Process Querying with Uncertainty

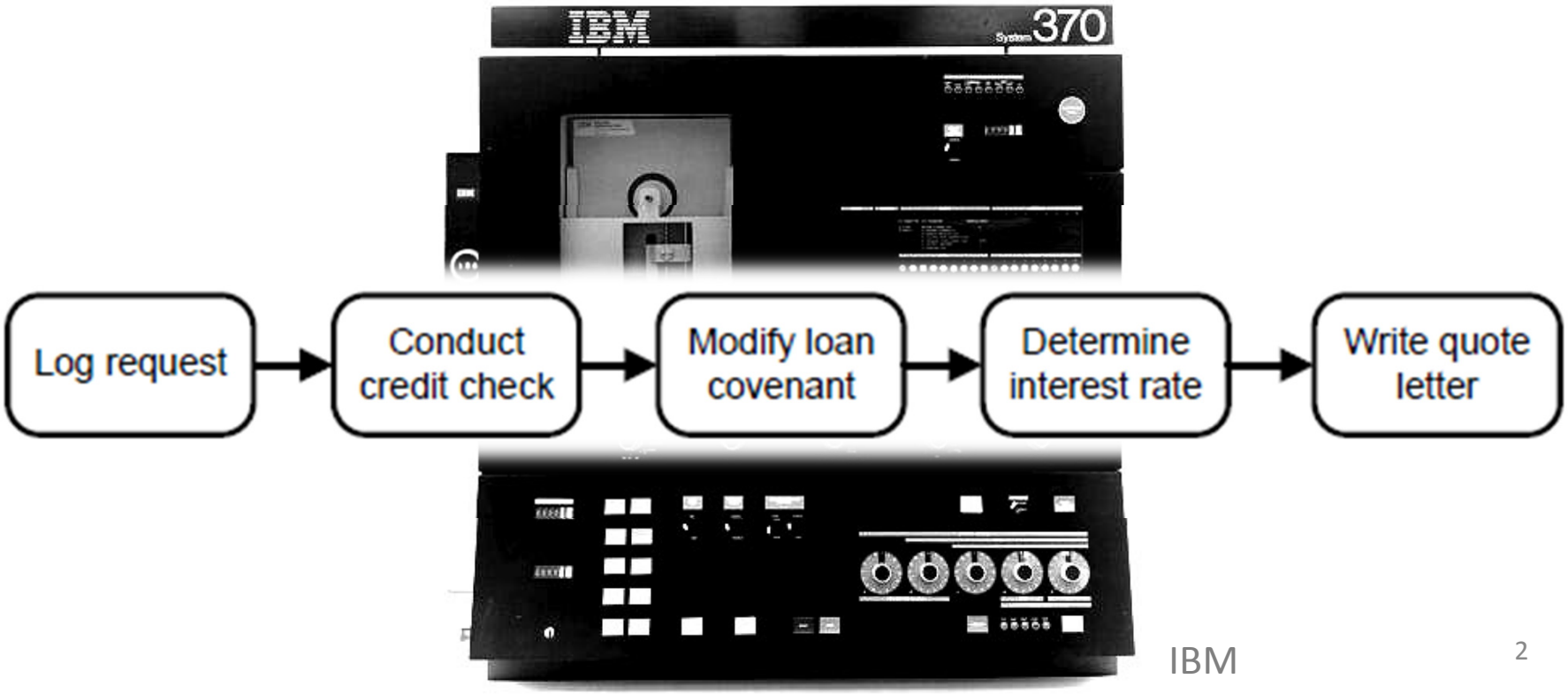
The Need for Process Matching

Avigdor Gal
Technion – Israel Institute of Technology

Based on tutorial notes, jointly authored with Matthias Weidlich, Humboldt Universität zu Berlin

Business Processes

”a set of logically related tasks performed to achieve a defined business outcome for a particular customer or market”
[Davenport 1992]

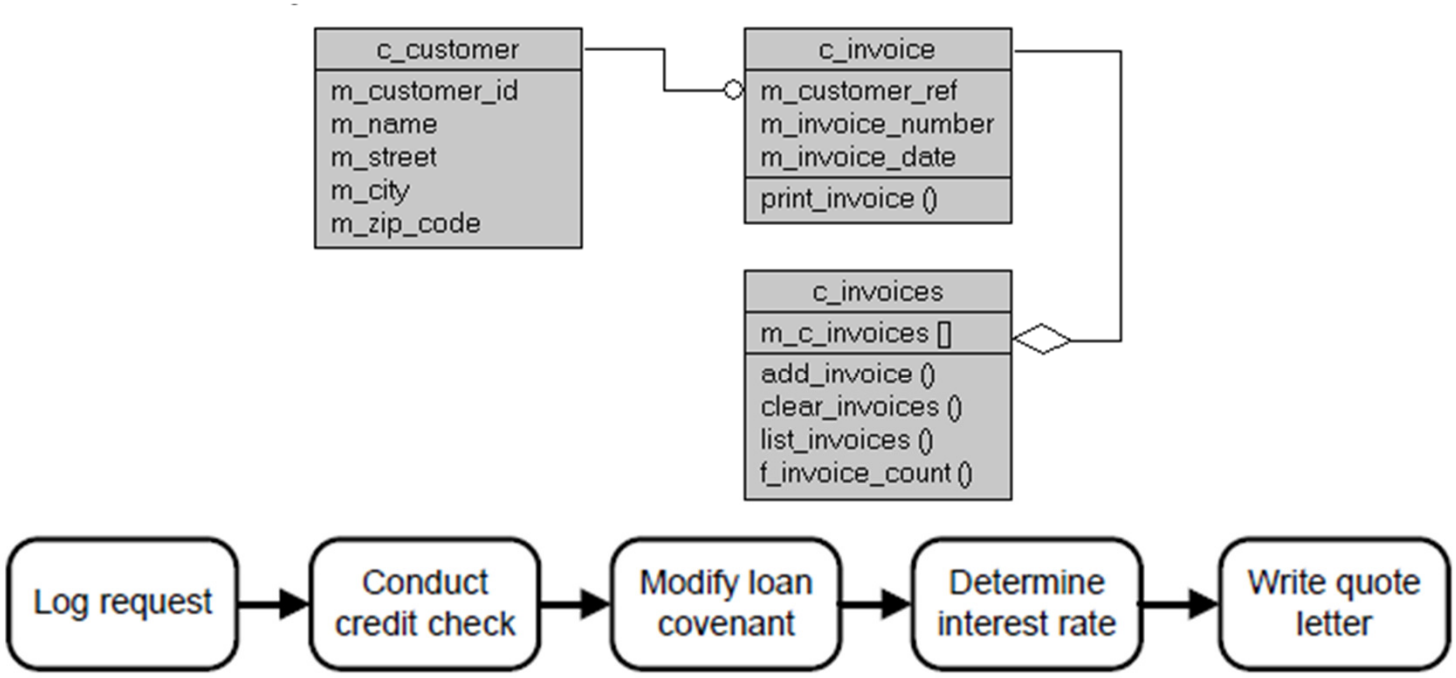


IBM

Conceptual Models

”a conceptual model is invented to provide an appropriate representation of the target system, appropriate in the sense of being accurate, consistent and complete.”

[Norman 1983]



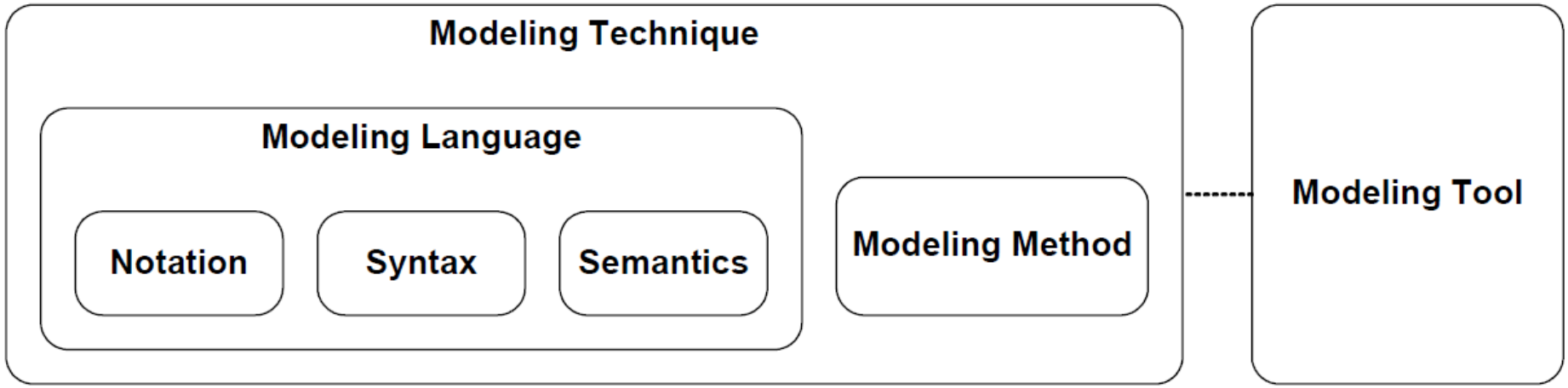
Drivers and Approaches

Technion
Israel Institute of Technology

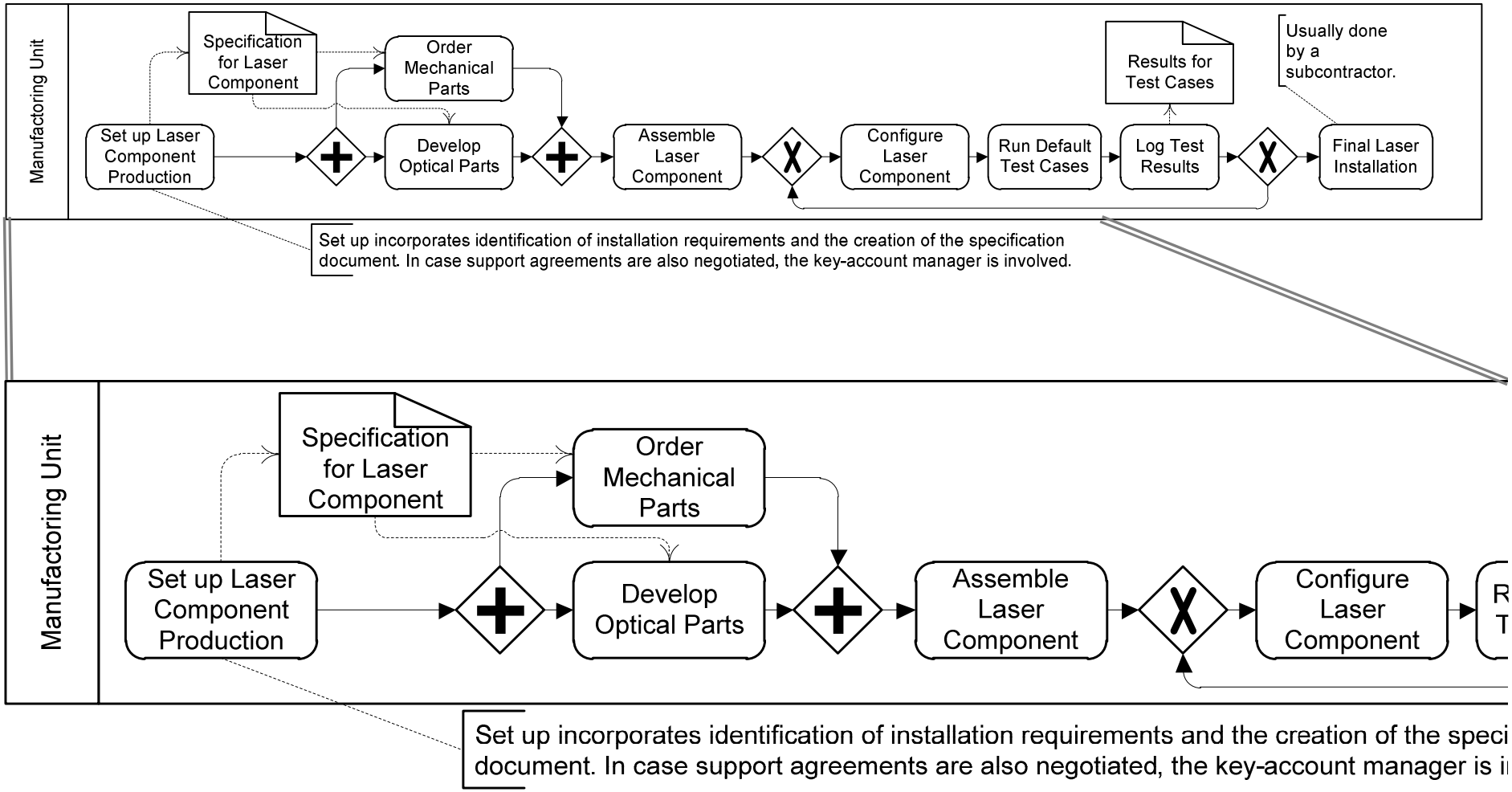
Drivers

- Documentation & standardization
- Workflow automation & system selection
- Staff planning
- Process simulation

Approach



Process Example



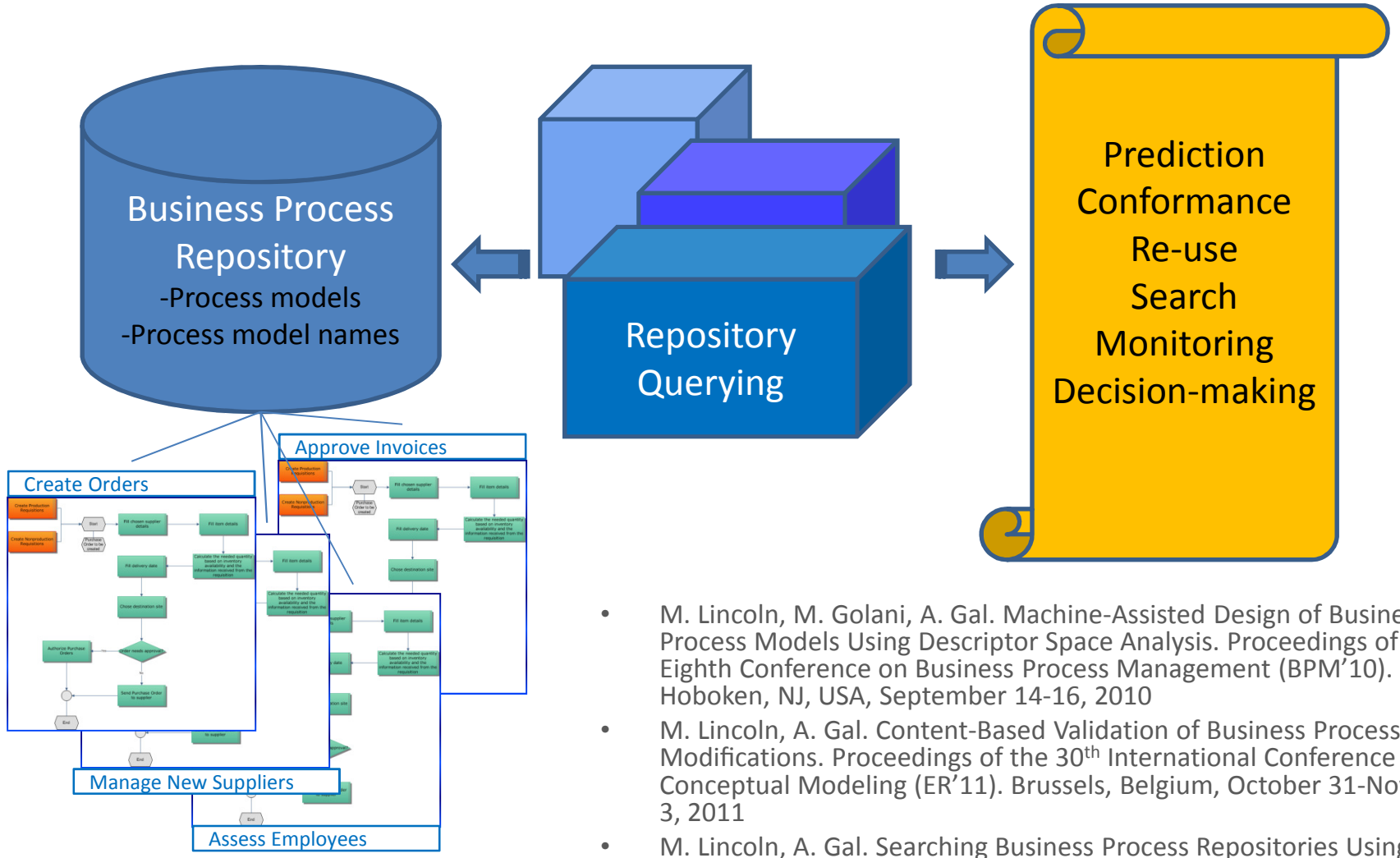
Process Repositories

Technion
Israel Institute of Technology

- Business Process Repositories describe the “know-how” of organizations
- Business Process Repositories can be used for:
  - Management of regulations and compliance enforcement
  - Management and control of IT systems
  - Analysis and improvement of processes
  - Documentation and training
  - Mergers and acquisitions planning
  - Performance monitoring

➡ Process Querying

Process Querying in Repositories



- M. Lincoln, M. Golani, A. Gal. Machine-Assisted Design of Business Process Models Using Descriptor Space Analysis. Proceedings of the Eighth Conference on Business Process Management (BPM'10). Hoboken, NJ, USA, September 14-16, 2010
- M. Lincoln, A. Gal. Content-Based Validation of Business Process Modifications. Proceedings of the 30th International Conference on Conceptual Modeling (ER'11). Brussels, Belgium, October 31-November 3, 2011
- M. Lincoln, A. Gal. Searching Business Process Repositories Using Operational Similarity. Proceedings of the 19th International Conference on Cooperative Information Systems (CoopIS 2011). Crete, Greece, Oct 19-21, 2011

Issues with Querying Process Repositories

Technion
Israel Institute of Technology

- Processes are created over time
  - Semantic drifts
- Processes are created independently
  - Designer biases
- Processes are created for various purposes
  - Varying granularity
- Repositories may be heterogeneous
  - Aligned with organization history

Need for process matching as a generic technique in process querying

Process Matching: Use Cases

Technion
Israel Institute of Technology

Company merger

- Align operations
- Identify commonalities and differences

© www.biojobblog.com

Models, the formalised representations of processes are matched

Take Away

Technion
Israel Institute of Technology

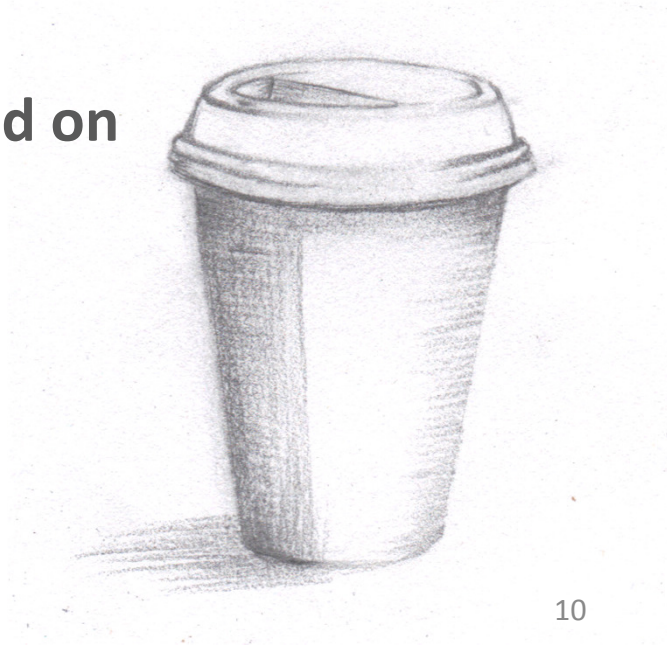
Process matching as a tool for various applications

Process matching is an uncertain process!

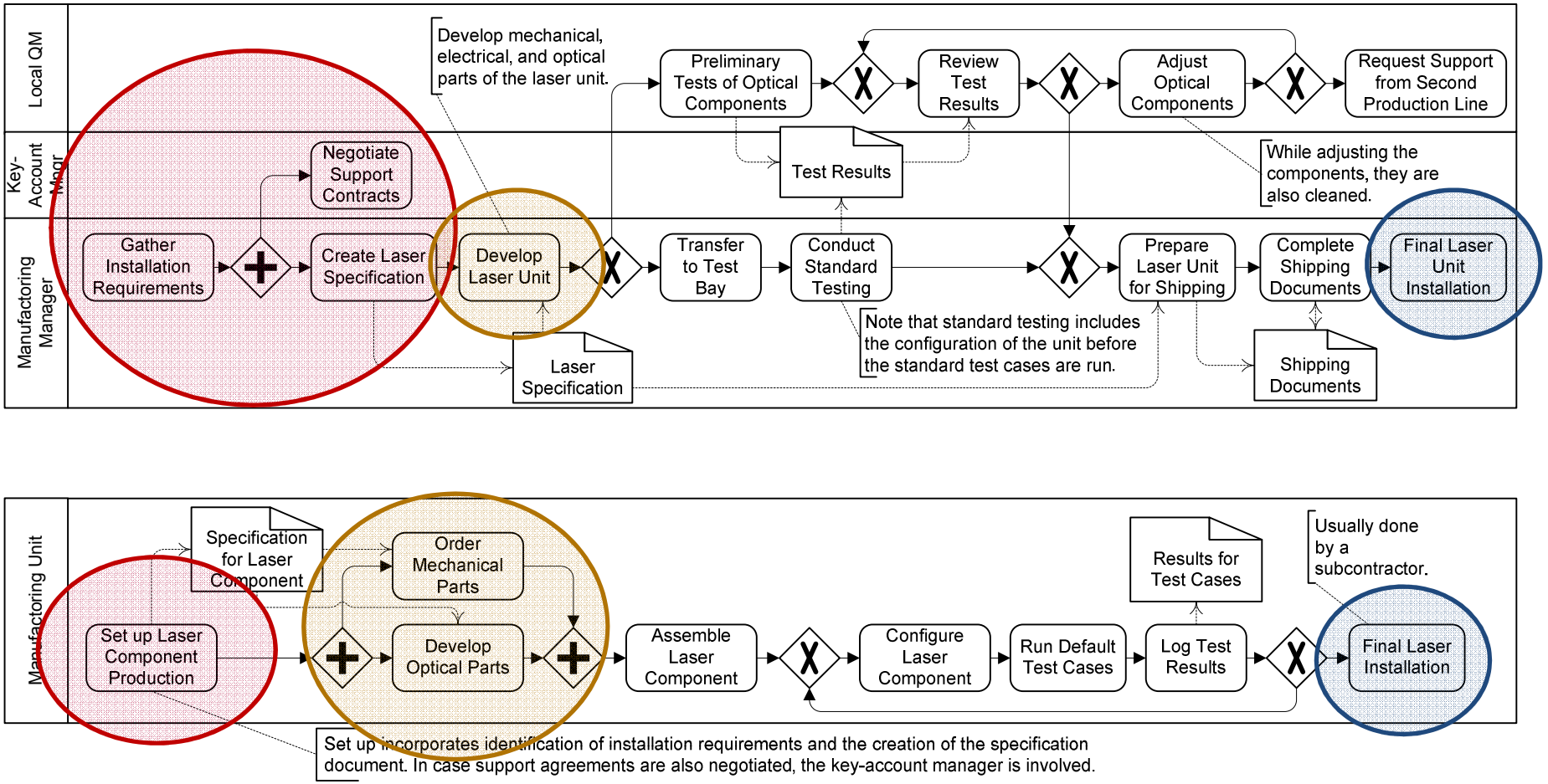
Basic measures to assess the similarity of the model entities

Structural and behavioural matching based on basic similarity measures

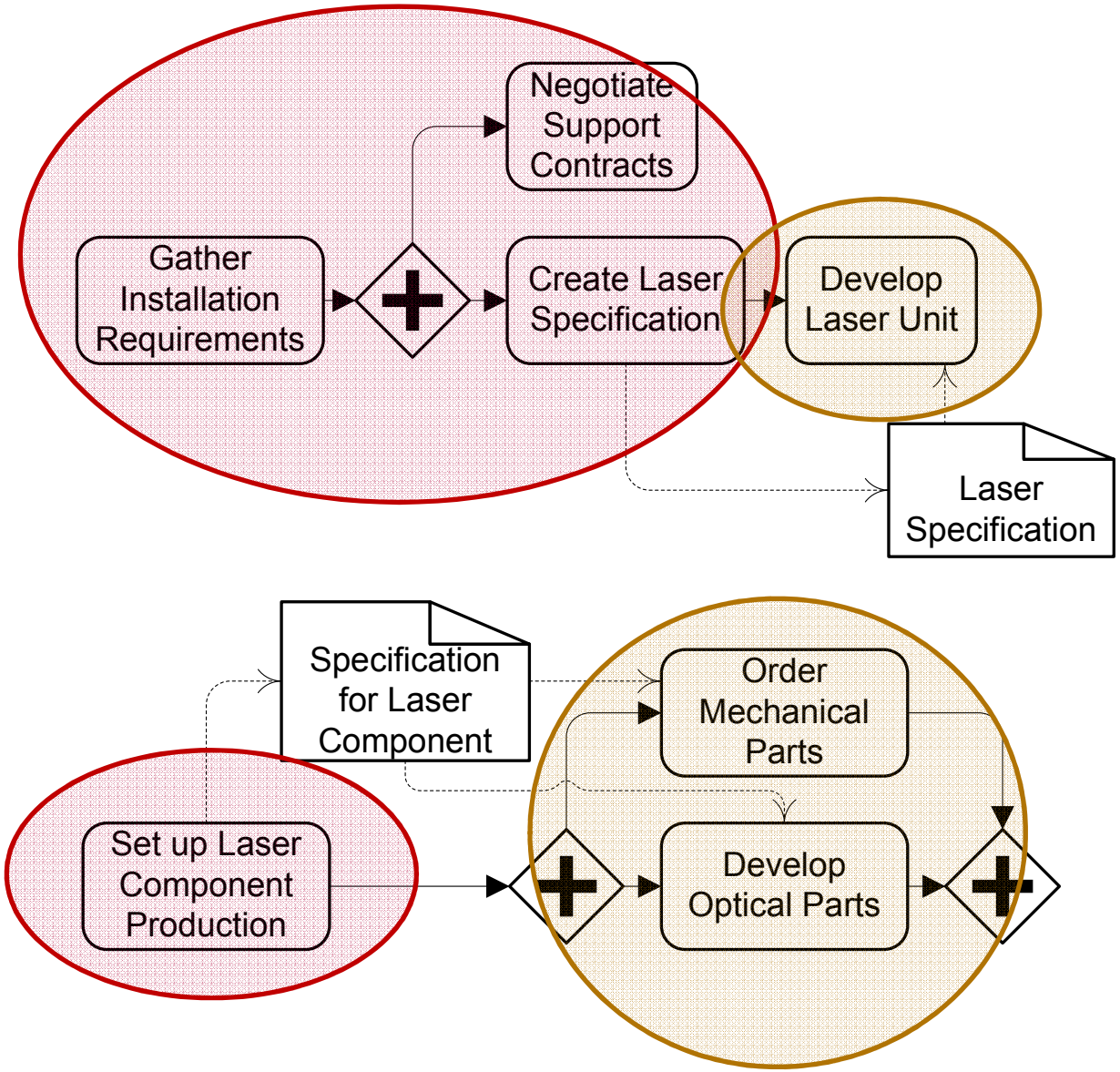
Active research field, e.g., complex correspondences and user input



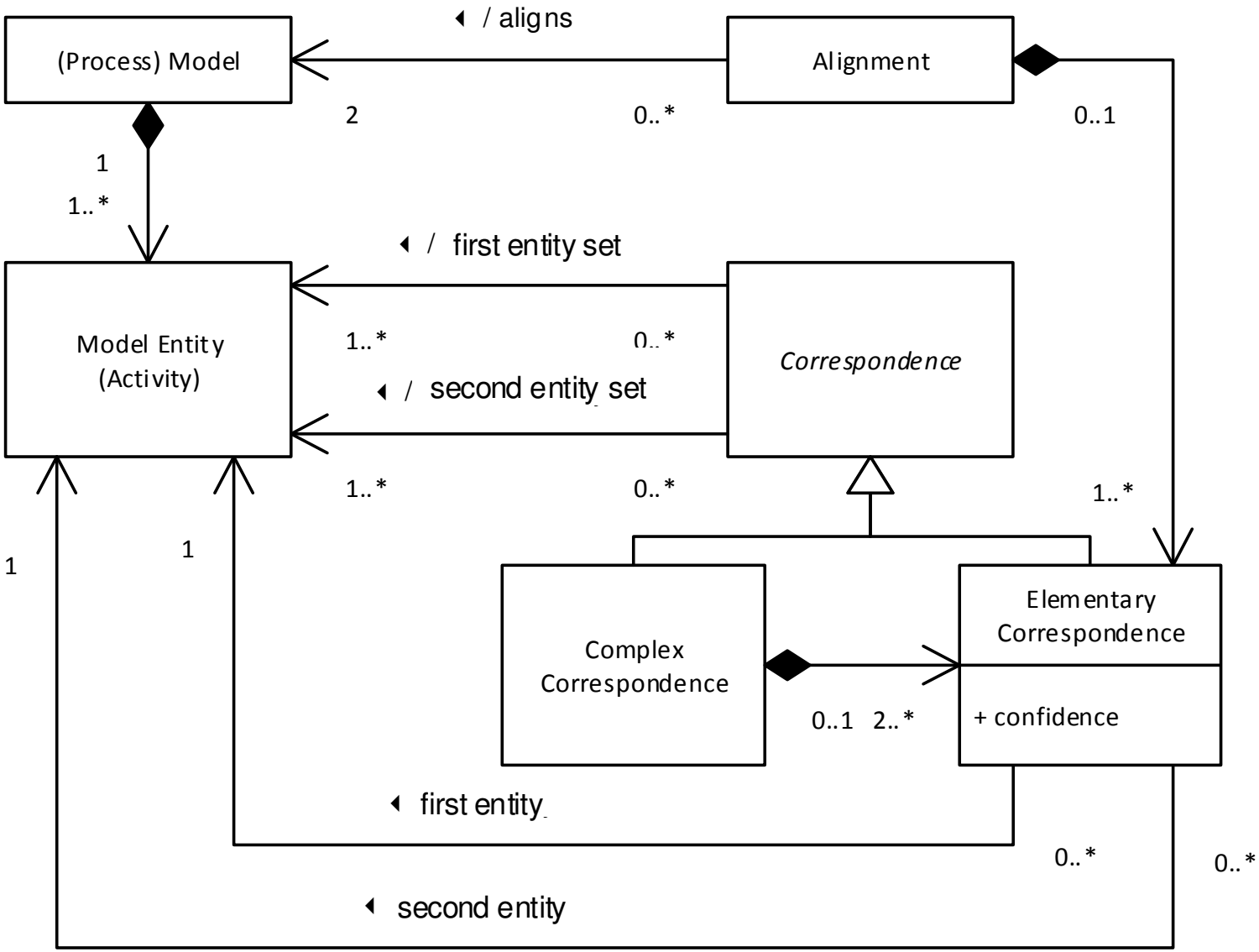
Matching Example



Example in Detail



Terminology



Challenges

On the model level

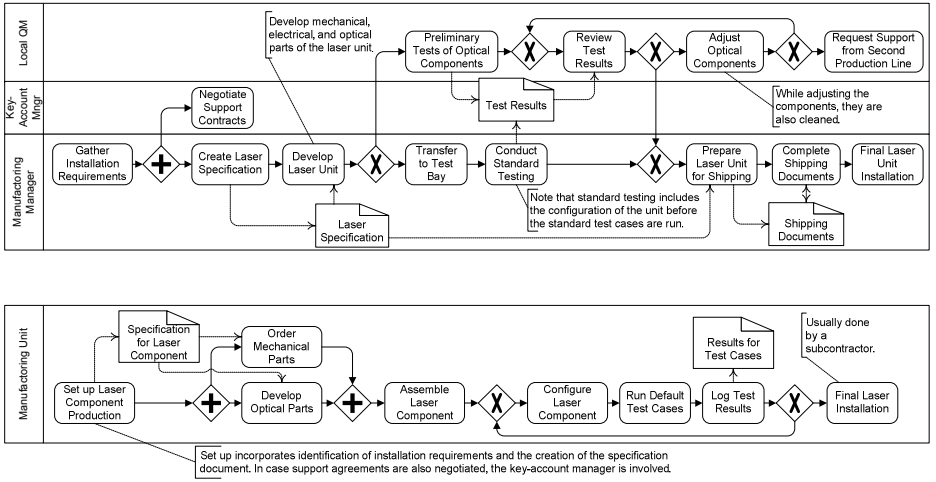
- Representational bias: multitude of modelling languages
- Differences in syntax, semantics and expressiveness
- Differences in applied vocabularies

On the language level

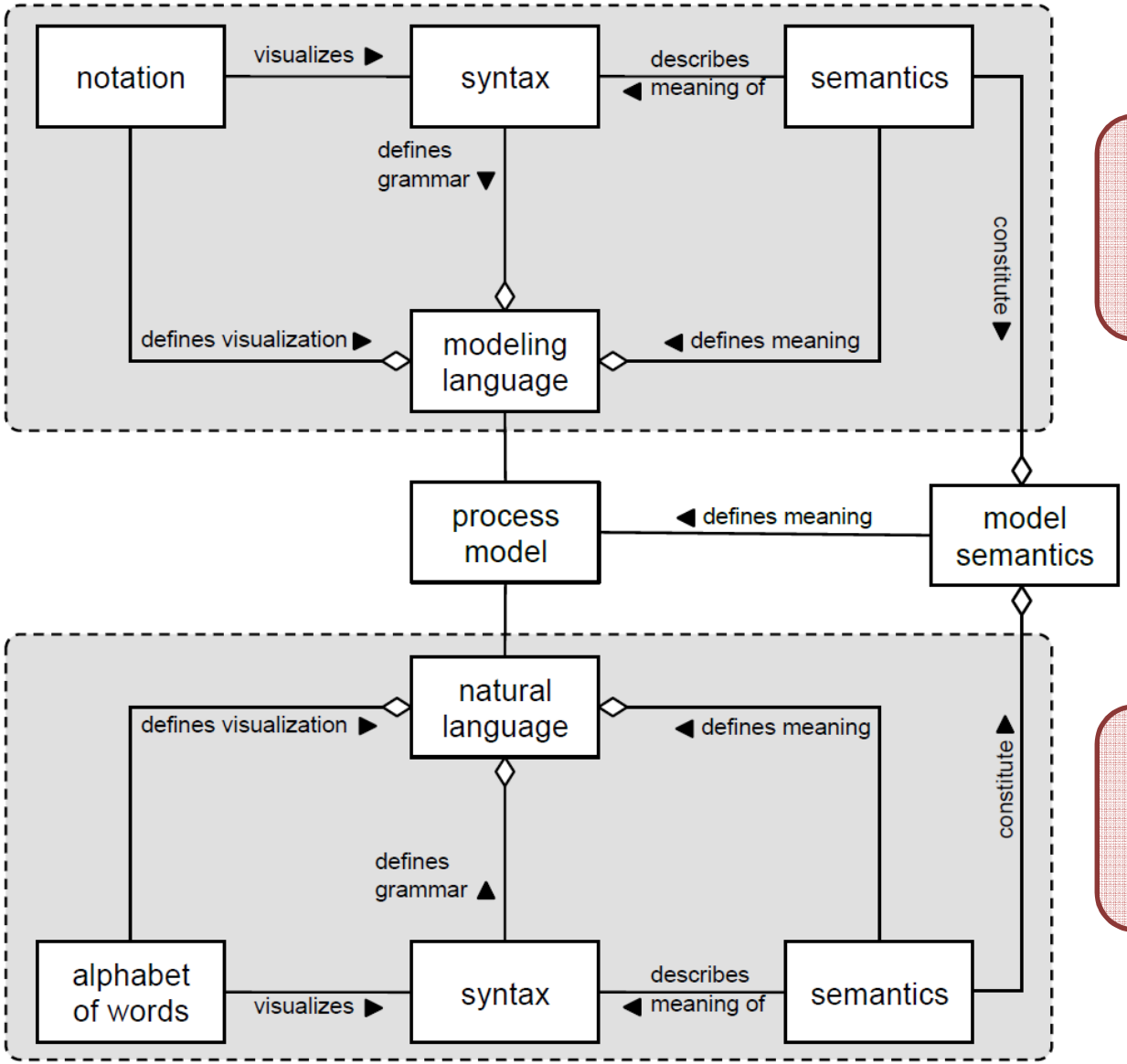
- Linguistic issues
- Short concept labels
- Refinement is hard to track on linguistic level

On the result level

- Result certainty
- Combinatorial issues



Model Perspectives



Structural and Behavioural Matching

Basic Similarity Measures

Agenda

Technion
Israel Institute of Technology

1) **Basic similarity measures**
   - Pre-processing of labels
   - Syntactic and semantic similarity
2) **Structural and behavioural matching**
   - Graph matching
   - Behavioural similarity measures
3) **Practical considerations**
4) **Research directions**

Axiomatic Approach towards Similarity

- Intuition 1: The similarity between A and B is related to their commonality. The more commonality they share, the more similar they are.

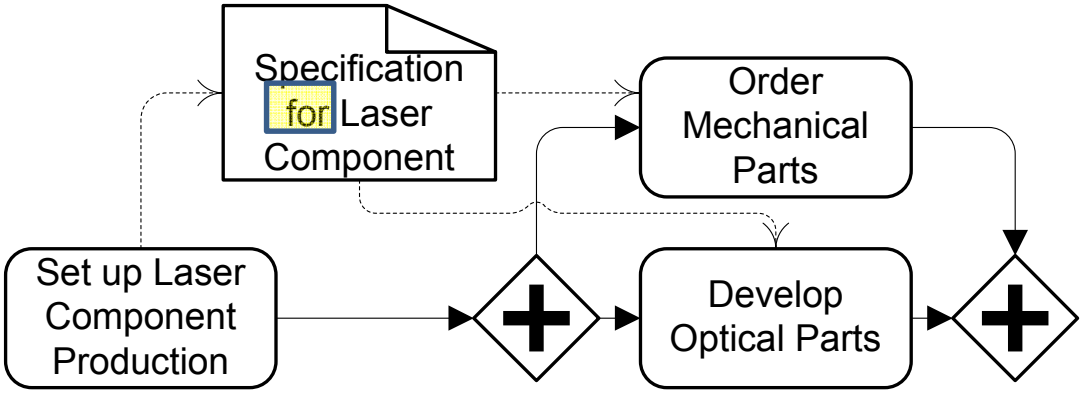
- Intuition 2: The similarity between A and B is related to the differences between them. The more differences they have, the less similar they are.

- Intuition 3: The maximum similarity between A and B is reached when A and B are identical, no matter how much commonality they share.

An Information-Theoretic Definition of Similarity, Dekang Lin, ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning, Pages 296-304

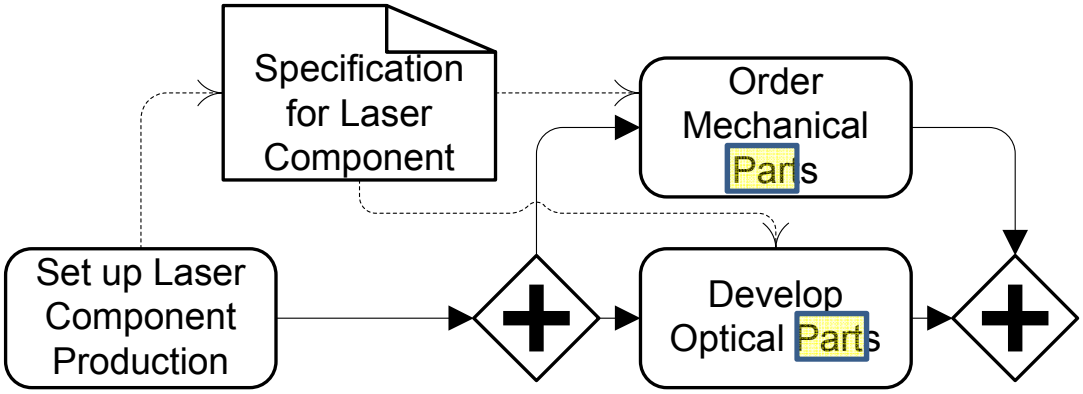
Toying with Words

- **Stop words**
  - Words that are filtered out before the matching process starts.
  - Typically, common words that have little potential of adding to the matching process.
  - Example:



Toying with Words (cont.)

- **Stemming:**
  - reducing words to their word stem, base or root form.
  - Example: Suffix-stripping algorithms



Toying with Words (cont.)

Technion
Israel Institute of Technology

- **Other (straightforward) techniques:**
  - White space elimination
  - Capitalization-based separation
  - De-capitalization
  - Acronym expansion

Between Syntax and Semantics

Technion
Israel Institute of Technology

- Syntactic similarity considers “only” the syntax of labels.
- Semantic similarity requires the notion of a “meaning”.
- In fact, a continuum:
  - Textual similarity assumes similar text = similarity
  - Structure similarity assumes similar structure = similarity
  - …

Syntactic Measures

- Syntactic similarity considers only the syntax of labels.
- String Edit Distance
  - The number of changes (addition, deletion and replacement of characters) necessary to turn one string into another.
  - For example:
    - “Develop Laser Unit” and “Develop Optical Parts”
    - After stemming: “Develop Laser Unit” and “Develop Optic Part”
    - String Edit Distance: 18 (white spaces ignored)
  - Efficiently computing in $O(k^2)$ using dynamic programming and caching.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8):707–710, 1966.

Syntactic Measures (cont.)

- What is a string?
  - Sometimes, a string is just a string
  - Words as tokens. For example:
    - “Develop Laser Unit” and “Develop Optical Parts”
    - After stemming: “Develop Laser Unit” and “Develop Optic Part”
    - Word Edit Distance: 4
  - N-grams as tokens. For example:
    - “Develop Laser Unit” and “Develop Optical Parts”
    - After stemming: “Develop Laser Unit” and “Develop Optic Part”
    - After white space elimination and decapitalization: “developlaserunit” and “developopticpart”
    - N-grams, with N=3:
      - dev, eve, vel, elo, lop, opl, pla, las, ase, ser, eru, run, uni, nit
      - dev, eve, vel, elo, lop, opo, pop, opt, pti, tic, icp, cpa, par, art
    - N-gram Edit Distance: 18

Syntactic Measures (cont.)

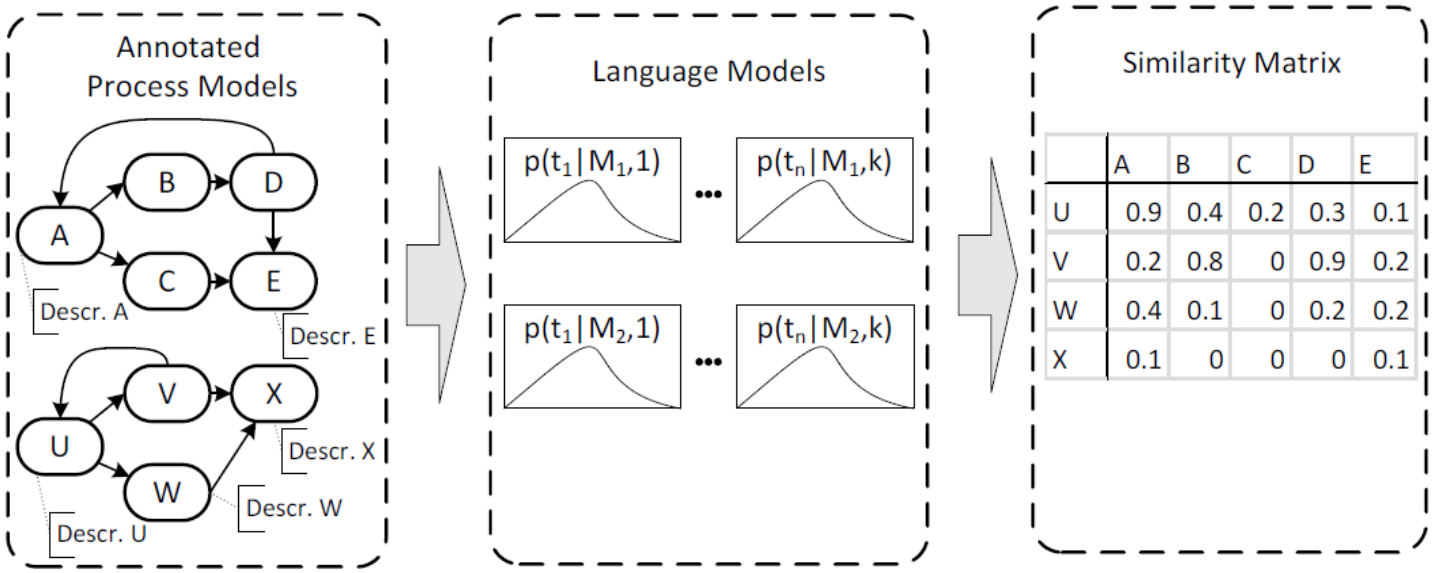
- Bag of terms:
  - Different terms may have different weights.
  - TF-IDF (TF = term frequency, IDF = inverse document frequency) scheme:
    - TF: # of times a term appears in a "document"
    - IDF: 1/(# of "documents" in which a term appears)
  - Notion of a corpus.
- Vector space:
  - Each term represents a dimension
  - Vector similarity using cosine similarity.

    Given two vectors T_1, T_2: $\cos\theta = \frac{T_1 \cdot T_2}{\|T_1\|\|T_2\|}$
  - For example:
    - "Develop Laser Unit" and "Develop Optical Parts"
    - 5D vector with dimensions: Develop, Laser, Unit, Optical, Parts
    - 2 Vectors: (1,1,1,0,0) and (1,0,0,1,1) (assuming equal weights)
    - Cosine similarity: $\frac{1}{3}$

Stephen Robertson and Karen S. Jones. Relevance weighting of search terms Journal of the American Society for Information Science, 27(3):129–146, 1976.
Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
Jonathan J. Webster and Chunyu Kit. Tokenization as the initial phase in NLP. In COLING, pages 1106–1110, 1992.
Gerard Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Commun. ACM, 18(11):613–620, 1975.

Syntactic Measures (cont.)

- Positional, Passage-based Language Models



Matthias Weidlich, Eitam Sheetrit, Moises C. Branco, and Avigdor Gal Matching Business Process Models Using Positional Passage-based Language Models. Proceedings of the 32nd International Conference on Conceptual Modeling (ER'13), Hong Kong, China, November 11-13, 2013.

Semantic Measures

Technion
Israel Institute of Technology

- Using thesaurus to create meaning:
  - Wordnet: Synsets are considered synonyms.
  - For example: “laser” and “optical maser” are in the same synset in Wordnet.
  - Stemming would give a match of “laser” with “optic”.
  - Synonyms may be weighted differently.

George A. Miller. WordNet: A lexical database for english. Commun. ACM, 38(11):39–41, 1995.
Marc Ehrig, Agnes Koschmider, and Andreas Oberweis. Measuring similarity between semantic business process models. In APCCM, pages 71–80, 2007.

Semantic Measures (cont.)

- Parts-of-Speech tagging:
  - Two-phase approach for refactoring of labels following action-noun style into verb-object labels
    - Style recognition: determine the label style.
    - Derivation phase: tag action, business object, and optional fragments.
  - Example:
    - “Develop Laser Unit”
      - “Develop" (verb, action) and
      - “Laser Unit“ (noun, business object)
    - “Develop Optical Parts”
      - “Develop" (verb, action) and
      - “Optical Parts“ (noun, business object)

Remco M. Dijkman, Marlon Dumas, Luciano Garcıa-Banuelos, and Reina Kaarik. Aligning business process models. In EDOC, pages 45–53, 2009.
Remco M. Dijkman, Marlon Dumas, Boudewijn F. van Dongen, Reina Kaarik, and Jan Mendling. Similarity of business process models: Metrics and evaluation. Inf. Syst., 36(2):498–516, 2011.
Henrik Leopold: **Natural Language in Business Process Models: Theoretical Foundations, Techniques, and Applications** . Lecture Notes in Business Information Processing, Vol. 168, Springer-Verlag, 2013.

Agenda

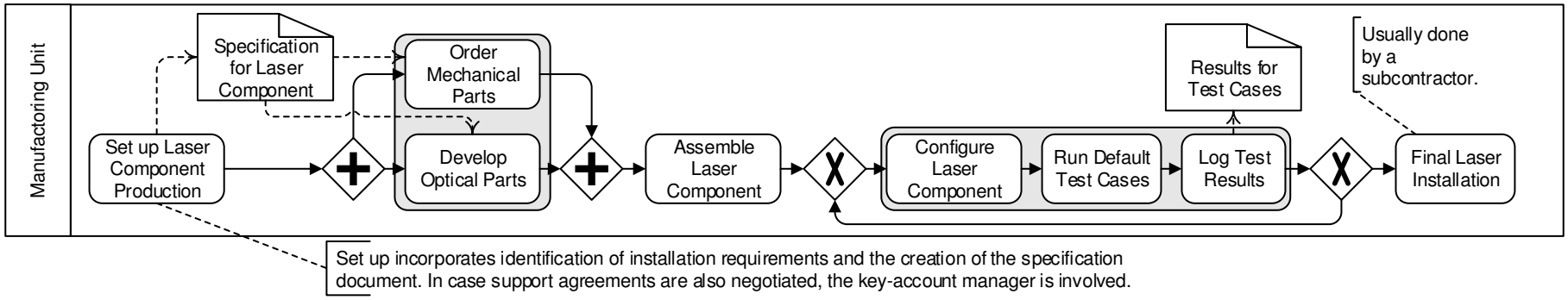
Technion
Israel Institute of Technology

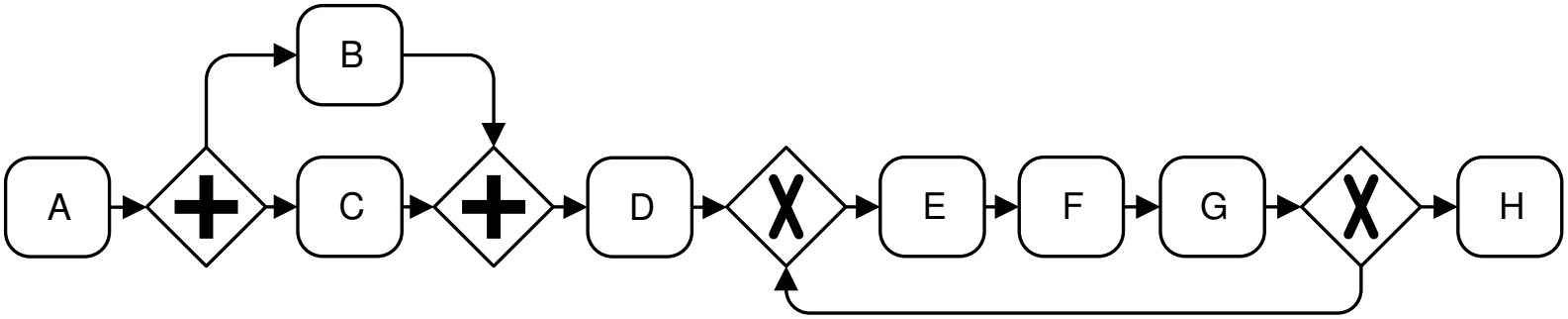
1) **Background on (process) model matching**
2) **Basic similarity measures**
   - Pre-processing of labels
   - Syntactic and semantic similarity
3) **Structural and behavioural matching**
   - Graph matching
   - Behavioural similarity measures
4) **Practical considerations**
5) **Research directions**

Graph Matching

Technion
Israel Institute of Technology

Process models are essentially activity graphs



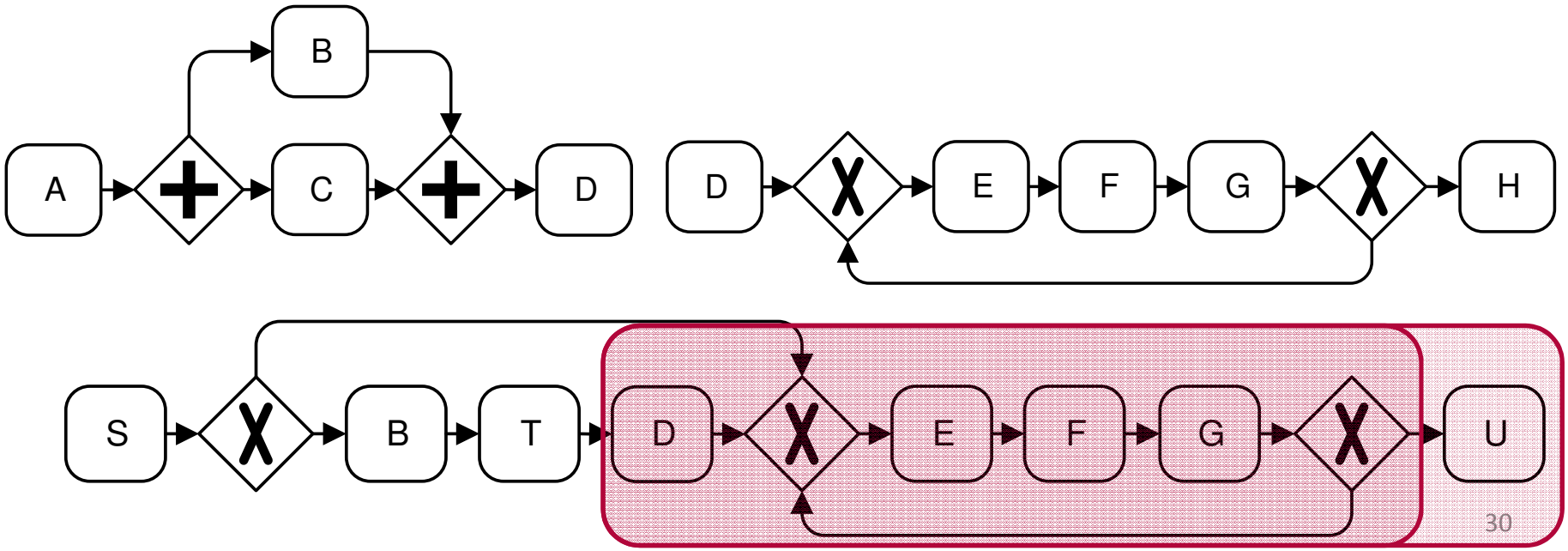


Subgraph Isomorphism

Given an activity graph, find isomorphic subgraphs in the other model

Isomorphism ensures equivalent control flow structures

Notion of isomorphism can be extended with node type equivalence and basic similarity measures



GED-based Similarity

Technion
Israel Institute of Technology

Graph edit distance

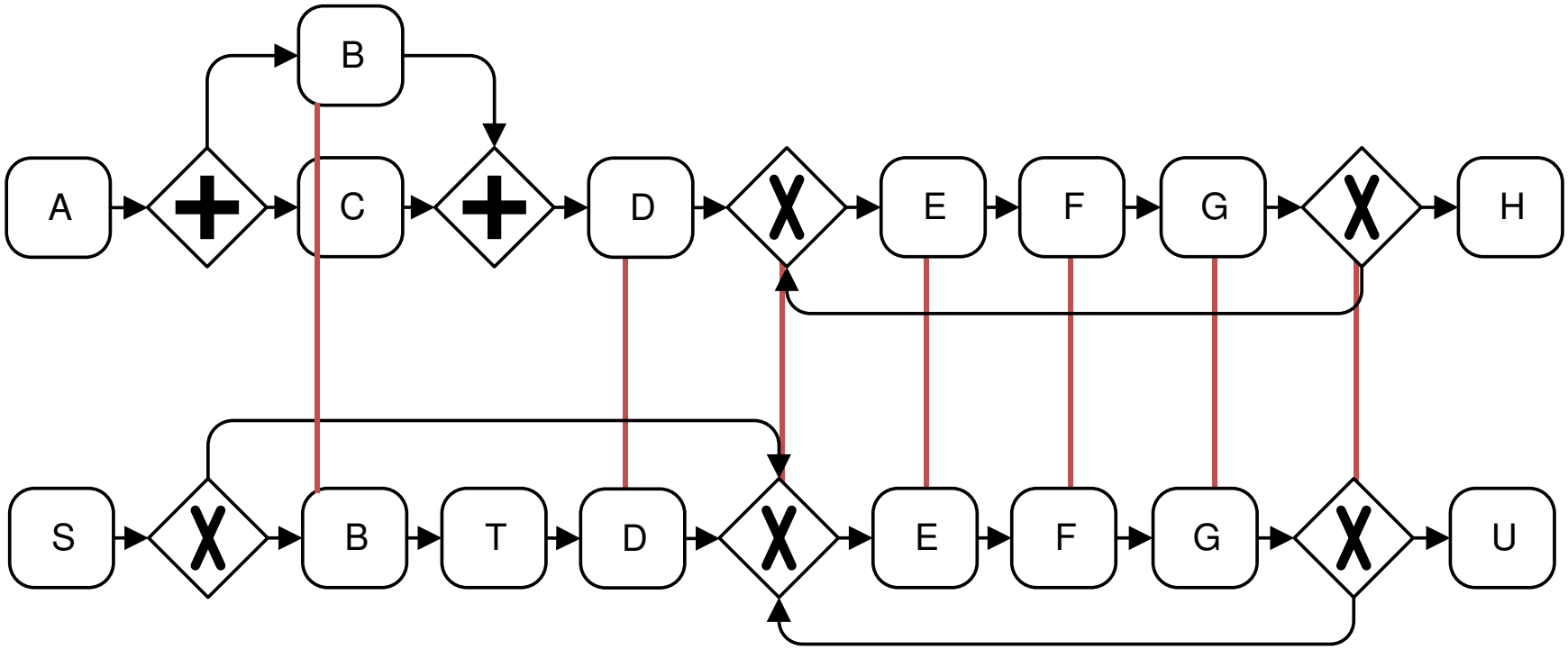
- Inspired by string edit distance
- Minimal number of graph operations needed to transform one graph into another
- Different sets of graph operations (insert/delete/substitute node/edge)
- Different criteria to judge substitution quality

Remco M. Dijkman, Marlon Dumas, Luciano Garcıa-Banuelos, and Reina Kaarik. Aligning business process models. In EDOC, pages 45–53, 2009.
Remco M. Dijkman, Marlon Dumas, Boudewijn F. van Dongen, Reina Kaarik, and Jan Mendling. Similarity of business process models: Metrics and evaluation. Inf. Syst., 36(2):498–516, 2011.

Graph Edit Distance (GED) Sim

Technion
Israel Institute of Technology

Consider the case that correspondences are given

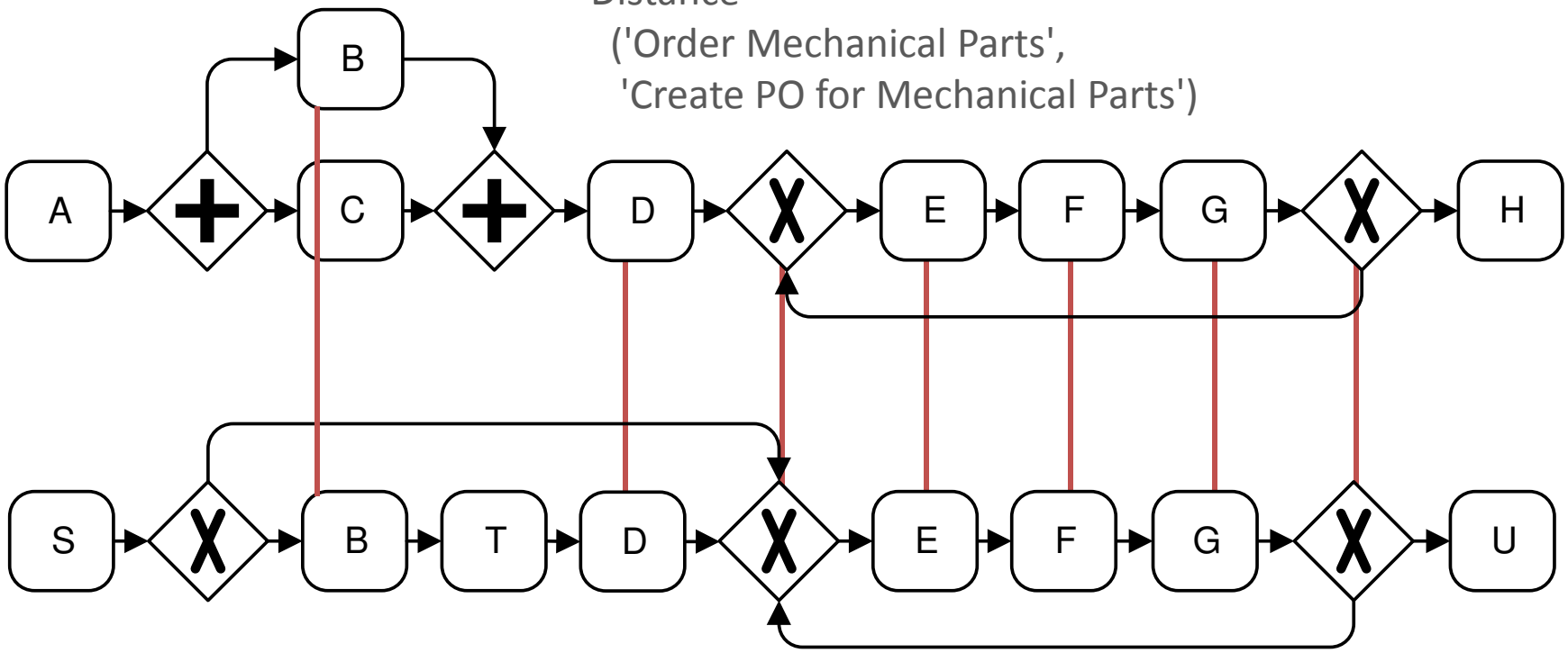


GED Sim cont.

Score substituted nodes

$$f_{subs} = \frac{\text{score}}{\#\text{substitutions}}$$

Distance
('Order Mechanical Parts',
'Create PO for Mechanical Parts')

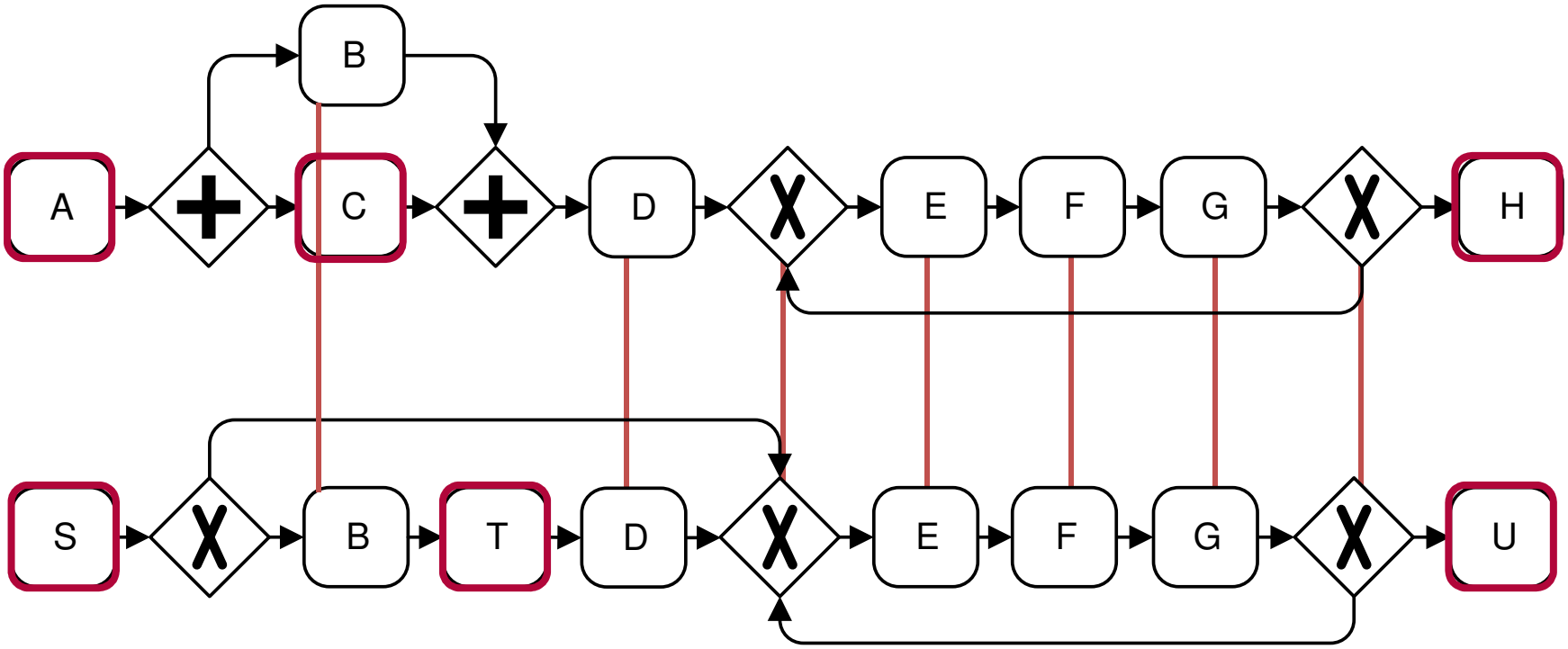


GED Sim cont.

Technion
Israel Institute of Technology

Score skipped nodes

$$f_{skipn} = \frac{score}{\#nodes}$$

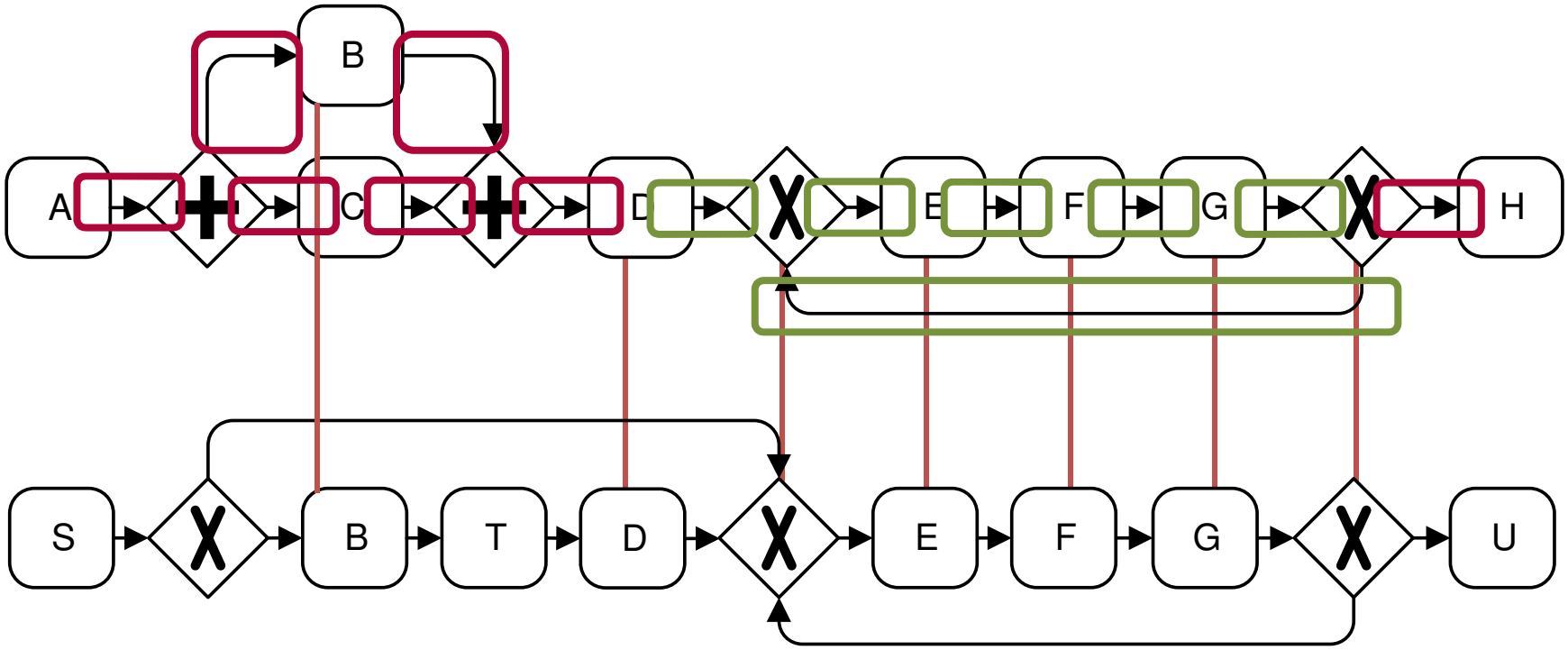


GED Sim cont.

Technion
Israel Institute of Technology

Score skipped edges

$$f_{skipe} = \frac{score}{\#edges}$$



GED Sim cont.

Technion
Israel Institute of Technology

Given correspondences:

1. Score substituted nodes (f_{subs})
2. Score skipped nodes (f_{skipn})
3. Score skipped edges (f_{skipe})

Similarity = 1.0 - weighted average of f_{subs}, f_{skipn}, f_{skipe}

Use for matching:

find correspondences that maximise similarity

Behavioural Similarity

Technion
Israel Institute of Technology

Process models first and foremost describe behaviour

Also, the same behaviour may be expressed with different structures

Hence, quantify behavioural similarity

- Different underlying models qualify to be the basis
- Different operationalisations of measures

Trace Similarity

Technion
Israel Institute of Technology

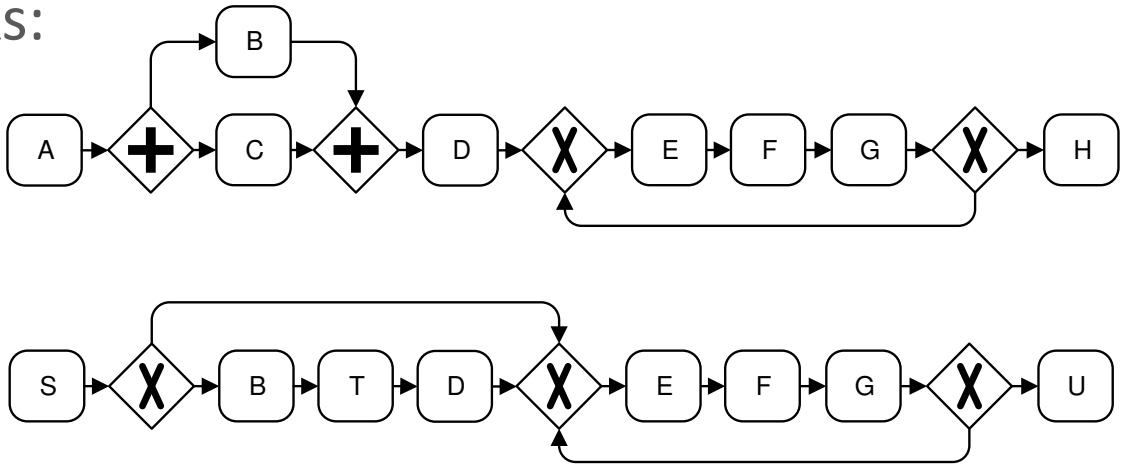
Assume that “behaviour” is interpreted as a set of traces

A straightforward measure:
Jaccard coefficient of two sets, the intersection relative to the union

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

But, obvious drawbacks:

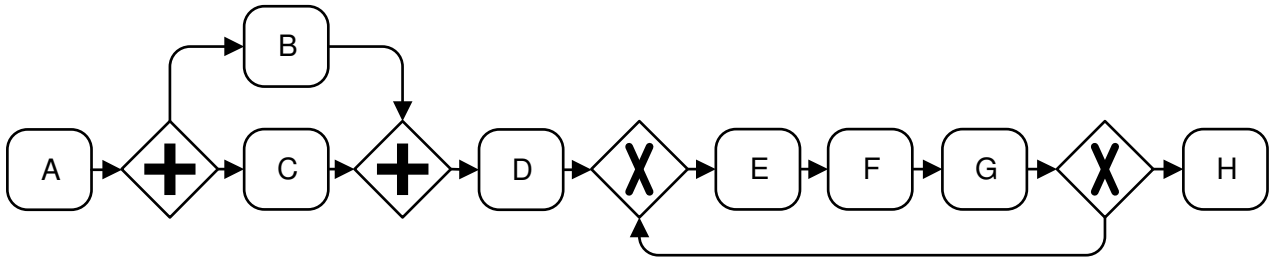
- Infinite sets of traces
- Minor deviation, but empty intersection



Trace Similarity cont.

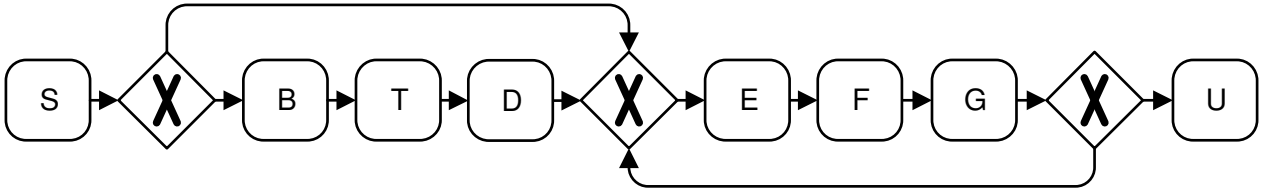
N-Grams to realise trace-based measures

- Compare sets of n-grams of traces
- Length of n-gram defines granularity of the measure



2-grams:

AC, CD, DE, EF, GE, GH, …



SB, SE, DE, EF, GE, GU, …

Andreas Wombacher, Maarten Rozie: Evaluation of Workflow Similarity Measures in Service Discovery. Service Oriented Electronic Commerce 2006:51-71

Sim with Behavioural Relations

Technion
Israel Institute of Technology

Idea: Similarity measure compares behavioural constraints of two process models

Different approaches may be followed to capture these constraints using trace semantics

- Direct-successorship as introduced in the context of the α-algorithm, also called *footprint*
- Indirect-successorship, also called *behavioural profile*

Footprint

Technion
Israel Institute of Technology

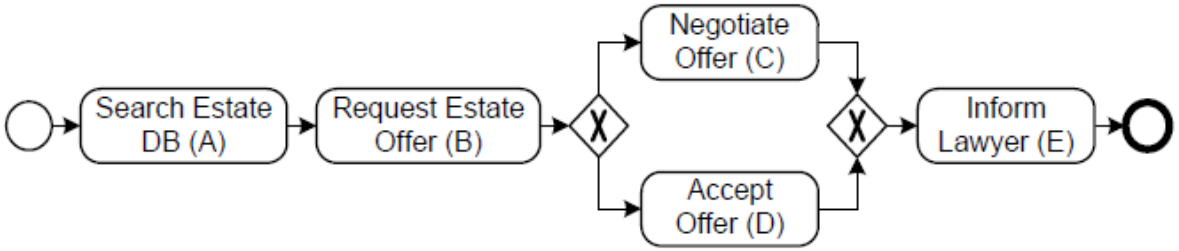
Apply the log-based ordering relations to traces of a process model

Again, a, b $\in$ A as two activities of a process model

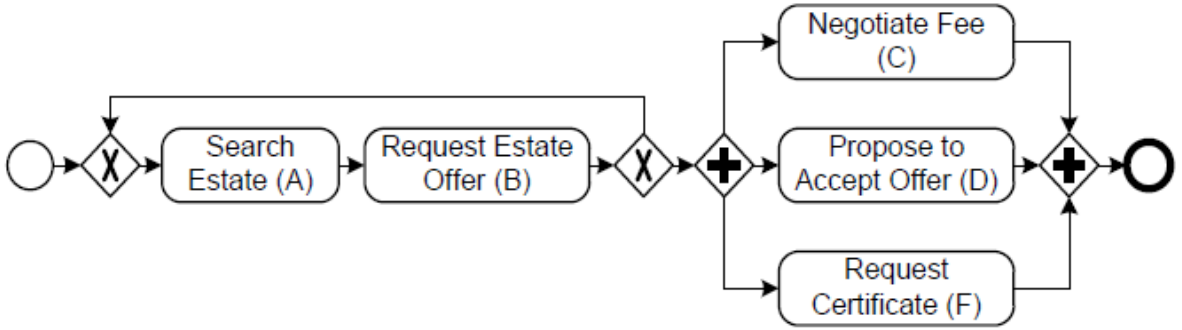
- Direct successor
  $a > b$ iff b directly follows a in a trace (e.g., “x a b y”)
- Causality
  $a \rightarrow b$ iff $a > b$ and not $b > a$
- Concurrency
  $a \parallel b$ iff $a > b$ and $b > a$
- Exclusiveness
  $a + b$ iff not $a > b$ and not $b > a$

Wil M. P. van der Aalst: Process Mining - Discovery, Conformance and Enhancement of Business Processes. Springer 2011

Footprint Example



(a)



(b)

	A	B	C	D	E
A	+	→	+	+	+
B	←	+	→	→	+
C	+	←	+	+	→
D	+	←	+	+	→
E	+	+	←	←	+

	A	B	C	D	F
A	+	||	+	+	+
B	||	+	→	→	→
C	+	←	+	||	||
D	+	←	||	+	||
F	+	←	||	||	+

Behavioural Profile

Technion
Israel Institute of Technology

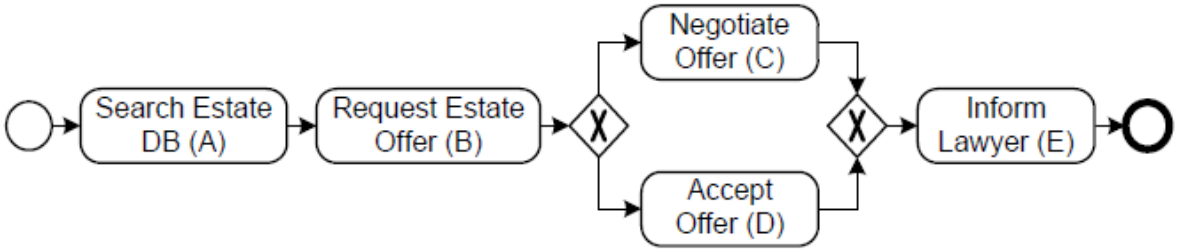
Obtain indirect relations by changing the underlying base relation from direct to indirect successorship

Again, a, b $\in$ A as two activities of a process model

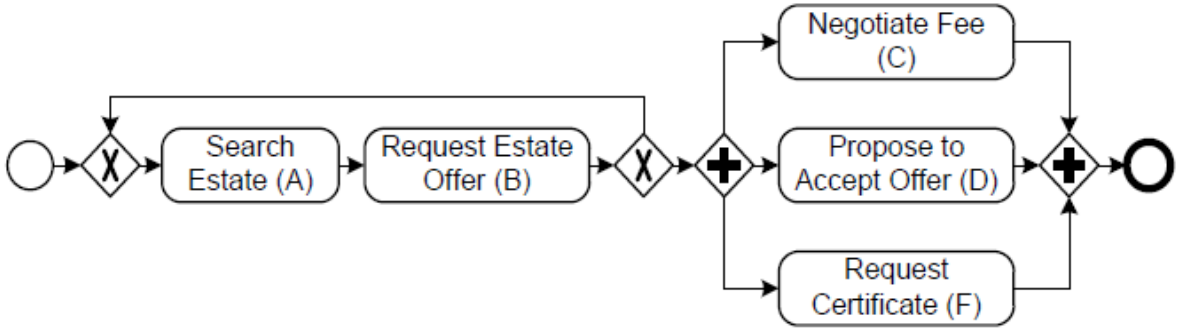
- Indirect successor
  $a > b$ iff b indirectly follows a in a trace (e.g., “x a y b z”)
- Order
  $a \rightarrow b$ iff $a > b$ and not $b > a$
- Interleaving
  $a \parallel b$ iff $a > b$ and $b > a$
- Exclusiveness
  $a + b$ iff not $a > b$ and not $b > a$

Matthias Weidlich, Jan Mendling, Mathias Weske: Efficient Consistency Measurement Based on Behavioral Profiles of Process Models. IEEE Trans. Software Eng. (TSE) 37(3):410-429 (2011)

Behavioural Profile Example



(a)



(b)

	A	B	C	D	E
A	+	→	→	→	→
B	←	+	→	→	→
C	←	←	+	+	→
D	←	←	+	+	→
E	←	←	←	←	+

	A	B	C	D	F
A	‖	‖	→	→	→
B	‖	‖	→	→	→
C	←	←	+	‖	‖
D	←	←	‖	+	‖
F	←	←	‖	‖	+

A Behavioural Similarity Metric

Technion
Israel Institute of Technology

Jaccard coefficient to quantify similarity of relations:

Exclusiveness Similarity $sim_{+}(\mathcal{R}_P, \mathcal{R}_Q) = \frac{|{+_P} \cap {+_Q}|}{|{+_P} \cup {+_Q}|}$

Order Similarity $sim_{\rightarrow}(\mathcal{R}_P, \mathcal{R}_Q) = \frac{|{\rightarrow_P} \cap {\rightarrow_Q}|}{|{\rightarrow_P} \cup {\rightarrow_Q}|}$

Disorder Similarity $sim_{||}(\mathcal{R}_P, \mathcal{R}_Q) = \frac{|\,||_P \cap ||_Q\,|}{|\,||_P \cup ||_Q\,|}$

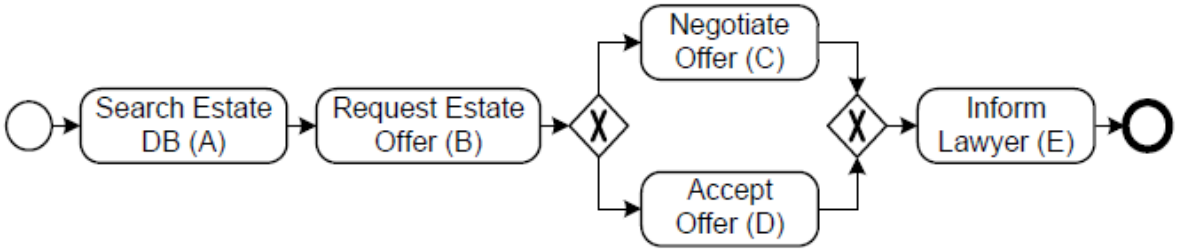
Aggregated similarity metric:

proven to be a metric (non-negativity, identity, symmetry, subadditivity)

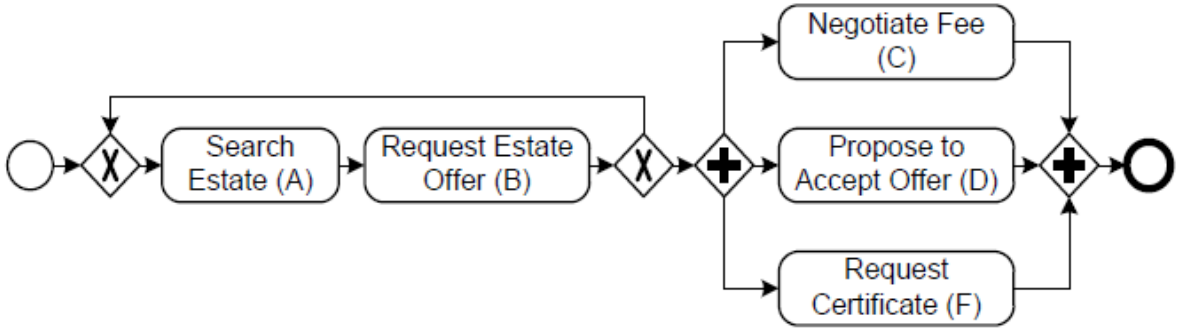
$$d_{\mathcal{R}}(\mathcal{R}_P, \mathcal{R}_Q) = 1 - \sum_i w_i \cdot sim_i(\mathcal{R}_P, \mathcal{R}_Q)$$

$$i \in \{+, \rightarrow, ||\} \qquad w_i \in \mathbb{R}, 0 < w_i < 1 \qquad \sum_i w_i = 1$$

Example



(a)



(b)

Footprint

$$sim_{+} \approx 0.733$$
$$sim_{\rightarrow} = 0.333$$
$$sim_{||} = 0$$

Profile

$$sim_{+} \approx 0.429$$
$$sim_{\rightarrow} = 0.666$$
$$sim_{||} = 0$$

Matthias Kunze, Matthias Weidlich, Mathias Weske: Behavioral Similarity - A Proper Metric. BPM 2011:166-181

Agenda

1) **Background on (process) model matching**
2) **Basic similarity measures**
   - Pre-processing of labels
   - Syntactic and semantic similarity
3) **Structural and behavioural matching**
   - Graph matching
   - Behavioural similarity measures
4) **Practical considerations**
5) **Research directions**

Tools and Libraries

Technion
Israel Institute of Technology

Several matchers have been proposed

- Most rely on combination of basic matching and structural or behavioural matching techniques
- Open source implementations available
- Integration into modelling environments typically not done

Next, two examples:

- RefMod-Mine/NSCM
- ICoP framework

Matcher: RefMod-Mine/NSCM

Technion
Israel Institute of Technology

Clustering of nodes based on word stem similarity

Approach

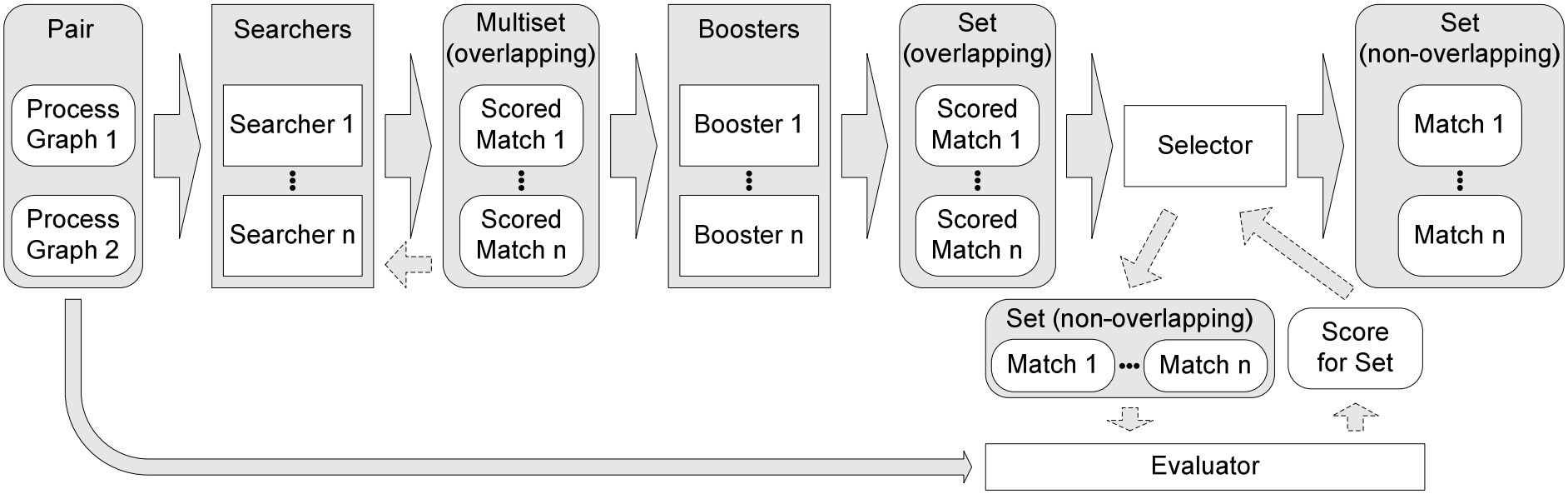
1. *Semantic error detection*: Identifies set of relevant nodes
2. *Similarity Measure:* Based on common word stems and taking into account antonyms / negation
3. *N-ary cluster matcher*: Agglomerative cluster creation based on similarity measure and definable threshold
4. *Binary matching extraction*: Extracts matches from node clusters

Implemented as a PHP command line tool using the Porter Stemmer and WordNet: https://code.google.com/p/refmodmine/

Matcher: ICoP Framework

Technion
Israel Institute of Technology

Architecture and a set of re-usable components for assembling concrete matchers



Implemented in Java: https://code.google.com/p/process-matching/

Process Model Matching Contest 2013

No common basis for evaluation of matching techniques

Process Model Matching Contest 2013 addressed this need

Comparative evaluation with two datasets

- University Admission Processes (UA)
- Birth Registration (BR) Processes

Characteristic	UA	BR
No. of labeled Transitions (min)	11	9
No. of labeled Transitions (max)	44	25
No. of labeled Transitions (avg)	22	17.9
No. of 1:1 Correspondences (total)	345	348
No. of 1:1 Correspondences (avg)	9.6	9.7
No. of 1:n Correspondences (total)	83	171
No. of 1:n Correspondences (avg)	2.3	4.75

Ugur Cayoglu, Remco M. Dijkman, Marlon Dumas, Peter Fettke, Luciano García-Bañuelos, Philip Hake, Christopher Klinkmüller, Henrik Leopold, André Ludwig, Peter Loos, Jan Mendling, Andreas Oberweis, Andreas Schoknecht, Eitam Sheetrit, Tom Thaler, Meike Ullrich, Ingo Weber, Matthias Weidlich: Report: The Process Model Matching Contest 2013. Business Process Management Workshops 2013:442-463

Results (UA)

No.	Approach	Precision	Recall	F-Measure
1	Triple-S	0.31	0.36	0.33
2	Business Process Graph Matching	***0.60***	0.19	0.29
3	RefMod-Mine/NSCM	0.37	0.39	0.38
4	RefMod-Mine/ESGM	0.16	0.12	0.14
5	Bag-of-Words Similarity	0.56	0.32	***0.41***
6	PMLM	0.12	***0.58***	0.20
7	The ICoP Framework	0.36	0.37	0.36

Results (BC)

No.	Approach	Precision	Recall	F-Measure
1	Triple-S	0.19	0.25	0.22
2	Business Process Graph Matching	0.55	0.19	0.28
3	RefMod-Mine/NSCM	***0.68***	0.33	***0.45***
4	RefMod-Mine/ESGM	0.25	0.18	0.21
5	Bag-of-Words Similarity	0.29	0.22	0.25
6	PMLM	0.19	***0.60***	0.29
7	The ICoP Framework	0.42	0.28	0.33

Agenda

Technion
Israel Institute of Technology

1) **Background on (process) model matching**
2) **Basic similarity measures**
   - Pre-processing of labels
   - Syntactic and semantic similarity
3) **Structural and behavioural matching**
   - Graph matching
   - Behavioural similarity measures
4) **Practical considerations**
5) **Research directions**

Further Directions

	Gathering	Managing	Analyzing	Visualizing
Volume				
Velocity				
Variety		Model Matching		
Veracity				

Further Directions

	Gathering	Managing	Analyzing	Visualizing
Volume				
Velocity				
Variety		Model Matching		
Veracity		Probabilistic Model Matching		

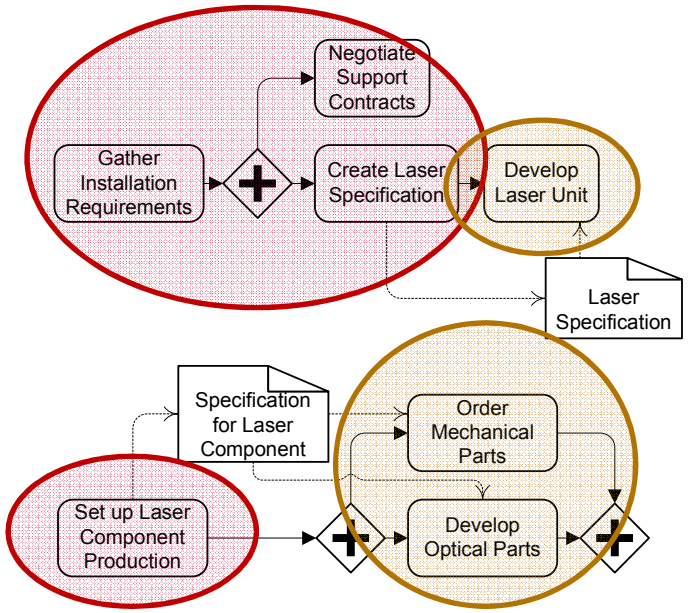
- From a deterministic “one match fits all”…
- … to multiple views, query dependent model matching

Open Issue: Complex Correspondences

Technion
Israel Institute of Technology

Getting back to the challenges:

- Refinement is hard to track on linguistic level
- Complex correspondences impose combinatorial challenges



Techniques for assessing completeness of correspondence

One direction: exploiting object lifecycles

Further Directions

Technion
Israel Institute of Technology

	Gathering	Managing	Analyzing	Visualizing
Volume	The Human in the loop			The Human in the loop
Velocity				
Variety		Model Matching		
Veracity		Probabilistic Model Matching		

- Visualizing large model matches
- New methods for gathering information from “experts”

Open Issue: Integration of User Feedback

Technion
Israel Institute of Technology

Matching is inherently uncertain

- Most scenarios require manual validation of correspondences
- Yet, complete manual validation may be infeasible

Directions

- Assess for which correspondences human input is most beneficial
- Go beyond validation of correspondences, but consider alignment of objects

Research Directions

	Gathering	Managing	Analyzing	Visualizing
Volume	The Human in the loop			The Human in the loop
Velocity				
Variety		Model Matching	Model Matching Prediction	
Veracity		Probabilistic Model Matching		

- Can I estimate the quality of my match without a golden standard?
- Generating an iterative process, where prediction guides the improvement

Take Away

Technion
Israel Institute of Technology

Process matching as a tool for various applications

Process matching is an uncertain process!

Basic measures to assess the similarity of the model entities

Structural and behavioural matching based on basic similarity measures

Active research field, e.g., complex correspondences and user input