Modeling, Measuring and Exploiting Concept Drift in the Labour Market Domain

Panos Alexopoulos & Spyretta Leivaditi

2nd Drift-a-LOD Workshop,

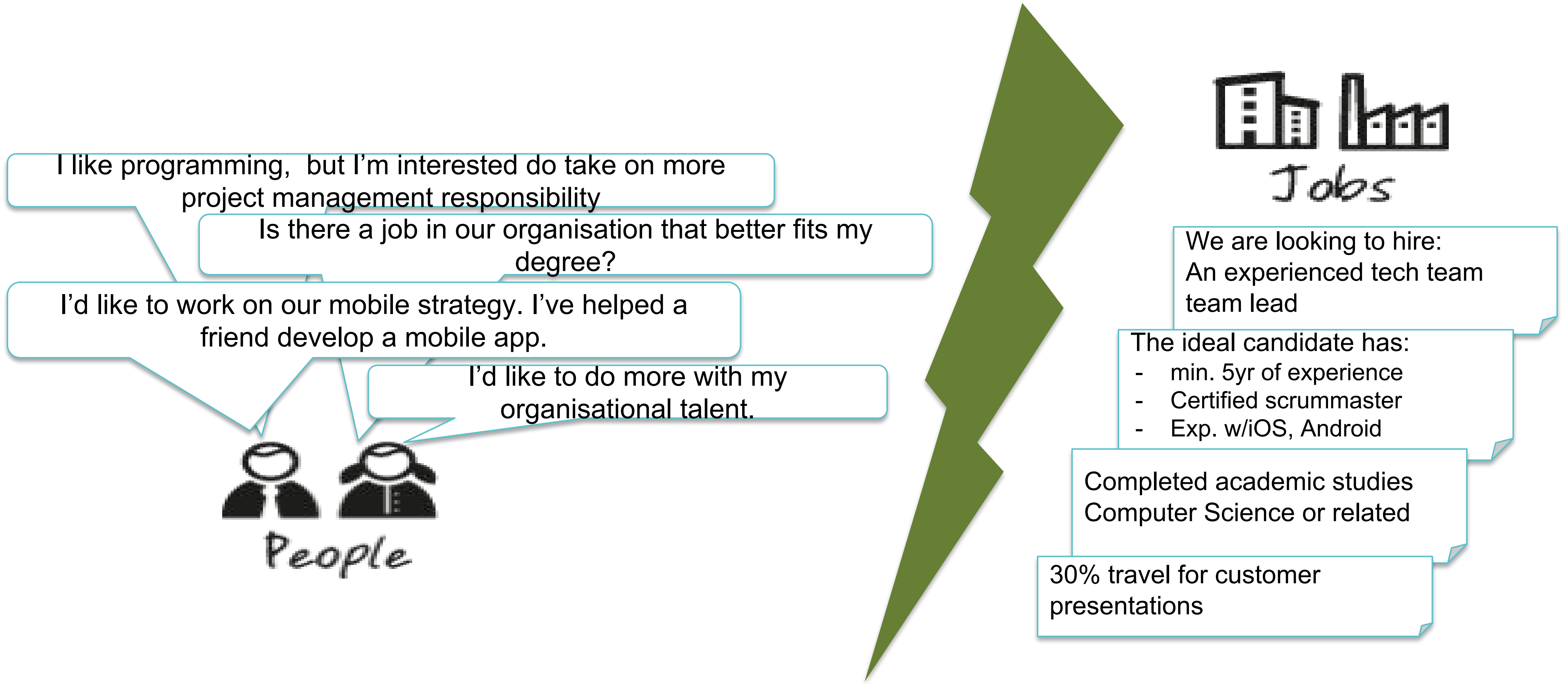
Amsterdam, Netherlands, September 11th 2017

textkernel

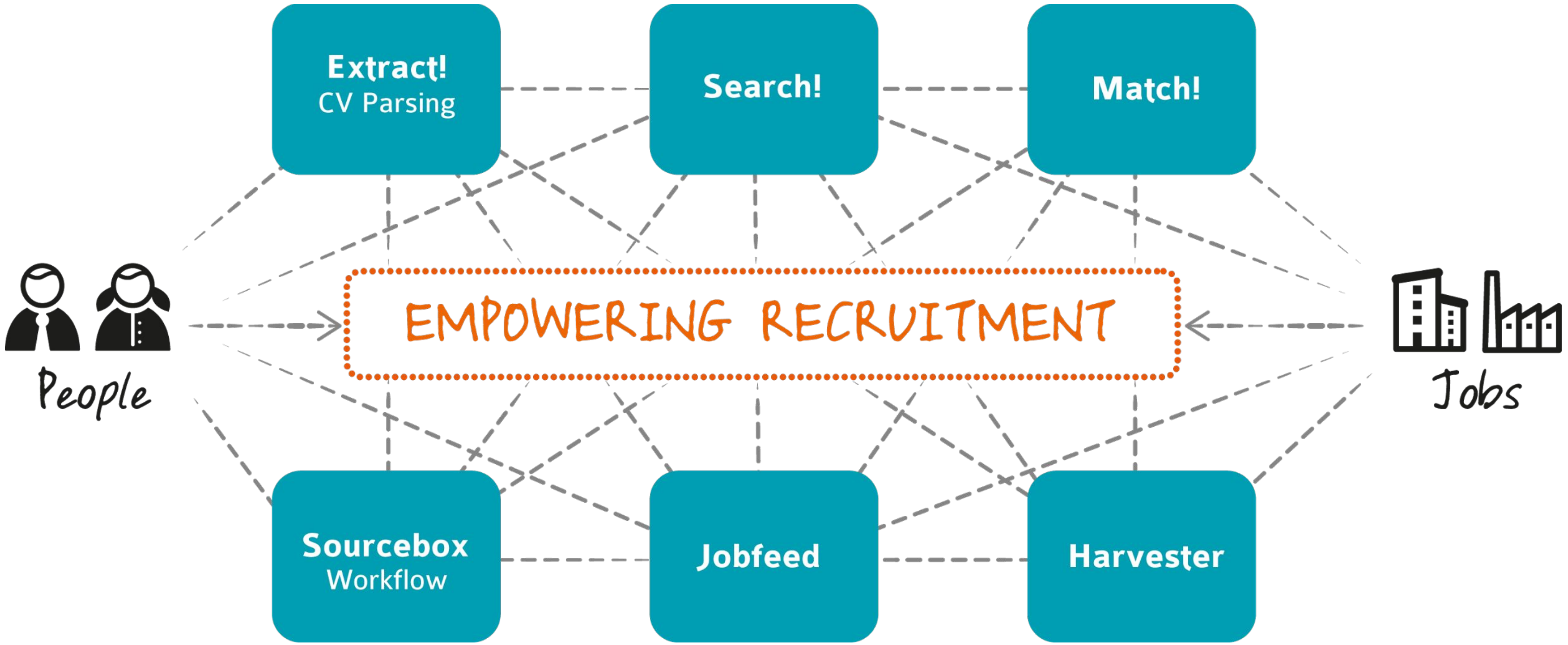
Machine Intelligence for People and Jobs

Who we are and what we do

We develop Technology to bridge the language and meaning gap between People and Jobs …



… through a family of sophisticated software products …



… that large organizations in the HR and Recruitment sector use…

usg people
StepStone
ORDINA
Adecco
better work, better life
CERN
MONSTER
RBS™
KFORCE®
AMSTERDAM RAI
Michael Page INTERNATIONAL
Rabobank
AEGON
ManpowerGroup™
VATTENFALL
randstad
AkzoNobel
Tomorrow's Answers Today
JobStairs
The Top Company Portal
BASF
The Chemical Company

Our Knowledge Graph

Schema Overview

- **Concept Types:**
    - Professions
    - Skills
    - Qualifications (Degrees, Certificates)
    - Organizations (Companies, Educational Institutes)
    - Industries

- **General relations:**
    - prefLabel and altLabel
    - broader and narrower

Domain-Specific Relations

- **Professions are linked to:**
  - Skills and activities they involve
  - Locations, organizations and industries where they are found.
  - Qualifications that are (formally or informally) required for their exercise
  - Similar professions
- **Skills are linked to:**
  - Similar skills and activities,
  - Professions and industries they are mostly demanded by
  - Qualifications that develop and verify them.
- **Qualifications are linked to:**
  - Skills that have as learning outcomes
  - Organizations that provide them
  - The educational levels they cover.

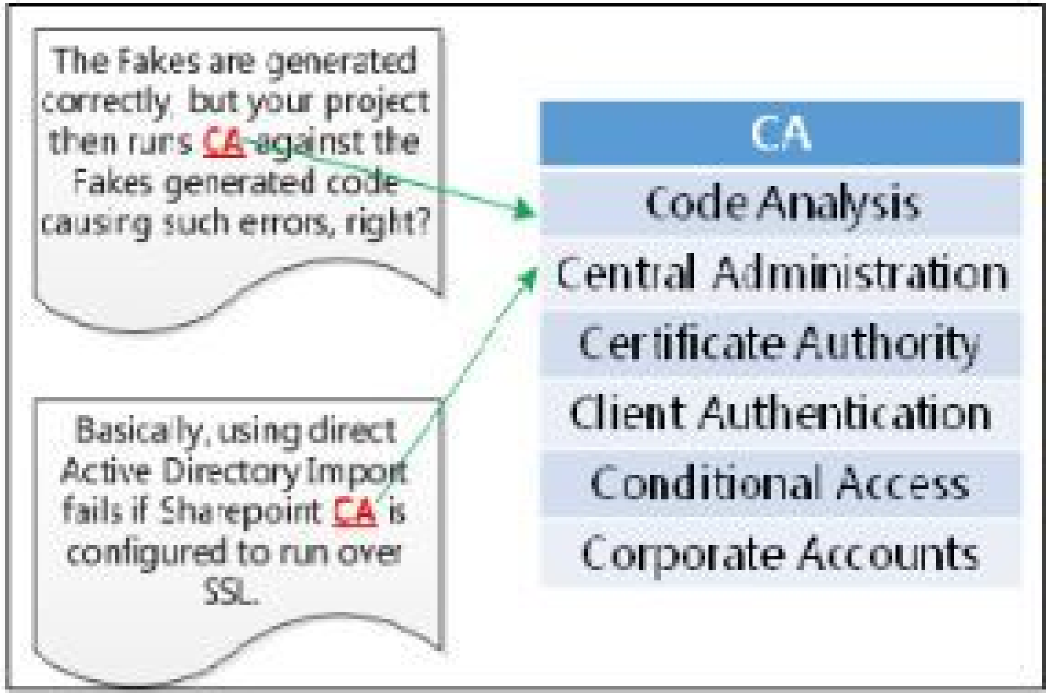
Relation Vagueness

- Many of the relations are vague (e.g., the importance of a skill for a profession).
- As such, their interpretation is ***subjective***, ***context-dependent,*** and usually ***a matter of degree***.
- For that, in our graph, such relations have the following three properties:
  - **Strength**: A number (typically from 0 to 1) indicating the strength/confidence of the relation.
  - **Applicability Context**: The contexts (location, language, industry etc) in which the relation has been discovered and considered to be true.
  - **Provenance**: Information about how the relation has been added to the graph (source, method, process).

How we use our Knowledge Graph

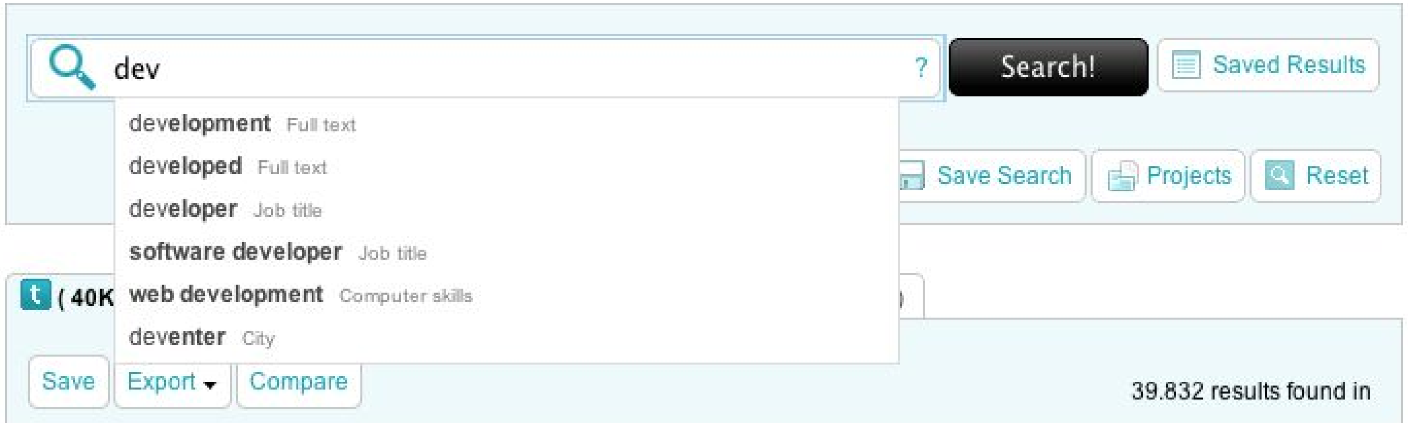
Extract! CV & Vacancy Parsing

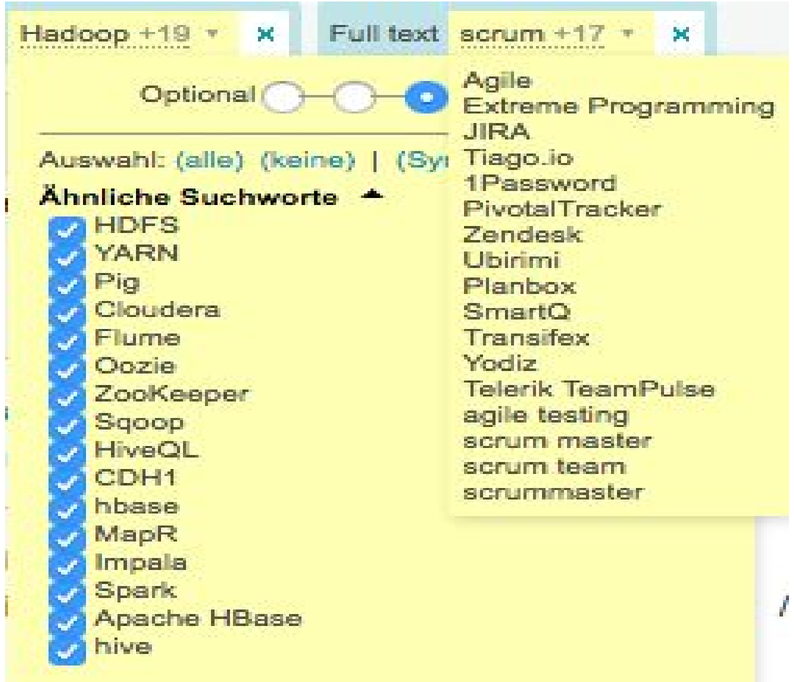
- As Gazeteer for Entity Detection (Skills, Professions, etc.)
- As Contextual Evidence for Entity Disambiguation



Search! Match!

- For auto-suggestion
- For expanding queries
- For calculating semantic similarity





Labour Market Analytics

- Supply-Demand Analysis
- Top Skills per Job
- Career Paths

How we model Concept Drift

General principles

- **Concept drift** is typically modeled by means of a concept’s:
  - **Labels**, i.e., the words used to express the concept
  - **Intension**, i.e., the concept characteristics as expressed via its properties and relations
  - **Extension**, i.e., the set concept's of the concept's instances

- **In our model:**
  - We do not consider extension
  - We do not consider all properties and relations, nor to the same extent

Drift in labels and broader/narrower relations

- Changes in labels matter only when they are not merely additions or removals of spelling and/or morphosyntactic variations of existing labels.
- Changes in preferred labels are slightly more important than alternative labels.
- Changes in a concept's broader and narrower relations are important, with broader changes suggesting a more fundamental drift.

Professions Drift

- Profession meaning is primarily defined by the skills and activities they involve.
- Essential skills for a profession are more important than optional skills.
- Profession meaning also changes, though to a lesser extent, when the industries it is found in change (e.g., journalists start working in the tech sector).
- A profession concept does not drift when the locations or companies it is most popular in, change.

Skills & Qualifications Drift

- **Skills**:
  - Meaning is primarily defined by their similar skills and activities, as these describe for what tasks and in what contexts a skill is used.
  - It also changes, though to a lesser extent, when it starts being applied in different professions and industries.

- **Qualifications**:
  - Meaning is primarily defined by the skills they develop and/or verify.
  - Secondarily, by the professions they regulate and/or are useful for.

How we measure Concept Drift

General approach

- Concept drift is typically detected and quantified by measuring the change in labeling, extension and intension over time.

- In our model:
  - For labeling we use set similarity instead of string one.
  - For vague relations we use metrics that can take in consideration their strength
  - We use parameters to define the particular characteristics of the drift we want to measure (e.g, target concept types, time scope, relation context and provenance.)

Measuring vague relations drift

- Given two versions of the same concept and a vague relation, we derive the top-N related concepts for each version (based on the strength score)
- Then we calculate their similarity using the generalized Kendall's tau that can measure distance between rankings.
- In that way, for example, if the "Data Scientist" profession continues having the same top 10 related skills but differently ranked, a drift will be detected.

Measuring drift for different parameters

- **Different parameter values** can yield **different drift**, not only in terms of **intensity** but also in terms of **interpretation**. For example:
  - Using CVs as a relation provenance, the drift reflects the change in the way the workforce side of the labour market interprets and uses the concept.
  - Using only Vacancies, we shall get an idea of how the same concept changes from the industry's perspective.
  - Using news articles, we will measure the change in the general perception of the concept.
  - Using more encyclopedic and definitional data sources (e.g. Wikipedia or specialized dictionaries) may indicate changes in more core aspects of the concept's meaning.

How we exploit Concept Drift

Engineering Dimension

- Measuring and monitoring drift helps us quantify and understand better the dynamics of our domain and our graph's content.
- This, in turn, enables us to plan and prioritize the maintenance and evolution of the knowledge graph much more effectively
- For example, we are able to identify highly volatile graph aspects that need more frequent updates, and allocating more resources for that.

Business Dimension

- Drift in our knowledge graph indicates to a large extent the changes that take in place in the labour market, especially the one that we derive from CVs and Vacancies.

- These changes we can then communicate to job seekers, candidate seekers, education and training providers, policy makers etc.

Wrapping Up

Main Conclusions

- The definition and modeling of **semantic drift** for a given knowledge graph **should take into account the graph's content, domain and application context**, and adapted accordingly.

- In order to be able to understand and interpret concept drift better, we need a **versatile measurement framework** that enables the **dynamic and highly configurable measurement and presentation of drift**.

Thank you!

Dr. Panos Alexopoulos
Head of Ontology

E-mail: alexopoulos@textkernel.nl

Web: http://www.panosalexopoulos.com

LinkedIn: www.linkedin.com/in/panosalexopoulos

Twitter: @PAlexop