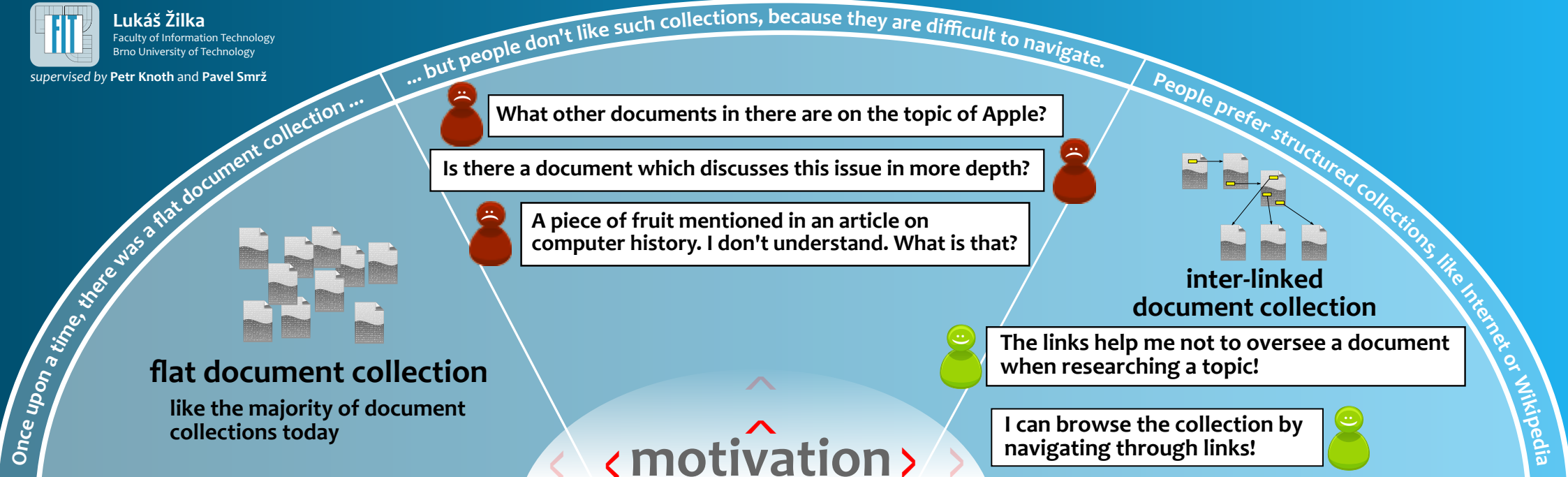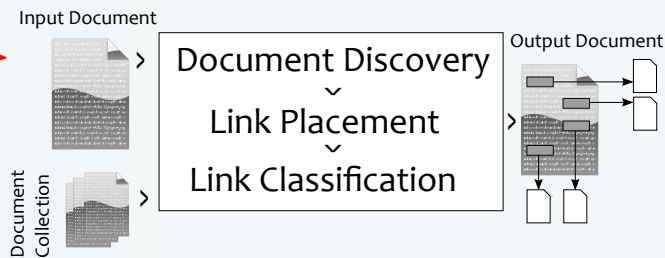# Using Explicit Semantic Analysis for Linking in Multi-lingual Document Collections

Lukáš Žilka
Faculty of Information Technology
Brno University of Technology

supervised by **Petr Knoth** and **Pavel Smrž**

Once upon a time, there was a flat document collection ...

... but people don't like such collections, because they are difficult to navigate.

People prefer structured collections, like Internet or Wikipedia

What other documents in there are on the topic of Apple?

Is there a document which discusses this issue in more depth?

A piece of fruit mentioned in an article on computer history. I don't understand. What is that?

flat document collection

like the majority of document collections today

inter-linked document collection

The links help me not to oversee a document when researching a topic!

I can browse the collection by navigating through links!

‹ **motivation** ›

ESADocLink is a system that automatically suggests links for documents in a multi-lingual document collection.

Input Document

Document Discovery
⌄
Link Placement
⌄
Link Classification

Output Document

Document Collection

ESA is the core of the system

1. Document Discovery

Select *documents* from the document collection that should be *linked* with the *input* document (different approaches):
a] semantically similar documents
b] from existing links in the most semantically similar documents
c] by average semantic representation
d] dimensions of the ESA vector

2. Link Placement

- look for a position in the input document, which should link to the target document
- searches the document for the possible matches with the title of the discovered documents

3. Link Classification

- using SVM with the following features:
  - relative link position,
  - link frequency,
  - generality,
  - ESA similarity (input vs. target, link text vs. target document, link text vs. target document title)

Explicit Semantic Analysis

iPad

	Concepts	ESA vector
	Tablet	0.2
	Screen	0.15
	Forest	0.001
	Toy	0.1