textkernel

Machine Intelligence for People and Jobs

NLP in the wild

Machine intelligence for matching people and jobs

Kasper Kok
Textkernel BV

Agenda

1. AI in industry versus academia
2. CV parsing
3. Matching and normalization
4. Knowledge graphs

Agenda

01 AI in industry versus academia

02 CV parsing

03 Matching and normalization

04 Knowledge graphs

Speaker

Kasper Kok, PhD

Product Manager
BSc AI
MSc CogSci
PhD Linguistics

textkernel

Machine Intelligence for Matching People and Jobs

- AI and Machine Learning
- Semantic Search and Match
- Document Understanding
- Web Mining
- Labor Market Intelligence

International market leader in AI for HR and Recruiting

Founded in 2001 | Headquarter in Amsterdam | 1.000+ clients worldwide | 145 full-time employees, majority R&D and development

textkernel

Textkernel product line

An AI Platform for Talent Acquisition and HR which transforms data into rich and actionable information, enabling you to understand, connect and analyze in a meaningful way.

Understand

- All documents
- Behavior & history
- Great details & nuances
- In any language

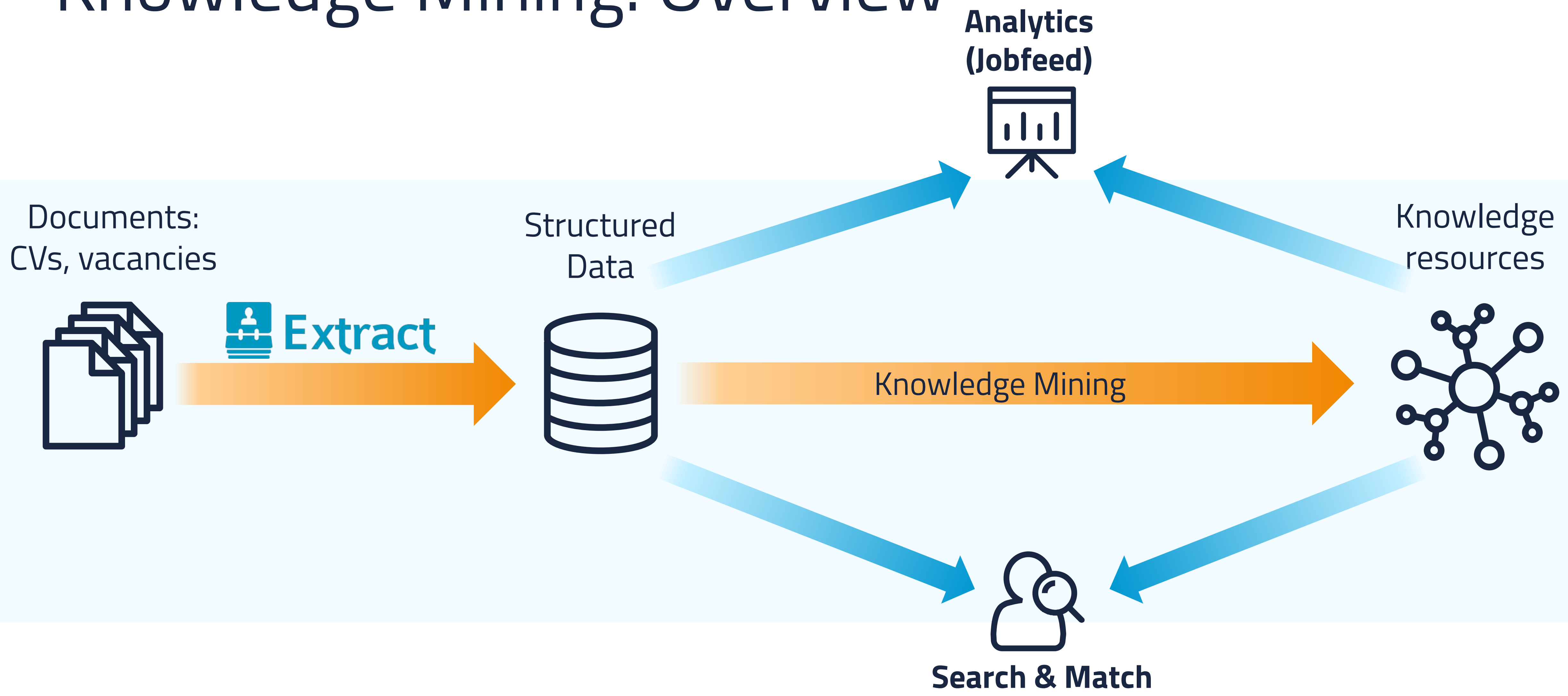
Connect

- All people & jobs
- Matching
- Recommendations
- Personalization

Analyze

- Supply & demand
- Public & private data

Knowledge Mining: Overview



textkernel

Textkernel in Numbers

20 Years of Jobs
History & Trends

48+
Countries

23
Languages

835 Million
CVs per year

140K
Job Titles

154K
Skills

100+
partners

1,3 Billion Jobs analyzed adding 350 million per year

1,000+
Customers

We serve 7 out of 10
Top Global Staffing Firms

ISO27001
Certified

AI in industry vs academia

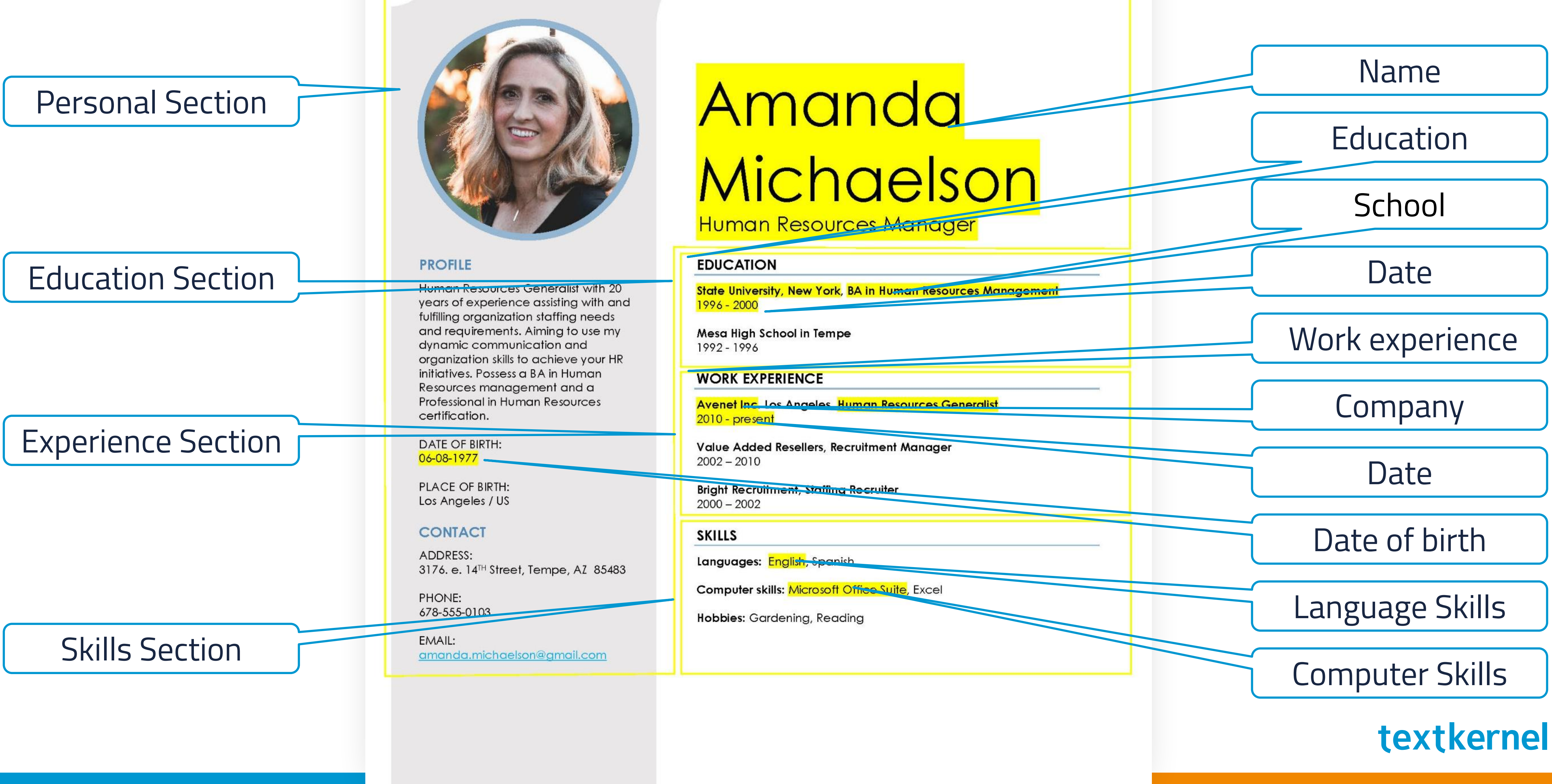
	Academia	Industry
Overall goal	Advance scientific understanding	$ (at least not make a loss)
Project goal	Publish a paper = Scientifically noteworthy research outcome	Make a viable product = Solve a customer problem
Data	Controlled benchmark datasets (usually)	Messy real world data, continuously evolving
Models	Latest and greatest	Whatever works to stay ahead of the competition.

Document understanding

textkernel

CV parsing demo

CV Parsing



textkernel

What customer problem do we solve?

250 applications are received for each corporate job offer -Glassdor 2019

40 seconds is the time it takes to **read a resume** Study Miratech 2018

1 min the time it takes to **select a candidate** after reading the cv- study Miratech 2018

46,3% of the applications received **are read** -Tilkee Study - 2017

textkernel

Breakout session

Task: how would you build a system that extracts the candidate name from a CV

- Rule-based
- Machine Learning

5 mins

Present your idea via a representative: 1-2 minutes

textkernel

Rule-based approach?

- Candidate name extraction
  - Context: words after “Name:”

```
Name: Mihai Rotaru
```

Rule-based approach?

- Candidate name extraction
    - Context: words after “Name:”
        - Until end of line

```
Name: Mihai Rotaru
```


textkernel

Rule-based approach?

- Candidate name extraction
    - Context: words after “Name:”
        - Until end of line

```
Name: Mihai Rotaru                    Amsterdam,Netherlands
```


textkernel

Rule-based approach?

- Candidate name extraction
  - Context: words after “Name:”
    - Until end of line
    - Stop when reaching a large space

```
Name: Mihai Rotaru                          Amsterdam,Netherlands
```


textkernel

Rule-based approach?

- Candidate name extraction
  - Context: words after “Name:”
    - Until end of line
    - Stop when reaching a large space

```
Name: Mihai Rotaru  phone: +31 20 494 2496 Amsterdam,Netherlands
```

Rule-based approach?

- Candidate name extraction
  - Context: words after “Name:”
    - Until end of line
    - Stop when reaching a large space
    - Stop when reaching a lowercase word

```
Name: Mihai Rotaru  phone: +31 20 494 2496 Amsterdam,Netherlands
```

Rule-based approach?

- Candidate name extraction
  - Context: words after “Name:”
    - Until end of line
    - Stop when reaching a large space
    - Stop when reaching a lowercase word

```
Name: Mihai van Rotaru
```


textkernel

Rule-based approach?

- Candidate name extraction
  - Context: words after “Name:”
    - Until end of line
    - Stop when reaching a large space
    - Stop when reaching a lowercase word
      - Allow certain lowercase words (von, van, de la)

```
Name: Mihai von Rotaru
```


textkernel

Rule-based approach?

- Candidate name extraction
  - Context: words after “Name:”
    - Until end of line
    - Stop when reaching a large space
    - Stop when reaching a lowercase word
      - Allow certain lowercase words (von, van, de la)

```
Father’s Name: Mihai Rotaru
```


textkernel

Rule-based approach?

- Candidate name extraction
  - Context: words after “Name:”
    - Until end of line
    - Stop when reaching a large space
    - Stop when reaching a lowercase word
      - Allow certain lowercase words (von, van, de la)
    - Nothing before “Name:”

```
Father's Name: Mihai Rotaru
```

Rule-based approach?

- Candidate name extraction
  - Context: words after “Name:”
    - Until end of line
    - Stop when reaching a large space
    - Stop when reaching a lowercase word
      - Allow certain lowercase words (von, van, de la)
    - Nothing before “Name:”
  - No context?

textkernel

Rule-based approach?

- Candidate name extraction
  - Context: words after “Name:”
    - Until end of line
    - Stop when reaching a large space
    - Stop when reaching a lowercase word
      - Allow certain lowercase words (von, van, de la)
    - Nothing before “Name:”
  - No context
    - List of first names

```
                     Mihai Rotaru
rotaru@textkernel.nl          William Street 12, Amsterdam
+31 20 494 2496                          The Netherlands
```

From rules to machine learning

Problem with rules

- Gets complex to accommodate for exceptions
- Coverage is limited

Every field in a CV is a brain teaser

- **Dates**
  - 2011
  - ‘11
  - 11
  - Mar 02
  - 01/2011
  - 31.01.2011
  - 2011-06-01
  - 120206

- **Date ranges**
  - 03-2011 - 05-2011
  - 03/05 2011
  - 060401-060930
  - 062006-072013

textkernel

Machine Learning to the rescue

- Problem with rules: not 100% sure signals
- Machine Learning:
  - Estimate the quality of signals (from annotated data)
  - Combine multiple signals

CV parsing: Name extraction

Start right after "Name:"

- Unless "Father"/"Mother" before

Stop:

- End of line OR
- Large white space OR
- Lower case word (unless: van, von, de, la, …)
- …

Combine signals

Signals (features)

textkernel

Why as a sequence?

- Context and order is important

```
1999-2001: Preparator at Universitatea de Vest
```


textkernel

Why as a sequence?

- Context and order is important
  - Pattern: DATE: JOB **at** COMPANY

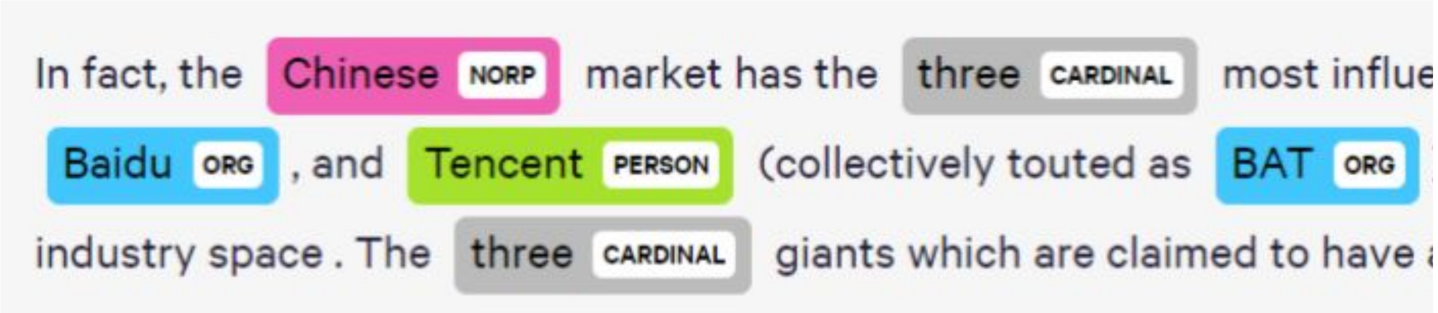
$$\overbrace{\texttt{1999-2001}}^{\text{Date}}\texttt{: }\overbrace{\texttt{Preparator}}^{\text{Job}}\ \textbf{at}\ \overbrace{\texttt{Universitatea de Vest}}^{\text{Company}}$$

textkernel

Modeling text as a sequence

Problem class: Sequence labeling

- Part of Speech Tagging, Named Entity Recognition
- Models: HMM, CRF, RNN/LSTM



textkernel

Pamela Woolley

PERSONAL INFORMATION

Address:	76, Millbrook Road East, Southampton, X7W2BB, Hampshire
Mobile:	07776-396738
e-mail:	Pamela@hotmail.com , pamela@gmail.com
Nationality:	American
Date of birth:	7 August, 1967

PROFESSIONAL EXPERIENCE

2003 - present **FREELANCE PROJECTS, Brussels**
Global Communications officer, *Huntsman Advanced Materials* (nine month contract)
Responsible for the global communication function post re-structuring
Activities include:
- Auditing intemal communications
- Preparation of internal and external communications for the president

1999 - 2003 **TOYOTA MOTOR EUROPE, Brussels**
Manager, *Organisational Identity and Brand Management*
Responsible for strategic development and implementation of the Toyota brand in Europe

1996 - 1999 **SCOTTISH INDUSTRIAL AND TRADE EXHIBITIONS, Edinburgh**
Sales and Marketing Assistant

EDUCATION

1994-1996 **LONDON BUSINESS SCHOOL**
MBA degree
Second year project in brand building for Maria Bland

1995-1995 **UNIVERSITY of Cologne**
Completed one term of *Business Administration (BWL)* degree

LANGUAGES

English	fluent (mother tongue)
French	fluent (spoken and written)
German	good (spoken)

COMPUTER SKILLS

Microsoft Office (Powerpoint, WORD, Excel, Outlook), C++, Perl

Personal section

Pamela Woolley

PERSONAL INFORMATION

Address:	76, Millbrook Road East, Southampton, X7W2BB, Hampshire
Mobile:	07776-396738
e-mail:	Pamela@hotmail.com , pamela@gmail.com
Nationality:	American
Date of birth:	7 August, 1967

Experience section

PROFESSIONAL EXPERIENCE

2003 - present **FREELANCE PROJECTS, Brussels**
Global Communications officer, *Huntsman Advanced Materials* (nine month contract)
Responsible for the global communication function post re-structuring
Activities include:

- Auditing internal communications
- Preparation of internal and external communications for the president

1999 - 2003 **TOYOTA MOTOR EUROPE, Brussels**
Manager, *Organisational Identity and Brand Management*
Responsible for strategic development and implementation of the Toyota brand in Europe

1996 - 1999 **SCOTTISH INDUSTRIAL AND TRADE EXHIBITIONS, Edinburgh**
Sales and Marketing Assistant

Education section

EDUCATION

1994-1996 **LONDON BUSINESS SCHOOL**
MBA degree
Second year project in brand building for Maria Bland

1995-1995 **UNIVERSITY of Cologne**
Completed one term of *Business Administration (BWL)* degree

Skill section

LANGUAGES

English	fluent (mother tongue)
French	fluent (spoken and written)
German	good (spoken)

Skill section

COMPUTER SKILLS

Microsoft Office (Powerpoint, WORD, Excel, Outlook), C++, Perl

Personal section

Pamela Woolley

PERSONAL INFORMATION

Address:	76, Millbrook Road East, Southampton, X7W2BB, Hampshire
Mobile:	07776-396738
e-mail:	Pamela@hotmail.com , pamela@gmail.com
Nationality:	American
Date of birth:	7 August, 1967

Experience section

PROFESSIONAL EXPERIENCE

Item 1

2003 - present **FREELANCE PROJECTS, Brussels**
Global Communications officer, *Huntsman Advanced Materials* (nine month contract)
Responsible for the global communication function post re-structuring
Activities include:

- Auditing internal communications
- Preparation of internal and external communications for the president

Item 2

1999 - 2003 **TOYOTA MOTOR EUROPE, Brussels**
Manager, *Organisational Identity and Brand Management*
Responsible for strategic development and implementation of the Toyota brand in Europe

Item 3

1996 - 1999 **SCOTTISH INDUSTRIAL AND TRADE EXHIBITIONS, Edinburgh**
Sales and Marketing Assistant

Education section

EDUCATION

1994-1996 **LONDON BUSINESS SCHOOL**
MBA degree
Second year project in brand building for Maria Bland

1995-1995 **UNIVERSITY of Cologne**
Completed one term of *Business Administration (BWL)* degree

Skill section

LANGUAGES

English	fluent (mother tongue)
French	fluent (spoken and written)
German	good (spoken)

Skill section

COMPUTER SKILLS

Microsoft Office (Powerpoint, WORD, Excel, Outlook), C++, Perl

experience date

company name, location

job title

textkernel

Extraction

- Typical pipeline (Machine Learning)
  - Preprocessing/OCR
  - Detection of CV pages [mostly for DE]
  - Section segmentation
  - Item segmentation
  - Phrase extraction

Parsing CVs and Jobs

Machine learning: **since 2001**

- Signals
  - Known header on the line
  - Starts with date
  - Email
  - Typical experience words
  - ...
- Combine signals
  - Hidden Markov Models
  - Conditional Random Fields

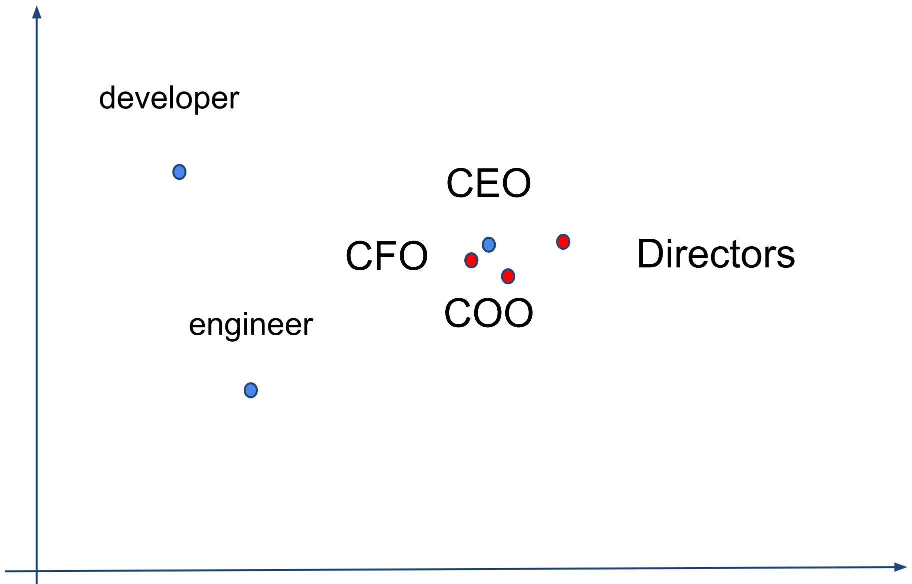
Deep Learning: **since 2014**

	Rule based	Machine Learning	Deep Learning
Signals (features)	People	People (ML engineers)	**Machine** (patterns in data)
Combine signals	People	**Machine** (based on training data)	**Machine** (based on training data)

textkernel

Deep Learning: words

- Multiple dimensions
    - Same level: CEO, CFO, etc
    - Same domain: Nurse, Doctor, Pharmacist, etc
- Word → vector (word2vec tool)
    - Feed unannotated documents



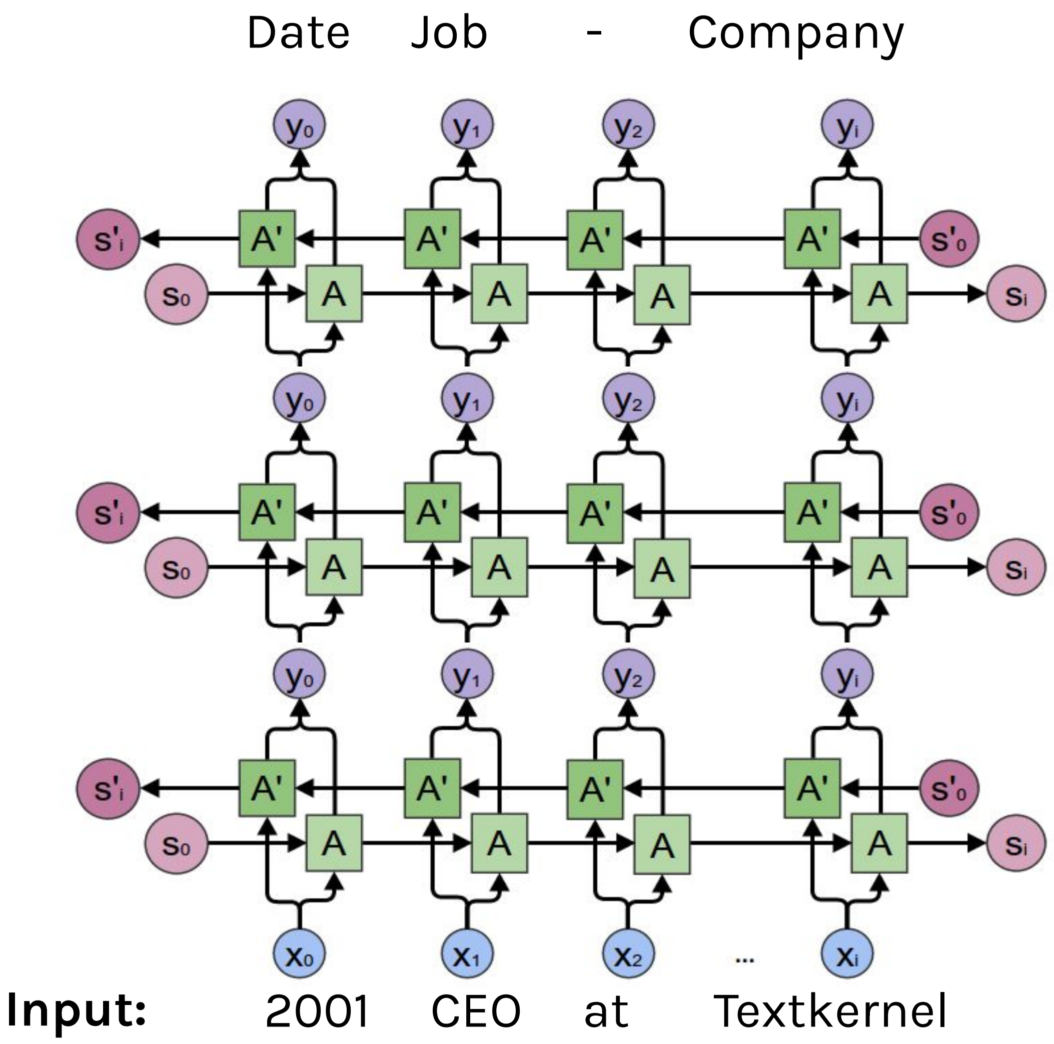
CEO

COO
CFO
SVP
CIO
EVP
VP
CTO
PRESIDENT
DIRECTORS
CHAIRMAN
C.E.O.

textkernel

Deep Learning: Parsing

Recurrent Neural Networks



textkernel

CRF/HMM → Deep Learning

Language	Personal section	Experience section	Education section
English	+25%	+20-30%	+10-20%
Dutch	+10-25%	+20-30%	+15-40%
French	+20%	+30%	+20%
German	+25%	+15%	+25%
Russian	+60%	+50%	+60%
Spanish	+15-30%	+60%	+50%
Swedish	+50%	+35%	+30%

Matching people to jobs and vice versa

textkernel

What customer problem do we solve?

250 applications are received for each corporate job offer -Glassdor 2019

40 seconds is the time it takes to **read a resume** Study Miratech 2018

1 min the time it takes to **select a candidate** after reading the cv- study Miratech 2018

46,3% of the applications received **are read** -Tilkee Study - 2017

44 hours is the average time taken to **consult an application file** -Robert Half 2017

textkernel

Vacancy Parsing

Human Resources Manager

For a growing Los Angeles IT firm, we are looking for an experienced Human Resources Manager. This position performs a wide variety of Human Resource responsibilities, including but not limited to talent acquisition, benefits administration, records maintenance and management, onboarding and offboarding, and employment law compliance.

Responsibilities:

- Responsible for talent acquisition including posting positions, screening of resumes, conducting telephone screens, and conducting background and reference checks.
- Prepare for all new hires. Conduct new hire onboarding and follow up.
- Responsible for the administration of benefit plans.
- Support Human Resources in planning, implementing, communicating and administering human resources programs, policies and practices.

A track-record to fit in perfectly

- Bachelor's degree in Human Resources or related field required.
- Minimum of 4 years prior Human Resources experience in a professional environment.
- Prefer candidates with PHR or SHRM-CP Certification.
- Knowledge of commonly used HR concepts, practices and laws.
- Proficient in Windows and Microsoft Office Suite.

What you will get for your efforts

- Annual salary between $105.000 and $130.000
- 40 hour work week
- 30 paid holidays
- Permanent contract
- Excellent pension scheme

Our agency is specialized in recruiting professionals in the IT, Global Energy and Natural Resources, Life Sciences, Supply Chain and Engineering. We can help you find temporary and permanent job opportunities in these industries. For more information, visit our website.

Job Description Section

Requirements Section

Benefits Section

Company info Section

Vacancy title

Education

Years experience

IT skills

Salary

Offer details

Additional info

textkernel

But the real world looks more like this

CV parsing

job=HR Consultant

experience=7 years

city=Noordwijk

skill=coordinated projects

Match?

Vacancy parsing

job=
Human Resources Adviser

Experience required: >5

city=Leiden

skill=project management

How to make a system that 'knows' that the fields on the left match the ones on the left?
Discuss 5 minutes: solution for each field

People

CV parsing

job=HR Consultant

experience=7 years

city=Noordwijk

skill=Project Management

Normalized data

job=23
branch=HR
experience=5-10
skill=X12
loc=52N4E

Match?

Normalized data

job=23
branch=HR
experience=5-10
skill=X12
loc=52N4E

Vacancy parsing

Jobs

job=
Human Resources Adviser

experience=>5 years

city=Leiden

skill=coordinated projects

URL

Location normalization: geographic coordinates

gps coordinates leiden

All Maps Images Shopping

About 99.800 results (0,68 seconds)

Leiden / Coordinates

52.1601° N, 4.4970° E

gps coordinates noordwijk

All Maps Images Shopping

About 140.000 results (0,93 seconds)

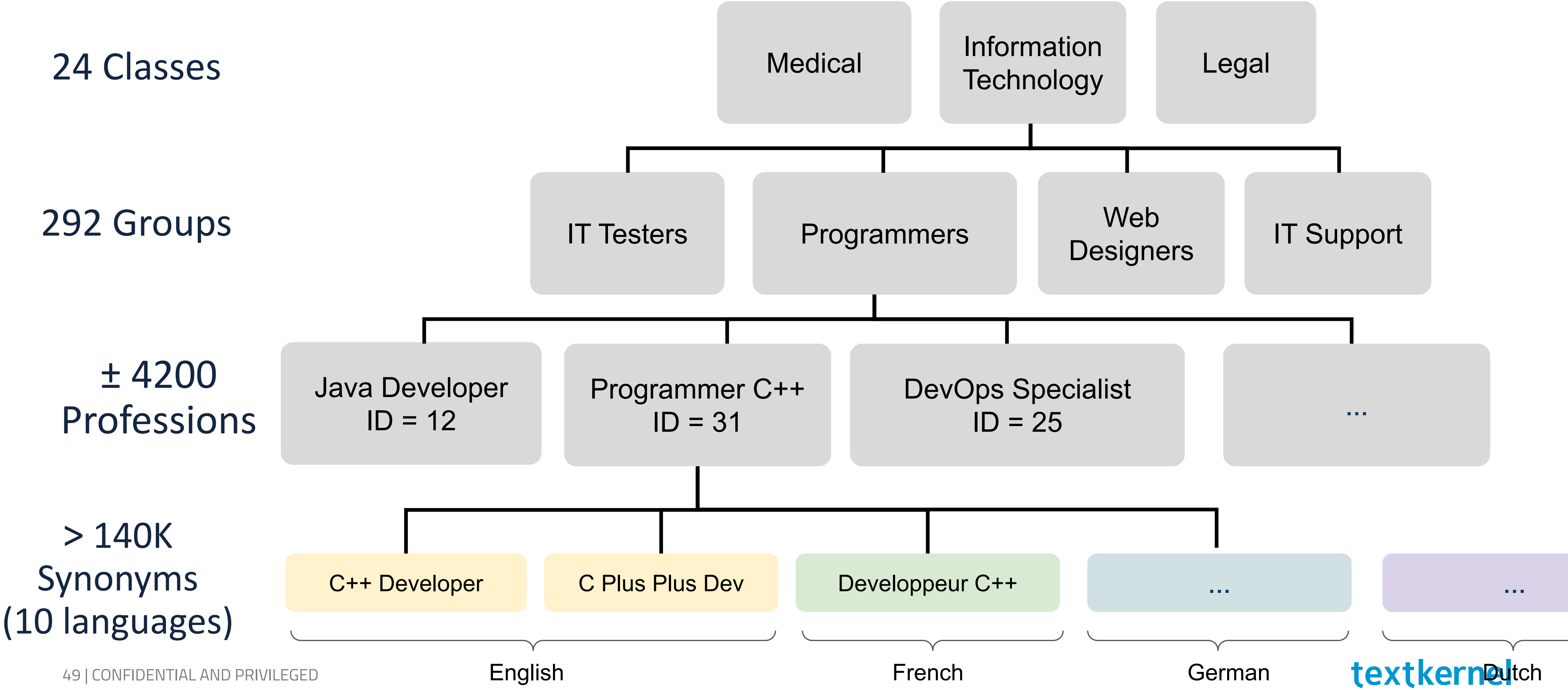
Noordwijk / Coordinates

52.2400° N, 4.4500° E

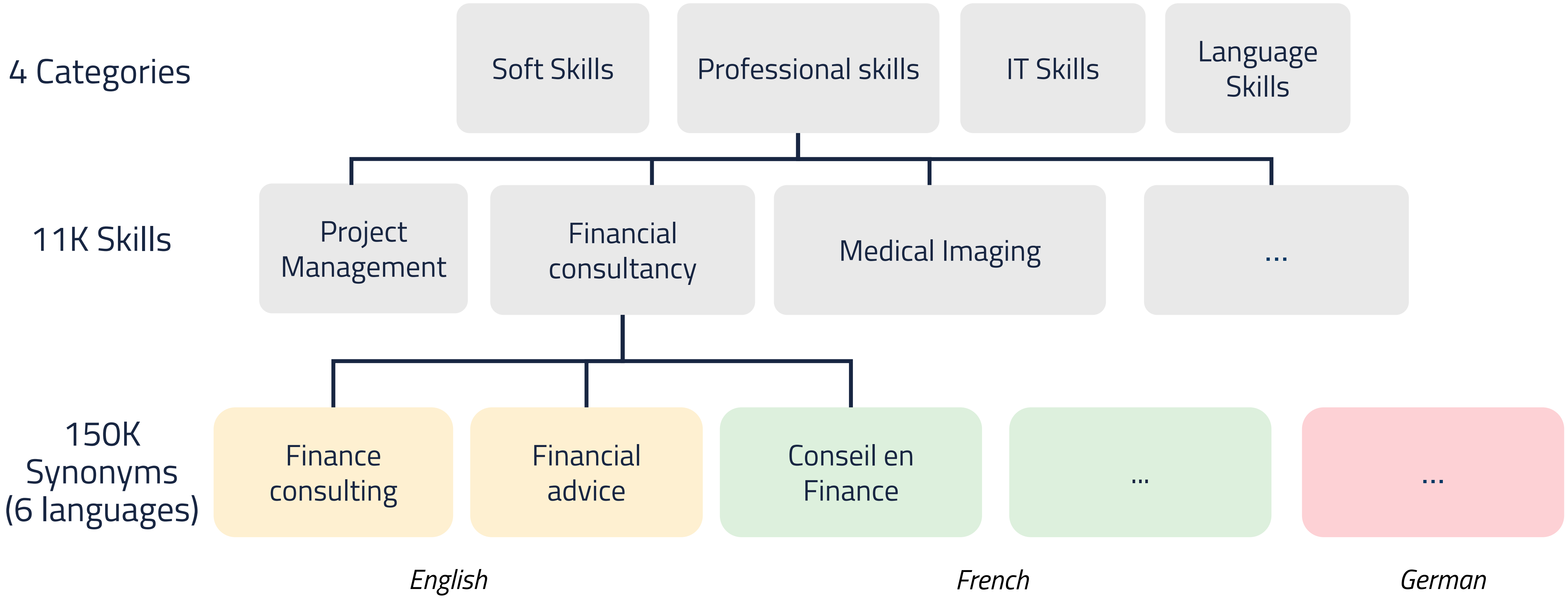
Leiden = Noordwijk +- 30 KM

textkernel

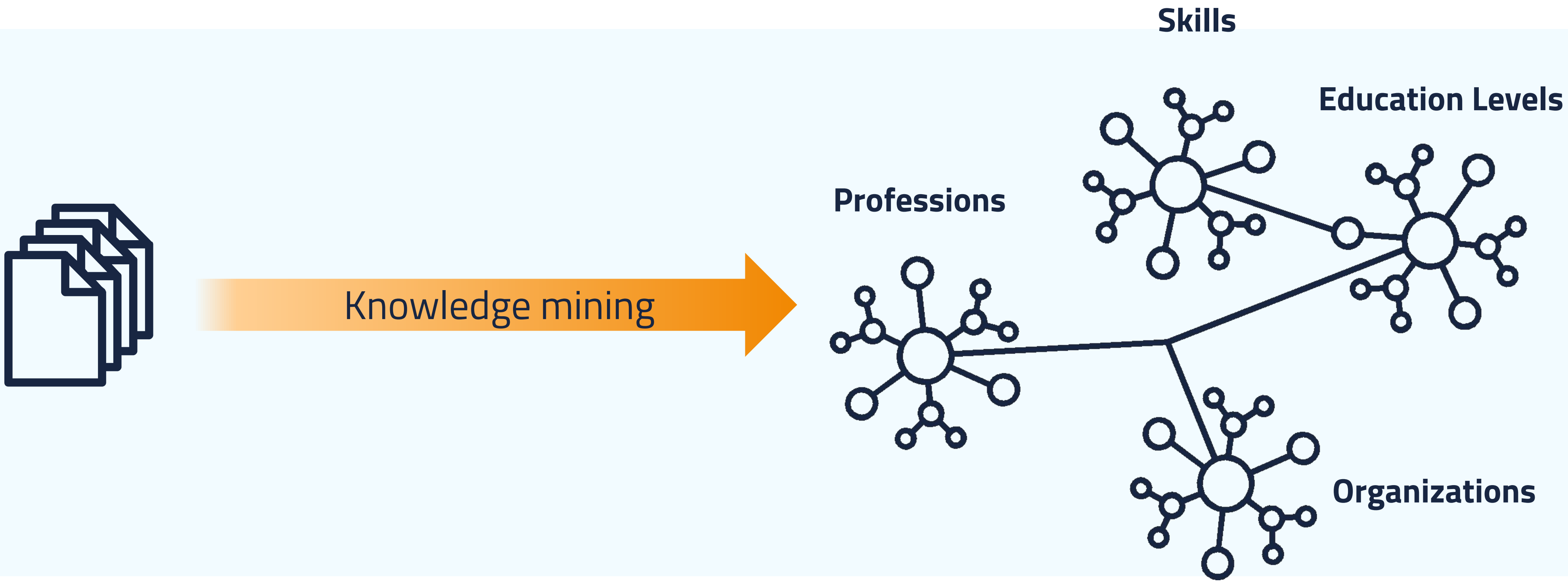
Profession normalization



Skills taxonomy



Textkernel knowledge graph



Comprehensive
Up-to-date
Multilingual

Knowledge mining process

Mine

Filter

Attach

Synonym detection techniques

Rule based

- Dictionaries and lexicons
- Context Heuristics
  - X a.k.a. Y
- Reversed translations

Unsupervised

- Word embeddings
  - But relatedness ≠ synonymy!
- Subword embeddings
  - Byte-pair encoding

Supervised

- Siamese networks

textkernel

Demo Search/Match

Next generation matching: deep learning

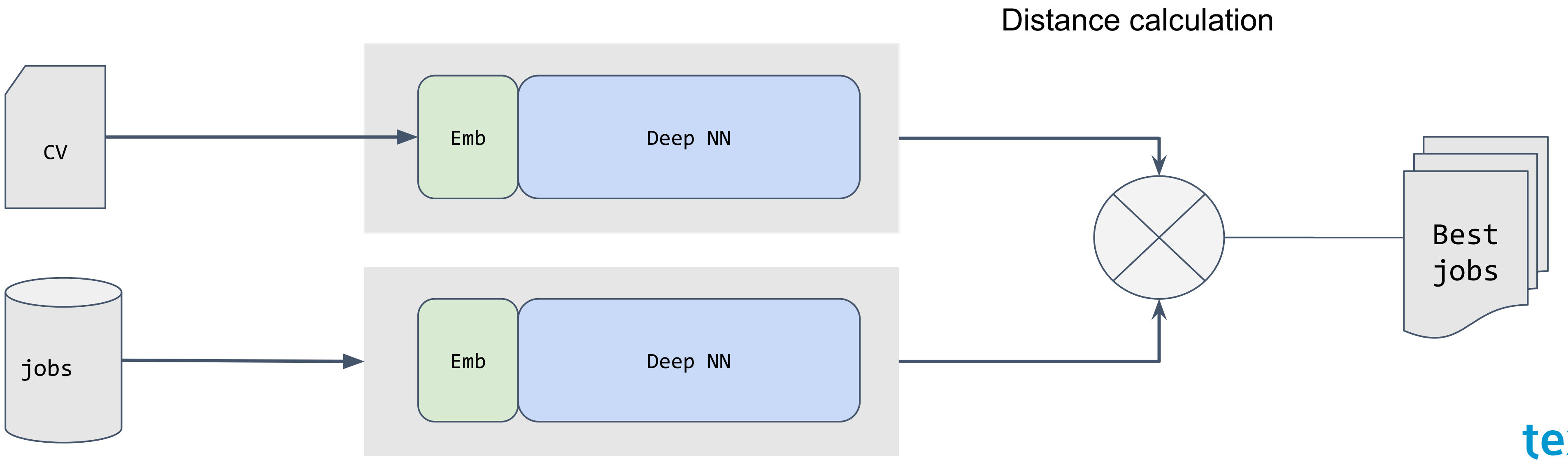
	Rule based	Machine Learning	Deep Learning
Signals (features)	People	People (ML engineers)	**Machine** (patterns in data)
Combine signals	People	**Machine** (based on training data)	**Machine** (based on training data)

textkernel

Document vectors (fingerprints)

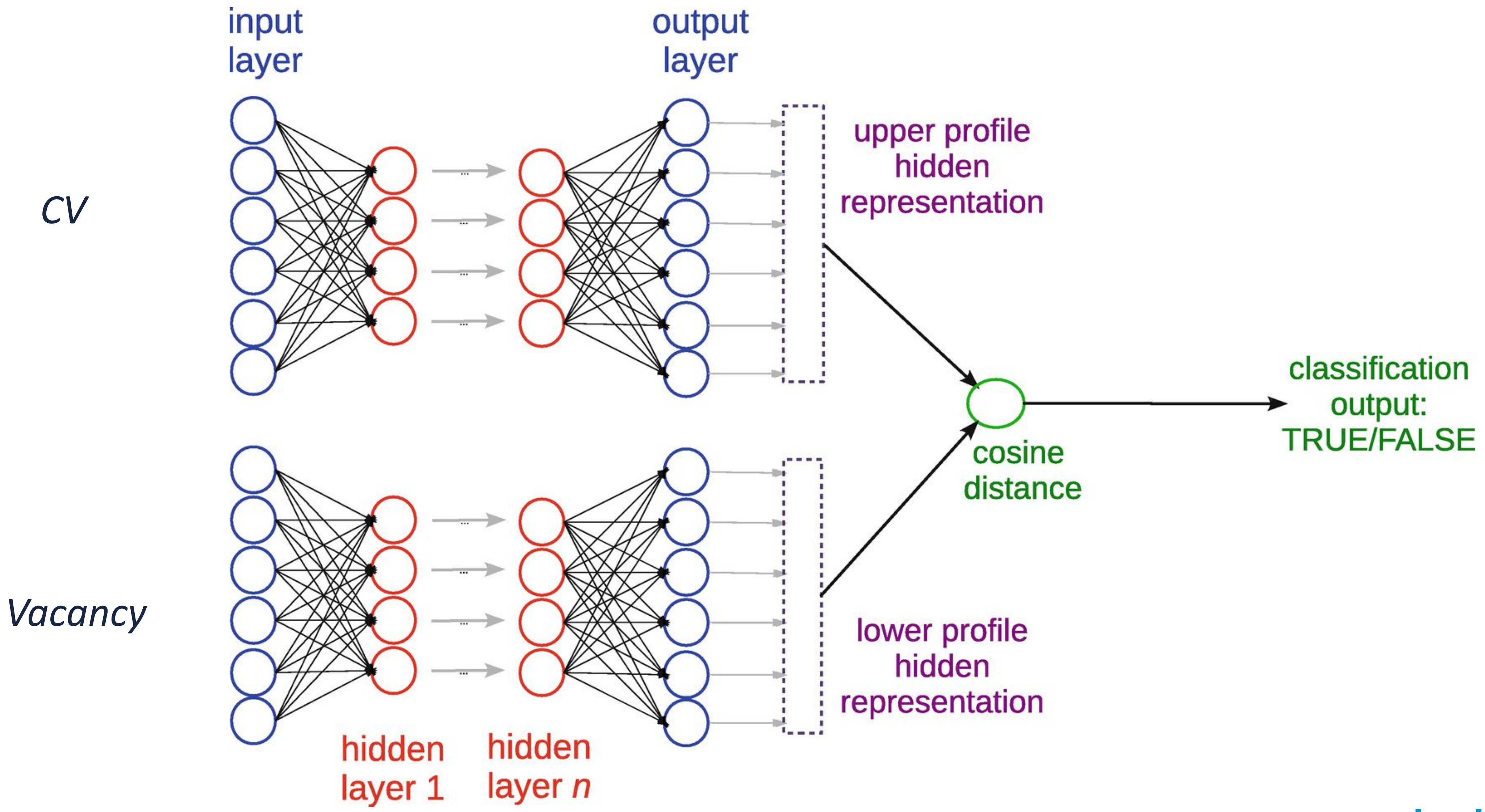
Transform for CVs and vacancies into vectors

- Relevant CVs are “close” to a job
- Capture semantics in a continuous holistic way
- Use wisdom of crowd: learn from people applying to jobs
  - No recruiter bias



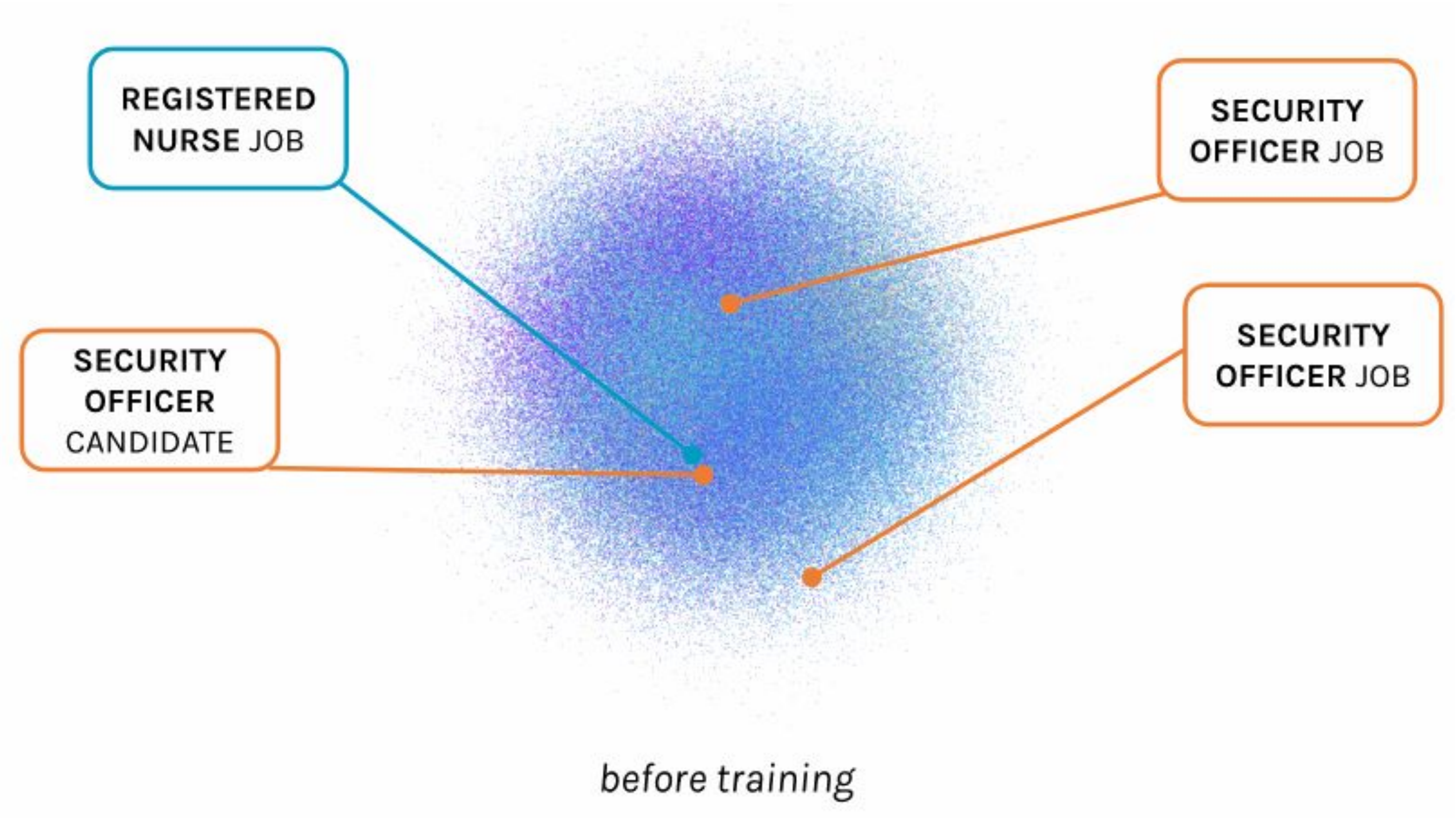
textkernel

Deep learning matcher

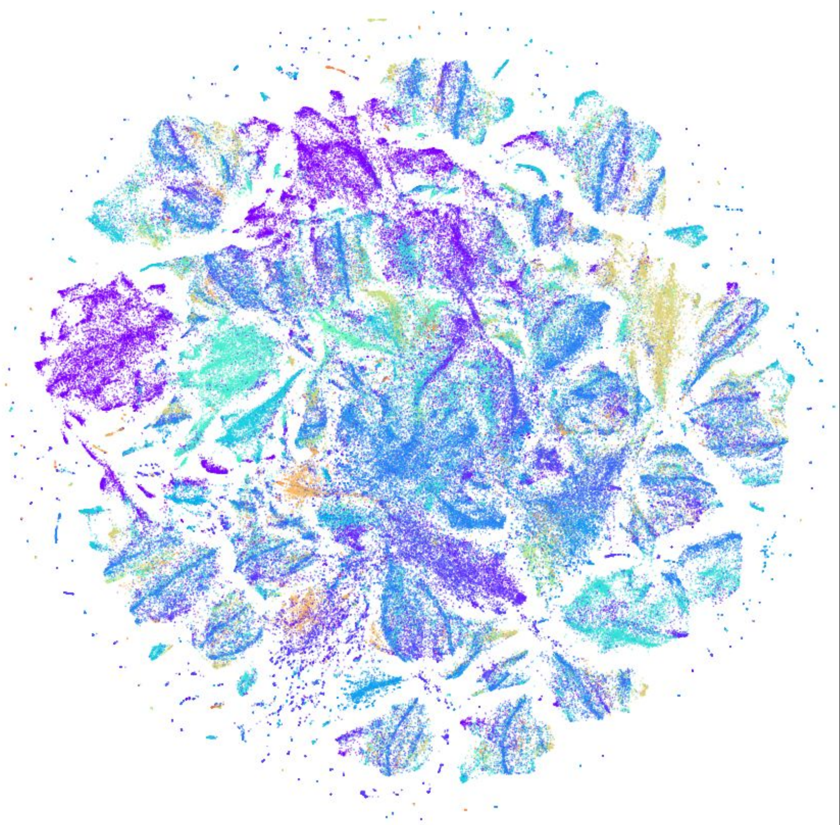


textkernel

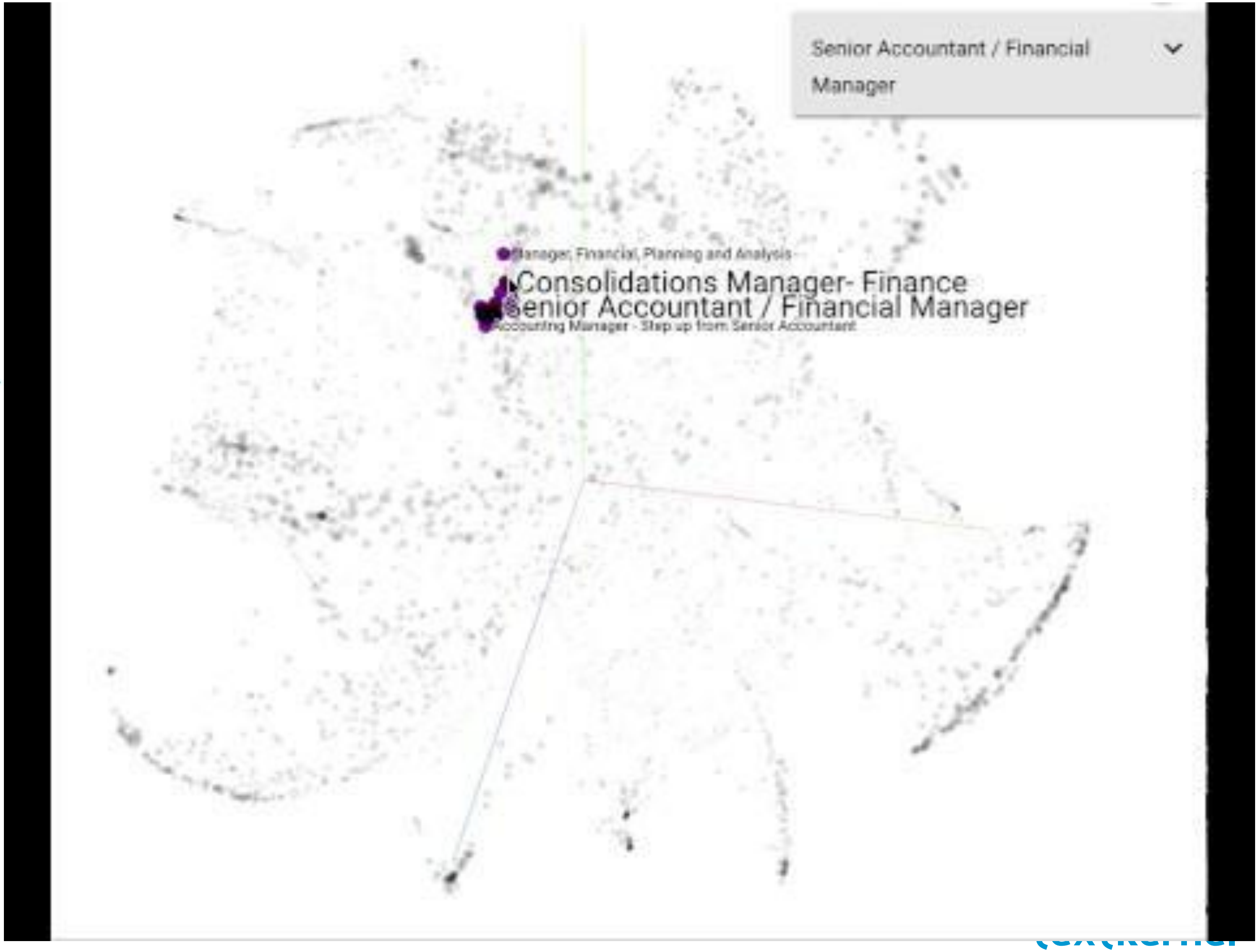
Deep learning matcher: training



Deep learning matcher

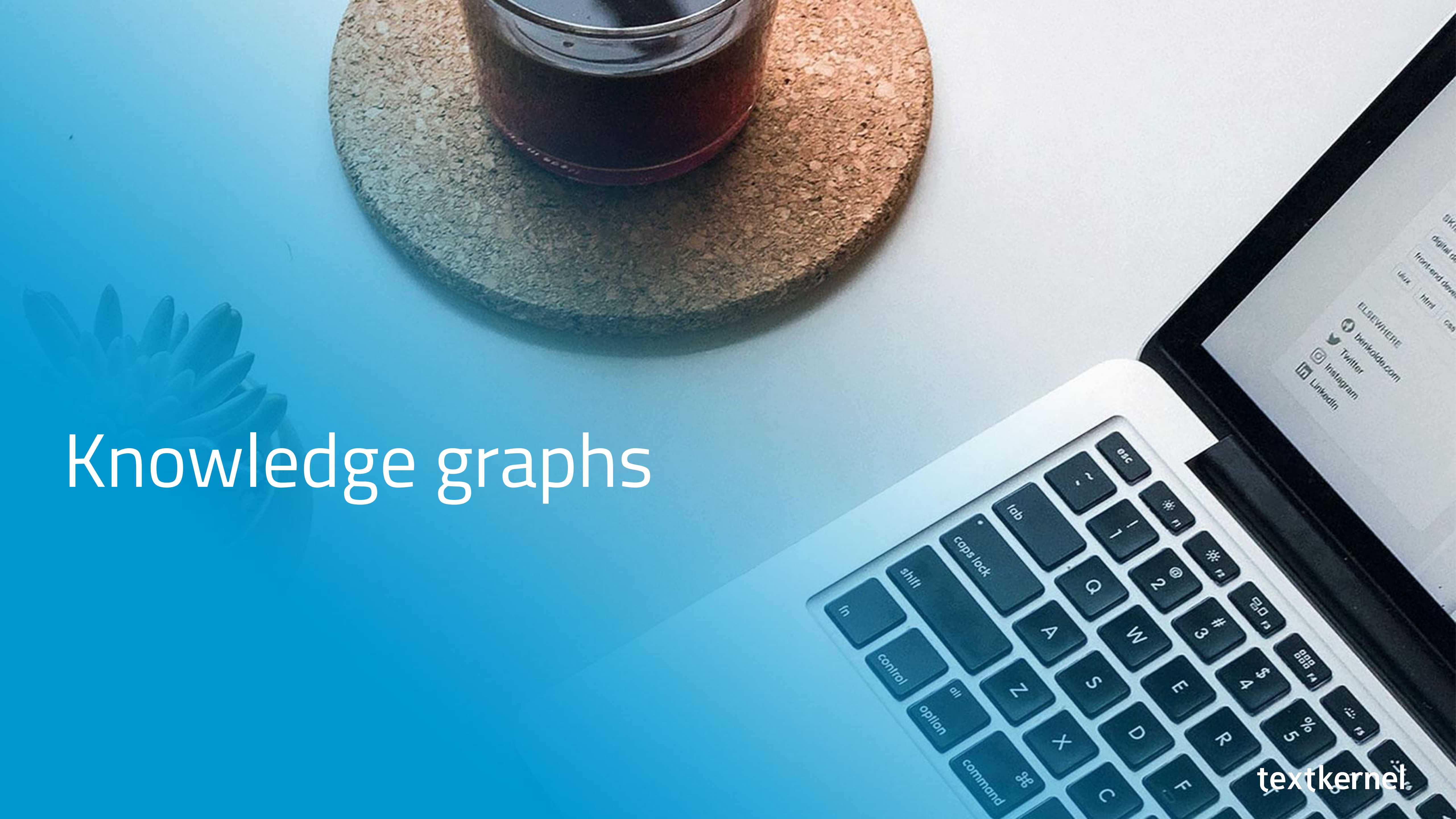


colors represent domains



Deep learning matching

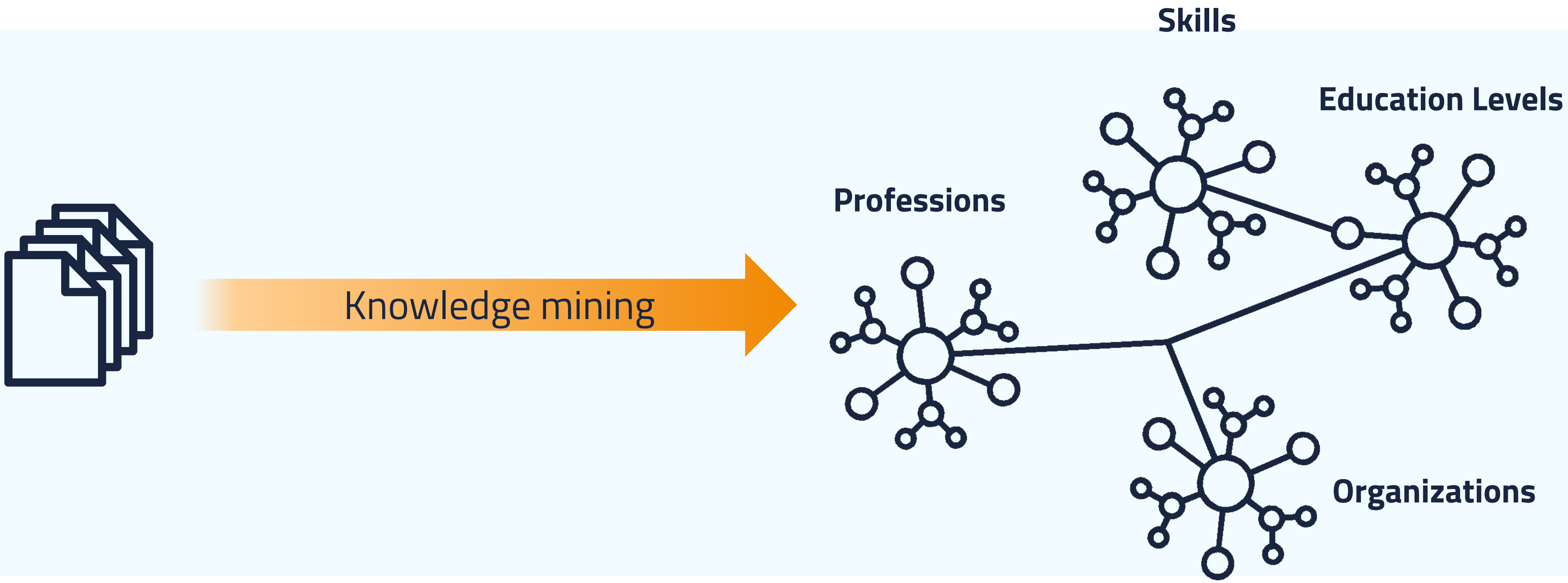
What could be the risks of this method, relative to ‘whitebox’ matching?



Knowledge graphs

textkernel

Back to the knowledge graph

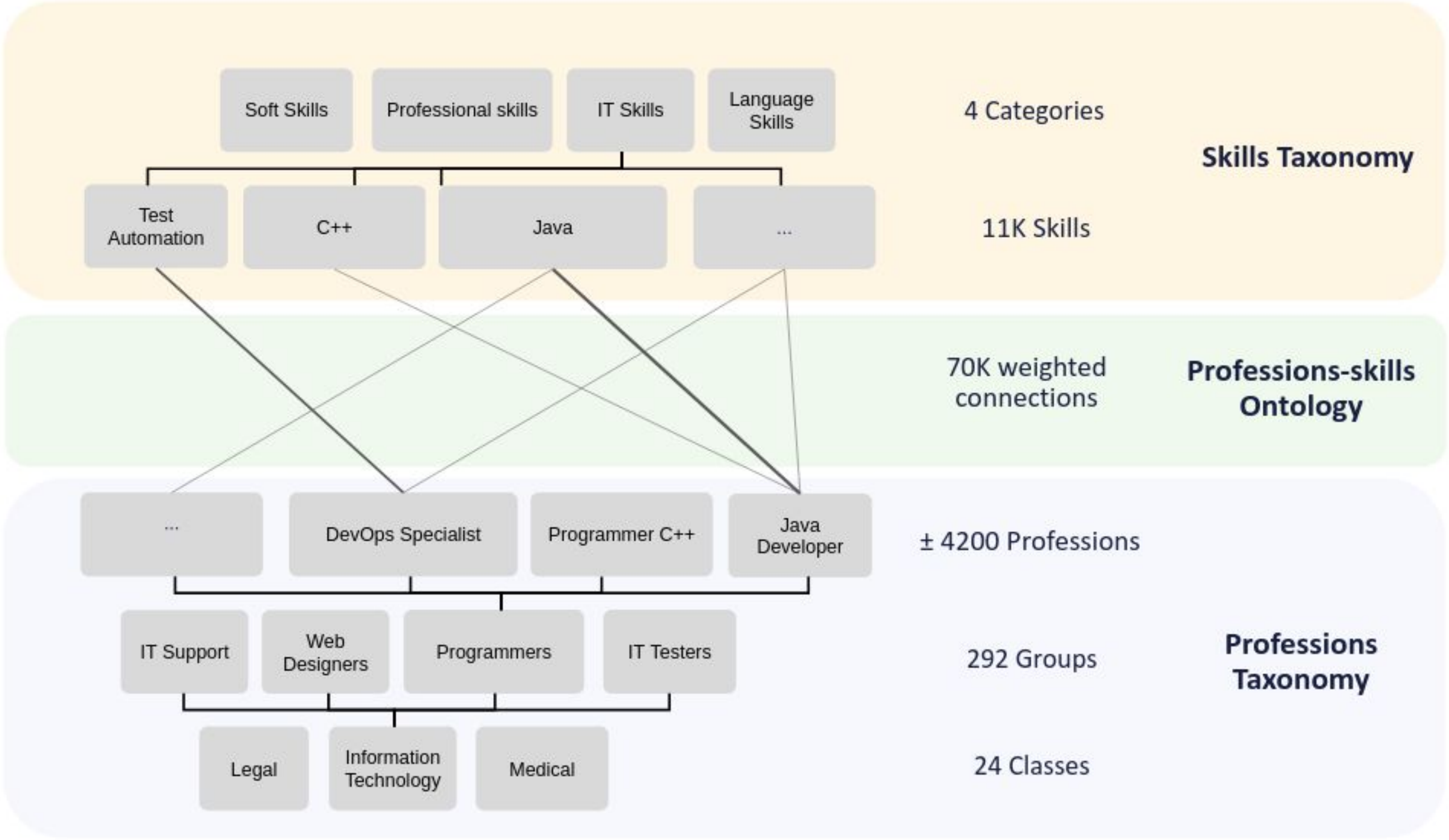


Comprehensive
Up-to-date
Multilingual

Customer request:
We'd like you to tell us which skills relate to which professions

- **Alternative job recommendations**
- **Offer "next job" advice to employees (internal mobility)**

From taxonomies to ‘ontology’



textkernel

Aggregating millions of parsed vacancies

Professional Skills

Marketing	3,124
Sales	2,160
Social Media	1,577
Campaigns	1,308
Digital Marketing	1,134
Marketing Management	999
Marketing Strategies	986
Brand Identity	911
Branding	807
Stakeholder Management	730

IT Skills

Data Analysis	608
Microsoft Excel	548
Databases	392
Microsoft Office	386
MS-Word	305
Google Analytics	290
Microsoft PowerPoint	289
Adobe Photoshop	256
Salesforce.Com	233
Marketing Automation	201

Soft Skills

Communication	2,131
Creativity	1,288
Passionate	1,231
Self Motivation	1,220
Success Driven	1,052
Team-working	1,032
Leadership	835
Attention To Detail	830
Analytical	691
Hardworking And Dedicated	645

Language Skills

English	475
German	86
French	47
Chinese	42
Spanish	29
Italian	18
Japanese	13
Korean	10
Dutch	9
Arabic	9

From *frequent* skills to *salient* skills

How to compute the saliency of skill X for profession Y?

Simplest approach:

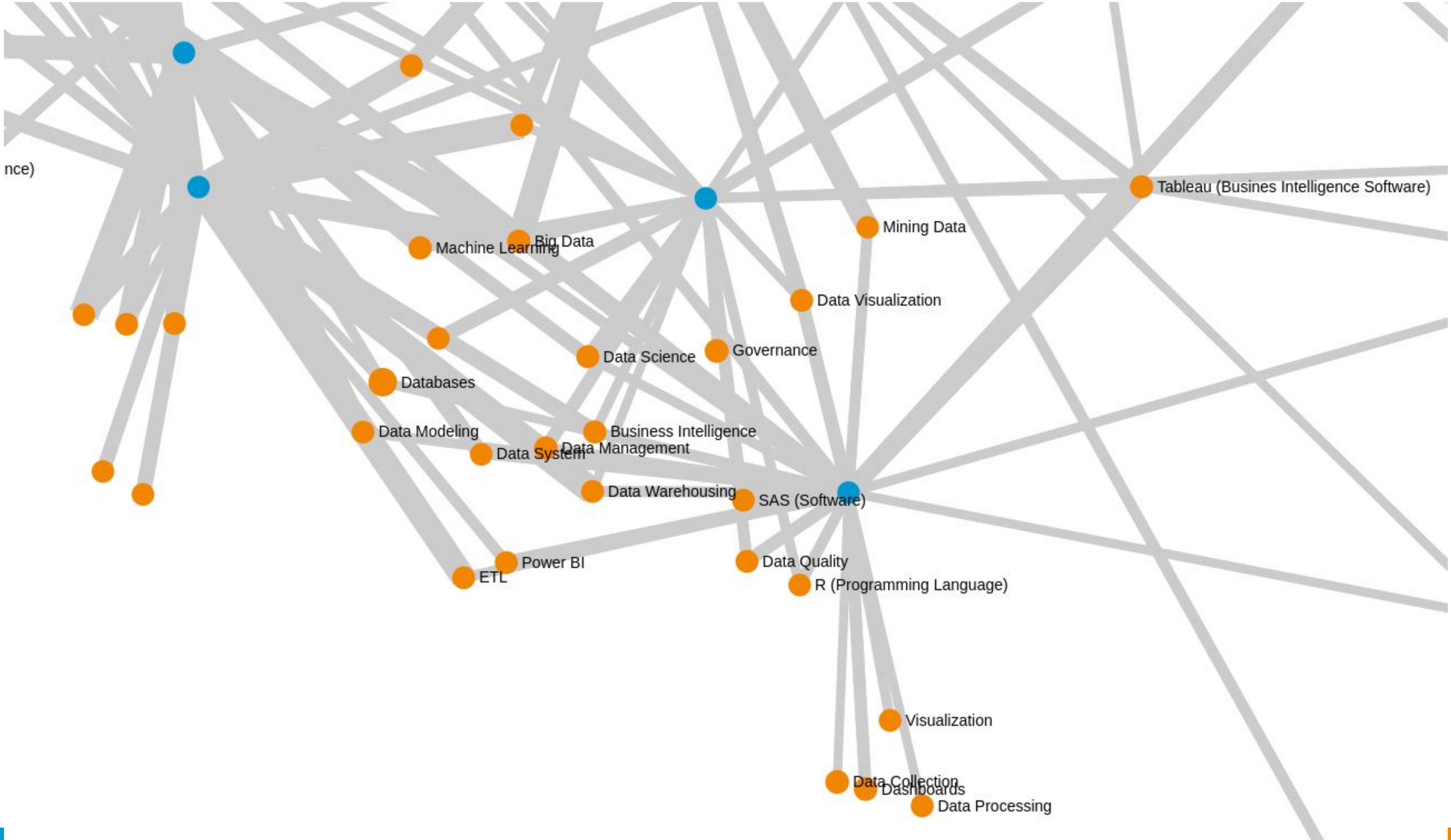
$$\frac{\text{\% of vacancies for profession Y with skill X}}{\text{\% all vacancies with skill X}}$$

Better approaches:

- Chi-square
- Mutual information

textkernel

Knowledge graph: demo



textkernel

textkernel

Machine Intelligence for People and Jobs

Thank you!

kok@textkernel.com

textkernel.careers