# An Approach to Named Entity Disambiguation based on Explicit Semantics

**Martin Jačala**
Supervised by Dr. Jozef Tvarožek

## Motivation

The constantly growing amount of human written textual content available on the web is a source of interesting and actual information about persons, organisations or places. One of the problems we face when analysing or querying in such content is the name ambiguity. The proper names in news articles comprise approximately 10% of text and many of them are highly ambiguous.

> Does the word **jaguar** mean the sports **car**, the **jungle animal** or something different? Which **Michael Jordan** does the text refer to?

In our work we propose an approach to answer such questions by disambiguating the named entities using explicit semantics extracted from a web-based corpora used as the background knowledge. We follow the Miller and Charles distributional hypothesis stating that similar entities appears in similar contexts even across multiple documents.

Find boundaries
↓
Extract candidate meanings
↓
Build document vectors
↓
Compute similarity

## Proposed Method

Our proposed method make use of the data provided by Wikipedia in various stages of the computation. Firstly, we use disambiguation pages, redirects and page titles to find any candidate meanings and associated pages for the analysed surface form. Then, we use Explicit Semantic Analysis to build vectors from analysed document and any of the candidate meaning documents. Finally, we compute the similarity between the vectors and rank all meanings according to the attained score.

We use ESA to construct an term-concept matrix (semantic space) where the matrix values are the tf-idf frequencies of the words extracted from given corpora. Traditional approaches, such as Latent Semantic Analysis use e.g. SVD to decrease the number of dimensions and discover hidden concepts. We assume that each Wikipedia article discuss one concept, therefore each dimension of ESA space corresponds to this "explicit concept".

## Evaluation and Results