A Reproducibility Study of Information Retrieval Models

Peilin Yang and Hui Fang
University of Delaware

Have you used these baselines?

- Okapi BM25
- Pivoted Document Length Normalization
- Dirichlet Language Model
- Divergence from Randomness Models (PL2)

More than 20 models proposed in SIGIR/CIKM papers have used these models as baselines…

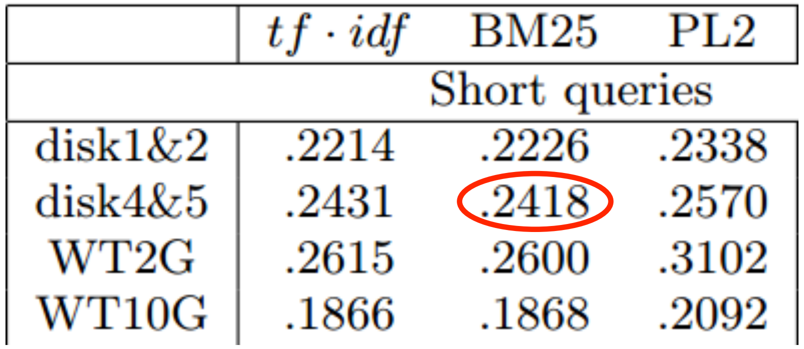
Steps of proposing an IR model…

1. Provide theoretical foundation
2. Implement the model/algorithm
3. Run, Evaluate against collections
4. Compare with other models (significant test)
5. Claim the advantage (yeah…)

The procedure is quite reasonable
However, problems exist in real world…

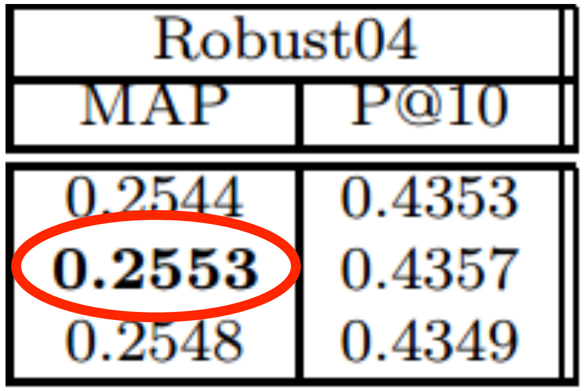
1. Implementation Variations

Screenshots from previous papers



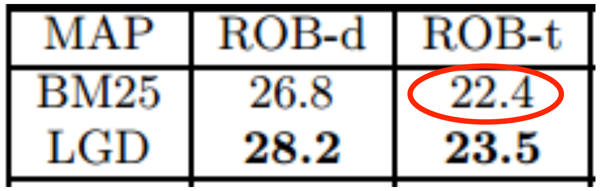
	tf · idf	BM25	PL2
	Short queries		
disk1&2	.2214	.2226	.2338
disk4&5	.2431	.2418	.2570
WT2G	.2615	.2600	.3102
WT10G	.1866	.1868	.2092

B. He and I. Ounis. SIGIR '05



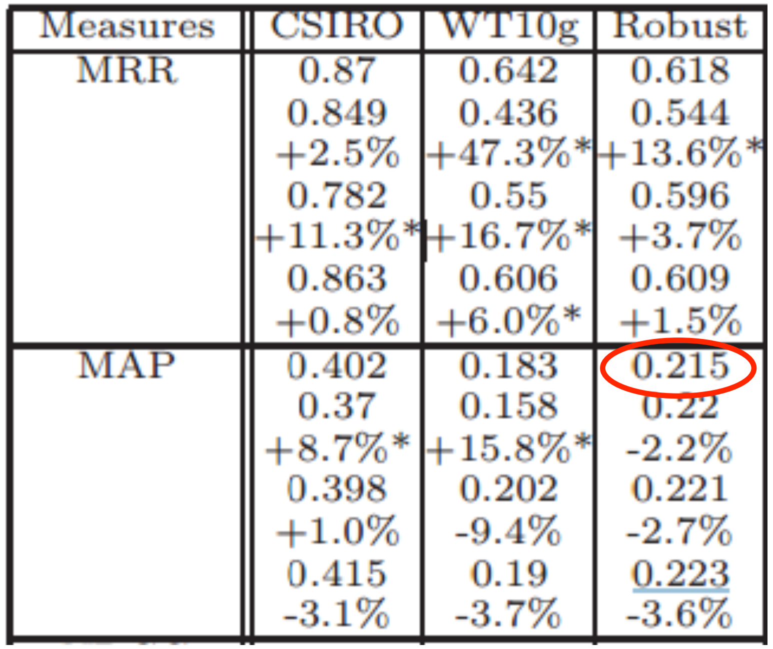
Robust04	
MAP	P@10
0.2544	0.4353
0.2553	0.4357
0.2548	0.4349

Y. Lv and C. Zhai. CIKM '11



MAP	ROB-d	ROB-t
BM25	26.8	22.4
LGD	**28.2**	**23.5**

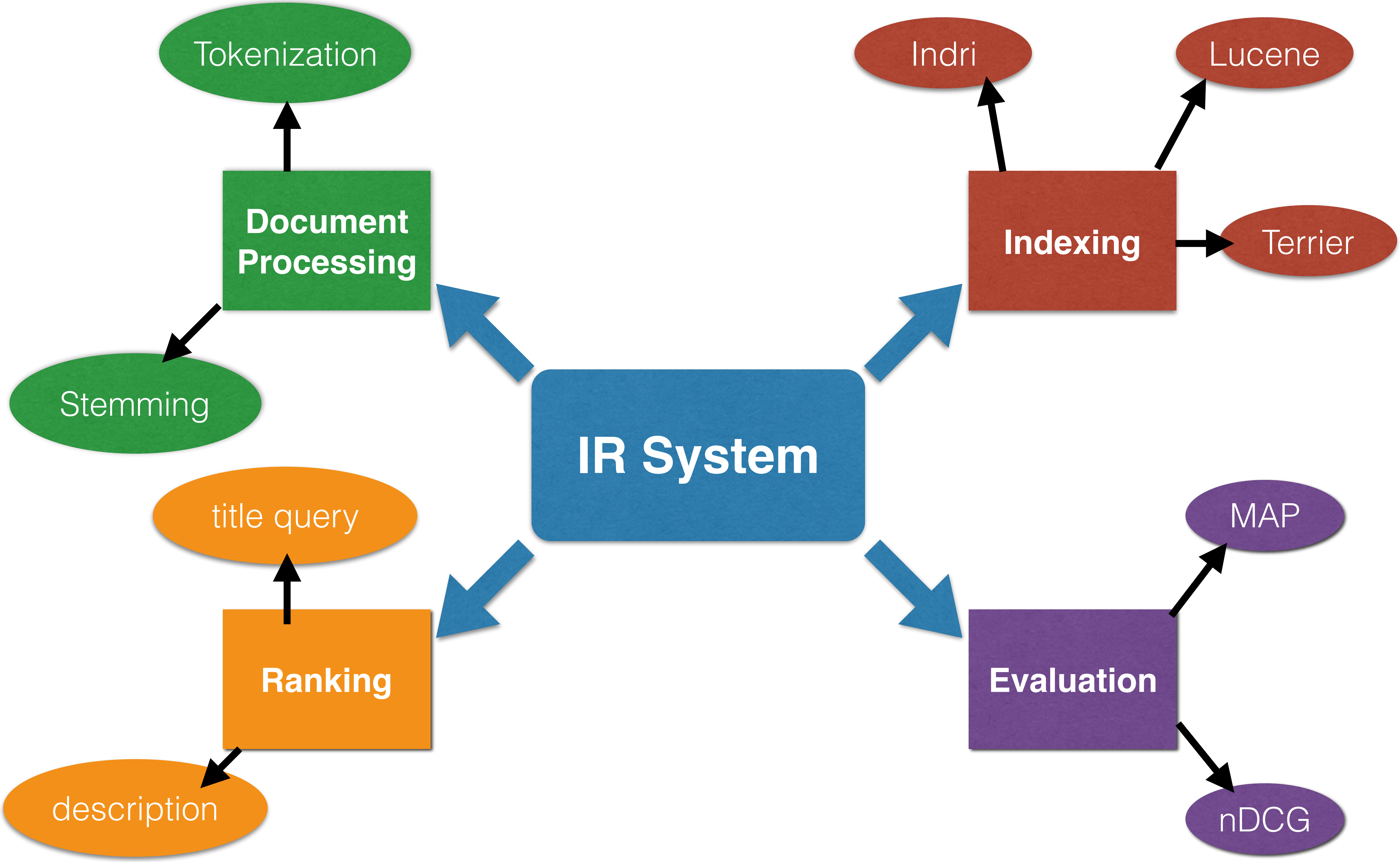
S. Clinchant and E. Gaussier. SIGIR '10



Measures	CSIRO	WT10g	Robust
MRR	0.87	0.642	0.618
	0.849	0.436	0.544
	+2.5%	+47.3%*	+13.6%*
	0.782	0.55	0.596
	+11.3%*	+16.7%*	+3.7%
	0.863	0.606	0.609
	+0.8%	+6.0%*	+1.5%
MAP	0.402	0.183	0.215
	0.37	0.158	0.22
	+8.7%*	+15.8%*	-2.2%
	0.398	0.202	0.221
	+1.0%	-9.4%	-2.7%
	0.415	0.19	0.223
	-3.1%	-3.7%	-3.6%

J. Zhu, J. Wang, I. J. Cox, and M. J. Taylor. SIGIR '09

Different researchers have different results for the same model!!!

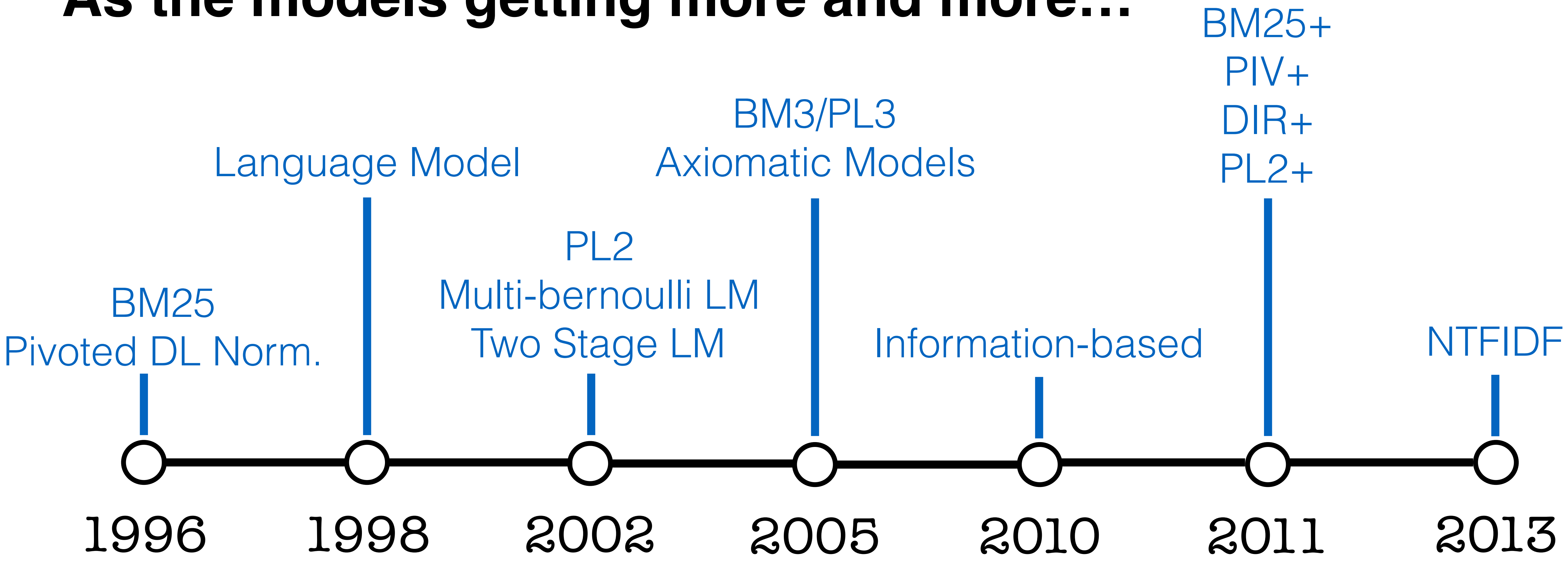


Options of the components are considered as the reason

2. Lack of Comprehensive Comparison

Only **4** models (out of 20 models we mentioned before) have used the baselines other than the popular ones.. and they are tf.idf and JM Language Model.

As the models getting more and more…



BM25
Pivoted DL Norm.
1996

Language Model
1998

PL2
Multi-bernoulli LM
Two Stage LM
2002

BM3/PL3
Axiomatic Models
2005

Information-based
2010

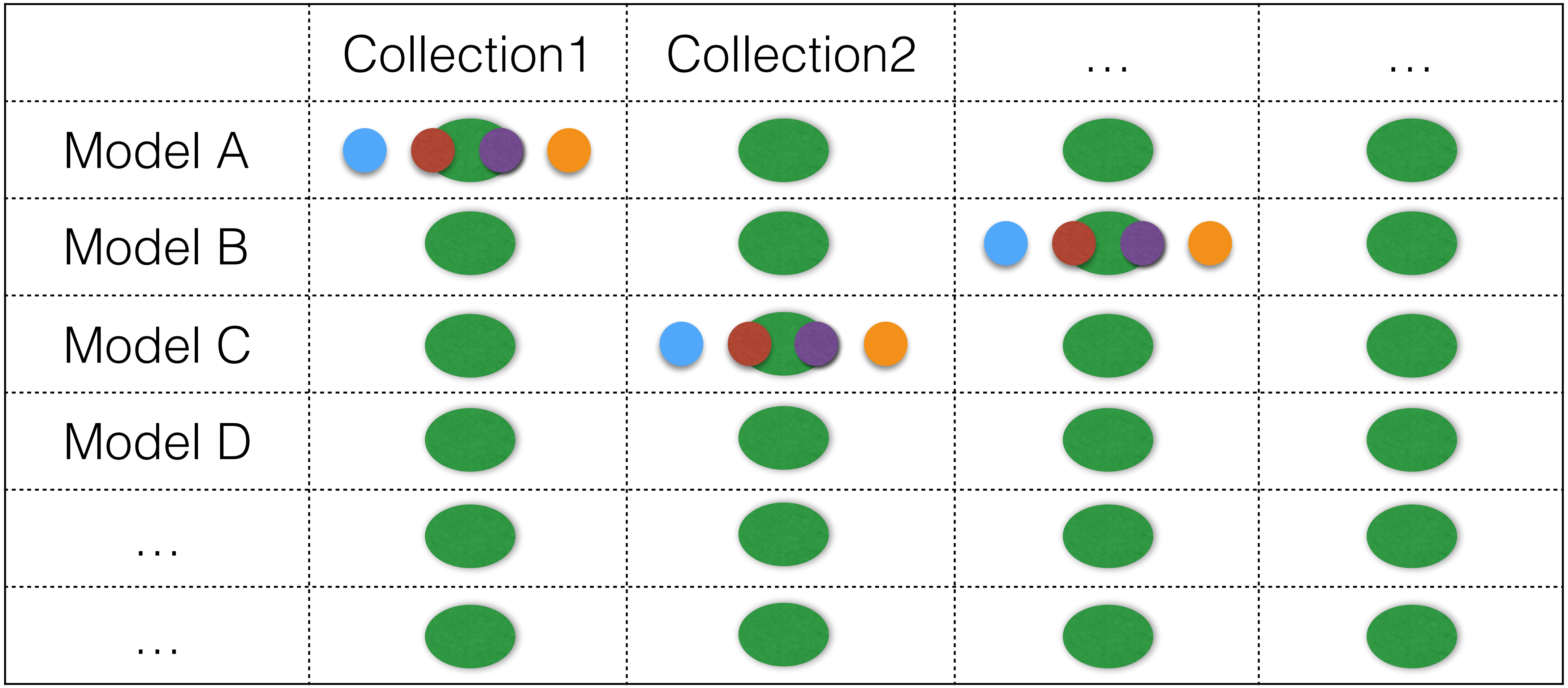
BM25+
PIV+
DIR+
PL2+
2011

NTFIDF
2013

It is harder and harder to re-implement all existing models but they should be included in the comparison

Questions:

1. What would be the performances of existing models if tested using a normalized environment?
2. How do existing models perform against the collections that were not reported?



	Collection1	Collection2	…	…
Model A	🔵 🔴 🟢 🟣 🟠	🟢	🟢	🟢
Model B	🟢	🟢	🔵 🔴 🟢 🟣 🟠	🟢
Model C	🟢	🔵 🔴 🟢 🟣 🟠	🟢	🟢
Model D	🟢	🟢	🟢	🟢
…	🟢	🟢	🟢	🟢
…	🟢	🟢	🟢	🟢

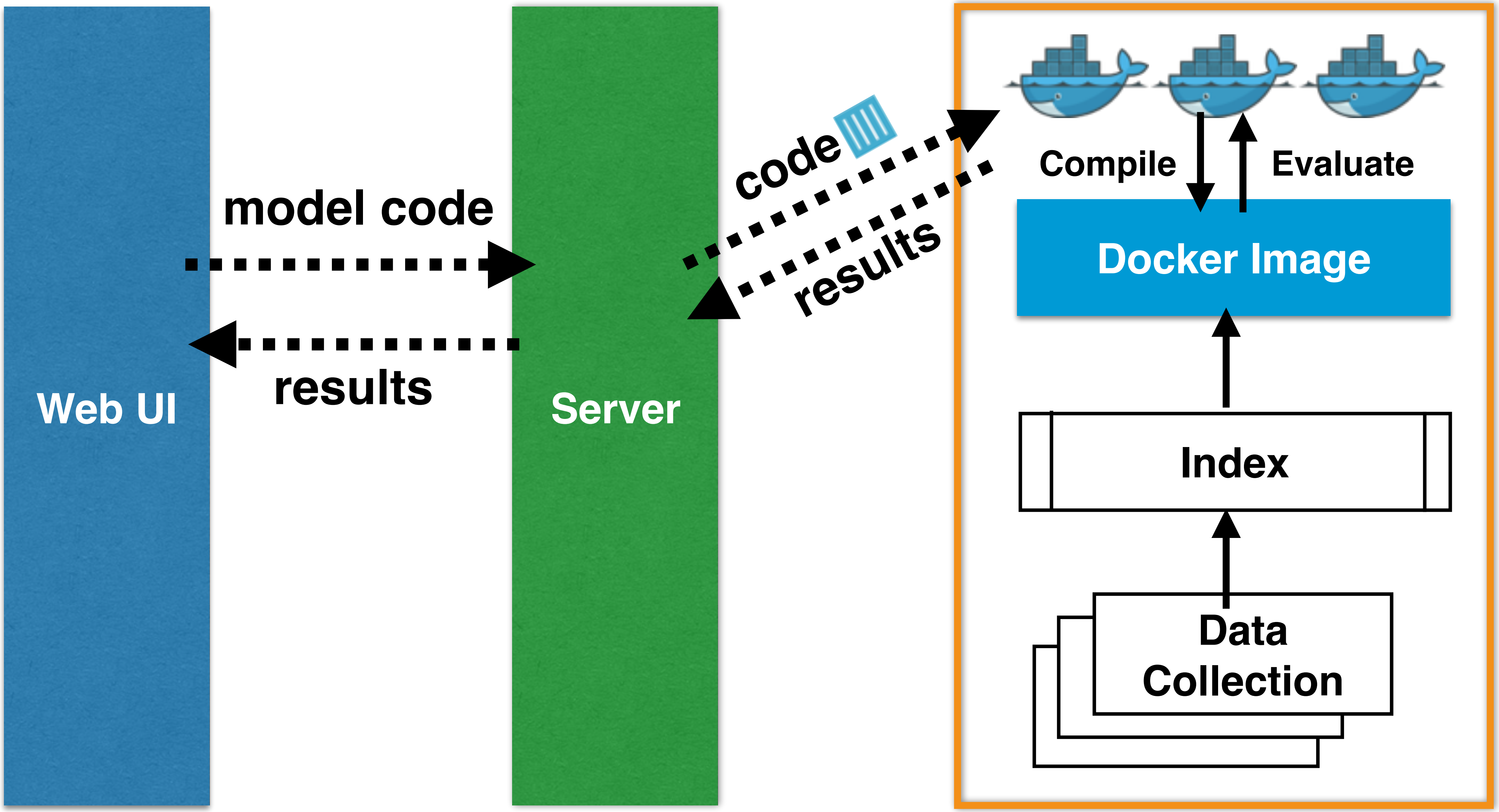
Previous Studies

- Privacy Preserved Evaluation (PPE) [Fang&Zhai SIGIR2014Workshop]
  - VIRLab
  - cooperation is not possible
- Evaluation as a Service (EaaS) [Rao&Lin ECIR2015, Lin ECIR2016]
  - Microblog domain
  - no web interface

Web-based
Reproducible Information retrieval
System Evaluation (RISE)

1. A unified environment for evaluating models
2. Easy management/cooperation of models

Web-based Reproducible Information retrieval System Evaluation (RISE)



RISE enables us to make:

- Reproducible study of published papers
- Comprehensive comparisons

REPRODUCIBILITY STUDY

Reproduced Models

BM25 Family	Pivoted Normalization Family	Language Models	Divergence from Randomness Models	Information-based Models
BM25	PIV	DIR	PL2	SPL
F2EXP	F1EXP	BLM	PL3	LGD
F2LOG	F1LOG	TSL	PL2+	
BM3	PIV+	F3EXP		
BM25+	NTFIDF	F3LOG		
		DIR+		

Collections and Queries

Collections	Topics	# of Documents	Average Document Length
TREC1 TREC2 TREC3	51-100 101-150 151-200	741,856	412.89
TREC6 TREC7 TREC8 ROBUST04	301-350 351-400 401-450 601-700	528,155	467.55
WT2G	401-450	247,491	1057.99
Terabyte04 Terabyte05 Terabyte06	701-750 751-800 801-850	25,205,179	937.25

Experimental Settings

Tools

Modified Indri-5.9

Queries

As reported in the original papers

Pre-processing of the collections

NO stop words removal

Porter Stemmer

Evaluation Method

MAP (using trec_eval)

Reproducibility Results

BM25 Family

Models	Mean	Std.
BM25	-2.08%	4.11%
F2EXP	+0.68%	2.18%
F2LOG	+0.22%	1.63%
BM3	-5.92%	0.74%
BM25+	-0.67%	1.19%

Pivoted Norm. Family

Models	Mean	Std.
PIV	-3.64%	4.67%
F1EXP	-6.62%	2.23%
F1LOG	-7.76%	2.79%
PIV+	-0.94%	2.31%
NTFIDF	-17.08%	4.71%

Language Models

Models	Mean	Std.
DIR	+1.03%	3.26%
TSL	+4.09%	6.18%
F3EXP	-2.65%	2.72%
F3LOG	-4.11%	3.74%
DIR+	-0.20%	0.20%

Divergence from Randomness

Models	Mean	Std.
PL2	+5.54%	16.19%
PL3	+0.59%	2.41%
PL2+	+0.35%	0.04%

Information-based Models

Models	Mean	Std.
SPL	-4.60%	3.42%
LGD	-2.04%	2.45%

- Within 5% for most Mean and Std.
- PL2 and NTFIDF have much larger Mean/Std.

Reproducibility Results

Details of PL2 and NTFIDF

Models	Collection	original	reproduced	DIFF
PL2	TREC1	0.207	0.257	+24.46%
	TREC2	0.238	0.285	+19.60%
	TREC3	0.271	0.327	+20.89%
	TREC6	0.257	0.233	-9.30%
	TREC7	0.221	0.196	-11.39%
	TREC8	0.256	0.228	-11.01%
NTFIDF	TREC678	0.234	0.209	-10.64%
	ROBUST04	0.302	0.245	-18.84%
	GOV2	0.317	0.248	-21.77%

- PL2 has different performances over collections
- NTFIDF is always worse
- Different tools might be the reason

COMPREHENSIVE COMPARISONS

Experimental Settings

Collections

Clueweb added

Queries

Title ONLY

Evaluation Method

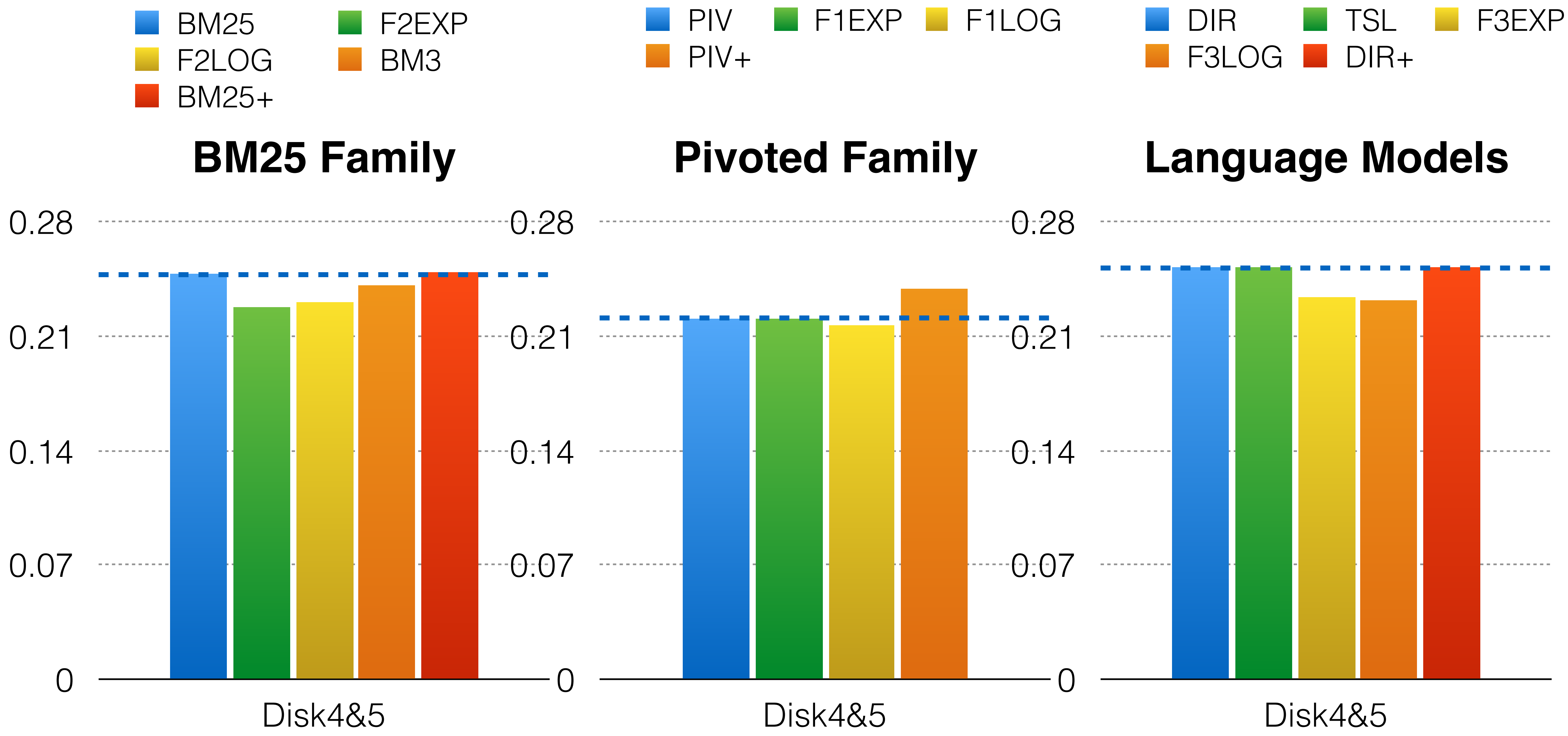
MAP & ERR

BM25 Family	Pivoted Normalization Family	Language Models	Divergence from Randomness Models	Information-based Models
BM25	PIV	DIR	PL2	SPL
F2EXP	F1EXP	BLM	PL3	LGD
F2LOG	F1LOG	TSL	PL2+	
BM3	PIV+	F3EXP		
BM25+	NTFIDF	F3LOG		
		DIR+		

🟩 Basic Models

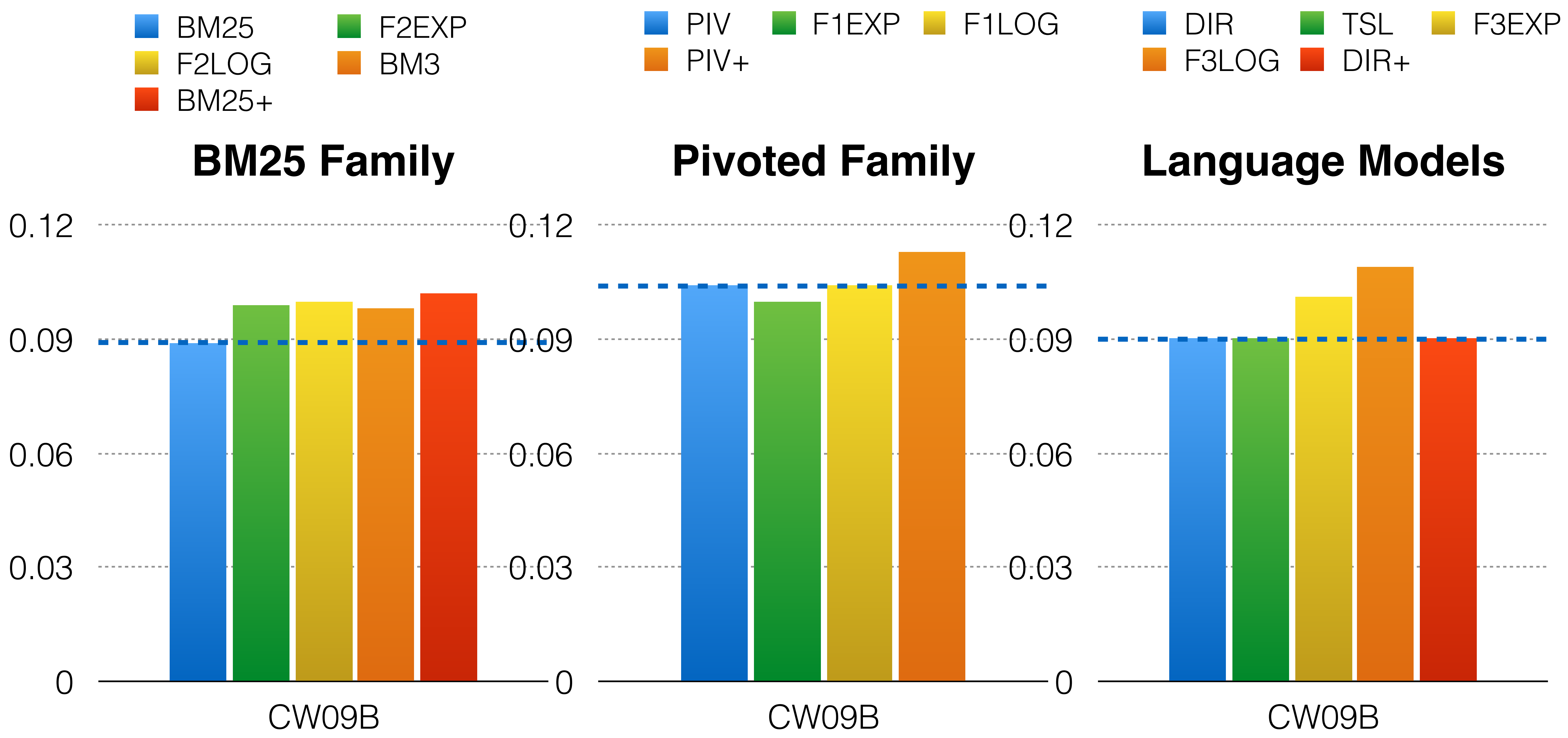
Variations are always better?

Disk4&5



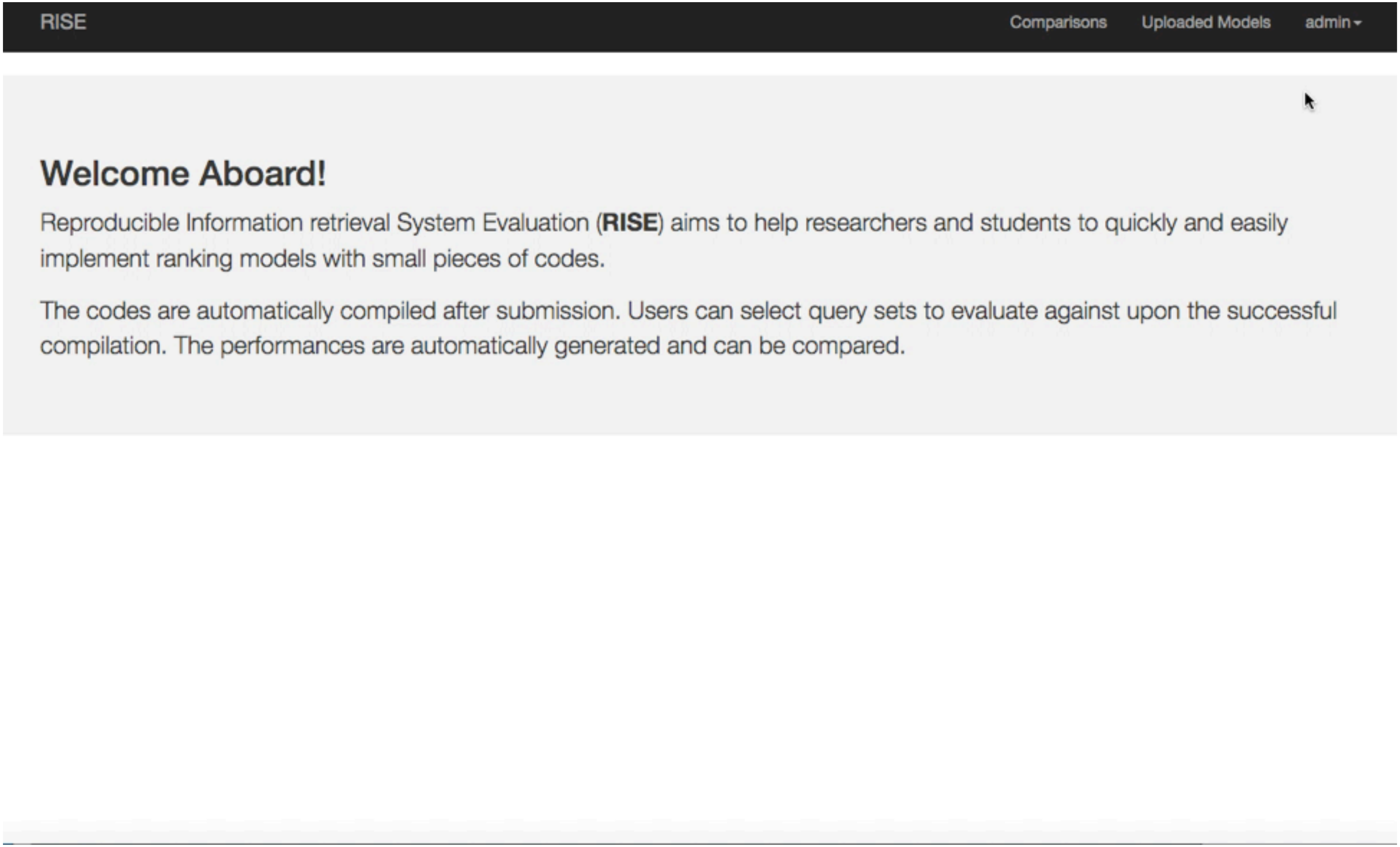
The variations are not necessarily better than the basic models

ClueWeb



- The variations are basically better except Pivoted family
- Optimal performances from families are comparable
- Please refer to our paper for more detailed results!

Demo



Welcome Aboard!

Reproducible Information retrieval System Evaluation (**RISE**) aims to help researchers and students to quickly and easily implement ranking models with small pieces of codes.

The codes are automatically compiled after submission. Users can select query sets to evaluate against upon the successful compilation. The performances are automatically generated and can be compared.

Open Sourced

- **RISE (system)**
  - http://rires.info:8080/
- **Web Server (code)**
  - https://github.com/Peilin-Yang/reproducibleIR
- **Docker (code)**
  - https://github.com/Peilin-Yang/RIRES_EVAL

Future Work

- More stats to RISE
- Split collections to training - testing
- Learning to Rank
- Parameter Tuning
- Listen to the community

Thank You!
Q & A