NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

*Ilya Kuzminov, Pavel Bakhtin,
Elena Khabirova, Maxim Kotsemir,
Alina Lavrinenko*

MAPPING THE RADICAL INNOVATIONS IN FOOD INDUSTRY: A TEXT MINING STUDY

BASIC RESEARCH PROGRAM
WORKING PAPERS

SERIES: SCIENCE, TECHNOLOGY AND INNOVATION

WP BRP 80/STI/2018

This Working Paper is an output of a research project implemented at the National Research University Higher School of Economics (HSE). Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE

Ilya Kuzminov[1], Pavel Bakhtin[2], Elena Khabirova[3],
Maxim Kotsemir[4], Alina Lavrinenko[5]

MAPPING THE RADICAL INNOVATIONS IN FOOD INDUSTRY: A TEXT MINING STUDY [6]

The article presents the results of the study of radical innovations in the global food industry which were obtained through semantic analysis of heterogeneous unstructured text data sources by applying innovative big data text mining system. The approach used allows performing rapid, yet comprehensive aggregation of the whole polyphony of existing knowledge of the technology development in any sector for traditional foresight, future oriented technology analysis, and horizon scanning studies. The sources for the analysis include research papers, patent applications with both full-text data and additional structured metadata, analytical reports by main international organizations and national key players, various media and news resources, including all the major technology innovation, disruption and venture capital news websites. Their processing with an introduced approach for trend- and technology-mapping helps to identify ongoing and emerging technology-related trends, weak signals on possible scientific breakthroughs in the global food industry, including most promising startup strategies and food innovation controversies. This kind of analysis can be performed on a regular basis owing to constant accumulation of textual data and serve as a framework for constant science and technology (S&T) monitoring for early warning on changing technology landscape and its implications on agriculture and food markets.

Keywords: radical innovations, trends, weak signals, big data, text mining, food industry.

JEL: C55, O1, O3

[1] Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow, Russian Federation. E-mail: ikuzminov@hse.ru

[2] Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow, Russian Federation. E-mail: pbakhtin@hse.ru

[3] Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow, Russian Federation. E-mail: etochilina@hse.ru

[4] Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow, Russian Federation. E-mail: mkotsemir@hse.ru

[5] Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow, Russian Federation. E-mail: alavrinenko@hse.ru

[6] The Working Paper was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of the subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

Introduction

The food industry is one of the main sectors in the modern economy, and a crucial one, as food and biological security issues are one of the central points in the policy agenda for virtually any government (Maggio et al. 2016). In addition, there are severe long-term risks for the stability of global food system against the backdrop of yet unsolved hunger (von Grebmer et al. 2017), malnutrition (Sahn, 2015) and agricultural sustainability problems across the globe (Pretty et al. 2010; Rockström et al. 2017). Only new technology revolutions in the sector can help solve these dire issues (King, 2017). Moreover, these disruptive changes are looming large, as more and more venture capital in developed countries is attracted to the food industry startups aiming to bring about the radical innovation, such as synthetic foods (World Preservation Foundation, 2017), synbio replacements of traditional processes and items (Hayden, 2015), and other disruptive changes, i.e. those threatening traditional markets, lifestyles, and cultural values (van der Boezem et al. 2015). The forecast of synthetic food future can be found in a publication by Winston Churchill (1932) who claimed that "we shall escape the absurdity of growing a whole chicken in order to eat the breast or wing, by growing these parts separately under a suitable medium". Earlier professor Berthelot in an interview in 1894 predicted that in the year 2000, "the epicure of the future is to dine upon artificial meat, artificial flour, and artificial vegetables" (Ferreira, 2017). After that time the issue of possible future forms of food heats the imagination of scholars, sci-fi writers and business especially on the waves of successes of chemistry, synthetic biology and other disciplines (see, for example, Belasco, 2006; Hubert et al. 2010; Selle & Barrangou, 2015; Shelomi, 2016; Buscemi, 2018 and many others).

Theme of innovations in food industry is growingly becoming controversial nowadays with international and non-governmental organizations, social movements, government and academia being in great concern about the way the food industry can feed growing global population (Springer & Duchin, 2014; Thurner & Zaichenko, 2015; Smith, 2015; Hanjra et al. 2016). According to Food and Agriculture Organization of the United Nations estimation (FAO, 2017), in 2050 when world population would reach 9.73 billion agriculture will face the requirement to produce almost 50 percent more food, feed and biofuel in comparison to 2012. Profound technological innovations offers an opportunity to meet future food needs sustainably. Food start-ups contribute to the discussion actively developing radical innovations that may change the sector (Chkaiban, 2016).

Innovations are traditionally viewed as radical depending on the perceived degree of novel knowledge incorporated in the innovation that must be grasped and the connected learning barriers (Kline & Rosenberg, 1986; Dewar & Dutton, 1986). Green et al (1995) in their classical research identified the following multiple dimensions of radical innovation: technological uncertainty (the degree to which the technology employed in the innovative undertaking is not well developed or understood in the general scientific community), technical inexperience (the degree to which the firm lacks the required technological experience and knowledge), business inexperience (the degree to which the firm lacks required business experience and knowledge) and the technology development cost. Companies and governments more often use technology foresight methods in order to anticipate and respond to radical innovations (Vasamo et al. 2016).

Traditionally food foresights are based on experts' interviews, panel surveys and workshops. Salo et al. (2004) report on foresight study of the Finnish food and drink industries up until 2015 which was based on semi-structured interviews, surveys and workshop for experts from industrial firms, research organizations and public agencies. One of the most complex and profound international studies in this sphere is Foresight (2011) Global Food and Farming Futures Project that involved around 400 leading experts and stakeholders from about 35 low-, middle- and high-income countries who contributed to investigation of the challenges for the global food system until 2050 and development of indicators and tools that policy makers need to take. Jermann et al. (2015) conducted research with the participation of food professionals from industry, academia and government in North America and Europe. Experts' task was to identify novel technologies either applied now or with the potential to be adopted in 5-10 years and evaluate commercialization factors, to enumerate associated regulations and limitations.

Application of modern information technologies including Big Data analysis and text mining gains popularity in innovation studies including in food sector. Zong et al. (2010) analyze application of data mining in the food manufacturing industry for production and logistics management, coding systems, sales and customer data management. Ways of applying data mining for food supply networks analysis and management turn into a highly dynamic research field (Beulens et al. 2006; Ting et al. 2014). Another prospective area in food industry involves Natural-Language Processing (NLP) and machine learning innovative instruments implementation for automated food ontology construction, for example, for food traceability control (Salampasis et al. 2012; Pizzuti et al. 2014). Li and Ko (2007) proposed automated food ontology construction mechanism for diabetes diet care. Text mining is also used for nutritional content analysis of products' labels (Do Nascimento et al. 2013) or for mobile apps examination of side effects of food additives and another components via semantic web (Ertuğrul, 2016).

However, there were little attempts conducted in horizon scanning of the food industry as a whole. Based on this rationale, the paper initiates to close the literature gap by complex mapping of innovations in food industry applying text mining and Big Data analysis techniques.

When text mining software is exploited for science and technology (S&T) information resources in order to inform technology management tech mining (TM) approach emerges. It can be defined as the combination of bibliometrics, patent analysis and different NLP tools to collect, process and represent competitive technological intelligence in the visually understandable form (Porter & Cunningham, 2004; Yoon B., 2008). Until recently simple unrelated bibliometrics and patent analysis methods based on official classifications and citation indexes dominated the industry and technology research studies. As a rule, researchers tend to scope their field of study on clearly defined small samples of documents.

The limitation of such approaches is that it is impossible to evaluate whether enough text corpus analyzed to make valid inference due to lack of confidence about data boundaries for fields. Moreover, different types of documents (for example, grants and patent applications, scientific papers and media articles) rarely automatically analyzed in complex in order to reveal synergies among this data. Traditional orientation of many researches towards structured metadata highly determined by methodological, computational and technical difficulties of integrated text mining. Using existing classifications not only prevents researchers from catching the young areas of innovation and science, yet not included in official taxonomies, but prevents them from optimally setting topic boundaries, as opposed to modern methods of topic clustering / topic modelling approaches, which became de-facto standard in text mining applications since mid-2000 (Duh and Kirchhoff, 2008; Chen, 2009; Ramage et al. 2011; Zhao et al., 2012; Chuang et al., 2013; Zhang et al., 2014). The approach based on bare keywords without semantic multipliers (such as word2vec based associative / synonymous series of terms) is even weaker than the use of official taxonomies, as the results are ultra-sensitive to meaningless fluctuations in word usage patterns, homonyms, and lexical hypes (see application of these approaches in Levy and Bullinaria, 2001; Varathan et al., 2011; Goldberg and Levy, 2014; Rong, 2014; Vidra, 2015). To find, download and process the most relevant documents that potentially contain information about emerging S&T areas, researchers apply iterative processes from general search categories and terms to more concrete words and phrases (Haung et al., 2015).

Taking into account these numerous drawbacks of such linear, non-machine learning (ML)-augmented approaches along with the absence of modern and complex horizon scanning exercises in the food industry, our research question sounds like: what are the potential and limitations of text mining approach for mapping radical innovations in food industry. Accordingly the paper proceeds as follows: first, it describes the research methodology, second, it presents the results obtained. They include maps of main thematic segments that were categorized in semi-automated way coupled with the grouping of the technologies and scientific research topics based on two-dimensional (impact and growth of impact[7]) clustering into four categories: weak signals, emerging trends, stable segments and niche activities together with additionally identified weak signals on possible scientific breakthroughs in the global food industry, including most promising startup strategies and food innovation controversies. Finally, the paper deals with the critical discussion of the potential and limitations of text mining approach for mapping radical innovations in food industry.

Methodology

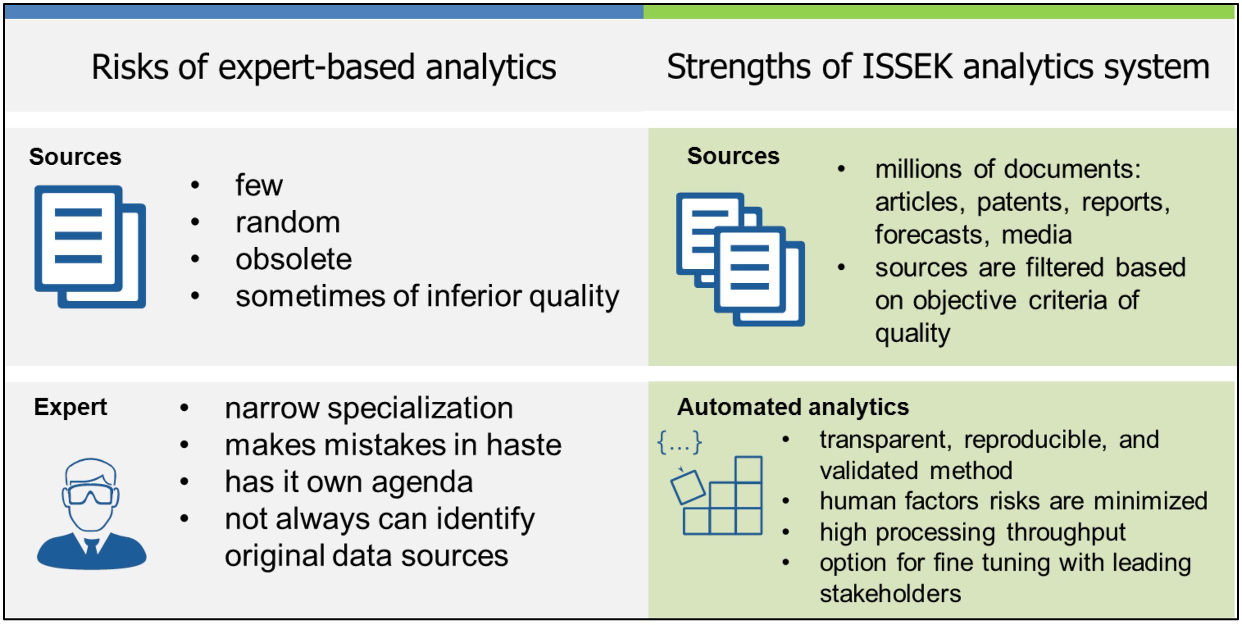
The acceleration of the development of science, technology and innovation entails a radical transformation of markets, business models (Baden-Fuller & Haefliger, 2013), methods of production and distribution of products and services, institutions and socio-cultural relations (Fountain, 2004; Kallinikos, 2010). Being highly dynamical and important, the innovation process in agriculture & food (A&F) sector is one that is needed to be monitored objectively, continuously and thoroughly (Hoholm, 2011). According to IBM (2017), 90% of the data in the world today has been created in just the last two years. The rapid growth of unstructured information, particularly heterogeneous textual data, which is not feasible to analyze by traditional methods, is becoming a serious challenge for the adoption of evidence-based and timely decisions (see e.g. Tsui et al., 2014; Grau et al., 2015; Bystrov et al., 2016; Singh et al., 2016; Engel 2017 as the examples of latest discussions on this topic). The importance of evidence-based approaches (in spotting critical technologies, existing and emerging trends, as well as forecasting future development and setting STI priorities) as the way to reduce sectoral experts' subjectivity, incompleteness of their knowledge and bias of views favoring familiar subject areas has been in the center of attention of many technology and strategy stakeholders for a long time (Davies, 1987; Tichy, 2004; Ecken et al. 2011).

However, recent progress in both hardware and software produced a number of enabling technologies and methods providing robust solutions to data analysis challenges mentioned earlier. The rapid development of big data requires prompt notification of key decision-makers in science, technology and innovation (STI). Therefore, in the era of big data the efficient work of decision makers is impossible without the use of automatic monitoring and analytical information systems, built in the paradigm of data mining; machine learning; self-improving deep-learning-based prediction models; natural language processing; knowledge discovery; agile user-friendly automatic ontology building; rich context insight-oriented user experience, etc. For example, de Miranda Santo et al. (2006) started applying text mining (TM) to scientific publications in order to spot major topics in the emerging and highly dynamic sphere of nanotechnology in order to make up the limited knowledge of experts. Altuntas et al. (2015) looked at patent data to assess technology readiness levels and their potential for further development relying on alternative approaches and indicators. Zhang et al. (2016) used National Science Foundation (NSF) awards data to identify technology topics, their dynamics and attempt to forecast future development in the computer science area of big data research.

In this context, our article studies the ongoing and emerging technological changes in the global agriculture and food sector (using the case study of food market innovations) by applying original big data augmented text mining system, developed in the National Research University Higher School of Economics, Institute for Statistical Studies and Economics of Knowledge (ISSEK HSE) (see *Figure 1* for details).

[7] More details on these indices are given in the methodology section of the paper.

Figure 1. ISSEK HSE Text Mining System vs. expert-based analytics



Risks of expert-based analytics

Sources

- few
- random
- obsolete
- sometimes of inferior quality

Expert

- narrow specialization
- makes mistakes in haste
- has it own agenda
- not always can identify original data sources

Strengths of ISSEK analytics system

Sources

- millions of documents: articles, patents, reports, forecasts, media
- sources are filtered based on objective criteria of quality

Automated analytics

- transparent, reproducible, and validated method
- human factors risks are minimized
- high processing throughput
- option for fine tuning with leading stakeholders

Source: National Research University Higher School of Economics

This text mining system for analytical tasks in the field of technology foresight responds to the following challenges:

1) Efficient processing the growing amount of information about the development of science, technology and innovation, which usually leads to an information overload of experts and analysts and cannot be adequately processed without the use of automated methods for the extraction and systematization of knowledge.

2) Answering the growing need for automated systems for the monitoring and analysis of data from various sources to ensure timely, science-based decisions in STI policy (on the state and corporate levels).

3) Reducing the growing risks of unfounded expert claims that may not always be adequately verified due to shortage of time and require, therefore, automated integrity control tools.

4) Answering the need for objective verification of experts involved in the preparation of materials and discussions for the formation of science and technology policy.

5) Counterbalancing the negative effects of increasing specialization and fragmentation of knowledge of individual scientists, researchers and industry experts, as well as specialists of public authorities in STI policy in the era of interdisciplinary research and convergence of technologies.

The article builds on recent Russian Agriculture and Food Industry Foresight 2030 (Gokhberg & Kuzminov, 2017; Gokhberg et al. 2017 a). It encompasses a wide range of quantitative and qualitative methods: in-depth interviews with managers and leading experts, survey with 400 participants from more than 50 specialized companies, research organizations and universities and workshops along with text mining of more than 12 thousand analytical and foresight documents for trends identification. Semantic analysis of heterogeneous unstructured text data sources was conducted and resulted in complex depiction of ongoing and emerging technology-related trends in the area of food innovation.

The study of radical innovations in the global food industry implied making an extensive, momentary snapshot of the situation in the area (tech trends, key players, prospects of development) with the use of original big data augmented text mining analysis of recent texts describing significant research and development results, as well as most recent newsfeeds and web publications. The sources for the analysis are heterogeneous and include vast amounts of research papers, patent and grants applications with both full-text data and additional structured metadata, analytical reports by main international organizations and national key players, various media and news resources, including all the major technology innovation, disruption and venture capital news websites. The choice of information sources for the analysis is based on the results of a reputation assessment of the quality, reliability of data sources (including taking into account the impact factor of scientific journals, the citation of scientific publications, the popularity of professional Internet resources, the ratings of analytical centers, etc.). Filling of the database was accomplished with the use of both subscription-based and free sources to the volume of petabyte scale (of bare text) with the use of continuous automated data accumulation techniques. The methodological principles of gathered data integration corresponds to standard practices in information management.

The general approach of the study draws on a wide array of tools from computer science, on the one hand, and the convergent methodology of technology foresight, future-oriented technology analysis, trend spotting and tech mining (see Popper 2008a, 2008b; Miles et al., 2016), on the other. The tools developed on this basis include the instruments for large-scale collection, extraction, and transformation of data; knowledge discovery; integrated statistical, syntactic, and semantic analysis of text and natural language processing.

The developed methodology combines statistical and semantic analysis of the full textual content of documents, as well as structured metadata with the elements of machine learning and expert-proved sectoral ontologies (e.g. a formal naming and definition of the types, properties, and interrelationships of the entities in a particular domain) developed in semi-automated way during the 10 years of trends and technologies mapping within the framework of the Long-Term Science and Technology Foresight of the Russian Federation (more methodological details see in Bakhtin et al. 2017; Gokhberg et al. 2017 b).

The principal components of the text-mining based sectoral trend and technology spotting rely on ngram identification and categorization, knowledge extraction based on syntactic and ontology relationships of hierarchical nested synonym sets and, finally, pattern recognition of strictly typed statements on the technology evolution milestones and future forecast estimates along with statistical indicators of dynamics and intensity.

All documents from all sources analyzed are processed using the same instrumental pipeline based on the use of open program libraries that integrate best practices to obtain reliable results. The primary processing of the natural language is carried out – the transformation of large arrays of unstructured textual information into structured tables, arrays and vector representations. For this each document is divided into separate sentences, words and phrases with different linguistic characteristics.

Foremost documents are split into sentences, sentences are tokenized (split into words) and words are lemmatized (all word forms used in the text are standardized to normal form), syntactic relationships are calculated using "property-function" approach. These connections that characterize the properties and functions describing nouns are considered to be the basis for the formation of meaningful terms.

Later, ngrams (phrases that combine *n* amount of words) are built using the syntactic relationships of governor word (usually noun) with other words through "adjective modifier" and "compound" dependencies (e.g. “innovative pig farming”, where farming – is a governor word, innovative – is an adjective modifier of farming, pig – compound dependency of farming). Then, based on the table of terms and their relation to the document, the matrix of co-occurrence is calculated based on terms that appear in the same document.

Once the co-occurrence matrix is built, the network analysis can be performed. Each term in the matrix represents a separate node of the graph. The amount of documents, where two terms occur, represents the weighted edge between nodes. Network analysis helps to calculate the *centrality* of each node (the amounts of edges that connect to the node with their weights).

Then, clustering algorithm is applied to divide graphs into subgraphs of the most closely related nodes. In other words, the more co-occurrence a group of terms have between each other, the more probability that they end up in the same cluster. Once clustering is done, it objectively divides all terms in several groups. These groups are perceived as S&T topics. Optionally, they can be named by *N* most central ngrams constituting them, like it is done in Web of Science research fronts (see more Kessler, 1960; Small, 1973, 1980; Garfield, 1990 on methodological basis of research fronts detection).

After all extracted terms are grouped into topics, their trending behavior can be analyzed. For that matter the study uses calculation of relative textual statistics, cauterization and classification of results based on dynamics and relative occurrence of terms in the documents of the study area. Such representation of data is perceived as time series where the main task is to divide all terms by their trend patterns as part of the dynamic clustering based on the following parameters.

The frequency index demonstrates the amount of times term occurs in documents. The more term occurs – the more popular it is, which in turn shows its general relevance to the topic of food. This index calculated and then normalized as:

$$f_{term} = \sum_{i}^{n} sentence\_occurrences_i$$

- f_{term} – term frequency;
- i – document's number;
- n – number of documents;
- $sentence\_occurrences_i$ – the amount of sentences in the i-th document, in which a term has occurred.

$$Z_{term} = \frac{f_{term} - \min(F)}{\max(F) - \min(F)} * 99 + 1$$

- Z_{term} – frequency index of a term;
- $min(F)$ – minimum value of term frequency;
- $\max(F)$ – maximum value of term frequency.

The trending index for each term is the proxy of the growth or reduction of general attention to a specific term. It means, the higher the index, the bigger the increase of occurrence in documents the term had from 2005 to 2015. *Trending index* is calculated based on the slope of the linear trend of a given term's statistics distribution over the years in the corpus of food-related documents using following formula of linear regression equation:

$$k = \frac{n * \sum(i * f_{term,i}) - \sum i * \sum f_{term,i}}{n * \sum(i^2) - (\sum i)^2}$$

- k – slope of linear trend built using linear regression (*trending index*);
- i – publication year of documents in which term occurs;
- n – number of documents' publication years included in the analysis;
- $f_{term,i}$ – term frequency based on occurrence in documents of a particular year.

Trend mapping helps to classify items based on frequency and trending indexes. When trend mapping is done for the clusters as a whole, proportion of high trending terms (by distribution quintiles) is used for discriminations. Trend mapping of terms used only terms representing objects of certain kinds (concepts, events, technologies, markets), and can also be enriched by trend mapping of terms representing named entities (companies, persons, geographies, etc.).

Trend maps are accompanied by *semantic maps* that show the thematic structure of the analyzed area: what narrow themes and special concepts does it consist, how it can be categorized. The end result of using the tool of semantic maps for knowledge discovery and further interpretation of the results by the experts are reasonable structural lists and rankings (based on described text-mining statistical proxy) of future directions or systemic problems in a particular field, as well as recommendations for priority areas of development.

Identification of semantic patterns is another type of used text mining tools. *Trends, weak signals and other data patterns* were identified using two text-mining-oriented approaches: quantitative and qualitative. Quantitative identification relies on term statistics of occurrence in documents over the years (trends = high dynamics + high frequency, weak signals = high dynamics + low frequency). Qualitative spotting is based on linguistic patterns in the sentences – statements that might describe trends, weak signals and other potentially relevant objects. For example, phrases with verbs in future tense or gerund form that describe some transformation, increase or decrease in values or other forms of dynamics can describe trends.

On the basis of described analysis major global trends, promising areas of science and technology development can be identified, as well as organizations and key experts being centers of radical innovations in each field can be mapped.

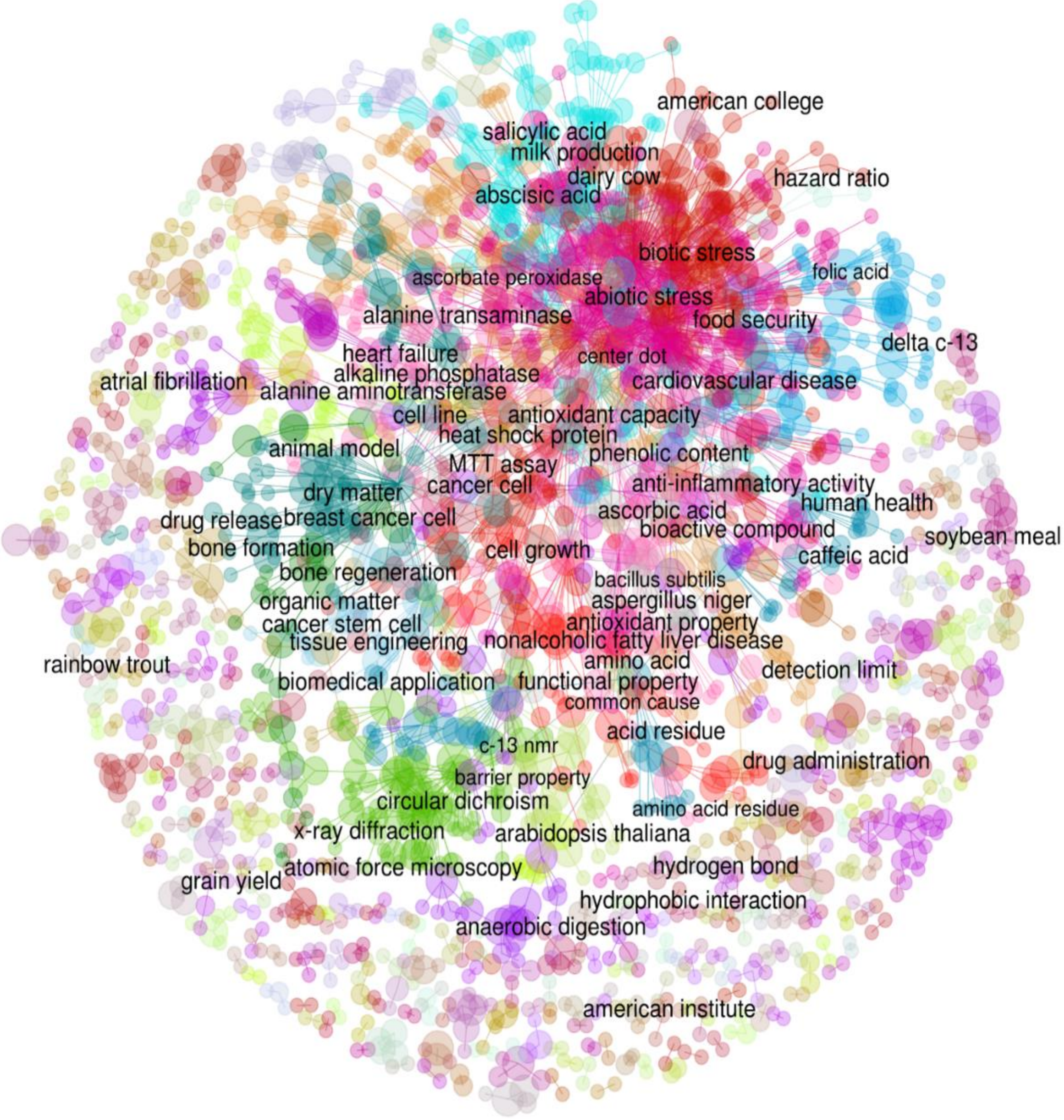
Results

Quantitative findings: extraction of main topics and calculating trends

In a first step, we plot intuitive, user-friendly visualizations of automatically generated ontologies of considered field – semantic maps. As mentioned earlier they show the most relevant topics within the studied corpus of proceeded documents (food-related documents in our case) and their interrelationships. Semantic maps are formed from processing of large arrays of text data from scientific publications, patents, industry analytical documents or popular news, but can be built from a small corpus of specially selected documents or pseudodocuments of various sizes (as a general rule, about 10 to 20 thousand separate topical sentences are enough for building a robust stably clustered graph visualized by a semantic map). Semantic maps are very useful in cases where there are no official classifications or established taxonomy for the studied objects. They are valuable tool for the structural characterization of areas of scientific and technological development, including emerging ones. Semantic maps also provide recommendations on effective search terms to further find relevant information in iterative, interactive way. In addition, they can be used for quick characterization of some corpus documents, allowing the user to make a quick decision about whether it is necessary to carefully examine the particular document.

Figure 2 is a semantic map of more than 3000 terms that occurred more than 2 times in the corpus of about 12 000 food-related texts: including research papers, patent and grants applications, analytical reports by main international organizations and national key players, various media and news resources, including all the major technology innovation, disruption and venture capital news websites. On this map we can see the key topics of scientific interest in the field of food innovation. Every single node (bubbles) on a semantic map represents a term derived from a corpus of the studied documents. Location, size, and color of each item have a strict mathematical sense. Bubble size is determined by number of term occurrences in the studied corpus documents. All terms are grouped into 12 clusters (with 20-50 to several hundred terms in each) based on their semantic closeness. The color of bubbles represents dynamically allocated clusters according to the results of solving an optimization problem of the classification of unstructured full-text data derived from the corpus food-related documents. Configuration of areas with high density of bubbles of one cluster (color) also has direct meaning. If this area is on the periphery and placed compactly it represent a narrow, niche field of research. If the area is located close to the center of a map and is scattered in different directions it means that a given topic (within food industry) has many interdisciplinary connections and is highly convergent. For a more clear understanding we present the most important terms from this scientific map in *Table 1*.

Figure 2. Semantic map of terms extracted from food-related documents



Note: Each bubble represents a single term derived from food-related documents. Bubble size is determined by number of term occurrences in the studied corpus of documents. All terms are grouped into 12 clusters based on their semantic closeness. Clusters are highlighted by different colors.

Source: National Research University Higher School of Economics's Text Mining System

Semantic analysis of food-related documents shows that the most frequently used term is “tissue engineering”. This term occurred 1627 times in the corpus of 12 000+ food-related documents. The other main topic clusters (named as centers of strongly connected colored groups of terms) are the following trends, issues and instruments applied in food innovations: "amino acid", "oxidative stress", "gene expression", "infrared spectroscopy", "mechanical property", "cell proliferation", "body weight", "molecular mechanism", "abiotic stress", "liquid chromatography", "insulin resistance", "sensitive detection", "food allergy". The most dynamic food-related terms (in term of growth of number of occurrences) is “protein complex” with term trending index value of 7.16 points. The other highly dynamic term is “food ingredient” (term trending index is 5.43). The most frequently occurred food-related terms in general have low values of term trending index.

Table 1. The most important scientific terms on the semantic map for food-related documents

Term	First year mentioned	last year mentioned	Term frequency	Term trending index
tissue engineering	2000	2016	1627	1.26
vitamin a	1994	2016	900	0.80
protein complex	2000	2016	715	7.16
dairy product	1998	2016	612	0.41
amino acid	1997	2016	610	0.34
stem cell	1999	2016	594	0.23
food additive	1999	2016	593	1.07
rainbow trout	1999	2016	524	0.99
membrane protein	1999	2016	451	1.24
milk beverage	2001	2015	415	0.99
soybean meal	2000	2016	405	1.63
poultry production	1998	2016	380	0.76
protein product	1992	2016	366	0.63
aroma chemical	2002	2016	348	3.28
staple food	1987	2016	282	0.52
vitamin c	1998	2016	240	1.09
food waste	1997	2016	220	3.61
fish consumption	1992	2016	213	1.31
vitamin b	1997	2016	209	1.08
GM food	1999	2016	168	1.21
soy protein	2000	2016	168	1.33
grain legume	1991	2016	160	0.85
yeast extract	1999	2016	119	3.68
food ingredient	1998	2016	108	5.43
wheat flour	1998	2016	101	0.50
protein alternative	2000	2016	93	3.72
milk powder	1995	2016	89	0.79
plant protein	1991	2016	88	0.85
protein synthesis	2000	2016	80	1.46
yeast cell	1999	2016	78	0.87
vitamin d	2000	2016	71	1.69
cow milk	1998	2016	67	2.13
fishmeal oil	1999	2016	60	1.79
lab grown meat	2008	2015	60	2.53
insect protein	2002	2016	57	2.09
food safety	1997	2016	54	0.11
protein digestibility	1998	2016	54	0.69
protein isolate	2000	2016	54	1.39
alternative protein sources	1999	2016	51	0.61

Notes: 1. This table shows the most frequently occurred terms in the corpus of 12 000+ food-related documents. 2. Term trending index shows the growth of number of this term occurrences. Highest values of term frequency and term trending index indicators are highlighted in green, the lowest values – in red.

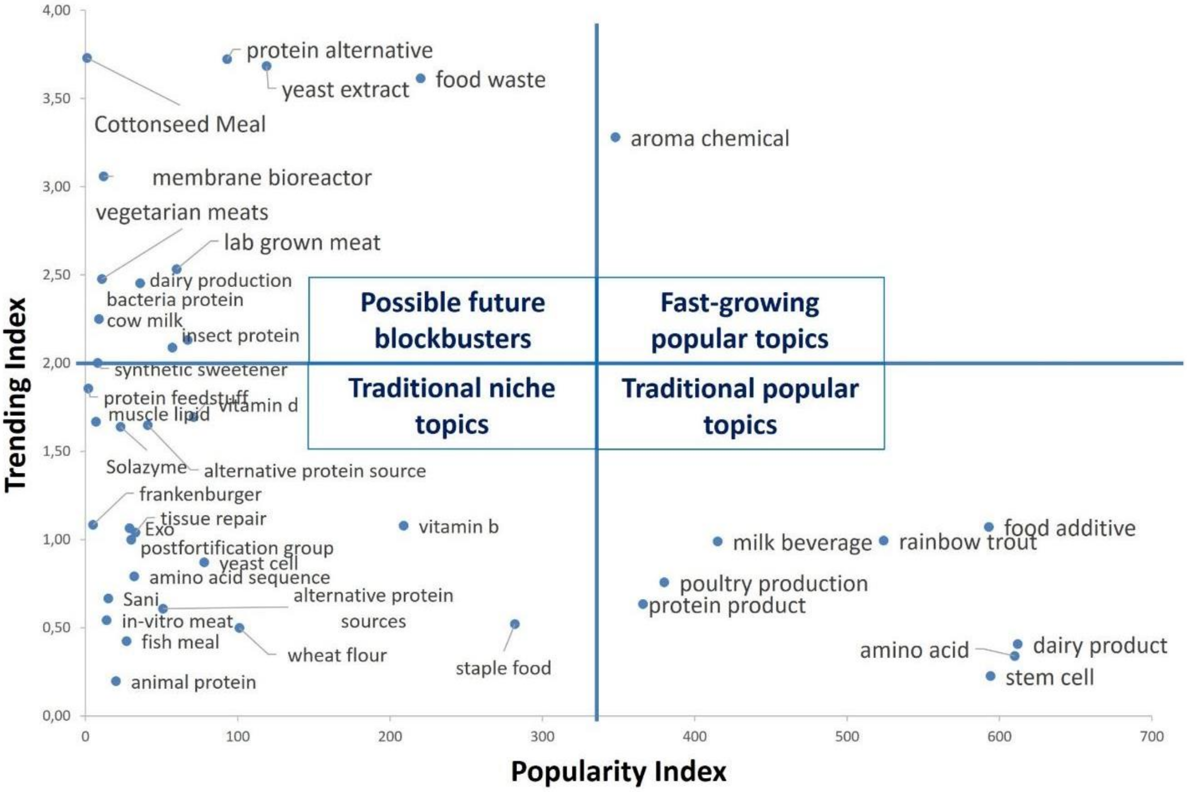
Source: National Research University Higher School of Economics’s Text Mining System

For more efficient identification of trends in the studied area, structural analysis of terms derived from food-related documents is complemented by the visualization of the significance (popularity index) and dynamism (trending index) of these terms. Such an analysis allows clearly identify in addition to trends also the stagnating topics of interest and so-called weak signals (for example, emerging technologies or potential future markets).

Figure 3 visualizes *trend map* or the scatterplot on the axes of popularity index and trending index for the terms, derived from food-related documents. Popularity index (horizontal axis) measures the mean intensity of the discussion of a given topic in the corpus of processed food-related documents. It is calculated as the term frequency of occurrence in the documents. Trending index (vertical axis) measures the growth of relative occurrence of a given term from year to year over a period of time.

Trend map is an intuitive visualization that allows us to divide the subjects of most important technologies, markets, companies, etc. on four quadrants depending on the value of popularity and trending index for each specific term – higher/below the median values of popularity and trending index for all terms derived from food-related documents.

Figure 3. Clusters of terms by research and development significance and trending rate



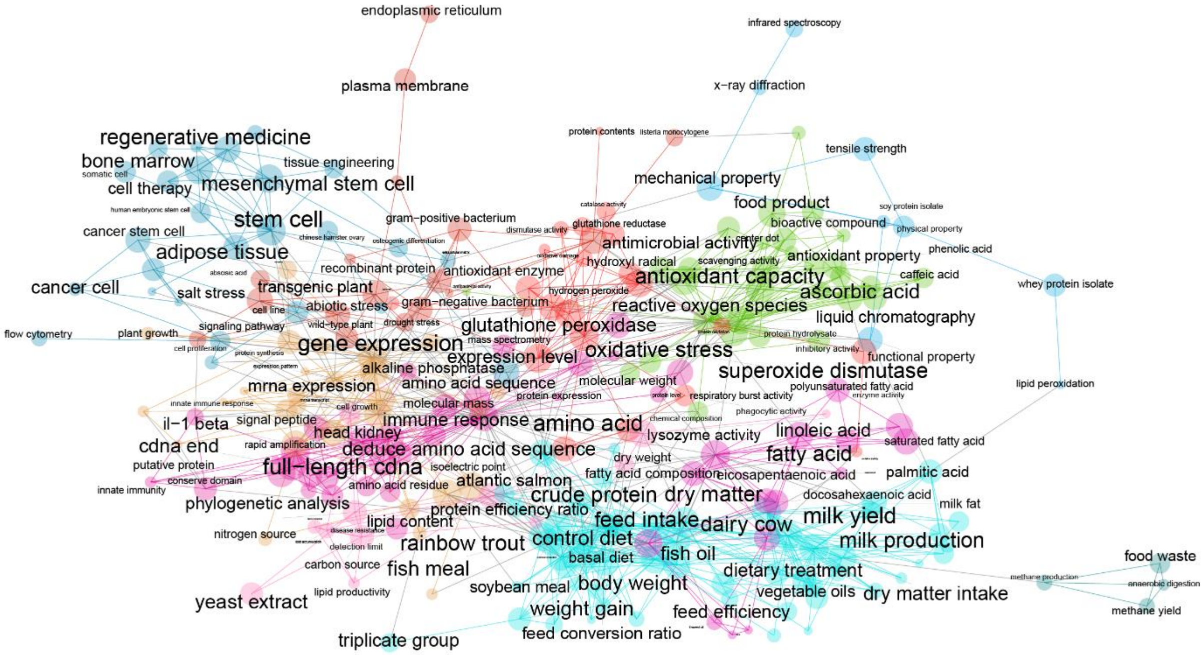
Source: National Research University Higher School of Economics's Text Mining System

In the upper right quadrant fast-growing popular topics are concentrated. Topics in the upper left are weak signals (possible future blockbusters that yet not very well-known but are extremely fast growing in overall popularity / impact). In the lower right quadrant popular but stagnant topics are concentrated. Finally the lower left quadrant is filled by traditional niche topics (they are neither very impactful, nor fast growing – of course, in relation to the leaders; but the mere fact of their appearance on a trend map while thousands of other terms have so low values that are out of bounds of the trend map – tells us that these topics are also very important).

Detailed analysis: future of food based on documents of 2016

In order to understand the latest developments in the sphere of food, we built separate semantic map and other analytical outputs for the documents published in 2016[8] (Figure 4). This allows for spotting the newest trends, weak signals, research areas, technologies and other relevant terms.

Figure 4. Core concepts for the future of food, 2016



Source: National Research University Higher School of Economics's Text Mining System

We proxy the importance of a term by comparing normalized popularity indexes in different sources of data (analytic reports, research papers or media and news). This way we distill a total of 386 terms most relevant to the field. For example, the term "stem cell" indeed marks a relevant trend, but it is enthusiastically used in media papers. In contrast to such buzzwords, terms like "food product", "food security", "food safety" and to a lesser extent "dairy product" and "dairy production" are terms that are specific for analytical reports. A high values of normalized indicators show that the term is relevant, impactful, significant, resilient[9] and thereby marks an important trend.

Table 2. Main related concepts for the future of food, 2016

Term	Normalized popularity index in analytic reports	Normalized popularity index in research papers	Normalized popularity index in media & news	Normalized popularity index in all corpuses
stem cell	41.24	100.00	100.00	100.00
amino acid	68.27	75.89	52.41	87.91
food product	100.00	14.11	49.47	55.02
vitamin A	42.07	42.10	70.33	54.03

[8] Data for 2017 is not full and incomplete.

[9] Stable or growing in significance irrelevant to hypes and fluctuations in new texts issued.

Term	Normalized popularity index in analytic reports	Normalized popularity index in research papers	Normalized popularity index in media & news	Normalized popularity index in all corpuses
food security	78.20	6.68	17.74	36.76
tissue engineering	13.04	29.64	9.81	25.65
food safety	42.05	7.11	28.61	23.25
membrane protein	13.46	21.92	3.94	19.42
protein complex	14.75	17.10	8.05	16.87
rainbow trout	11.66	17.85	2.47	15.49
dairy product	25.92	6.43	13.04	14.21
protein content	18.37	11.09	4.23	13.49
vegetable oil	18.15	10.03	10.40	13.34
amino acid sequence	16.13	10.92	6.29	12.66
protein synthesis	11.72	13.38	5.70	12.60
vitamin D	3.92	15.00	14.51	11.54
protein product	13.77	9.19	8.93	10.71
Atlantic salmon	9.08	9.92	2.76	8.59
vitamin C	6.76	9.44	12.75	8.42
food additive	14.64	3.30	8.34	6.63
cow milk	5.81	6.61	11.28	5.74
tissue repair	5.01	7.13	5.41	5.13
food waste	9.76	3.30	11.58	4.98
plant protein	3.72	7.66	4.53	4.88
protein source	10.41	3.01	7.76	4.58
drug carrier	3.47	7.54	1.59	4.33
staple food	12.37	1.70	5.99	4.25
animal protein	9.85	2.42	6.58	3.78
yeast cell	4.07	5.33	7.76	3.66
protein isolate	3.94	5.48	3.06	3.18
poultry production	10.30	1.54	3.06	2.92
food ingredient	7.07	2.80	5.70	2.79
milk protein	2.36	5.28	3.94	2.48
fish consumption	5.43	3.58	2.76	2.35
food supplement	5.45	2.80	5.70	2.14
protein digestibility	3.56	4.10	2.76	1.96
soybean meal	4.40	3.25	4.82	1.92
membrane bioreactor	1.00	5.57	1.00	1.80
protein food	3.56	3.70	3.64	1.77
vitamin b	3.00	3.73	5.41	1.76
wheat flour	4.36	2.70	3.64	1.35
milk powder	5.41	2.16	3.06	1.33

Term	Normalized popularity index in analytic reports	Normalized popularity index in research papers	Normalized popularity index in media & news	Normalized popularity index in all corpuses
food assistance	6.85	1.00	4.23	1.19
yeast extract	4.52	2.68	1.59	1.16
drug delivery application	2.09	3.83	1.29	1.00
…	…	…	…	…

Source: National Research University Higher School of Economics's Text Mining System

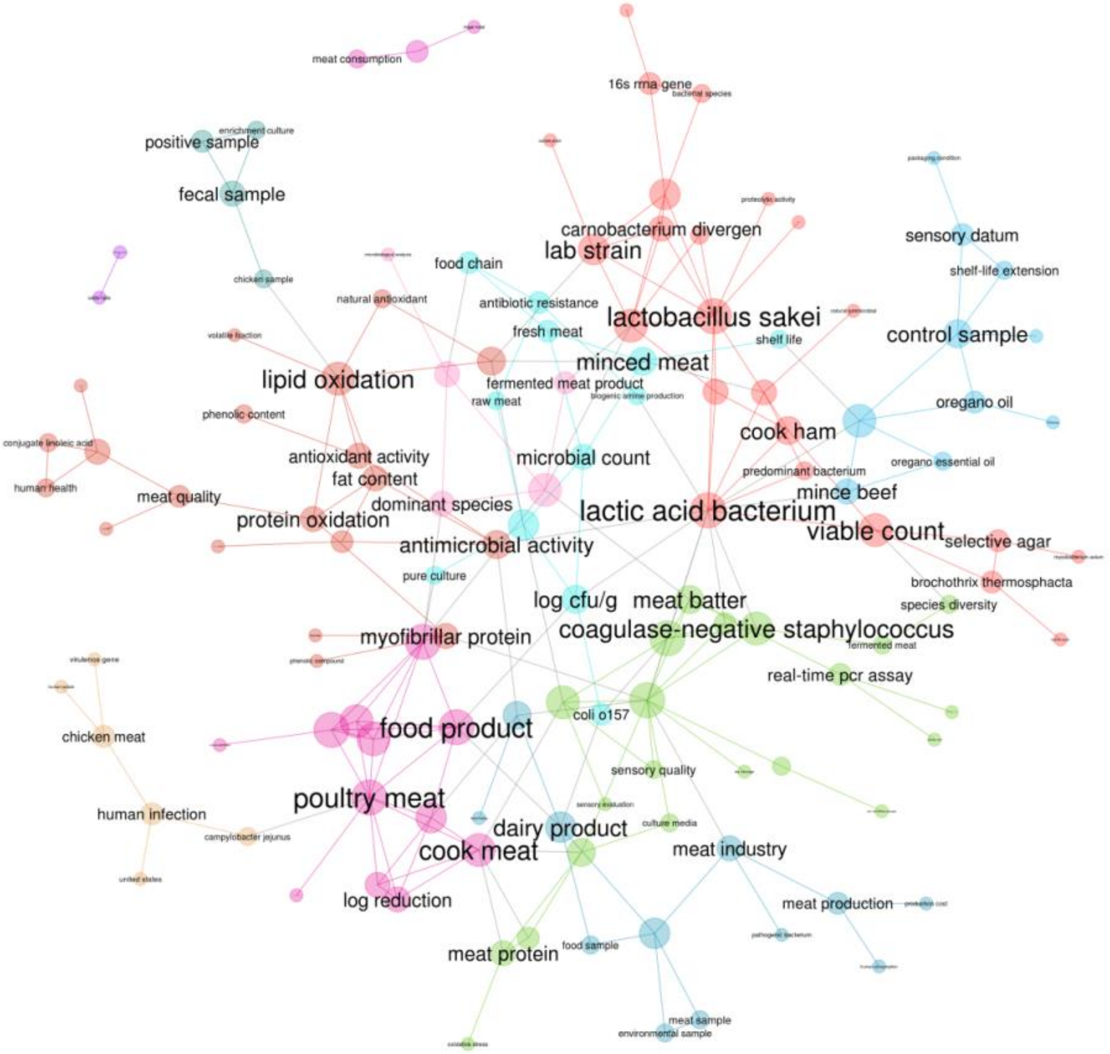
Looking onto the values of calculated rates we can also detect "purely scientific" terms – those that have high values of specificity for scientific papers in relation to media, news and analytical reports. These terms are e.g. "membrane protein", "protein complex", "rainbow trout", "membrane bioreactor" and some other. They are either topics that cannot be efficiently communicated to wide society from science by now (too sophisticated research, early research, or research irrelevant to current and prospective challenges).

Case study: cultured meat

All clusters of terms identified in during semantic analysis show general distribution of topics of studied research area with all interlinkages. However, content of each cluster can be assessed in bigger detail. Terms of one cluster can be further clustered into even smaller groups to show hierarchy of lower levels.

For further discussion, we chose the cluster for "cultured meat" – one of the clusters on the big map with term “cultured meat” being its center (describing the main theme of the spotted topic) (Figure 5). Cultured meat is also known as cell-cultured meat, in-vitro meat, artificially cultured meat, lab meat, lab-grown meat or synthetic meat. This cluster contains about 600 terms related to food innovation in general (all irrelevant terms and stop words were excluded). However, only 96 of these terms have indeed a high impact (occurrence greater than 10).

Figure 5. Fragment of semantic map for "cultured meat" cluster



Note: Each bubble represents a single term derived from food-related documents. Bubble size is determined by number of term occurrences in the studied corpus. All terms from “cluttered meat” cluster are again grouped into subclusters based on their semantic closeness. Subclusters are highlighted by different colors.

Source: National Research University Higher School of Economics’s Text Mining System

In the center of the semantic map for artificial meat are the thematic areas with highest impact and with the greatest involvement in global interdisciplinary context. Out of all concepts the followings trends can be identified: fortified foodstuffs, vitamin-rich soft drinks, plant-protein-based meat, milk and eggs substitutes, artificially-grown meat, personalized food and many other.

Qualitative findings: extraction of meaningful statements about future of food

Knowledge extraction based on syntactic and ontology relationships of hierarchical nested synonym sets and pattern recognition of strictly typed statements on the technology evolution milestones and future forecast estimates enriches the insights acquired by the use of semantic maps and trend maps. All the statements of the knowledge extraction module are relevant finished sentences with their neighboring sentences constituting the context of the statements, and various forms of their automatic and semi-automatic, rule-based and ML-based aggregation and generalizations.

One of the most useful types of statements to be extracted are quantitative forecast estimates. Quantitative and qualitative forecast estimates are those statements in the texts that describe the prospects for the development of any markets, market segments or products in quantitative terms (average annual growth rate of the market, its volume, etc.) or in qualitative (textual description of possible scenarios for market development). In addition to market information, forecasts can also include any information about the strategic plans of companies and countries, adopted laws, expected phenomena (like global warming), etc. We can automatically identify such statements by marker terms, dictionaries of market indicators and named entities (names of companies and their products, geographical locations, etc.) along with their syntactic links to other terms in the sentences. Box 1 and Box 2 provide the examples of global food projections extracted from food-related documents.

Box 1. Key identified global food projections before 2030

Feeding the growing world population will require cheap fats and protein sources

- With the world's population is expected to increase to 8 billions in 2028 from 6.5 billions in 2005, **only palm oil has the potential to be the source of fats and vegetable oil to feed the people around the globe** (Tan et al., 2009).
- **World trade in soybean meal climbs** by more than 12 million tons (about 21 percent) in the projections to 2020 (USDA, 2011).

Global agriculture cannot produce necessary amounts of grain to sustainably support meat livestock farming

- **The global demand for grain in 2030 would reach 5.1 billion tones if the world's population adopted approximately the same diet enjoyed by Europeans.** This would create a global grain shortfall of 2.3 billion tones, compared with estimated 2030 production of 2.8 billion tones. Demand in 2030 will not reach 5.1 billion tones because many people in the world will still lack the necessary income to increase their consumption of animal protein.

Fish food production to play major role in global protein supply growth

- **World per capita fish food consumption is projected to reach 20.6 kg in 2022**, up from nearly 19 kg per capita, on average, in 2010-12 (OECD-FAO, 2013).
- Not only is world per capita fish consumption to rise by 16 percent by 2021 from 2011 levels, but overall **global seafood production is predicted to rise from 154 million metric tones in 2011 to 172 million tones by 2021** (FAO, 2014).

Global demand for milk and milk substitutes will steadily grow

- At a global level, the demand for dairy products **will expand by 23% over the ten year projection period**, approaching 48 Mt by 2024 (OECD-FAO, 2015).

Box 2. Key identified global food projections after 2030

World’s population will grow further, which will require new industrial methods of food production

- The world’s population is expected to grow to over 9 billion people by 2050; there will be a need to **raise food production by some 70 percent globally** and by almost 100 percent in developing countries (FAO, 2012 a).

Climate change poses additional risks to traditional agriculture systems

- If global average temperature increases **by more than 2 degrees Celsius, it will be difficult to secure enough food** for the global population, which is projected to increase to 9 billion people by 2050 (FAO, 2012 b).

Growing livestock sector can become unsustainable, which will require synthetic meat production instead

- The share of livestock production (meat, milk and dairy products and eggs) in total world production would increase from 36 percent in 2005/2007 **to 39 percent in 2050** (from 30 to 35 percent in the developing countries) (FAO, 2012 c).
- In response to the projected demand growth, production of meat and dairy products in the developing world is projected to increase substantially by 2050 compared with 1999/2001 volumes. Between those years, total production of meat and dairy products by developing country farmers is projected to **grow by 206M tones and 410M tones respectively** (Boland et al., 2013).

Aquaculture will remain one of the key solutions to the global food problem

- Given a current fisheries and aquaculture production for human consumption of about 136.2 million tones (animals from capture fisheries and aquaculture), with annual per capita fish consumption remaining at 19.2kg, a similar proportion of fish going into fishmeal, fish oil and other non-food uses as today, and a world population of 9.6 billion people**, approximately 47.5 million additional tones of food fish will be needed in 2050** (FAO, 2014).

Forecasts, estimates, retrospective milestones, enumerations of technologies, skills, etc. are good for general foresight tasks. However, we also have in our arsenal other instruments, suitable for strictly practical market forecasts and business intelligence tasks. Among them are statements on innovative companies’ strategies. We show this on the case study of food industry startups business models that can shape the future structure of food markets.

The Text Mining System derived startup strategies from the corpus of food-related documents. To detect them we apply the developed technique of special words and syntactic models identification in media and news texts that radically increase the probability of a sentence containing them to characterize startup strategies. The practical application of this techniques derives the list of more than 100 startups from food industry. Some of them are presented below (see Box 3)

Box 3. Identified food industry startup strategies

- **Impossible Food's** mission is to make a plant-based version of meat that's just as good as real, animal-based meat while being more sustainable.

  URL: http://www.businessinsider.com/pat-brown-on-why-impossible-foods-didnt-sell-to-google-2016-6 (date last accessed 22/01/18).
- Through bioengineering, **Modern Meadow** is figuring out ways to brew leather from a few cells in a laboratory, instead of raising livestock.

  URL: https://techcrunch.com/2016/04/27/bolt-threads-modern-meadow-ceos-to-speak-at-disrupt-ny-2016/ (date last accessed 22/01/18).
- San Francisco-based startup **Memphis Meats** made its public debut, with the team's ambitious plan to grow beef and pork in laboratory bioreactors—and to be the first company to bring lab-grown meat to market.

  URL: https://gizmodo.com/this-biotech-startup-promises-lab-grown-pork-within-fiv-1756365159 (date last accessed 22/01/18).
- San Francisco-based **Hampton Creek** is a food technology startup with a focus on developing plant-based products. Last month, the Food and Drug Administration (FDA) ordered Hampton Creek to change the name of its Just Mayo line of plant-based mayonnaise, saying that it violates federal regulations that require mayonnaise-branded products to contain eggs.

  URL: https://www.theverge.com/2015/9/3/9254261/just-mayo-egg-lobby-fda-usda-hampton-creek (date last accessed 22/01/18).
- **Ginkgo** took the investment to help it move beyond its original production of synthetic fragrances to cosmetics, nutritional products and health and consumer products.

  URL: https://techcrunch.com/2015/07/23/ginkgo-bioworks-takes-on-zymergen-with-45-million-in-series-b-funding/ (date last accessed 22/01/18).
- **Clara Foods** Cooks Up $1.7 Million In Funding To Make Egg Whites From Yeast Instead Of Chickens.

  URL: https://techcrunch.com/2015/07/09/clara-foods-cooks-up-1-7-million-in-funding-to-make-egg-whites-from-yeast-instead-of-chickens/ (date last accessed 22/01/18).
- **Enviroflight** (in the USA), **Ynsect** (in France) and **Protix** (in the Netherlands) have built large-scale insect production facilities.

  URL: http://www.bbc.com/future/story/20141014-time-to-put-bugs-on-the-menu (date last accessed 22/01/18).
- **One Lab** discovered a type of seaweed that tastes just like fried bacon (and it's two times healthier than kale).

  URL: https://www.engadget.com/2015/11/01/concept-cars-betting-big-on-solar/ (date last accessed 22/01/18).
- **Nestlé Purina** launched a personalized dog food product that will allow pet owners to order food tailored to their pet's unique needs.

  URL: http://time.com/3543378/nestle-personalized-dog-food/ (date last accessed 22/01/18).
- Instead of using cashews and almonds to replicate the curdy backbone as some alternative cheese makers do, **Muufri** is bioengineering yeast to produce authentic milk proteins, which will give it the same taste and nutrition as regular milk. Muufri's GMO process starts by adding cow DNA sequences into the yeast cells.

  URL: https://newatlas.com/muufri-synthetic-milk/34415/ (date last accessed 22/01/18).
- **Nutrinsic**, a Colorado-based company formerly known as **Oberon FMR**, which just raised $12.7 million for its technology extracting high quality proteins from industrial food and agricultural waste streams.

  URL: https://techcrunch.com/2014/08/26/early-stage-investor-artiman-ventures-raises-350-million/ *(date last accessed 22/01/18).*

Further our text mining system allows to detect innovation controversies. We determine as innovation controversy the following statements about ongoing conflict, radical discrepancy between opinions of key stakeholders or other series of events featuring diverging interests and points of view, confrontation with unclear future results. To detect innovation controversy from a separate document we use sentiment analysis with the use of custom rules and dictionaries. Some examples of about 300 innovation controversies derived from news items about food innovations are presented below (Box 4).

Box 4. Identified food innovation controversies

- The American Egg Board (AEB) may have violated federal laws by using public funds to **try to obstruct Hampton Creek from selling its eggless mayonnaise alternative "Just Mayo."** Emails obtained under a Freedom of Information Act (FOIA) request show that the American Egg Board (AEB), egg industry executives, and an official from the US Department of Agriculture (USDA) discussed several strategies to bring down Hampton Creek and its Just Mayo mayonnaise alternative, which outgoing AEB president Joanne Ivy described as **"a crisis and major threat to the future of the egg product business"** in an August 2013 email.

  URL: https://techcrunch.com/2015/10/26/the-american-egg-board-may-have-used-public-funds-to-conspire-against-hampton-creek/ (date last accessed 22/01/18).

- And why not **label all pesticides used** on food, whether synthetic or organic, if awareness is important? ... Inconvenient Truth: There Are Synthetic Pesticide Residues On Organic. ... However, the other 40 of the 41 different pesticides detected on the organic foods were synthetic chemicals that are not approved for use on organic. **Finding synthetic pesticide residues on organic is not unprecedented.**

  URL: https://www.forbes.com/sites/stevensavage/2016/02/08/inconvenient-truth-there-are-pesticide-residues-on-organics/#b3d1724683b9 (date last accessed 22/01/18).

- The food and supplement industries have been very **willing to promote the fear of scarcity of these functional nutrients, in order to construct consumer demand for their fortified foods and nutritional supplements**, such as cholesterol-lowering margarine fortified with plant sterols, omega-3 enriched orange juice and probiotic yogurts.

  URL: https://www.huffingtonpost.com/gyorgy-scrinis/diet-and-nutrition_b_4524915.html (date last accessed 22/01/18).

- Technically, the Food and Drug Administration **frowns on the practice of fortifying snack foods or carbonated beverages with added nutrients** ... The FDA is planning Experimental Studies on Consumer Responses to Nutrient Content Claims on Fortified Food – that means they want to find out **whether fortifying snack foods with vitamins and noting its nutritional content on labels would convince people to swap out regular old junk food** with a slightly less unhealthy form of junk food. Your 'federal family' at work, supposedly to protect you, again? No, this is about changing a 20 year old rule regarding labels on foods that have been modified to be a little healthier.

  URL: https://www.huffingtonpost.com/michael-f-jacobson/fortified-soda_b_2204583.html (date last accessed 22/01/18).

- Gluten-free food is certainly not needed for 99 percent of the public but it has become a $5 billion industry by slapping gluten-free labels on meat. **Stoli targets health-conscious drinkers with gluten-free vodka.**

  URL: https://uk.reuters.com/article/us-stolichnaya-gluten/stoli-targets-health-conscious-drinkers-with-gluten-free-vodka-idUKKCN0WI2NX (date last accessed 22/01/18).

Finally our text mining system allow determining weak signals on possible scientific breakthroughs by identifying certain syntactic constructions with special verb tenses and certain modal verbs used. Statements with adverbs, adjectives and other words that display low level of certainty (might, may, can) along with various values may display weak signals. Weak signals refer to "the early signs of possible but not confirmed changes that may later become more significant indicators of critical forces for development, threats, business and technical innovation" (Saritas & Smith, 2011). They represent the first signs of paradigm shifts, or future trends, drivers or discontinuities. In this respect, the study of weak signals aims at collecting and analyzing data for the purpose of providing early indications of potential developments in particular field. The anticipatory intelligence gathered through the scanning of weak signals is used to provide stakeholders opportunities to develop early responses to capitalize on, protect against, or mitigate the impact of potential disruptions in the future. Some examples are presented in Box 5.

Box 5. Identified weak signals on possible scientific breakthroughs

- Researchers tested three nontraditional **bioplastic materials – albumin, whey and soy proteins –** as alternatives to conventional petroleum-based plastics that pose risks of contamination.
  URL: https://news.uga.edu/recipe-for-antibacterial-plastic-plastic-plus-egg-whites-0315/ (date last accessed 22/01/18).
- A new study suggests that bread from certain wheat varieties have differentiated sensory properties and that could mean **customized breeding for more personalized food** in the future.
  URL: http://www.science20.com/news_articles/sensory_properties_may_inform_wheat_modification-155597 (date last accessed 22/01/18).
- A team from Northwestern University in Evanston, Illinois, have found inspiration for the new compound in enzymes called phosphotriesterases. Usually produced by bacteria, **these proteins deactivate some pesticides** – and nerves gases – in milliseconds.
  URL: https://www.gizmodo.com.au/2015/03/a-new-synthetic-compound-can-neutralise-chemical-weapons-in-minutes/ (date last accessed 22/01/18).
- The Food and Agriculture Organization of the UN recently published a paper marshaling evidence that **insect protein can be used to make things like sausage**.
  URL: https://www.wired.com/2013/09/fakemeat/ (date last accessed 22/01/18).
- **Parents have gastric bypass; children's DNA may receive the benefits** ... Gastric bypass surgeries would, at first glance, seem to tackle the problems of obesity through simple physics: with a smaller stomach, there's only so much food a person can ingest.
  URL: https://arstechnica.com/science/2013/05/parents-have-gastric-bypass-childrens-dna-may-receive-the-benefits/ (date last accessed 22/01/18).
- Human diarrheal diseases claim the lives of 1.8 million children around the world and impair the physical and mental development of millions more and these findings offer hope that **genetically fortified milk could eventually help prevent such diseases**. In the study, researchers fed young pigs milk from goats that were genetically modified to produce higher levels of lysozyme, which occurs naturally occurs in the tears, saliva and milk of all mammals.
  URL: http://www.science20.com/news_articles/genetically_modified_milk_cures_diarrhea_could_save_millions_kids_annually-105860 (date last accessed 22/01/18).
- **Peptide depots and DNA tattoos could deliver drugs in the future** ... Proteins manufactured in bacteria can supplement natural proteins; DNA-based vaccines can make immunity more specific while lasting for months at room temperature.
  URL: https://arstechnica.com/science/2013/01/peptide-depots-and-dna-tattoos-for-future-drug

Discussion and conclusions

Timeliness of conducted research is noticeable in conditions of rapid growth of unstructured information about highly dynamical and important innovation process in agriculture and food sector, which is becoming not feasible to analyze by traditional methods in order to make evidence-based and timely decisions. The results of the study with an application of big data text mining approach include the multifaceted lists and maps of ongoing and emerging technology-related trends, along with weak signals on possible scientific breakthroughs in the global food industry coupled with most promising startup strategies and food innovation controversies.

Among the identified trends in food sector innovations the most influential one is tissue engineering strongly connected with artificially-grown (cultured) meat which was viewed in more details. As the most dynamic topics were named those related with proteins and food chemistry. Issues of consumers' obesity prevention and food allergy responses were also included in the map of radical innovations directions in global food industry. Par to the course these topics were found in the text-mining-derived startup strategies from the corpus of food-related documents which are mostly connected with synthetic food innovations and personalized food. In addition to this picture it was illustrated that most food innovation controversies are related to legal status improvement of radical innovations, food labeling and consumers' anxieties about fortified and synthetic foods. Identified weak signals on possible scientific breakthroughs include advance of packaging materials with special properties for keeping food fresh for longer time, approach to more personalized food production through customized breeding, enriched proteins application and food-related health problems prevention. Summing-up, conducted research allowed to map a complex picture of sectoral radical innovations and future development directions, in contrast with previous research traditionally focused on particular narrow aspects and separate branches of food industry.

Mostly it was achieved owing to an introduced approach for trend- and technology-mapping based on big data text mining. As our research question was about the potential of text mining methods for mapping the radical innovations in food industry, we demonstrated that their advantages include possibility of combination of broad scope of information sources in a complex analysis with high velocity and reproducibility based on standardized data processing and analysis procedures. It enhances transparency and objectivity of the analysis as well as possibilities of periodic monitoring which is crucial in mapping of radical innovations in food industry.

Along with reproducibility and high level of formalization, text mining approach brings speed as one of the main benefits for application in many areas in both private and public sectors. This type of analysis can be performed on a regular basis (up to weekly periodicity) based on constant accumulation of textual data and serves as a framework for constant S&T monitoring for early warning on changing technology landscape and its implications on markets.

The instrument is also suitable for express testing of the expert perspectives, hypotheses, insights and desk-research-based analytics and reports based on them, mainly for gap analysis on the important sectoral topics covered. The methodology implemented programmatically in a radically new (and highly specialized) interactive information system brings to an expert an integrated analytical environment. This environment augments the expert's capabilities by bringing attention to gaps in overviews of existing trends and in recommendations on the future actions, and by suggesting comprehensive yet comprehensible lists of high-level constructs based on long traditions and time-proven foresight / future studies methodologies. Developed approach allows performing rapid aggregation of the whole polyphony of existing expert views on the technology development in virtually any area. The application of the implemented hybrid expert-machine learning interaction reduces to a very high extent subjectivity, bias and incompleteness of expert knowledge as opposed to bare (un-augmented) desk-research and expert-panel approaches traditional in technology foresight studies.

Together with the advantages described, our research also allows to figure out certain limitations of text mining approach for horizon scanning tasks. First of all, any text mining study faces the challenge of comprehensive and sufficient coverage of information sources to be processed and analyzed. As all sorts of empirical research are sensitive to input data quality, text mining especially delicate in documents sampling procedures as far as biased corpus analysis can result in inaccurate output, illustrating only one side of the considered complex issue. Therefore the most promising strategy there is the usage of big data samples which can guarantee necessary heterogeneity of raw input data.

Strongly connected with this aspect is another challenge for productive big data augmented text mining approaches – technical limitations of efficient storage and high-performance analysis of big data to make possible the fast and flexible output delivery. These problems demand an application of advanced methods and algorithms developed in modern computer science along with high-performance server infrastructure entailing appreciable capital investments and recruitment of top-class specialists. An alternative to development of proprietary text mining system more relevant for small-scale studies and assuming less possibilities of flexible calibration and adaptation of analytical tools is the purchase of license for private text mining software like VantagePoint[10] or TechWatchTool[11]. However, working with these tools usually presuppose a lot of manual cleaning, filtering and grouping of keywords in order to evaluate the presented information. Furthermore, the majority of systems are limited to standard data representations (lists, co-occurrence matrices and factor/cluster maps), which do not allow to work deeply with dependencies between different types of data (for example, building multidimensional relationships).

Based on obtained results it can be supposed that further studies can develop the approach proposed by adding complexity to big data text mining proxies of sectoral radical innovations on the basis of flexible combination of various information sources. For instance, proposed trend map quadrants method has a specific limitation due to dependence of term popularity on generalization level and several other factors connected with peculiar properties of each type of information sources. For example, it might be the case that a term is often used and at the same time often criticized, being popular in negative terms rather than illustrating prospective area for company development or investment. Another example can be connected with the excessive attention in media sources to overestimated but lacking in prospects so called hype concepts and products being promoted by PR agents and other interested stakeholders. Additional analytical tools like sentiment analysis or life cycle analysis are needed for more reliable evaluation. Further development of the research field will be connected with the necessity to look at the radical innovations in food industry in the broader context, taking into account along with sectoral tendencies also a big picture of global challenges and trends.

Alongside the sectoral researchers and future-oriented studies methodologists, the results of the study can also be used in strategic planning activities by governments and corporations in the agriculture and food sector, as well as some adjoining industries, such as agricultural machine building, mineral fertilizers and other agrochemicals, biofuels industry, fisheries, environment protection, information & communication, space and robotics for smart & precision agriculture. Another area of application is setting the priorities of STI policy in food industry and supporting / regulating the advanced food technologies, in the paradigm of evidence-based policy-making. Furthermore, results create the basis for new directions of venture investments and sectoral best practices adaptation, as they include extracted data on successful technology startup strategies covered by specialized sectoral media sources.

[10] URL: https://www.vantagepointsoftware.com/ (date last accessed 09/02/18).

[11] URL: https://www.dfki.de/lt/publication_show.php?id=5541 (date last accessed 09/02/18).

References

Altuntas, S., Dereli, T., & Kusiak, A. (2015). Forecasting technology success based on patent data. *Technological Forecasting and Social Change*, *96*, 202-214.

Baden-Fuller, C., & Haefliger, S. (2013). Business models and technological innovation. *Long range planning*, 46(6), 419-426.

Bakhtin, P., Saritas, O., Chulok, A., Kuzminov, I., Timofeev, A. (2017). Trend Monitoring for Linking Science and Strategy. *Scientometrics*. 111 (3), 2059-2075.

Belasco, W. (2006). Meals to come: A history of the future of food (Vol. 16). Univ of California Press.

Beulens, A. J., Li, Y., Kramer, M. R., & van der Vorst, J. G. (2006). Possibilities for applying data mining for early Warning in Food Supply Networks. In *CSM'06 20th Workshop on Methodologies and Tools for Complex System Modeling and Integrated Policy Assessment*, 6-7.

Boland, M. J., Rae, A. N., Vereijken, J. M., Meuwissen, M. P., Fischer, A. R., van Boekel, M. A., ... & Hendriks, W. H. (2013). The future supply of animal-derived protein for human consumption. *Trends in Food Science & Technology*, 29(1), 62-73.

Buscemi, F. (2018). Today – The Future: Meat Forecast. In From Body Fuel to Universal Poison (pp. 127-149). Springer, Cham.

Bystrov, I. I., Kozichev, V. N., & Tarasov, B. V. E. (2016). Conceptual basis for the unstructured information automated processing in perspective control systems. *Sistemy i Sredstva Informatiki [Systems and Means of Informatics]*, *26*(4), 162-171.

Chen, B. (2009, April). Latent topic modelling of word co-occurence information for spoken document retrieval. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference* (pp. 3961-3964). IEEE.

Chkaiban, M. (2016). The routes beneath the roots: a system map for prospective food innovators striving for sustainable disruption. Doctoral dissertation, Massachusetts Institute of Technology.

Chuang, J., Gupta, S., Manning, C., & Heer, J. (2013). Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of the 30th International Conference on machine learning (ICML-13)* (pp. 612-620).

Churchill W. (1932). Fifty Years Hence. *Popular Mechanics*. URL: http://rolandanderson.se/Winston_Churchill/Fifty_Years_Hence.php (date last accessed 10/01/18).

Davies, M. F. (1987). Reduction of hindsight bias by restoration of foresight perspective: Effectiveness of foresight-encoding and hindsight-retrieval strategies. *Organizational Behavior and Human Decision Processes*, 40(1), 50-68.

De Miranda Santo, M., Coelho, G. M., dos Santos, D. M., & Fellows Filho, L. (2006). Text mining as a valuable tool in foresight exercises: A study on nanotechnology. *Technological Forecasting and Social Change*, *73*(8), 1013-1027.

Dewar, R.D., & Dutton, J.E. (1986). The Adoption of Radical and Incremental Innovations: An Empirical Analysis. *Management Science*, 32 (11), 1422-1433.

Do Nascimento, A. B., Fiates, G. M. R., dos Anjos, A., & Teixeira, E. (2013). Analysis of ingredient lists of commercially available gluten-free and gluten-containing food products using the text mining technique. *International journal of food sciences and nutrition*, 64(2), 217-222.

Duh, K., & Kirchhoff, K. (2008, July). Learning to rank with partially-labeled data. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 251-258). ACM.

Ecken, P., Gnatzy, T., & Heiko, A. (2011). Desirability bias in foresight: consequences for decision quality based on Delphi results. *Technological Forecasting and Social Change*, 78(9), 1654-1670.

Engel, J. (2017). Improving retrieval of structured and unstructured information: Practical steps for better classification, navigation and search. *Business Information Review*, *34*(2), 86-95.

Ertuğrul, D. Ç. (2016). Foodwiki: A mobile app examines side effects of food additives via semantic web. *Journal of medical systems*, 40(2), 41.

FAO (2012 a). Sustainable diets and biodiversity: directions and solutions for policy, research and action. Rome: FAO.

FAO (2012 b). Peatlands. Rome: FAO.

FAO (2012 c). World agriculture towards 2030/2050: the 2012 revision. Rome: FAO.

FAO (2014). The State of World Fisheries and Aquaculture. Rome: FAO.

FAO (2017). The future of food and agriculture. Trends and challenges. Rome. URL: http://www.fao.org/3/a-i6583e.pdf (date last accessed 22/01/18).

Ferreira, A. (2017). Future Food in Utopia: A View from the Twenty-Second Century. *Cadernos de Literatura Comparada*, 36.

Foresight, U. K. (2011). The future of food and farming. *Final Project Report, London, The Government Office for Science*. URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/288329/11-546-future-of-food-and-farming-report.pdf (date last accessed 10/01/18).

Fountain, J. E. (2004). Building the virtual state: Information technology and institutional change. Brookings Institution Press.

Garfield E G. (1990). KeyWords Plus: ISI's breakthrough retrieval method. Part 1. Expanding your searching power on Current Contents on Diskette. *Essays of an Information Scientist*, 13, 295- 299.

Gokhberg L., Kuzminov I. F., Chulok A. A., Thurner T. (2017 a). The future of Russia's agriculture and food industry between global opportunities and technological restrictions. *International Journal of Agricultural Sustainability*, 15 (4), 457-466.

Gokhberg, L., & Kuzminov, I. (2017). Technological future of the agriculture and food sector in Russia. In *The Global Innovation Index 2017. Innovation Feeding the World* (pp. 135-141).

Gokhberg, L., Kuzminov, I., Bakhtin, P., Tochilina, E., Chulok, A., Timofeev, A., & Lavrynenko, A. (2017 b). Big-Data-Augmented Approach to Emerging Technologies Identification: Case of Agriculture and Food Sector. National Research University Higher School of Economics Working Paper. URL: https://www.hse.ru/mirror/pubs/lib/data/access/ram/ticket/18/1516715550da8e00d0ae35bdbf1a1b9a2a6863333c/76STI2017.pdf (date last accessed 22/01/18).

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint,* arXiv: 1402.3722.

Grau, B., Ligozat, A. L., & Gleize, M. (2015). Precise information retrieval in structured and unstructured information sources: Challenges, approaches and hybridization. *Traitement Automatique des Langues*, 56(3), 75-99.

Green, S. G., Gavin, M. B., & Aiman-Smith, L. (1995). Assessing a multidimensional measure of radical technological innovation. *IEEE transactions on engineering management*, 42(3), 203-214.

Hanjra, M. A., Noble, A., Langan, S., & Lautze, J. (2016). Feeding the 10 Billion within the Sustainable Development Goals Framework. Food Production and Nature Conservation: Conflicts and Solutions, Chapter 15.

Hayden, E. C. (2015). Tech investors bet on synthetic biology: once hesitant, Silicon Valley venture capitalists are warming to the idea of engineered cells. *Nature*, 527(7576), 19-20.

Hoholm, T. (2011). The Contrary Forces of Innovation: An Ethnography of Innovation in the Food Industry. Palgrave Macmillan.

Hubert, B., Rosegrant, M., Van Boekel, M. A., & Ortiz, R. (2010). The future of food: scenarios for 2050. *Crop Science*, 50, S33.

IBM (2017). 10 Key Marketing Trends for 2017. URL: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN (date last accessed 07/02/18).

Jermann, C., Koutchma, T., Margas, E., Leadley, C., & Ros-Polski, V. (2015). Mapping trends in novel and emerging food processing technologies around the world. *Innovative Food Science & Emerging Technologies*, 31, 14-27.

Kallinikos, J. (2010). Governing through technology: Information artefacts and social practice. Springer.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *Journal of the Association for Information Science and Technology*, 14(1), 10-25.

King, A. (2017). Technology: The Future of Agriculture. *Nature*, 544(7651), S21-S23.

Kline, S. J., & Rosenberg, N. (1986). An overview of innovation. The positive sum strategy: Harnessing technology for economic growth, 275, 305.

Levy, J. P., & Bullinaria, J. A. (2001). Learning lexical properties from word usage patterns: Which context words should be used. In *Connectionist models of learning, development and evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop* (p. 273 - 282).

Li, H. C., & Ko, W. M. (2007, August). Automated food ontology construction mechanism for diabetes diet care. In *2007 International Conference on Machine Learning and Cybernetics* (Vol. 5, pp. 2953-2958). IEEE.

Maggio, A., Maggio, A., Van Criekinge, T., Van Criekinge, T., Malingreau, J. P., & Malingreau, J. P. (2016). Global food security: assessing trends in view of guiding future EU policies. *foresight*, 18(5), 551-560.

Miles, I., Saritas, O., & Sokolov, A. (2016). Foresight for Science, Technology and Innovation. Springer.

OECD/FAO (2013). OECD-FAO Agricultural Outlook 2013-2022. OECD/FAO.

OECD/FAO (2015). OECD-FAO Agricultural Outlook 2015. OECD/FAO.

Pizzuti, T., Mirabelli, G., Sanz-Bobi, M. A., & Goméz-Gonzaléz, F. (2014). Food Track & Trace ontology for helping the food traceability control. *Journal of Food Engineering*, 120, 17-30.

Popper, R. (2008a). How are foresight methods selected? *Foresight*, 10(6), 62-89.

Popper, R. (2008b). Foresight methodology. *The handbook of technology foresight*, Springer 44-88.

Porter, A., Cunningham, S. (2004). Tech Mining: Exploiting New Technologies for Competitive Advantage. John Wiley & Sons, Inc.

Pretty, J., Sutherland, W. J., Ashby, J., Auburn, J., Baulcombe, D., Bell, M., Bentley, J., ... & Pilgrim, S. (2010). The top 100 questions of importance to the future of global agriculture. *International Journal of Agricultural Sustainability*, 8(4), 219-236.

Ramage, D., Manning, C. D., & Dumais, S. (2011, August). Partially labeled topic models for interpretable text mining. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 457-465). ACM.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Rockström, J., Williams, J., Daily, G., Noble, A., Matthews, N., Gordon, L. ... & de Fraiture, C. (2017). Sustainable intensification of agriculture for human prosperity and global sustainability. *Ambio*, 46(1), 4-17.

Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint*, arXiv:1411.2738.

Sahn, D. E. (Ed.). (2015). The fight against hunger and malnutrition: the role of food, agriculture, and targeted policies. OUP Oxford.

Salampasis, M., Tektonidis, D., & Kalogianni, E. P. (2012). TraceALL: a semantic web framework for food traceability systems. *Journal of Systems and Information Technology*, 14(4), 302-317.

Salo, A., Konnola, T., & Hjelt, M. (2004). Responsiveness in foresight management: reflections from the Finnish food and drink industry. *International Journal of Foresight and Innovation Policy*, 1(1-2), 70-88.

Saritas, O., & Smith, J. E. (2011). The big picture–trends, drivers, wild cards, discontinuities and weak signals. *Futures*, 43(3), 292-312.

Selle, K., & Barrangou, R. (2015). CRISPR-Based Technologies and the Future of Food Science. *Journal of food science*, 80(11).

Shelomi, M. (2016). The meat of affliction: Insects and the future of food as seen in Expo 2015. *Trends in Food Science & Technology*, 56, 175-179.

Singh, S. K., Mani, N., & Singh, B. (2016). A Framework for Extracting Reliable Information from Unstructured Uncertain Big Data. In *Intelligent Decision Technologies 2016* (pp. 175-185). Springer International Publishing.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology*, *24*(4), 265-269.

Small, H. (1980). The ABCs of cluster mapping. Part 1. Most active fields in the life sciences in 1978. *Essays of an Information Scientist*, 4, 634 – 641.

Smith, P. (2015). Malthus is still wrong: we can feed a world of 9–10 billion, but only by reducing food demand. *Proceedings of the Nutrition Society*, 74(3), 187-190.

Springer, N. P., & Duchin, F. (2014). Feeding nine billion people sustainably: conserving land and water through shifting diets and changes in technologies. E*nvironmental science & technology*, 48(8), 4444-4451.

Tan, K. T., Lee, K. T., Mohamed, A. R., & Bhatia, S. (2009). Palm oil: addressing issues and towards sustainable development. *Renewable and sustainable energy reviews*, 13(2), 420-427.

Thurner, T. W., & Zaichenko, S. (2015). The Feeding Of The Nine Billion—A Case For Technology Transfer In Agriculture. *International Journal of Innovation Management*, 19(02), 1550026.

Tichy, G. (2004). The over-optimism among experts in assessment and foresight. *Technological Forecasting and Social Change*, 71(4), 341-363.

Ting, S. L., Tse, Y. K., Ho, G. T. S., Chung, S. H., & Pang, G. (2014). Mining logistics data to assure the quality in a sustainable food supply chain: A case in the red wine industry. *International Journal of Production Economics*, 152, 200-209.

Tsui, E., Wang, W. M., Cai, L., Cheung, C. F., & Lee, W. B. (2014). Knowledge-based extraction of intellectual capital-related information from unstructured data. *Expert systems with Applications*, 41(4), 1315-1325.

USDA (2011). USDA Agricultural Projections to 2020 (OCE-2011-1). URL: http://usda.mannlib.cornell.edu/usda/ers/94005/2011/OCE111.pdf (date last accessed 22/01/18).

van der Boezem, T., Schobe, G., Pascucci, S., & Dries, L. (2015). Startups: key to open innovation success in the Agri-Food sector. URL: http://www.compete-project.eu/fileadmin/compete/files/working_paper/COMPETE_Working_Paper_21_Start-ups_and_Open_Innovation.pdf (date last accessed 22/01/18).

Varathan, K. D., Sembok, T. M. T., Omar, N., & Kadir, R. A. (2011, June). Retrieving answers from multiple documents using semantic skolem indexing. In *Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on* (pp. 53-56). IEEE.

Vasamo, A. L., Baiyere, A., & Saukkonen, J. (2016, March). Corporate Technology Foresight Methods in Anticipation of Disruptive Innovations. In ISPIM Innovation Symposium (p. 1 – 19). The International Society for Professional Innovation Management (ISPIM).

Vidra, J. (2015). Implementation of a Search Engine for DeriNet. In *ITAT* (pp. 100-105).

von Grebmer, K., Bernstein, J., Hossain, N., Brown, T., Prasai, N., Yohannes, Y., ... & Foley, C. (2017). 2017 global hunger index: The inequalities of hunger. Intl Food Policy Res Inst.

World Preservation Foundation (2017). Business Report – The Future of Food. URL: http://worldpreservationfoundation.org/wp-content/uploads/2017/09/WPF-Business-Doc-2017.pdf-.pdf (date last accessed 22/01/18).

Yoon, B. (2008). On the development of a technology intelligence tool for identifying technology opportunity. *Expert Systems with Applications*, 35, 124–135.

Zappacosta, M., & Gomez y Paloma, S. (1999). Food for the future. *Foresight*, 1(6), 575-582.

Zhang, W. N., Liu, T., Yang, Y., Cao, L., Zhang, Y., & Ji, R. (2014). A topic clustering approach to finding similar questions from large question and answer archives. *PloS one*, 9(3), e71511.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting and Social Change*, 105, 179-191.

Zhao, J., Li, X., & Jin, P. (2012). A time-enhanced topic clustering approach for news web search. *International Journal of Database Theory and Application*, 5(4), 1-10.

Zong, C. M., Wang, H. B., & Du, X. J. (2010). The Application of Data Mining in the Food Manufacturing Industry. *Applied Mechanics and Materials*, 20, 856-860.

Elena Khabirova

Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow, Russian Federation. E-mail: etochilina@hse.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.