Instituto Superior Técnico
Universidade Técnica de Lisboa

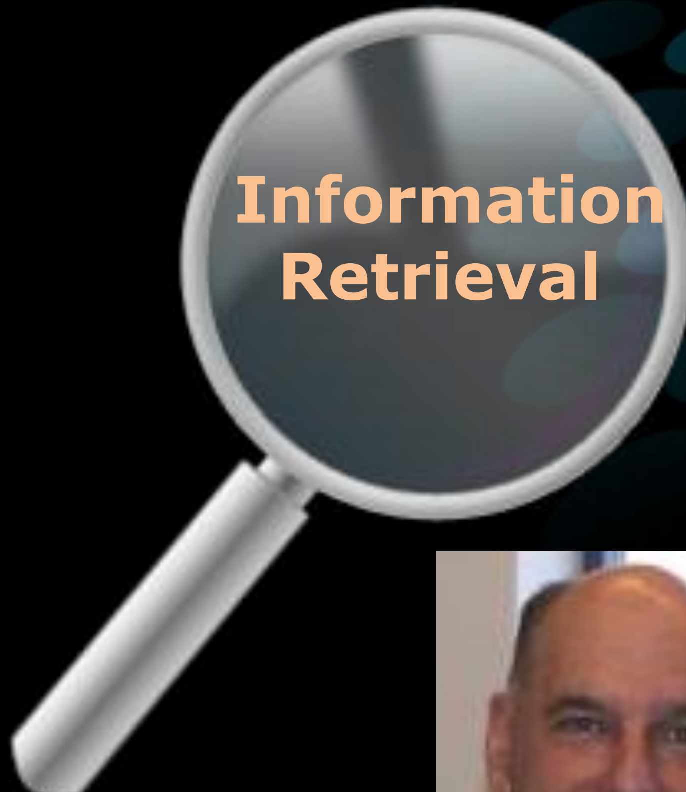
Using Rank Aggregation for Expert Finding

Catarina Moreira, Bruno Martins and Pável Calado

Outline

- ✓ Introduction
- ✓ Related Work
- ✓ Rank Aggregation
- ✓ Features
- ✓ Dataset
- ✓ Experimental Results
- ✓ Future Work

Expert Finding



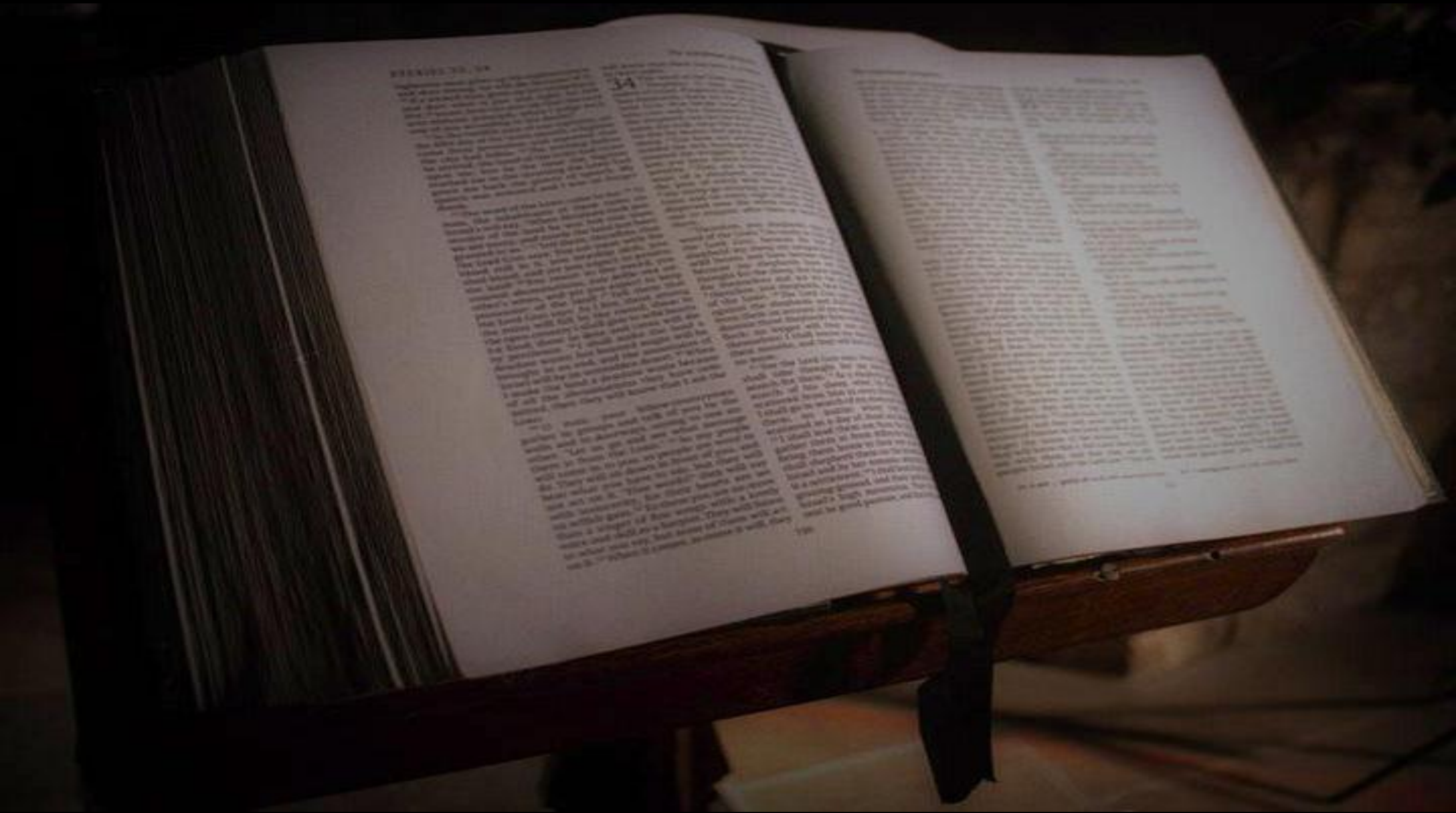
Why Expert Finding?

Too many documents

Information is dispersed

Need answers quickly

Related Work



Candidate Centric Approach

1. Gather documents associated to a candidate
2. Merge documents into a single profile document
3. Rank the profile according to the query

Document Centric Approach

1. Gather documents containing query topics

2. Uncover candidates and rank them

Problems?

Generative Probabilistic Models

Simple heuristics

Heuristics do not reflect expertise

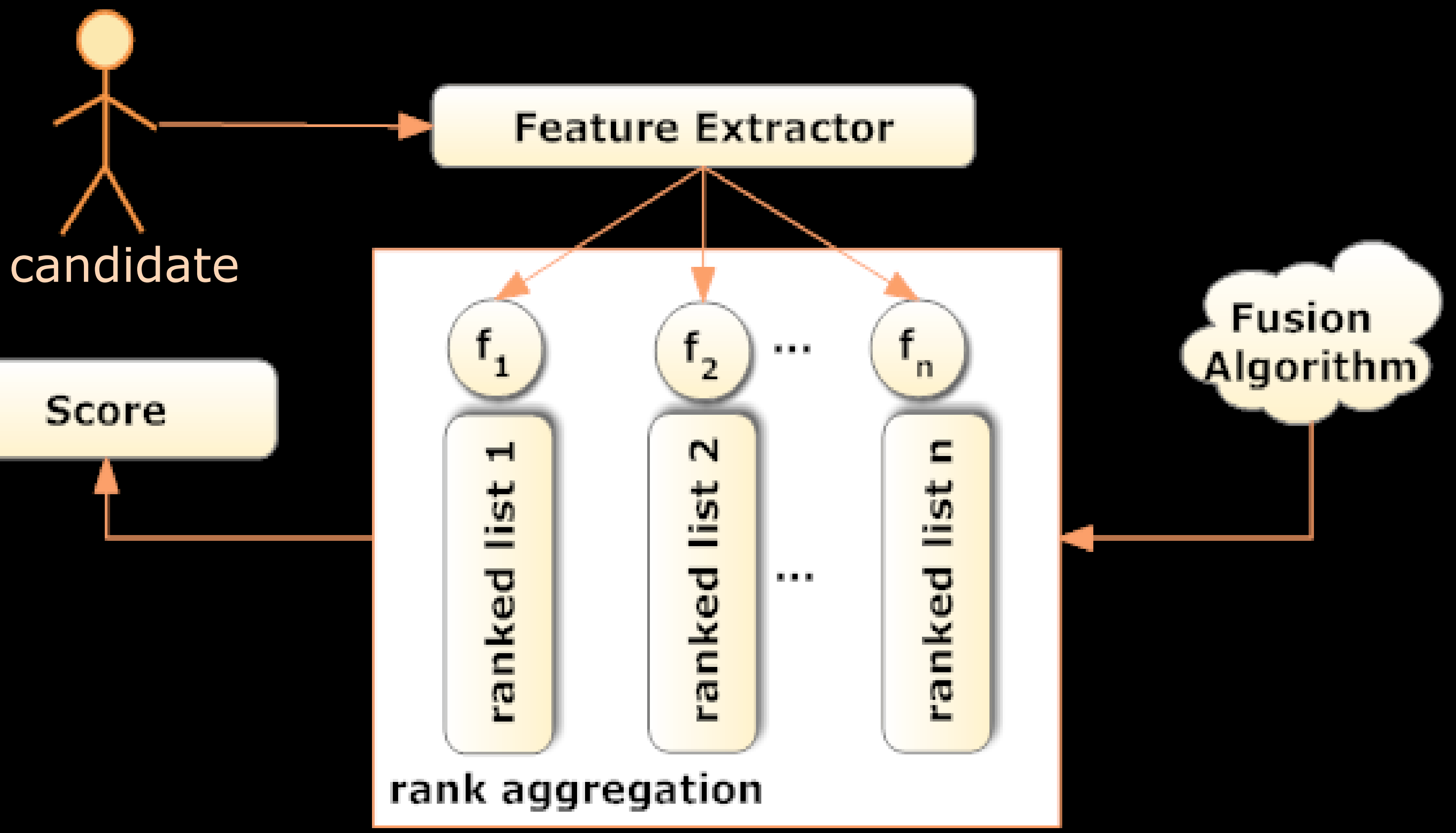
Only based on textual contents

Our Approach

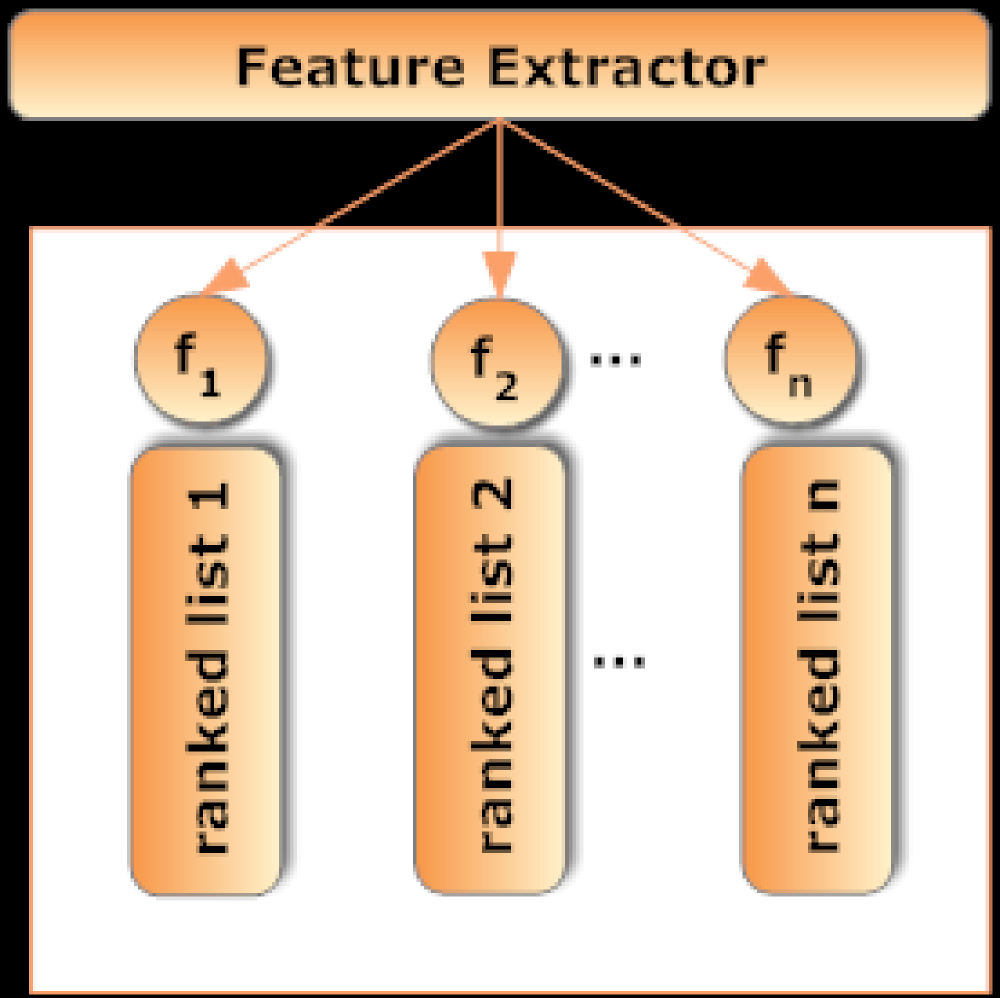
A set of features to estimate expertise

Features combined in a rank aggregation framework

Rank Aggregation



Feature Extractor

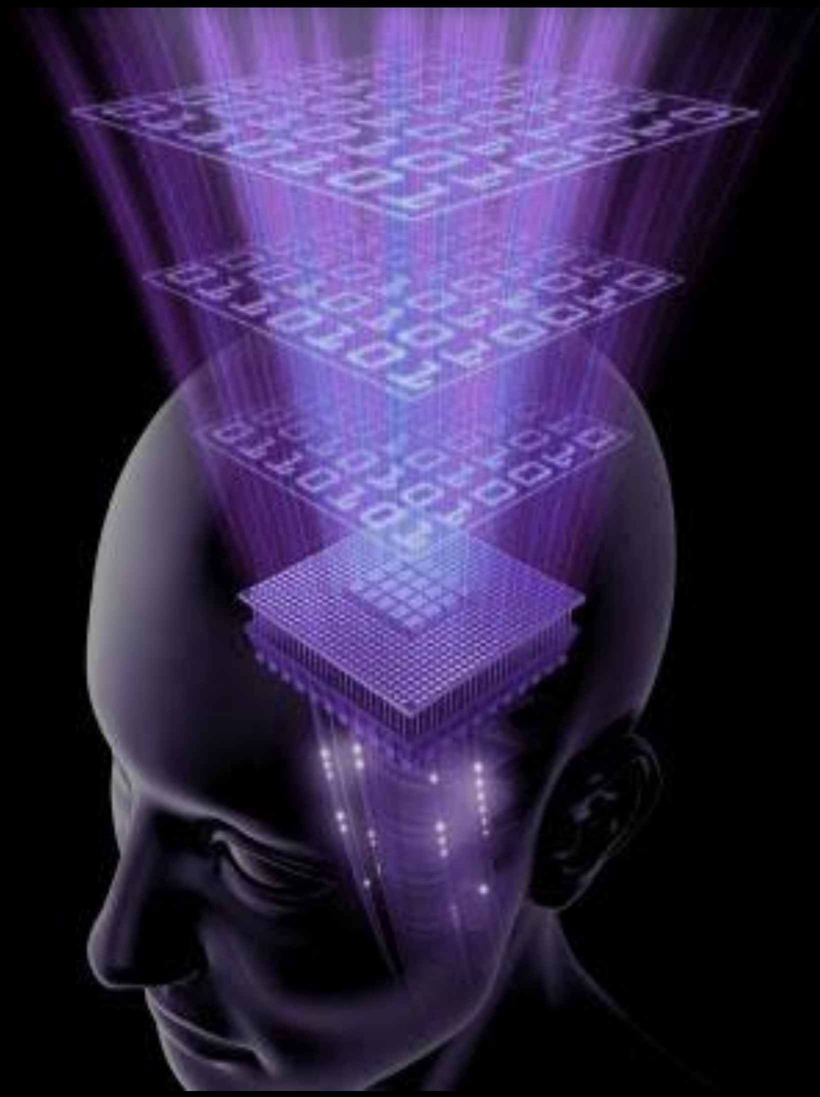


Features

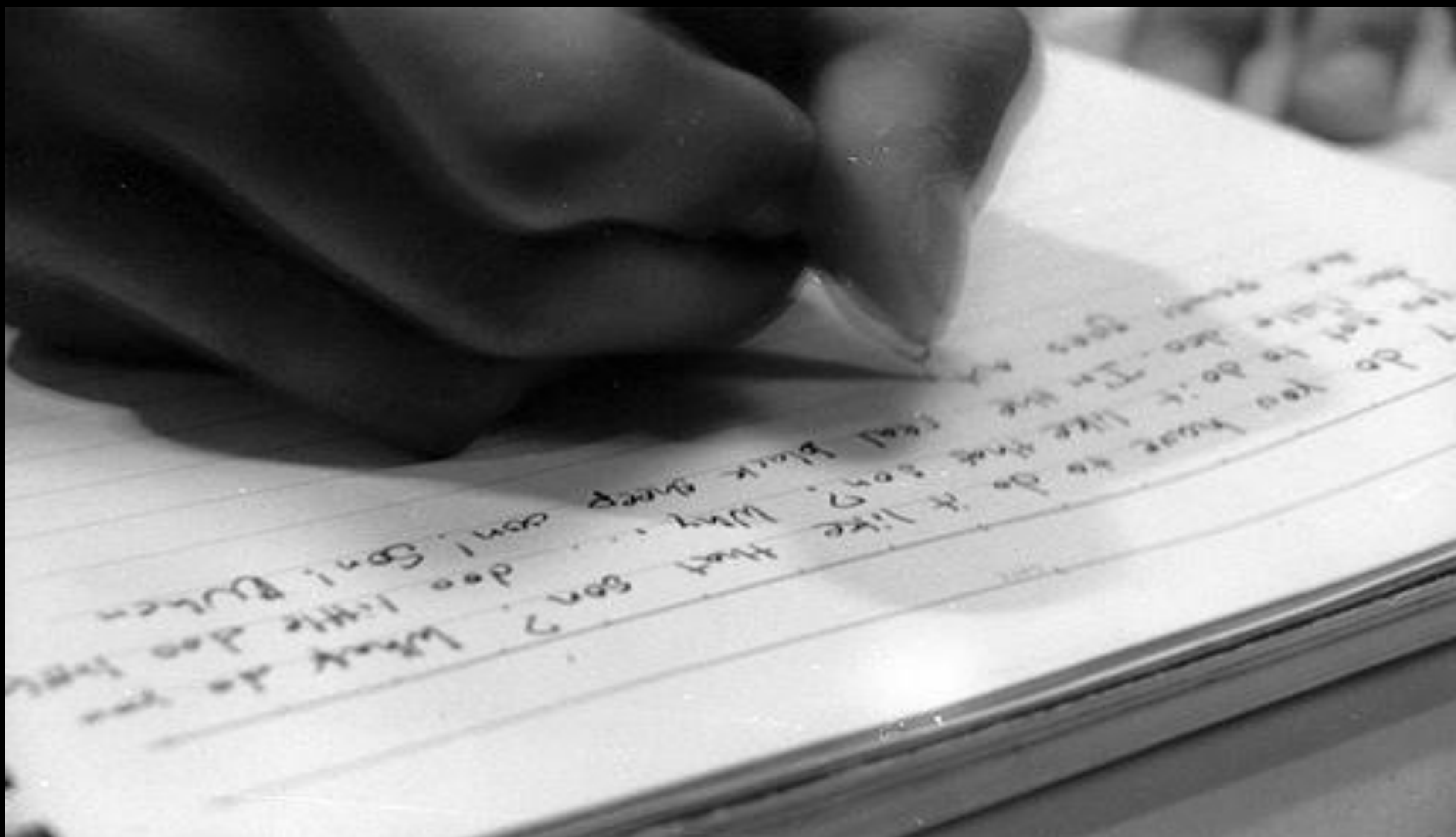
Textual Similarities

Profile Information

Graph Structure



Textual Features



Textual Features

TF

$$TF_{q,a} = \sum_{j \in Docs(a)} \sum_{i \in Terms(q)} \frac{Freq(i, d_j)}{|d_j|}$$

IDF

$$IDF_q = \sum_{i \in Terms(q)} \log \frac{|D|}{f_{i,D}}$$

BM25

$$BM25_{q,a} = \sum_{j \in Docs(a)} \sum_{i \in Terms(q)} \log \left(\frac{N - Freq(i) + 0.5}{Freq(i) + 0.5} \right) \times \frac{(k_1 + 1) \times \frac{Freq(i,d_j)}{|d_j|}}{\frac{Freq(i,d_j)}{|d_j|} + k_1 \times (1 - b + b \times \frac{|d_j|}{\mathcal{A}})}$$

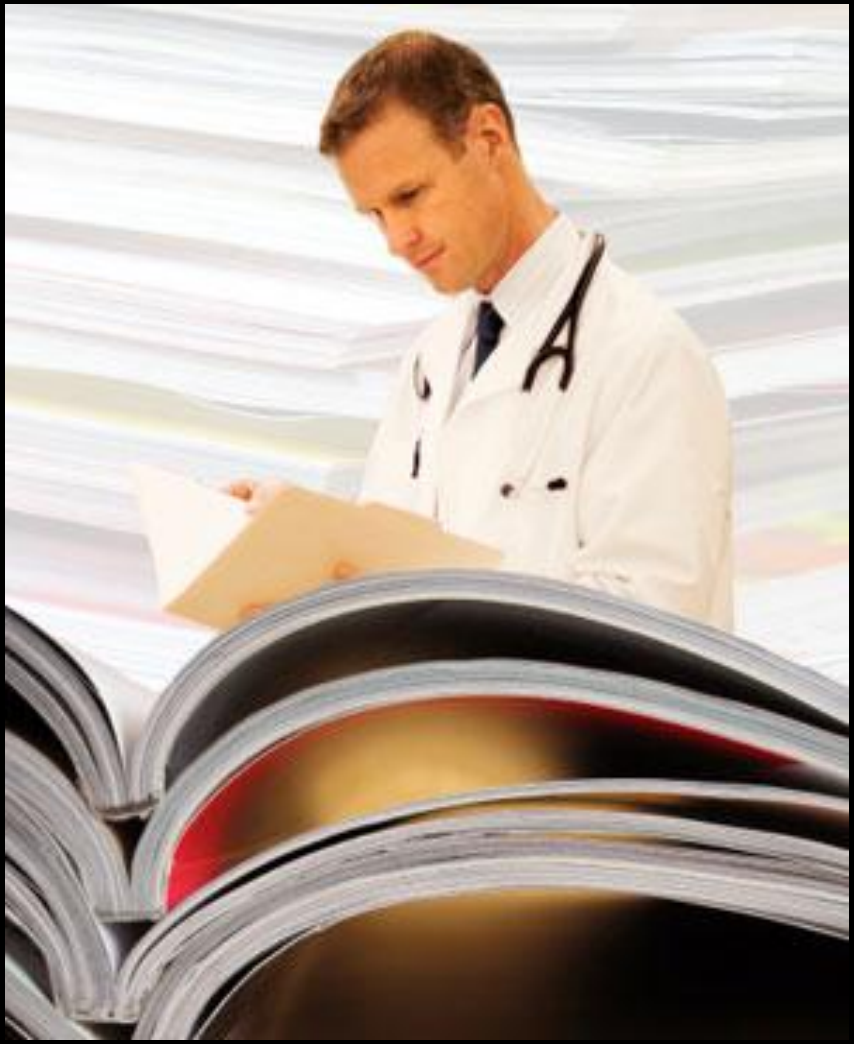
Profile Features

Profile Features

Number of Publications

Years Between Publications

Number of Articles



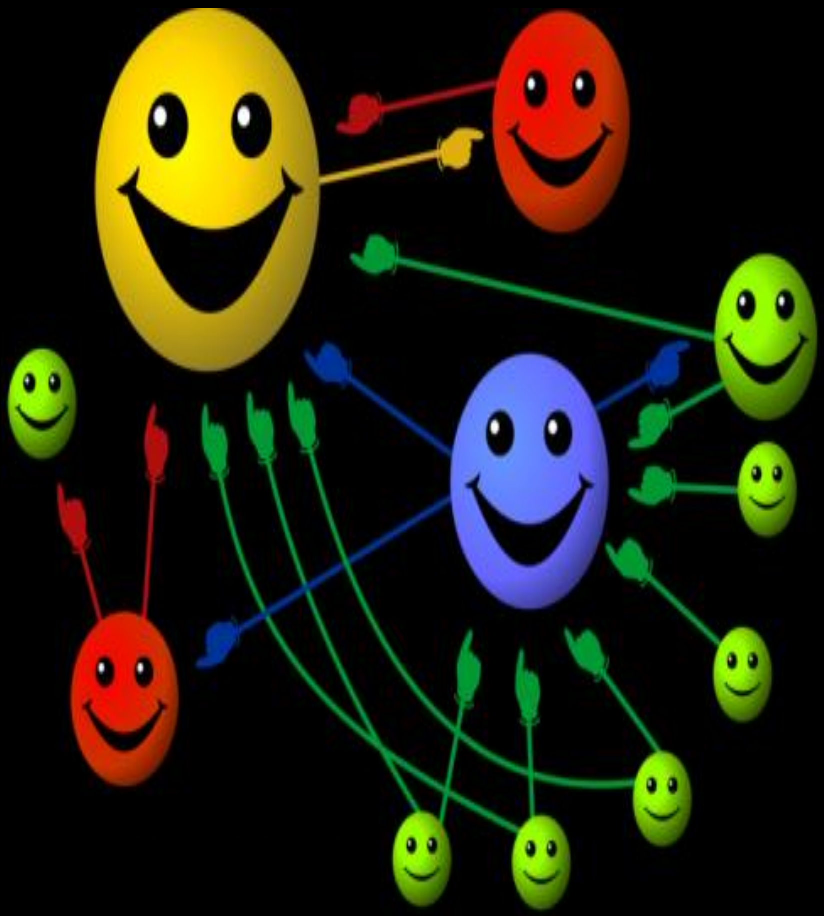
Graph Features

Graph Features

Citations Graphs

Co-authorship Graphs

Academic Indexes



Academic Indexes Measure Scientific Impact!

Academic Indexes

H-Index

G-Index

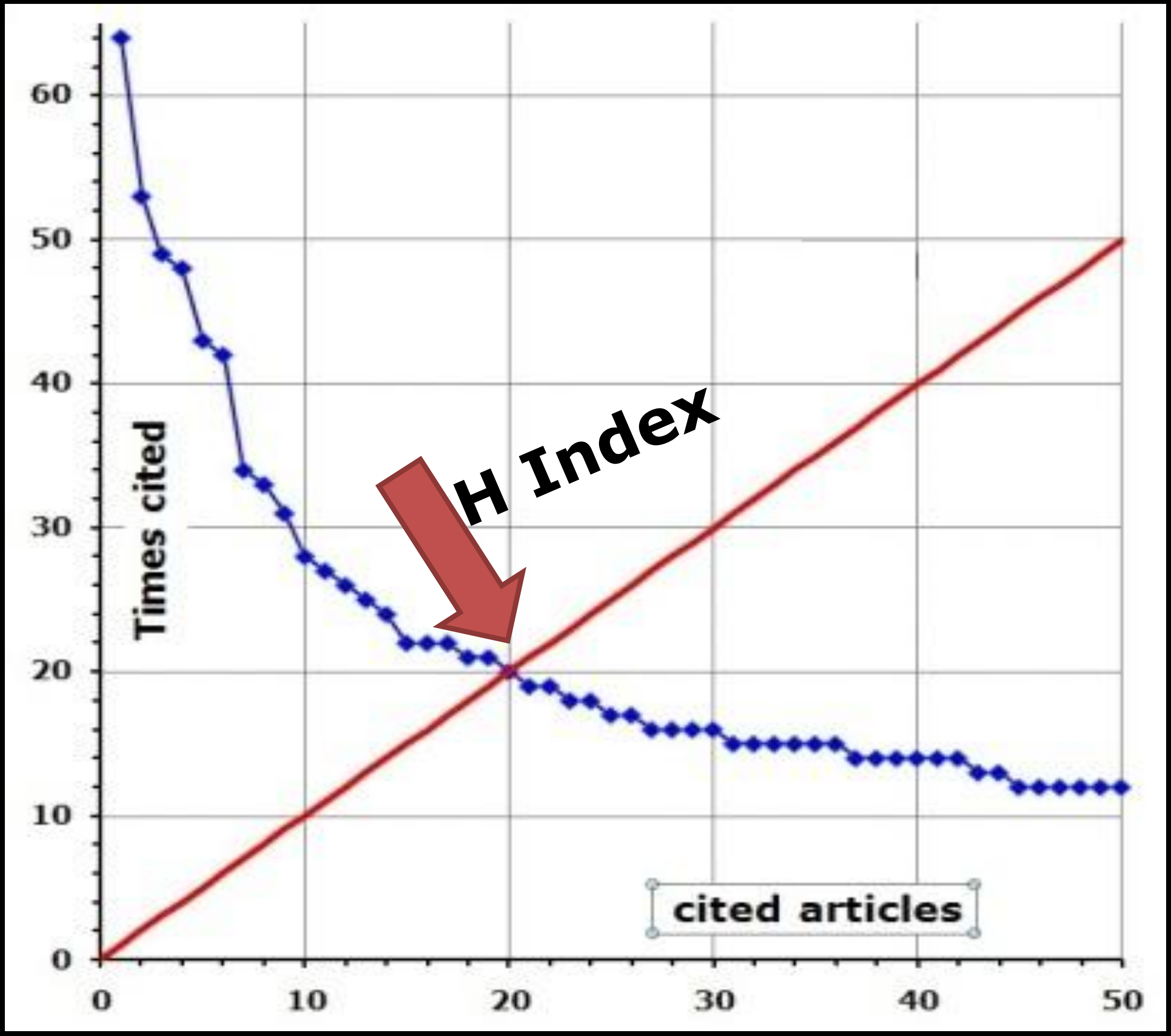
A-Index

H Index

A given author has a Hirsch Index of *h*, if *h*

of his *N* papers have at least *h* citations each

H Index - Example



G Index

Is the largest number such that the top *g* papers received on average at least *g* citations each

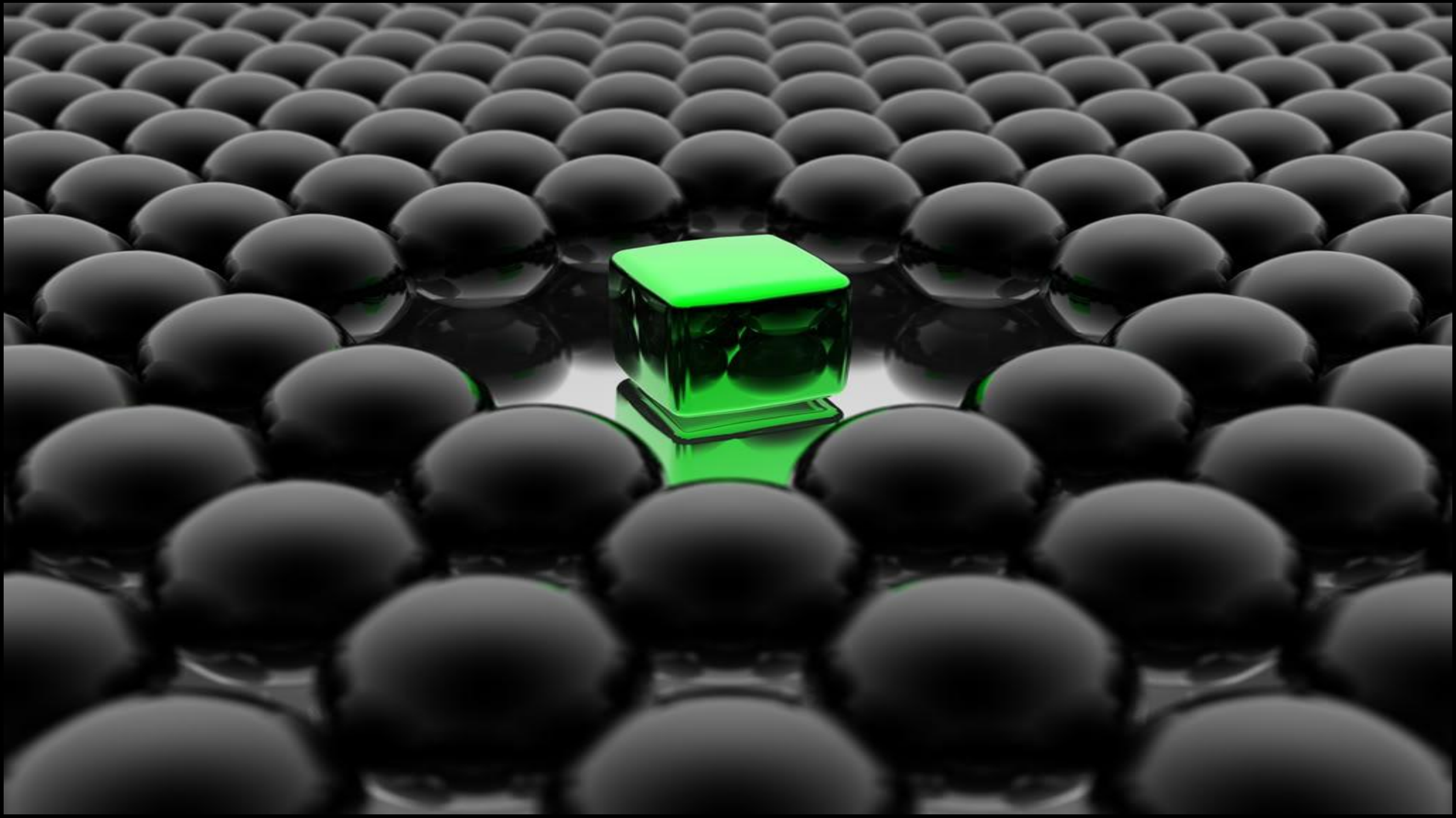
a Index

Measures the maginitude of the most influential

papers of a given author

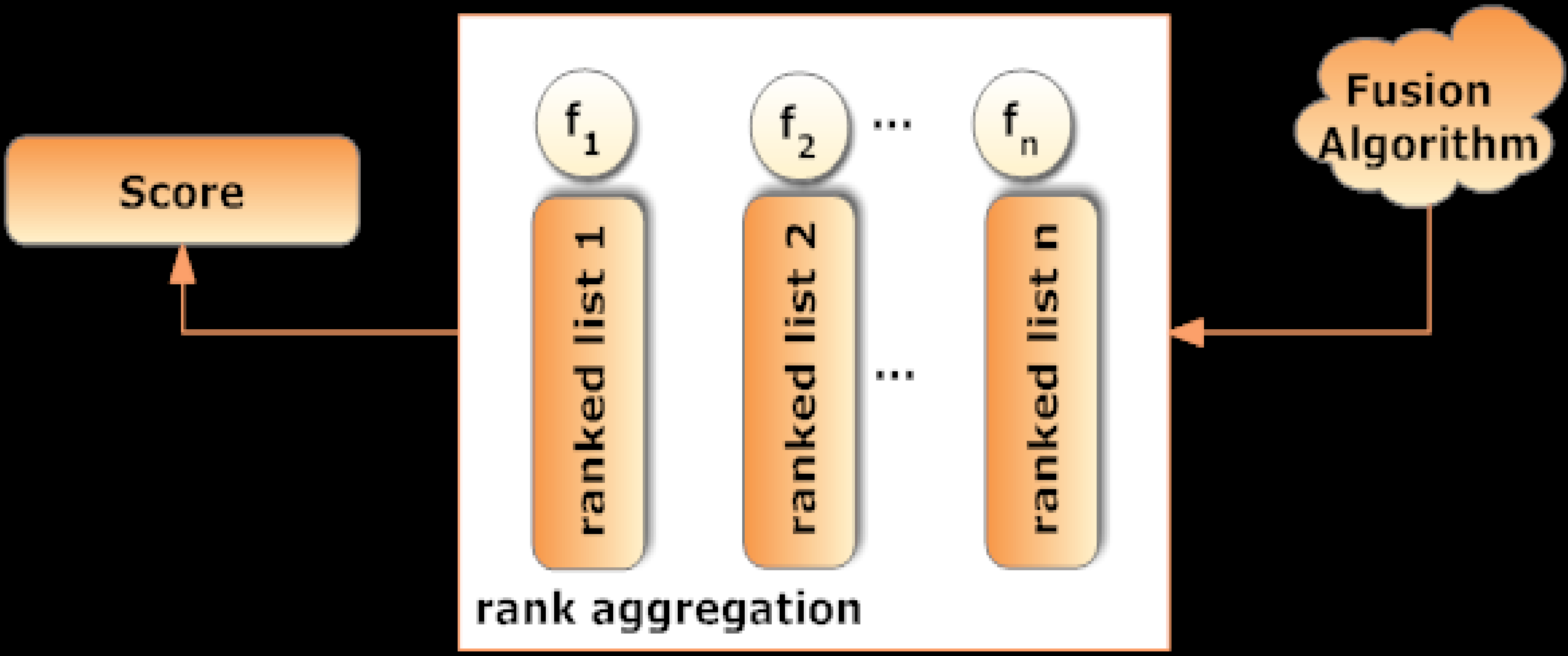
$$a = N_{c,tot}/h^2$$

First work using academic indexes



for Expert Retrieval!

Fusion Algorithms



Fusion Algorithms

CombSUM

$$CombSUM(e,q) = \sum_{j=1}^{k} score_j(e,q)$$

CombMNZ

$$CombMNZ(e,q) = CombSUM(e,q) \times r_e$$

Normalization

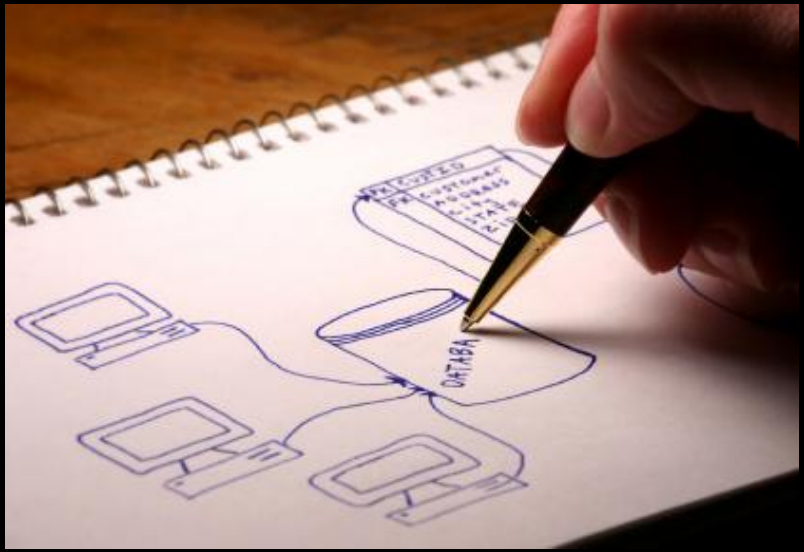
CombSUM and CombMNZ require normalized scores

$$NormalizedValue = \frac{Value - minValue}{maxValue - minValue}$$

Dataset

DBLP Computer Science Bibliography

- Covers journal and conference publications
- Contains publication abstracts
- Contains citation links



Dataset for Validation

Arnetminer

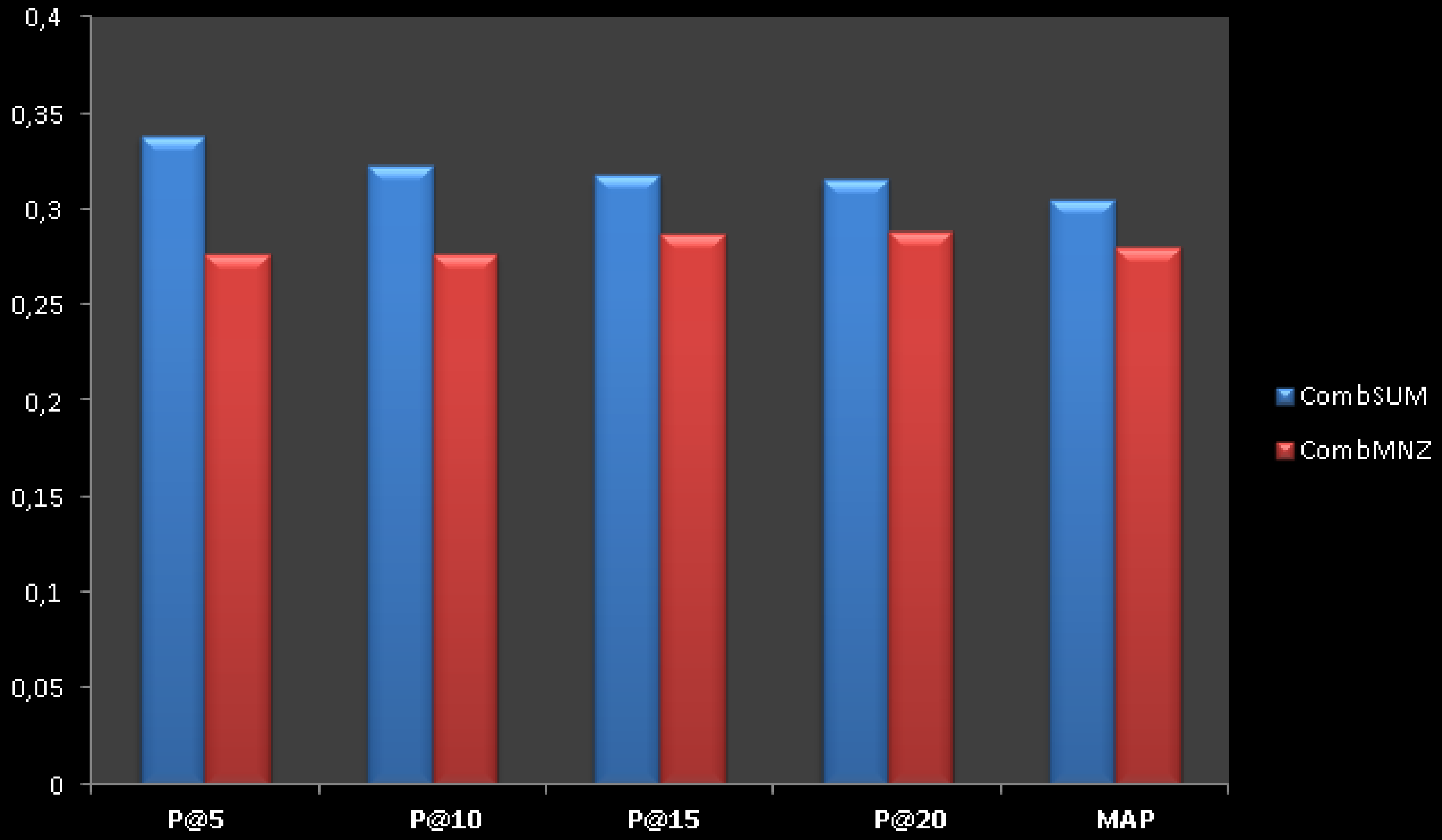
Contains a set of people considered experts

Contains 13 different query topics

Based on people from program committees of important conferences

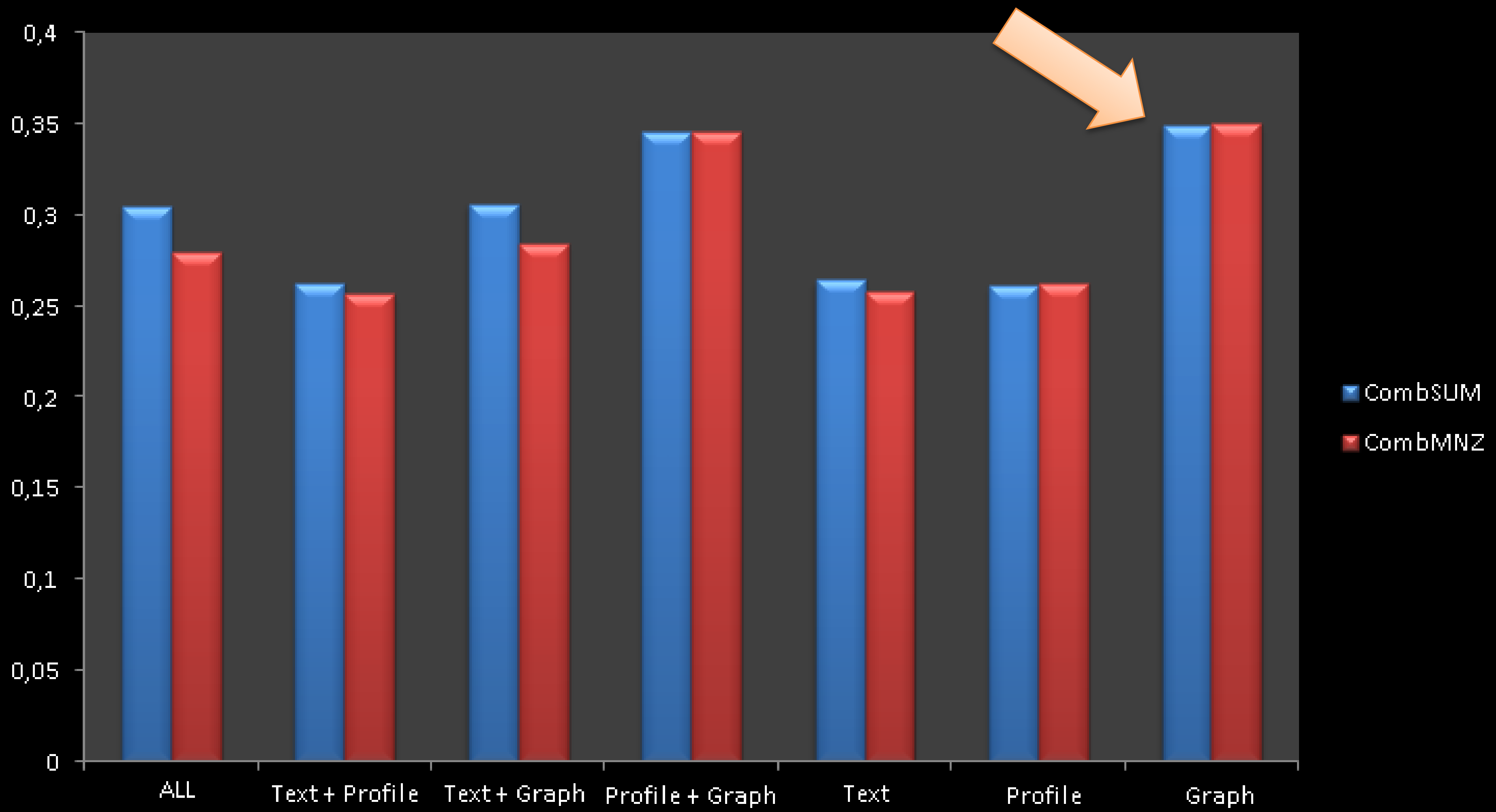
Experimental Results

CombSUM Wins!



Impact of the Features?

Graph + Academic Features are the Best!

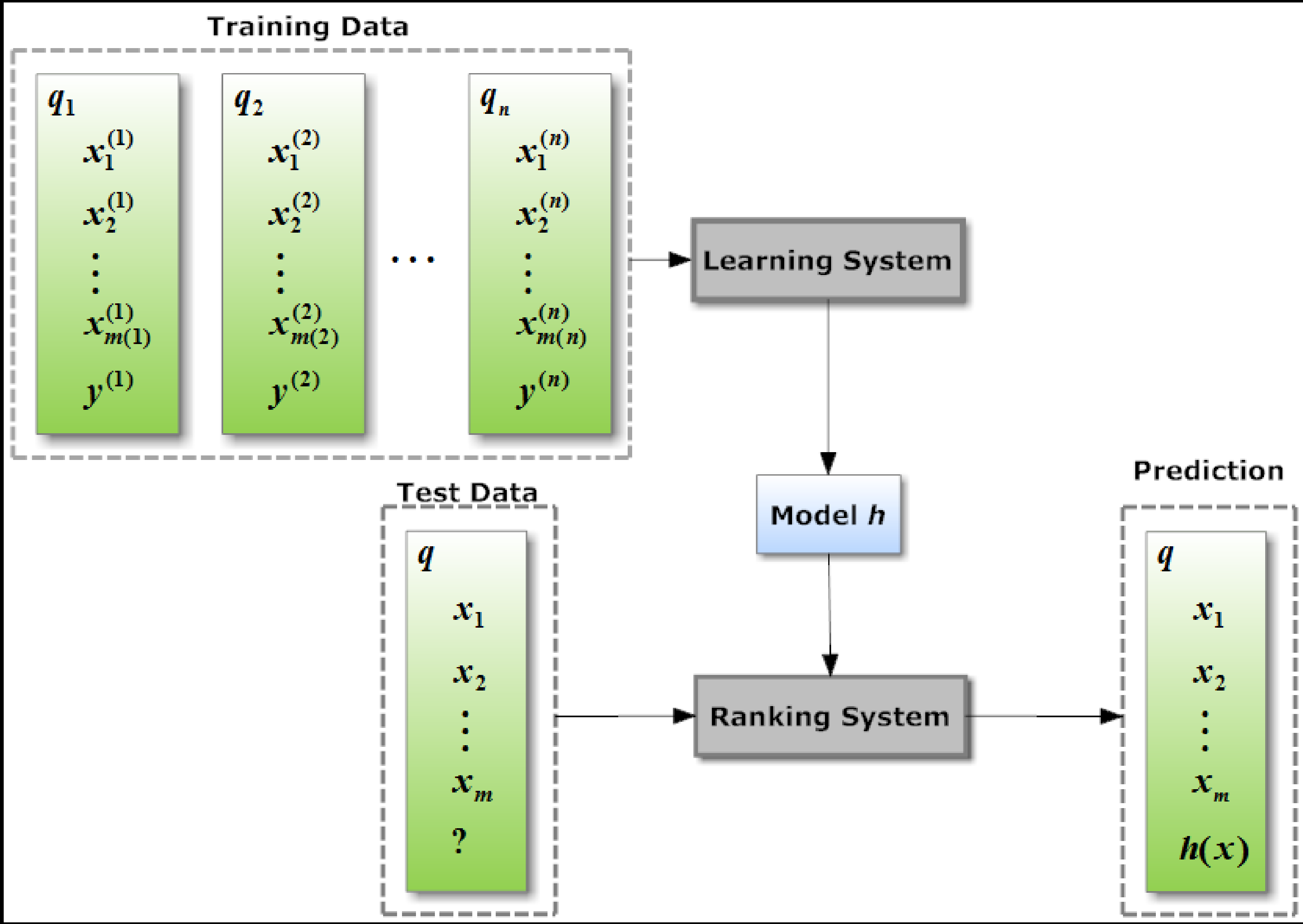


Future Work

The set of features defined in this work are effective!

But, how to combine them in an **optimal way**?

Learning to Rank



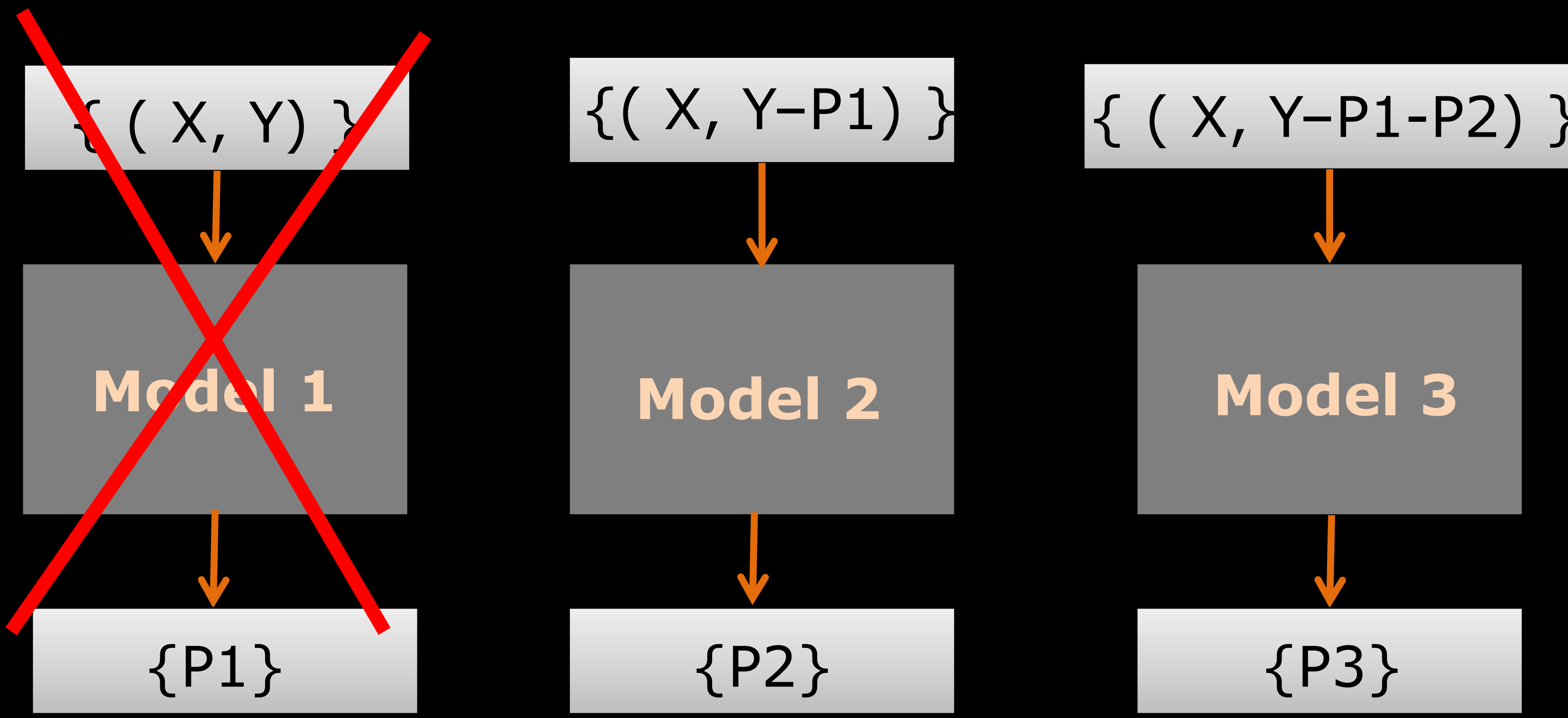
Learning Algorithms

Additive Groves by Daria Sorokina

Additive Groves

Training Set: { (X , Y) }

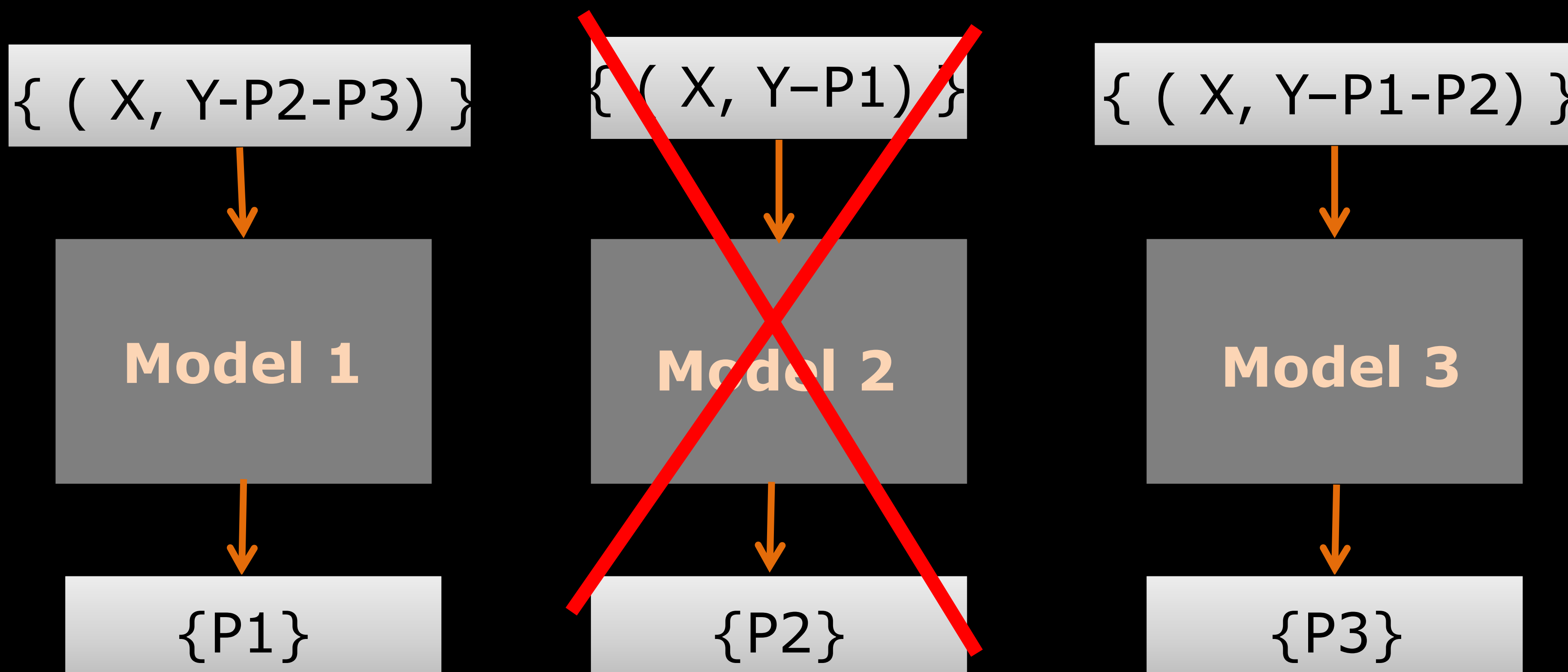
Goal: model h = P1 + P2 + P3



Additive Groves

Training Set: { (X , Y) }

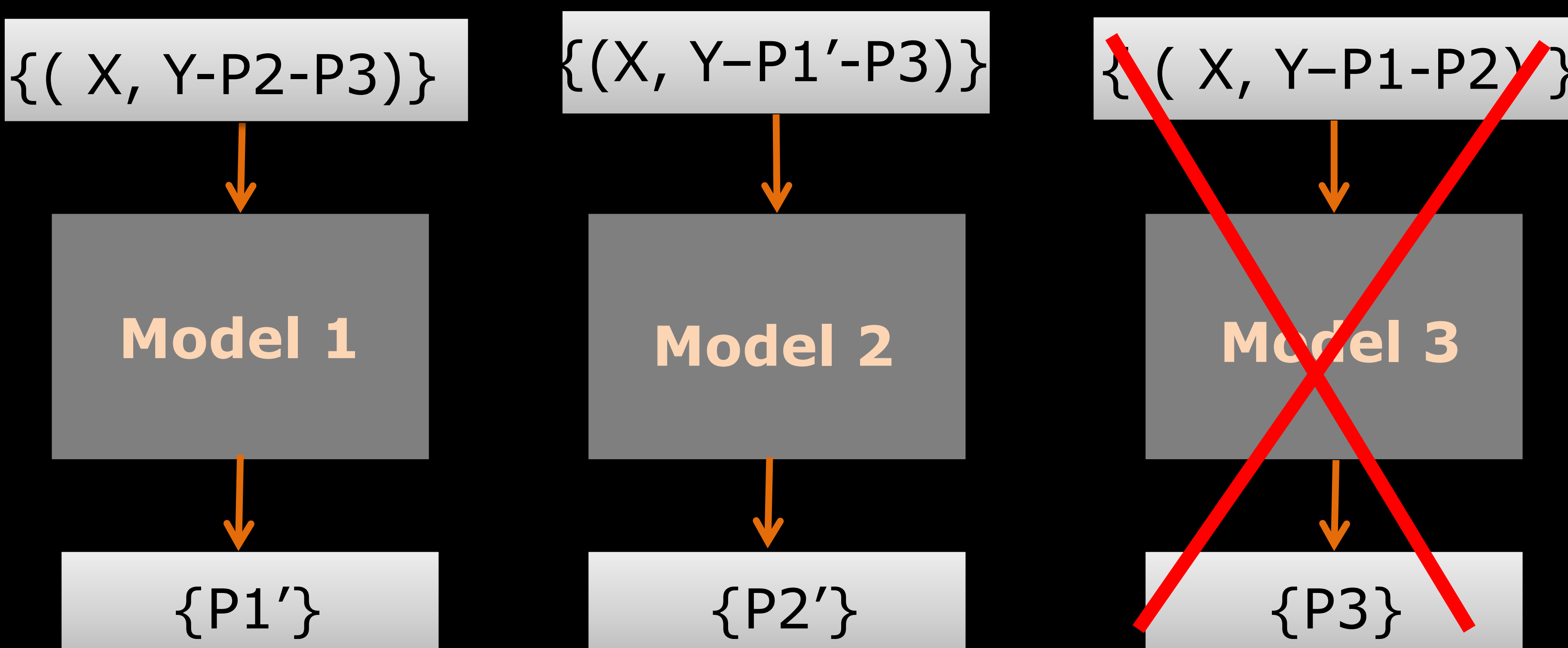
Goal: model h = P1 + P2 + P3



Additive Groves

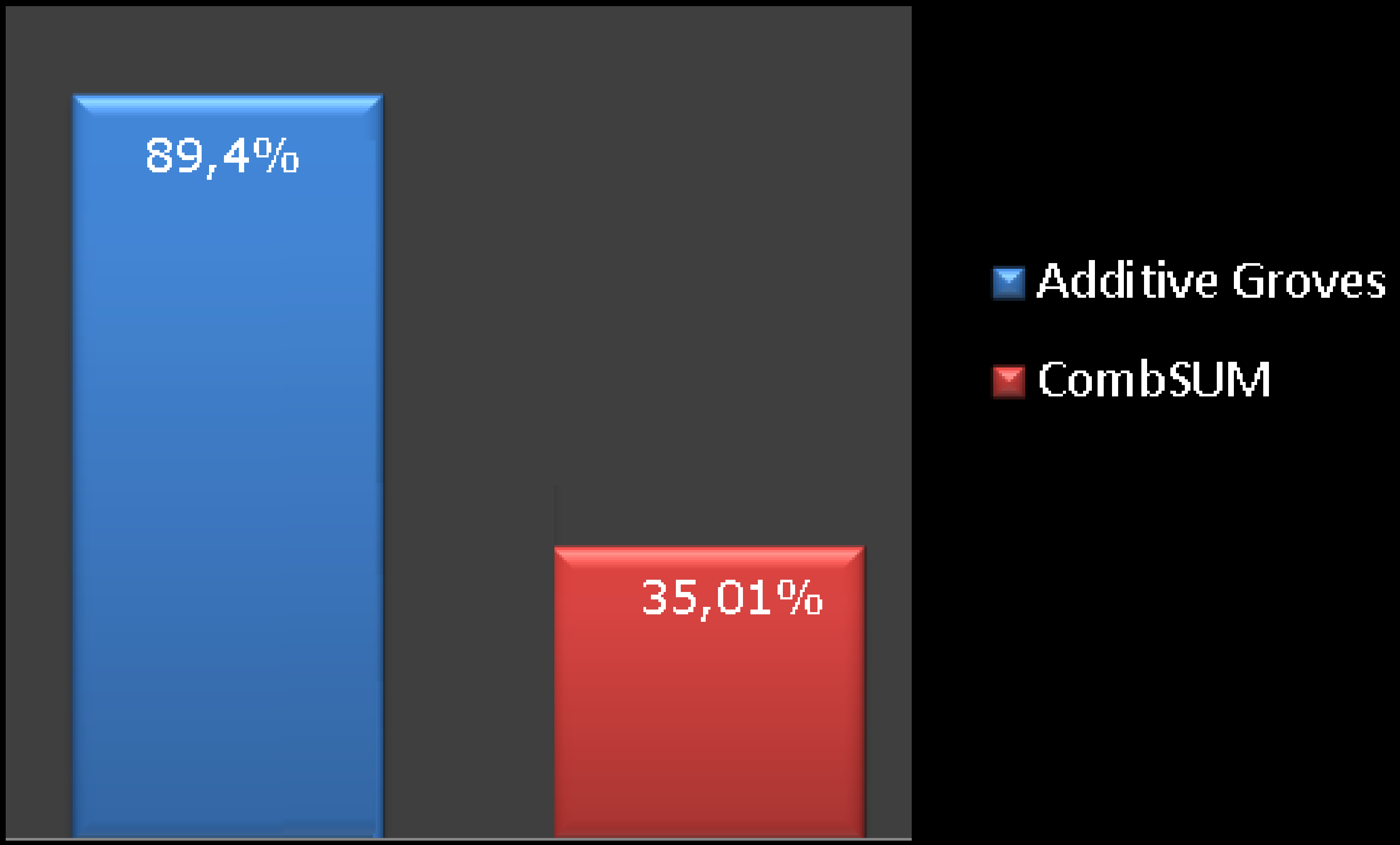
Training Set: { (X , Y) }

Goal: model h = P1 + P2 + P3



Experimental Results

Additive Groves vs CombSUM




Thank You!!