How machines help in finding the right career

Recommendation Systems

lmc

Value for customers

- find things that are interesting
- narrow down the set of choices
- help to explore the space of options
- discover new things
- entertainment

Value for provider

- unique personalized service for the customer
- increase trust and customer loyalty
- increase sales, click rates, conversion etc.
- opportunities for promotion, persuasion
- obtain more knowledge about customers

Basic Concepts

Data objects

- user
- item

User properties

- user_id
- name, work, skill

Relations between them

- detail view
- purchase

Item properties

- item_id
- title, description
- salary, availability

Two Major Approaches

Content-based recommenders

- user preferences and interests
- user profile
- matching with the item attributes

Collaborative filtering recommenders

- user-item interactions
- predicting future interactions

Two Major Approaches

Content-based recommenders

- user preferences and interests
- user profile
- matching with the item attributes

Collaborative filtering recommenders

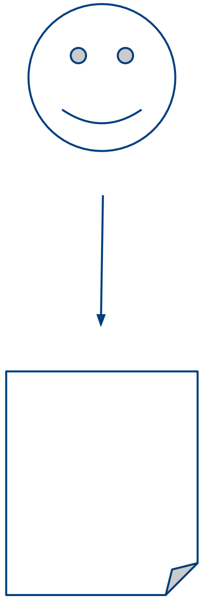
- user-item interactions
- predicting future interactions

Collaborative Filtering

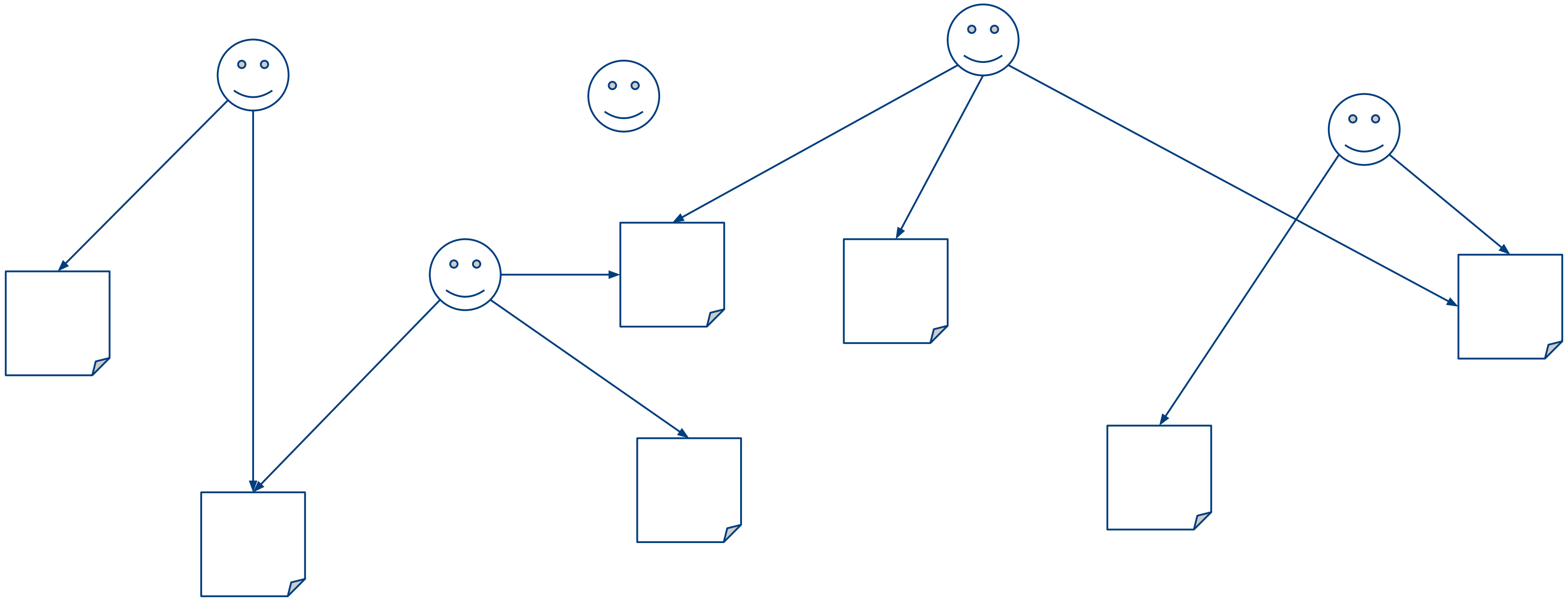
Collaborative Filtering

User-Item Interactions

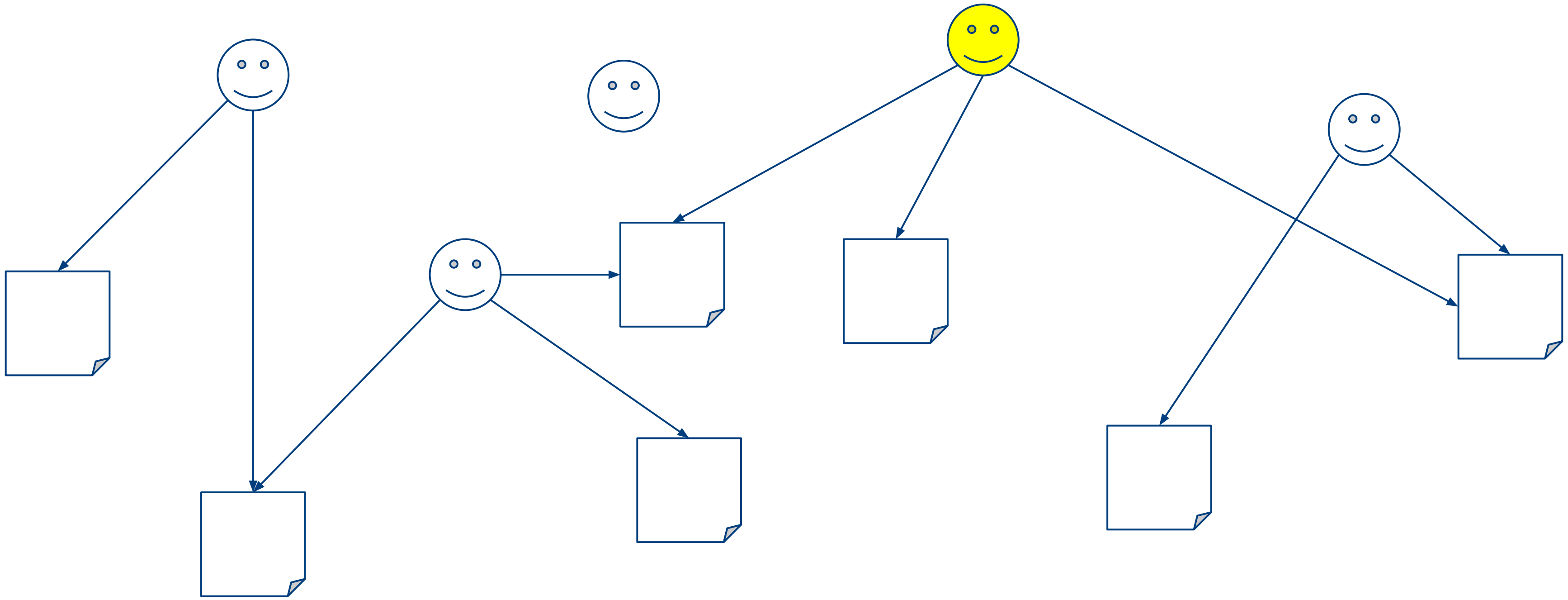
- explicit
  - users rate items
- implicit
  - detail view
  - cart addition
  - purchase



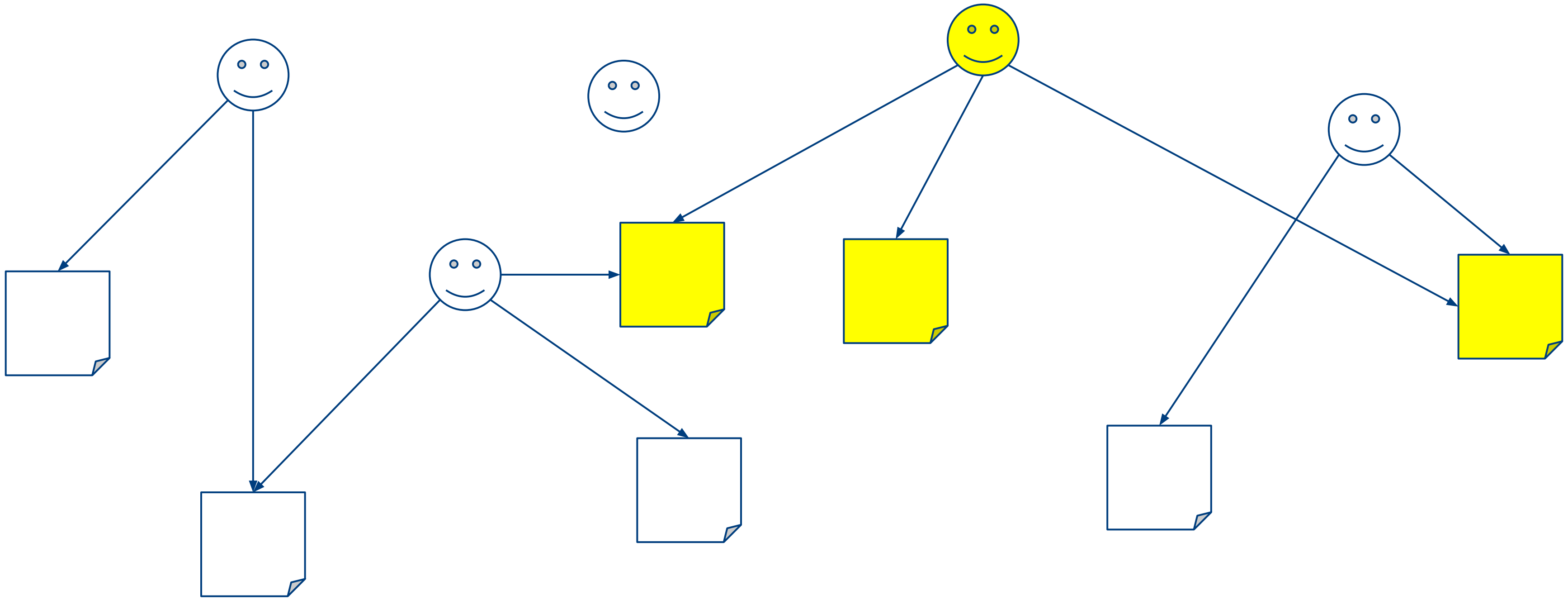
Collaborative Filtering | User-based Recommendation



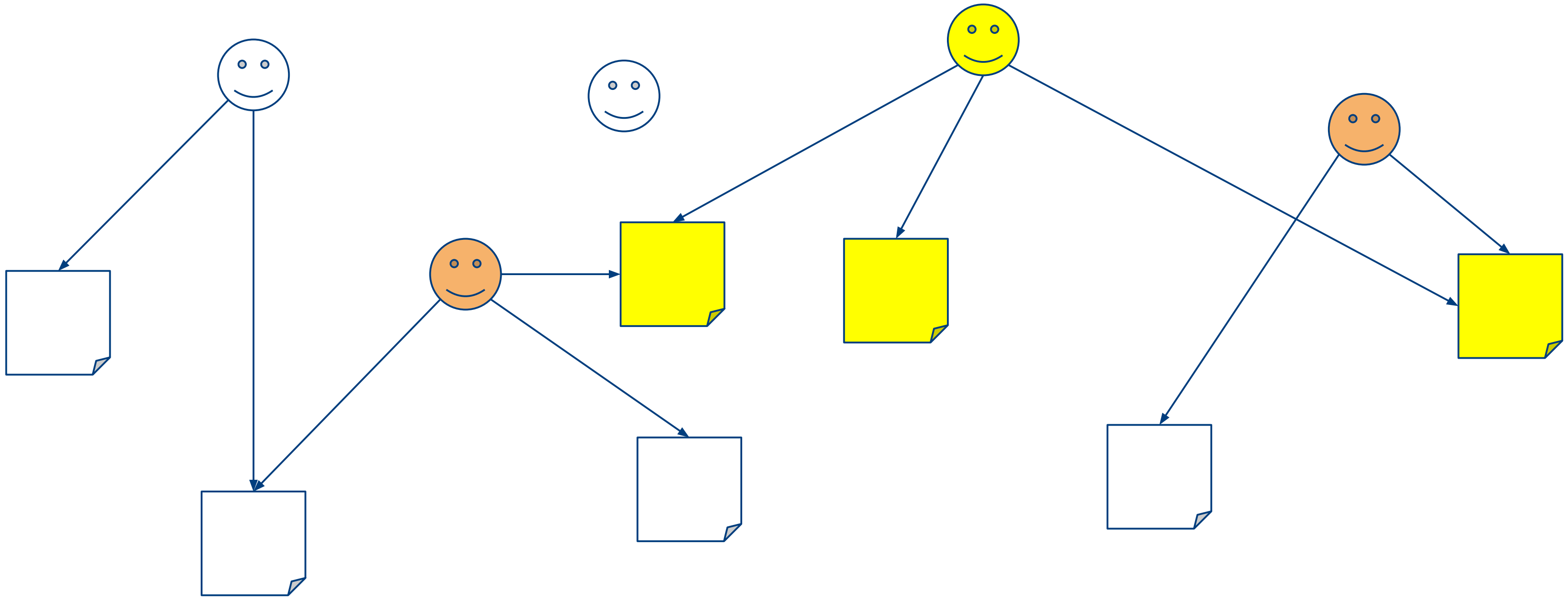
Collaborative Filtering | User-based Recommendation



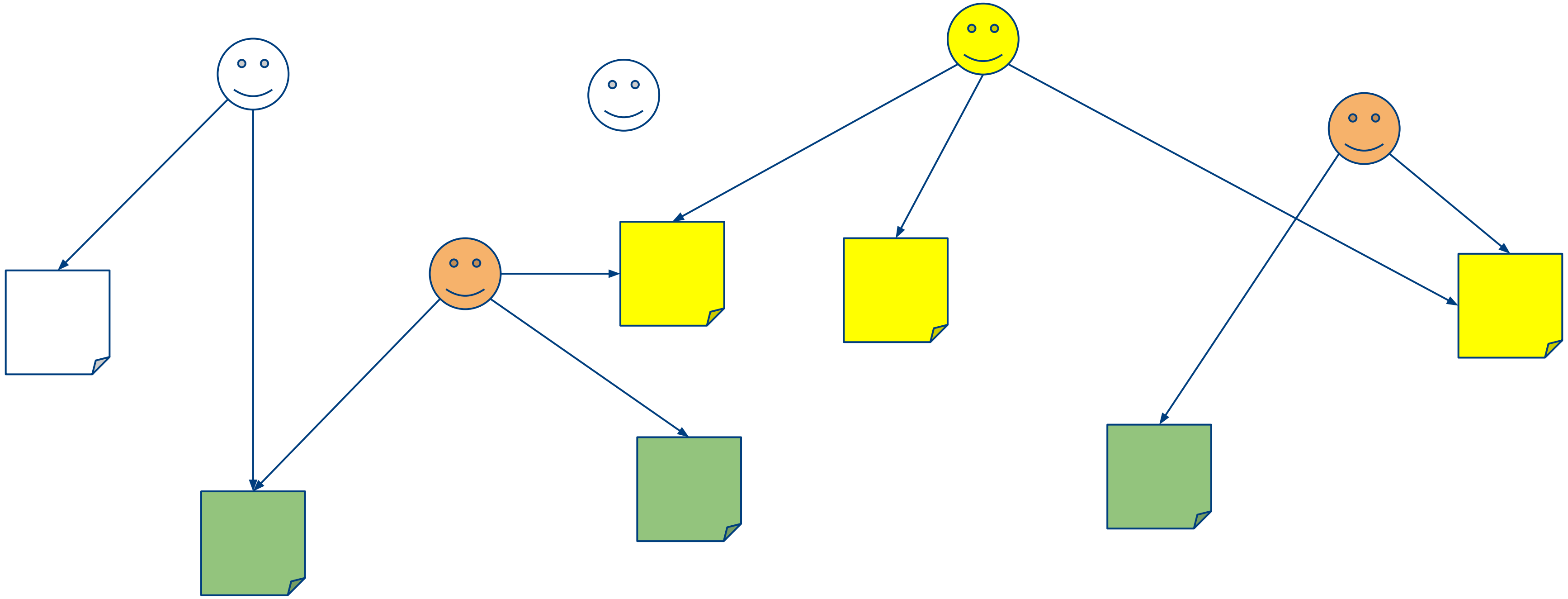
Collaborative Filtering | User-based Recommendation



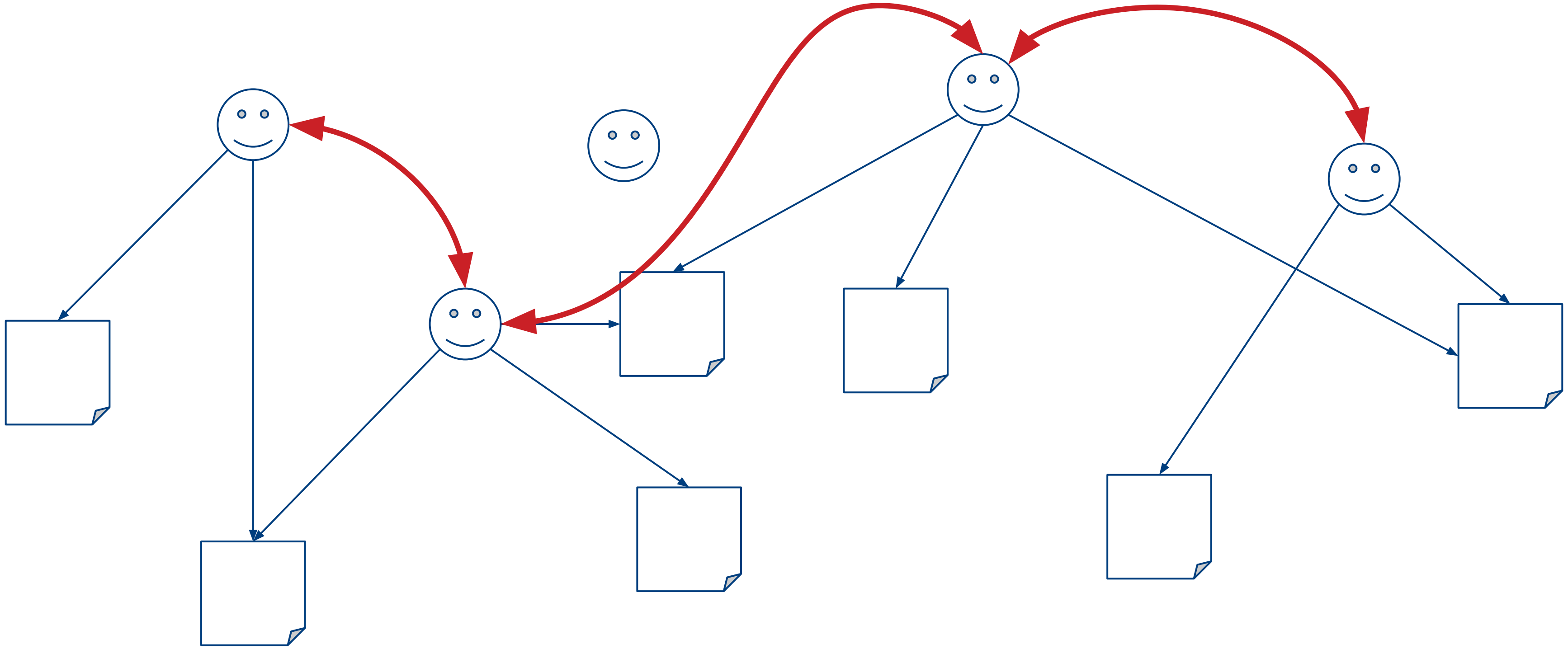
Collaborative Filtering | User-based Recommendation



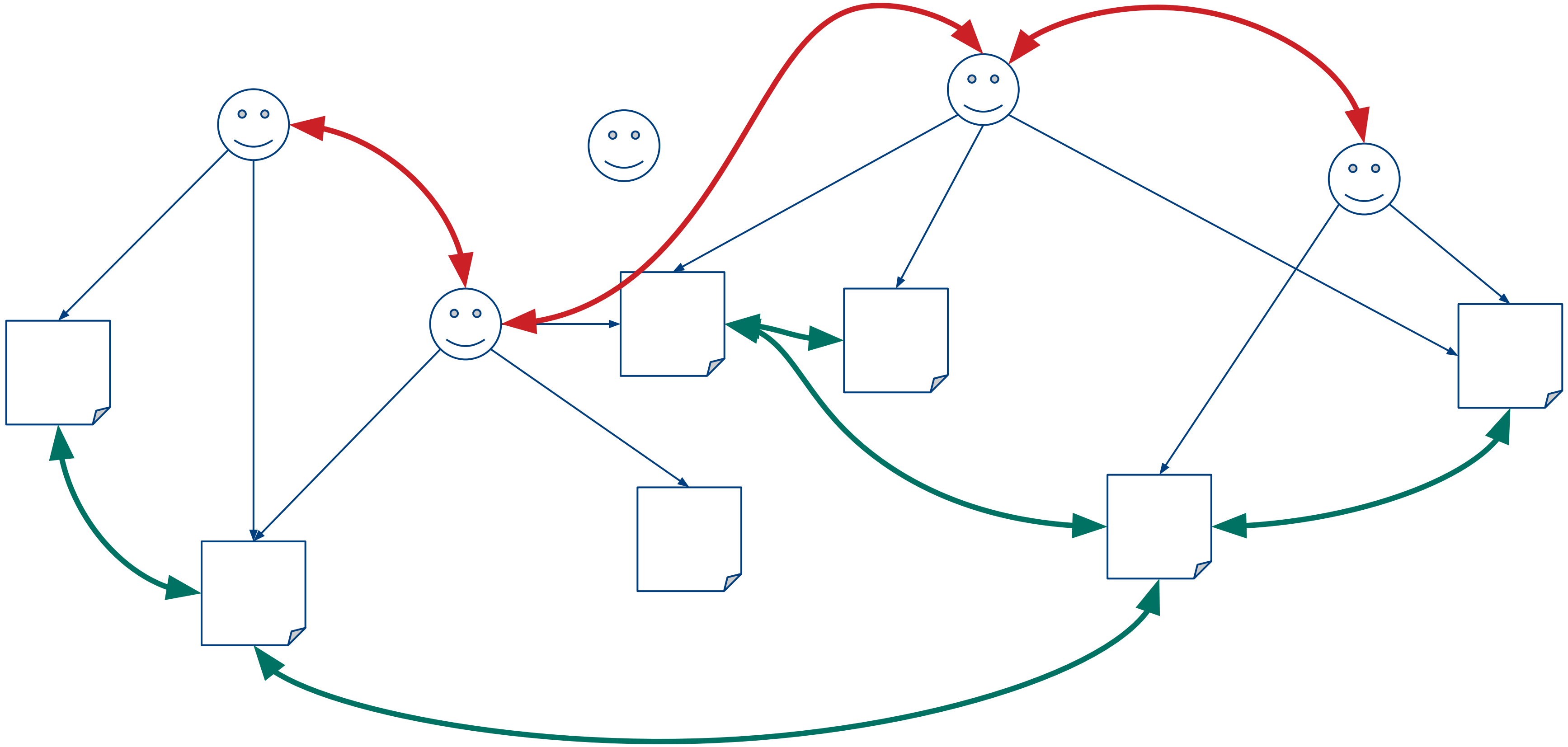
Collaborative Filtering | User-based Recommendation



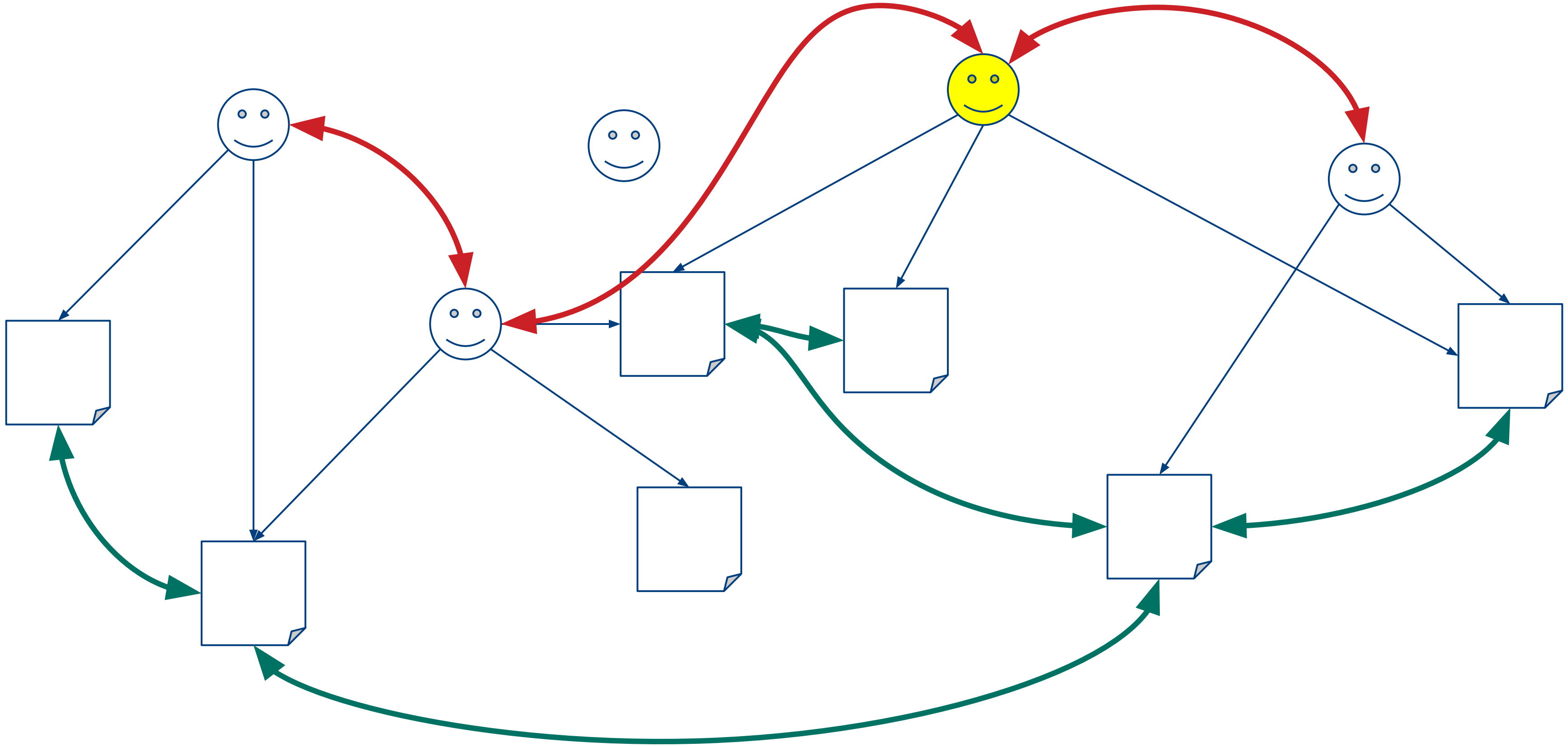
Collaborative Filtering | User Neighbours



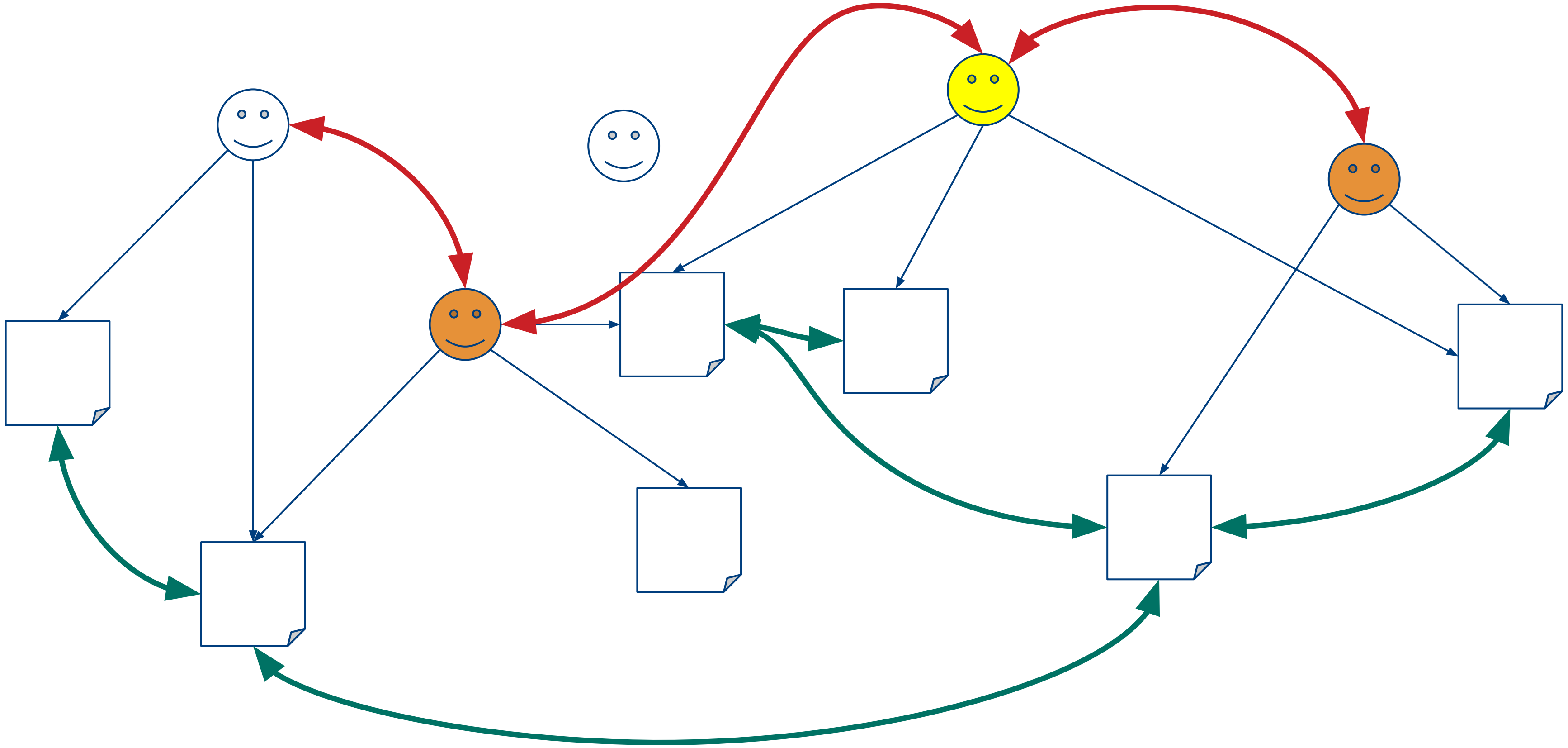
Collaborative Filtering | Item Neighbours



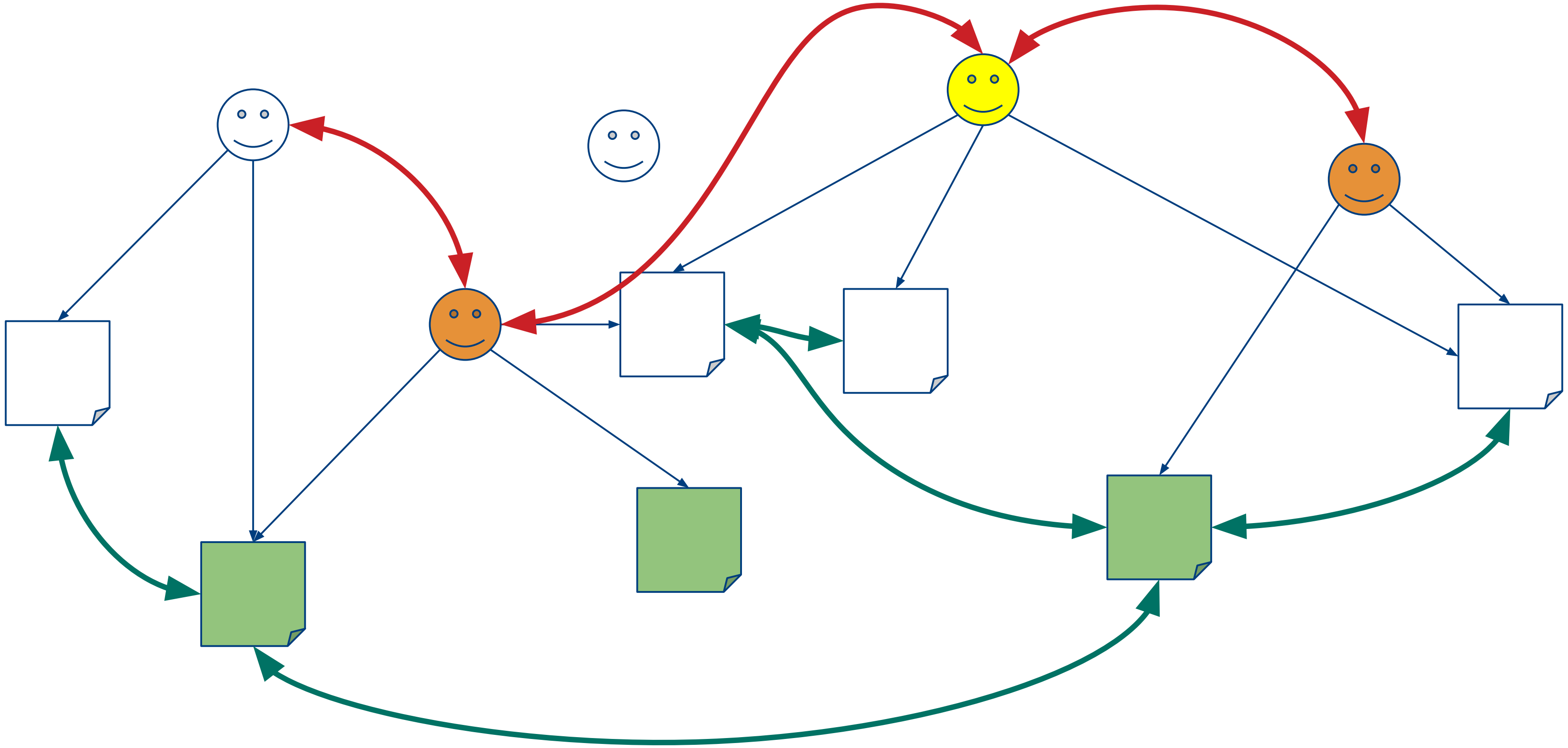
Collaborative Filtering | Neighbour-based Recommendation



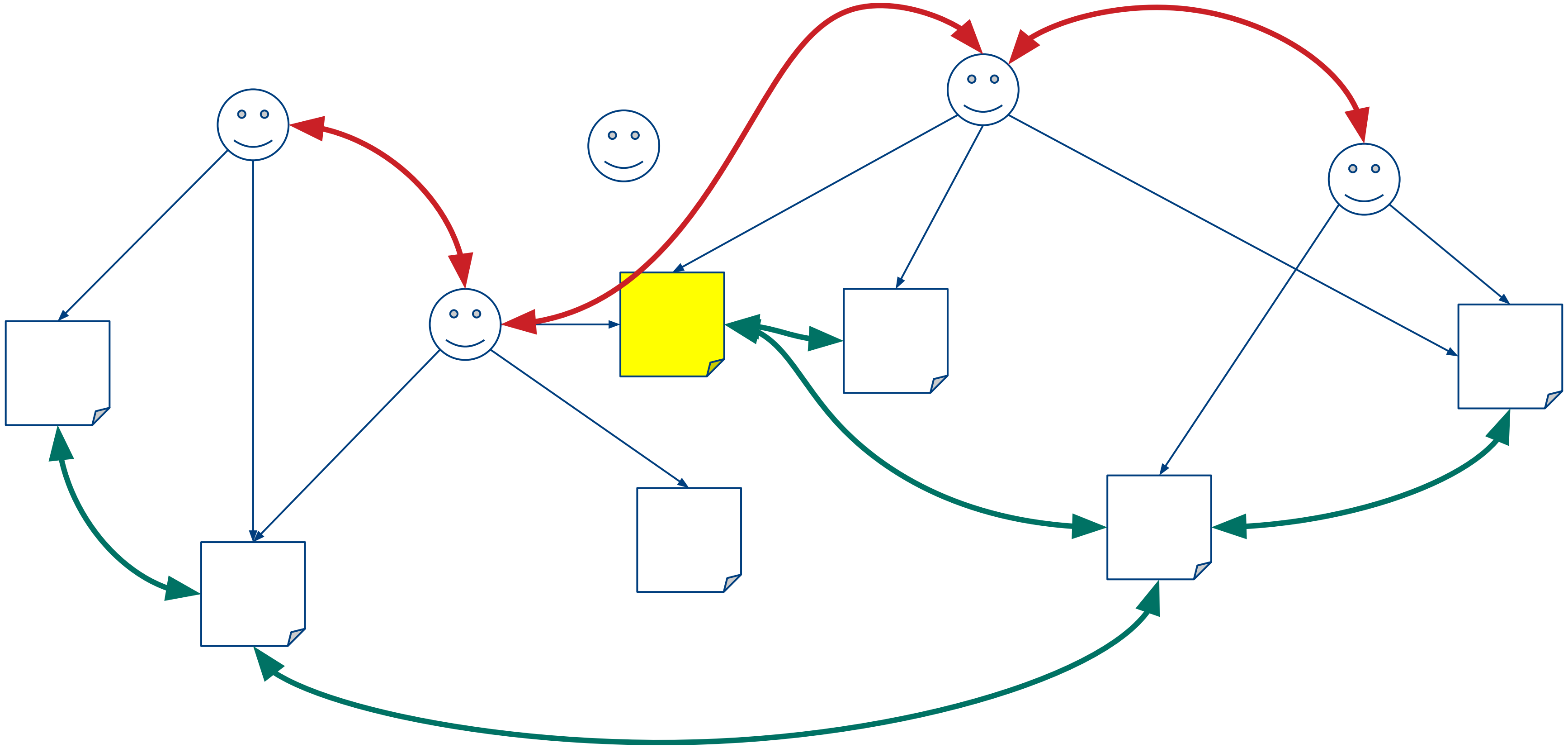
Collaborative Filtering | Neighbour-based Recommendation



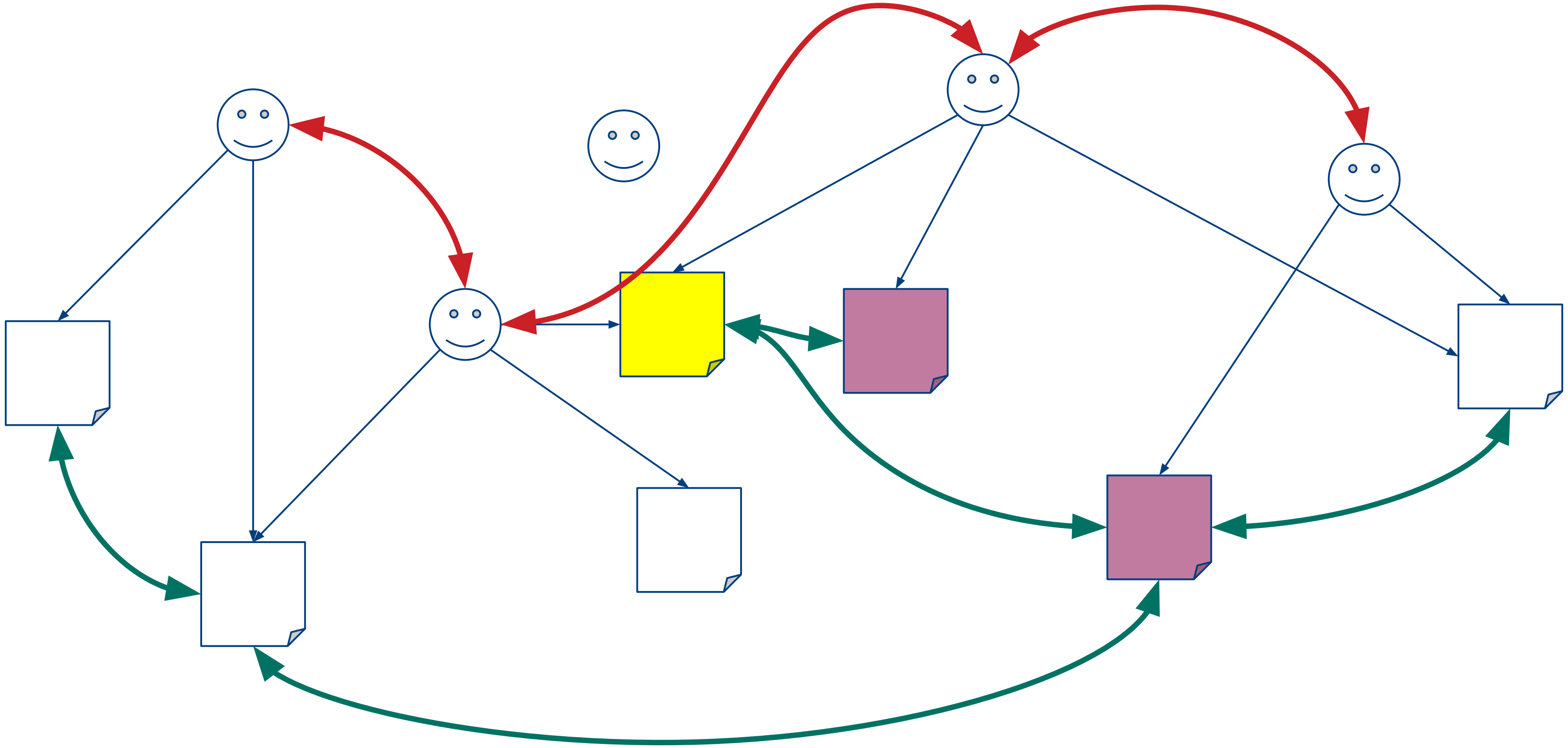
Collaborative Filtering | Neighbour-based Recommendation



Collaborative Filtering | Item-based Recommendation



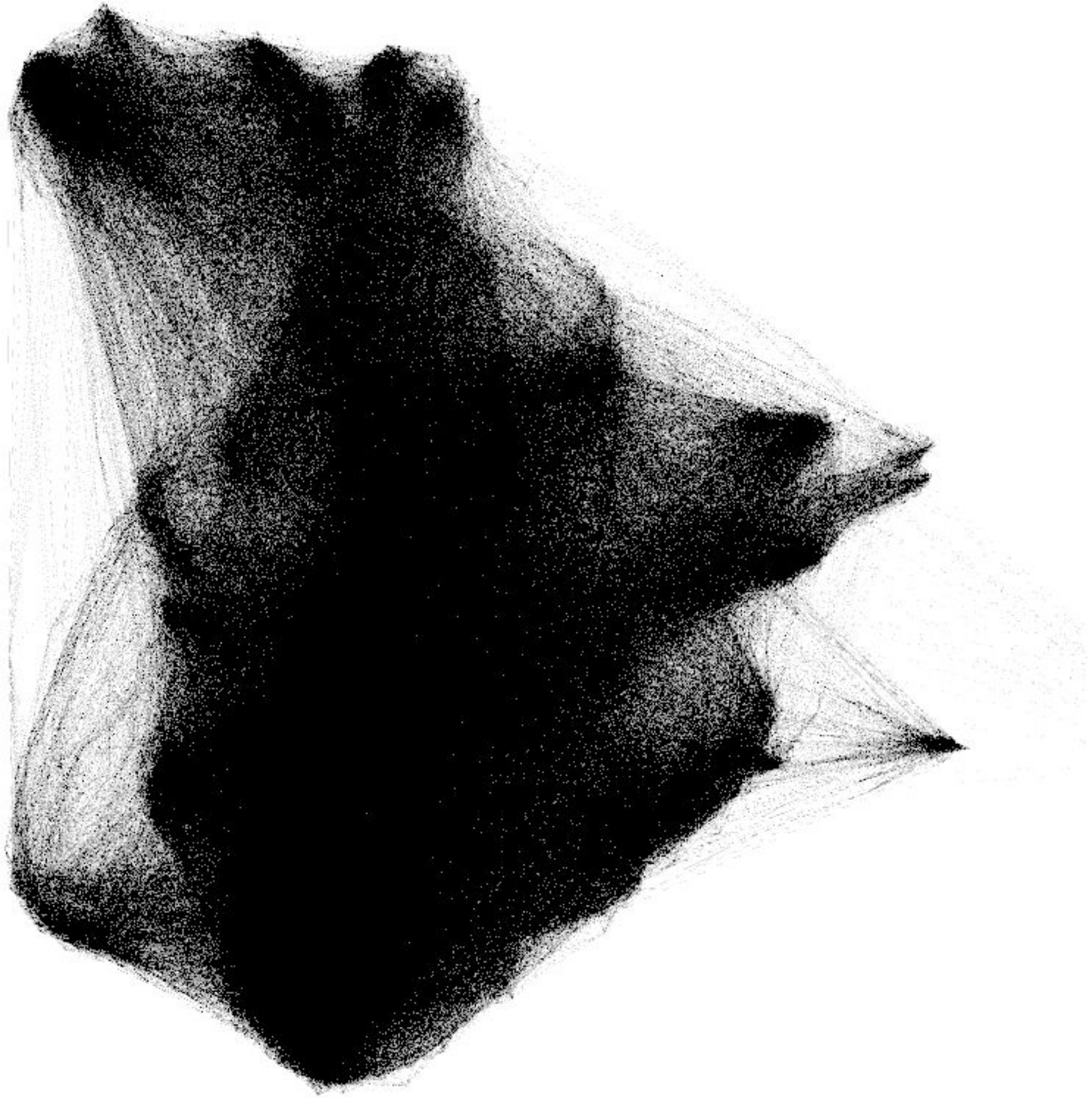
Collaborative Filtering | Item-based Recommendation



Collaborative Filtering

Well, in fact… :-)

- millions of users
- thousands of items



Text processing

The Document Parser

Document parser

- advanced **machine learning** techniques
- currently supported are **CVs and job offers** in **Czech language**
- framework for parsing **any type** of documents in **any language**

Document parser

Input:

- different formats (PDF, DOC, DOCX, ODT, TXT)
- different layouts
- (almost) free texts

Document parser

Output:

- structured data
- uniformed and normalized

Document parser

Personal

- Name
- Birthday
- Marital Status

Contacts

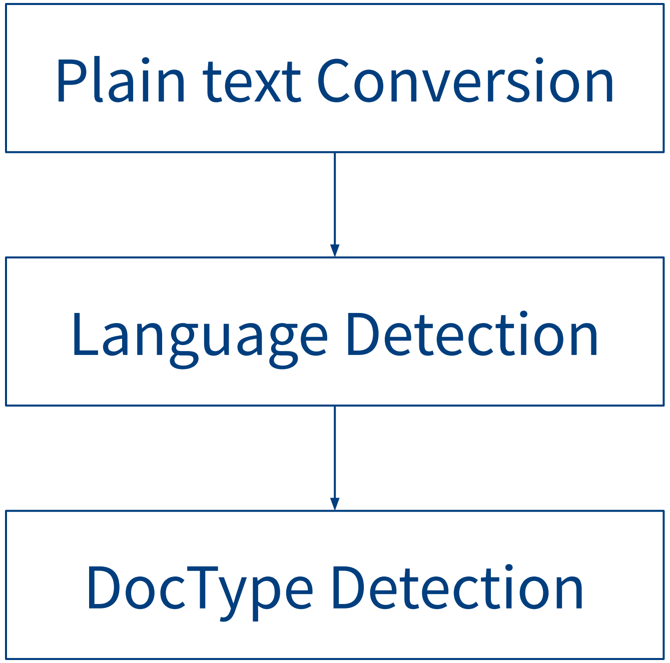
- Postal address
- Phones
- Emails

Work Experience

- From/to
- Company
- Job title

Document Parsing Strategy

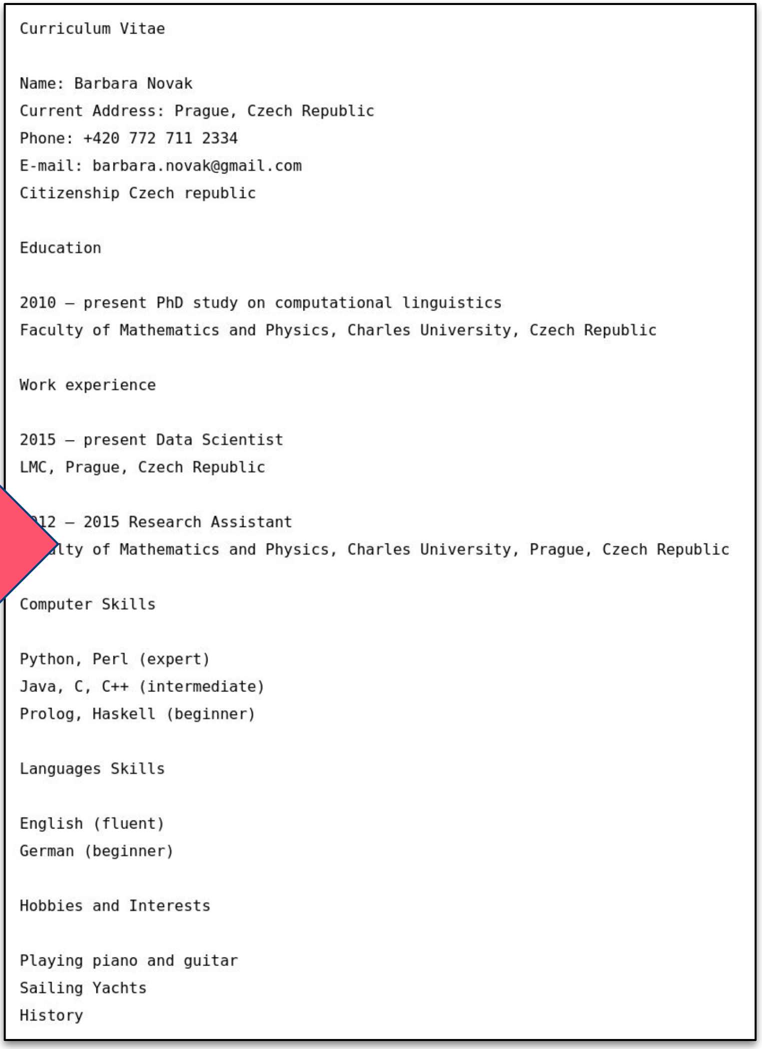
Document parser Processing Pipeline



Plain text Conversion

Language Detection

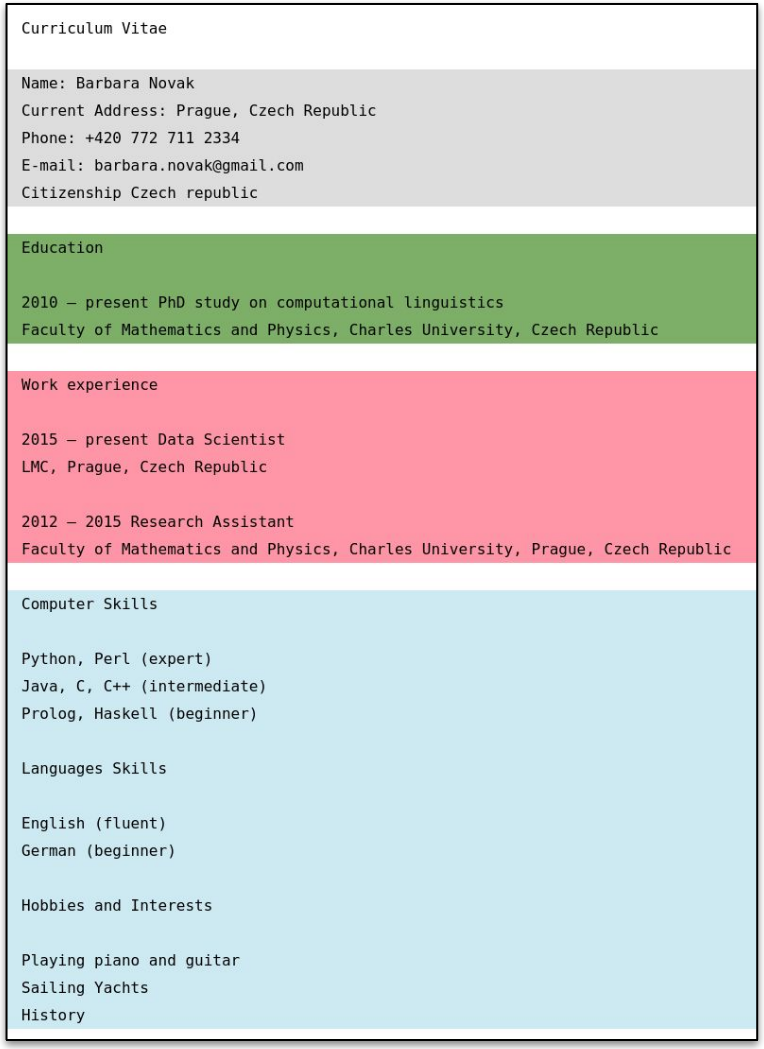
DocType Detection



```
Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 – present PhD study on computational linguistics
Faculty of Mathematics and Physics, Charles University, Czech Republic

Work experience

2015 – present Data Scientist
LMC, Prague, Czech Republic

12 – 2015 Research Assistant
lty of Mathematics and Physics, Charles University, Prague, Czech Republic

Computer Skills

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

Languages Skills

English (fluent)
German (beginner)

Hobbies and Interests

Playing piano and guitar
Sailing Yachts
History
```

Document parser Processing Pipeline

Section Detection



Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 – present PhD study on computational linguistics
Faculty of Mathematics and Physics, Charles University, Czech Republic

Work experience

2015 – present Data Scientist
LMC, Prague, Czech Republic

2012 – 2015 Research Assistant
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Computer Skills

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

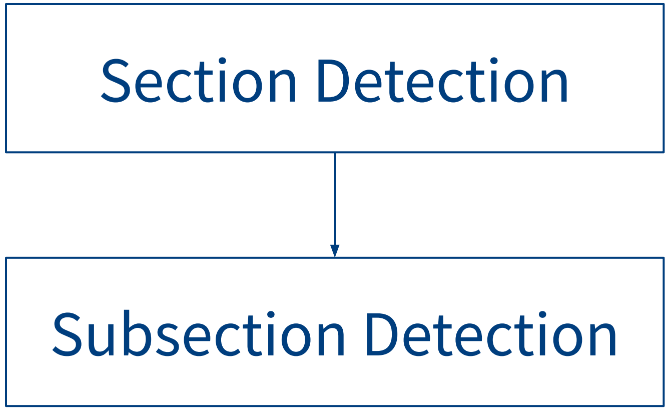
Languages Skills

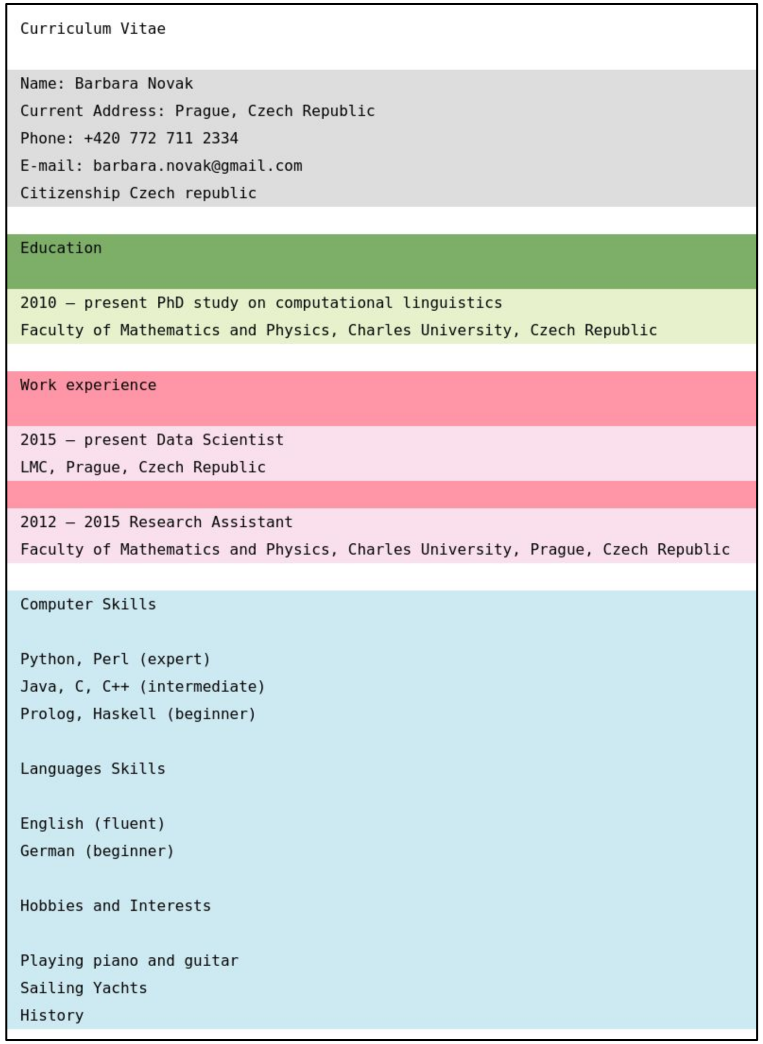
English (fluent)
German (beginner)

Hobbies and Interests

Playing piano and guitar
Sailing Yachts
History

Document parser Processing Pipeline





Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 – present PhD study on computational linguistics
Faculty of Mathematics and Physics, Charles University, Czech Republic

Work experience

2015 – present Data Scientist
LMC, Prague, Czech Republic

2012 – 2015 Research Assistant
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Computer Skills

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

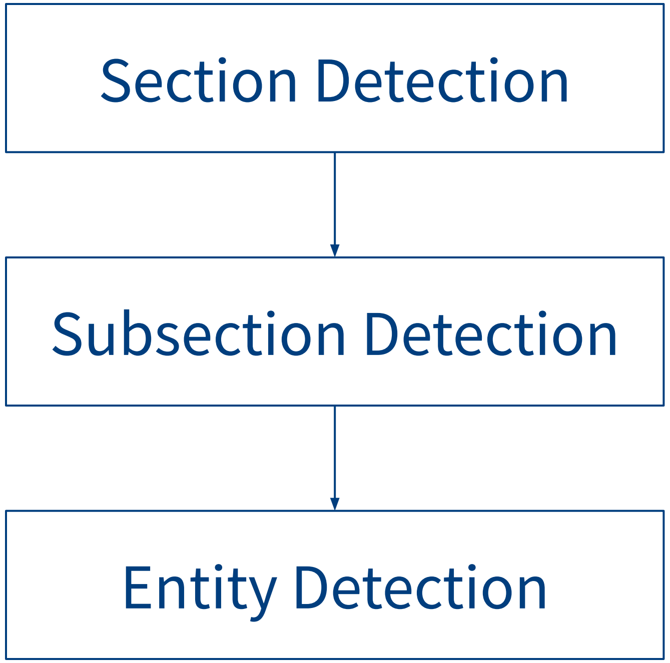
Languages Skills

English (fluent)
German (beginner)

Hobbies and Interests

Playing piano and guitar
Sailing Yachts
History

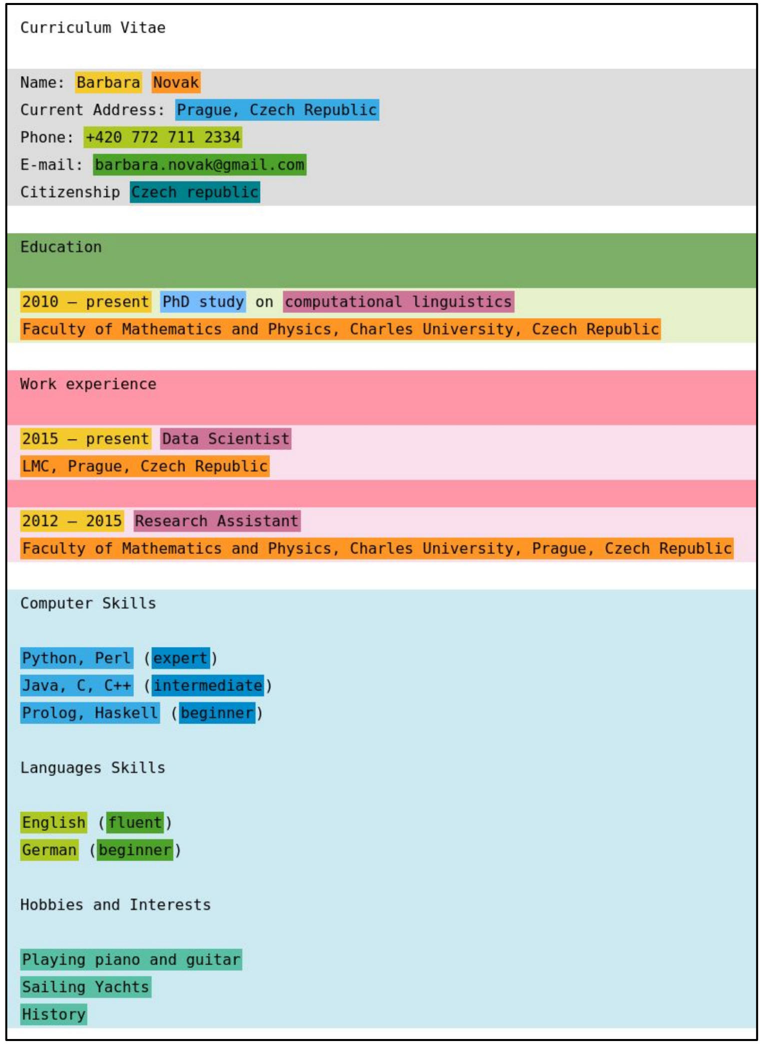
Document parser Processing Pipeline



Section Detection

Subsection Detection

Entity Detection



Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 – present PhD study on computational linguistics
Faculty of Mathematics and Physics, Charles University, Czech Republic

Work experience

2015 – present Data Scientist
LMC, Prague, Czech Republic

2012 – 2015 Research Assistant
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Computer Skills

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

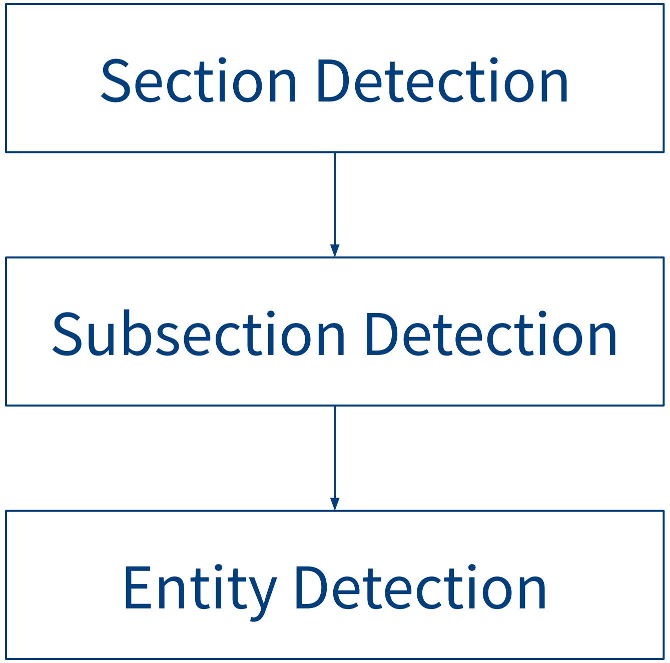
Languages Skills

English (fluent)
German (beginner)

Hobbies and Interests

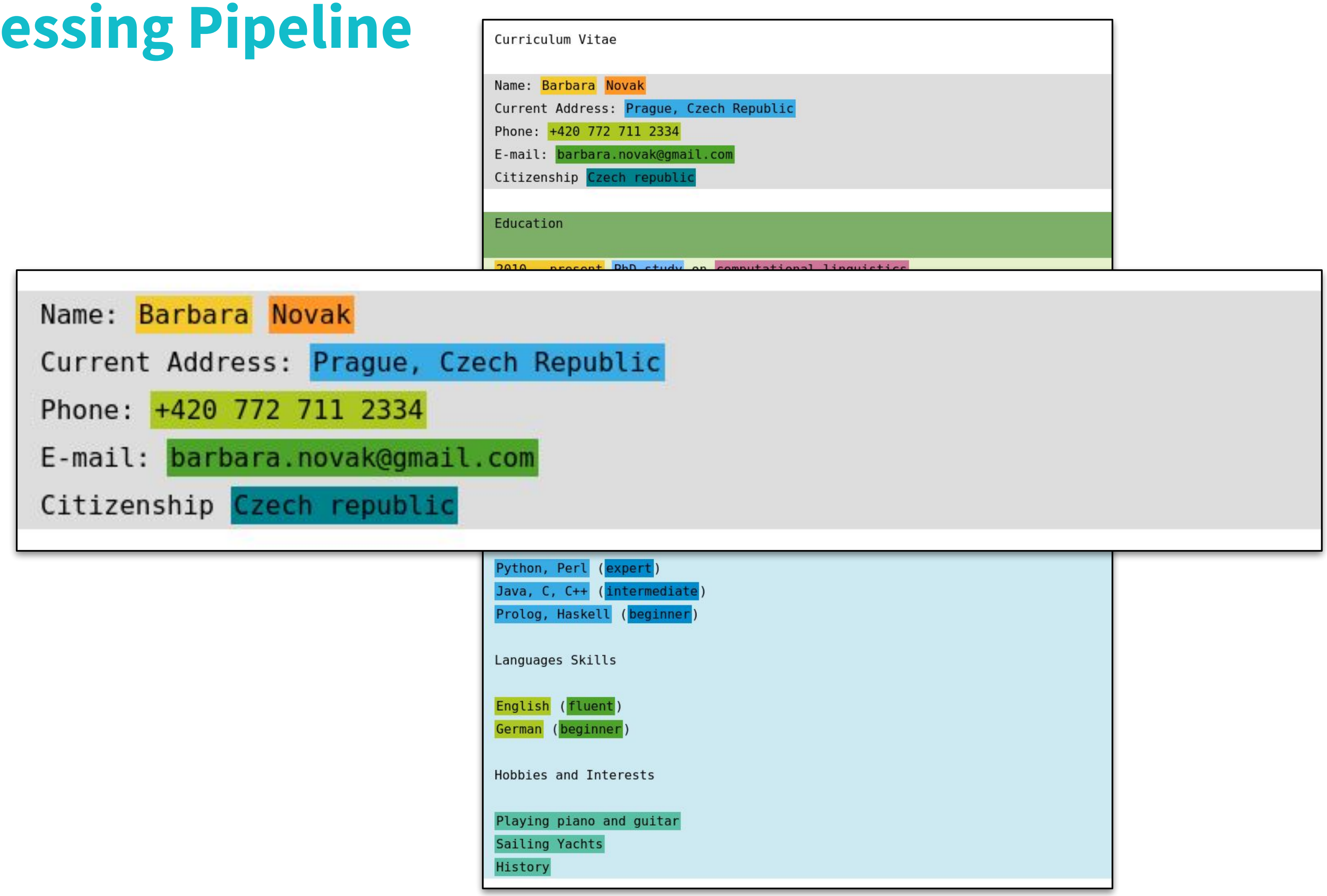
Playing piano and guitar
Sailing Yachts
History

Document parser Processing Pipeline


Section Detection

Subsection Detection

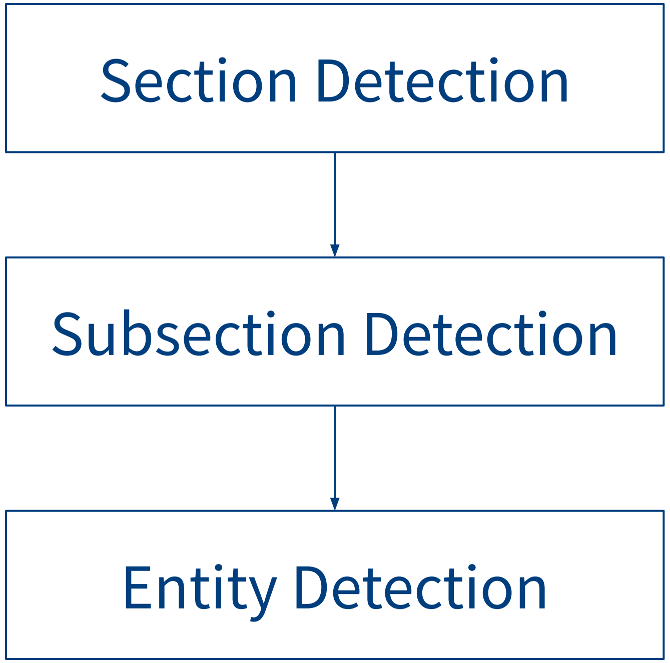
Entity Detection

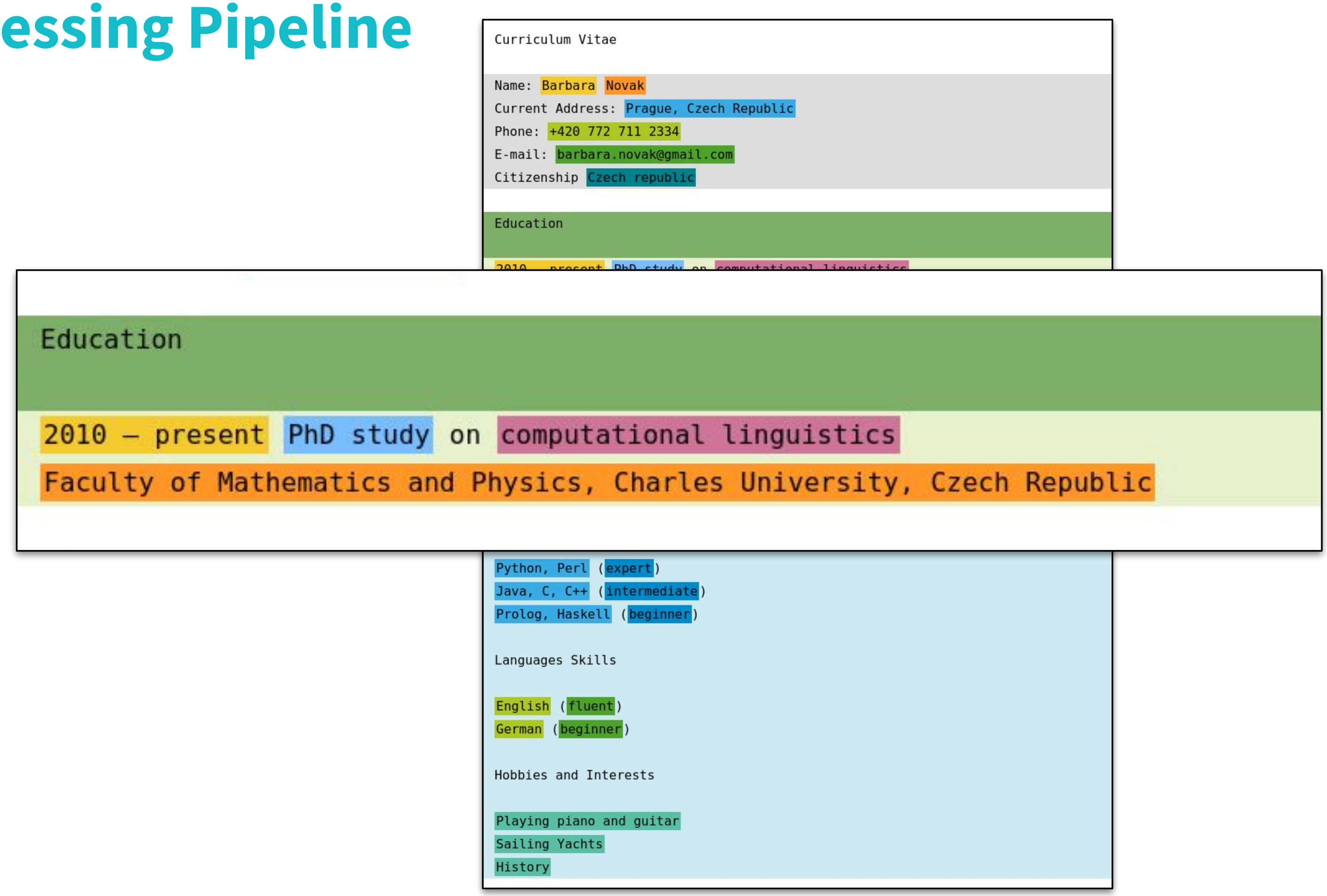

```
Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 – present PhD study on computational linguistics
```

```
Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic
```

```
Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

Languages Skills

English (fluent)
German (beginner)

Hobbies and Interests

Playing piano and guitar
Sailing Yachts
History
```

Document parser Processing Pipeline





Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 – present PhD study on computational linguistics
Faculty of Mathematics and Physics, Charles University, Czech Republic

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

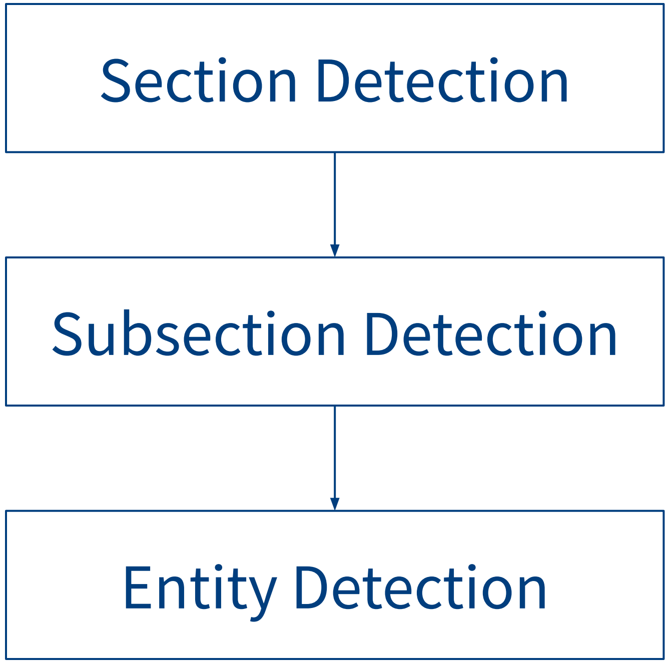
Languages Skills

English (fluent)
German (beginner)

Hobbies and Interests

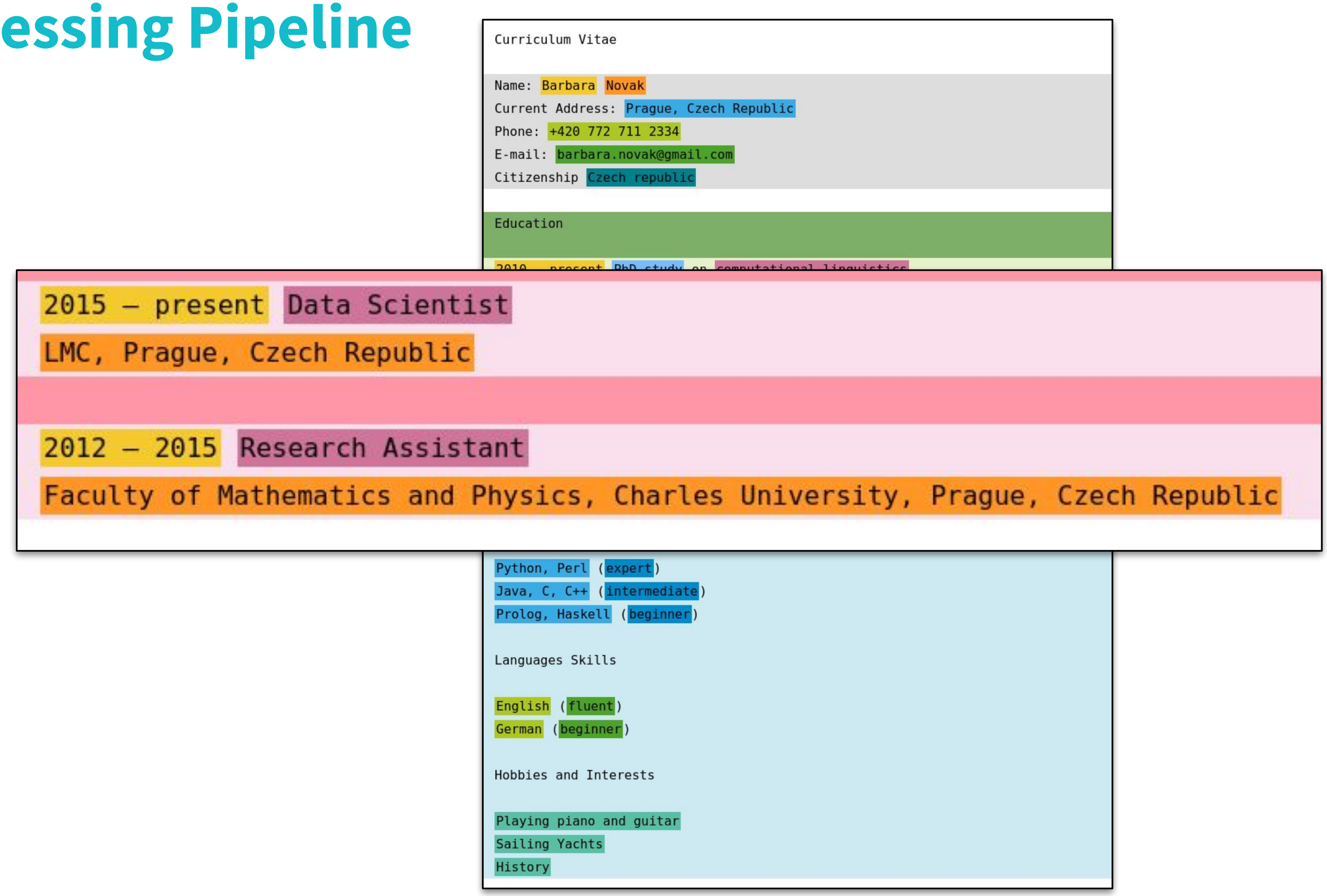
Playing piano and guitar
Sailing Yachts
History

Document parser Processing Pipeline


Section Detection

Subsection Detection

Entity Detection


Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 – present PhD study on computational linguistics

2015 – present Data Scientist
LMC, Prague, Czech Republic

2012 – 2015 Research Assistant
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

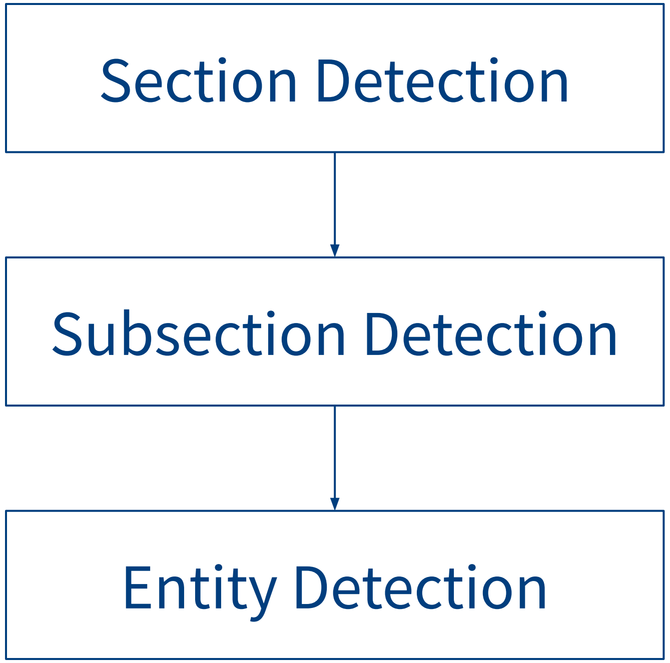
Languages Skills

English (fluent)
German (beginner)

Hobbies and Interests

Playing piano and guitar
Sailing Yachts
History

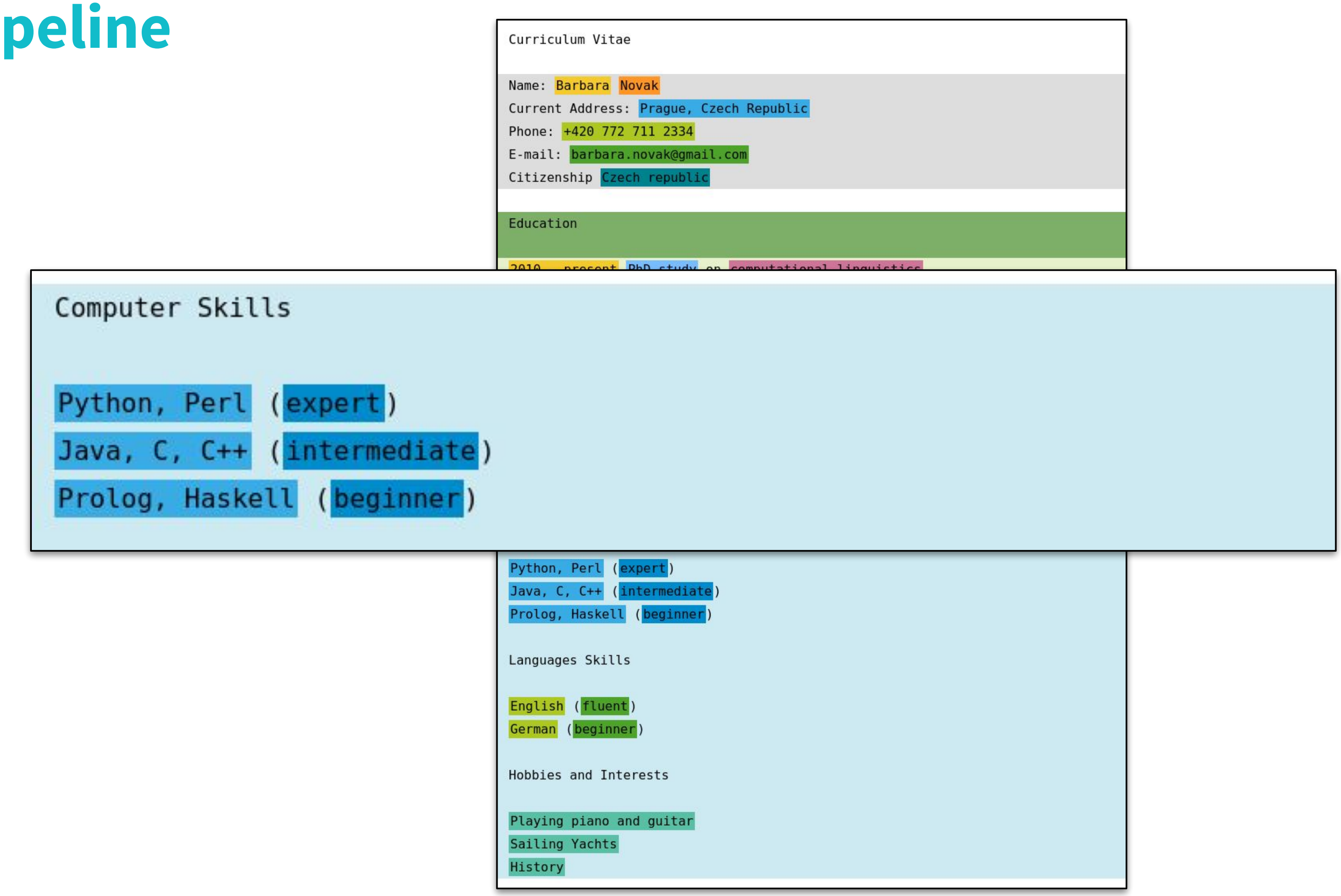
Barbara Processing Pipeline



Section Detection

Subsection Detection

Entity Detection



Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 – present PhD study on computational linguistics

Computer Skills

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

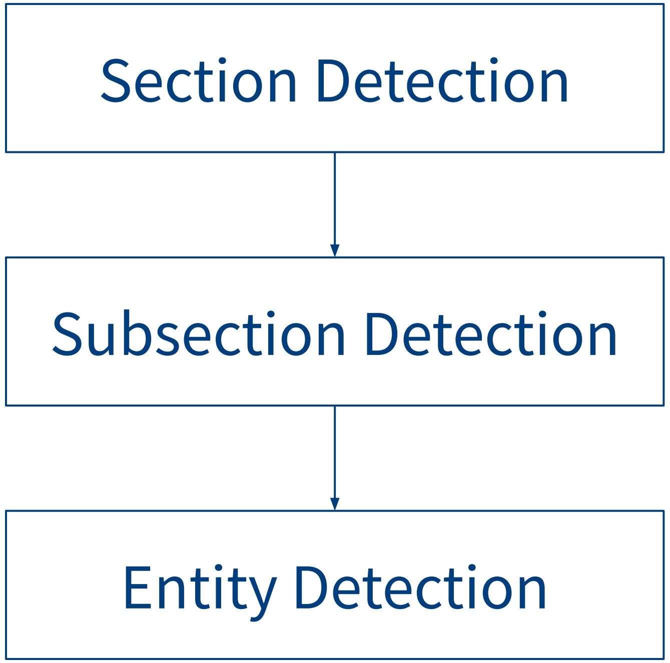
Languages Skills

English (fluent)
German (beginner)

Hobbies and Interests

Playing piano and guitar
Sailing Yachts
History

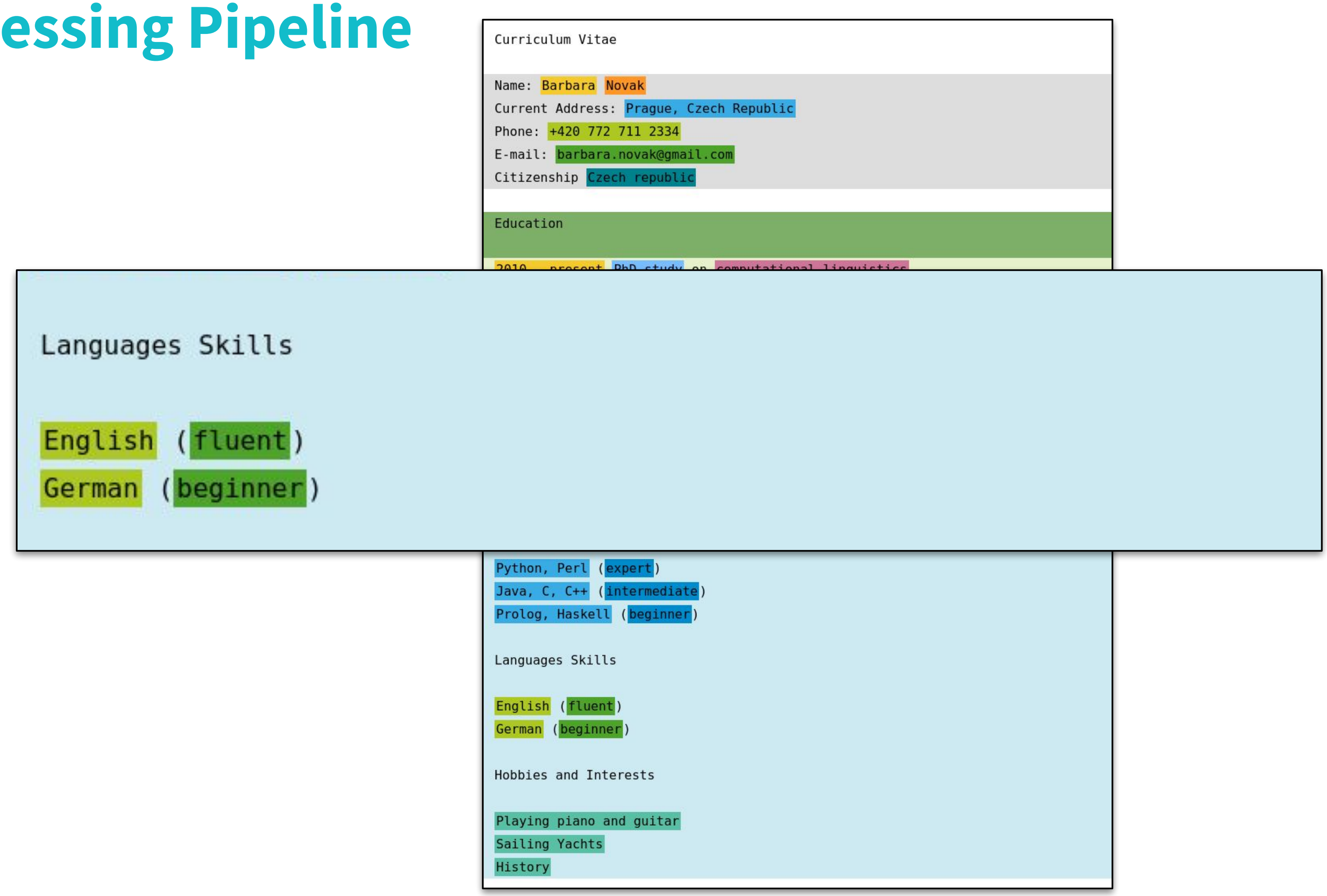
Document parser Processing Pipeline



Section Detection

Subsection Detection

Entity Detection



Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 – present PhD study on computational linguistics

Languages Skills

English (fluent)
German (beginner)

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

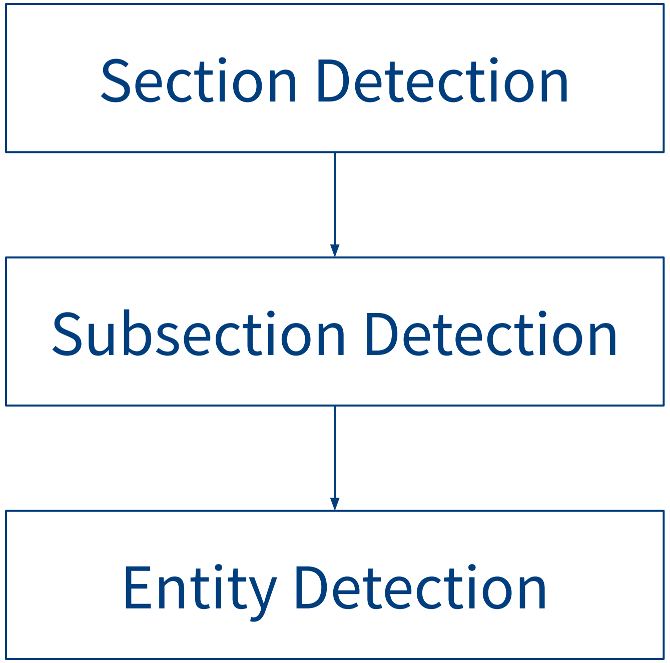
Languages Skills

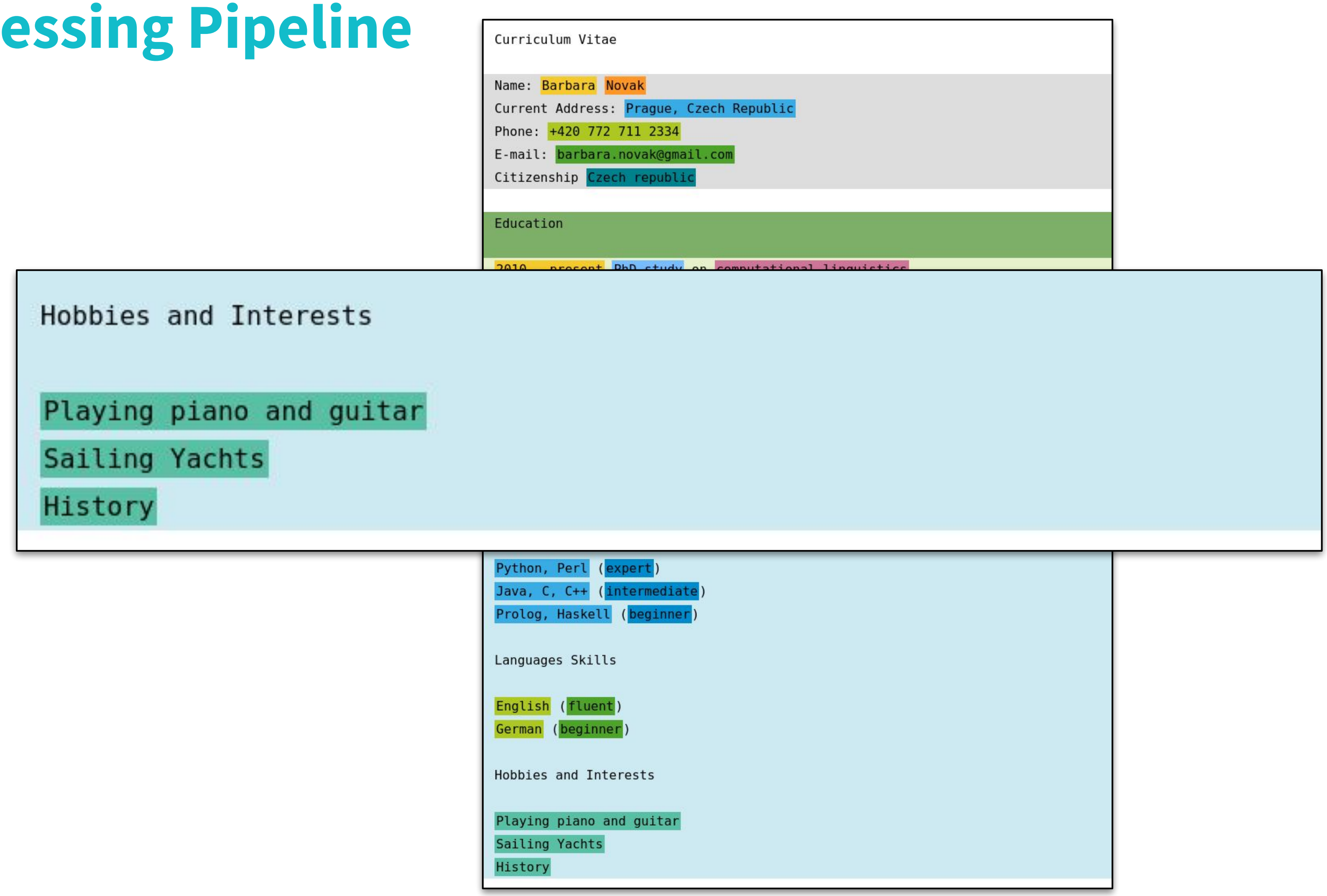
English (fluent)
German (beginner)

Hobbies and Interests

Playing piano and guitar
Sailing Yachts
History

Document parser Processing Pipeline





Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 – present PhD study on computational linguistics

Hobbies and Interests

Playing piano and guitar
Sailing Yachts
History

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

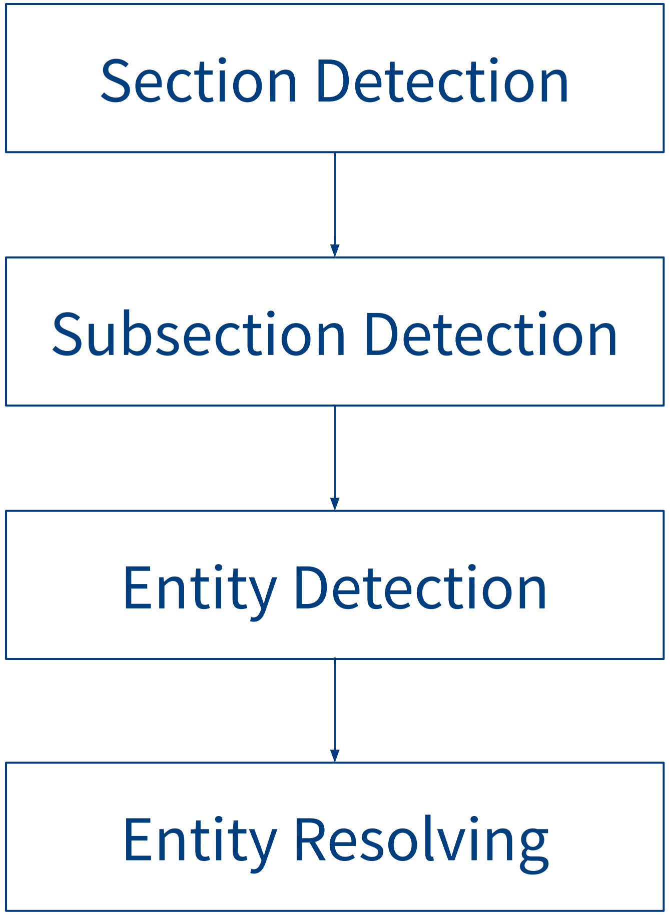
Languages Skills

English (fluent)
German (beginner)

Hobbies and Interests

Playing piano and guitar
Sailing Yachts
History

Document parser Processing Pipeline

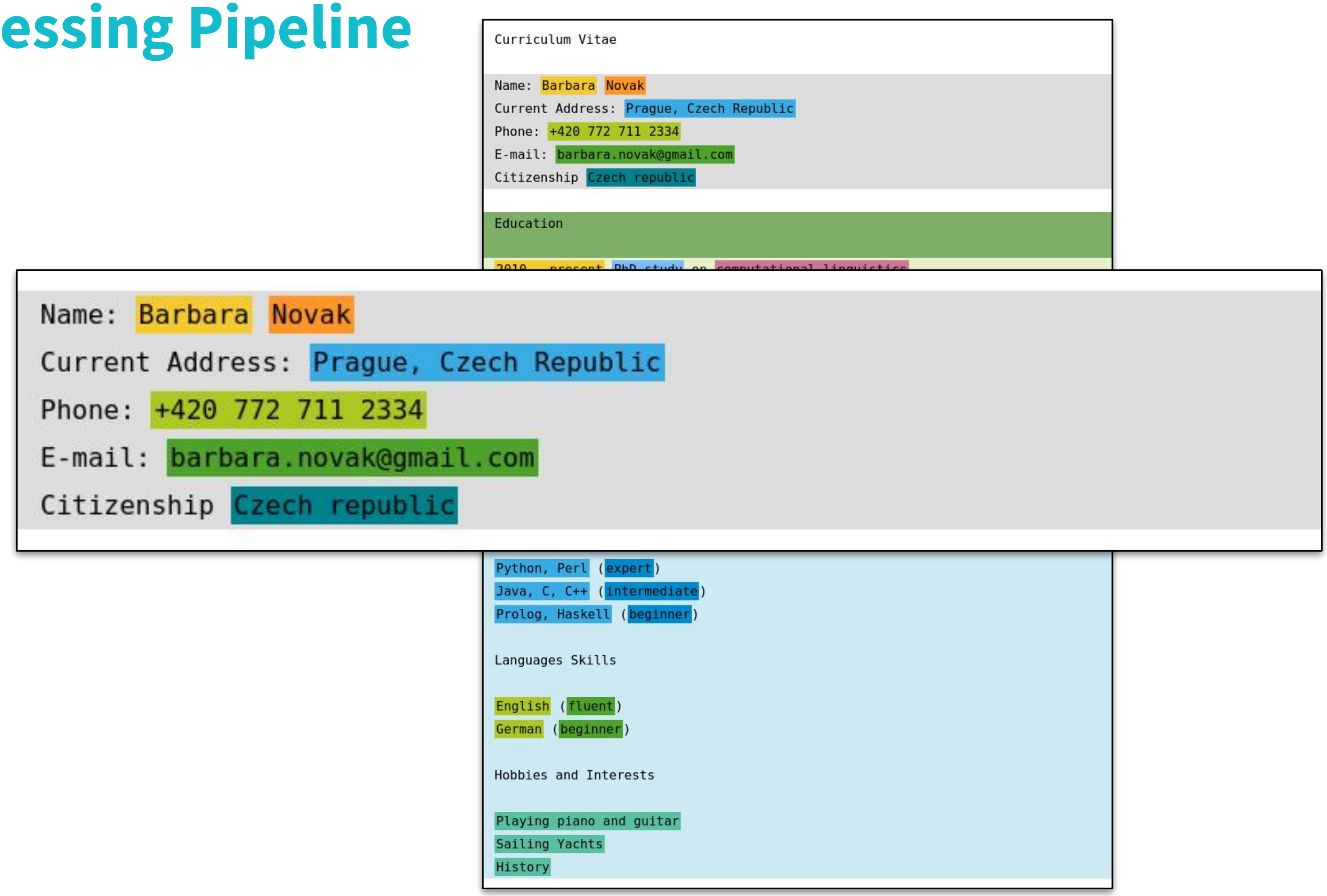


Section Detection

Subsection Detection

Entity Detection

Entity Resolving



Curriculum Vitae

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Education

2010 - present PhD study on computational linguistics

Name: Barbara Novak
Current Address: Prague, Czech Republic
Phone: +420 772 711 2334
E-mail: barbara.novak@gmail.com
Citizenship Czech republic

Python, Perl (expert)
Java, C, C++ (intermediate)
Prolog, Haskell (beginner)

Languages Skills

English (fluent)
German (beginner)

Hobbies and Interests

Playing piano and guitar
Sailing Yachts
History

Document parser ML Framework

Document parser Machine Learning Framework

Divide and Conquer Strategy

Sequence Labeling Models

- (sub)sections detectors
- entity detectors (one per section)

Normalization

- simple rules

Document parser Machine Learning Framework

Training from Examples

- language and type specific
- cca 10k documents for **Czech CVs** → reliable performance (>95%)

Document parser Machine Learning Framework

Manual Document Annotation

- our own web application for annotators

Data Quality

- annotation manual (20 A4 pages with instructions for Czech CVs)
- annotators selection
- on-line helpdesk for questions

Document parser Performance on Czech CVs

Detector	Performance
Personal Section Detection	97 %
Experience Section Detection	98 %
Work Experience Subsection Detection	96 %
Personal Entity Detection	94 %
Work Experience Entity Detection	93 %

Question time.

Diar Masri
Diar.Masri@lmc.eu