

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

Vlad Sandulescu

Senior Data Scientist @ Adform, Denmark

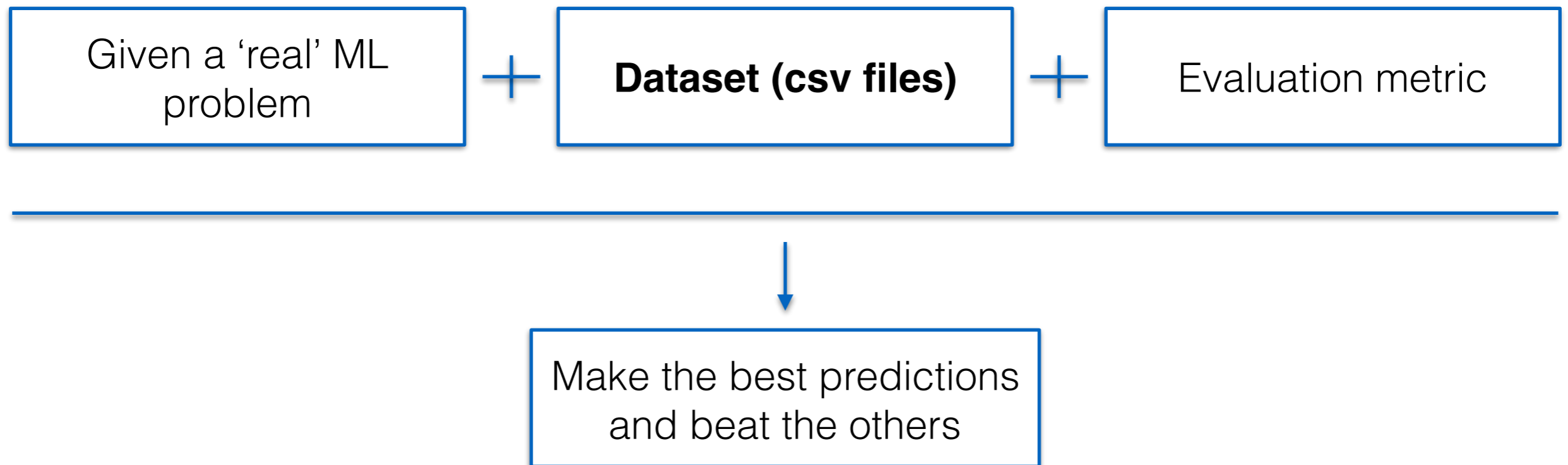
 @vladsandulescu / vladsandulescu.com

joint work with **Mihai Chiru**, from Bitdevelop in Sweden

2017.01.30 - Data Science Day 2017, Copenhagen, Denmark

ML Competitions

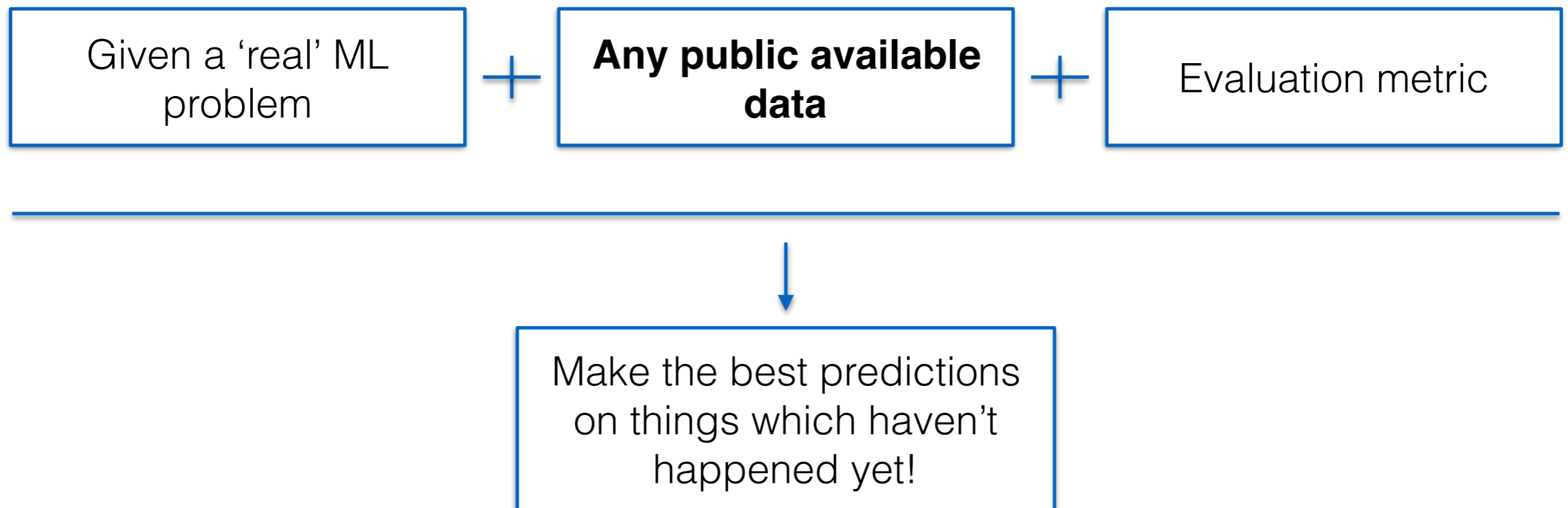
kaggle™



- Public/Private Leaderboard (public - 30% of test data, private - 70%)
- Competitors have diverse backgrounds, some are serial Kagglers with an already available personal modeling toolkits
- Many times the winners ensemble and stack a looooot of models => monster model

The KDD Cup 2016

- Happens each year, part of the ACM SIGKDD conference on Knowledge Discovery and Data Mining (KDD)
- More than 550 teams participated this year
- Top 3 prizes: \$10,000, \$6500 and \$3500



- More academics participate in the KDD Cup
- Competing against others but also against future real-world events

The KDD Cup 2016

The task?


Rank research institutions by predicting how many of their full research papers will be accepted at future academic conferences.

conferences = [SIGIR, SIGMOD, SIGCOMM], [KDD, ICML], [FSE, MobiCom, MM]

research institutions = *Google, Stanford, Oxford, Microsoft, etc.*

full research papers = different than workshop papers, poster papers, tutorials, etc.

Predictions



Rank	Affiliation	#papers
1	Google	9
2	Microsoft	8
3	CMU	7
...
20	Yahoo	1

The KDD Cup 2016

A paper has multiple authors, possibly from different affiliations.

So:

- Each accepted paper has an equal vote (i.e., they are equally important).
- Each author has an equal contribution to a paper.
- If an author has multiple affiliations, each affiliation also contributes equally.

Paper 1

author 1: affiliation 1
author 2: affiliation 2
author 3: affiliation 3

Paper 2

author 1: affiliation 1
author 4: affiliation 3
affiliation 4

	affiliation 1	affiliation 2	affiliation 3	affiliation 4
paper 1	1/3	1/3	1/3	0
paper 2	1/2	0	1/4	1/4
...				
relevance	0.83	0.33	0.58	0.25
ranking	1	3	2	4

The KDD Cup 2016

The evaluation metric?

$NDCG@n=20$ (the gist= penalizes the predicted ranking versus the ideal ranking)

$$DCG_n = \sum_{i=1}^n \frac{rel_i}{\log_2(i+1)} \quad NDCG_n = \frac{DCG_n}{IDCG_n}$$

Suppose we submit these predictions...

i	Affiliation
1	Microsoft
2	UIUC
3	IBM
4	Yahoo!

The KDD Cup 2016

The evaluation metric?

NDCG@20 - n = 20

Suppose we submit these predictions... then we find out the true ranking.

i	Affiliation	true i	true rel _i	log ₂ (i+1)	rel _i /log ₂ (i+1)
1	Microsoft	1	8.46	1	8.46
2	UIUC	3	4.18	1.585	2.64
3	IBM	4	2.51	2	1.25
4	Yahoo!	2	6.23	2.32	2.68

$$DCG_4 = \sum_{i=1}^4 \frac{rel_i}{\log_2(i+1)} = 8.46 + 2.64 + 1.25 + 2.68 = 15.03$$

$$IDCG_4 = \frac{8.46}{1} + \frac{6.23}{1.585} + \frac{4.18}{2} + \frac{2.51}{2.32} = 15.57$$

$$NDCG_4 = \frac{DCG_4}{IDCG_4} = \frac{15.03}{15.57} = 0.965$$

The Microsoft Academic Graph

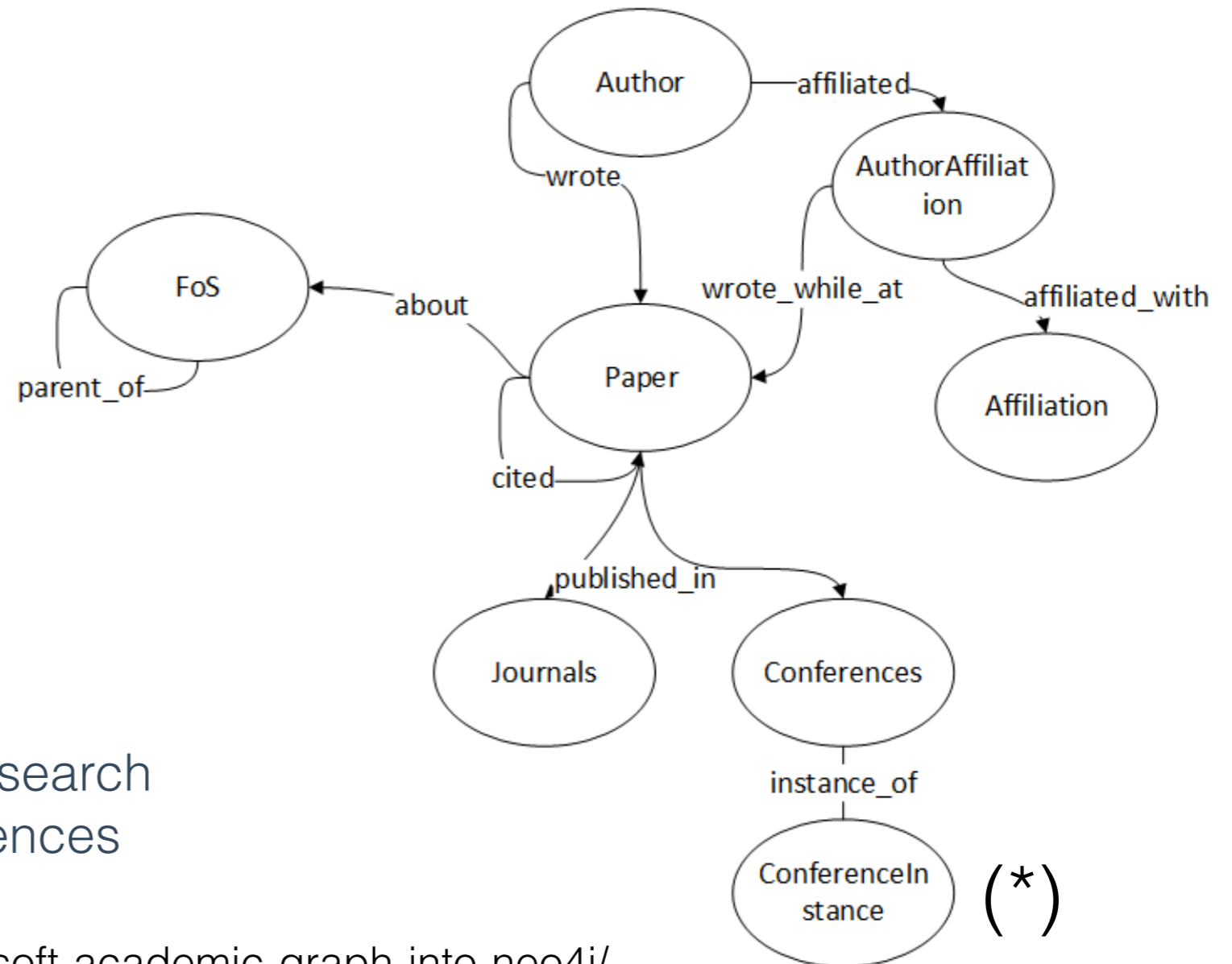
The data?

Participants could use any publicly available information on the Web.

MAG - large heterogenous graph of academic papers metadata (large csv files ~30GB)

- Papers
- Authors
- Affiliations
- Conferences
- Papers-Authors-Affiliations
- Papers Keywords
- Fields of study

- Microsoft provided a list of **full** research papers accepted at the 8 conferences



(*) - <http://www.markcosta.net/load-the-microsoft-academic-graph-into-neo4j/>

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.)

Phase 1

Start simple

#1: build a dataset

#2: explore the dataset

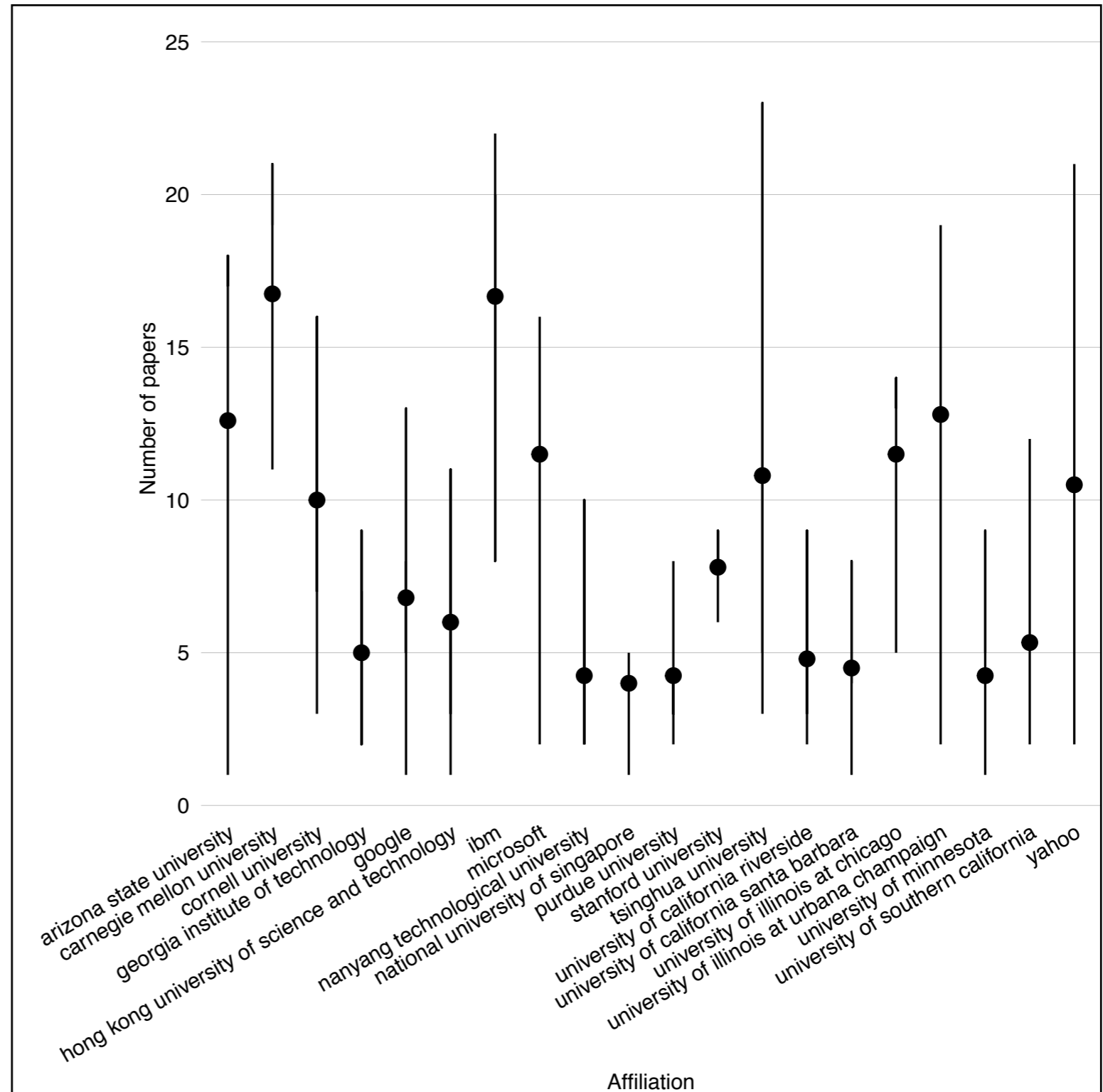


Figure: Full range and mean value of the number of accepted full research papers for top 20 affiliations at KDD between 2011 and 2015

Phase 1

Start simple

#1: build a dataset

#2: explore the dataset

#3: set up a baseline

Baseline model:

Compute the probability of a full research paper is published by an affiliation at each conference using historical 2011-2015 data

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

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, so why not predict this instead of the number of papers!

Phase 2

Start improving the baseline

#1: predict relevance directly

#2: explore the dataset even more

#3: try GBDT & Mixed models

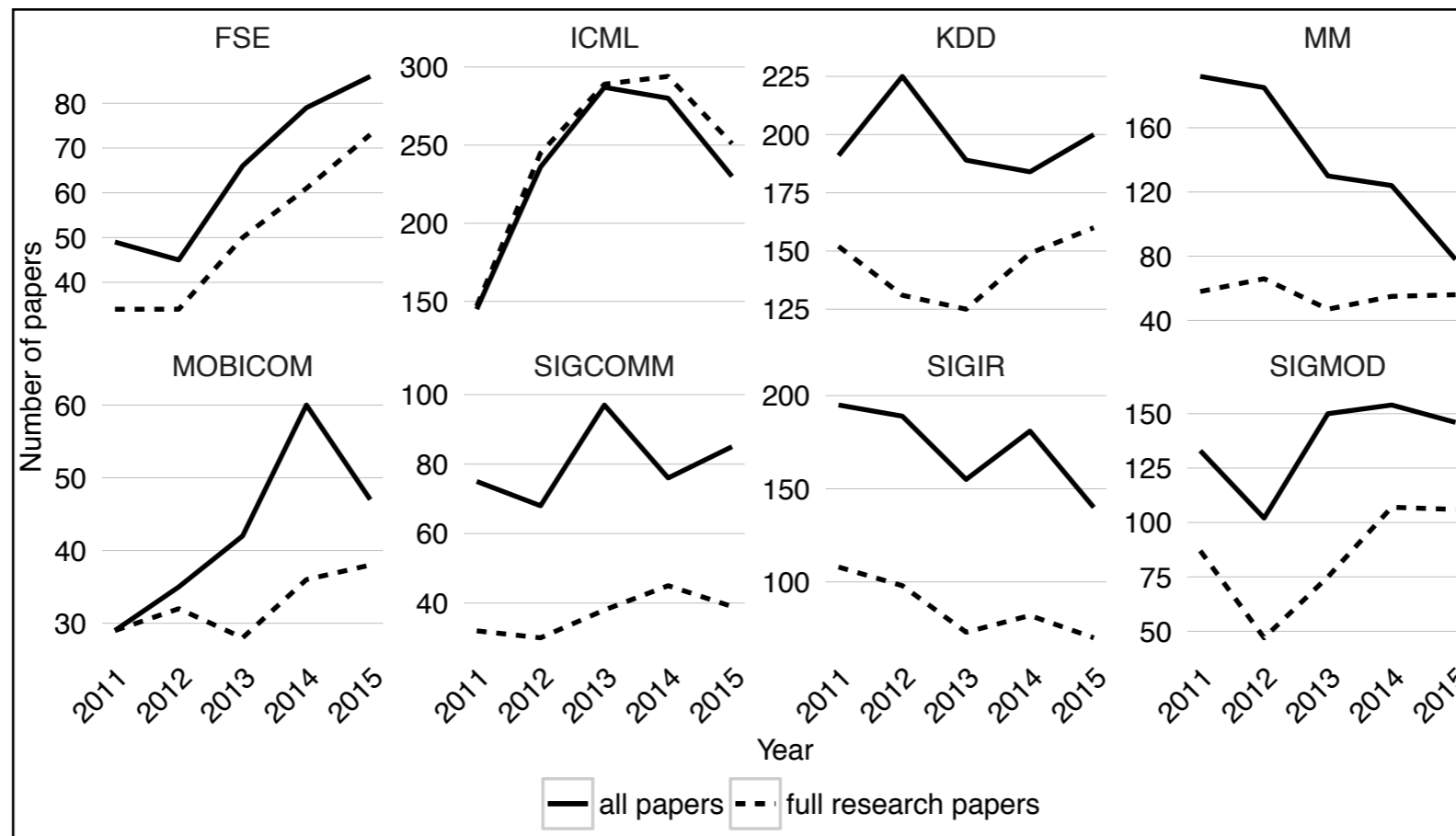


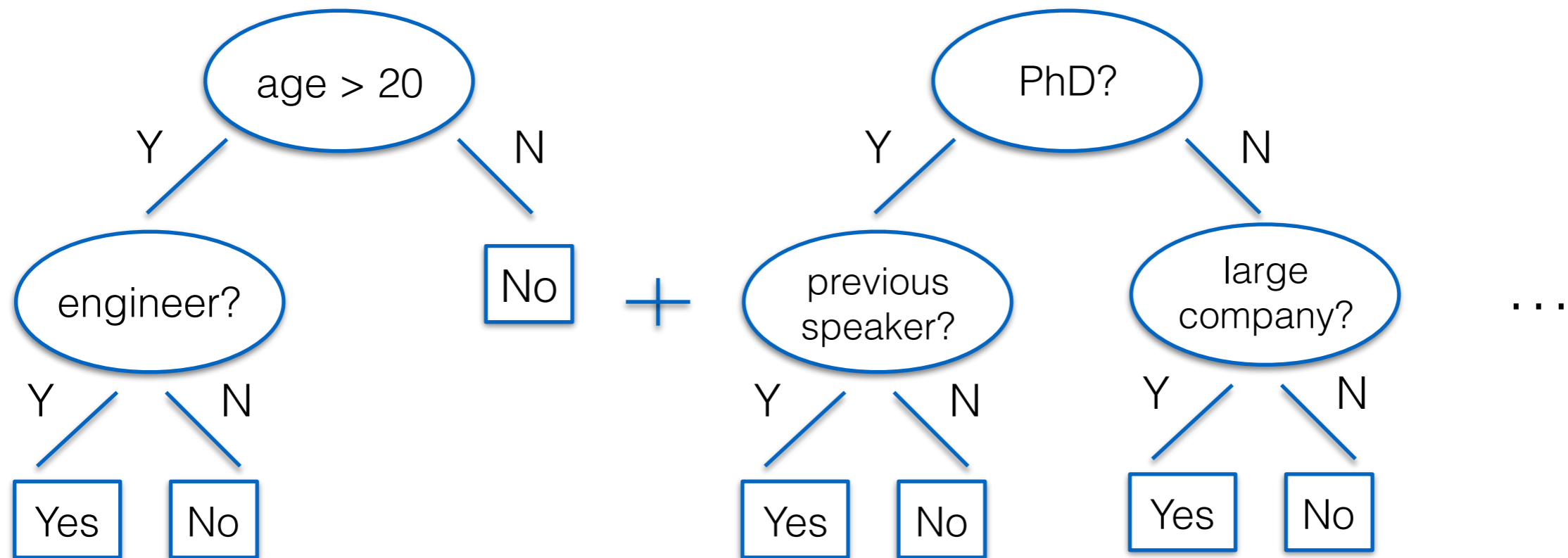
Figure: Number of all the papers vs the full research papers for all the conferences in the competition for 2011-2015

Gradient boosted decision trees and mixed models

Should I attend this presentation at the conference?

Decision trees

Input: age, engineer, company size, previous speaker, PhD, etc.



- additive model of weak learners, each improving on the residuals of the previous tree
- mixed models - just think linear regression per conference per affiliation

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**

	# 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**

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 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
- $\text{sim} = (\text{\#common authors}) / (\text{\#all authors})$

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
Phase 1: 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

More Features

- 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

GBDT model:

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

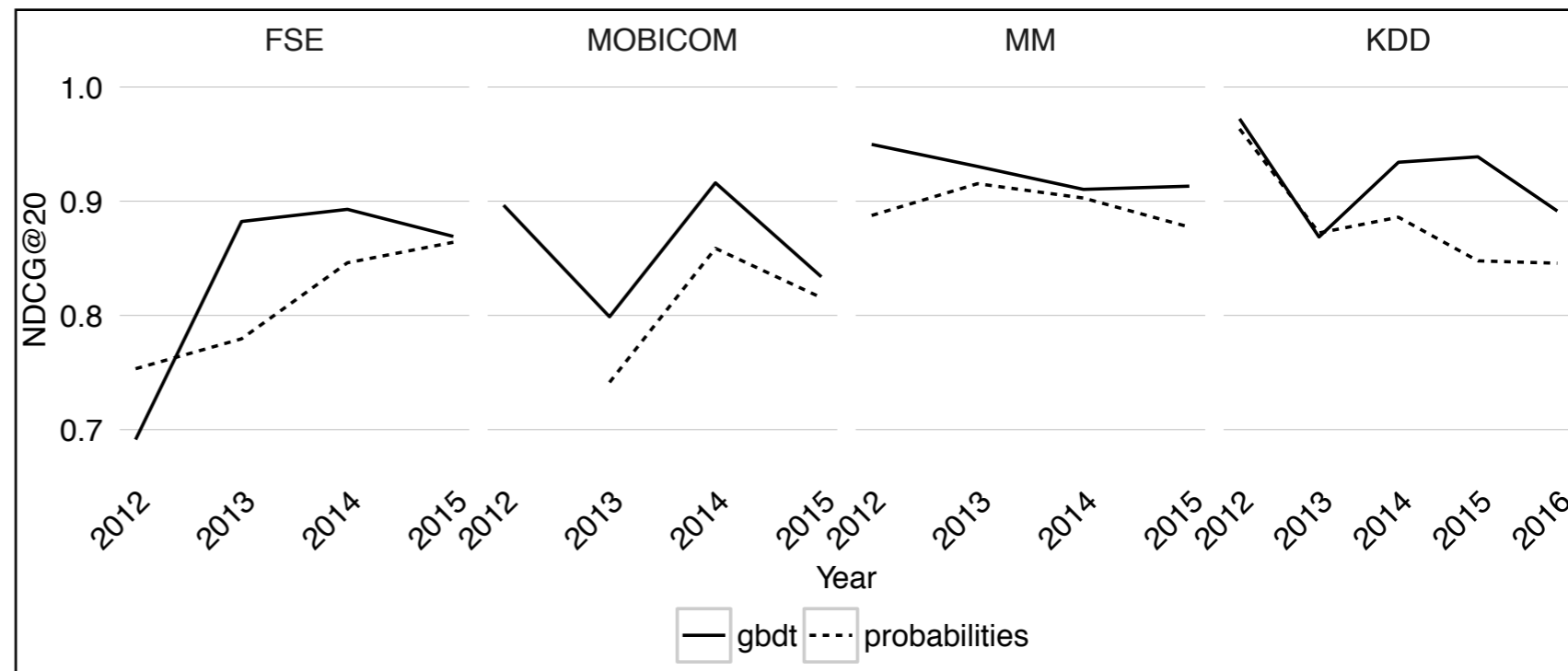


Figure: Results for the best features configuration

Key points

- Data exploration is crucial to get good models
 - Always have a baseline you know works well
 - Start simple and gradually improve your models
 - More data is always better
 - Teamwork is a natural ensemble
-
- **Won the cup!**

THANK YOU

questions?

Slides @ vladsandulescu.com

Link to paper

<https://arxiv.org/pdf/1609.02728v1.pdf>