European Centre for the Development
of Vocational Training

EN

Understanding technological change and skill needs

Big data and artificial intelligence methods

Cedefop practical guide 2

cedefop

Understanding technological change and skill needs

Big data and artificial intelligence methods

Cedefop practical guide 2

Luxembourg: Publications Office of the European Union, 2021

Please cite this publication as:
Cedefop (2021). *Understanding technological change and skill needs: big data and artificial intelligence methods. Cedefop practical guide 2.* Luxembourg: Publications Office. http://data.europa.eu/doi/10.2801/144881

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (http://europa.eu).

Luxembourg:
Publications Office of the European Union, 2021

PDF
ISBN: 978-92-896-3238-6
doi:10.2801/144881
TI-05-21-028-EN-N

Designed by Missing Element Prague

The **European Centre for the Development of Vocational Training** (Cedefop) is the European Union's reference centre for vocational education and training, skills and qualifications. We provide information, research, analyses and evidence on vocational education and training, skills and qualifications for policy-making in the EU Member States.

Cedefop was originally established in 1975 by Council Regulation (EEC) No 337/75. This decision was repealed in 2019 by Regulation (EU) 2019/128 establishing Cedefop as a Union Agency with a renewed mandate.

Europe 123, Thessaloniki (Pylea), GREECE
Postal address: Cedefop service post, 57001 Thermi, GREECE
Tel. +30 2310490111, Fax +30 2310490020
Email: info@cedefop.europa.eu
www.cedefop.europa.eu

Jürgen Siebel, *Executive Director*
Barbara Dorn, *Chair of the Management Board*

Foreword

Cedefop has been at the forefront of developing robust skills anticipation methods and skills intelligence tools for the European Union for more than a decade. The European skills forecast and the European skills and jobs survey shed light on how the labour market, skill needs and jobs are developing and help signal potential skills bottlenecks. Cedefop's big data analysis of online job advertisements provides detailed and real-time skills intelligence capturing which skills have currency in job markets. Cedefop has used skills foresight to develop stakeholder-backed policy roadmaps aimed at strengthening national skills anticipation and matching systems. Complementing quantitative skills analysis and intelligence, qualitative insight into skills policies and measures also contributes to evidence-based policy-making.

The continuing development of national skills intelligence systems and approaches has helped strengthen the feedback loops between the labour market and vocational education and training (VET) and skills policy. In the coming years, we need to be more ambitious. Our vision for 'Skills intelligence 2.0' is information that is more actionable: detailed and relevant, better contextualised, timelier, and better communicated. Making sense of trends and fostering capacity to act on them means combining sources and approaches – skill surveys, skills forecasting, skill foresight, big data analyses, and others – and exploring synergies. This gives policy-makers the means to separate noise from signal and supports employers and citizens in making decisions in line with the new realities in the world of work.

It is no surprise that skills intelligence is a key priority in the 2020 European skills agenda. Reliable and fit-for-purpose labour market and skills intelligence has enormous value in times of rapid change and transformation. In a context of fast-paced digital advancements, such as artificial intelligence and advanced robotics, and other megatrends such as population ageing and the green transition, VET and skills policies should become more proactive. To prepare new generations of learners and to support people in making and shaping career transitions, reliable skills intelligence is indispensable.

This publication is the second in a series of practical skills anticipation guides for policy-makers and analysts. The guides present a rich mosaic of conventional and emerging methods for identifying technological change and its impact on skills. Systematically presenting the merits and disadvantages of different methods, they show no single approach can provide all the answers. Apart from reliable data and sound methods, creativity, holistic thinking and using collective wisdom to actively shape the future are key building blocks of skills intelligence 2.0.

This second guide focuses on automated methods of anticipating changing technologies and skill demands: big data and AI-driven analyses. We trust the practical insights it provides will prove to be useful in your context.

Jürgen Siebel
Executive Director

Antonio Ranieri
Ad interim head of department for skills and labour market

Acknowledgements

This publication was produced by Cedefop, Department for skills and labour market, under the supervision of Antonio Ranieri, Head of department ad interim, and Jasper van Loo. Konstantinos Pouliakas developed the concept and edited this guide as part of Cedefop's *Digitalisation, AI and future of work and Anticipating and matching skills* projects. The publication was peer-reviewed by Cedefop expert Jasper van Loo.

The Cedefop project Governance of EU skills anticipation and matching systems: in-depth country reviews (contract No 2016-FWC7/AO/DSL/KPOUL-JVLOO/skills-governance/005/16), under which this guide was developed, was managed by Cedefop experts Konstantinos Pouliakas and Jasper van Loo.

Cedefop wishes to acknowledge the research and services of its contracted institution, the Fondazione Giacomo Brodolini (FGB) (lead partner) in consortium with Economix Research and Consulting, Cambridge Econometrics Ltd, and Panteia B.V. Cedefop wishes to thank Liga Baltina (FGB), Terence Hogarth (FGB and University of Warwick) and the following experts for their contribution to drafting the report: Riccardo Apreda, Filippo Chiarello, Gualtiero Fantoni and Pietro Manfredi.

Contents

Tables, figures and boxes

Tables

Figures

Boxes

CHAPTER 1.

Conventional and automated skills anticipation

1.1. Technological change and skill needs

Technological change and digitalisation are transforming the nature of the employment relationship (for example rise of platform work, see Cedefop, 2020) and drive the automation of work (Frey and Osborne, 2013; Arntz et al., 2017; Nedelkoska and Quintini, 2018; Pouliakas, 2018). They also alter the skill content of those jobs that remain (Deming and Noray, 2020) and are drivers of new task and job creation (Acemoglu and Restrepo, 2019; McGuinness et al., 2019; Freeman et al., 2020).

Popular media and the vocational education and training (VET) and skills policy discourse highlight that the world of work is being impacted by a fourth industrial revolution. It is being transformed by industry 4.0, advanced robotics, artificial intelligence (AI), the internet of things (IoT) and other emerging technologies in a way that is more profound than previous waves of change. These trends have been impacting the labour market increasingly in the past decade. Cedefop's first European skills and jobs survey (ESJS) revealed, back in 2014, that 43% of EU adult employees had recently experienced new technologies at work, such as new machines and information and communications technology (ICT) systems.

Several other macrotrends are driving the future demand for skills. Climate change and the trend towards greening the economy, demographic change and migration are also reshaping the world of work. Notwithstanding this, looking at how technological progress and innovation impacts skills needs is important, as it is widely viewed as the most dynamic megatrend shaping the future of work (Brynjolfsson and McAfee, 2014). Policy has also become increasingly concerned with emerging (digital) skill gaps and skills obsolescence affecting workers and the need to step up investment in lifelong learning to mitigate inequalities due to the growing digital divide (Cedefop, 2016).

1.2. Skills assessment and anticipation methods

To understand the extent to which technology is transforming the world of work, it is necessary to measure its magnitude and impact on skills demand. Labour market and skills intelligence (LMSI – often referred to as skills intelligence) provides such information and – provided that it is based on sound approaches and methods – can serve the needs of those responsible for reacting to changing skill needs.

While analysts and experts have a range of different skills assessment and anticipation methods at their disposal, identifying and anticipating the pace of technological change in labour markets – in particular in times of rapid change – is challenging. With the process of predicting the future becoming more complicated and perhaps less certain, the range of methods and tools available to those involved in such exercises has become more varied and sophisticated.

Table 1. **Tools for carrying out skills assessment and anticipation**

Type of activity	Data collected
Descriptive statistics/ stock-taking	Estimates of overall demand and supply of skills and technology use, often based on collating data from various sources (e.g. sector skill studies)
Quantitative forecasting	Forecasting or projecting future demand for skills, typically using econometric modelling
Skills and jobs surveys (questionnaire surveys)	Assessments of demand for, and supply of, skills and technology use, usually with an assessment of the extent to which demand and supply are in balance
Graduate tracer studies	Using matched administrative datasets or surveys to track people through education and the labour market to see how the former influences the latter
Qualitative research	Use of non-quantitative techniques to gauge in-depth information about current and future skill demand/supply and technology trends, e.g. via company case studies, use of focus groups
Foresight	Critical thinking about the future of skills supply/demand and technology trends, using participatory methodologies
Big data	Use of web sourcing, combined with text mining and machine learning approaches, to collect and classify data about skills, vacancies, technologies, etc.

Source: Cedefop classification.

Table 1 summarises some of the main methods that can be used to gather information on skills needs. Four are particularly important. These are those that:

(a) rely on asking questions of key stakeholders (questionnaire surveys of employers' and employees' skill needs and experience of technological change);
(b) produce quantitative estimates of future skill demands, by extrapolating past trends and modelling expected developments;
(c) source big data on new technologies and skills from a variety of online sources (e.g. job portals, CVs, social media, patents, scientific databases);
(d) use non-quantitative techniques, relying mostly on participatory stakeholder approaches to gauge in-depth information about the state of current and future skill demand and supply.

1.3. Purpose of this guide

This second Cedefop practical guide on understanding the impact of technological change on skill demand ([1]) focuses on big data and AI-driven methods for analysing current and emerging technologies and skill needs. Apart from looking at online job advertisements, it also describes how patent data, scientific databases and online course websites can be used to derive information on technological change and emerging skill needs.

With the increasing use of big data and AI analysis – essentially the application of text mining, natural language processing and machine learning techniques – analysts have at their disposal a new method in their skills anticipation toolkit. Big data-driven skill analysis is a so-called non-participatory approach to skills anticipation, as it does not involve stakeholders in deriving estimates of the impact of technological change on skills. Cedefop's skills online vacancy analysis tool for Europe (Skills-OVATE) presents skills information gathered from online job portals and classified

([1]) This guide is the second of a series. See the other two guides:
Cedefop (2021a) *Understanding technological change and skill needs: skills surveys and skills forecasting*. *Cedefop practical guide 1.* Luxembourg: Publications Office.
http://data.europa.eu/doi/10.2801/212891
Cedefop (2021b). *Understanding technological change and skill needs: technology and skills foresight. Cedefop practical guide 3.* Luxembourg: Publications Office.
http://data.europa.eu/doi/10.2801/307925

using international skills and jobs classifications. Presenting detailed granular information on skills demanded in jobs in countries, regions, sectors and occupation in (quasi-) real time, the tool showcases the potential of such information for policy purposes. It can enrich and complement conventional skills forecasting approaches, which typically provide more aggregate information at sectoral or occupational level.

This second 'how-to' guide builds on the first one on conventional skills assessment and anticipation methods, which covers skills forecasts and skills (employer or worker) surveys. These latter methods rely on the collection of representative labour market information and analysis, using statistically robust techniques. This guide also complements the third one on participatory technology and skill foresight methods. Such qualitative methods heavily rely on stakeholder involvement in assessing and reflecting on the scientific evidence relating to technology's impact on skills.

The guide is structured as follows. Chapter 2 discusses the techniques, tools and processes of implementing automated technology and skills analyses. It explains how knowledge can be extracted from texts and documents to detect emerging technologies and skill needs. Chapter 3 presents several applications of such web-based skills anticipation methods. Sources covered are online job advertisements, patent data, scientific repositories and information on online courses offered by providers. Chapter 4 concludes with a review of the advantages and pitfalls of big data and AI methods. It also provides reflection on which skills anticipation methods are most suited in particular situations and the reasons why this is the case.

Box 1. **Cedefop's 'how-to' guides on understanding technological change and skills demand**

> The purpose of Cedefop's short 'how-to' guides is to provide those charged with a responsibility for undertaking skills assessment and anticipation with the means to deal with the uncertainty of technological change and its impact on skill needs.
> As the process of predicting the future becomes more complicated and less deterministic, the range of tools available to those involved in skills anticipation has become more varied and sophisticated. The Cedefop guides aim to showcase to policy-makers and interested analysts how various techniques or methodological tools can be readily applied by carefully considering the associated pitfalls and rewards of doing so.

The guides provide targeted information on how interested analysts adopt and can implement either conventional labour market and skills intelligence methods, such as skills surveys and skill forecasts; automated methods reliant on big data and artificial intelligence techniques; or technology foresight methods, to detect emerging skill needs related to technological change. Implicit in the guides is recognition that no one methodology is likely to provide all the answers and the challenge for analysts is to bring together outputs from different approaches to skills anticipation. The guides build on the existing compendium of guides on skills anticipation produced by the ETF, Cedefop and the ILO, as well as several previous Cedefop reports on skills anticipation methods ([2]). But they are distinct from previously published methodological handbooks or guides, in that they are explicitly concerned with the process of identifying technological (digital) change, a key driver of changing skill needs.

Source: Cedefop.

([2]) For instance, see Cedefop, 2013; 2015 and Cedefop's project Anticipating and matching skills. The 2021 publication *Perspectives on policy and practice: tapping into the potential of big data for skills policy* of the inter-agency group on technical and vocational education (IAG-TVET) complements this guide and provides useful reflections for both developed and developing countries.

CHAPTER 2.

Automated skills anticipation methods: opportunities, challenges and techniques

Using automated methods to analyse (quasi-) real-time labour market and skills information, extracted from online sources, can be of considerable value for training providers, policy-makers, employers and employees:

(a) policy-makers need timely insights on future technologies and skill needs to ensure that policies and measures are in line with changing labour market demand;
(b) training providers are keen on having fast access to data on emerging technologies, skill demand and trending and emerging jobs or occupations to inform training programme design and updates;
(c) employers need information on the skills their employees need to adapt to impending and future technological changes;
(d) jobseekers/career counsellors can benefit from information on skills needs associated with technological change.

Chapter 2 provides a critical and structured overview of how automated knowledge extraction techniques can be used to identify and analyse emerging technologies and skill needs. These techniques (for example machine learning, web scraping and text mining) can be used to extract information from documents and texts. Chapter 3 presents examples of how such techniques are applied to online job advertisement data and other data types (such as patent or scientific bibliographic databases).

Sections 2.1 and 2.2 that follow explain how to use knowledge extraction software efficiently to discover emerging technologies and skill needs, covering which data sources can be exploited and how, and what the strengths and drawbacks are of different approaches. The focus is on the knowledge extraction process and not on the application of data collection methods, such as web crawling, scraping or access to proprietary databases.

2.1. Extracting knowledge from text

Skills surveys and skills forecasts provide information that is grounded in the past (Cedefop, 2021a). Surveys (among employers and employees) tend to ask about recent or expected changes and skills forecasting extrapolates past trends to provide an image of the future. But what if the past is not such a good guide to the future as it once was?

Such claims should not be overestimated. Firms tend to be cautious about investing in new technologies and their implementation takes time and effort: the pace of technological change can be slower than some commentators suggest. Nevertheless, technological breakthroughs can introduce structural breaks in time series of labour market aggregates. It is also important to have the possibility to identify new jobs and new skills, especially those not captured in existing classifications such as European skills, competences and occupations (ESCO) and international standard of classification of occupations (ISCO). One example is the growing interest in the skills implications of carbon footprint reduction technologies and the diffusion of AI in workplaces. It is clear that green or AI skills, however they might be defined, are not well reflected in classifications, such as ESCO/ISCO.

Automated techniques, based on textual knowledge extraction, can inform the process of anticipating future change in the labour market. The knowledge extraction software used for this purpose contains a class of algorithms which can:

(a) decompose any text written in natural language in its basic components;
(b) identify and extract those entities that the analyst has indicated as relevant (for example skills and competences related to a particular technology);
(c) recognise relationships and correlations (for example, synonymy) between the selected objects.

Human intervention is limited to training the software how to recognise the elements of interest, for instance by setting rules or by providing examples. Since such techniques can be applied to large volumes of documents, there is an overlap with the topic of big data analysis.

The main contribution of such methods is their ability to find a signal among the noise. Until recently, the performance of knowledge extraction algorithms was inadequate for identifying future skill needs. The reliability of this kind of software has steadily increased over time as a consequence of:

(a) progress in natural language processing (NLP) techniques, which has reached a precision of 95% in tasks related to understanding human language;
(b) the availability of larger databases which improve signal detection and training;
(c) more powerful computers, which enable the application of so-called 'deep' learning/neural networks, techniques which can tackle tasks where the rules are not known *a priori*.

The high precision achieved makes knowledge extraction software an interesting and powerful skills anticipation tool. The automated techniques made possible by knowledge extraction software:
(a) allow processing huge amounts of data quickly and in a cost-effective way. Such analysis would be difficult, if not impossible, using manual techniques (such as conventional content analysis) and, in any case, would be prohibitively expensive;
(b) allow identifying even non-obvious and hidden patterns, connecting pieces of information scattered among many different and distant sources, and detecting weak or emerging signals. Some of those tasks are not achievable by human analysts due to the large number of documents involved;
(c) can deliver standardised and reliable results. Findings based on other methodologies would not necessarily be consistent over time and across sectors or may be operator-dependent;
(d) make it easier to manage, understand and share data. The software can detect and distil the knowledge that humans have included in many different documents (for example reports, papers, online job advertisements, online CVs) and condense it in concise indicators, infographics or knowledge bases;
(e) deliver objective signals and correlations, although it is important to be aware of limitations, such as data completeness, reliability and representativeness of sources. Human expert statements may be subject to cognitive biases.

As with any scientific method, approach or tool, knowledge extraction software also has its limitations. Specifically:
(a) any information of interest must be contained in written documents in digital format;

(b) the documents must be accessible so that they can be processed;
(c) the available data may not be suitable at face value for finding meaningful signals. They must be cleaned, complete and relevant, otherwise results may be misleading (garbage in, garbage out);
(d) there may be intrinsic bias linked to the type of sources available, even if the issues stated above can be solved or mitigated. There is, for example, plenty of written information about technology-related jobs and competences, but much less on relatively low-level or generic skills that form part of jobs which are also affected by technological change.

Automated techniques and tools, therefore, cannot substitute human experts. They provide a complementary point of view which needs to be cross-checked and integrated with results based on other approaches.

2.2. A primer in automated data analysis

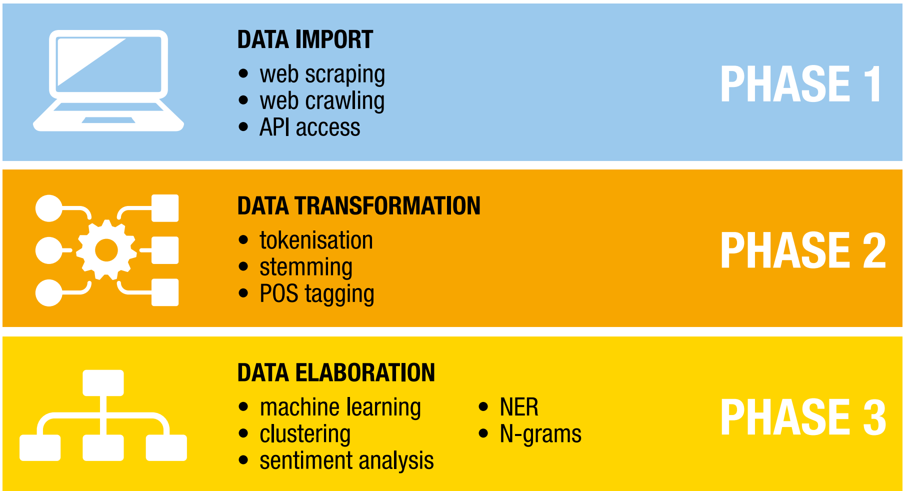
Most contemporary automated techniques can be classified and explained according to the phase in which they are used:
(a) data import phase: the collection of data to be processed in subsequent phases;
(b) data transformation phase: turning unstructured data into structured, processed data;
(c) data elaboration phase: extracting information of interest, such as trends, semantic relations between words, groups of words gathered according to specific similarities.

Figure 1 shows the clusters of techniques contained in each phase.

Data imported can be of two types:
(a) structured (numbers, dates, author keywords);
(b) unstructured (videos, audios, texts).

The process of importing data is strictly connected to the data source used. The way data are downloaded or extracted depends on how well-structured a data source is in the first instance. For example, retrieving information from websites can be complicated, since not all have a clear structure behind them and not all give access through application programming interfaces (APIs).

Figure 1. **Main phases of automated data analysis**



Source: Cedefop.

There are three online data retrieval techniques (Cedefop, 2019a):

(a) web scraping is used to extract structured data from websites [3]. Web scraping implies that data are already in a structured form on the web page and can be extracted by knowing the exact position of the information;

(b) web crawling uses a programmed robot (crawler) to browse web portals systematically and download their pages. Web crawling is a main component of web scraping, fetching pages for later processing. Although crawlers can collect websites in an automated way, they also gather much noise (irrelevant content) and more effort is needed to clean the data before further processing;

(c) API access gives the possibility of bulk downloading of information and online content from websites. It requires a formal agreement with the website owner, which is not always granted. Data collected via an API typically have higher quality compared to web scraping.

Data transformation includes format conversions, transforming unstructured into structured data and data cleaning.

---

(3) Readers interested in web scraping can visit: https://medium.com/the-andela-way/learn-how-to-scrape-the-web-2a7cc488e017

This guide focuses on the transformation of textual data, as this data type is most commonly used for automated analysis related to new technologies and skills. Natural language processing tools can be used to help computers interpret and manipulate human language (Bird et al., 2009). Considering the large amount of unstructured data that is generated online every day, automated methods, such as NLP are central to analysing textual data efficiently.

NLP also helps in resolving ambiguities in human language. Human beings tend to express themselves in various ways, using different dialects and abbreviations when speaking and writing. Written documents may contain misspelling and punctuation may be omitted; when speaking, there may be problems related to broken speech patterns and borrowed terms from other languages (Copestake, 2004) ([4]).

The most common tasks NLP is capable of are (Bai et al., 2009):

(a) tokenisation: splitting text into single words which are commonly defined as tokens. These single words can then be used as input to other analysis, such as understanding the existing syntactic relationships present in the text;
(b) stemming: chopping off the ends of words, considering a list of common suffixes. This depends on the different forms a word can have for grammatical reasons;
(c) lemmatisation: returning the dictionary form of text, known as lemma (for example was→ be, better→ good);
(d) part-of-speech tagging: also known as POS tagging, this task refers to the process of associating tags to words, based on their definitions or roles inside the text or phrase. A tag, for example, can be 'noun' or 'verb', but this is not always a straightforward task, since a word could have different meanings or tags, depending on the context and word order.

Once data has been structured and is in a format that can be processed, various techniques can be applied during data elaboration to extract information.

Machine learning (ML) algorithms undertake the major part of this process nowadays. ML is based on the idea that systems can learn from

([4]) Further information on NLP and how it works: https://www.sas.com/it_it/insights/analytics/what-is-natural-language-processing-nlp.html data, identifying patterns that can later be used to analyse previously unseen data (Michie, 1968).

ML techniques can be supervised or unsupervised. Supervised ML relies on a training set of data. For example, if the aim is to train a ML model to recognise images of dogs and cats, the computer needs pictures labelled as 'dog' and 'cat'. Supervised learning algorithms are particularly useful for document classification.

Classifiers are many and can be positioned in a spectrum of accuracy (the quality of the classification task) and interpretability (the possibility for a human to understand the decisions of the algorithm). The following classification algorithms can be found at either extreme of this spectrum:

(a) decision-tree classifiers ([5]): high interpretability and relatively low accuracy. In order to classify documents, this ML algorithm uses possible test questions and conditions organised in a tree structure. In the decision tree, the root and internal nodes contain attribute test conditions to separate documents with different characteristics. Based on the answers given to the test questions, a document is then assigned to a specific class;

(b) neural-network classifiers ([6]): high accuracy and low interpretability. A neural network is made up of nodes, denoted as neurons, which receive input data and turn it into output. Each neuron has a specific weight, depending on the input data. Thanks to a process, called learning phase, they can adjust their weights in such a way that new documents can be classified. If there is more than one layer of neurons, the process falls into the category of deep learning (Huang and Lippmann, 1988; Goodfellow et al., 2016).

Machine learning classifiers that lie in between the two extremes of the spectrum include rule-based algorithms, support vector machines and Bayesian classifiers (Domingos, 2015; James et al., 2013).

Unsupervised ML refers to methods whereby the learning algorithm is deprived of labelled data. Clustering algorithms, for instance, separate input data into groups, based on particular features that are not manually

([5]) Further information on how a decision tree classifier works: http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/decisionTree.html

([6]) Further information on how neural network classifiers work: https://towardsdatascience.com/classification-using-neural-networks-b8e98f3a904f annotated but automatically identified by the algorithm itself. Given the unique characteristics of textual data, traditional clustering algorithms, such as K-means or hierarchical methods, do not typically work well. Specialised techniques have been designed for text clustering. In this case, the groups of text are formed-based, for example, on keywords, occurrences and cooccurrences of words and semantic relationships between terms. This is commonly referred to as topic modelling, as it discovers the abstract topics that occur in a set of documents or in text (Rosell and Csc, 2008).

Figure 2. **Example of named entity recognition application**

temp.py*

```
#importing the necessary package for NER
import spacy
text = "In 2001, Willian S. Cleveland introduced data science"
spacy_nlp = spacy.load('en')
document = spacy_nlp(text)

print('Original Sentence: %s' % (text))

for element in document.ents:
    print('Type: %s, Value: %s' % (element.label_, element))


Original Sentence: In 2001, William S.Cleveland introduced data science
Type: DATE, Value: 2001
Type: PERSON, Value: William S. Cleveland
```

NB: The figure provides a simple example of named entity recognition (NER) code using the Python programming language to extract entities from the phrase: 'In 2001, William S. Cleveland introduced data science'.
Source: Cedefop.

One of the methods most commonly used to undertake the task of text elaboration using ML techniques includes named entity recognition (NER). NER is a supervised learning approach that allows the identification of entity names, such as people, organisations, places, temporal expressions or numerical expressions. NER, as can be seen in Figure 2, identifies the type of entity within text; for example, '2001' is an entity of the type DATE, while 'William S. Cleveland' is an entity of the type PERSON. Not all entities are always recognised. The algorithm would have problems if the name of a person was written in lowercase, since this entity usually has initial capital letters.

CHAPTER 3.
Applications of automated skills anticipation methods

3.1. Identifying emerging technologies and skills from online job advertisements

Tools that analyse millions of online job postings to provide (quasi-) real time data on technologies, tools and the skills recruiters want have been subject to growing interest. Such tools typically apply modern big data and AI analysis techniques, such web scraping, NLP and other ML algorithms ([7]).

The data production system (DPS) of such tools typically has four phases (Figure 3):

(a) data collection;
(b) pre-processing;
(c) information extraction;
(d) data classification.

Data collection has two sub-phases:

(a) selecting the right websites/portals to extract job vacancies. To maximise the quality of the information extracted, websites are ranked and prioritised according to the information they provide (for example vacancy release dates; frequency and regularity of vacancy updates; territorial, sectoral, and occupational coverage of vacancies);
(b) downloading data from the identified websites, using web scraping, web crawling or via APIs.

Since bulk downloading of data from websites is not always possible without permission, it is good practice to inform online job advertisement portal owners about the intended data collection. In some cases, formal agreements between the user and the portals should be concluded.

([7]) See, for instance, Cedefop's Skills-OVATE (Cedefop, 2019a), a web tool using job ads collected from portals in all EU Member States to present (quasi-) real-time skills intelligence.

After downloading the data, the language of each job ad should be detected, especially when information from different countries is being extracted. Some online job ads are published in a language different from (one of) the official language(s) of the country. Developing an algorithm to recognise and process languages is a key step in the process.

Since recruitment websites are not designed to provide data suitable for analysis but aim to attract the most suitable candidates for a job position, other essential parts of data pre-processing are:

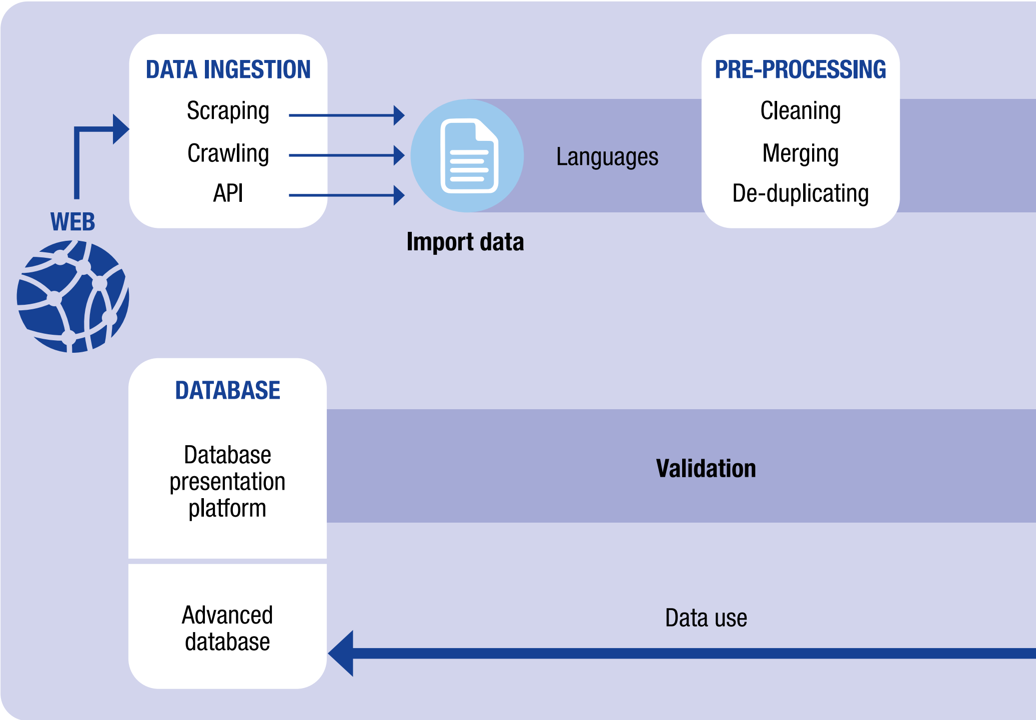
(a) data cleaning to remove unnecessary information, such as advertisements, unticked options in drop-down menus, company profile presentations, layout elements and logos, pictures;
(b) de-duplication to remove (after first merging) duplicate job offers, given that many job ads are posted on several portals. A job advertisement may be considered a duplicate if the job location and the description are the same as another posting on a different website.

Information extraction and classification typically takes place once data have been pre-processed. ML algorithms are used to match the content of the downloaded job advertisements to education, skills and labour market ontologies/classifications (for example, ESCO/ISCO for occupations and skills, statistical classification of economic activities in the European Community (NACE) for sectors, nomenclature of territorial units for statistics (NUTS) for places of work, international standard classification of education (ISCED) for education levels). Alternatively, custom ontologies can be developed from the information in the job offers (such as contract type, skills, salary). These can be used to develop terms and synonyms which are not yet included in existing ontologies, providing valuable information on emerging technologies, jobs or skills.

The ML algorithms have to be trained using suitable training sets (for instance, previously labelled occupation groups corresponding to different job titles) to fit best the variables and the language used. Once the ML model is trained, its accuracy must be tested. It is good practice to involve experts tasked with regularly validating the results of the ML classification ('man in the loop') and making corrections to improve the algorithm's accuracy.

Once the data classification process is completed, it is possible to store the processed data in a multidimensional database for data navigation and analysis. This can include developing insights into current and emerging technologies, jobs and skills in demand by employers.

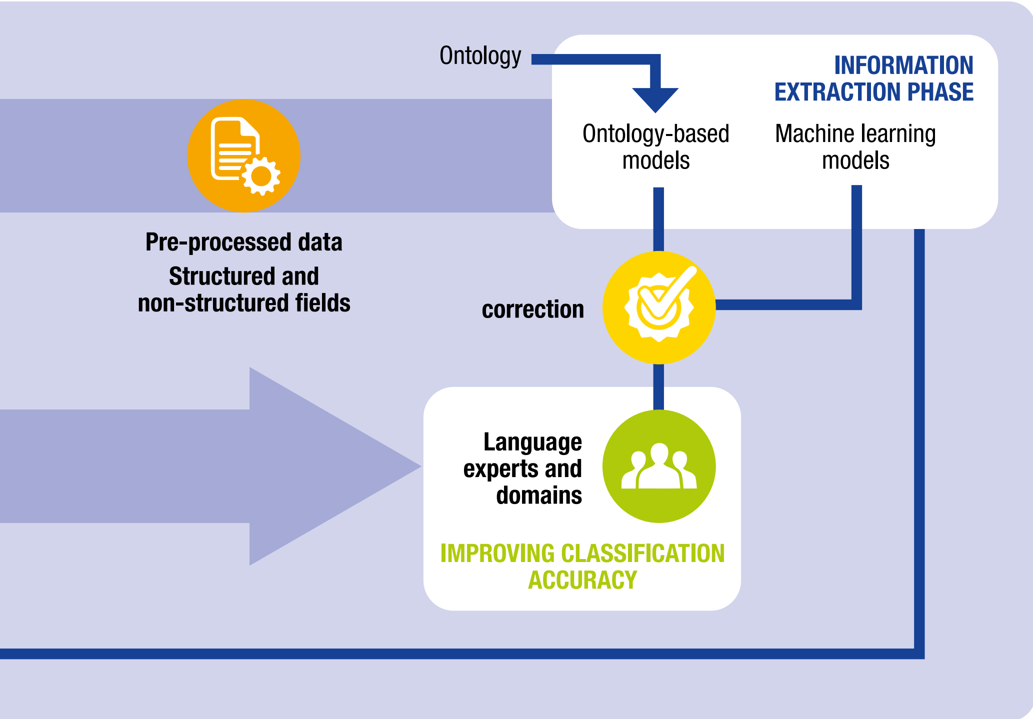
Figure 3. **A big data production system**



Source: Cedefop.

Box 2. **Representativeness challenges of big data based on online job advertisements**

- Not all jobs are advertised online and not all job advertisements lead to actual job openings. The nature and maturity of the online labour market is shaped by the size of the informal economy, cultural factors and digital divides in internet connectivity and digital skills.
- Employers tend to use occupation-specific hiring strategies. High-level professionals are often recruited via dedicated or privately owned portals or job-hunting. Public employment service (PES) portals are typically used for medium- or low-skilled jobs.
- Some jobs are rarely advertised online at all, because word of mouth or a notice in a shop window are more effective and cheaper solutions to recruit staff.



- Some portals restrict access to particular groups, such as the registered unemployed in the case of several national PES portals in the EU.
- Skills requested in the online job advertisements (OJAs) are not skills profiles. Employers emphasise the skills that give candidates a competitive edge and those that may help reduce the pool of available applicants. Lack of common standards and tools for describing skills in OJAs causes selectivity and variation in the skills indicated.

Source: Cedefop, 2019a; 2019b.

It is important to bear in mind that the use of automated knowledge extraction techniques, applied to online job advertisement data, provides only a piecemeal and non-representative picture of emerging technologies and skills in labour markets (Box 2). The information extracted reflects the type of technologies, tools and skills that employers request from job applicants, but this knowledge is likely to be bounded by the technology they currently use. It is a well-established fact that the capacity of a heterogeneous group of employers to foresee new technological developments is imperfect and that they have an incentive to overrepresent their skill needs (Gambin et al., 2016). Moreover, data extracted from online job advertisements is often fraught with statistical and selection biases and may only be loosely related to the actual skill needs in jobs. Not all skills are listed in vacancies, since job-specific skills may be taken for granted and transversal skills may be emphasised instead. Online job postings also serve the function of a 'beauty contest', to attract potential job applicants to the recruitment stage and to overcome the adverse-selection problem associated with their unobserved abilities (Cedefop, 2019b; Akerlof, 1970). It is therefore possible for online recruiting to encourage superfluous vacancy postings by employers and inferior skills matching outcomes due to a large share of unsuitable applicants per vacancy (Gürtzgen et al., 2021).

3.2. Analysis of patents and scientific papers

In addition to online job advertisement data, researchers have used other types of documents, such as patents, scientific papers and Wikipedia, to gain insights into new or future-oriented technologies and skills.

Patents granted and patent applications are a unique source of technological information. Patent data are publicly available free of charge and there are services, such as the European Patent Office (EPO) Espacenet, which collects patent information worldwide and stores it in a single easily accessible repository. Patents are also retrievable from Google patents or free patents online.

About 80% of the technical information contained in patent documents is not available elsewhere (Kütt and Schmiemann, 1998; Terragno, 1979). Even though the proliferation of the internet has probably reduced this proportion, patent data remain a source of information that complements the traditional scientific literature. According to the European Patent Office, reasons why information published in patent documents is not available in scientific documents or elsewhere are (Golzio, 2012):

(a) the early publication of information on inventions can irreparably compromise patentability;

(b) the content of a scientific paper differs from that required for patent documents. A scientific paper usually requires more detailed disclosure of information, which gives competitors an advantage and creates opportunities to reproduce the invention and modify it (generating new patents);
(c) sometimes the technical information present in a patent may not trigger the interest of a journal and its editorial committee.

Although there is an established literature on technological maps and foresight based on patent analysis (Daim et al., 2006; Cagnin et al., 2013), much less is known about the relationship between patents and skills. The study by Hwang and colleagues (2015), which developed a methodology to assess the suitability of patent information and analyse patent applications to identify future skills in the information security sector, is a notable exception.

Patents are a significant cost for companies and are filed to protect only technological innovations that would otherwise be at risk of poaching or replication. Enterprises applying for a patent possess skills related to that technology, for example research and development/logistics/manufacturing skills. Therefore, the correlation between patents and the skills requested is expected to be strong, providing a potentially biased view of emerging technologies across the wider population of businesses in an economy.

To extract the correct information from patents, it is important to select the right subset to be processed. In order to avoid losing knowledge, it is necessary to consider all pertinent documents: in information retrieval, such a parameter of completeness is called recall. It is also crucial to exclude unrelated documents to avoid including misleading information (maximising the parameter called precision).

To process patent data and extract valuable information on technologies, text-mining algorithms can be applied to identify keywords that embed the required information. These words can be found in any part of the document, from the title itself to the least important of references. It is important to consider the position of information within the document and the relationship with other text strings. This is typically done using named entity recognition techniques (Figure 4).

Figure 4. **Finding technology-related keywords in patent data**

Laser cutting and processing of display glass compositions

Abstract

The present invention relates to a laser cutting technology for cutting and separating thin substrates of transparent materials, for example to cutting of display glass compositions mainly used for production of Thin Film Transistors (TFT) devices. The described laser process can be used to make straight cuts, for example at a speed of >0.25m/sec, to cut sharp radii outer corners (<1 mm), and to create arbitrary curved shapes including forming interior holes and slots. A method of laser processing an alkaline earth boro-aluminosilicate glass composite workpiece includes focusing a pulsed laser beam into a focal line. The pulsed laser produces pulse bursts with 5-20 pulses per pulse burst and pulse burst energy of 300-600 micro Joules per burst. The focal line is directed into the glass composite workpiece, generating induced absorption within the material. The workpiece and the laser beam are translated relative to each other to form a plurality of defect lines along a contour, with adjacent defect lines have a spacing of 5-15 microns.

Classifications

C03B33/0222 Scoring using a focussed radiation beam e.g. laser

Source: Hackert, 2016.

The identification of keywords is an important issue in text mining, since it requires high-quality technology to detect them automatically and to measure their relevance reliably in any type of document. There are many programmes available for this task, such as R studio, Python natural language toolkit (NLTK) library, IBM SPSS, RapidMiner and Google cloud natural language.

To gain better understanding of the keywords extracted with the help of one of these programmes, it is important to use clustering algorithms to group keywords and identify correspondences, similarities and associations to particular concepts.

Automated analysis of patents can be used to identify trends that help to understand whether a particular technology is growing or decreasing over time. A critical step in this process is setting the time frame to be used. Using less than a 10-year period is likely to result in distorted and inexact findings, since there would not be enough data for the software to detect trends. It is also important to be aware of the 18-month gap between filing a patent and its publication (and hence visibility). Consequently, when making analyses of patents it is important not to consider the final two years, since they would lead to incorrect or distorted results.

To understand how time series graphs can be generated to extract relevant information on future technologies, an illustrative case study is presented in Table 2. In this example, a limited number of patents related to four technologies is considered and their growth in numbers is recorded. To determine the trend in a technology, patents must be downloaded and stored in a table in chronological order (Table 2).

Table 2. **Data tabulation of patents**

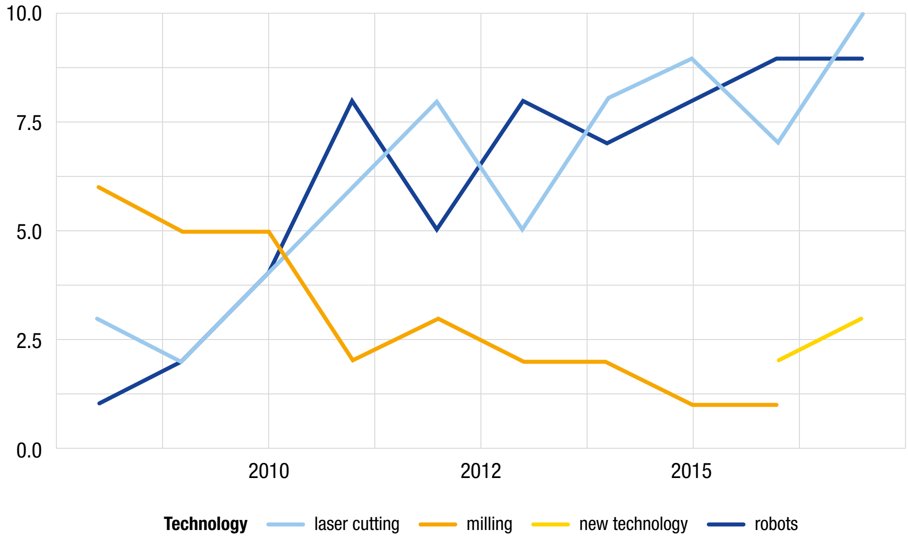
Technology	Number of patents	Year
Laser cutting	3	2008
Robots	1	
Milling	6	
New technology	0	
Laser cutting	2	2009
Robots	2	
Milling	5	
New technology	0	
Laser cutting	4	2010
Robots	4	
Milling	5	
New technology	0	
Laser cutting	6	2011
Robots	8	
Milling	2	
New technology	0	

Source: Cedefop.

After collection, the data can be visualised (Figure 5) ([8]). Apart from understanding how the use of a particular technology develops over time, these time-series graphs can be used to determine which skills might increase in demand. In the example, there will likely be an increasing demand for skills associated with laser cutting and robot technologies, but not for milling.

([8]) Figure 5 was generated using the R studio and its associated data visualisation package 'ggplot'. Further information at: http://r-statistics.co/R-Tutorial.html, section ggplot2.

Figure 5. **Number of patents published over time for each technology group**



Source: Cedefop.

Trend analysis can also help identify emerging technologies, in the early stages of development, which could affect the demand for new skills or the mix of new and old skills, resulting in the creation of new job profiles. The S-curves ([9]) visualise, in general terms, the diffusion of technology over time. They depict the correspondence between the number of patents in time and the concept of maturity and growth in performance of a specific technology. The figure illustrates the theoretical situation where a new technology eventually supersedes an existing one.

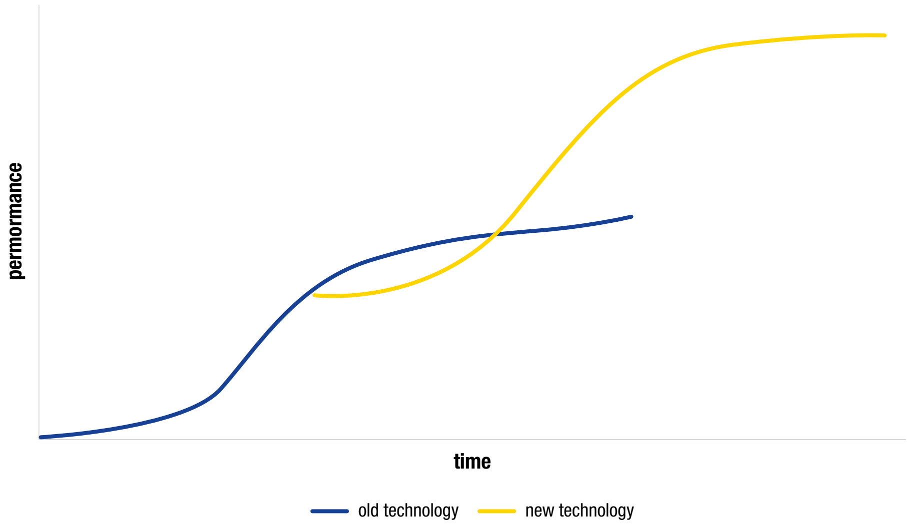
3.3. Analysis of scientific literature

Similar to patents, existing and possible new technologies and skills can be detected from information contained in scientific papers. An indicator based

---

([9]) For an explanation of the S-curve and its associated theory see: https://www.youtube.com/watch?v=Rm1v7lI2iMk

Figure 6. **Technology S-curve**



Source: Adner and Kapoor, 2016.

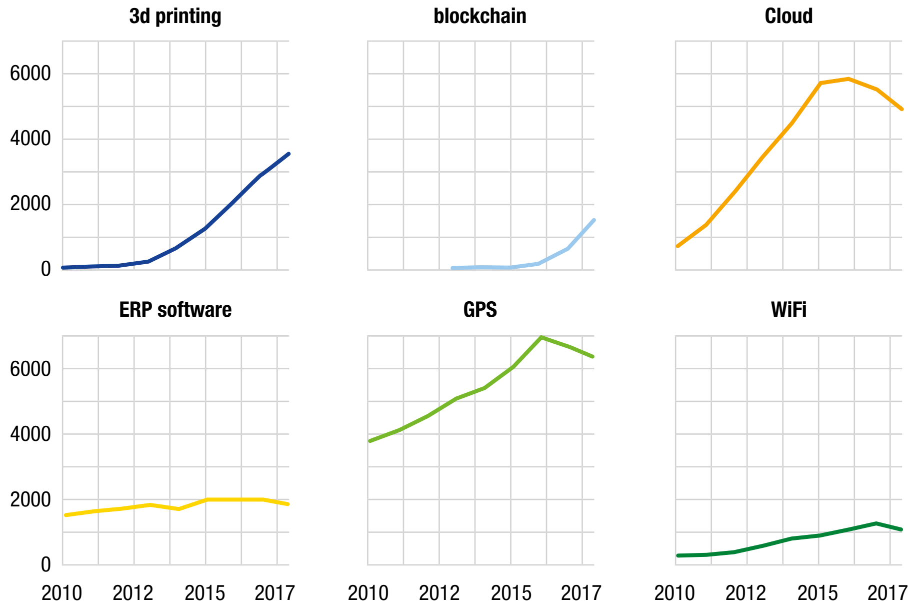
on a growing trend in the number of papers related to a specific technology over time could be constructed. The related skills could be identified using text analysis to search for terms which resemble a competence linked to that technology. Web of science, Google scholar and Scopus are among the many scientific paper databases with an API.

Given their similar structure, the process applied to patents can be replicated for scientific papers. The latter can provide a complete mapping of technologies even though, compared with patent data, they might not give as much insight into emerging technologies. This is the case because:

(a) the publication of a scientific paper may or may not guarantee an industrial application;
(b) a paper may or may not bring out an invention;
(c) papers could simply be a review of previous research or discoveries.

Figure 7 provides a case study showing how to map technology trends over time from scientific databases and how they can be interpreted applying a future-oriented perspective. The analysis is based on scientific papers that contained some reference to the industry 4.0 paradigm. The main objective is to identify technologies in use and track change. The approach used consists of extracting terms related to industry 4.0 technologies from the corpus of the retrieved articles and counting their occurrence over time. For the sake of illustration, only six technologies are shown.

Figure 7. **Identifying industry 4.0 technologies and related trends using scientific papers**



NB: ERP stands for enterprise resource planning; GPS stands for global positioning system.
Source: Cedefop.

Similar to the patent analysis case study, there are some technologies, such as 3d printing, cloud and GPS systems, for which trends are well defined. For instance, 3d printing, which is a vital component of additive manufacturing technologies, is also considered an essential ingredient in the industry 4.0 paradigm, due to growth in mass product customisation. In comparison, it is apparent that blockchain appears only after 2012. This could be interpreted as a signal that blockchain is an emerging technology.

Another case study illustrates how scientific papers can be used to extract skills related to robotic systems (Table 3). Applying NLP and text mining algorithms to the text in the scientific papers gives the extracted phrases containing information about robotic skills.

Table 3. **Identifying skills related to robotic systems**

Knowledge of big data analytics	Knowledge of data mining techniques	Knowledge of machine vision
Knowledge of cloud computing	Knowledge of deep learning algorithms	Knowledge of natural language processing tools
Application of clustering techniques	Knowledge of feature selection algorithms	Knowledge of neural networks
Application of computer vision	Knowledge of image processing and image segmentation	Ability to apply object detection algorithms
Knowledge of POS tagging (part of speech annotation)	Application of pose estimation algorithms	Ability to do speech recognition
Knowledge of support vector machines	Use of text mining	Knowledge of word embedding

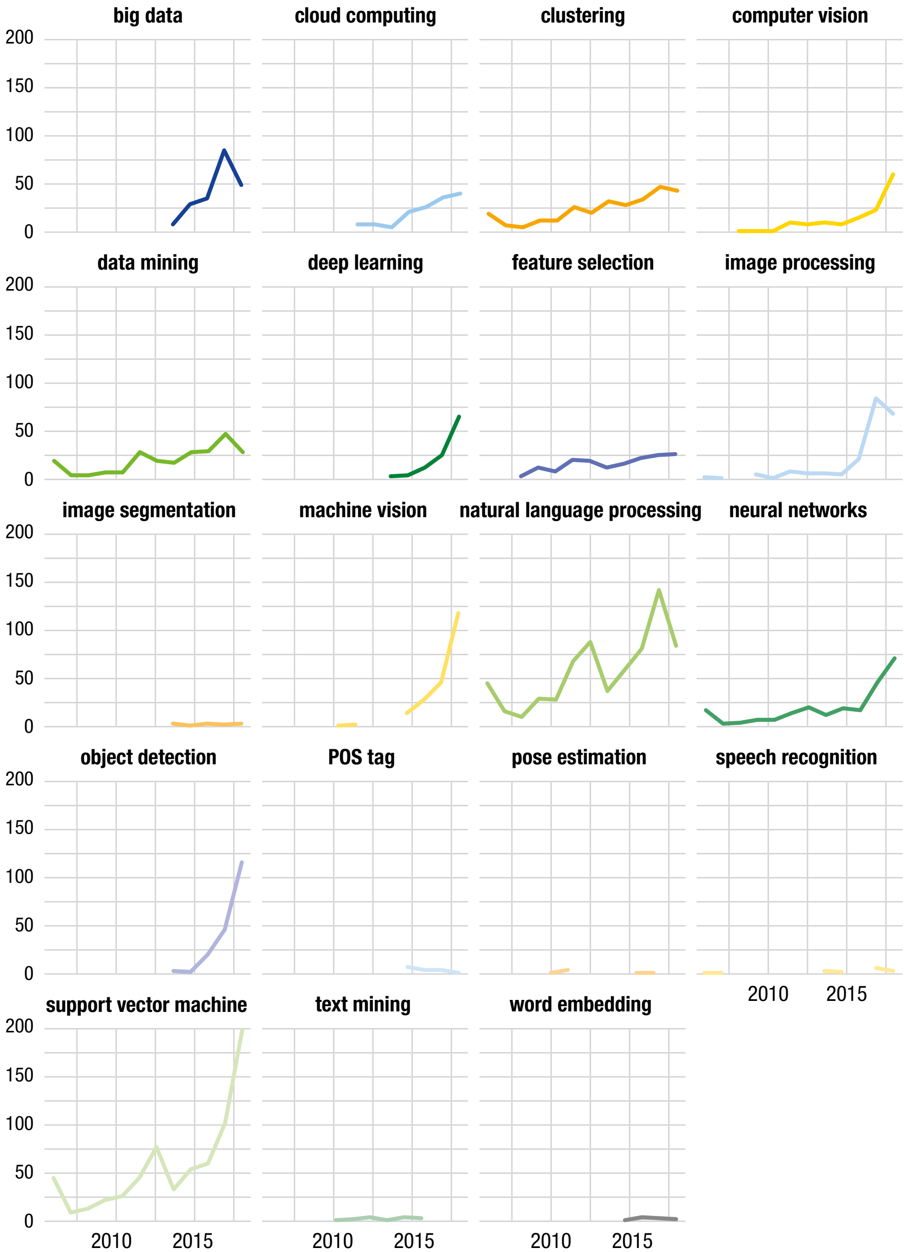
Source: Cedefop.

By calculating the yearly frequency of each extracted skill in the scientific paper database, the time trends can be visualised (Figure 8). Such trends can provide valuable and dynamic information on ongoing technological changes but may not be very useful from a future-oriented perspective. Knowing which skill is linked to specific technologies does not provide a basis for assessing how likely are future developments. Some insight on future change can be deduced from the direction (the slope) of the curves. In the case study, support vector machines, object detection, deep learning and natural language processing exhibit growing trends. It is fair to conclude that there is a high probability that these skills will keep growing in the years to come, which may translate into a greater demand for occupations in need of them.

3.4. Wikipedia analysis

Wikipedia is a free-of-charge, unrestricted encyclopaedia, made of interconnected knowledge whose corpus is frequently updated and modified by its users. All pages have internal links (hyperlinks) which allow the reader

Figure 8. **Skills trends associated with robotic systems**



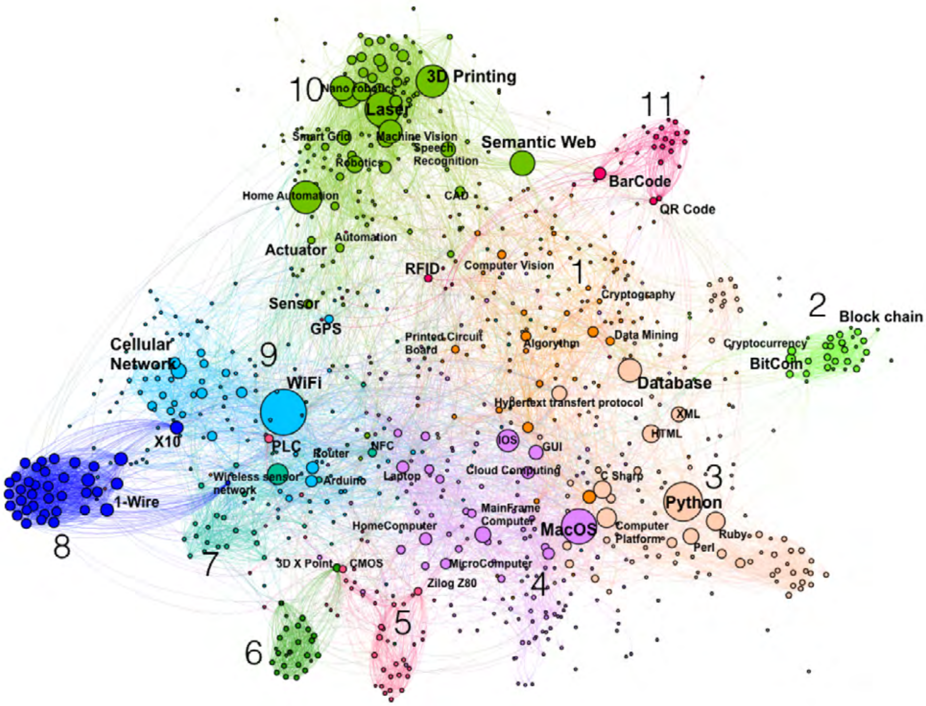
Source: Cedefop.

to move easily from one page to another. Apart from being useful for readers, these links also make it possible to use clustering techniques to group Wikipedia pages by topic.

Knowledge extraction techniques applied to Wikipedia can be used to map technologies and the skills associated with them. Via the hyperlinks contained in a Wikipedia page, related to a given technology or field of interest, it is possible to extract all related concepts.

Using Wikipedia as the source to analyse technological evolution related to industry 4.0, Figure 9 (Fantoni et al., 2018) shows how the technique works in practice. Starting from the page entitled industry 4.0, every hyperlink of the page is extracted using Wikipedia's API. For every sub-Wiki page hyperlinked to the industry-4.0-page, associated technologies are collated using text mining algorithms. The results of this exercise can be summarised in a network diagram (Figure 9).

Figure 9. **Network diagram of industry 4.0-related technologies identified using Wikipedia**



Source: Fantoni et al., 2018.

Network diagrams can provide valuable information complementing trend analyses. No coding skills are required; programmes such as Gephi or VOSviewer create them automatically. In network diagrams, every node represents a technology. In this case study, an arc between two nodes denotes hyperlinks connecting them. Colours are used to identify clusters of similar technologies. Such data visualisation is particularly insightful in research on (existing and growing) technologies, as it shows how technologies are interconnected and gives an impression of their relative importance. The closer two technologies are, the stronger the relationship between them; the bigger the node, the greater the importance of that technology.

3.5. Analysis of massive online open courses

The examples in Sections 3.1-3.4 look at technologies to make inferences about skill demands. It is also possible to look more directly at skills in relation to specific technologies by collating data from skill-specific or skill-intensive training sources. Massive online open courses (MOOCs), such as those provided by the platforms Coursera, Udemy, edX and Udacity, among others, are gaining increasing attention, since they give learners the opportunity to study a subject without taking a university course.

Descriptions of MOOCs contain information on:

(a) what it is about (title, description, etc.);
(b) skills gained or learning outcomes upon completion;
(c) duration (flexible schedule);
(d) language of instruction;
(e) course level;
(f) names of instructors.

The sections course description and skills gained or learning outcomes tend to contain skills or skills-related information. Analysing these sections aids understanding of the scope of the course and background knowledge and skills required. With MOOCs focusing on particular technologies (for example AI), which are common, it is possible to uncover a direct link between technologies and skills.

Table 4 provides an example of using MOOC descriptions and syllabi to extract skills related to cloud computing.

Table 4. **Extract of information from cloud computing online course descriptions using web scraping**

Title of the course	Description of the course
Learn cloud computing from scratch	The course will start with basic introduction to cloud concepts like SAAS, PAAS and IAAS. You will also learn how Linux systems is changing the infrastructure landscape worldwide. You will then learn to use popular cloud technologies, like Google compute engine, Amazon AWS and Redhat open shift.
Cloud computing basics: enhance your career as cloud engineer	In this course I will take you through all the basics and jargon used in the cloud computing industry and these will be explained in layman's term. So you don't need any prior knowledge on cloud computing to enrol for this course. After completing this course you will be able to comprehend cloud computing related discussion happening around and all set to start a career or manage a team in this field.
Introduction to cloud computing	In this hands-on VTC course, you will access a variety of cloud services and work with different cloud providers, such as Apple, Microsoft, Google and Amazon. You will set up virtual servers, work with cloud file storage, learn about a variety of cloud collaboration options and much more. This practical course will help you make the transition to working in the cloud from any device, anywhere, anytime! To begin learning today, simply click on the movie link.

Source: Cedefop.

Once data collection is completed, NLP is applied to transform syllabus text (course descriptions) into structured data that can be processed. Every course description is first split into single phrases, i.e. tokens. Subsequently, a simple text mining algorithm which searches for clues, such as 'use', 'learn', 'understand', 'apply' (words that identify a skill) can be used to identify and extract skills mentioned (Table 5).

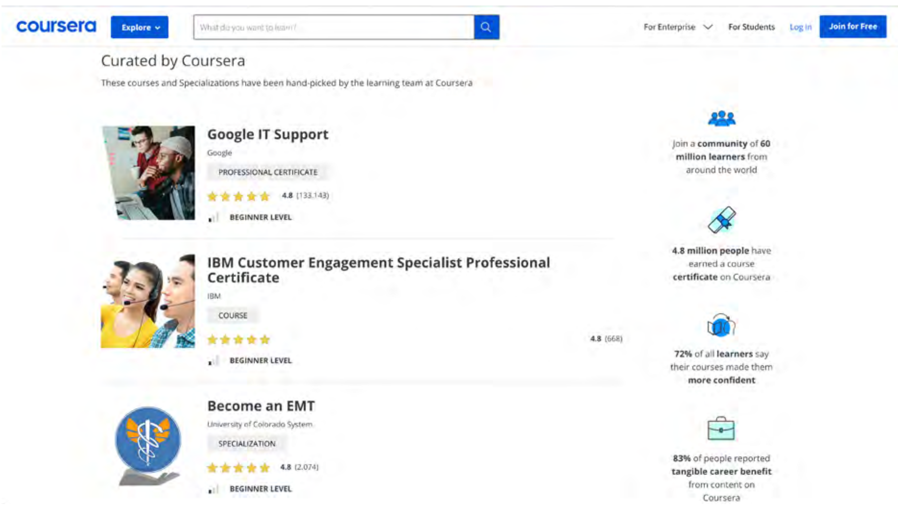
Extracting other labour-market-relevant information, such as the job profiles found in online courses, is another analysis opportunity. Some online course providers stratify their courses by skills, job profiles, difficulty levels and other variables (Figure 10). On top of extracting information on skills (for example knowledge of Python programming, knowledge of web scraping), it is also possible to determine which occupation(s) online courses target. This can be done by combining information on job titles and skills information (Table 6).

Table 5. **Cloud-computing-related skills extracted from MOOCs**

Technology	Extracted skills
Cloud computing	Use cloud platforms
	Understand virtualisation and its use in infrastructure development
	Understand cloud computing concepts and technologies
	Apply the learning to build cloud infrastructure
	Set up virtual servers
	Work with cloud file storage
	Use cloud collaboration options
	Use Dropbox
	Use cloud print
	Use private cloud model
	Use public cloud model
	Understand about cloud computing architecture
	Planning cloud computing

Source: Cedefop.

Figure 10. **MOOC website section containing information on skills and job titles**



Source: Coursera.com.

Table 6. **Skills and associated job profiles based on analysis of a MOOC website**

Skill	Associated job profile
Python programming	Software engineer
	PHP developer
	Data scientist
Web scraping	Java engineer
	Applications developer
Marketing	Brand manager
	Web manager
Team management	Development coordinator
Team building	Sales operations analyst
Communication	Account manager
	Agile coach

Source: Cedefop.

It is also possible to use MOOC descriptions to assess the portability of skills, which is particularly relevant in the context of up- and reskilling policies. Network diagrams, showing occupations and their commonalities in terms of required skills, can be used to visualise skills portability. As an example, three job profiles that have certain skills in common, as extracted from a MOOC website, are shown in Table 7.

Gephi (a software for visualising network data), can be used to create a network diagram (Figure 11). In the example shown, transversal skills can be identified by looking at the nodes which are shared by more than one occupational profile. For instance, data integration, database design and data modelling are skills which the occupations ETL developer, data quality analyst and finance data analyst have in common (hence 'portable' skills). Knowledge of information privacy and survey design skills is only relevant for the occupation data quality analyst, implying these skills have low portability.

Table 7. **Three job profiles and their skills requirements based on analysis of a MOOC website**

Job profile	Skills in…
ETL developer	Data modelling
	Big data
	Unix
	Bash
	Data visualisation
Data quality analyst	Data integrity
	Survey design
	Version control
	Simulation
Finance data analyst	Pie chart
	Covariance
	Time series
	Decision theory
	Forecasting

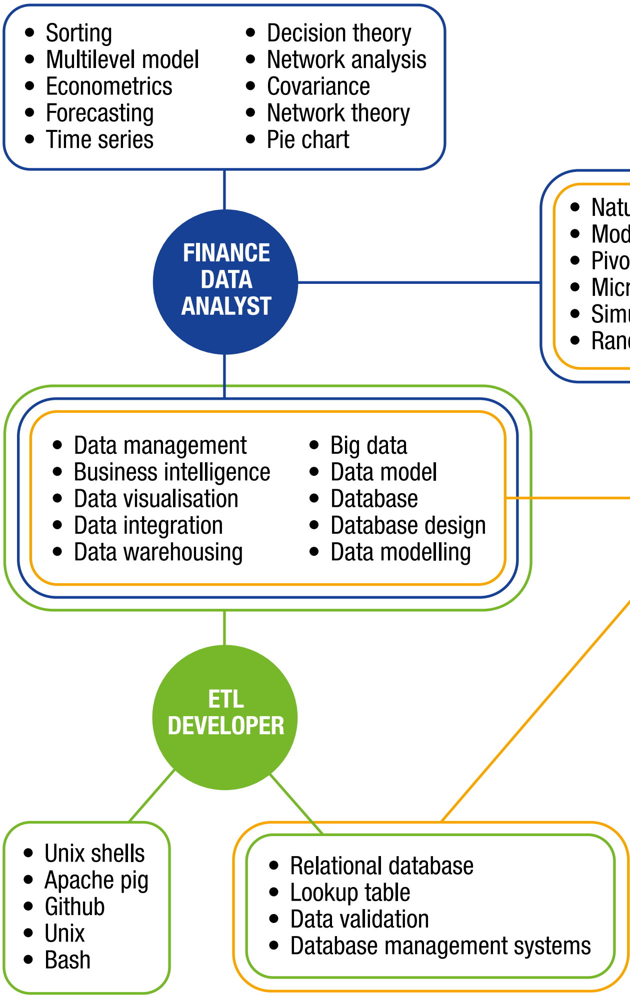
NB: ETL stands for 'extract, transform, load'.
Source: Cedefop.

3.6. Analysis using occupational databases

In addition to identifying the skill profile of jobs, it is also possible to look at jobs/occupations where a particular skill is evident. ESCO and the occupational information network developed under the sponsorship of the US Department of Labour (O*NET) are online databases with descriptions and classifications of occupations/jobs. In ESCO and O*NET, every occupation has a general description and a list of skills or tasks routinely undertaken. This information can be very useful, as it allows a technology identified using the approaches described above to be matched to job profiles.

The example below extracts ESCO occupations related to machine learning skills and NLP skills. It is possible to identify the job profiles in ESCO that mention these two skill terms most frequently. A potential problem is that the skills terms might not be explicitly mentioned in the database, making it necessary first to identify the equivalent skill terms in ESCO (Table 8). Subsequently, ESCO skills can be matched to ESCO occupations (Table 9).

Figure 11. **Analysis of shared skills between ETL developer, data quality analyst and finance data analyst**



Source: Cedefop.

Table 8 **Skills matched to the corresponding ESCO skills**

Skill	ESCO skill
natural language processing	natural language processing
machine learning	utilise machine learning

Source: Cedefop.

Table 9. **Matching skills to ESCO occupations**

Skill	ESCO skill	ESCO occupation	Role in occupation
natural language processing	natural language processing	language engineer	essential
		ICT intelligent systems designer	
		knowledge engineer	
		user interface designer	optional
		application engineer	
machine learning	utilise machine learning	application engineer	
		software developer	

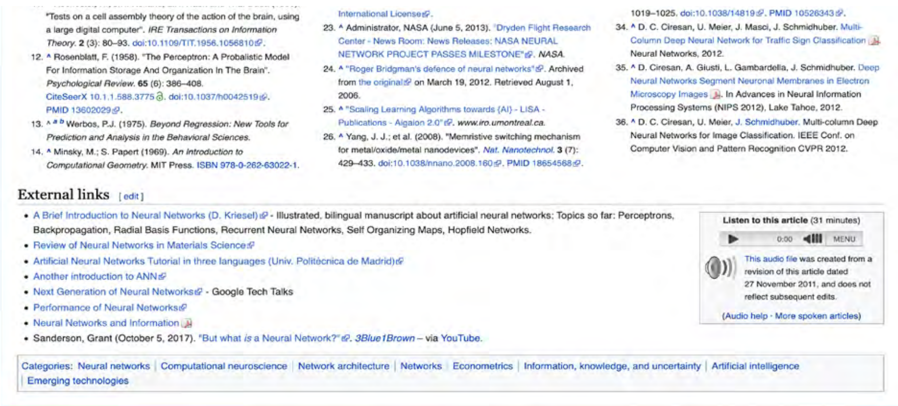
Source: Cedefop.

Unfortunately, ESCO and O*NET, do not always contain information at sufficient level of detail. For instance, ESCO does not include skills, such as knowledge of deep learning or knowledge of artificial neural networks. A possible solution is to identify for non-matched skills an overarching category to which they belong. These macro-categories can be found, for example, in the category section in Wikipedia pages (Figure 12).

3.7. Automated methods trained on expert input

Although not strictly considered an automated knowledge extraction technique, extrapolating insights on automatable or future jobs and skills by training ML algorithms based on expert opinions has received much attention in literature. Examples are the studies by Frey and Osborne (2017) and Bakhshi and colleagues (2017). Identifying which jobs and skills are resilient or in danger of becoming obsolete by training ML models using expert consultations, the findings pointed towards automation leading to

Figure 12. **Wikipedia categories related to neural networks**



"Tests on a cell assembly theory of the action of the brain, using a large digital computer". *IRE Transactions on Information Theory.* **2** (3): 80–93. doi:10.1109/TIT.1956.1056810.

12. ^ Rosenblatt, F. (1958). "The Perceptron: A Probalistic Model For Information Storage And Organization In The Brain". *Psychological Review.* **65** (6): 386–408. CiteSeerX 10.1.1.588.3775. doi:10.1037/h0042519. PMID 13602029.
13. ^ a b Werbos, P.J. (1975). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.*
14. ^ Minsky, M.; S. Papert (1969). *An Introduction to Computational Geometry.* MIT Press. ISBN 978-0-262-63022-1.

International License.

23. ^ Administrator, NASA (June 5, 2013). "Dryden Flight Research Center - News Room: News Releases: NASA NEURAL NETWORK PROJECT PASSES MILESTONE". *NASA.*
24. ^ "Roger Bridgman's defence of neural networks". Archived from the original on March 19, 2012. Retrieved August 1, 2006.
25. ^ "Scaling Learning Algorithms towards {AI} - LISA - Publications - Aigaion 2.0". *www.iro.umontreal.ca.*
26. ^ Yang, J. J.; et al. (2008). "Memristive switching mechanism for metal/oxide/metal nanodevices". *Nat. Nanotechnol.* **3** (7): 429–433. doi:10.1038/nnano.2008.160. PMID 18654568.

1019–1025. doi:10.1038/14819. PMID 10526343.

34. ^ D. C. Ciresan, U. Meier, J. Masci, J. Schmidhuber. Multi-Column Deep Neural Network for Traffic Sign Classification. Neural Networks, 2012.
35. ^ D. Ciresan, A. Giusti, L. Gambardella, J. Schmidhuber. Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. In Advances in Neural Information Processing Systems (NIPS 2012), Lake Tahoe, 2012.
36. ^ D. C. Ciresan, U. Meier, J. Schmidhuber. Multi-column Deep Neural Networks for Image Classification. IEEE Conf. on Computer Vision and Pattern Recognition CVPR 2012.

External links [edit]

- A Brief Introduction to Neural Networks (D. Kriesel) - Illustrated, bilingual manuscript about artificial neural networks; Topics so far: Perceptrons, Backpropagation, Radial Basis Functions, Recurrent Neural Networks, Self Organizing Maps, Hopfield Networks.
- Review of Neural Networks in Materials Science
- Artificial Neural Networks Tutorial in three languages (Univ. Politécnica de Madrid)
- Another introduction to ANN
- Next Generation of Neural Networks - Google Tech Talks
- Performance of Neural Networks
- Neural Networks and Information
- Sanderson, Grant (October 5, 2017). "But what *is* a Neural Network?". *3Blue1Brown* – via YouTube.

Listen to this article (31 minutes)
0:00 MENU
This audio file was created from a revision of this article dated 27 November 2011, and does not reflect subsequent edits.
(Audio help · More spoken articles)

Categories: Neural networks | Computational neuroscience | Network architecture | Networks | Econometrics | Information, knowledge, and uncertainty | Artificial intelligence | Emerging technologies

Source: Wikipedia.

massive job destruction. In Frey and Osborne's study, the main research aim was to estimate the probability of automation replacing people employed in 702 occupations. The study of Bakhshi and colleagues (2017) analysed future changes in job profiles by looking at multiple drivers of change – not just automation – and inferring emerging skill needs.

Both studies collected inputs from expert panel workshops. Relying on the expertise of ML experts, Frey and Osborne manually labelled 70 occupations as follows:

(a) label 1 was given to occupations which were subjectively considered fully automatable: where relevant tasks contained in the US O*NET database (for example finger or manual dexterity, creative or social intelligence) could be performed by state-of-the-art computer-controlled equipment;

(b) label 0 was given to occupations with tasks that could not be feasibly automated by ML algorithms.

An ML classifier was constructed, which could learn from the manually labelled occupations and predict the probability of automation for the other (non-labelled) occupations. More precisely, after labelling the set of 70 occupations with ones and zeros, it was divided into a training set and a test set. The first was used to train the classifier which was then used to predict the probability of automation of occupations belonging to the test set.

After the classifier was evaluated and validated, it was used to predict the probability of computerisation for a total number of 702 occupations.

Although their work reviewed several drivers of change expected to shape industry and the labour market and their interaction, Bakshi and colleagues (2017) used a similar approach. Seven relatively stable and clearly directed change drivers were selected:
(a) environmental sustainability;
(b) urbanisation;
(c) increasing inequality;
(d) political uncertainty;
(e) technological change;
(f) globalisation;
(g) demographic change.

During workshops, the experts were encouraged to debate future scenarios for a set of 30 occupations. Similar to the Frey and Osborne analysis, labels were assigned to occupations according to their future employment prospects (grow, stay the same, or shrink). The labelled data was subsequently used to train an ML classifier, which was deployed to generate employment predictions for all occupations.

While the above studies have received widespread attention in the academic community and the popular media, it is important to be aware that using experts as a primary source in skill anticipation carries a risk of arriving at distorted or exaggerated results, as their opinions can be subject to significant bias.

CHAPTER 4.

Choosing a method

This second Cedefop guide on methods for identifying technological change and its impact on skill requirements has looked at the potential of big data or AI-driven approaches for analysing technological trends and skills anticipation. Using text mining, NLP and ML techniques to analyse information in open online databases has made it possible to collate granular data on skills and technologies that would have been unimaginable in the recent past. Exploiting new analysis possibilities using data sourced from online job portals, patent repositories, scientific databases and online course providers has made it possible to develop insight into emerging technologies and skill needs that can typically not be achieved with conventional LMSI methods (for example skill surveys and skills forecasting).

Automated knowledge extraction analysis can be very effective in providing valuable information on continuing and newly emerging technological change and future skill demand. Almost any online document or text is a potential data source for analysis. There are several advantages to using automated analysis to develop LMSI:

(a) increased opportunities to provide up-to-date information on emerging skill needs;
(b) data are future-looking in the sense that they can identify emerging technologies which may well take off and become widely used in workplaces;
(c) skill needs linked to particular technologies can be identified;
(d) results can be delivered at a highly disaggregated level to provide the level of granularity that policy-makers typically require.

It is equally important to be aware of the limitations of using automated big data/AI-powered analysis. These include:

(a) data collected will be in large part driven by programming, which specifies which terms to search for;
(b) uncertainty with respect to whether an identified technology or skill is important, in terms of how likely it is to shape the future of work and skills;

(c) some identified technologies and skills may be of transitional importance (disappear quite quickly);
(d) difficulties evaluating the likely scale or pervasiveness of change, for example how many people will require a particular skill and whether supply of that skill is sufficient;
(e) the dependence on classifications that are quite dated (for example ISCO) to collate evidence on skills and jobs, which undermines the future-looking advantage automated techniques have over conventional approaches;
(f) the black-box nature of findings based on machine or deep learning models. High uncertainty and possible distortion in the datasets used for training purposes can have significant impacts. Such uncertainty is amplified when the training dataset is unstructured or of low quality, has limited number of observations or when feature variables recurrently change.

It is sometimes more challenging to utilise big data methods for analysis, relative to conventional techniques of skills assessment and anticipation. The reasons are that the underlying data extracted from web sources are unstructured and not generated for research use. As a result, any repurposing, or data classification and analysis carries uncertainties and limitations.

Another major problem is representativeness. For example, online job portals do not cover most vacancies which tend to be filled via word of mouth. Representativeness varies by occupation and coverage of different labour market segments tends to be linked to data source type: for example, high-skilled jobs on private web portals; blue-collar jobs on public employment service portals. As flow data, it is not clear whether vacancy postings are representative of current employment with respect to skill requirements. Jobs with above-average turnover will be overrepresented relative to their employment share. Single ads can represent multiple vacancies, or even no vacancy at all, given the low cost of posting online job ads and the phenomenon of some employers posting jobs online to see which potential candidates are available on the labour market (including so-called 'ghost' vacancies) (Cedefop, 2019a).

There are also challenges with the skills information collected via big data sources itself. A survey will use a common set of questions to all respondents in the universe and score responses on a common scale. Online job ads tend to focus on occupation-specific skills rather than transversal skill concepts. These skills can be quite specific and difficult to aggregate into broader skills concepts because they are qualitatively diverse and can usually not easily be mapped into a level of complexity framework, as is typically done in surveys employing a job-skill requirements approach (Cedefop, 2021a).

What is most commonly possible is coding the presence or absence of a specific skill requirement (for example commercial truck driver's licence, strong problem-solving skills, biochemistry, work with robots), or counting the number of skills of a given class (for example ICT-related skills) as they appear in job ads (for example, the number of computer programmes required). However, online job advertisements typically do not specify all important skills and technologies, because many are not mentioned; they are widespread and implicitly expected. Most online job advertisement databases also have little or no data on the characteristics of workers actually hired to fill jobs, which may differ from employers' stated preferences in job ads, for instance in terms of education credentials, experience, and specific skills.

Moreover, the algorithms for scraping and processing online postings, as well as the original websites that are sourced, tend to evolve, so trend studies will need to apply safeguards (for example stable sources) so that real change can be distinguished from statistical artefacts. By contrast, surveys can be repeated following standard procedures.

When considering using big-data-powered technology and skills analysis, the following checklist with key issues that will need to be considered can be used.

Big data are a rich source of information and automated analysis methods will become increasingly important in coming years. But such techniques should not replace conventional skills forecasts and surveys (see first guide: Cedefop, 2021a) or skills foresight methods designed to address particular policy-relevant questions with a longer-term horizon (see third guide: Cedefop, 2021b). In many respects the challenge is to make effective use of the wide variety of data and information available.

Participatory and quantitative, non-participatory methods are not mutually exclusive. Ideally, they should support one another so that they can potentially form an iterative process whereby the participatory process of stakeholder engagement can shape data collection and analyses in the non-participatory ones (and vice versa). Such interaction makes it possible to develop views on how the future will unfold and how informed skills policies and actions need to develop.

Table 10. **Issues to consider before undertaking big data analysis**

Issue	Detail to be considered
Initial check	• Are data already available from existing sources? • What is the big-data analysis designed to add value to?
Types of question that can be addressed regarding technologies	• Which particular technologies are coming on stream? • What are the technologies that comprise concepts, such as the internet of things, industry 4.0, etc.?
Types of question that can be addressed about skills	• What are the particular skills associated with a specific technology? • How are skills clustered together with respect to a particular technology? • What are the new skills emerging within existing jobs/occupations?
Requirements	• Identification of data in a form that contains comprehensive information about a particular technology • Identification of data in a form that contains comprehensive information about the skills that may be associated with a particular technology • Capacity to develop the text mining algorithms required to collect the data needed
Analysis steps	• Specification of technologies of interest • Data collection (text mining) • Data transformation (tokenisation; stemming, POS tagging) • Data elaboration (machine learning, n-grams, etc.) • Classifying data on skills using available classifications (e.g. ISCO and ESCO) • Enhancing/adding to existing skills classifications • Identifying new/emerging occupations/jobs not yet classified
Selected further information	• Skills-OVATE: skills online vacancy analysis tool for Europe

Source: Cedefop.

Table 11 provides a summary to guide policy-makers and analysts in understanding when to use the approaches covered by the short Cedefop guides. To learn more about conventional labour market and skills intelligence approaches (skills surveys, skills forecasting) as well as technology foresight approaches, readers are referred to the other two Cedefop 'how-to' guides.

Table 11. **A menu of skills assessment and anticipation choices**

Type of approach	When to use	Capacity to predict the future	Timeliness
Quantitative, non-participatory approaches			
Surveys and other primary data collections	When there is a relatively well-developed understanding of the technologies and associated skills of interest. Surveys will tend to provide information on the extent of use of skills and technologies, extent to which skills are available, efforts taken to fulfil skill needs etc.	Tend to be good at collecting information about recent past and impending changes. Not well suited to anticipating future technological changes and future skill needs.	Can be time-consuming to undertake – design of questionnaires, conducting fieldwork, cleaning data, producing findings.
Skills forecasting	Where time series data are available on skill needs (based on qualification and occupation), and where there is an underlying macroeconomic model that can provide robust estimates of future employment demand by sector, skills forecasts can provide a robust means of providing quantitative projections of future skill demand (circa 10 years ahead).	Skills forecasting models tend to provide a projection of future demand, based on an extrapolation of past trends and/or current policy. The assumption is that the future is based on a continuation of things as they are currently. Scenarios provide some basis for varying this to some extent, to account for continual technological change.	If the model already exists, analysis can be undertaken over a relatively short space of time. But setting up the initial model and ensuring regular updates of the results can be time-consuming and resource-intensive.
Big data analysis	Particularly useful where views about the future may not be well developed: where there is uncertainty about either the types of technology that are likely to become dominant or commonplace, and/or the skills associated with those technologies. Can also provide the detailed level of analysis that forecasting and surveys struggle to provide.	Can provide relatively real-time information on technological change and skill needs. By identifying those technologies that are at the point of take-off, there is scope to gauge likely future skill needs. There are uncertainties about how representative data are of a given population and about how much 'noise' can be removed from any analysis or their inability to provide standardised information on skills complexity.	Can be time-consuming to develop initial search algorithms, but, once established, can be undertaken in a relatively fast manner. It needs to be borne in mind that coding/classifying of technology and skills data can be time-consuming. Maintenance and operational costs are also non-trivial.

Type of approach	When to use	Capacity to predict the future	Timeliness
Participatory approaches			
Technology foresight	Where there is a large amount of information that needs synthesising to develop actions to ensure that skills needs associated with particular technologies can be met. Where there is limited data and information and where expert groups can address the lack of information.	Can provide a view of the future and, importantly, an indication of how the future might be shaped for the benefit of society as a whole. Is dependent upon the availability of expert groups who can provide key input and a process in place to develop a degree of consensus about the future direction of change.	Depends upon the scale of the exercise. Full-scale foresight involving a large number of participants is likely to prove time-consuming. But it is possible, and at times advisable, to conduct foresight with smaller groups over a relatively short-time span.

Source: Cedefop.

Acronyms

AI	artificial intelligence
API	application programming interface
Cedefop	European Centre for the Development of Vocational Training
DPS	data production system
EPO	European Patent Office
ESCO	European skills, competences and occupations
ESJS	European skills and jobs survey
EU	European Union
FGB	Fondazione Giacomo Brodolini
GPS	global positioning system
IAG-TVET	inter-agency group on technical and vocational education
ICT	information and communications technology
IoT	internet of things
ISCO	international standard of classification of occupations
LMSI	labour market and skills intelligence
ML	machine learning
MOOC	massive online open course
NER	named entity recognition
NLP	natural language processing
NLTK	natural language toolkit
OJA	online job advertisements
PES	public employment service
POS	part of speech
Skills-OVATE	skills online vacancy analysis tool for Europe
SPSS	statistical package for social sciences
VET	vocational education and training

References

[URLs accessed 14.2.2021]

Acemoglu, D.; Restrepo, P. (2019). Automation and new tasks: how technology displaces and reinstates labor. *Journal of Economic Perspectives*, Vol. 33, No 2, pp. 3-30.

Adner, R.; Kapoor, R. (2016). Right tech, wrong time. *Harvard Business Review*, Vol. 94, No 11, pp. 60-67.

Akerlof, G.A. (1970). The market for 'lemons': quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, Vol. 84, No 3, pp. 488-500.

Arntz, M.; Gregory, T.; Zierahn, U. (2017). Revisiting the risk of automation. *Economics Letters*, Vol. 159, pp. 157-160.

Bai, B. et al. (2009). Polynomial semantic indexing. Advances in neural information processing systems, pp. 64-72. 23rd Annual conference on neural information processing systems 2009. *Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.* https://www.researchgate.net/publication/221619566_Polynomial_Semantic_Indexing

Bakhshi, H. et al. (2017). *The future of skills: employment in 2030*. London: Pearson and Nesta.

Bird, S.; Klein, E.; Loper, E. (2009). *Natural language processing with Python: analysing text with natural language toolkit*. Sebastopol: O'Reilly Media.

Brynjolfsson, E.; McAfee, A. (2014). *The second machine age: work, progress, and prosperity in a time of brilliant technologies*. New York: WW Norton & Co Inc.

Cagnin, C.; Havas, A.; Saritas, O. (2013). Future-oriented technology analysis: its potential to address disruptive transformations. *Technological Forecasting and Social Change*, Vol. 80, pp. 379-385.

Cedefop (2013). *Piloting a European employer survey on skill needs: illustrative findings*. Luxembourg: Publications Office. Research paper; No 36. https://www.cedefop.europa.eu/files/5536_en.pdf

Cedefop (2015). *Skills, qualifications and jobs in the EU: the making of a perfect match? Evidence from Cedefop's European skills and jobs survey*. Luxembourg: Publications Office. Cedefop reference series; No 103. https://www.cedefop.europa.eu/files/3072_en.pdf

Cedefop (2016). *The great divide: digitalisation and digital skill gaps in the EU workforce*. Thessaloniki: Greece. #ESJsurvey Insights; No 9. https://www.cedefop.europa.eu/files/esj_insight_9_digital_skills_final.pdf

Cedefop (2019a). *Online job vacancies and skills analysis: a Cedefop pan-European approach*. Luxembourg: Publications Office. https://www.cedefop.europa.eu/files/4172_en.pdf

Cedefop (2019b). *The online job vacancy market in the EU: driving forces and emerging trends*. Luxembourg: Publications Office. Cedefop research paper; No 72. https://www.cedefop.europa.eu/files/5572_en.pdf

Cedefop (2020). *Developing and matching skills in the online platform economy: findings on new forms of digital work and learning from Cedefop's CrowdLearn study*. Luxembourg: Publications Office. Cedefop reference series; No 116. http://data.europa.eu/doi/10.2801/588297

Cedefop (2021a). *Understanding technological change and skill needs: skills surveys and skills forecasting. Cedefop practical guide 1.* Luxembourg: Publications Office. http://data.europa.eu/doi/10.2801/212891

Cedefop (2021b). *Understanding technological change and skill needs: technology and skills foresight. Cedefop practical guide 3.* Luxembourg: Publications Office. http://data.europa.eu/doi/10.2801/307925

Copestake, A. (2004). *Natural language processing*. Lecture synopsis. https://www.cl.cam.ac.uk/teaching/2004/NatLangProc/lectures.pdf

Daim, T.U. et al. (2006). Forecasting emerging technologies: use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, Vol. 73, No 8, pp. 981-1012.

Deming, D.J.; Noray, K.L. (2020). Earnings dynamics, changing job skills and STEM careers. *Quarterly Journal of Economics,* Vol. 135, No 4, pp. 1965-2005. https://academic.oup.com/qje/article/135/4/1965/5858010

Domingos, P. (2015). *The master algorithm: how the quest for the ultimate learning machine will remake our world*. New York: Basic books.

Fantoni, G. et al. (2018). Defining industry 4.0 professional archetypes: a data-driven approach. In: Hogarth, T. (ed.). *Economy, employment and skills: European, regional and global perspectives in an age of uncertainty*. Quaderni series. Rome: Fondazioni Giacomo Brodolini, pp. 75-95.

Freeman, R.B.; Ganguli, I.; Handel, M.J. (2020). Within-occupation changes dominate changes in what workers do: a shift-share decomposition, 2005-15. *AEA papers and proceedings 2020*, Vol. 110, pp. 394-399.

Frey, C.; Osborne, M. (2013). *The future of employment: how susceptible are jobs to computerisation*. Oxford: University of Oxford. https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf?link=mktw

Frey, C.; Osborne, M. (2017). The future of employment: how susceptible are jobs to computerisation. *Technological Forecasting and Social Change*, Vol. 114, pp. 254-280.

Gambin, L. et al. (2016). *Research to understand the extent, nature and impact of skills mismatches in the economy*. London: Department for Business Innovation and Skills. BIS Research Paper; No 265.

Golzio, D. (2012). WWWWWHOW Read a patent! *EPO presentation, Pisa, 12 April 2012*. https://www.unipi.it/ricerca/phdarchivi/edizioni/phd2012/Dispense-d/golzio/how_read.pdf

Goodfellow, I.J.; Bengio, Y.; Courville, A. (2016). *Deep learning*. Boston: MIT Press.

Gürtzgen, N. et al. (2021). *Does online search improve the match quality of new hires?*, IZA discussion paper; No 14031.

Hackert, T. (2016). *Worldwide patent No WO2016154284A1*. www.google.it/patents/WO2016154284A1?cl=en

Huang, W.Y.; Lippmann, R.P. (1988). Neural net and traditional classifiers. In: Anderson, D.Z. (ed.). *Neural information processing systems*. New York: American Institute of Physics, pp. 387-396. https://cutt.ly/qkVroHv

Hwang, G. et al. (2015) Use of patent analysis for the future skills-needs in information security. *Asian Journal of Innovation and Policy*, Vol. 4, No 3, pp. 307-327.

James, G. et al. (2013). *An introduction to statistical learning*. New York: Springer.

Kütt, W.; Schmiemann, M. (1998). Quick scan: a novelty search service in the framework of Euro-R&D programmes. *World Patent Information*, Vol. 2, No 20, pp. 146-147.

McGuinness, S.; Pouliakas, K.; Redmond, P. (2019). *Skills-displacing technological change and its impact on jobs: challenging technological alarmism*. IZA discussion paper; No 12541.

Michie, D. (1968). Memo functions and machine learning. *Nature*, Vol. 218, No 5136, p. 19.

Nedelkoska, L.; Quintini, G. (2018). *Automation, skills use and training*. Paris: OECD Publishing. OECD social, employment and migration working papers; No 202.

Pouliakas, K. (2018). The risk of automation in EU labour markets: a skils-requirement approach. In: Hogarth, T. (ed.). *Economy, employment and* *skills: European, regional and global perspectives in an age of uncertainty*. Rome: Fondazioni Giacomo Brodolini. https://www.fondazionebrodolini.it/sites/default/files/pubblicazioni/file/q61_x_web_cor_0.pdf

Rosell, M.; Csc, K. (2008). *Introduction to text clustering*. https://www.semanticscholar.org/paper/Introduction-to-Text-Clustering-Rosell-Csc/1dbc0966e98d2a21e892850a9c381fb69ac2522d

Terragno, P. (1979). Patent as technical literature. *IEEE Transaction on Professional Communication*, Vol. PC-22, No 2, pp. 101-104.

Weblinks
[URLs accessed 15.2.2021]

https://www.coursera.org/
https://cloud.google.com/natural-language/
https://www.edx.org/
https://worldwide.espacenet.com/
https://ec.europa.eu/esco/portal/occupation
http://www.freepatentsonline.com/
https://gephi.org/
https://www.google.com/?tbm=pts
https://scholar.google.com/
https://www.ibm.com/analytics/spss-statistics-software
https://medium.com/the-andela-way/learn-how-to-scrape-the-web-2a7cc488e017
http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/decisionTree.html
https://www.nltk.org/
https://www.onetonline.org/
https://rapidminer.com/
http://r-statistics.co/R-Tutorial.html
https://rstudio.com/
https://www.sas.com/it_it/insights/analytics/what-is-natural-language-processing-nlp.html
https://www.scopus.com/
https://towardsdatascience.com/classification-using-neural-networks-b8e98f3a904f
https://www.udacity.com/
https://www.udemy.com/
https://www.vosviewer.com/
https://www.webofknowledge.com
https://www.wikipedia.org/
https://en.wikipedia.org/wiki/Data_science
https://www.youtube.com/watch?v=Rm1v7Il2iMk

EN

Understanding technological change and skill needs

Big data and artificial intelligence methods

The world of work is being impacted by a fourth industrial revolution, transformed by artificial intelligence and other emerging technologies. With forecasts suggesting large shares of workers, displaced by automation, in need of upskilling/reskilling, the design of active skills policies is necessary.

Conventional methods used to anticipate technological change and changing skill needs, such as skill surveys and forecasting, have limited scope to provide insights into emerging trends. With the increasing use of big data and AI methods, analysts have new 'real-time' tools at their disposal. Skill foresight techniques are also increasingly used to gauge in-depth stakeholder information about future technologies and skill needs.

A series of Cedefop guides aims to inform analysts and policy-makers about available skills anticipation methods used to navigate through the uncertainty of changing technologies and skill demands. This second practical guide focuses on automated skills intelligence methods: big data and AI-driven analyses.

CEDEFOP European Centre for the Development of Vocational Training

Europe 123, Thessaloniki (Pylea), GREECE
Postal address: Cedefop service post, 570 01 Thermi, GREECE
Tel. +30 2310490111, Fax +30 2310490020, Email: info@cedefop.europa.eu

visit our portal www.cedefop.europa.eu

Publications Office
of the European Union

4198 EN – TI-05-21-028-EN-N – doi:10.2801/144881

ISBN 978-92-896-3238-6
9 789289 632386