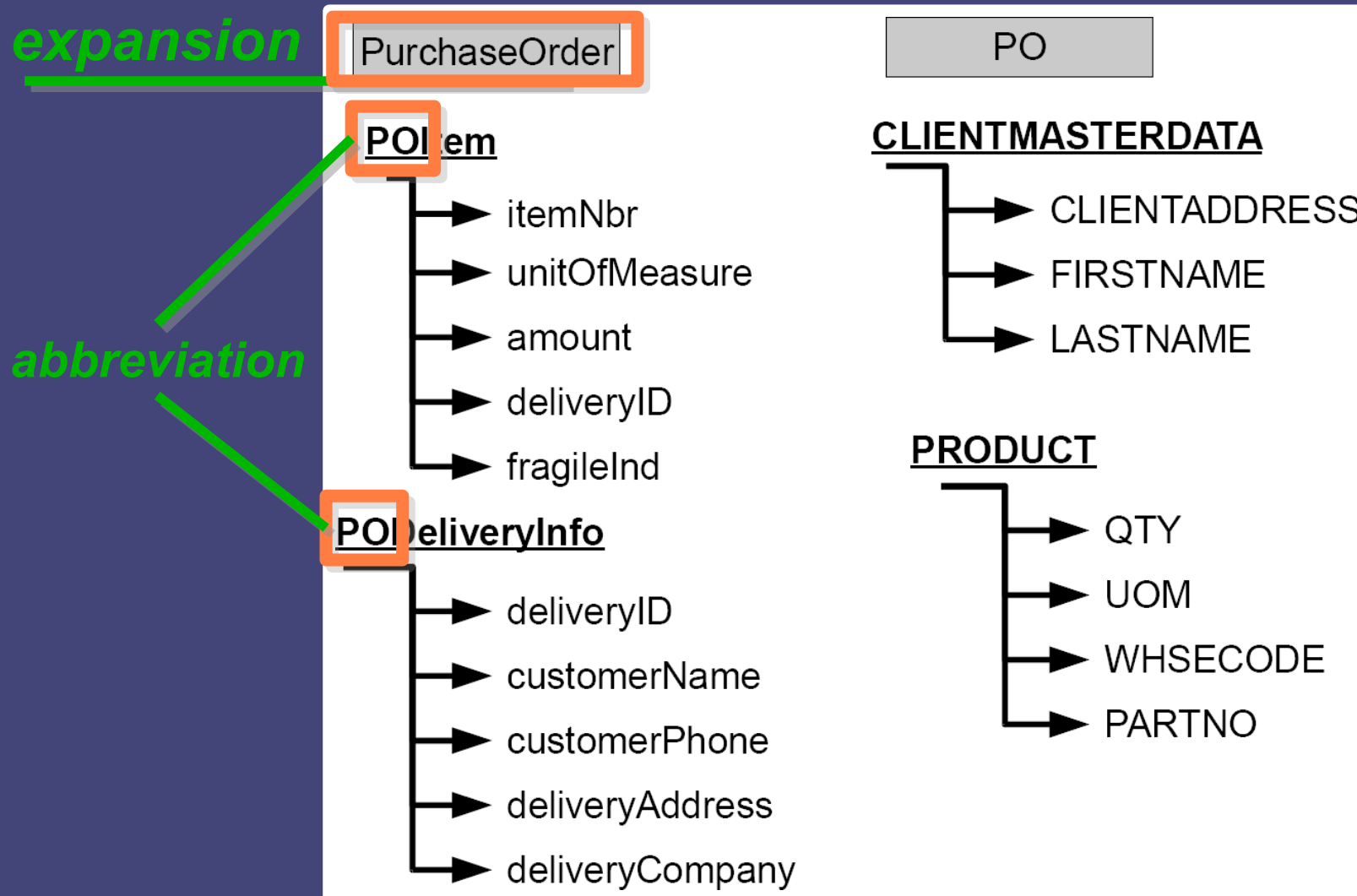
- We did manual expansion of abbrs. in several open-source schemata
- Observations
 - for **standard schema** abbreviations:
 - user-defined dictionary
 - for **standard domain** abbreviations:
 - online abbreviation dictionary
 - for **ad hoc** abbreviations:
 - context of abbreviation
 - complementary schema

} *external sources*

} *internal sources*

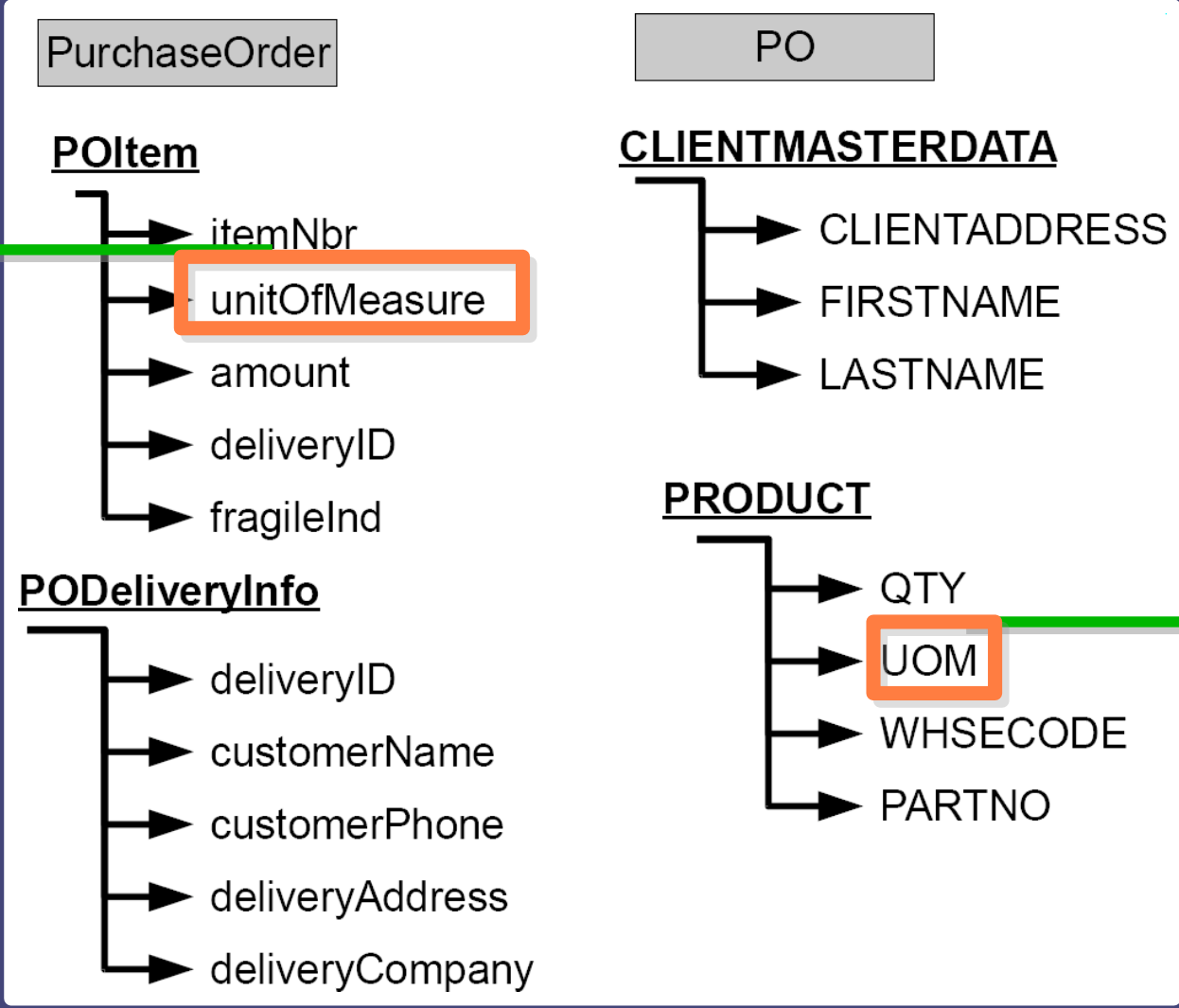
Internal Sources: Context Source

- Label of **containing class** (for *attribute*) or **schema** (for *class*)



Internal Sources: Complementary Schema

expansion



abbreviation

External Sources: Online Abbreviation Dictionary

popularity of
expansion in
given category

expansion

category, where abbr.
and expansion co-occur



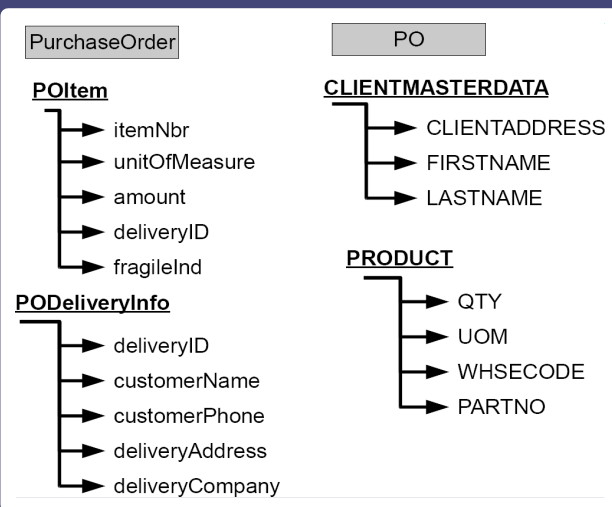
decreasing

| | | |
|--------|------------------|------------------------------|
| ☞ + PO | Post Office | Governmental » US Government |
| ☞ + PO | Post Office | Community » Media |
| ☞ + PO | post office | Miscellaneous » USPS |
| ☞ + PO | Purchase Order | Business » Accounting |
| ☞ + PO | Purchase Order | Governmental » Military |
| ☞ + PO | Purchase Order | Community » Educational |
| ☞ + PO | Portland, Oregon | Governmental » State & Local |
| ☞ + PO | Parents Of | Community » Law |
| ☞ + PO | Police Officer | Community » Law |

Online Abbreviation Dictionary: Selecting Expansion

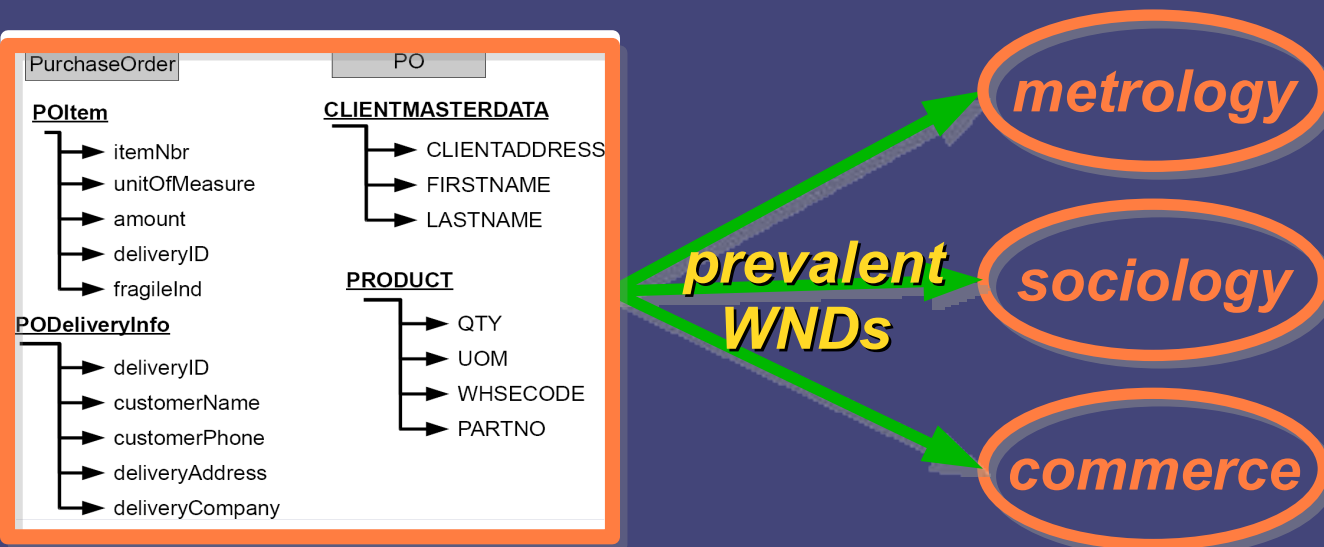
- Expansion is more relevant when
 - it is more *popular*?
 - it *shares* more *domains* of usage with both schemata?
 - it is more more popular in domains of usage shared with both schemata!

Online Abbreviation Dictionary: Scoring Relevance of Expansion



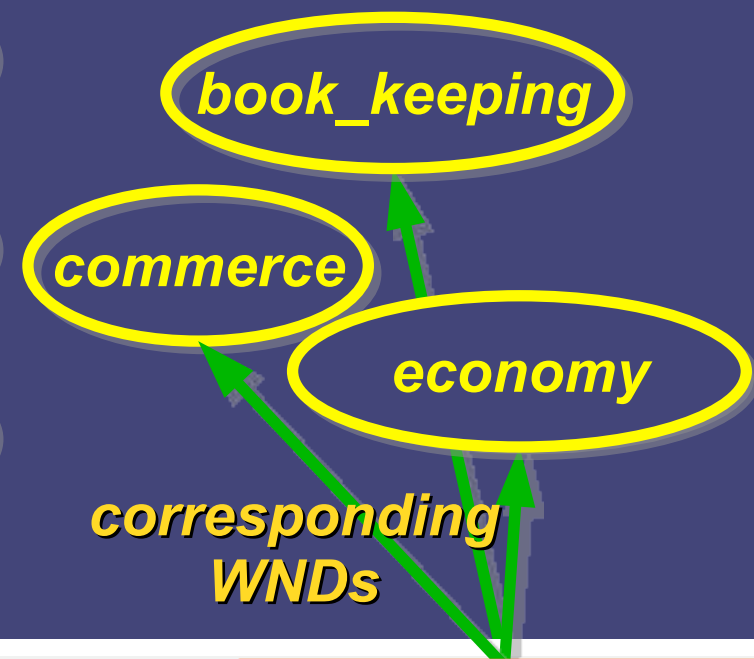
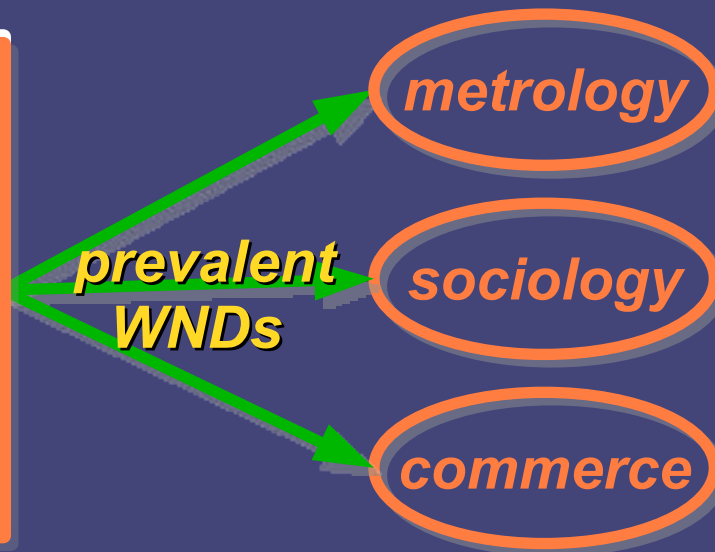
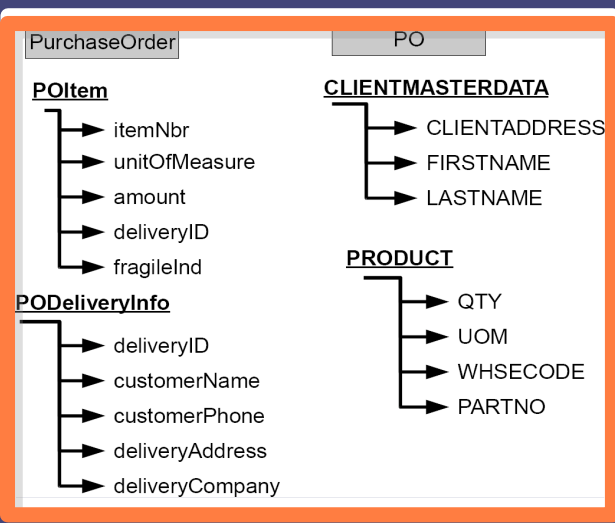
Online Abbreviation Dictionary: Scoring Relevance of Expansion

1. Compute schema prevalent WordNet Domains [Bergamaschi 2008]



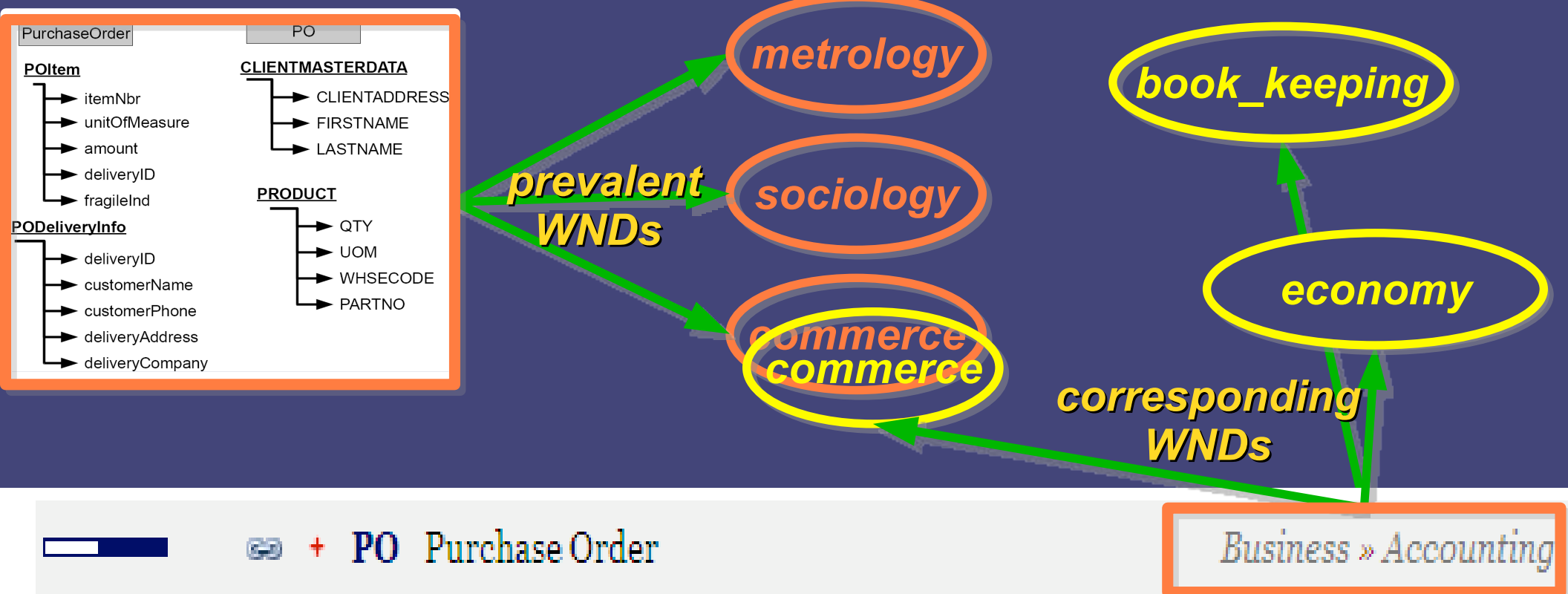
Online Abbreviation Dictionary: Scoring Relevance of Expansion

1. Compute schema prevalent WordNet Domains [Bergamaschi 2008]
2. Get WordNet Domains of expansion



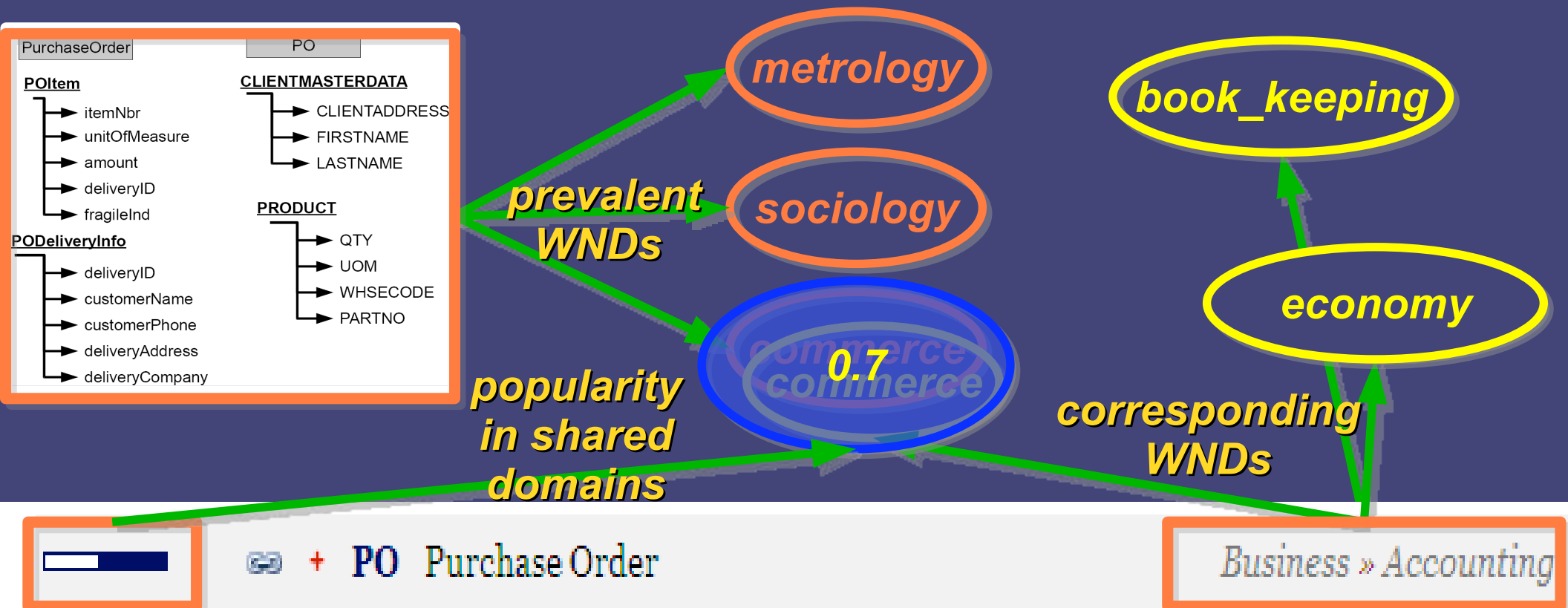
Online Abbreviation Dictionary: Scoring Relevance of Expansion

1. Compute schema prevalent WordNet Domains [Bergamaschi 2008]
2. Get WordNet Domains of expansion
3. Discover shared domains between schemata & expansion



Online Abbreviation Dictionary: Scoring Relevance of Expansion

1. Compute schema prevalent WordNet Domains [Bergamaschi 2008]
2. Get WordNet Domains of expansion
3. Discover shared domains between schemata & expansion
4. Sum up popularity of expansion in shared domains



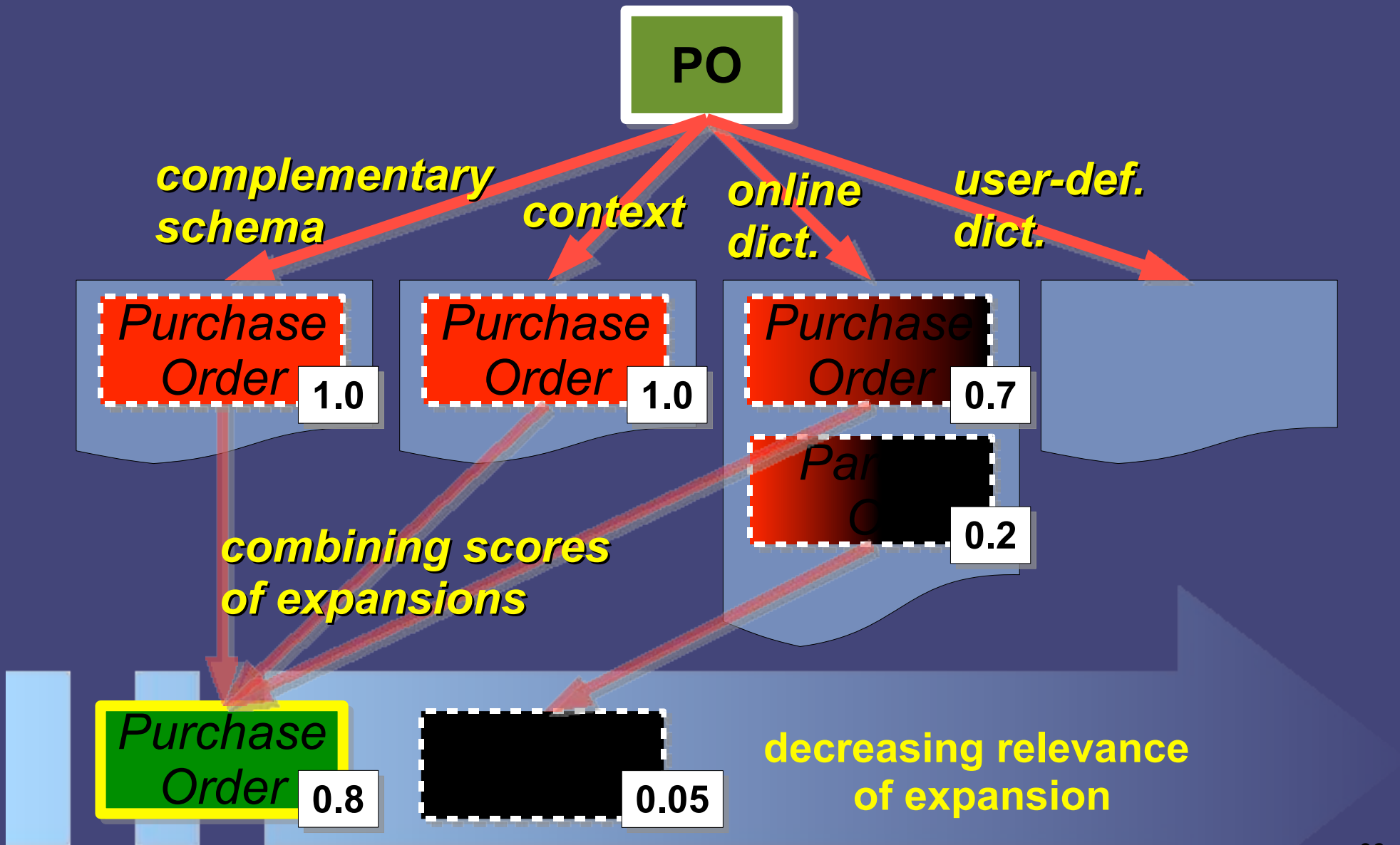
Combining Sources Together

- Sources are complementary in providing expansions
- **No objective criteria for distinguishing ad hoc abbrs. from (domain) standard abbrs!**
- *However*
 - *some types of abbreviations may be more relevant in general*
 - *and thus corresponding sources may be considered as more relevant!*

Relevance of Sources

- Assumption #1
 - **Standard schema abbreviations** should be always expanded to the same expansion
 - **User-defined dictionary** is the most relevant
- Assumption #2
 - **Ad hoc abbreviations** are more frequent than domain-specific abbreviations
 - **Context** and **complementary schema** reflects better user-intention than online dictionary

Example of Expansion



Evaluation Methodology

- Implemented on the top of MOMIS data integration system [Bergamaschi 1999]
- Dataset
 - 2 relational schemata of **Amalgam** integration benchmark
 - www.cs.toronto.edu/~miller/amalgam
 - 168 labels with 52 abbreviations
- Evaluation of identification and expansion methods done separately
 - output of identification gives different input for expansion

Evaluation Criteria of Identification

- Variable
 - **CORRECTNESS**: % of correctly identified labels
- Correctly identified label
 - correctly tokenized
 - all abbreviations identified
- Reference for output
 - **manually** tokenized labels and identified abbrs.

Experiments for Identification

- 3 experiments with different tokenization method used
 - **simple** (**ST**)
 - **greedy + WordNet** (**GT/WN**) dictionary to identify dict. words during tokenization
 - **greedy + Ispell** (**GT/Ispell**) English words list to identify dict. words during tokenization
- All experiments used WordNet for classifying abbreviations!

Results: Identification Correctness

- ST (92%) ~ GT/IsPELL (93%)
 - reason: relatively few labels in dataset without word boundaries, e.g. *bktittle*
- GT/WN much worse (70%)
 - reason: WordNet contains many short abbreviations forcing incorrect tokenization, e.g. *au* (gold) in *authID*
- General problem: legitimate English words!
 - e.g. *Pub* is used for *Publication* but is a dictionary word and it is not a standard schema abbreviation

Evaluation Criteria of Expansion

- Variable
 - **CORRECTNESS**: % of correctly expanded abbrs.
- Input
 - manually tokenized labels and identified abbreviations
- Reference for output
 - **manually** expanded

Experiments for Expansion

- Single sources
- External sources
- Internal sources
- All sources together

Results: Expansion Correctness

- Single source: user-defined dictionary: **42%** correct
 - errors in domain and ad hoc abbreviations
- Single source: online abbreviation dictionary: **19%** correct
 - errors in ad hoc and standard schema abbrs.
- Internal sources: **25%** correct
 - very good in ad hoc abbreviations
- All sources: **83%** ($\sim 42\% + 19\% + 25\%$) correct
 - constituent sources are complementary

Conclusions

- Abbreviations:
 - obstacle for data integration
- Solution
 - complementary sources of expansions for different types of abbreviations
- Results
 - **83% of correct expansion** (42% -- when only user-defined dictionary!) and **better scalability**
- Detailed experimental results and data used
 - <http://www.ibspan.waw.pl/~gawinec/abbr>

Thank you!