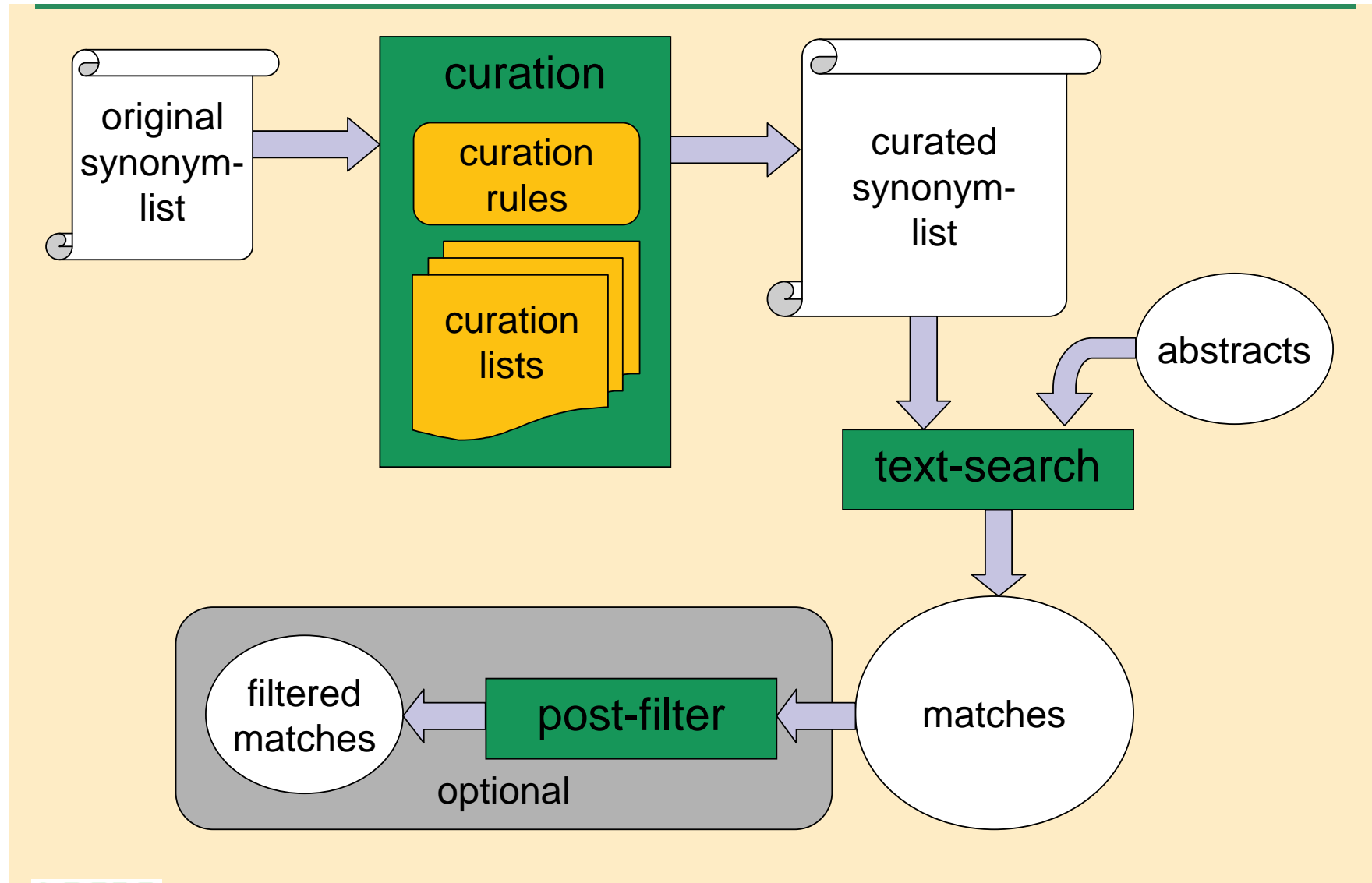


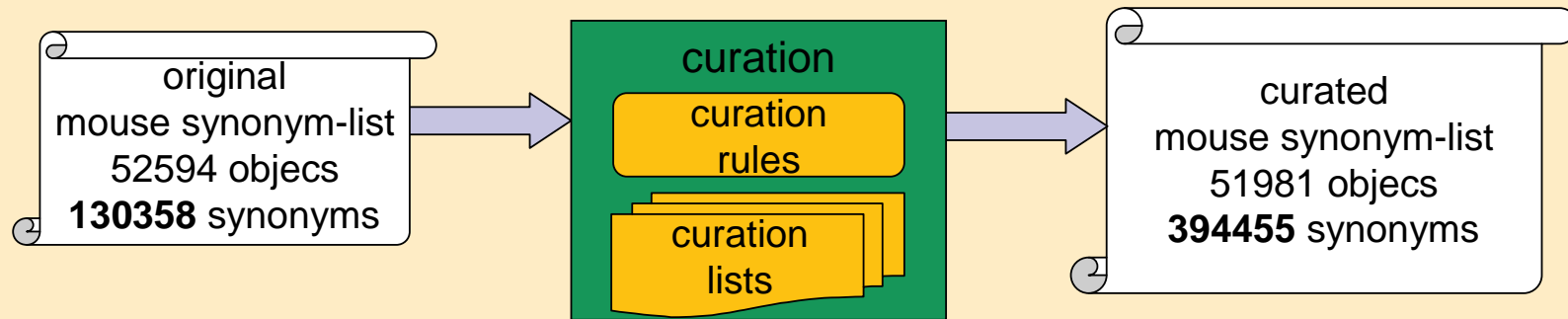
A simple approach for protein name identification

(Task 1b, user 24)

Approach



Curation – e.g. mouse



Addition of synonyms

- alternative subtype specifier:

a, b, c, ... ⇔ alpha, beta, gamma, ... (3698 synonyms)

1 ⇔ I (26168 synonyms)

- spelling variants:

Igf1 ⇔ Igf-1 ⇔ Igf 1 (60497 synonyms)

- removal of subtype specifier if unique:

mannose phosphate isomerase 1 ⇔ mannose phosphate isomerase

- expansion short names ⇔ long names:

IL ⇔ Interleukin (2801 synonyms)

Curation – e.g. mouse

Removal of unspecific and inappropriate synonyms

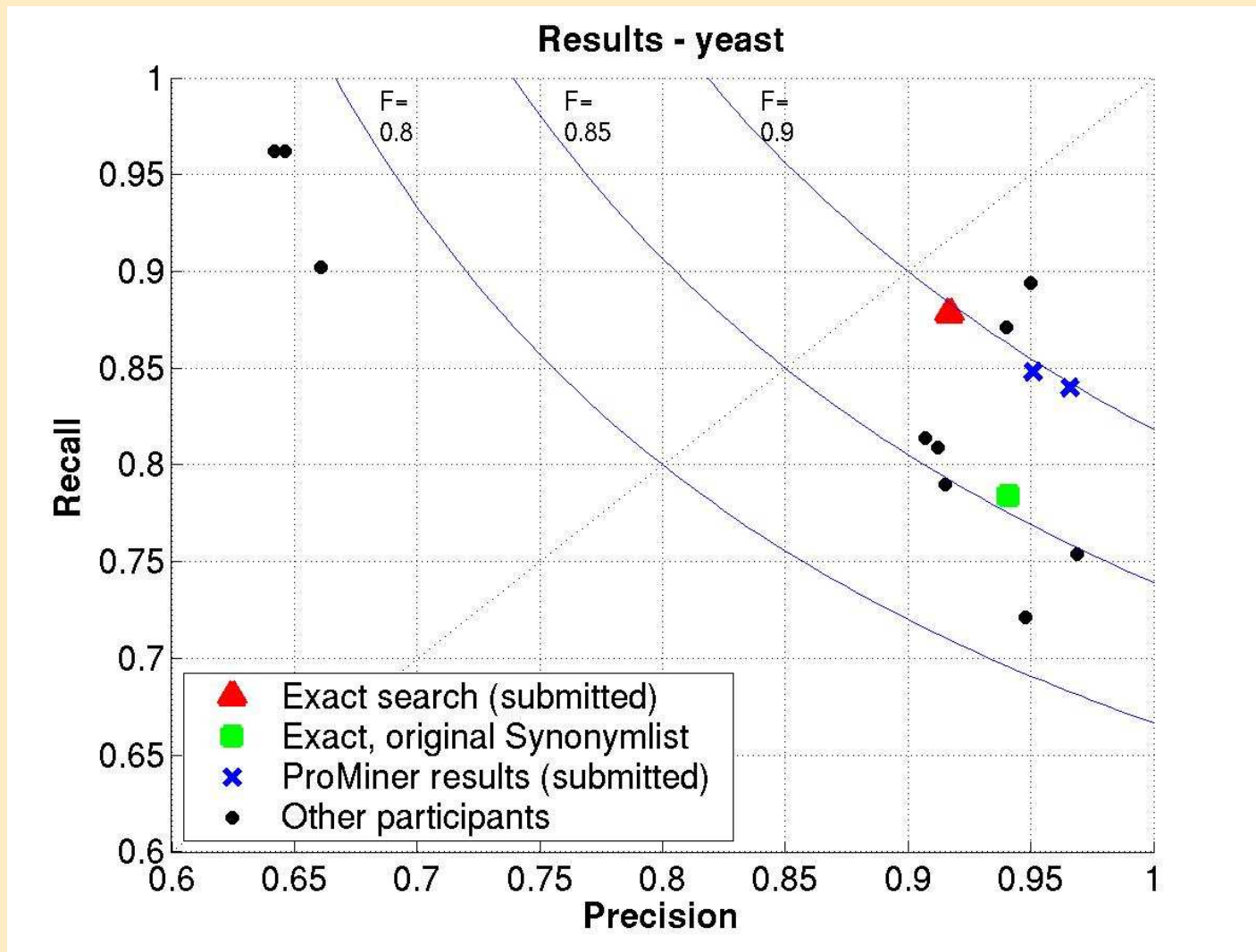
- length ≤ 2 (688 synonyms)
- composition solely of numbers and/or special characters (27 synonyms)
- common english words (258 synonyms)
- ambiguous synonyms (2641 synonyms)
- regular expressions based on token classes
 - token classes:
 - measuring units (kDa, Da, mg,...)
 - common words (if, and, as, at, for, of, ORF, EST...)
 - descriptions (tRNA, Ser, Tyr,...)
 - regular expressions:
 - common word + number (As 1, For 3, If 4, in 76) (161 synonyms)
 - number + measuring unit (35 kDa) (227 synonyms)
 - ...

Text-search

Text-search

- exact matching of synonyms in texts
- case insensitive if length of synonym is above a certain threshold or if synonym contains numbers
- case sensitive otherwise
- implemented in Perl
- ~750 lines of code
- runtime ~45 seconds for yeast training-set (5000 abstracts)

Results

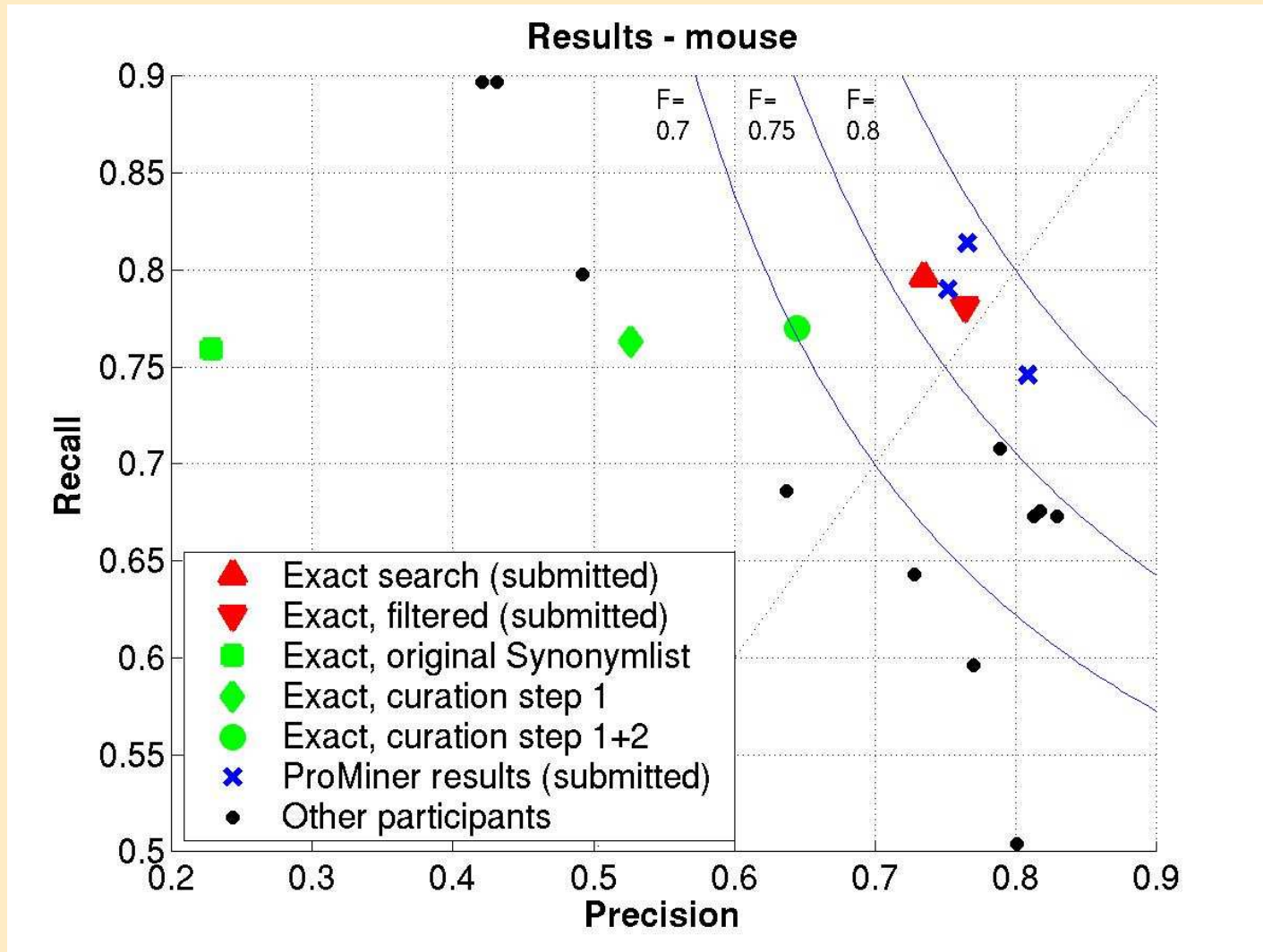


Rule-based Post-filter

Rule-based Post-filter

- applied on mouse results
- modifier appearing close to a putative match
 - cells, cell type, domain, DNA binding domain, ...
 - HEK cells, Sp1 binding sites
- short synonyms in parentheses
 - check of words right ahead of parentheses for significant overlap with long names of putative synonym
 - polymorphonuclear (PMN) infiltration ⇔ PMN ⇔ progressive motor neuropathy
 - diethylnitrosamine (DEN) ⇔ Den ⇔ denuded
- increases precision by 2.9%, decreases recall by 1.5%

Results



LiMB - Literature Mine Browser

LiMB – Literature Mine Browser

- Interactive text-mining
- Browsing through result-sets
- Manual curation of synonym-lists
- Testing of modifications of synonym-lists

The screenshot displays the LiMB Literature Mine Browser interface. The top navigation bar includes links for Home, Analysis, Upload, Query, myProfile, and Logout. The main content area is divided into a left sidebar and a main panel. The sidebar contains a 'JTM Database' section with options like 'Create new Database', 'Edit Database', and 'Browse Database', and a 'Synonym Eval' section. The main panel shows search results for 'Synonym(s)'. A search bar is present with a 'search' button. Below the search bar, there are four tables of results. The first table shows a synonym key 'L0000791' with 1 hit 'HML'. The second table shows 'L0000792' with 2 hits 'HMR'. The third table shows 'L0001112' with 4 hits: 'MIN1', 'PPMIN1', 'SCMIN1', and 'MIN1'. The fourth table shows 'L0001312' with 4 hits: 'ORIS', 'ORIS5P', 'PPORIS', and 'SCORIS'. A green circle highlights the 'Sentence' link in the navigation bar of the third table.

Synonym Key	Hit(s)	Synonym(s)
L0000791	1	HML
[Cooc in Abstract] [Cooc in Sentence] [Sentence] [Abstract]		

Synonym Key	Hit(s)	Synonym(s)
L0000792	2	HMR
[Cooc in Abstract] [Cooc in Sentence] [Sentence] [Abstract]		

Synonym Key	Hit(s)	Synonym(s)
L0001112	0	MIN1
	0	PPMIN1
	0	SCMIN1
	0	MIN1
[Cooc in Abstract] [Cooc in Sentence] [Sentence] [Abstract]		

Synonym Key	Hit(s)	Synonym(s)
L0001312	1	ORIS
	0	ORIS5P
	0	PPORIS
	0	SCORIS

LiMB - Literature Mine Browser

LiMB

Literature Mine Browser

Powered by **Struts**

Home Analysis Upload Query myProfile Logout

JTM Database

- Create new Database
- Edit Database
- Browse Database
 - Browse
 - Use Filter
 - Create Filter
 - Manage Filter
- Analyse Database

Synonym Eval

Sentence(s)

[Root]

[prev] [next]

Synonym Key	Hit(s)	Synonym(s)
S0003490	0	SCRAD2
	0	DNA REPAIR PROTEIN RAD2
	0	PPRAD2
	1	RAD2P
	0	RAD2
	0	YGR-258C
	0	YGR258C

[Cooc in Abstract] [Cooc in Sentence] [Sentence] [Abstract]

Occurs in:

PMID	Sentence
yeast_00018_testing; 0	Exo1p is a member of the Rad2p family of structure-specific nucleases that contain conserved N and I nuclease domains.
yeast_00018_testing; 4	Our study indicates that Exo1p shares similar, but not identical structure-function relationships to other characterized members of the Rad2p family in the N and I nuclease domains.

Copyright © 2003 - Daniel Guettler

SVM-based Post-filter

Fly

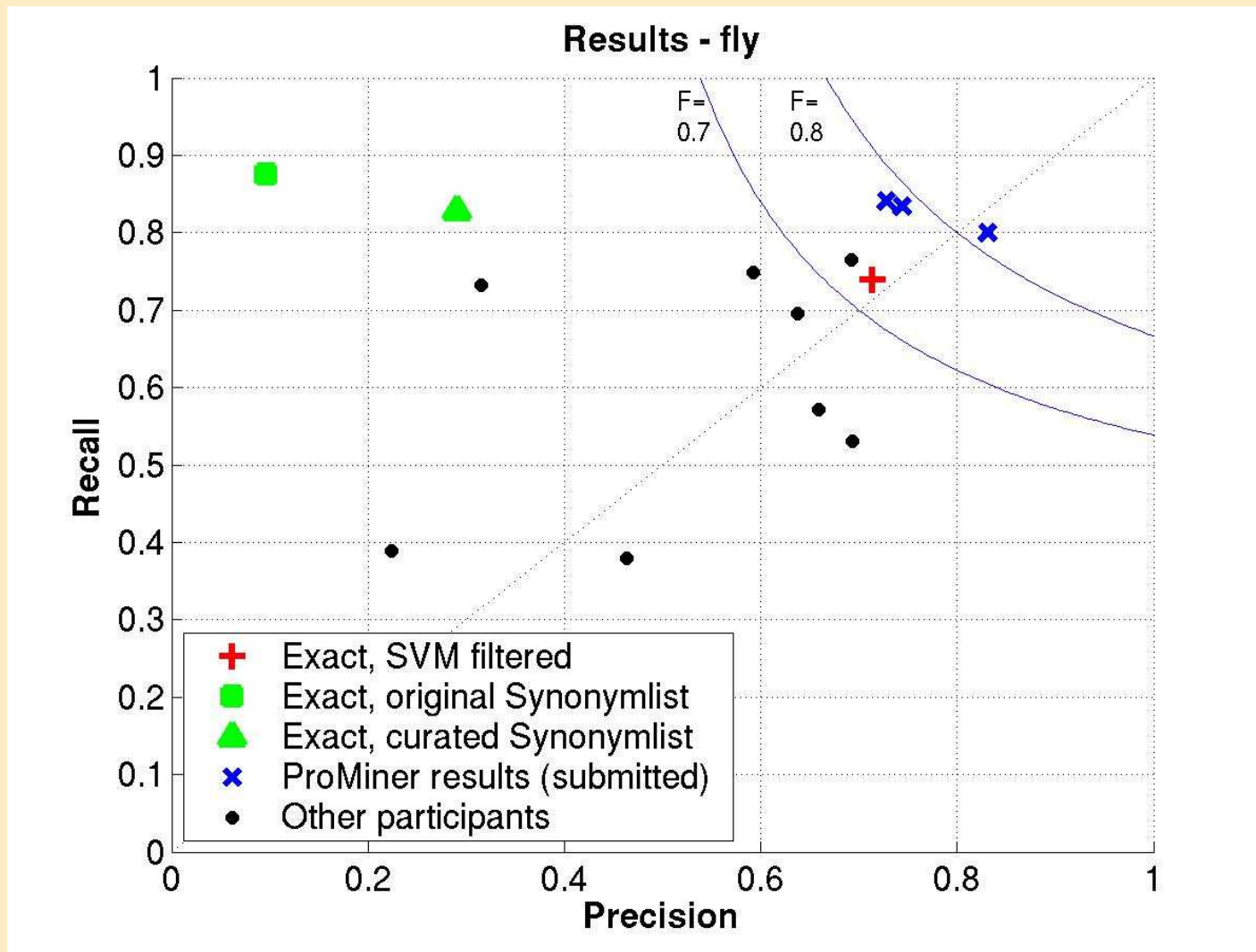
- synonyms show significant overlap with common english words
- extensive post-filtering is crucial
- our approach was not ready for achieving satisfying results in BioCreative

Support Vector Machine (SVM)-based Post-filter

Features:

- surface keys of synonym
 - length, numbers, capitals,...
- part of speech tags of synonym and adjacent words
- scores for proximate nouns/verbs
 - scores determined from fly trainings set
- scores for adjacent words
 - scores determined from a search of mouse list against ~700000 medline abstracts

fly - Post-evaluation



Conclusions

Conclusions

- exact text matching is appropriate for protein name identification
- extensive curation of synonym list is important
- depending on the characteristics of the synonyms post-filtering can be important
- the Literature Mine Browser is useful for fast curation of synonym lists
- SVMs proved to be effective for post-filtering synonym occurrences

Acknowledgement

- Daniel Güttler
- Joannis Apostolakis
- Daniel Hanisch
- Juliane Fluck
- Ralf Zimmer