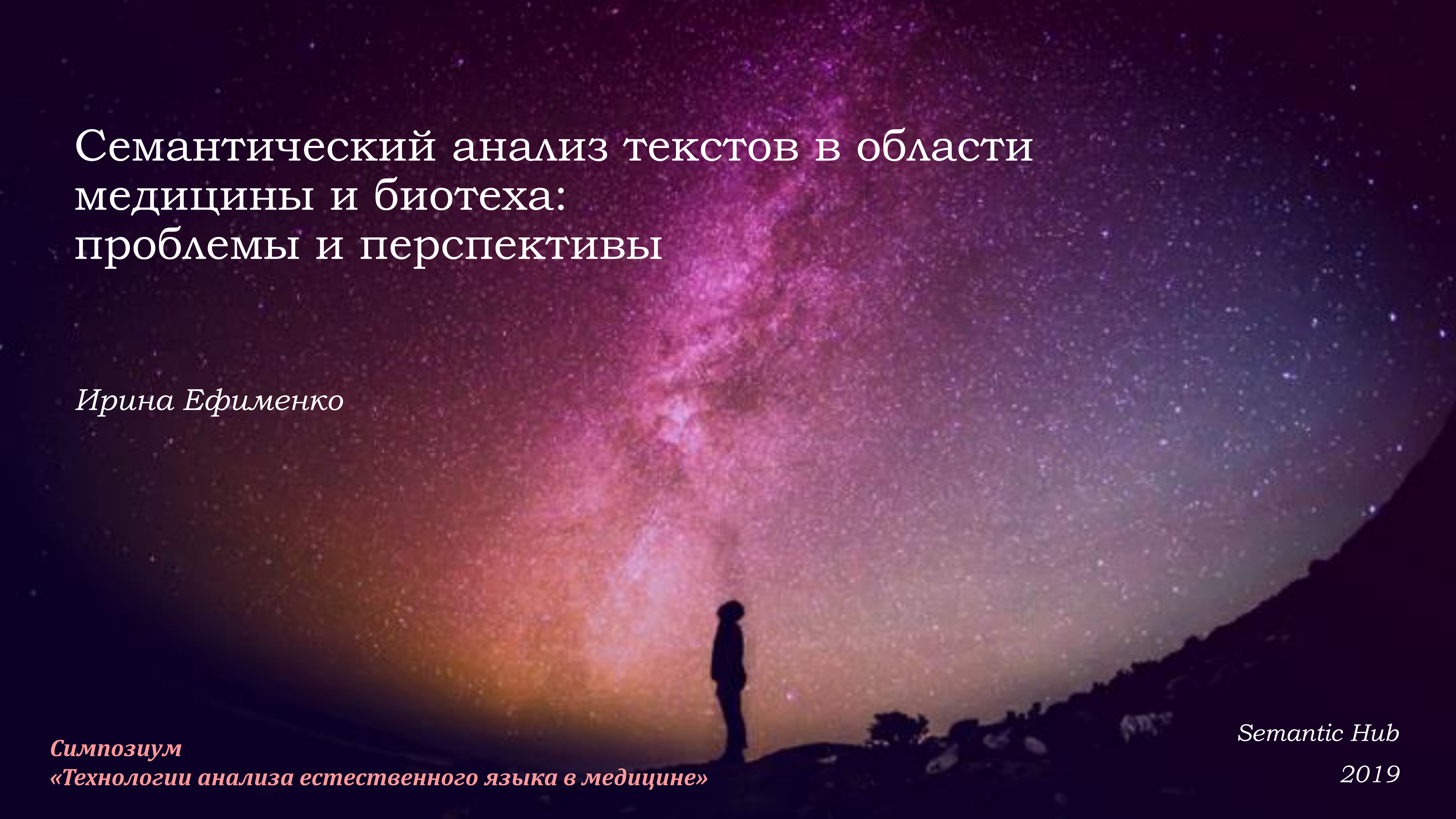


Семантический анализ текстов в области медицины и биотеха: проблемы и перспективы

Ирина Ефименко

Симпозиум
«Технологии анализа естественного языка в медицине»

Semantic Hub

2019

Semantic Hub solutions

Technologies aimed at processing large volumes of unstructured information (**Big text data**) and semantic analysis (**natural language understanding**), which extract knowledge from heterogeneous information sources. The software provides:

- Automated scanning of web resources and databases
- Deep semantic analysis of potentially relevant texts
- Generation of **insights** on all the factors, which are important for success of an innovative drug

Products	Technologic scouting (Scientific due diligence)	Gathering RWE on patients and physicians in social media	Identification of patients for rare diseases
For whom	BD and R&D experts	Experts in Marketing and Medical Affairs	Experts in Medical Affairs and Market Access, BU Heads
Description	Semantic Hub helps select the **most promising assets** out of the universe of candidates as a potential target for investment or as an emerging threat	Semantic Hub provides the "**reality check**" on patient journey and patient archetypes based on **100 000+** real patient stories	Semantic Hub helps **find patients** with rare diseases and build **the landscape** of the patient experience in the country of interest
Details	Millions of documents processed as an input (papers, CT results, patents, news). Assets qualified with a variety of success and risk factors (100+ criteria such as PKPD, toxicity, MoA, etc.)	Knowledge extraction from the millions of user posts in patient forums, professional social networks for physicians, health-related QA portals, etc. Evaluation of the real patient journey	Full-scale screening of the Internet and identification of patients in patient forums, professional social networks, health-related QA portals. Understanding patient needs, journey through the local healthcare system
Advantages	- Search for **red & amber flags** - Assets prioritization - **Easy-to-use** visualization of results	**Unbiased data** about drivers and barriers in choosing therapy, outcomes, unmet needs, awareness and opinions, emotional aspects, mentioned HCP, KOLs, clinical centers, etc.	- Finding patients who are already diagnosed or **potentially** having a disease - The solution can be used to find patients for **Clinical Trials**

What is unique about Semantic Hub:

- Not just data, but the **support** of your decisions
- AI which works in synergy with your experts
- Multilingual analysis for **various countries**
- Easy and inexpensive regular update
- **Adaptive design** of research
- **Compliance and security**

We have **100+ years of total experience in:**

- Semantic technologies
- Natural language processing
- Text mining
- Artificial intelligence
- Technology intelligence
- Decision support systems

We have successfully implemented **50+ projects (decision support systems based on "Big Text Data")** in various industries: healthcare, oil and gas, etc. Since 2016, we have focused on **Pharma as our one and only**

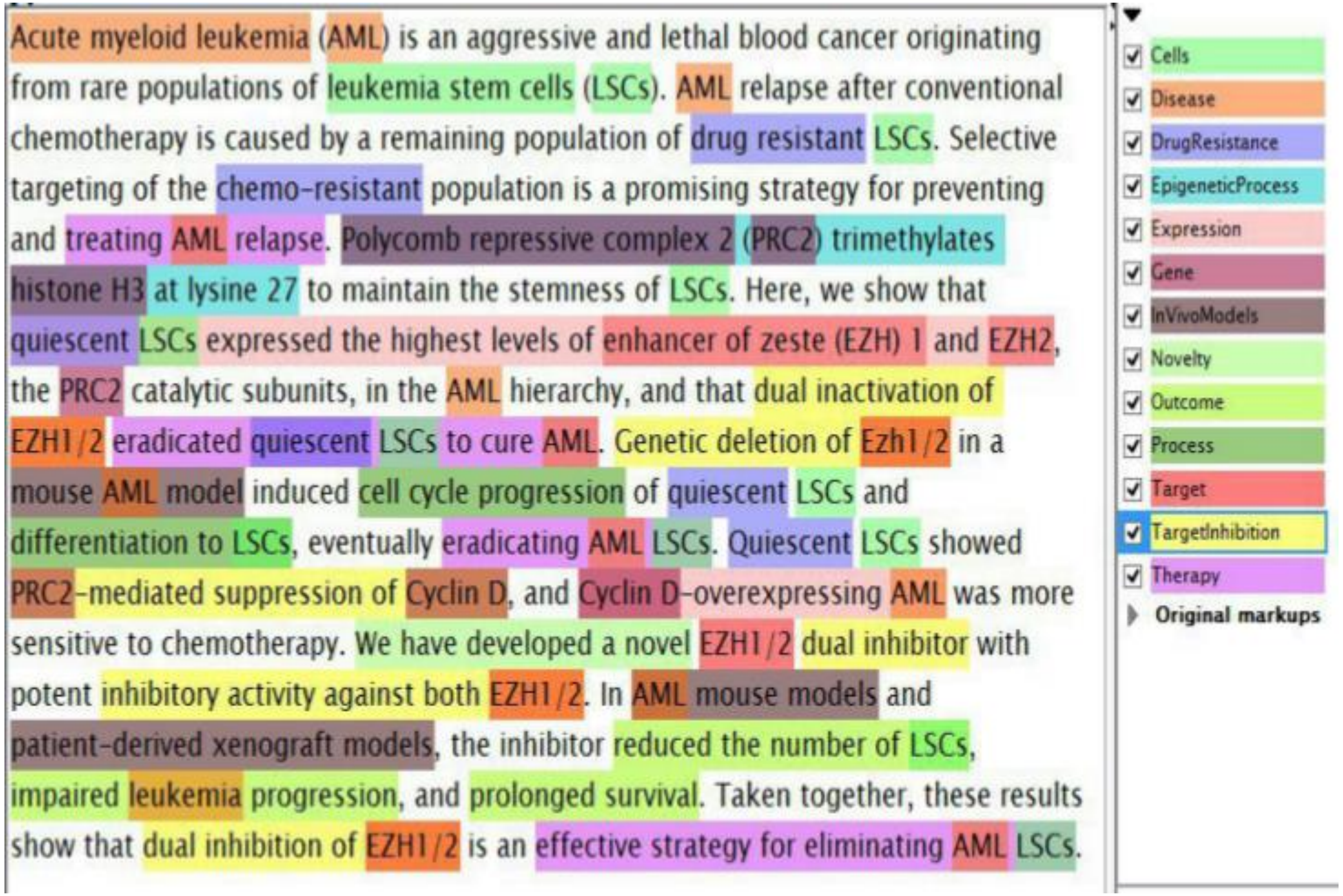
Проблемы…

Естественный язык – это очень сложно…

«Дайте мне мышку,

и я покажу, как умирали крыски»

Биотех и медицина – это очень сложно…



Acute myeloid leukemia (AML) is an aggressive and lethal blood cancer originating from rare populations of leukemia stem cells (LSCs). AML relapse after conventional chemotherapy is caused by a remaining population of drug resistant LSCs. Selective targeting of the chemo-resistant population is a promising strategy for preventing and treating AML relapse. Polycomb repressive complex 2 (PRC2) trimethylates histone H3 at lysine 27 to maintain the stemness of LSCs. Here, we show that quiescent LSCs expressed the highest levels of enhancer of zeste (EZH) 1 and EZH2, the PRC2 catalytic subunits, in the AML hierarchy, and that dual inactivation of EZH1/2 eradicated quiescent LSCs to cure AML. Genetic deletion of Ezh1/2 in a mouse AML model induced cell cycle progression of quiescent LSCs and differentiation to LSCs, eventually eradicating AML LSCs. Quiescent LSCs showed PRC2-mediated suppression of Cyclin D, and Cyclin D-overexpressing AML was more sensitive to chemotherapy. We have developed a novel EZH1/2 dual inhibitor with potent inhibitory activity against both EZH1/2. In AML mouse models and patient-derived xenograft models, the inhibitor reduced the number of LSCs, impaired leukemia progression, and prolonged survival. Taken together, these results show that dual inhibition of EZH1/2 is an effective strategy for eliminating AML LSCs.

☑ Cells
☑ Disease
☑ DrugResistance
☑ EpigeneticProcess
☑ Expression
☑ Gene
☑ InVivoModels
☑ Novelty
☑ Outcome
☑ Process
☑ Target
☑ TargetInhibition
☑ Therapy
▸ Original markups

Semantic Hub

Helping pharma see around the corner

Пациенты – это очень сложно…

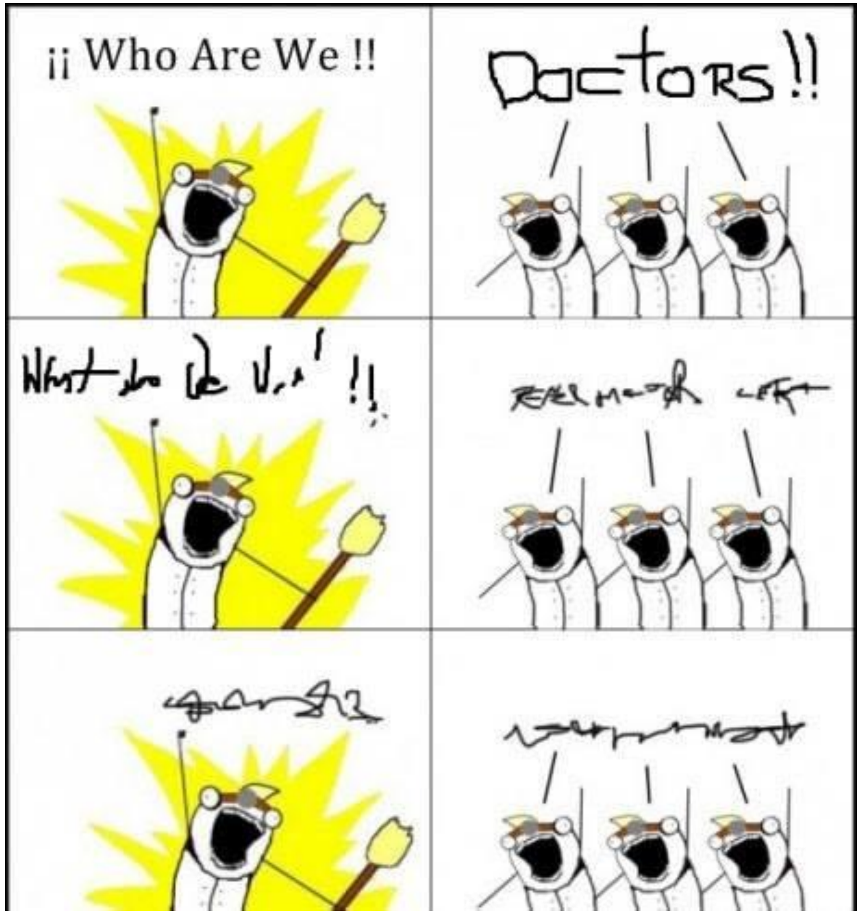
- **«Палится ли она флюшкой?»**
- **«Стаж на аналоге – 10 лет, в итоге парение»**

(И не только они)

https://www.oreilly.com/ideas/lessons-learned-building-natural-language-processing-systems-in-health-care

Triage Notes

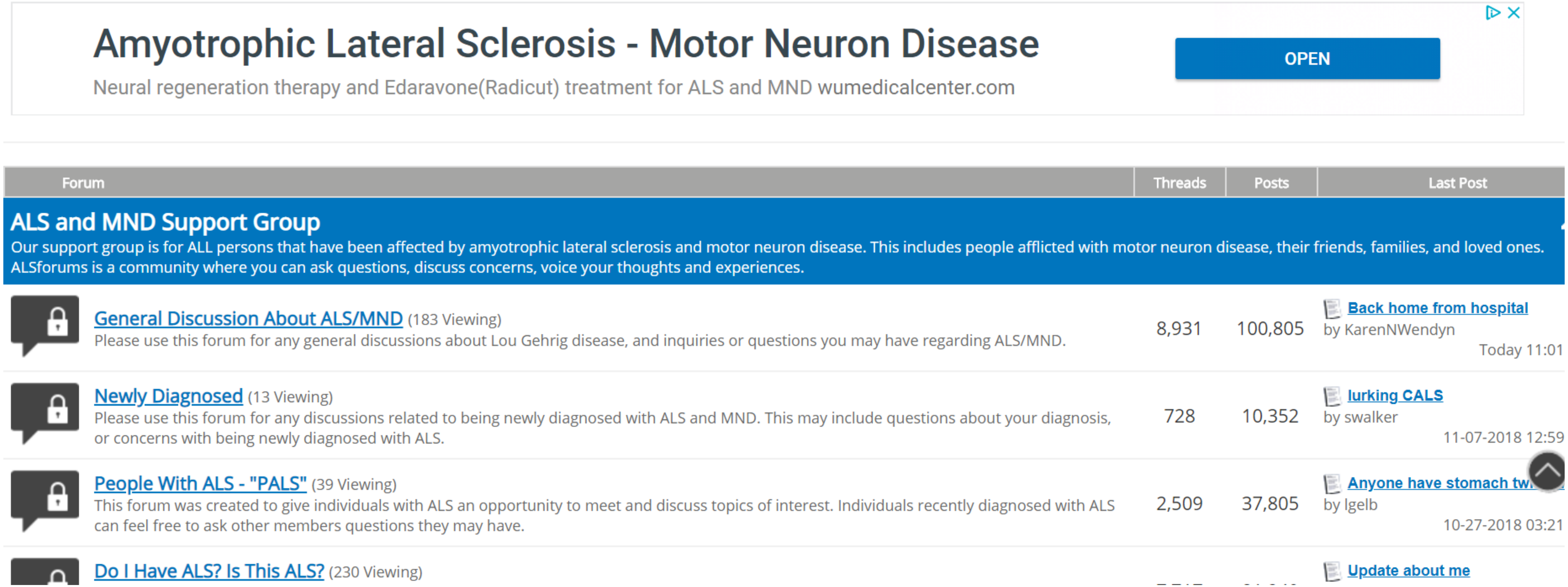
- *“states started last night, upper abd, took alka seltzer approx 0500, no relief. nausea no vomiting*
- *Since yesterday 10/10 "constant Tylenol 1 hr ago. +nausea. diaphoretic. Mid abd radiates to back*
- *Generalized abd radiating to lower x 3 days accompanied by dark stools. Now with bloody stool this am. Denies dizzy, sob, fatigue”*



Semantic Hub

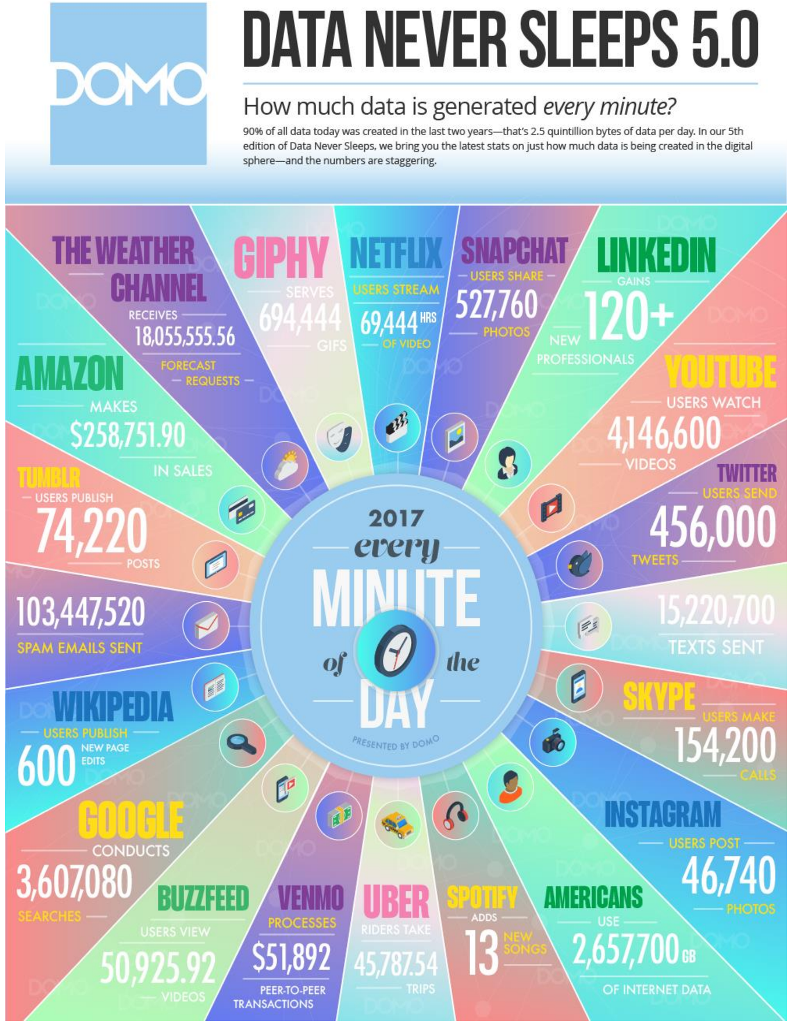
Helping pharma see around the corner

…Но ценно



Forum	Threads	Posts	Last Post
ALS and MND Support Group Our support group is for ALL persons that have been affected by amyotrophic lateral sclerosis and motor neuron disease. This includes people afflicted with motor neuron disease, their friends, families, and loved ones. ALSforums is a community where you can ask questions, discuss concerns, voice your thoughts and experiences.			
General Discussion About ALS/MND (183 Viewing) Please use this forum for any general discussions about Lou Gehrig disease, and inquiries or questions you may have regarding ALS/MND.	8,931	100,805	Back home from hospital by KarenNWendyn Today 11:01
Newly Diagnosed (13 Viewing) Please use this forum for any discussions related to being newly diagnosed with ALS and MND. This may include questions about your diagnosis, or concerns with being newly diagnosed with ALS.	728	10,352	lurking CALS by swalker 11-07-2018 12:59
People With ALS - "PALS" (39 Viewing) This forum was created to give individuals with ALS an opportunity to meet and discuss topics of interest. Individuals recently diagnosed with ALS can feel free to ask other members questions they may have.	2,509	37,805	Anyone have stomach tw by lgelb 10-27-2018 03:21
Do I Have ALS? Is This ALS? (230 Viewing)			Update about me

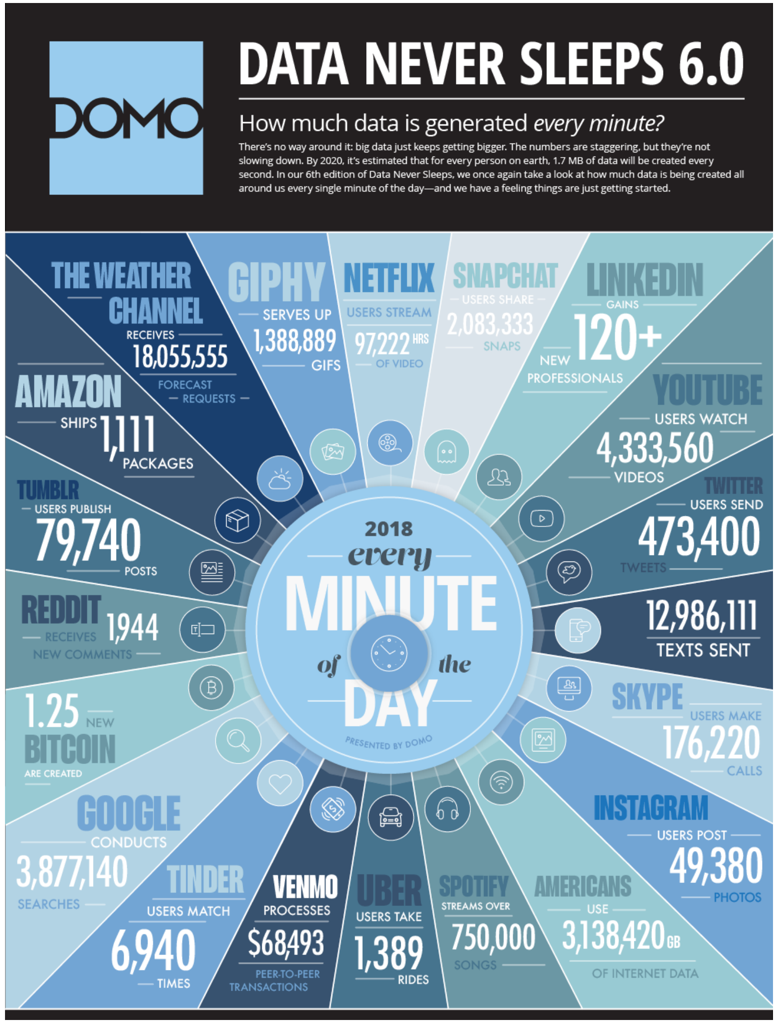
…И неизбежно



DATA NEVER SLEEPS 5.0

How much data is generated *every minute?*

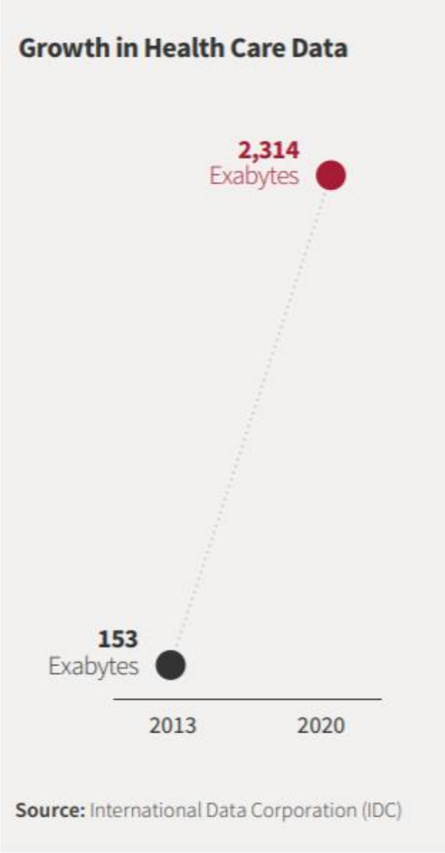
90% of all data today was created in the last two years—that's 2.5 quintillion bytes of data per day. In our 5th edition of Data Never Sleeps, we bring you the latest stats on just how much data is being created in the digital sphere—and the numbers are staggering.



DATA NEVER SLEEPS 6.0

How much data is generated *every minute?*

There's no way around it: big data just keeps getting bigger. The numbers are staggering, but they're not slowing down. By 2020, it's estimated that for every person on earth, 1.7 MB of data will be created every second. In our 6th edition of Data Never Sleeps, we once again take a look at how much data is being created all around us every single minute of the day—and we have a feeling things are just getting started.



Growth in Health Care Data

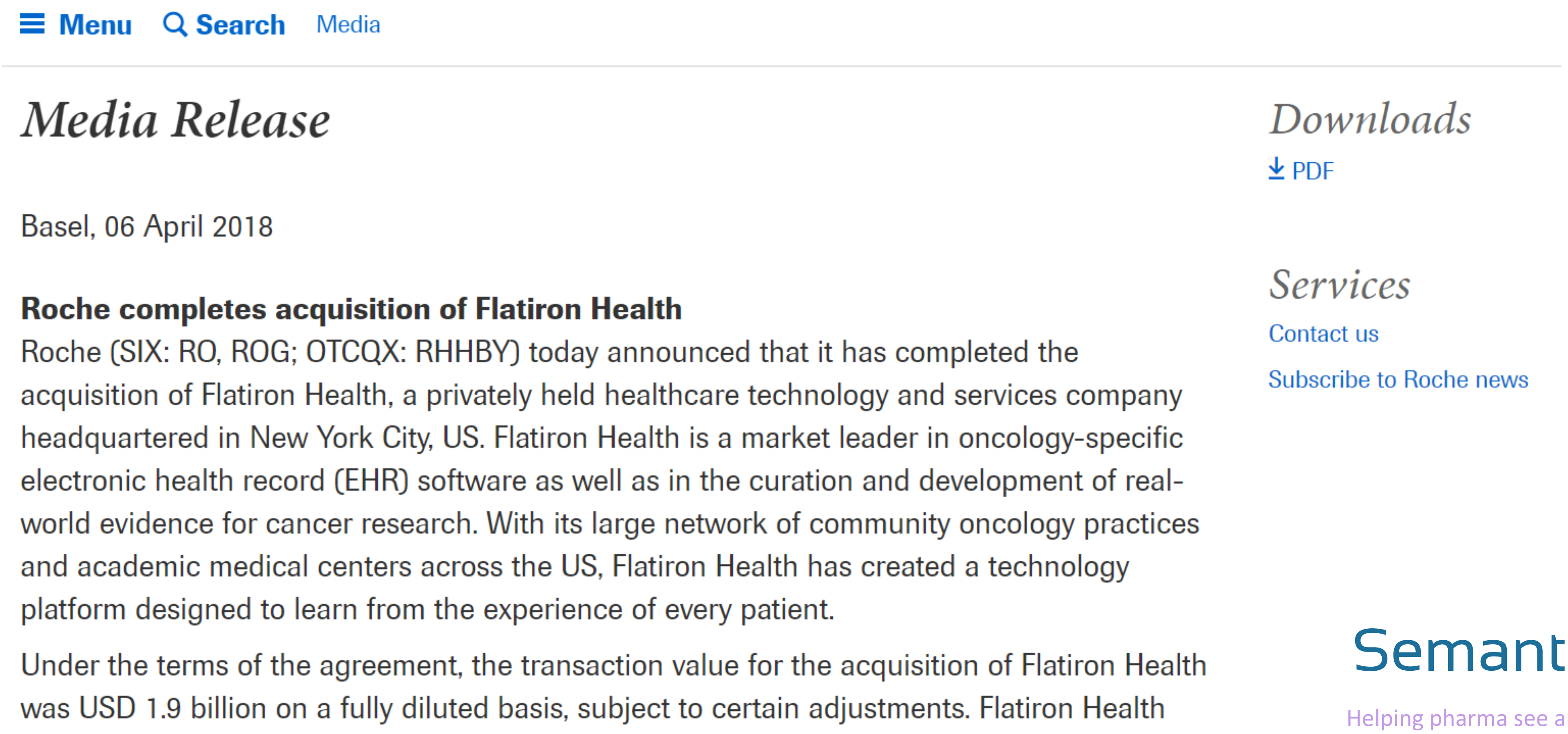
Source: International Data Corporation (IDC)

Semantic Hub

Helping pharma see around the corner

Что-то происходит

According to the US Centers for Disease Control and Prevention, more than 36 million hospital admissions and 1.3 billion ambulatory care visits are documented per year in the USA
(https://www.researchgate.net/publication/256101117_State_of_the_Art_in_Clinical_Informatics_Evidence_and_Examples)



Media Release

Basel, 06 April 2018

Roche completes acquisition of Flatiron Health
Roche (SIX: RO, ROG; OTCQX: RHHBY) today announced that it has completed the acquisition of Flatiron Health, a privately held healthcare technology and services company headquartered in New York City, US. Flatiron Health is a market leader in oncology-specific electronic health record (EHR) software as well as in the curation and development of real-world evidence for cancer research. With its large network of community oncology practices and academic medical centers across the US, Flatiron Health has created a technology platform designed to learn from the experience of every patient.

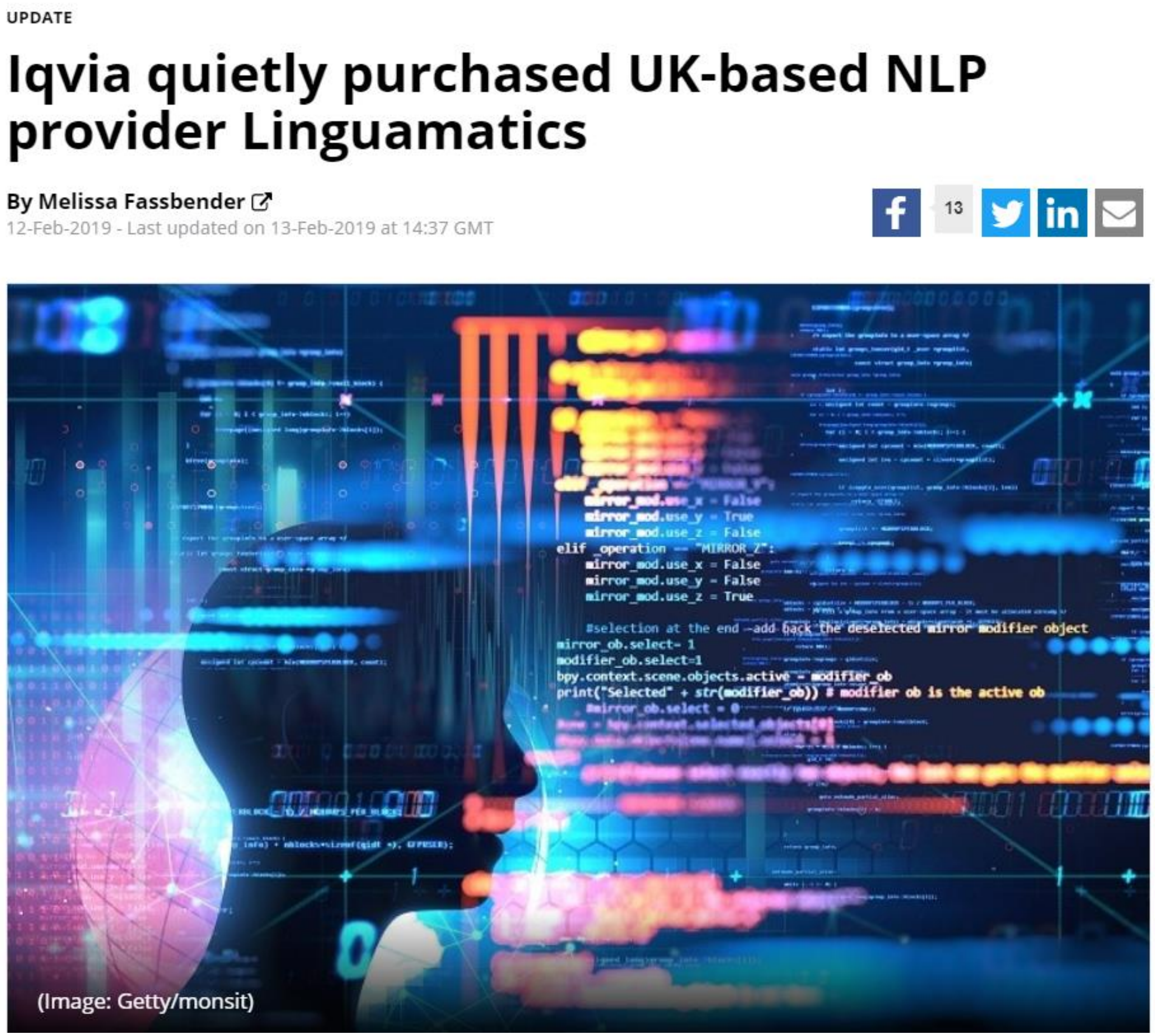
Under the terms of the agreement, the transaction value for the acquisition of Flatiron Health was USD 1.9 billion on a fully diluted basis, subject to certain adjustments. Flatiron Health

Downloads

Services

Semantic Hub
Helping pharma see around the corner

Что-то происходит



UPDATE

Iqvia quietly purchased UK-based NLP provider Linguamatics

By Melissa Fassbender

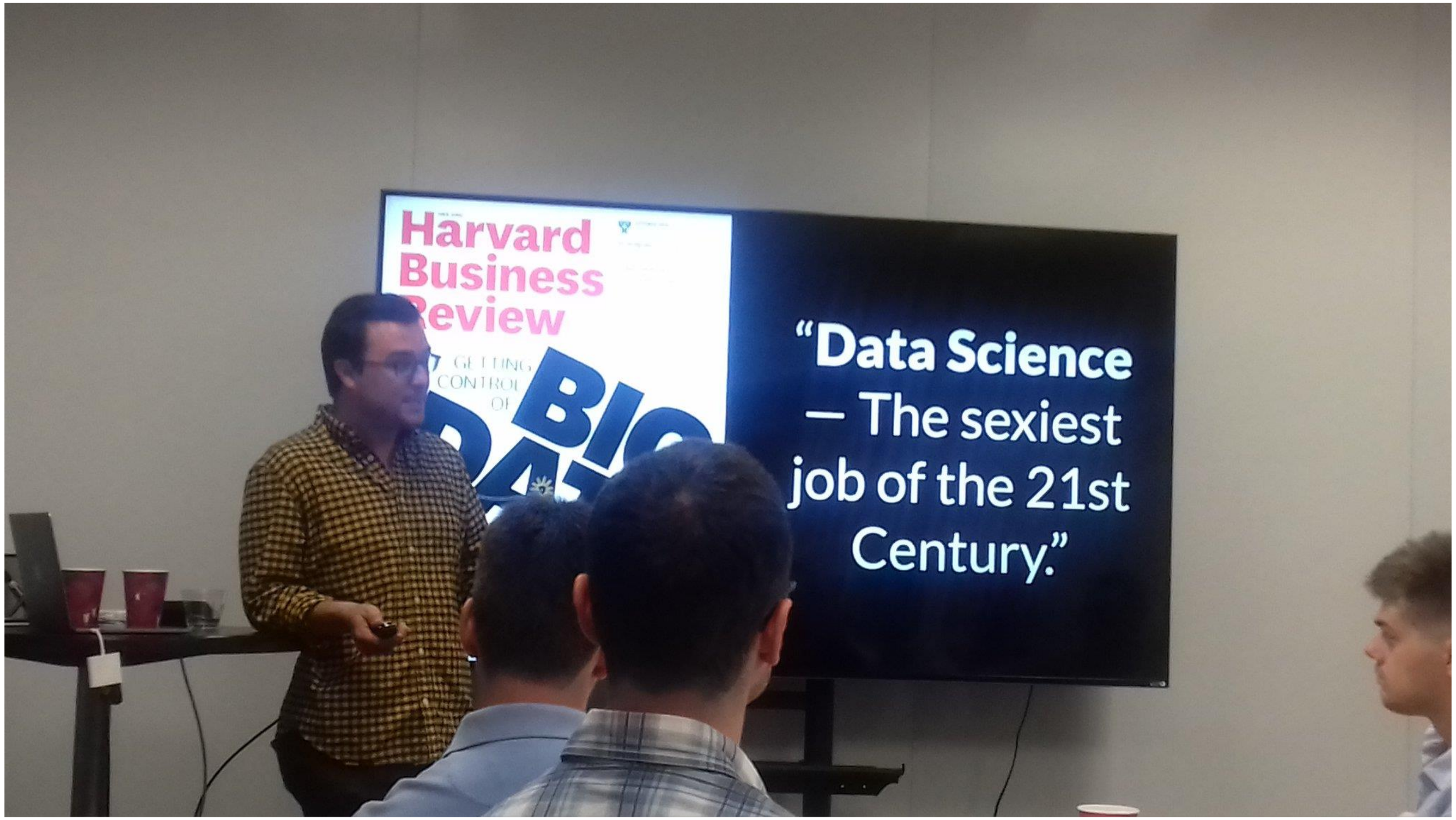
12-Feb-2019 - Last updated on 13-Feb-2019 at 14:37 GMT

(Image: Getty/monsit)

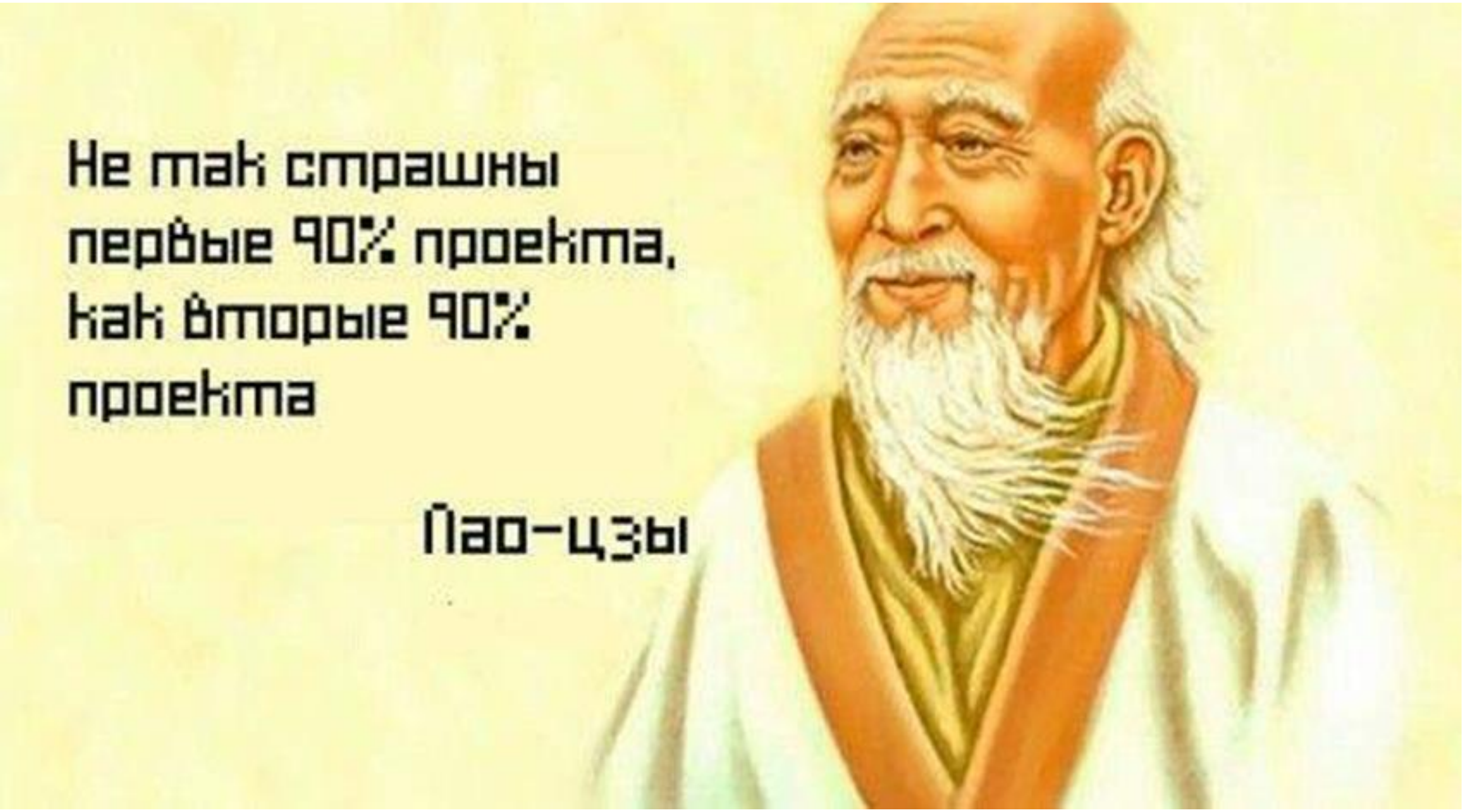
Semantic Hub

Helping pharma see around the corner

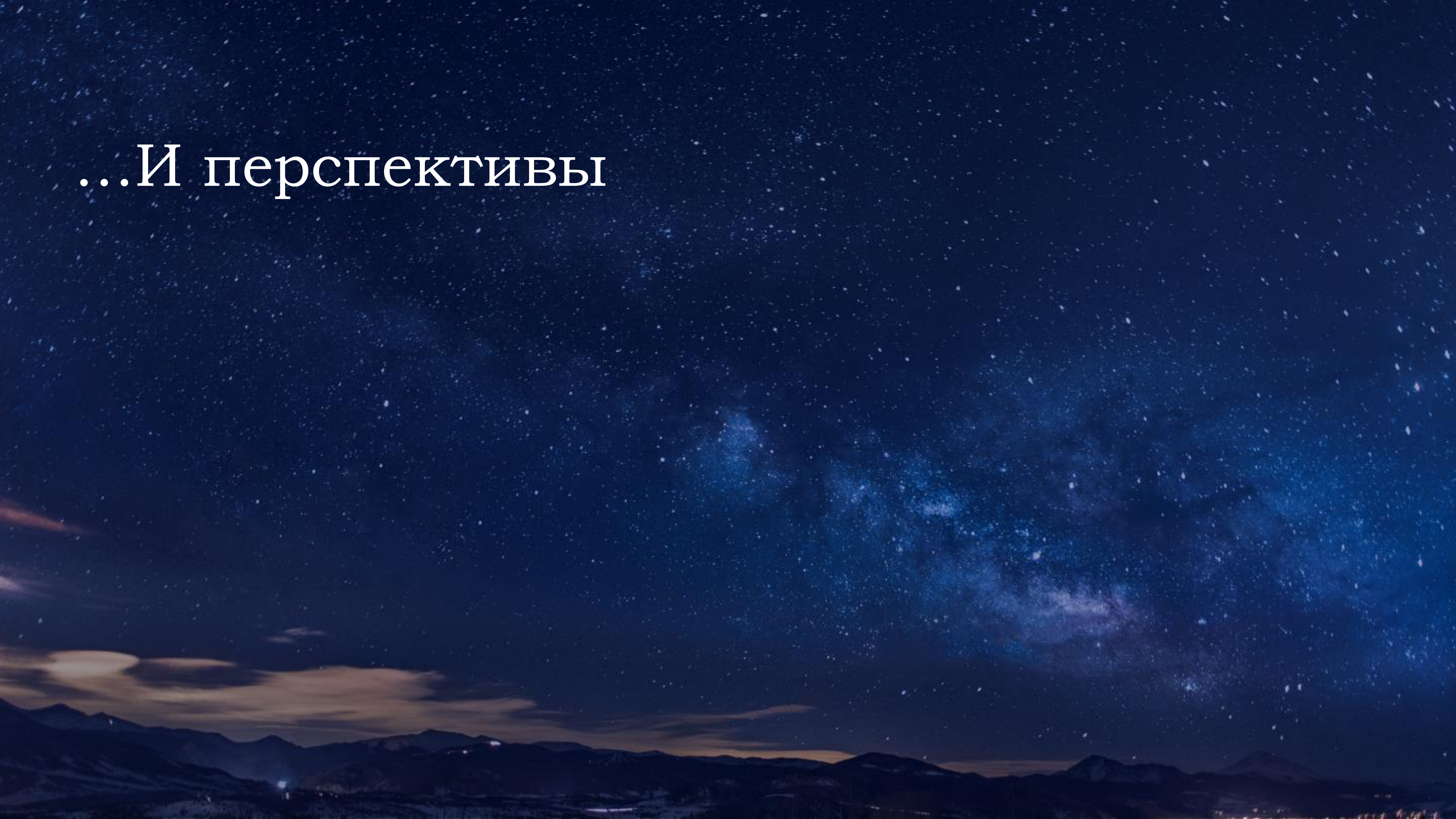
Что-то происходит



Что-то происходит



…И перспективы



Методы и подходы к анализу текстов в медицине

2 основных источника методов и инструментов:

- Biomedical natural language processing (BioNLP) methods – включая специализированные соревнования
- Методы анализа социальных сетей и user-generated content

Основополагающие подходы:

- Rule-based (pattern-based) approaches
- *Linguistic* ontologies

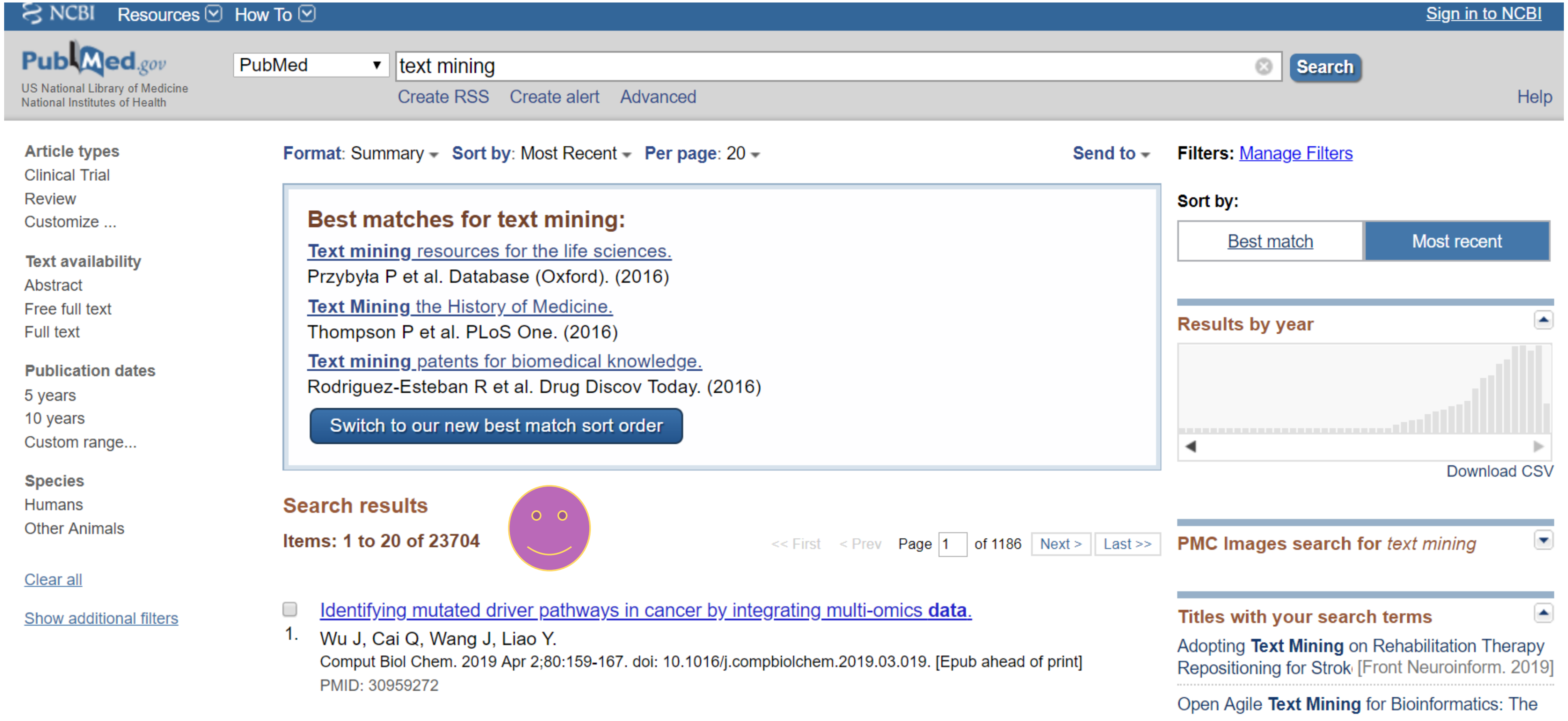
(semantic resources such as ontologies and controlled vocabularies)

- Statistical approaches

Пути развития:

- Shallow vs. deep approach
- Hybrid approach
- Ontology-driven approach (*domain* ontology)

Методы и подходы к анализу текстов в медицине



Best matches for text mining:

Text mining resources for the life sciences.
Przybyła P et al. Database (Oxford). (2016)

Text Mining the History of Medicine.
Thompson P et al. PLoS One. (2016)

Text mining patents for biomedical knowledge.
Rodriguez-Esteban R et al. Drug Discov Today. (2016)

Search results

Items: 1 to 20 of 23704

1. Identifying mutated driver pathways in cancer by integrating multi-omics data.
Wu J, Cai Q, Wang J, Liao Y.
Comput Biol Chem. 2019 Apr 2;80:159-167. doi: 10.1016/j.compbiolchem.2019.03.019. [Epub ahead of print]
PMID: 30959272

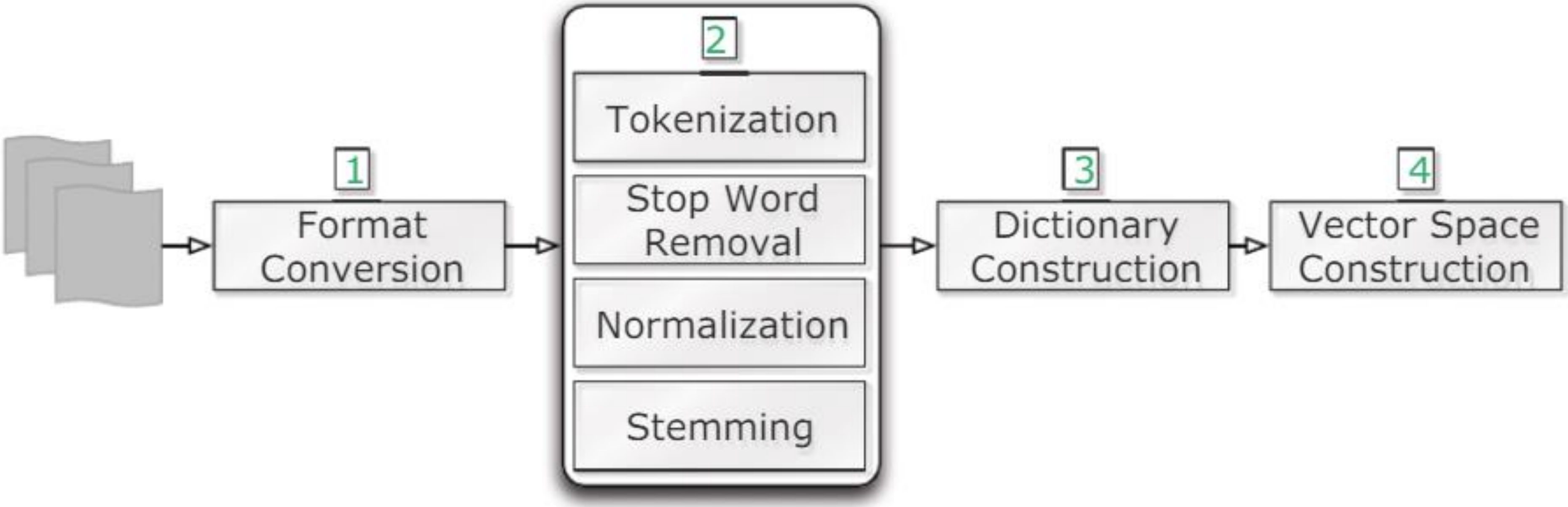
Adopting **Text Mining** on Rehabilitation Therapy Repositioning for Strok [Front Neuroinform. 2019]

Open Agile **Text Mining** for Bioinformatics: The

Твиттер, описания медицинских изображений и парсинг имен авторов научных статей…
Не всякий text mining – семантический?

Semantic Hub
Helping pharma see around the corner

Методы и подходы к анализу текстов в медицине



https://www.researchgate.net/publication/263292911_Biomedical_Text_Mining_State-of-the-Art_Open_Problems_and_Future_Challenges

Semantic Hub

Helping pharma see around the corner

Методы и подходы к анализу текстов в медицине

Классика:

Vector Space Model, tf-idf, cosine similarity и другие схемы

Примеры в медицине:

- content based image retrieval, text summarization of medical articles

Достоинства:

- The model is simple and clear

Ограничения:

- Each document is seen as a bag of words, words are considered to be statistically independent. The meaning of the word sequence is not reflected in the model
- Assumption a single term represents exactly one word sense, which is not true for natural language texts, which contain synonymous and polysemous words. Methods like word sense disambiguation have to applied in the pre-processing step

Развитие:

- The Semantic Vector Space Model (SVSM) which is a text representation and searching technique based on the combination of Vector Space Model (VSM) with heuristic syntax parsing and distributed representation of semantic case structures

Semantic Hub

Helping pharma see around the corner

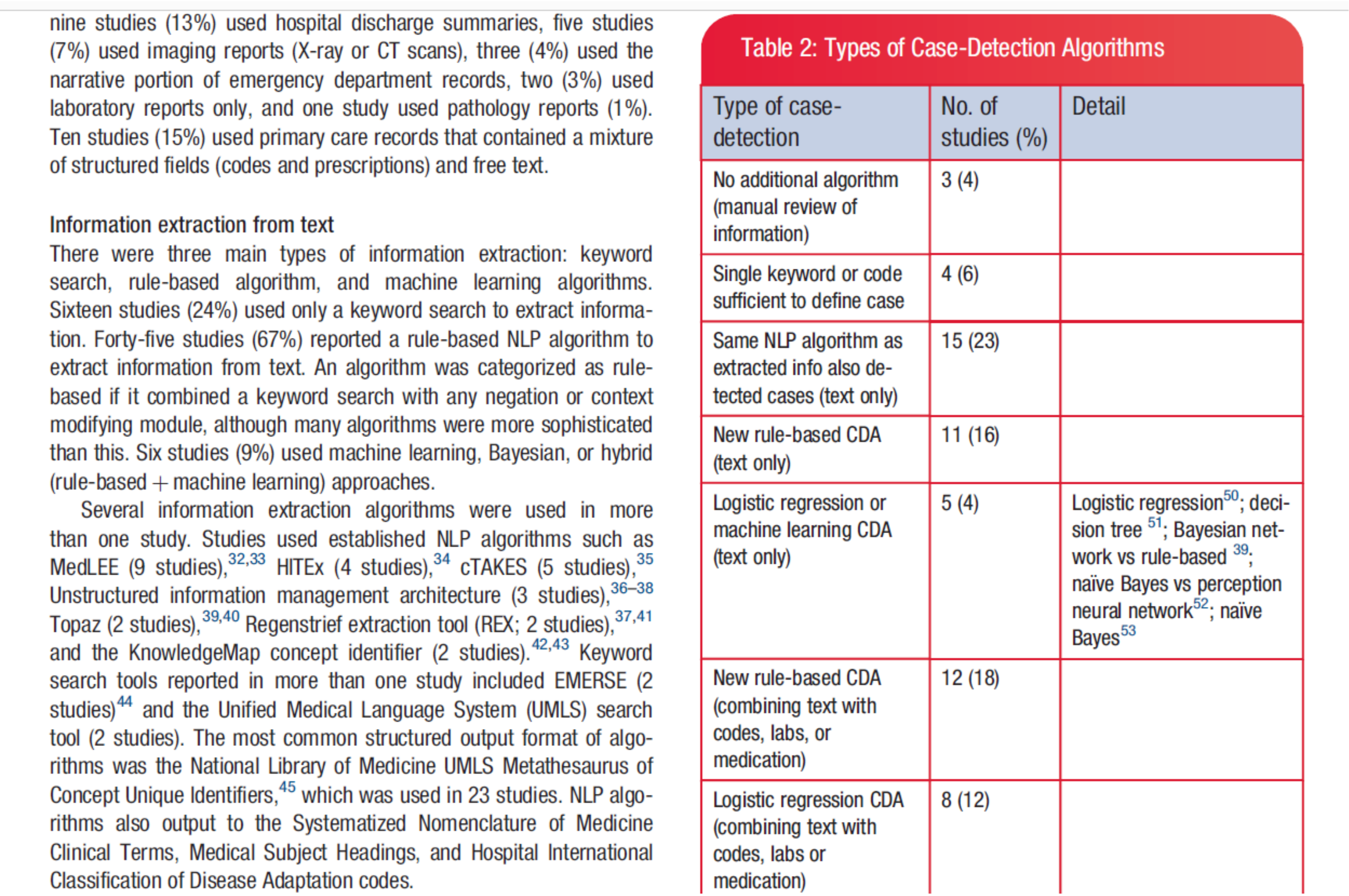
Методы и подходы к анализу текстов в медицине

Statistical approaches:

- Latent Semantic Analysis (LSA)
- Probabilistic latent semantic analysis (PLSA)
- Latent Dirichlet allocation (LDA)
- Hierarchical Latent Dirichlet Allocation (hLDA)
- Semantic Vector Space Model (SVSM)
- Latent semantic mapping (LSM)
- Principal component analysis (PCA)

https://www.researchgate.net/publication/263292911_Biomedical_Text_Mining_State-of-the-Art_Open_Problems_and_Future_Challenges

Методы и подходы к анализу текстов в медицине



nine studies (13%) used hospital discharge summaries, five studies (7%) used imaging reports (X-ray or CT scans), three (4%) used the narrative portion of emergency department records, two (3%) used laboratory reports only, and one study used pathology reports (1%). Ten studies (15%) used primary care records that contained a mixture of structured fields (codes and prescriptions) and free text.

Information extraction from text

There were three main types of information extraction: keyword search, rule-based algorithm, and machine learning algorithms. Sixteen studies (24%) used only a keyword search to extract information. Forty-five studies (67%) reported a rule-based NLP algorithm to extract information from text. An algorithm was categorized as rule-based if it combined a keyword search with any negation or context modifying module, although many algorithms were more sophisticated than this. Six studies (9%) used machine learning, Bayesian, or hybrid (rule-based + machine learning) approaches.

Several information extraction algorithms were used in more than one study. Studies used established NLP algorithms such as MedLEE (9 studies),[32,33] HITEx (4 studies),[34] cTAKES (5 studies),[35] Unstructured information management architecture (3 studies),[36–38] Topaz (2 studies),[39,40] Regenstrief extraction tool (REX; 2 studies),[37,41] and the KnowledgeMap concept identifier (2 studies).[42,43] Keyword search tools reported in more than one study included EMERSE (2 studies)[44] and the Unified Medical Language System (UMLS) search tool (2 studies). The most common structured output format of algorithms was the National Library of Medicine UMLS Metathesaurus of Concept Unique Identifiers,[45] which was used in 23 studies. NLP algorithms also output to the Systematized Nomenclature of Medicine Clinical Terms, Medical Subject Headings, and Hospital International Classification of Disease Adaptation codes.

Table 2: Types of Case-Detection Algorithms

Type of case-detection	No. of studies (%)	Detail
No additional algorithm (manual review of information)	3 (4)	
Single keyword or code sufficient to define case	4 (6)	
Same NLP algorithm as extracted info also detected cases (text only)	15 (23)	
New rule-based CDA (text only)	11 (16)	
Logistic regression or machine learning CDA (text only)	5 (4)	Logistic regression[50]; decision tree [51]; Bayesian network vs rule-based [39]; naïve Bayes vs perception neural network[52]; naïve Bayes[53]
New rule-based CDA (combining text with codes, labs, or medication)	12 (18)	
Logistic regression CDA (combining text with codes, labs or medication)	8 (12)	

Extracting information from text of electronic medical records to improve case detection: a systematic review. Elizabeth Ford et al., 2016

Semantic Hub

Helping pharma see around the corner

Возвращаясь к Emergency rooms (Lessons learned building natural language processing systems in healthcare)

As a philosopher or linguist, you might argue that this still does not constitute a “different language” in the typical sense of the word. However, if you’re a data scientist or NLP practitioner, there shouldn’t be any doubt that it is:

- **It has a different vocabulary**. The Unified Medical Language System (UMLS) includes more than 200 vocabularies for English alone, covering more than three million terms. In contrast, the Oxford English Dictionary of 1989 had 171,476 words (although, that should be roughly tripled to include derivatives that UMLS directly lists)
- **It has a different grammar**. The text has its own definition of what sentences are and what parts of speech are. Statements like “+nausea” and “since yesterday 10/10” are grammatical structures that don’t exist anywhere else
- **It has different semantics**. “Sob” means “shortness of breath” (and not the other meaning you had in mind). “Denies” means the patient says they don’t have the symptom, although the clinician thinks they might
- **It goes beyond jargon**. Jargon refers to the 100-200 new words you learn in the first month after you join a new school or workplace. In contrast, understanding health care language takes people as long as it takes to master day-to-day Italian or Portuguese

https://www.oreilly.com/ideas/lessons-learned-building-natural-language-processing-systems-in-health-care

Semantic Hub
Helping pharma see around the corner

Возвращаясь к Emergency rooms (Lessons learned building natural language processing systems in healthcare)

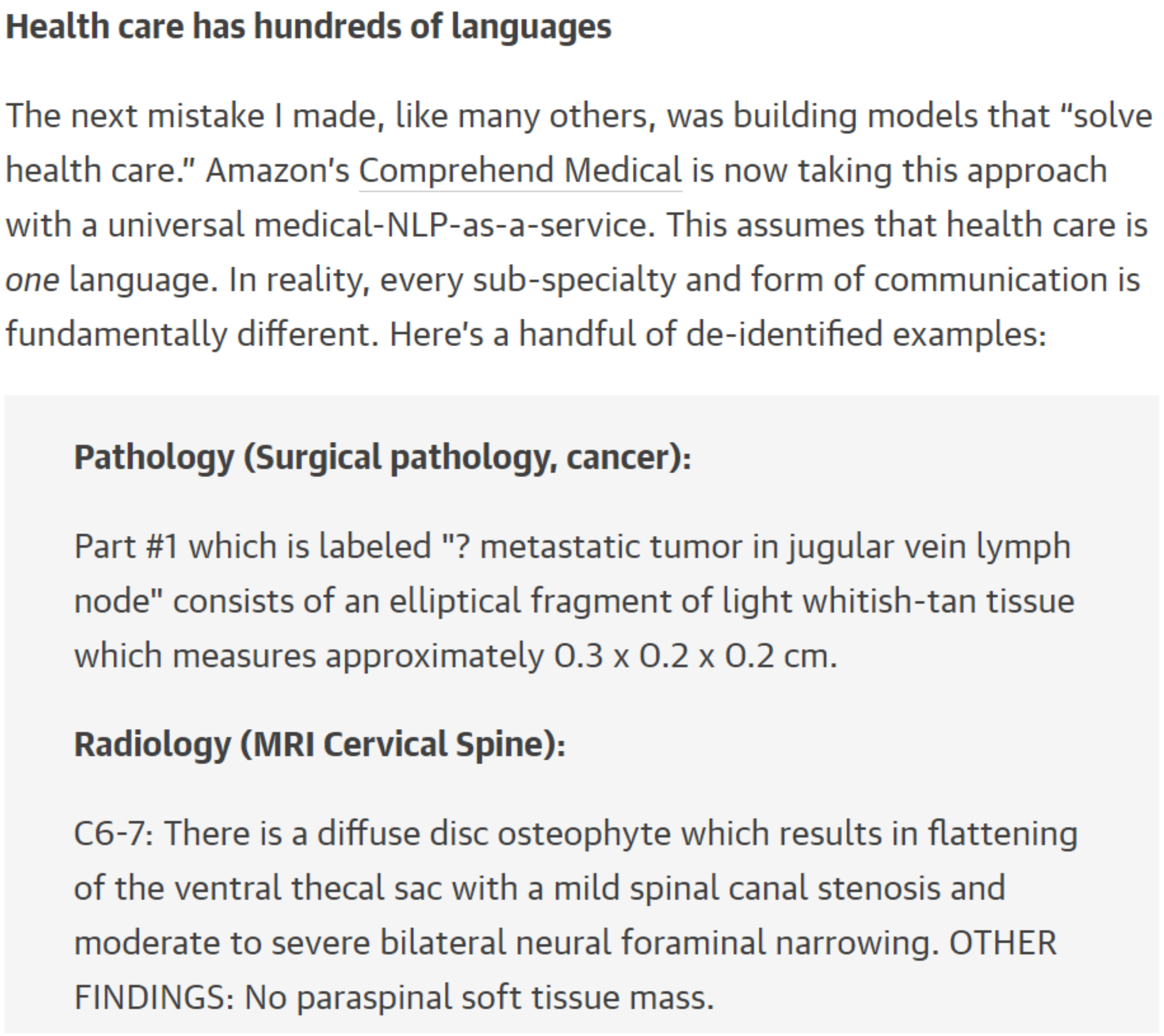
- **Lesson #1: Off-the-shelf NLP models don’t work** (не коммодити!) Not only will named entity recognition or entity resolution models fail, but even basic tasks such as tokenization, part of speech tagging, and sentence segmentation will fail for the majority of sentences (**впрочем, это верно для реальных текстов на ЕЯ в целом**)
    - *Google Cloud Natural Language*
    - *IBM Watson NLU*
    - *Azure Text Analytics*
    - *spaCy Named Entity Visualizer*
    - *Amazon Comprehend (offline)*
    - *Stanford Core NLP*

In a test done during December 2018, of the six engines, the only medical term (which only two of them recognized) was Tylenol as a product

Semantic Hub

Helping pharma see around the corner

Возвращаясь к Emergency rooms (Lessons learned building natural language processing systems in healthcare)



Health care has hundreds of languages

The next mistake I made, like many others, was building models that "solve health care." Amazon's Comprehend Medical is now taking this approach with a universal medical-NLP-as-a-service. This assumes that health care is *one* language. In reality, every sub-specialty and form of communication is fundamentally different. Here's a handful of de-identified examples:

Pathology (Surgical pathology, cancer):

Part #1 which is labeled "? metastatic tumor in jugular vein lymph node" consists of an elliptical fragment of light whitish-tan tissue which measures approximately 0.3 x 0.2 x 0.2 cm.

Radiology (MRI Cervical Spine):

C6-7: There is a diffuse disc osteophyte which results in flattening of the ventral thecal sac with a mild spinal canal stenosis and moderate to severe bilateral neural foraminal narrowing. OTHER FINDINGS: No paraspinal soft tissue mass.

Semantic Hub

Helping pharma see around the corner

Возвращаясь к Emergency rooms (Lessons learned building natural language processing systems in healthcare)

- **Lesson #3: Start with labeling ground truth** (начни с команды клиницистов…, которые должны еще договориться между собой)

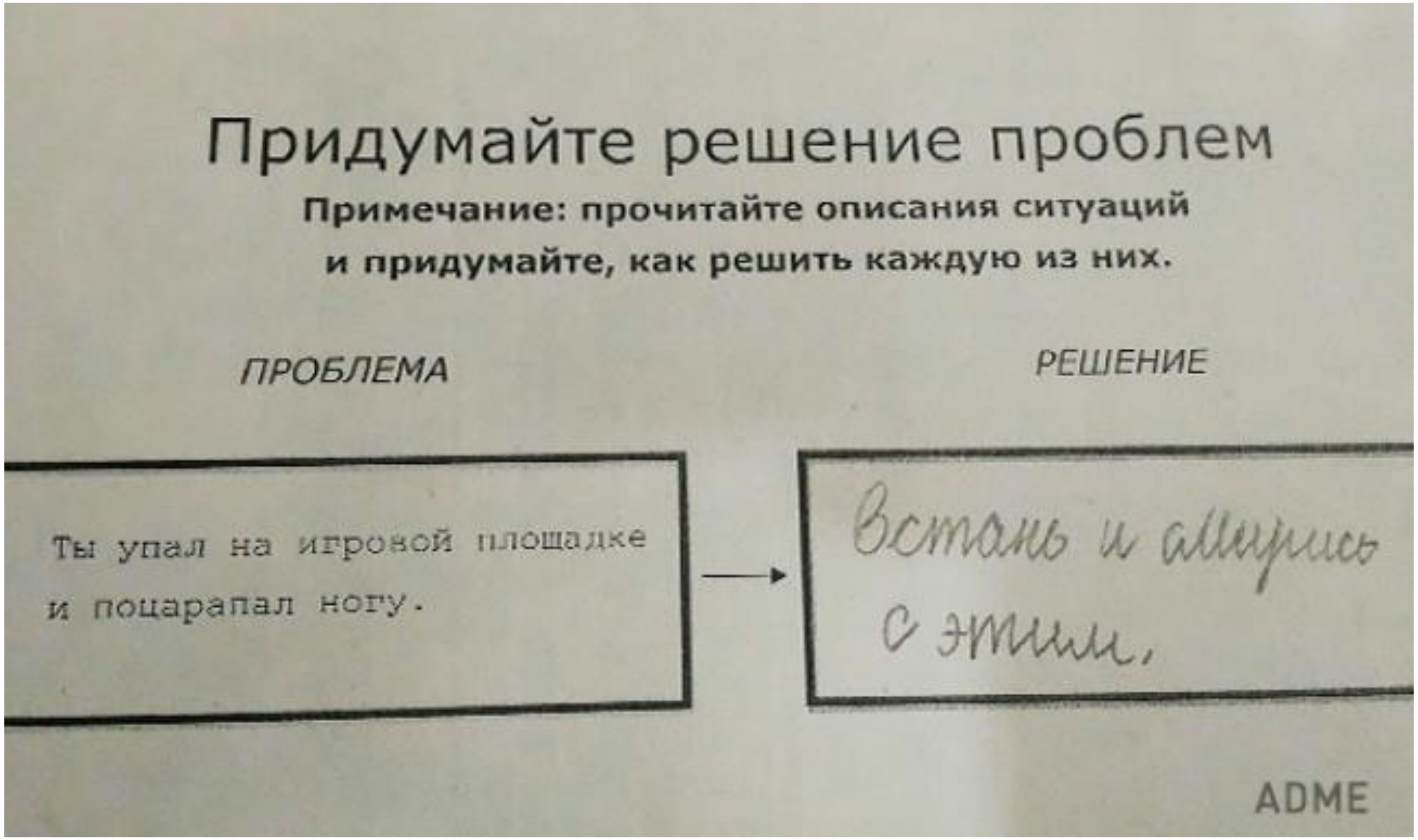
- **Lesson #2: Build trainable NLP pipelines**

Возвращаясь к Emergency rooms (Lessons learned building natural language processing systems in healthcare)

- **Lesson #3: Start with labeling ground truth** (начни с команды клиницистов…, которые должны еще договориться между собой)

- **Lesson #2: Build trainable NLP pipelines**

Методы и подходы к анализу текстов в медицине



Придумайте решение проблем

Примечание: прочитайте описания ситуаций и придумайте, как решить каждую из них.

ПРОБЛЕМА		*РЕШЕНИЕ*
Ты упал на игровой площадке и поцарапал ногу.	→	Встань и смирись с этим.

Методы и подходы к анализу текстов в медицине

Пути развития

- Shallow vs. deep approach
- Hybrid approach
- Ontology-driven approach (*domain* ontology)

Нет серебряной пули, но есть killer apps!

Semantic Hub

Helping pharma see around the corner

Спасибо за внимание!

ie@semantic-hub.com
+7 916 101 4840
www.semantic-hub.com