Predicting the future relevance of research institutions - The winning solution to the KDD Cup 2016

Vlad Sandulescu

Adform

Denmark

Mihai Chiru

Bitdevelop

Sweden

2016.08.14 - KDD 2016, San Francisco, USA

Phase 1

Start simple

#1: **build a dataset**

#2: explore the dataset

Dataset

- only used MAG
- for all the authors of the full research papers between 2011-2015, get all their papers starting with 2000
- for all these papers, get all related info (references, keywords, fields of study, etc.)

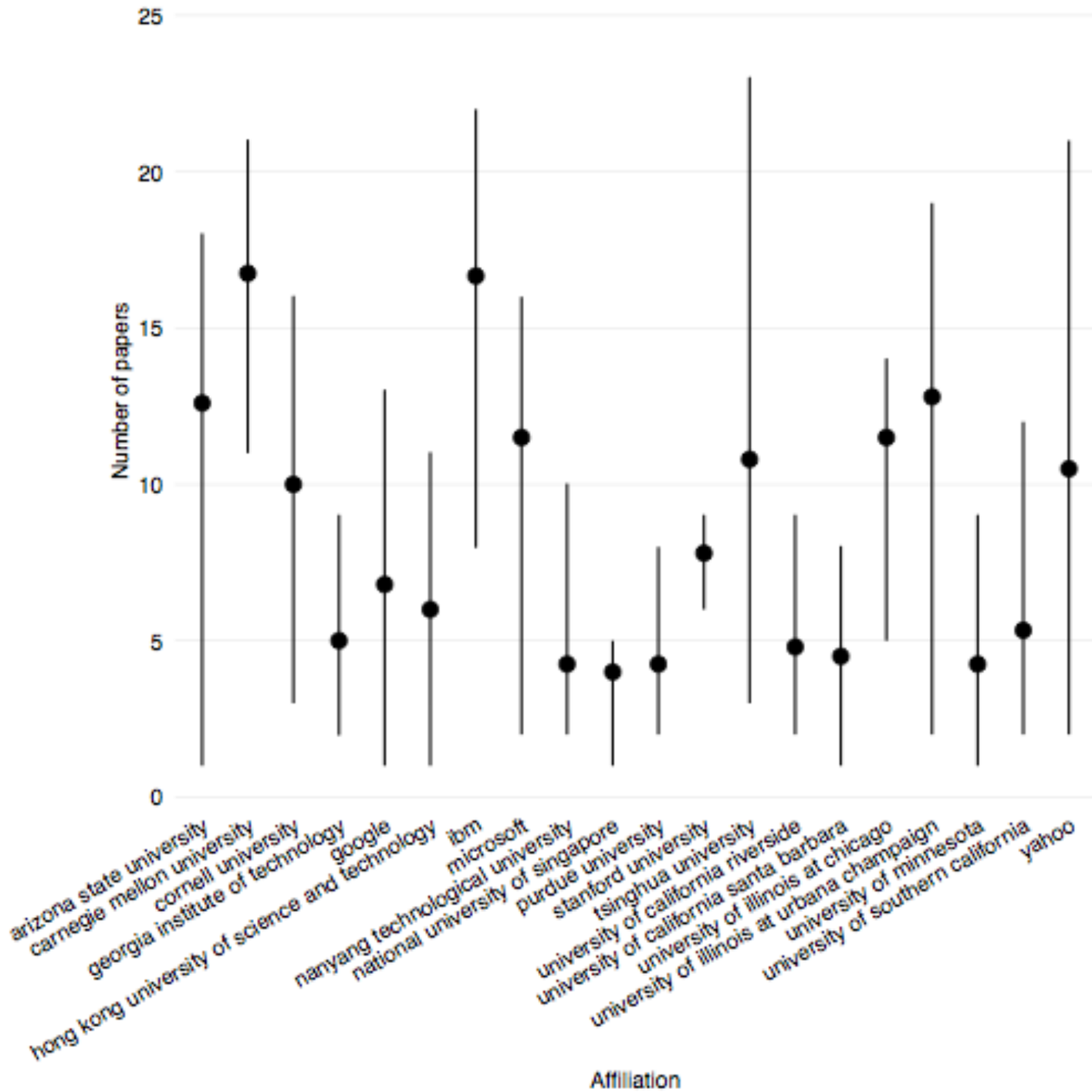

Figure: Full range and mean value of the number of accepted full research papers for top 20 affiliations at KDD between 2011 and 2015

Phase 1

Start simple

#1: build a dataset

#2: **explore the dataset**

Dataset

- only used MAG
- for all the authors of the full research papers between 2011-2015, get all their papers starting with 2000
- for all these papers, get all related info (references, keywords, fields of study, etc.)

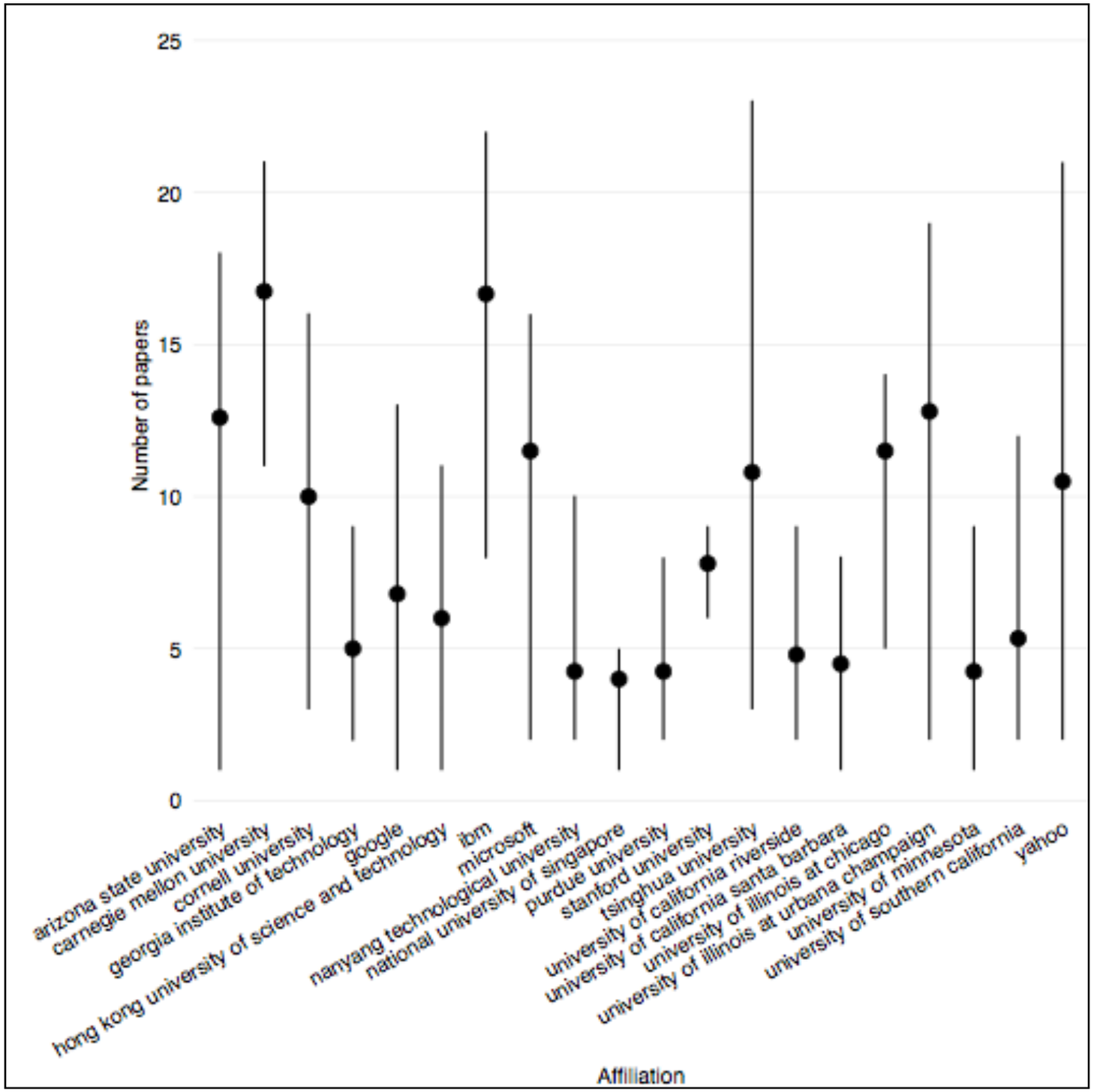

Figure: Full range and mean value of the number of accepted full research papers for top 20 affiliations at KDD between 2011 and 2015

Phase 1

Start simple

#1: build a dataset

#2: explore the dataset

#3: set up a baseline

Baseline model:
Count the number of full research papers an affiliation published at each conference in 2011-2015
Compute the probability of a full research paper is published by an affiliation

Conference	2015	2014	2013
SIGIR	0.95	0.94	0.89
SIGMOD	0.87	0.94	0.82
SIGCOMM	0.93	0.95	0.77

Table: NDCG@20 results for the probabilities model in phase 1 for 2013, 2014 and 2015

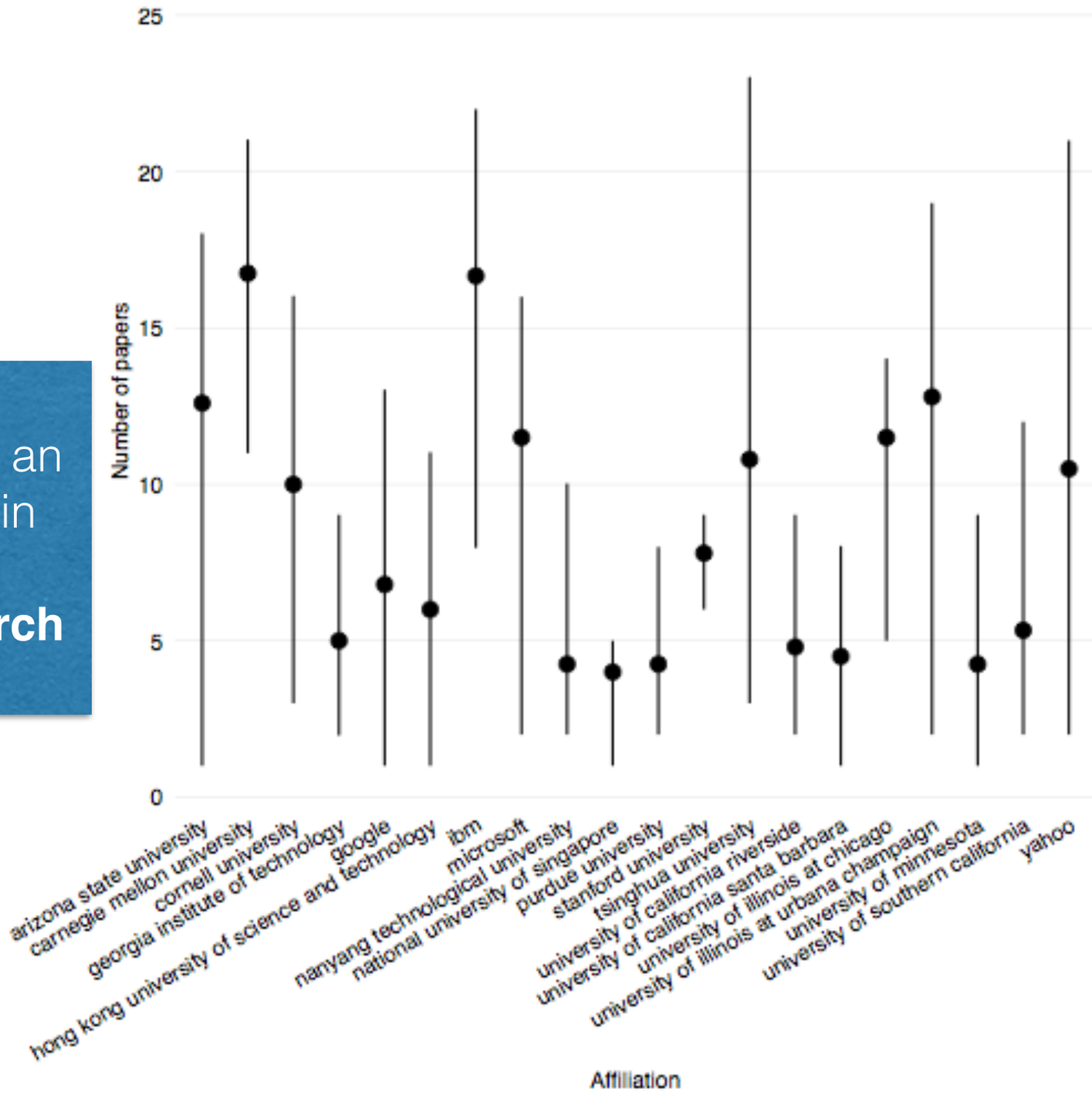

Figure: Full range and mean value of the number of accepted full research papers for top 20 affiliations at KDD between 2011 and 2015

Phase 2

Start improving the baseline

#1: **predict relevance directly**

#2: explore the dataset even more

#3: try GBDT & Mixed models

- The evaluation metric directly uses the relevance

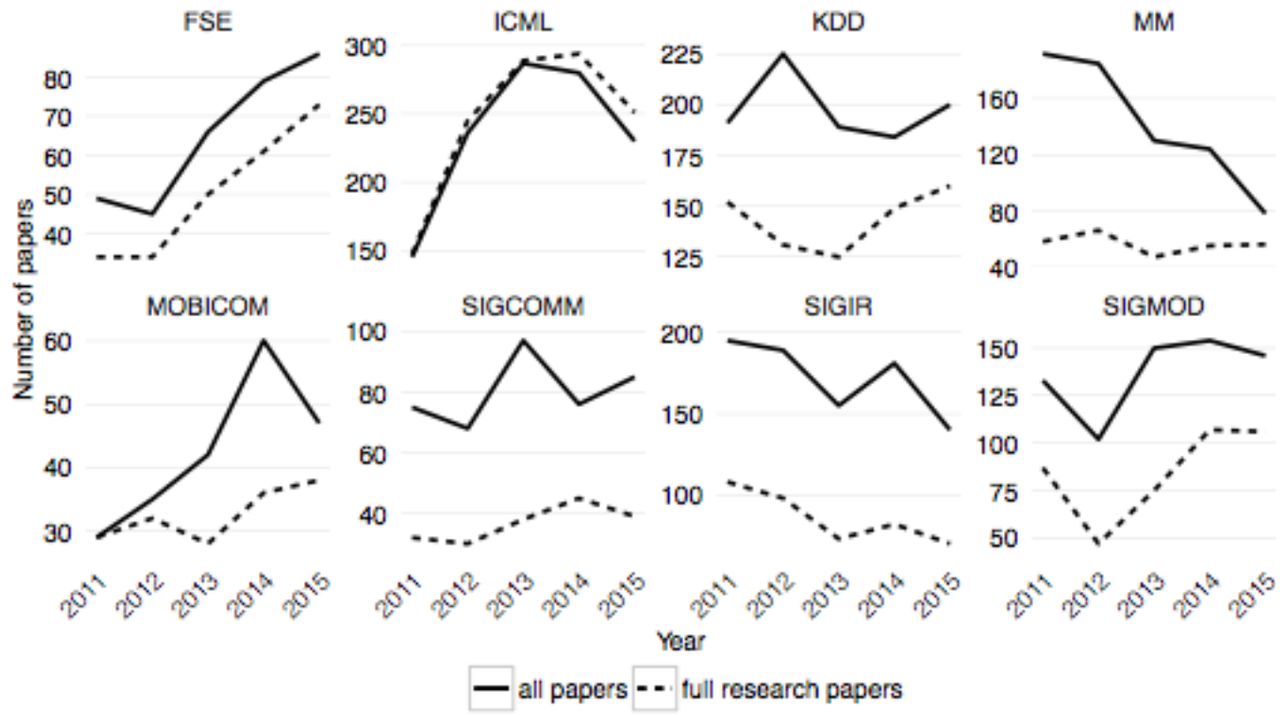

Figure: Number of all the papers vs the full research papers for all the conferences in the competition for 2011-2015

Phase 2

Start improving the baseline

#1: predict relevance directly

#2: **explore the dataset even more**

#3: **try GBDT & Mixed models**

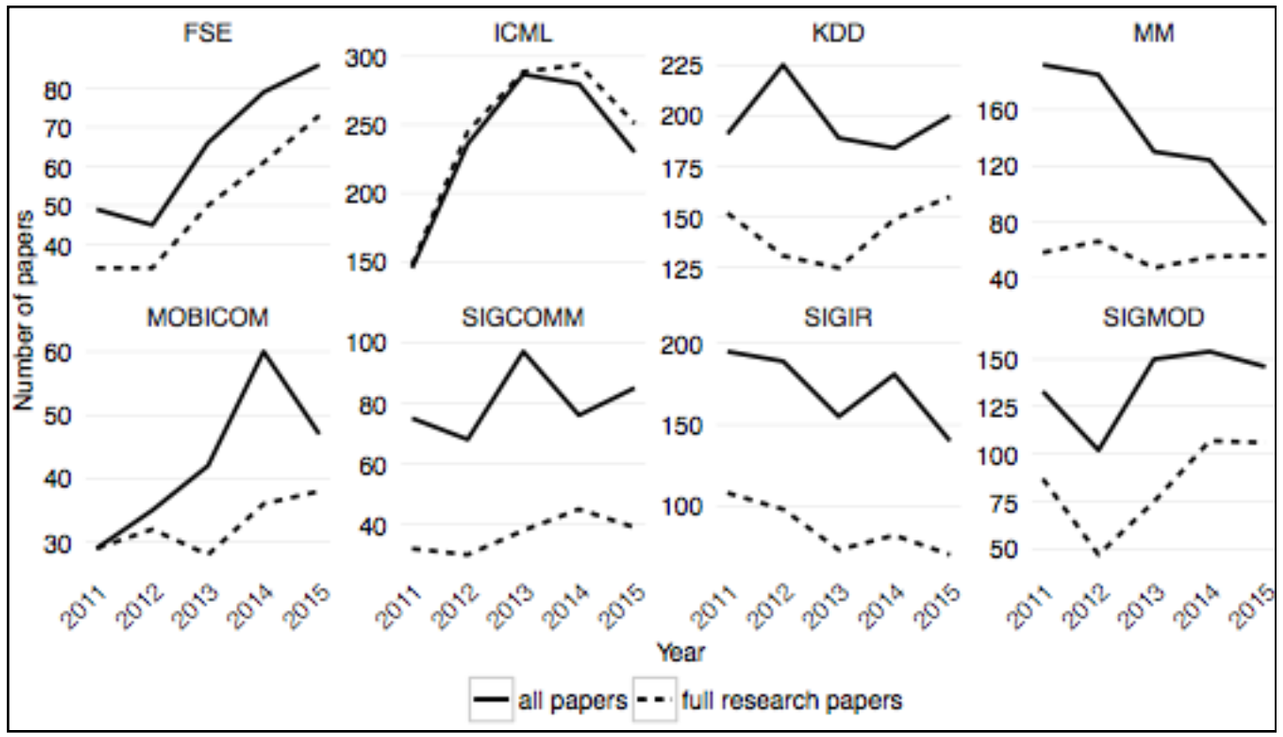

Figure: Number of all the papers vs the full research papers for all the conferences in the competition for 2011-2015

- The evaluation metric directly uses the relevance

Phase 2

Start improving the baseline

#1: predict relevance directly

#2: explore the dataset even more

#3: try GBDT & Mixed models

#4: **expand the dataset**

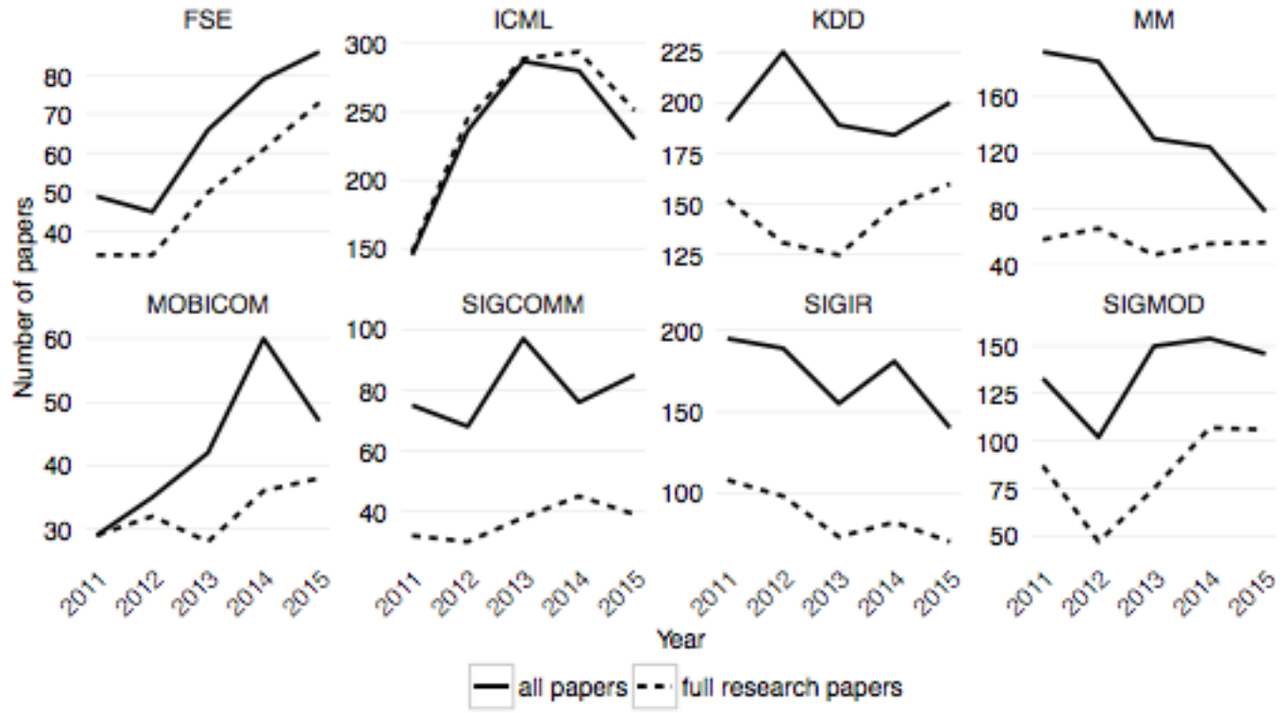

Figure: Number of all the papers vs the full research papers for all the conferences in the competition for 2011-2015

- The evaluation metric directly uses the relevance

	# samples
Full research papers	3,677
Phase1: probabilities	1,296
Phase 2: full research papers	**8,605**
Phase 2: all papers	**10,900**

Table: Dataset evolution between phases

Conference x Affiliation x Year x [features]

Phase 2

Start improving the baseline

#1: predict relevance directly

#2: explore the dataset even more

#3: try GBDT & Mixed models

#4: expand the dataset

#5: engineer features

- The evaluation metric directly uses the relevance

	# samples
Full research papers	3,677
Phase1: probabilities	1,296
Phase 2: full research papers	**8,605**
Phase 2: all papers	**10,900**

Table: Dataset evolution between phases

Features

- Stats of all previous relevance scores (std, sum, mean, median, min, max)
- Previous relevance scores in windows from previous year up to 4 years ago
- Weighted moving-average of previous relevance scores in windows from previous year up to 4 years ago
- Stats of AIF metrics (std, sum, mean, median, min, max)

GBDT model:
Using all papers consistently improved our predictions across all conferences
Predict the relevance of each affiliation in 2016 using all the papers from 2011-2015

Phase 3

Improve the model further

#1: find related conferences

- Authors submit papers to similar conferences
- Jaccard similarity using authors & keywords
- sim = (#common authors) / (#all authors)

By authors	By keywords
ICDM	CIKM
CIKM	ICDM
WWW	WWW
AAAI	SIGIR
ICML	SIGMOD
SDM	ICML
PAKDD	AAAI
ICDE	NIPS

Table: Conferences related to KDD

Phase 3

Improve the model further

#1: find related conferences

#2: expand the dataset even more

- Expand the dataset with papers starting with the year 2000

	# samples
Full research papers	3,677
Phase1: probabilities	1,296
Phase 2: full research papers	8,605
Phase 2: all papers	10,900
Phase 3: FSE + 5 related conferences	**25,136**
Phase 3: MOBICOM + 5 related conferences	**21,872**
Phase 3: MM + 10 related conferences	**92,762**

Table: Dataset evolution between phases

Phase 3

Improve the model further

#1: find related conferences

#2: expand the dataset even more

#3: refine the engineered features

Features

- Stats of all previous relevance scores (std, sum, mean, median, min, max)
- Previous relevance scores computed in windows from previous year up to 4 years ago
- Stats of previous relevance scores (std, sum, mean, median, min, max) computed in **windows** from previous year up to 4 years ago

Phase 3

Improve the model further

#1: find related conferences

#2: expand the dataset even more

#3: refine the engineered features

Features

- Stats of all previous relevance scores (std, sum, mean, median, min, max)
- Previous relevance scores computed in windows from previous year up to 4 years ago
- Stats of previous relevance scores (std, sum, mean, median, min, max) computed in **windows** from previous year up to 4 years ago

---

- Drift trend of previous relevance scores
- Exponential weighted moving average of previous relevance scores with estimated smoothing parameter
- Exponential weighted moving average of previous relevance scores, computed with a fixed smoothing parameter

Phase 3

Improve the model further

#1: find related conferences

#2: expand the dataset even more

#3: refine the engineered features

#4: tune the models

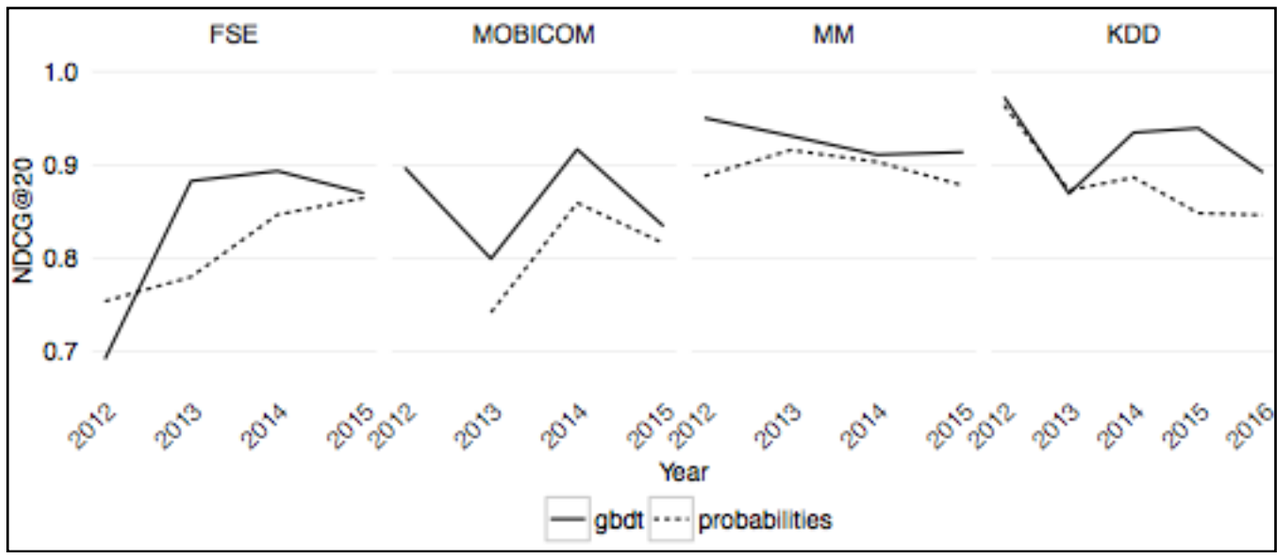

Figure: Results for the best features configuration

GBDT model:
Choose the model configuration (feature sets + #related conferences) which outperforms the baseline across all the years

Features

- Stats of all previous relevance scores (std, sum, mean, median, min, max)
- Previous relevance scores computed in windows from previous year up to 4 years ago
- Stats of previous relevance scores (std, sum, mean, median, min, max) computed in **windows** from previous year up to 4 years ago

---

- Drift trend of previous relevance scores
- Exponential weighted moving average of previous relevance scores with estimated smoothing parameter
- Exponential weighted moving average of previous relevance scores, computed with a fixed smoothing parameter

Key points

- Used only the publicly available Microsoft Academic Graph data
- Started simple and gradually improved our models
- Grew our dataset in each phase and this improved the predictability
- Tuned the models so they outperformed an already known good baseline

- **Won the cup!**

THANK YOU

questions?