International Symposium on Integrative Bioinformatics 2010

The LAILAPS Search Engine

-

A Feature Model for Relevance Ranking in Life Science Databases

M Lange, K Spies, C Colmsee, S Flemming, M Klapperstück, U Scholz

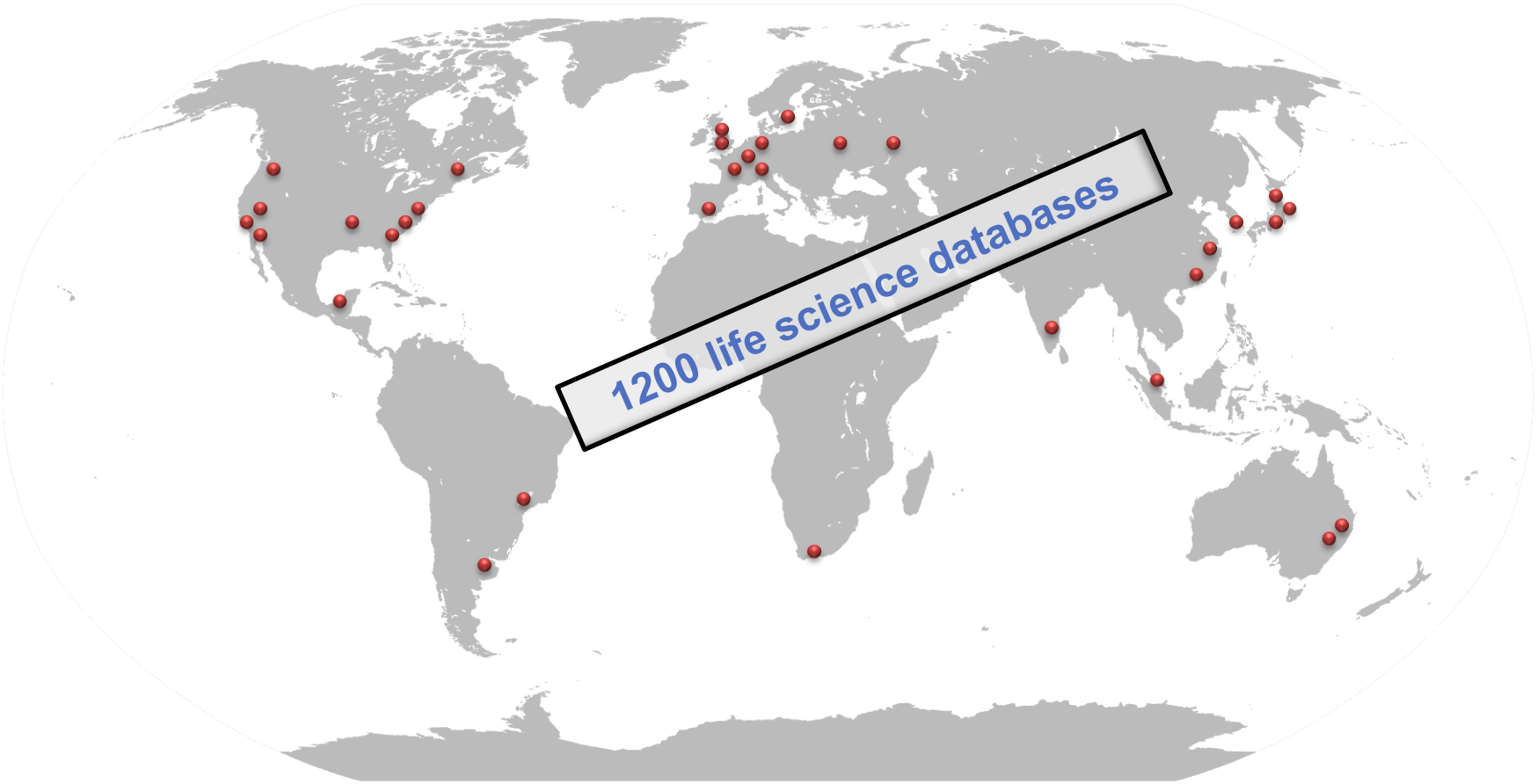
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

lange@ipk-gatersleben.de
IPK Gatersleben, March 22, 2010

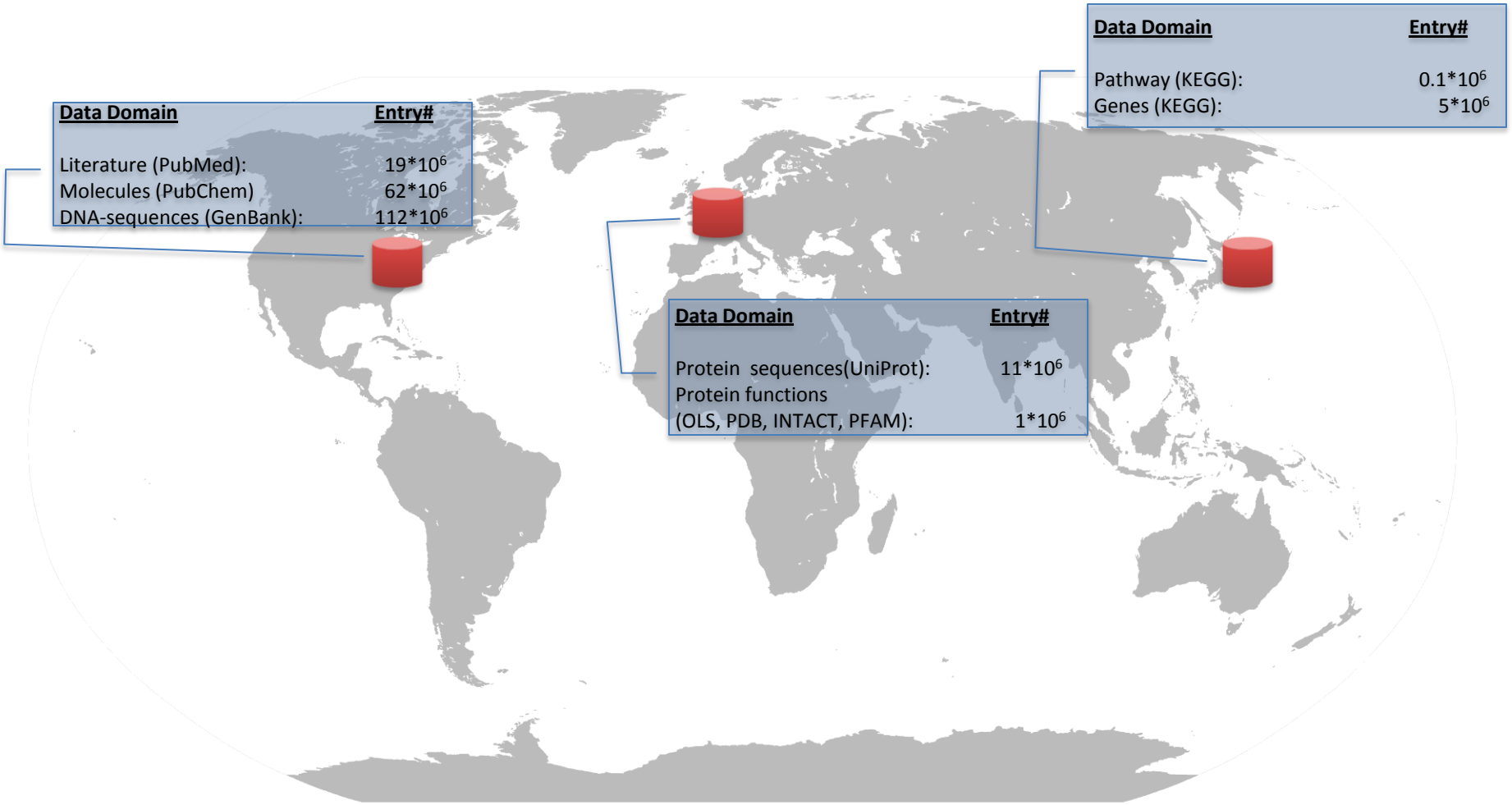
Outline

- Information Retrieval in Live Science
- The LAILAPS Approach
- Query Engine & Feature Scoring
- Relevance Prediction
- The LAILAPS System

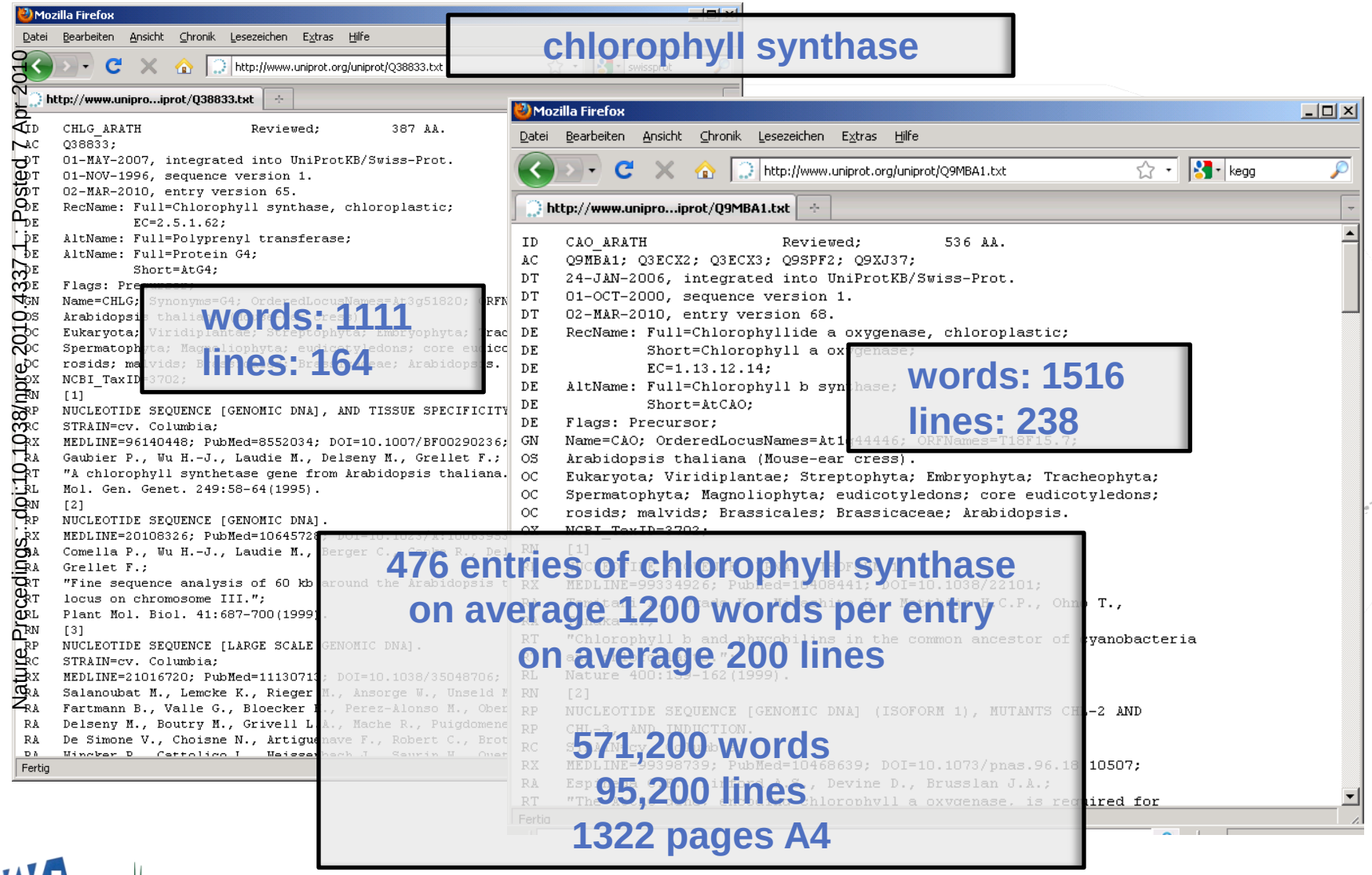
Life Science Data Universe



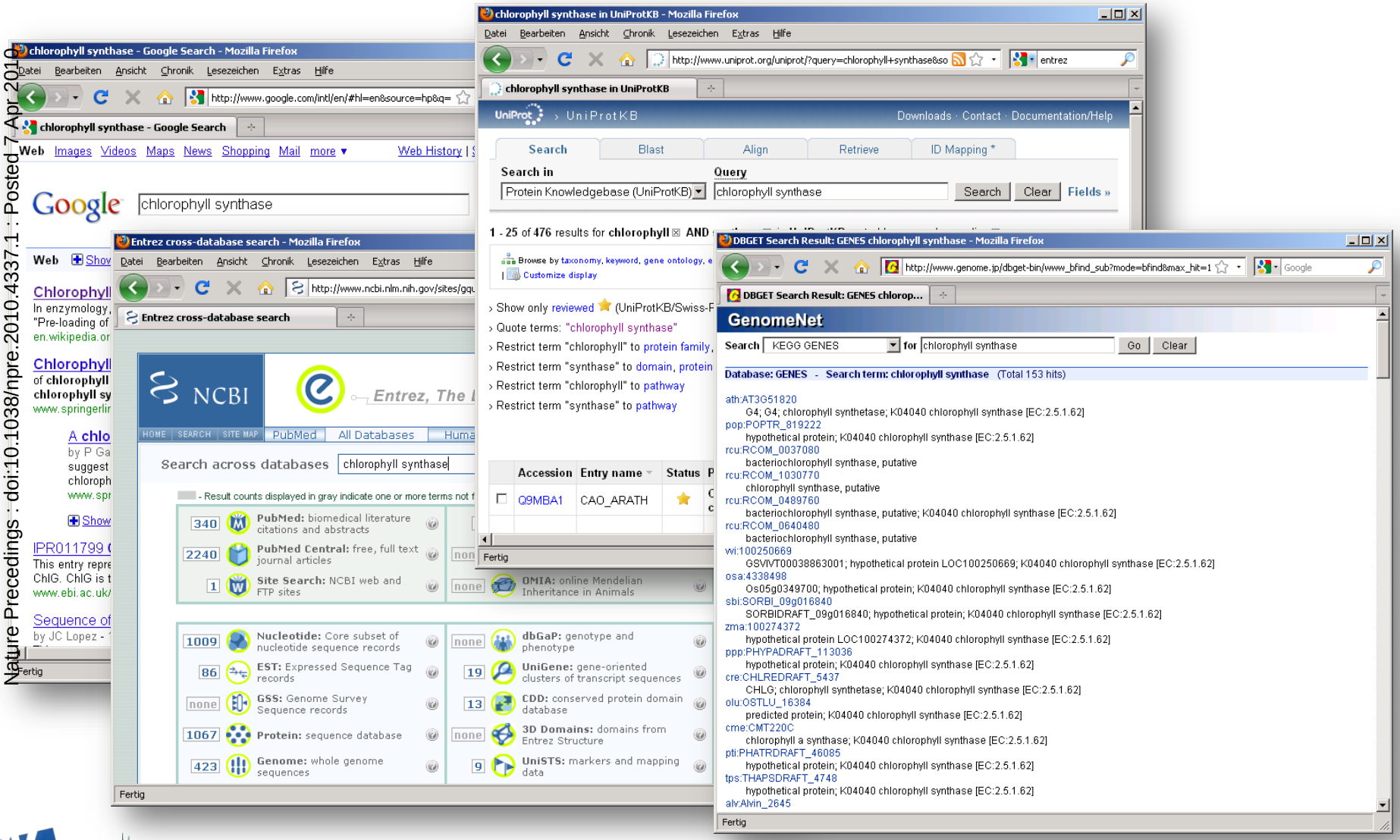
Life Science Data Universe



Life Science Data Universe



Life Science Data Universe - Search



The Story of Searching Data

1. 37% of life science scientists spending over 80% of working time in front of a computer
2. 47% of all scientist make daily use of search engines

Divoli A, **Hearst MA**, **Wooldridge MA.**, **Evidence for showing gene/protein name suggestions in bioscience literature search interfaces.,** Pac Symp Biocomput. 2008:568-79

When you are doing your work, approximately what percentage of the time involves your using a computer?			
0-20 %		2	5%
20-40 %		7	18%
40-60 %		8	21%
60-80 %		7	18%
80-100 %		14	37%
	Total	38	100%

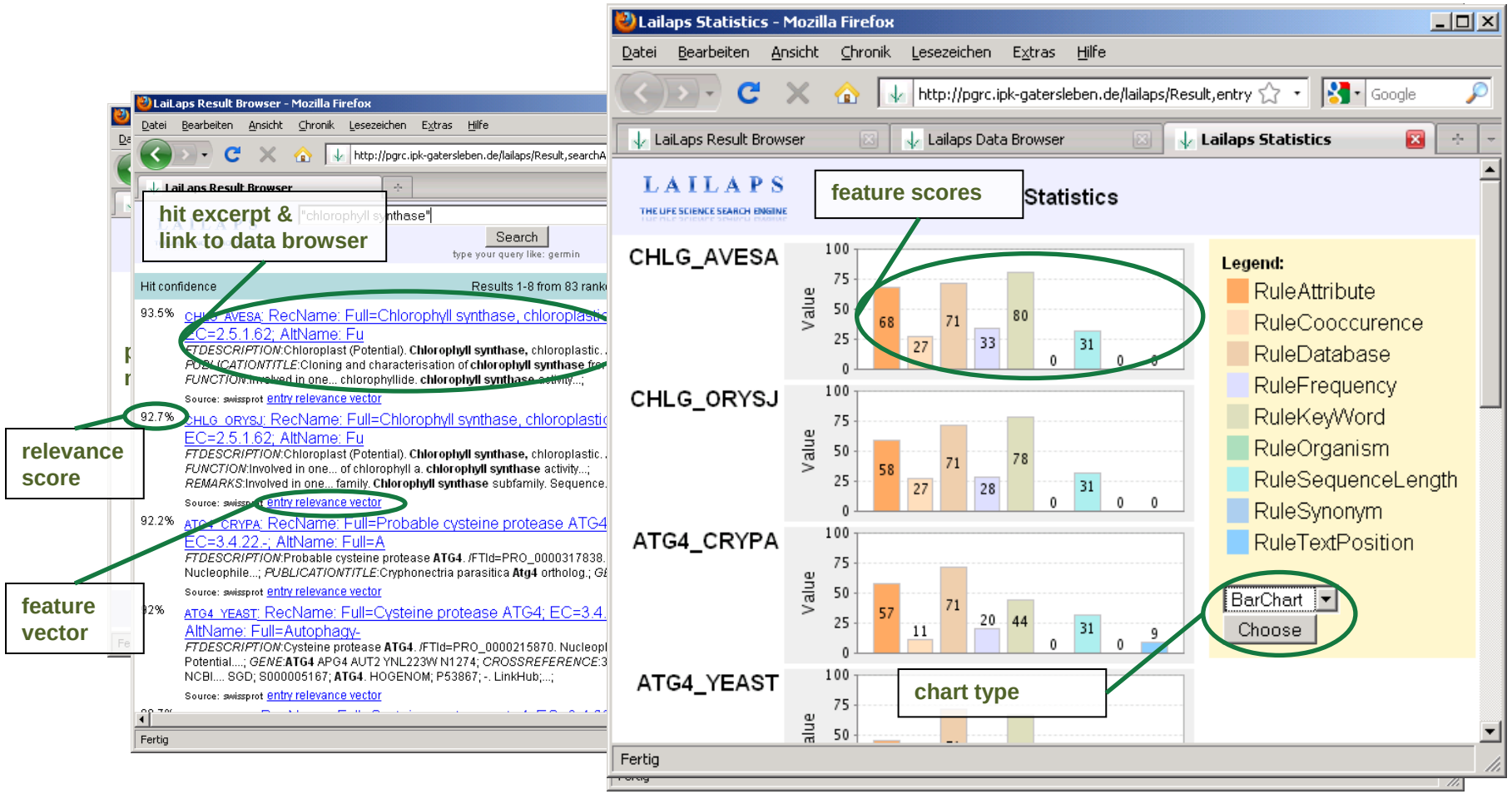
How often do you search the biomedical literature?			
Every day		18	47%
Every week		14	37%
Every month		3	8%
Rarely		3	8%
Never		0	0%
	Total	38	100%

The LAILAPS Approach

Motivation for LAILAPS

- platform for information retrieval over isolated databases
- content sensitive relevance ranking
- user profiles for relevance estimation
- estimation of relevance factors by tracking user behavior

LAILPAS Search Engine



Nature Precedings : doi:10.1038/npre.2010.4337.1 : Posted 7 Apr 2010

LAILAPS Feature Model

1. attribute in the record
2. database in which the hit was found
3. frequency of hits per attribute and record
4. co-occurrence of query terms
5. keyword surrounding the hit position
6. organism that is reported in the record
7. sequence length
8. text position of the hit in the record structure
9. synonym that produced the hit

LAILAPS Ranking Workflow

query
$Q = t_1 \wedge \ldots \wedge t_n$

chlorophyll synthase

find:
using a text
index system

query result
R = (t,d,{(a,p)})

('chlorophyll', 'Q9MBA1', {('DE', 83), ('RT',2)}
('synthase', 'Q9MBA1', {('DE', 97)}

score:
using features
$F_1 \ldots F_9$

entries feature scores
$(\omega_1 \ldots \omega_9)$

$score_{freq}$= 3
$score_{co\text{-}ocurence}$= 1-(distance/maxdist) *
{cor: 1; wrong: 0.3}
. . .

relevance $(\omega_1 \ldots \omega_9)$:
using a feed forward
neural network

relevance score
τ

relevance ranking
$\tau_1 \ldots \tau_n$

$N_{feedforward}$:(neurons = 20, input =9, output =1, edges = 7 x 4)
τ_{Q9MBA1} = N(68, 27,71,33,80,0,31,0,0) = 0.935

rank:
sorting of
relevance scores

Query Engine: Text Index

- decompose text into textual atoms
  - avoid line breaks in tokens
  - do not decompose number groups
  - keep chars in token _-/*+.
  - ignore chars in token ()[]<>'',:'
  - token delimiter whitspace and =;\t
- stop words (ignore frequent but useless tokens)
  - standard in natural language
    - but, his, nor, than, was, by, how, not, that, we, can, however, of, the, were …
  - more in life sciences
    - product, locus_tag, db_xref, region, source, DB, organism, CDS, xref, interpro, submission, submitted, pfam, binding, table, embl, geneid, genomic, transl, cdd, sequence, pubmed, taxon, IEA, NIH, genbank, data, here …
- synonyms (from databases Brenda, UniProt)
  - Synonyms Relations: 67521
  - Protein names: 76869

Relevance Prediction

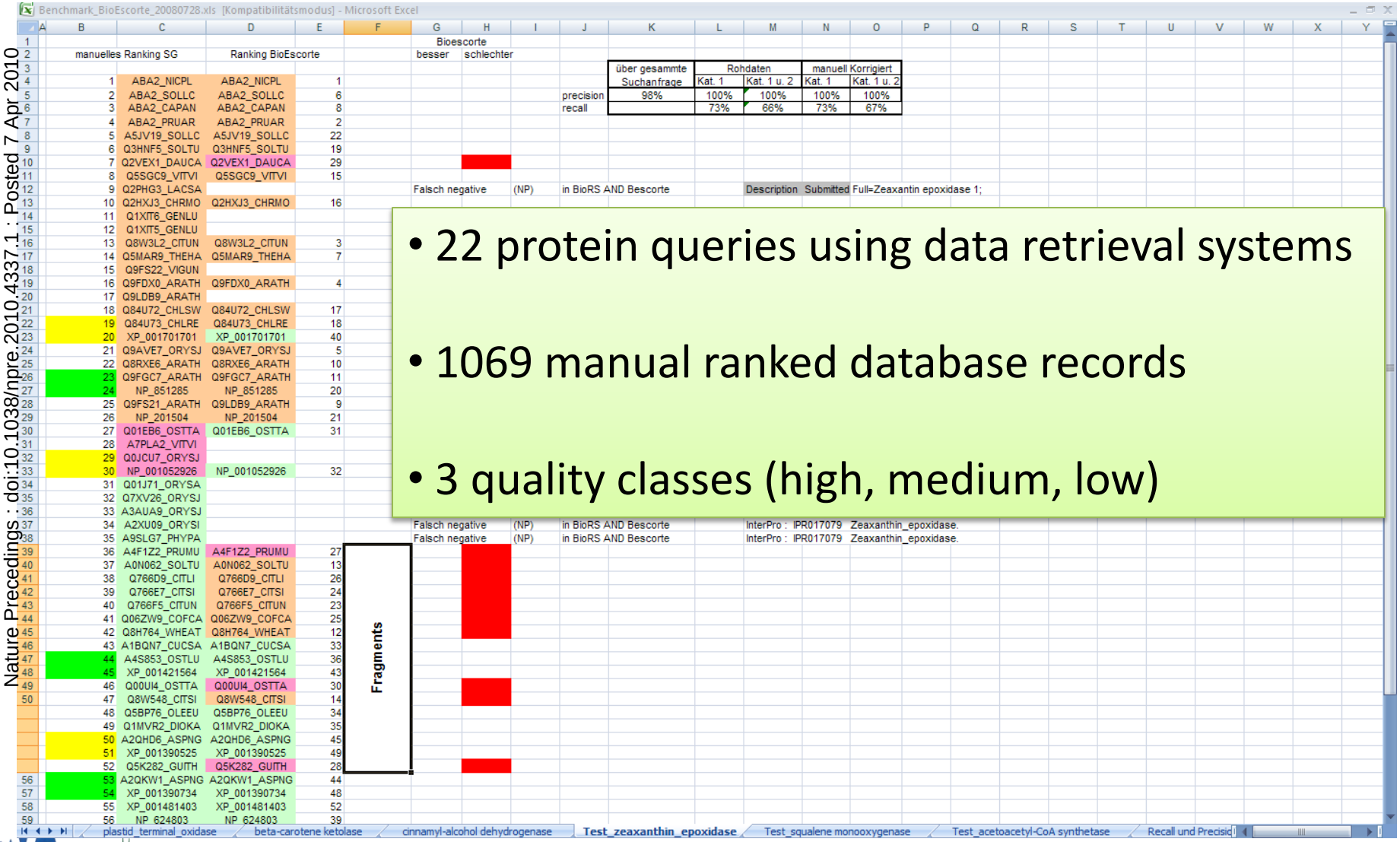
Relevance Prediction: Confidence Classes

confidence class	accession	curated rank	lailaps rank	relevant	problem
1	Acc1	1	5	no	wrong ranking
	Acc2	2	2	yes	
	Acc3	3	3	yes	
2	Acc4	-	4	-	additional hit (e.g. synonm)
	Acc5	4	5	yes	
	Acc6	5	-	no	database version
	Acc7	6	6	yes	
3	Acc8	7	7	yes	
	Acc9	8	-	no	text decomposition
	Acc10	9	8	yes	
	Acc11	10	1	no	wrong ranking

Relevance Prediction: Training Data

Query Text	Size	Category Split-Up (hi/me/lo)
industrial use case 1	20	6 / 4 / 10
”pinene synthase”	18	10 / 3 / 5
industrial use case 2	39	8 / 13 / 18
industrial use case 3	64	14 / 32 / 18
”gamma tocopherol methyltransferase”	38	21 / 9 / 8
”ent-kaurene synthase”	65	17 / 38 / 10
”chlorophyll synthase”	77	17 / 54 / 6
industrial use case 4	134	35 / 68 / 31
”cinnamyl-alcohol dehydrogenase”	214	45 / 36 / 133
industrial use case 5	17	3 / 4 / 10
”dihydrokaempferol 4-reductase”	65	9 / 29 / 27
”l-ascorbate peroxidase”	100	69 / 12 / 19
”morphine 6-dehydrogenase”	35	2 / 15 / 18
”zeaxanthin epoxidase”	51	21 / 2 / 28
”squalene monooxygenase”	84	24 / 30 / 30
”acetoacetyl-coa synthetase”	68	14 / 36 / 18

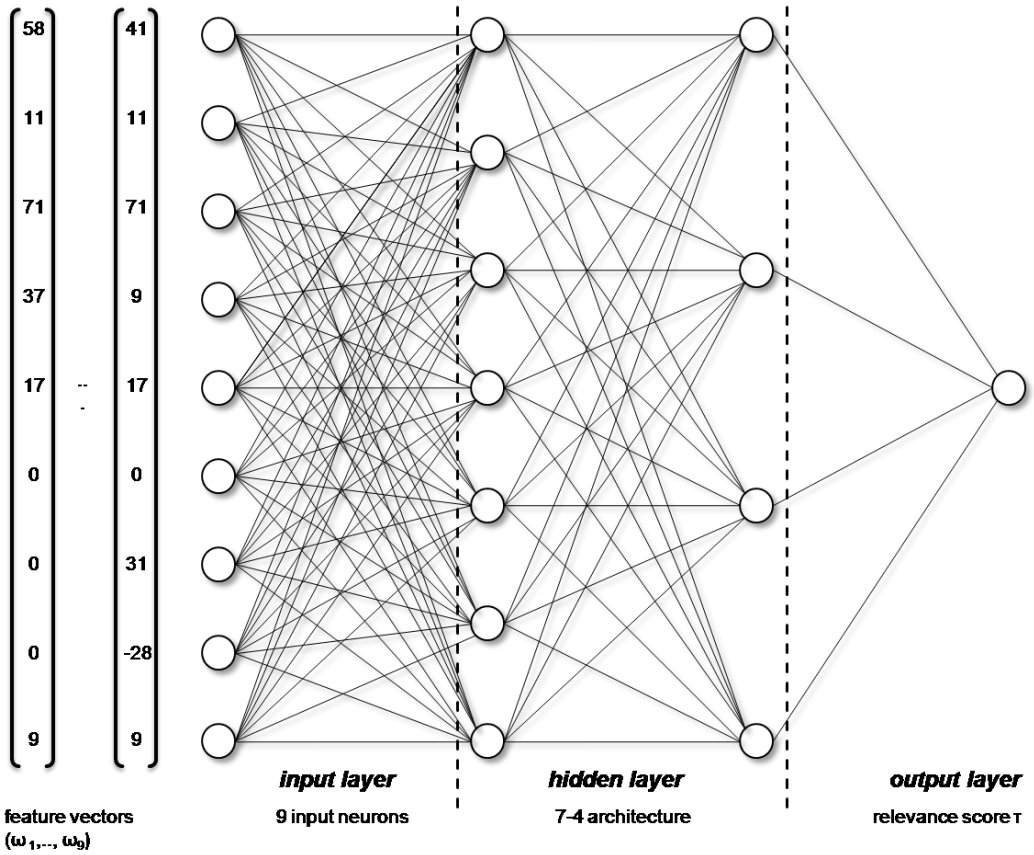
Relevance Prediction: Training Data



- 22 protein queries using data retrieval systems
- 1069 manual ranked database records
- 3 quality classes (high, medium, low)

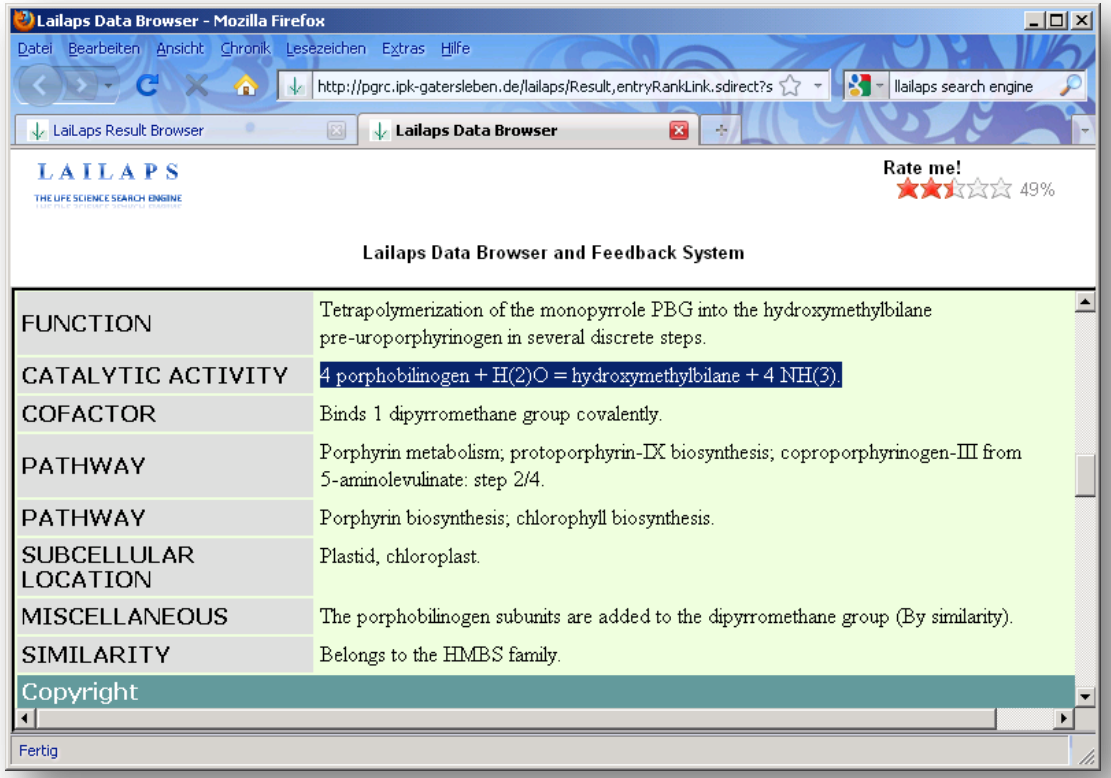
Relevance Prediction: Neural Network Training

- split of training/validation data: 80/20
- training epochs: 500
- mean square error: 0.33
- type: feed forward
- architecture: 7-4

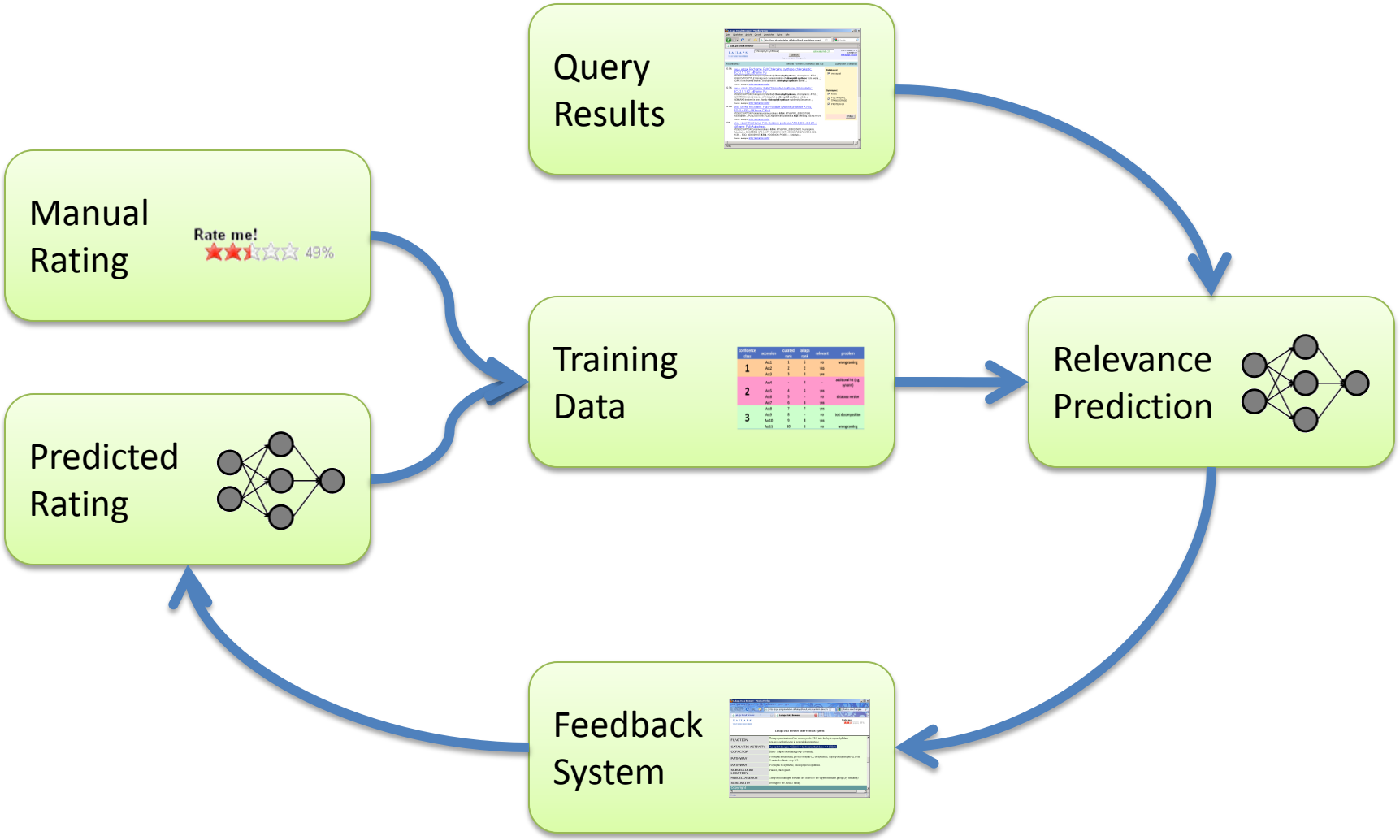


User Ranking Profile: Relevance Rating

- Explicit rating
- Implicit rating: tracking browsing behavior using JavaScript
  - clicked result entries
  - clicked entries above, below
  - activity time
  - scroll amount
  - mouse movement
  - lost / got focus
  - text selection

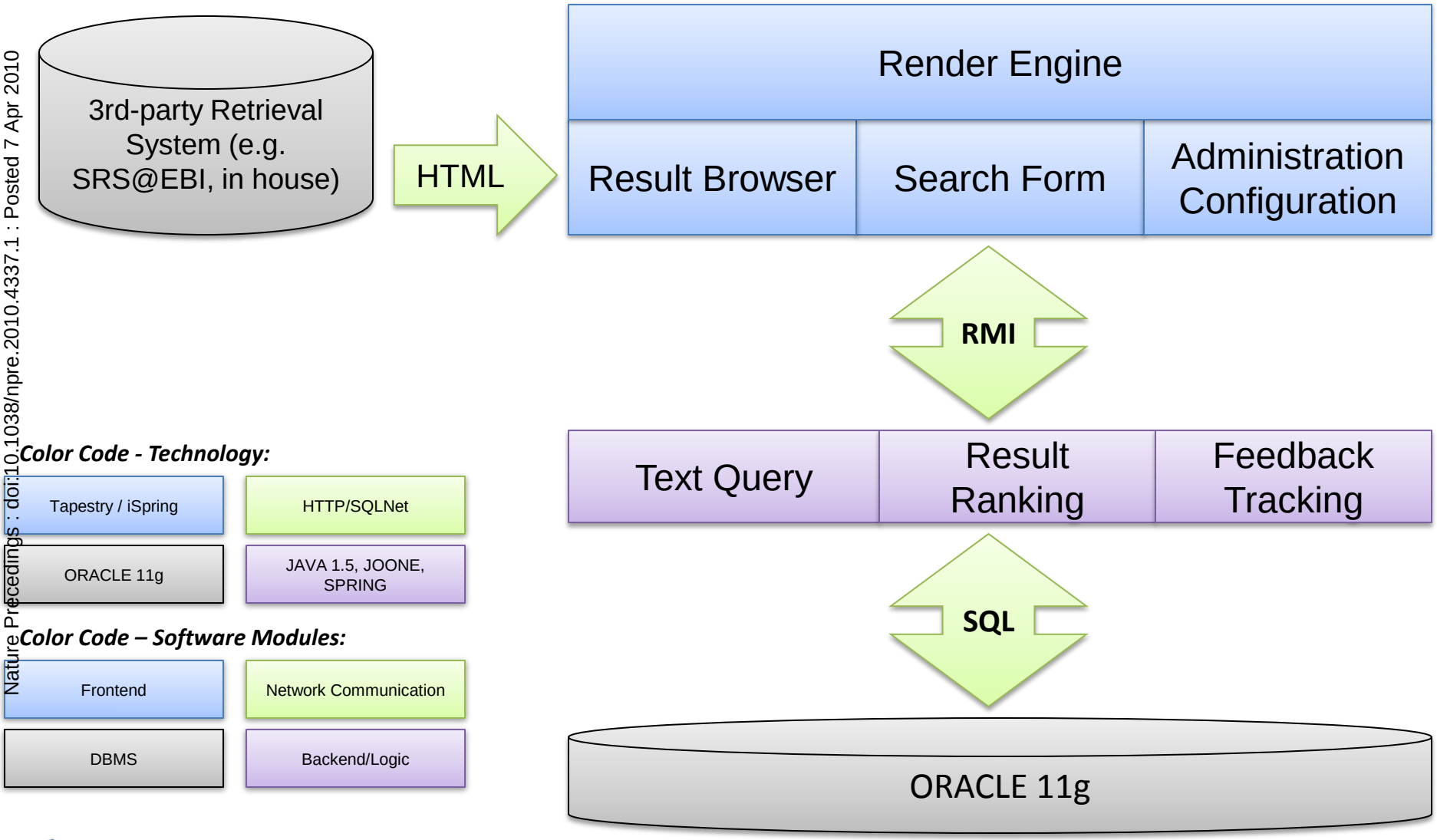


Relevance Prediction: Workflow



The LAILAPS System

LAILAPS – Technical Background



LAILAPS Customization

- database to be indexed

_ID_	AccNumber	Authors	Complex	Compound	Crossreference	CrystalConditions
100D	C.BAN	B.RAMAKRISHNAN M.SUNDARALINGAM	C.BAN	B.RAMAKRISHNAN M.SUNDARA		
101D	D.S.GOODSELL	M.L.KOPKA R.E.DICKERSON	D.S.GOODSELL	M.L.KOPKA R.E.DIC		
101M	R.D.SMITH	J.S.OLSON G.N.PHILLIPS JR	R.D.SMITH	MOL_ID: 1; 2 MOLEC		
102D	C.M.	NUNN S.NEIDLE	C.M.NUNN S.NEIDLE	MOL_ID: 1; 2 MOLECULE: DNA (		
102L	D.W.	HEINZ B.W.MATTHEWS	D.W.HEINZ W.A.BAASE F.W.DAHLQUIST B.W.MATTHE			
102M	R.D.	SMITH J.S	JR R.D.SMITH	MOL_ID: 1; 2 MOLEC		
103D	H.CH	OU L.ZHU B.R.REID	S.H.CHOU L.ZHU B.R.REID	MOL_ID: 1; 2 MOLEC		
103L	D.W.					
103M	R.D.					
104D	L.ZH					
104L	D.W.					
104M	R.D.					
105D	L.LE					
105M	R.D.					
106D	L.LE					
106M	R.D.					
107D	C.H.					
107L	M.BL					
107M	R.D.					
108D	H.P.					
108L	M.BL					
108M	R.D.					
109D	A.CZ					

- synonyms

TERM	SYNONYM
HEME-BINDING PROTEIN 1	P22HBP
EC 1.14.13.95	CYPVIIIB1
EC 1.14.13.95	CYTOCHROME P450 8B1
EC 1.14.13.95	STEROL 12-ALPHA- HYDROXYLASE
EC 1.14.13.95	7-ALPHA-HYDROXY-4-CHOLESTEN-3-ONE 12-ALPHA-HYDROXYLASE
EC 1.14.13.95	7-ALPHA- HYDROXYCHOLEST-4-EN-3-ONE 12-ALPHA-HYDROXYLASE
EC 1.14.13.95	CYP12
EC 1.14.13.95	
CADHERIN-2	
CADHERIN-2	
CADHERIN-2	
INTERLEUKIN-9 RECEPTOR PRE	
INTERLEUKIN-9 RECEPTOR PRE	
GRIFIN	
EC 1.13.99.1	

- configuration

```
- <node name="impl">
    <map />
  - <node name="QueryExecutionManagerImpl">
    - <map>
        <entry key="CONTAINS_THRESHOLD" value="30" />
        <entry key="POSITIONS_COMPUTE_THRES" value="50" />
        <entry key="MAX_SQL_RETRIEVE" value="10000" />
      </map>
    </node>
  - <node name="SynonymManagerImpl">
    - <map>
        <entry key="DEFAULT_MAX_HITS" value="5000" />
      </map>
    </node>
  - <node name="URLManagerImpl">
    - <map>
        <entry key="linkTemplate" value="http%3A%2F%2Fsrs.ebi.ac.uk%2Fsrsbin%2Fcgi-bin%2Fwgetz%3F-noSession%2B-e%2B%
          5B%7BSWALL_SP_REFSEQP_SP_PDB%7D-id%3AA7PS64_VITVI%5D" />
      </map>
    - <node isMap="true" name="databaseMaping">
      - <map>
          <entry key="pfamb" type="java.lang.String" value="SWALL_SP_REFSEQP_SP_PDB" />
          <entry key="refseq_protein_xp" type="java.lang.String" value="SWALL_SP_REFSEQP_SP_PDB" />
          <entry key="geneseq_prot" type="java.lang.String" value="SWALL_SP_REFSEQP_SP_PDB" />
          <entry key="sptrembl" type="java.lang.String" value="SWALL_SP_REFSEQP_SP_PDB" />
```

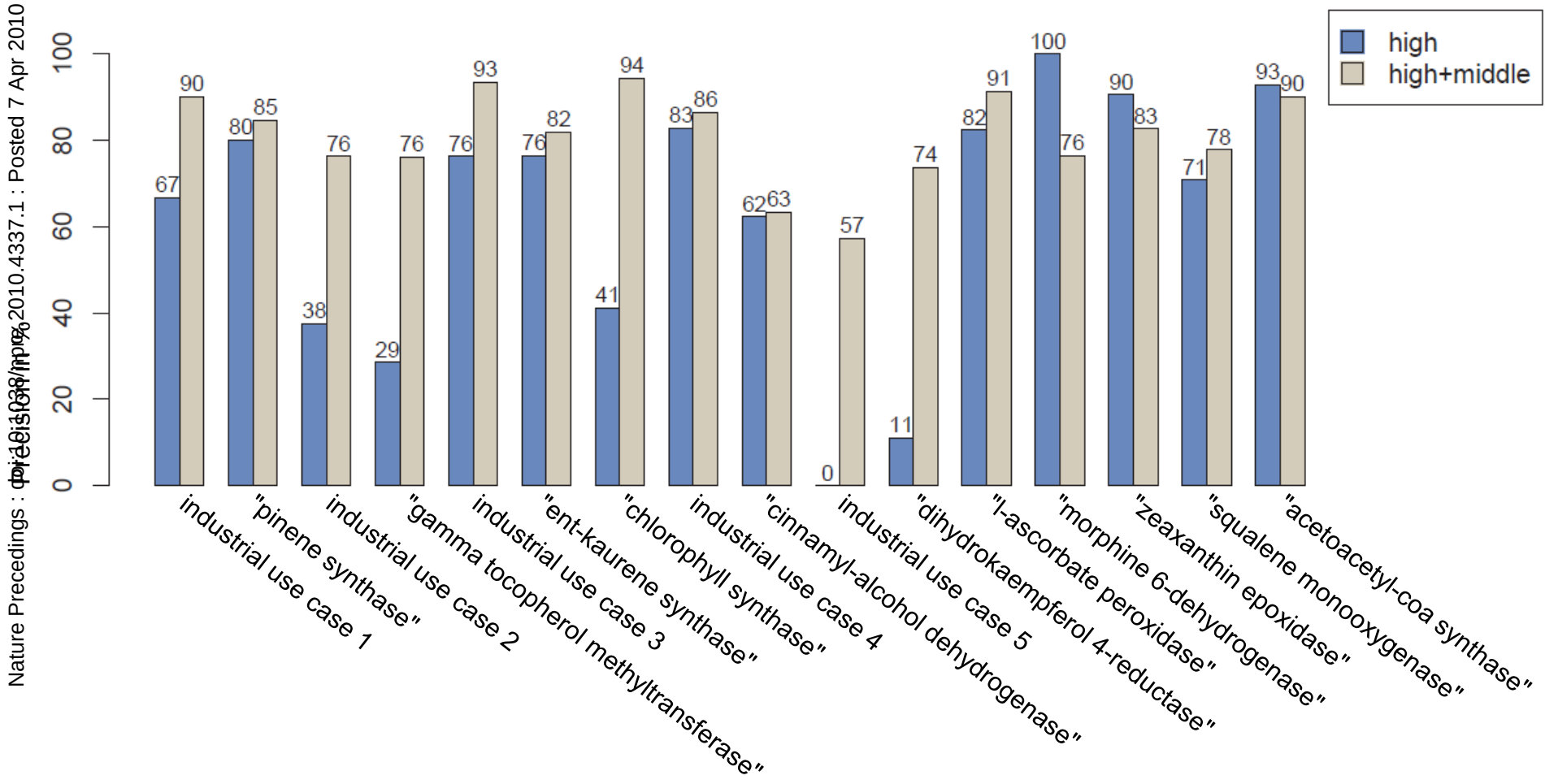
LAILAPS Benchmark

error type	LAILAPS benchmark semantics
true positive (TP)	the rank position is in the same window as in the reference set
false positive (FP)	the rank position is in a different window as in the reference set
false negative (FN)	the database entry was not found by LAILAPS

confidence class	precision	recall	F_1
"HIGH"	62%	100%	76
"MEDIUM" $\bigcup$ "HIGH"	81%	100%	90

$$Pr = \frac{TP}{TP + FP}; \quad Re = \frac{TP}{TP + FN}; \quad F_1 = \frac{2 * Pr * Re}{Pr + Re}$$

Benchmark of Neuronal Network Ranking



Conclusion

1. LAILAPS as search engine for life science dabases
2. „objective subjectiveness“ – learning of human relevance classes
3. cost-reduction of scientific information retrieval
4.

# database records	estimated effort in person day
- 50	0,5
51 – 100	1
101 - 250	> 1
> 250	objective rating hardly not possible

5. deterministic quality of database research

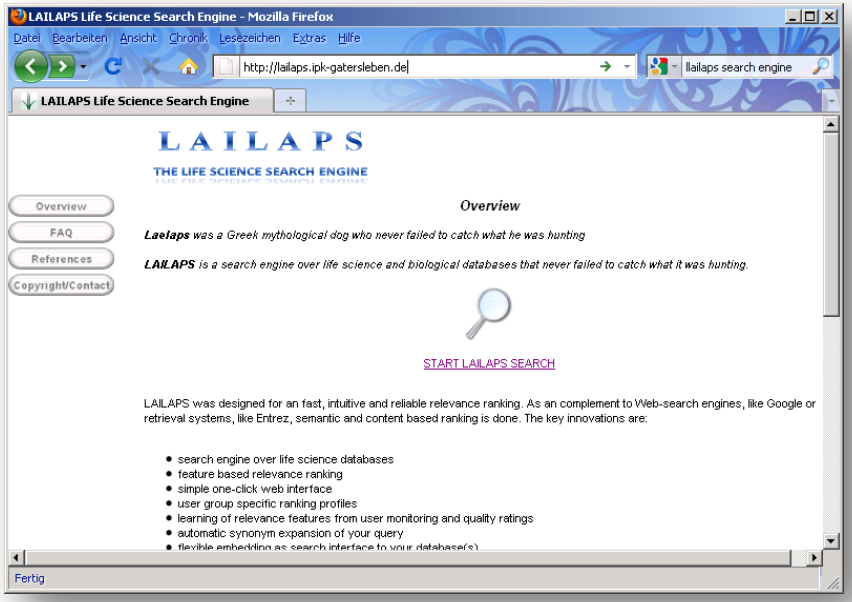
Thanks for your Attention

LAILAPS Project:

http://lailaps.ipk-gatersleben.de

Contact:

lange@ipk-gatersleben.de



Acknowledgement

Jinbo Chen (IPK Gatersleben)

Mandy Weißbach (IPK Gatersleben)

Gregor Haberhauer (BASF)

Michael Leps (BASF Plant Science)

Jens Stein (BASF Plant Science)

Röbbe Wünschiers (FH Mitweidae)