

Added value of semantic technologies in the chemical industry

... and what's missing?

Text Mining Symposium

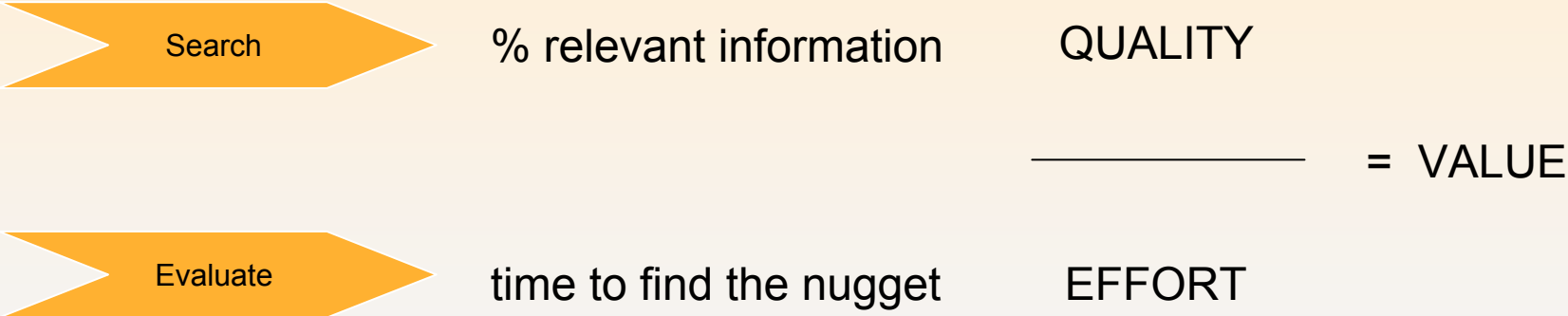
Bonn, 29.-30.09.2008

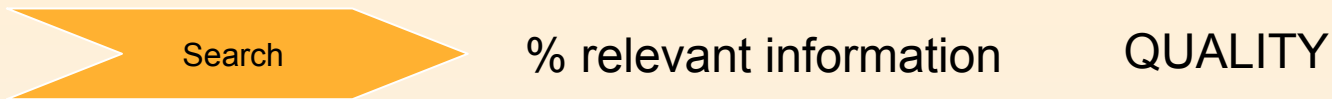
Dr. Heinz-Gerd Kneip, BASF SE, BASF Group Information Center
Head of Specialty Chemicals and Biotechnology Information

Agenda

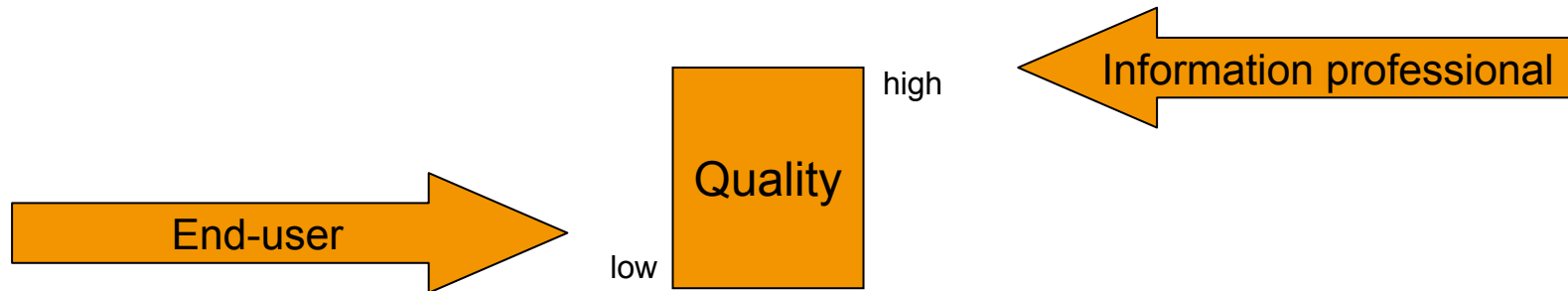
- Motivation ... for managers
- Added value ... for beginners
- Challenges ... for professionals
... for providers of semantic technologies

Search and evaluation: key parameters

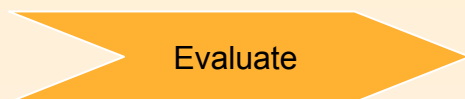




- The belief of end-users that Google offers everything in one place
- Often single word searches, rarely refinement of searches
- The belief of end-users that Google ranks the most important to the top
- ...

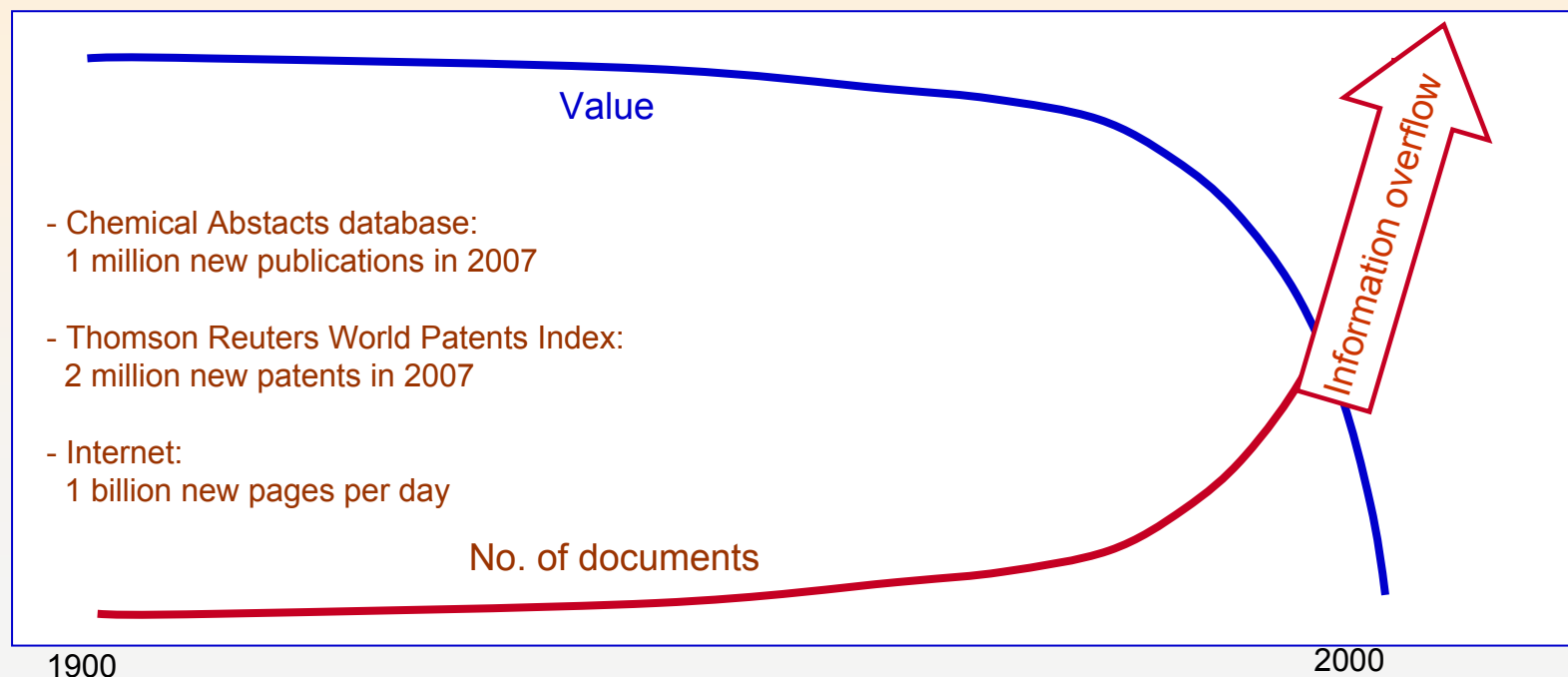


Evaluation & effort

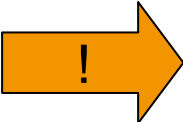
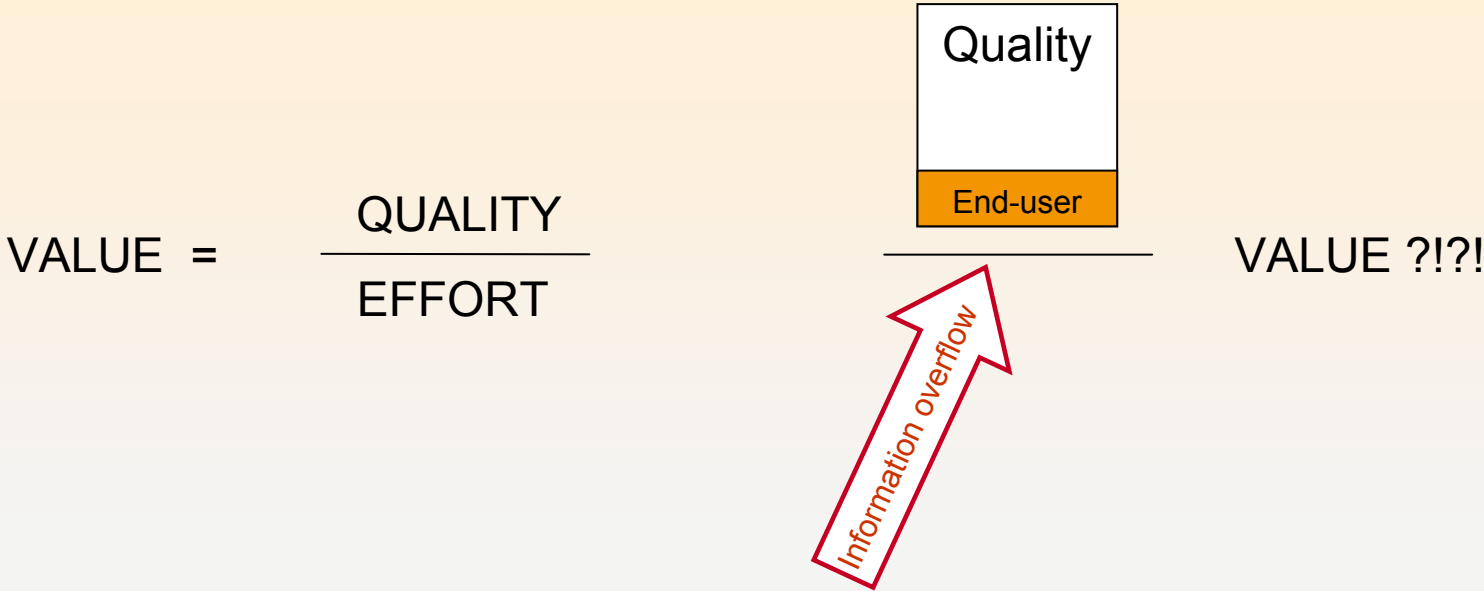


time to find the nugget

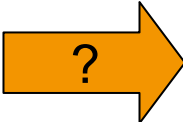
EFFORT



The challenges

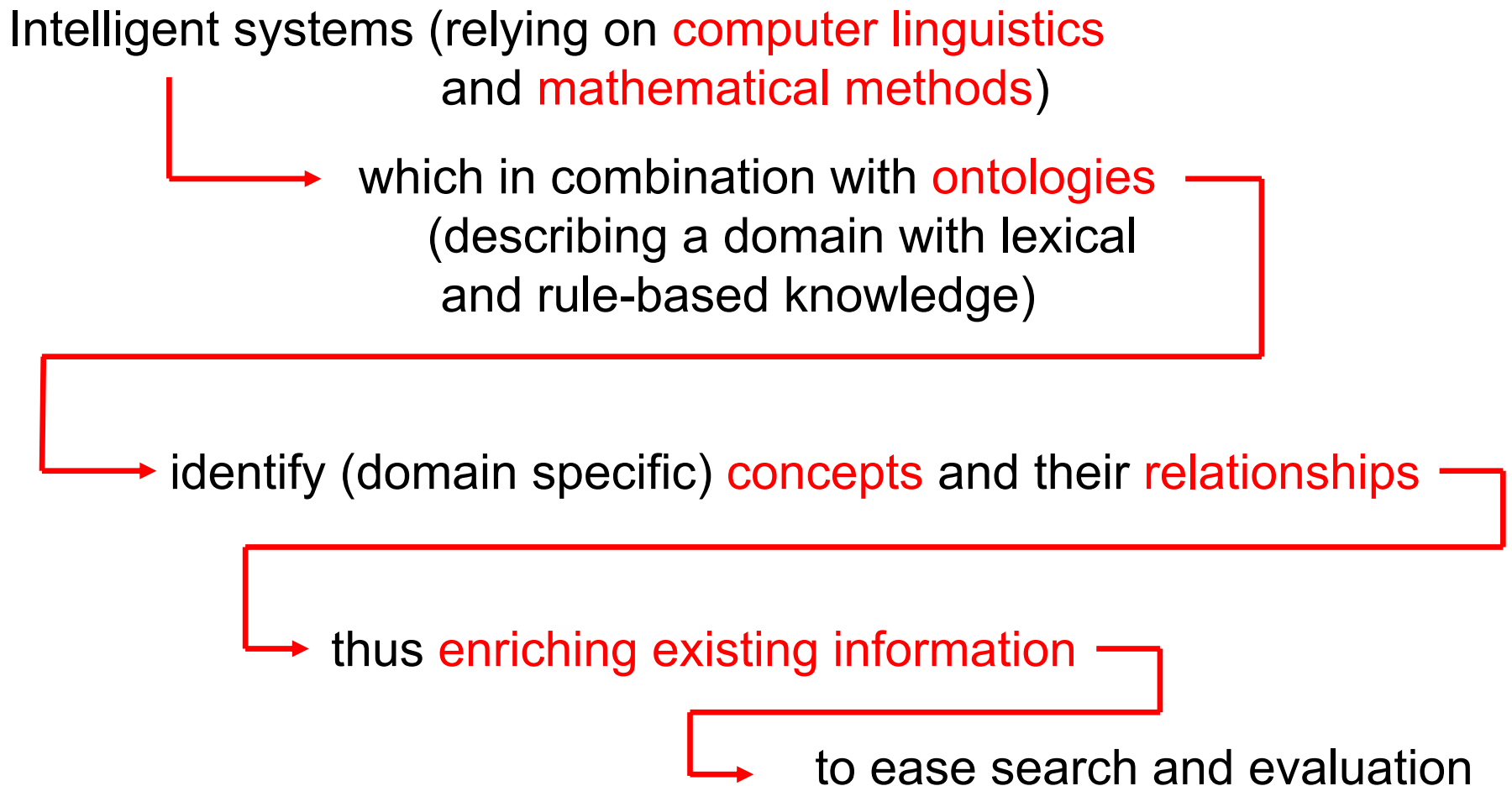


Time for solutions



Semantic technologies

Practical definition of semantic technologies



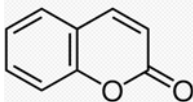
Range of applications (I): Entity identification & enrichment

- Identify domain specific vocabulary in text and add additional knowledge from other sources:

- *food additive* → nutrition → fields of application

- *leaf senescence* → plant trait → plant biotechnology

- *haloalkane dehalogenase* → EC 3 hydrolases → enzymes

- *coumarin* →  → chemical entity

- *aspirin* → acetyl salicylic acid

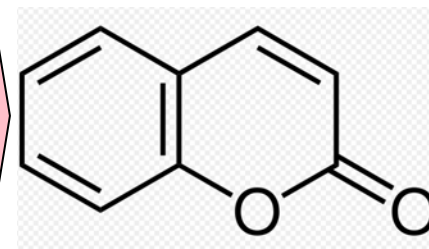
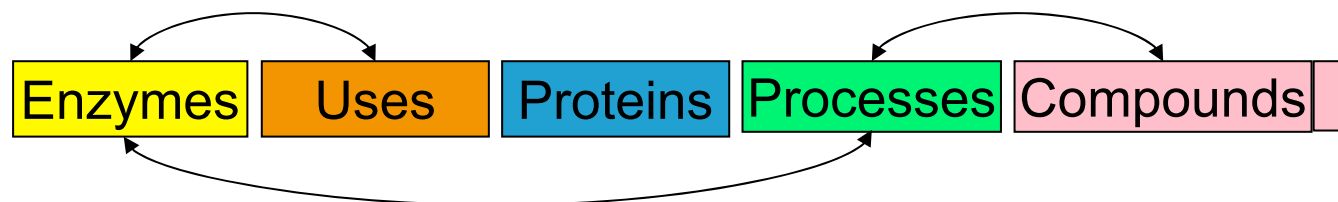
Which enzymes are used in nutrition?

Range of applications (I): Entity identification & enrichment

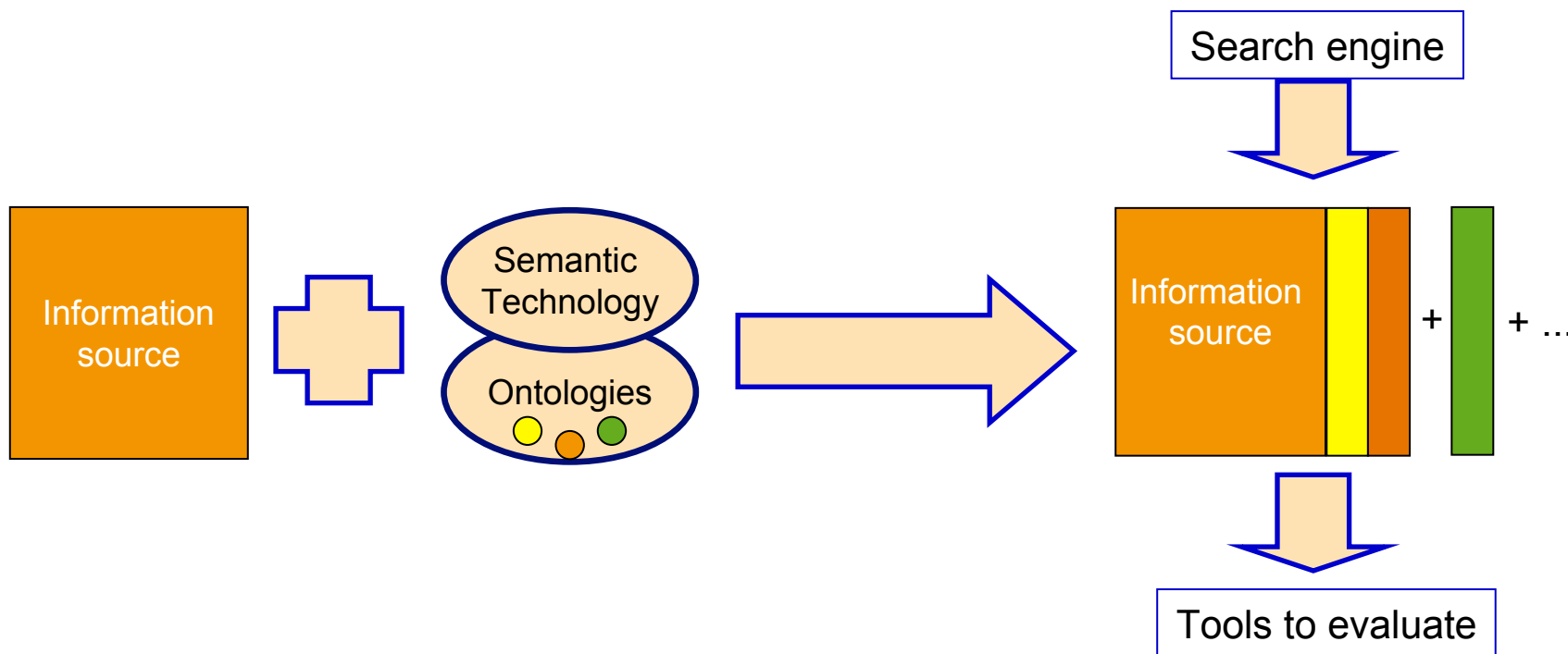
■ TITLE:
MANUFACTURE AND APPLICATION OF SHARK FIN PEPTIDE

■ ABSTRACT:

The shark fin peptide is manufd. by the steps of: (1) using fresh shark fin or dunked dry shark fin as raw material, freeze-drying, and pulverizing into powders, and (2) degrading the shark fin protein with ficin and bromelin in the presence of coumarin to obtain the final product. The shark fin protein can be used as medicinal raw material, medicine, health care material, food and food additive. The method has the advantages of simple process, low cost, low investment, rapid action, and high added value, and is convenient for mass prodn.



Range of applications (I): Entity identification & enrichment



Domain knowledge added to databases
for systematic re-use in search and evaluation

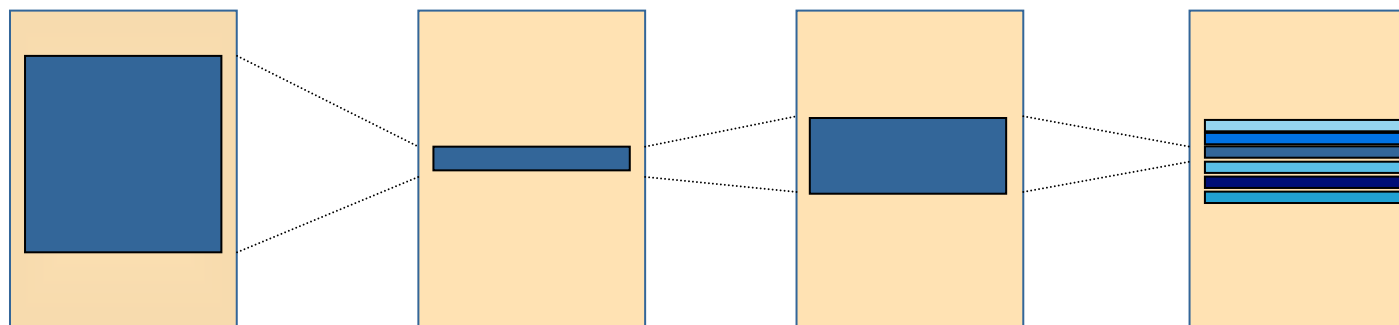
Range of applications (II): Query specification / expansion / refinement

Simple search:
one word

Query
specification

Query
expansion

Query
refinement



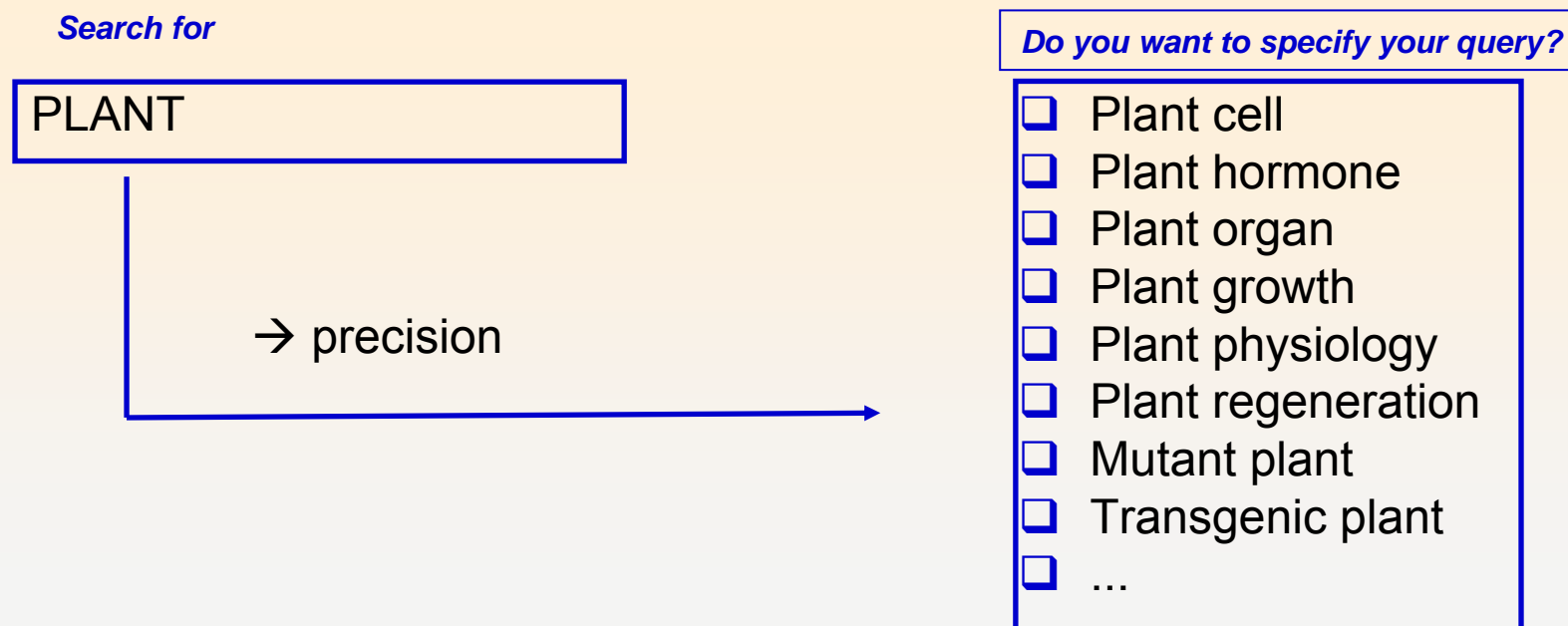
Recall: --
Precision: --

Recall: --
Precision: +

Recall: ++
Precision: +

Recall: ++
Precision: ++

Range of applications (II): Query specification



- Specific concepts of a search term = phrases containing the search term
- Standard linguistic processing of text to generate phrases → inverted index

Range of applications (III): Query expansion

Search for

Vitamin B1

→ recall

Do you want to expand with synonyms?

- Thiamin
- Aneurin
- Beivon
- Bethiamin
- Oryzanin
- Vitaneurin
- 49857-19-4
- 59-43-8
- ...

- Add synonyms, classification numbers, ...
- Ontologies / cartridges

Range of applications (IV): Query refinement

Search for

Magnetic layer

→ precision

Do you want to refine your search result?

- Coating process
- Coercive force
- Laminating
- Thickness
- Machine-readable
- Density
- Transparency
- ...

- Offer co-occurrent / related concepts
- Co-occurrence / statistical analysis / linguistic analysis

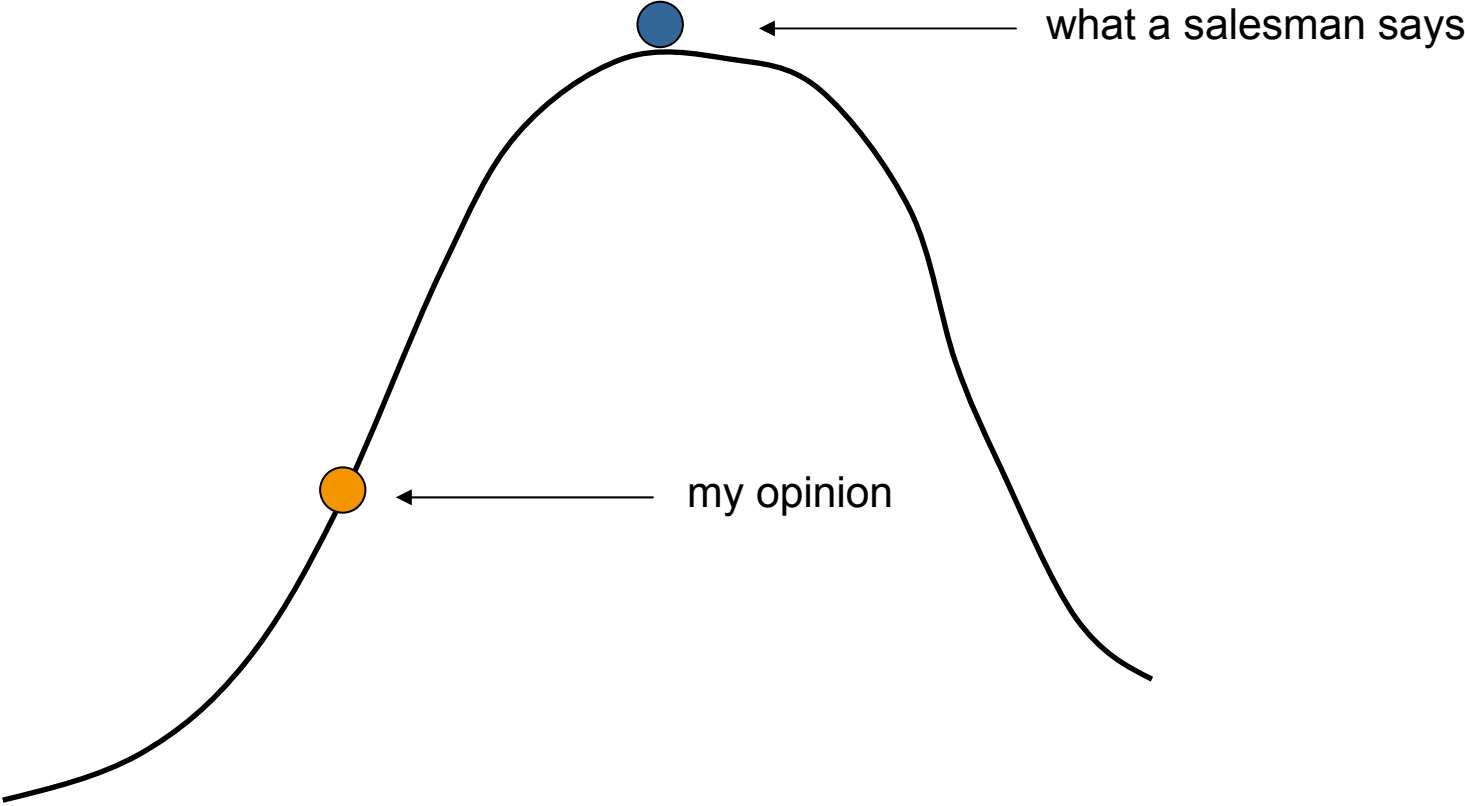
- TEMIS Luxid® licensed
- Luxid (off-the-shelf) used as a tool for information professionals to postprocess search results
- Integration of Luxid technologies into Fast® search engine via UIMA
- Luxid technologies used to
 - to build ontologies / cartridges focussing on BASF topics (Skill Cartridge Manager®, ...)
 - to semantically annotate text (Annotation Factory®, ...)
- Integration of heterogenous sources on a semantic level to enrich information for end-users (→ search & evaluation)

What's missing?

- R&D
- Ontologies
- Co-operation

...important to mention: experiences described on the following pages do refer to results from different vendors of text-mining solutions

Positioning of semantic technologies on a development curve

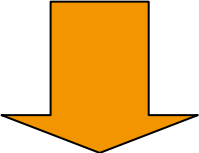


Some of today's unsatisfactory and unresolved issues

- Special characters and abbreviations
- Sentence boundaries
- Grammatical assignment of technical terms
- Lemmatisation of technical terms
- Verbs as nouns
- Anaphoric resolution
- Directions in relationships
- Disambiguation
- English \leftrightarrow patent jargon
- Negations
- Facts vs. hypotheses
- Enumerations
- Line breaks and formatings
- Inverted terms
- False positives
- Weighting of entities (esp. in full text)
- Structures from images
- Markush recognition

Some everyday examples: Sentence boundaries vs. abbreviations

The majority of available PFP appears to be best used as an alc. fermn. base, with the possibility of using the carotenoids (byproduct) as a food colorant.



rule-based extraction of uses



- an alc.

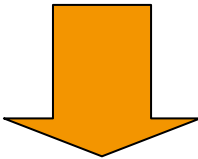


- alc. fermn. base
- alcoholic fermentation base



Some everyday examples: Linguistic rules vs. meaning

Curcumin suppressed protein expression of the NF-κB-regulated IκB



Extraction of biological relationships



- Curcumin suppresses protein
- Curcumin suppresses protein expression



- Curcumin suppresses IκB (expression)
- NF-κB regulates IκB



Some everyday examples: Chemical entity recognition

Text

di(2,2,6,6-tetramethylpiperidin-4-yl) N-(2,2,6,6-tetramethylpiperidin-4-yl)-.beta.-aminodipropionate

3-(4'-methoxy-benzylidene)-1,7,7-trimethyl-bicyclo[2.2.1]heptan-2-one

neopentyl glycol

Cetyl/stearyl alcohol, etherified with 12 mol of ethylene oxide

Mixture of p-hydroxybenzoic acid esters

Extracted chemical entity



n-(2,2,6,6-tetramethylpiperidin-4-yl)-.beta.-aminodipropionate



heptan-2-one



glycol



Ethylene



p-hydroxybenzoic acid



Some everyday examples: Anaphoric resolution

TITLE

MODIFICATION OF POLY (ETHYLENE OXIDE)-POLY (PROPYLENE OXIDE)-
POLY(ETHYLENE OXIDE) WITH RGD PEPTIDE

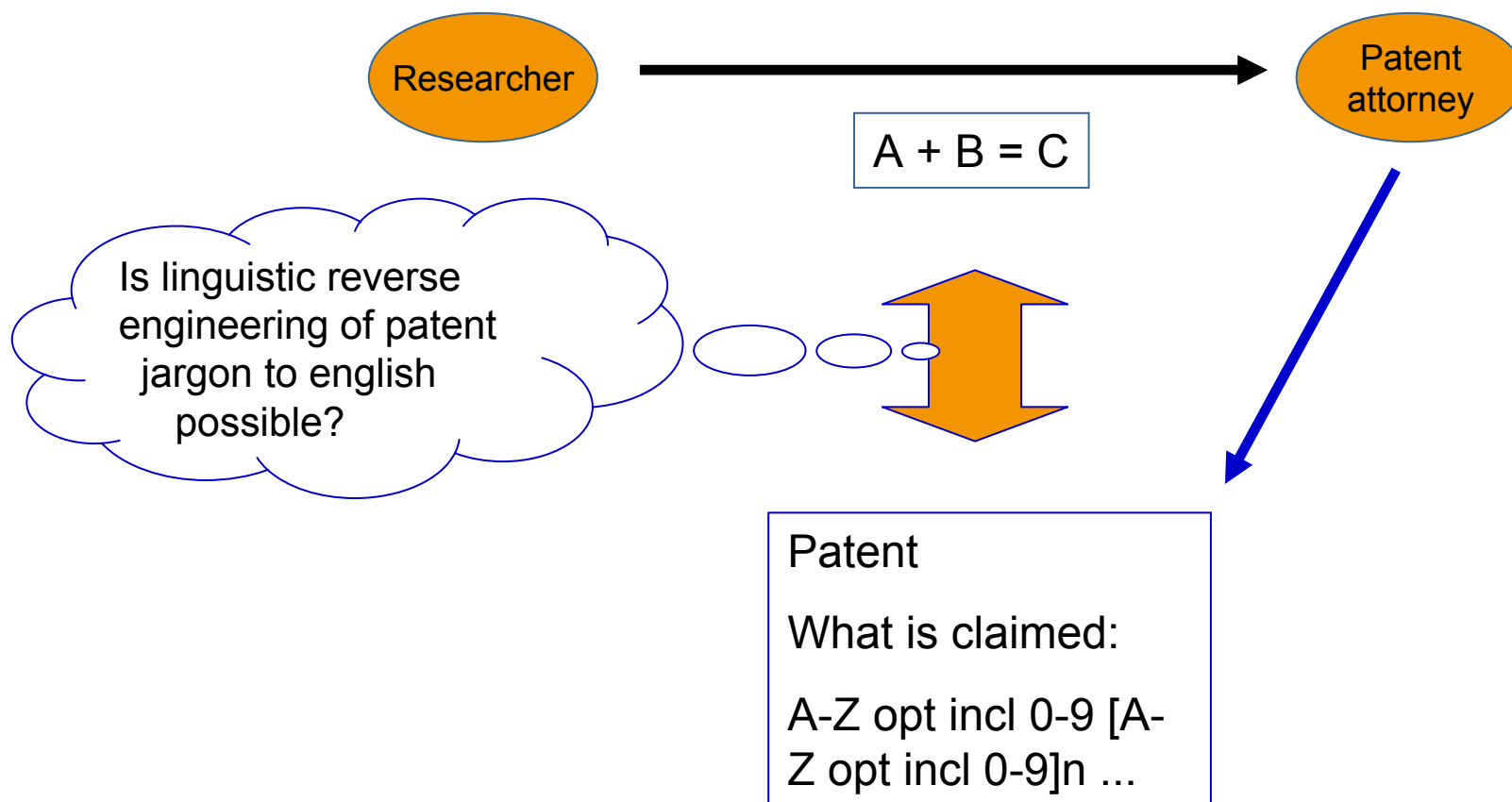
ABSTRACT

A novel method for the modification of Pluronic P105 with RGD peptide (= formula 1) for the possible application in drug delivery systems is described. GRGD was successfully conjugated with PEO-PPO-PEO by the novel method suggested. Formula 1 could serve as an injectible hydrogels for drug delivery. The modified polymers can potentially be used to target cancer cells that expresses RGD specific integrins.

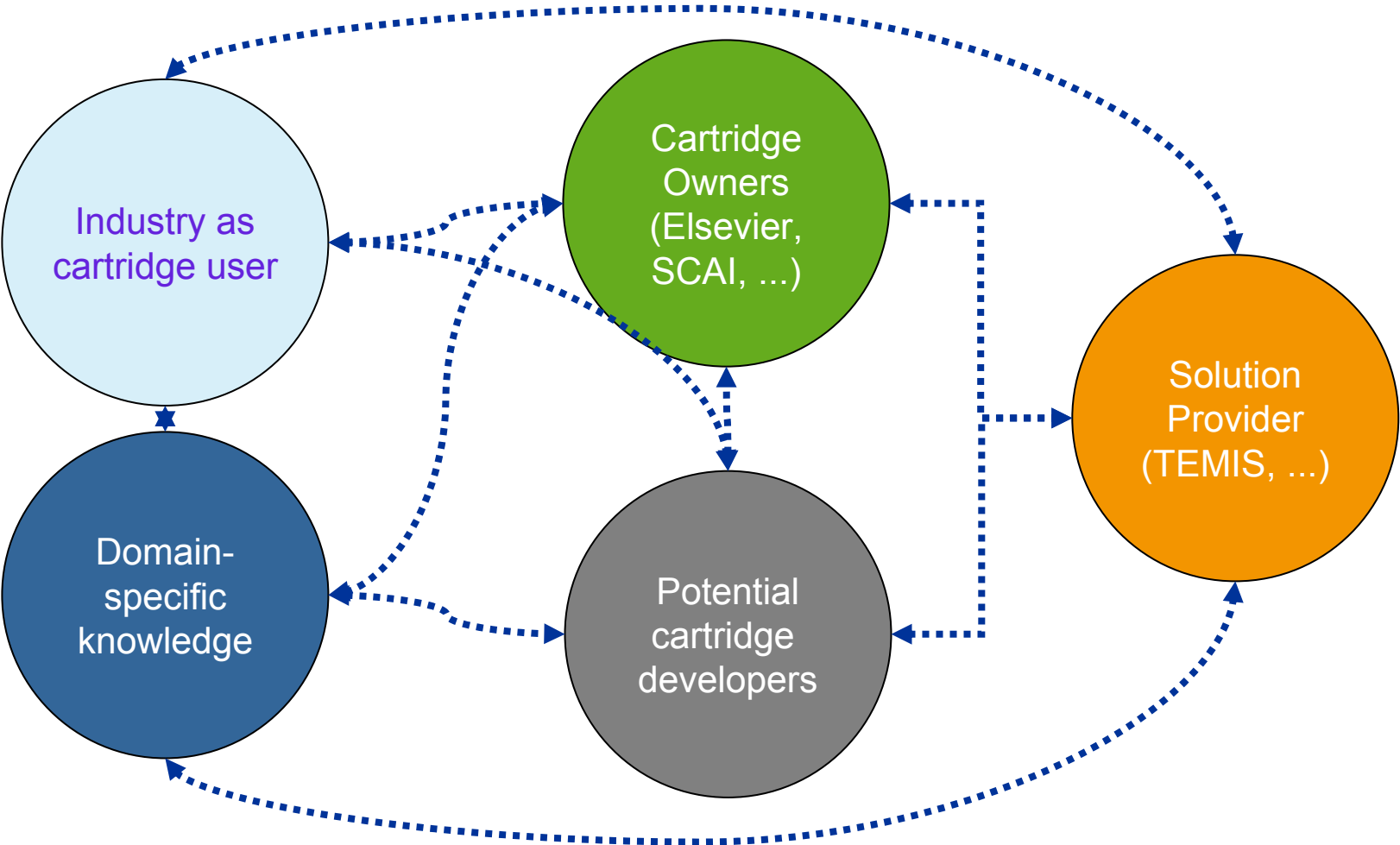
Formula 1 = modification of Pluronic P105 with RGD peptide



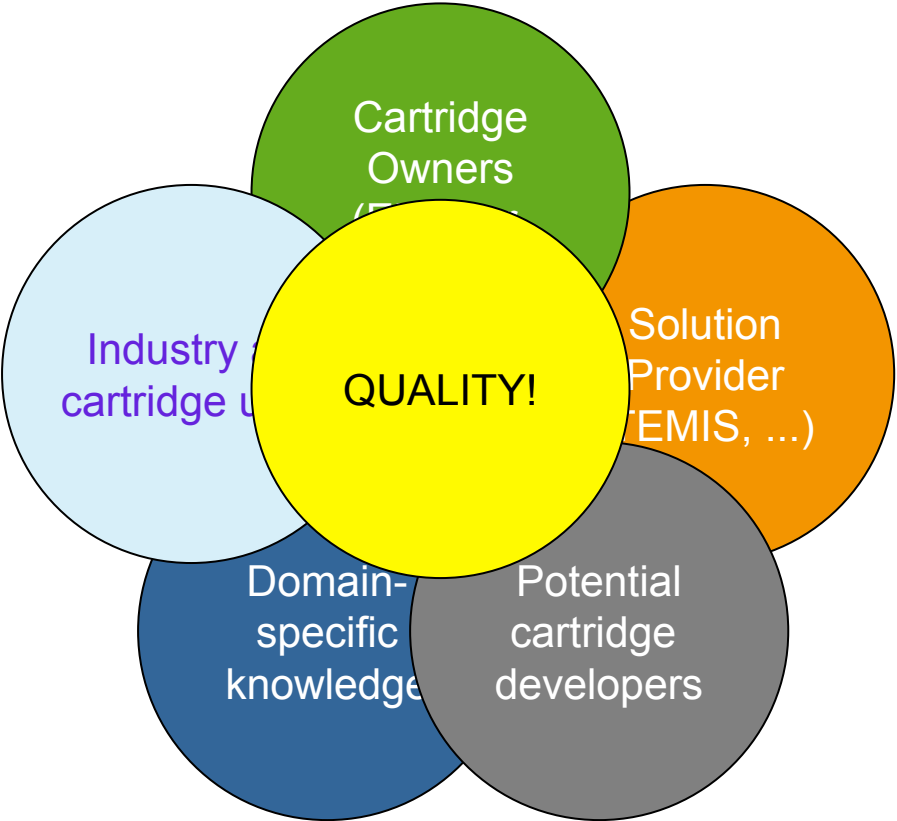
English \leftrightarrow patent jargon



What's missing: co-operation



Co-operation: the ideal situation



- In using semantic technologies we can make a significant step forward

- Semantic technologies can still go a step further:
 - Market leaders should be drivers for R&D!

 - Missing ontologies – who is able to offer the right quality?

 - Co-operate! Co-operate! Co-operate!

Thank you very much for your attention!

Time for discussion!