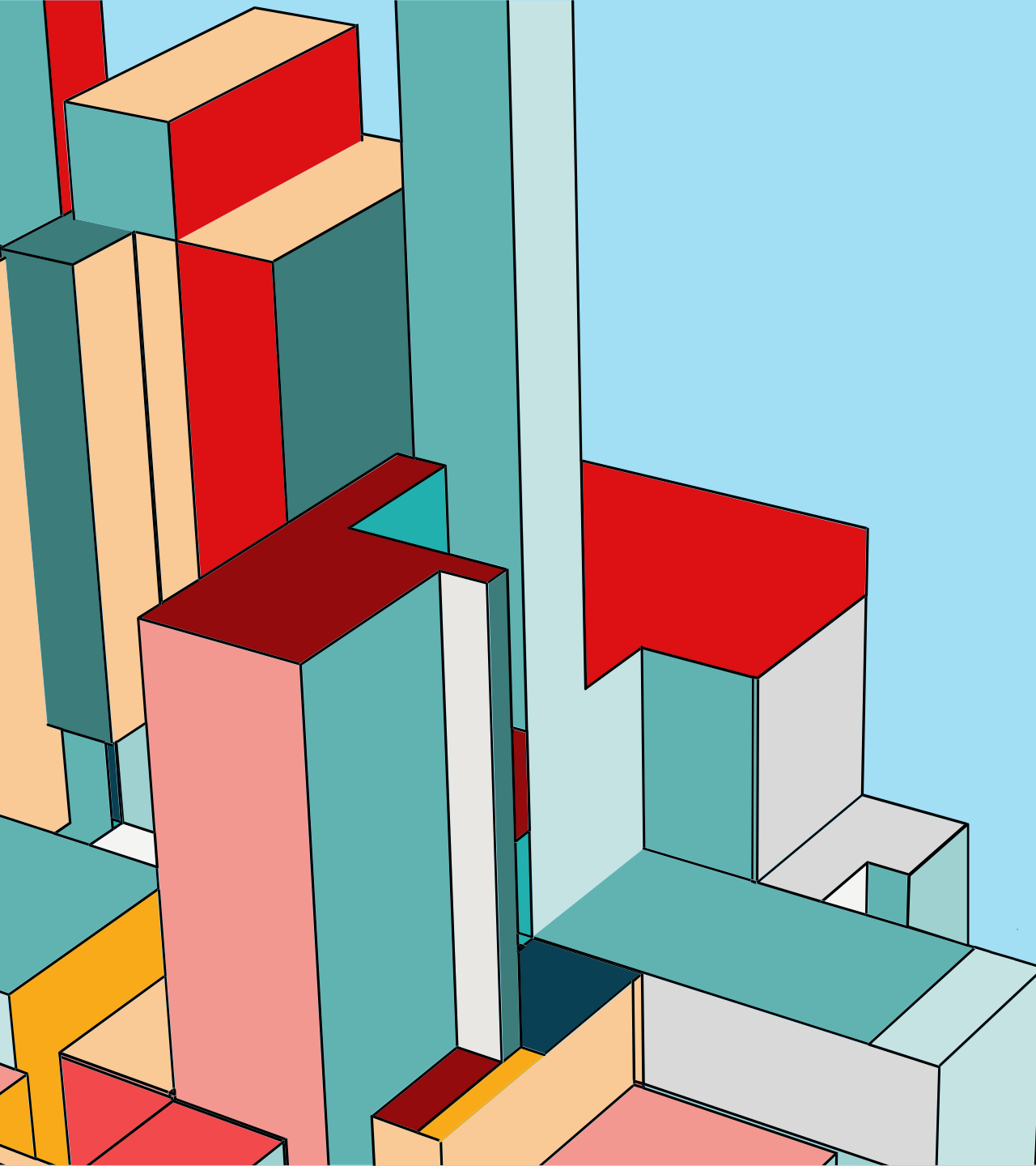




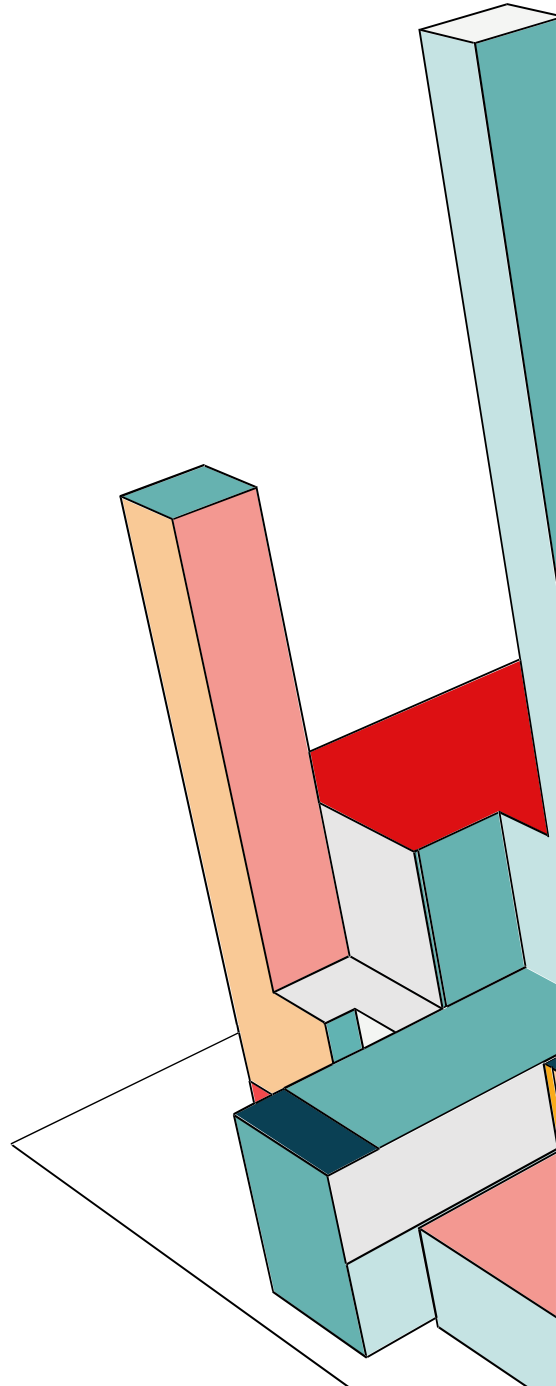
ВІДКРИТА НАУКА І ШТУЧНИЙ ІНТЕЛЕКТ ЯК ФАКТОРИ ТРАНСФОРМАЦІЇ НАУКОВО-ТЕХНІЧНИХ БІБЛІОТЕК

- Вадим Тульчинський*
- Марина Маслова
- Петро Тульчинський

*Інститут кібернетики імені
В.М. Глушкова НАН України*

ПЛАН

- Бібліотека і життя
- Пол Гінспарґ – arXiv.org
- Олександра Елбакян – Sci-Hub
- Георгій Перельман – громадянська наука
- Ілля Суцкевер – ChatGPT
- Що буде, що робити



БІБЛІОТЕКА І ЖИТТЯ

Прийнято 23 травня

Оприлюднено 15 червня

У мене – 3 серпня

nature medicine

Resource

<https://doi.org/10.1038/s41591-025-03790-9>

Deep phenotyping of health–disease continuum in the Human Phenotype Project

Received: 1 August 2024

Accepted: 23 May 2025

Published online: 15 July 2025

 Check for updates

A list of authors and their affiliations appears at the end of the paper

The Human Phenotype Project (HPP) is a large-scale deep-phenotype prospective cohort. To date, approximately 28,000 participants have enrolled, with more than 13,000 completing their initial visit. The project is aimed at identifying novel molecular signatures with diagnostic, prognostic and therapeutic value, and at developing artificial intelligence (AI)-based predictive models for disease onset and progression. The HPP includes longitudinal profiling encompassing medical history, lifestyle and nutrition, anthropometrics, blood tests, continuous glucose and sleep monitoring, imaging and multi-omics data, including genomics, transcriptomics, microbiome (gut, vaginal and oral), metabolomics and immune profiling. Analysis of these data highlights the variation of phenotypes with age and ethnicity and unravels molecular signatures of disease by comparison with matched healthy controls. Leveraging extensive dietary and lifestyle data, we identify associations between lifestyle factors and health outcomes. Finally, we present a multi-modal foundation AI model, trained using self-supervised learning on diet and continuous-glucose-monitoring data, that outperforms existing methods in predicting disease onset. This framework can be extended to integrate other modalities and act as a personalized digital twin. In summary, we present a deeply phenotyped cohort that serves as a platform for advancing biomarker discovery, enabling the development of multi-modal AI models and personalized medicine approaches.

In the landscape of modern medicine, the multifactorial nature of complex diseases demands an understanding that goes beyond mere cross-sectional snapshots of health conditions. It is crucial to develop large-scale prospective cohorts and biobanks that monitor health outcomes over extended periods¹. These large-scale cohorts have become essential resources^{2–4}, providing dynamic insights into how genetic, environmental and lifestyle factors⁵ interact to influence health outcomes, disease risk and progression.

Creating a large-scale longitudinal cohort with sufficient phenotypic depth is inherently challenging; strategic decisions are crucial to navigate the trade-offs imposed by limited research resources⁶. Although data gathered in electronic health records (EHRs) can provide valuable clinical insights at scale, they often lack the granularity

Alternatively, longitudinal deep phenotyping measurements, although more resource-intensive and, in most cases, resulting in fewer participants, can achieve a more detailed view of biological and molecular dynamics across multiple biological scales, enhancing our ability to identify potential new biomarkers or disease pathways. This method is especially valuable in the study of complex diseases or subtle health variations, for which insightful discoveries can often emerge from strong signals detected in smaller, deep-phenotyped cohorts.

Here we introduce the HPP, an in-depth, international, dynamic, prospective longitudinal study and biobank aiming to deeply characterize a large number of individuals, starting in Israel and subsequently expanding to other regions in the world, including the United Arab Emirates (UAE) and Japan. By merging diverse data collection with

to explore the variations in disease susceptibility, clinical phenotypes and aging patterns, ultimately contributing to the advancement of precision medicine.

Results

The HPP cohort

Established in 2018, the HPP originally aimed to recruit 10,000 individuals and was thus named the 10K cohort. The characteristics of the original cohort and the inclusion and exclusion criteria have been described previously⁷. Since the study's launch, 27,916 participants have signed up; as of December 2024, 13,072 in Israel had completed their first visit. Selected characteristics of participants are provided in Extended Data Table 1.

The data-collection process of the HPP cohort is presented in Fig. 1a. Volunteers enroll themselves on the trial website (<https://www.project10k.org.il/en>), on which they are shown consent modules and given a screening survey. After giving consent, participants receive baseline health surveys and are invited to the research center for a detailed health evaluation (Supplementary Table 1), which includes anthropometric measurements, imaging studies, funduscopy, cardiovascular function assessments, grip-strength measurement and voice recording. Following this visit, participants provide a stool sample through an at-home collection kit and are equipped with sensors for continuous health monitoring, including a home sleep-apnea test device for three nights and a continuous glucose monitoring (CGM) device for 2 weeks. During this period, participants log their diet, exercise and medication intake in a designated app. Complete blood count tests are conducted, and blood samples are preserved for future research. Extensive multi-omics profiling is performed, encompassing genomics, transcriptomics, proteomics, metabolomics and profiling of the gut, oral and vaginal (if applicable) microbiomes, and immune-system analysis (Extended Data Table 2). Future plans include incorporating additional tests such as wearable devices (for example, Fitbit or Samsung Watch), hormone assays and an expanded multi-omics dataset encompassing single-cell RNA sequencing (scRNA-seq), whole-genome sequencing and a broader metabolomics portfolio.

A 25-year longitudinal follow-up is planned, with annual questionnaires and bi-annual visits at the research center. The HPP integrates longitudinal data from EHRs, including medical diagnoses and laboratory test results, providing a temporal view of health, further enhanced by Israel's universal healthcare system⁸. The longitudinal design of this study provides substantial benefits for monitoring medical conditions, menopausal transition and aging over time. It allows for the collection of prospective exposure data before disease onset, reducing recall bias, and facilitates the examination of factors that impact disease progression and treatment efficacy. So far, 5,800 participants have completed the follow-up at the second year, 356 have withdrawn and 10 have died.

Ethnic variation in clinical and multi-omics data

Incorporating populations from diverse ethnic backgrounds is essential for ensuring the generalizability of research findings. However, many ethnic groups remain substantially under-represented in observational cohorts and clinical trials⁹, posing challenges in achieving representation. The HPP includes deep phenotypic information spanning more than 30 modalities (Fig. 1a,b) and is ethnically diverse, reflected by the wide geographical origins of the participants' ancestors (Fig. 1c and Supplementary Table 2). This enables the investigation of multi-omics and clinical phenotypic variation across ethnicities. Whereas some data types exhibit minimal variation with ethnicity, others show significant differences, which we show using a machine learning (ML) algorithm (LightGBM¹⁰; see Methods) to predict ethnicity from different data

data, provided the most accurate predictions for ethnicity, whereas microbiome data alone had random predictive power. In addition, whereas principal component analysis (PCA) analysis revealed clear separation in the genetics data between different ethnic groups, no separation was observed in microbiome composition (Fig. 1d), in line with previous studies demonstrating that environmental factors have a greater impact on microbiome composition than does genetics¹¹. By contrast, we revealed differences across various clinical parameters between the Ashkenazi ethnic group, the largest in our cohort, and other ethnic groups, when participants were matched by sex, age and body mass index (BMI) ($n = 2,226$) (Fig. 1e and Supplementary Table 4). These include eating habits, lifestyle behaviors, blood test results and vital signs. Although some of these differences can be attributed to different cultural backgrounds, others might reflect physiological variability between ancestries. These findings highlight the importance of having ethnically diverse cohorts to ensure that findings are generalizable. The HPP is expanding globally (Fig. 1b), so the ethnic diversity of the cohort will increase, enabling further research into the ethnic variation of multi-omics data and clinical phenotypes.

Clinical and multi-omics variations with age

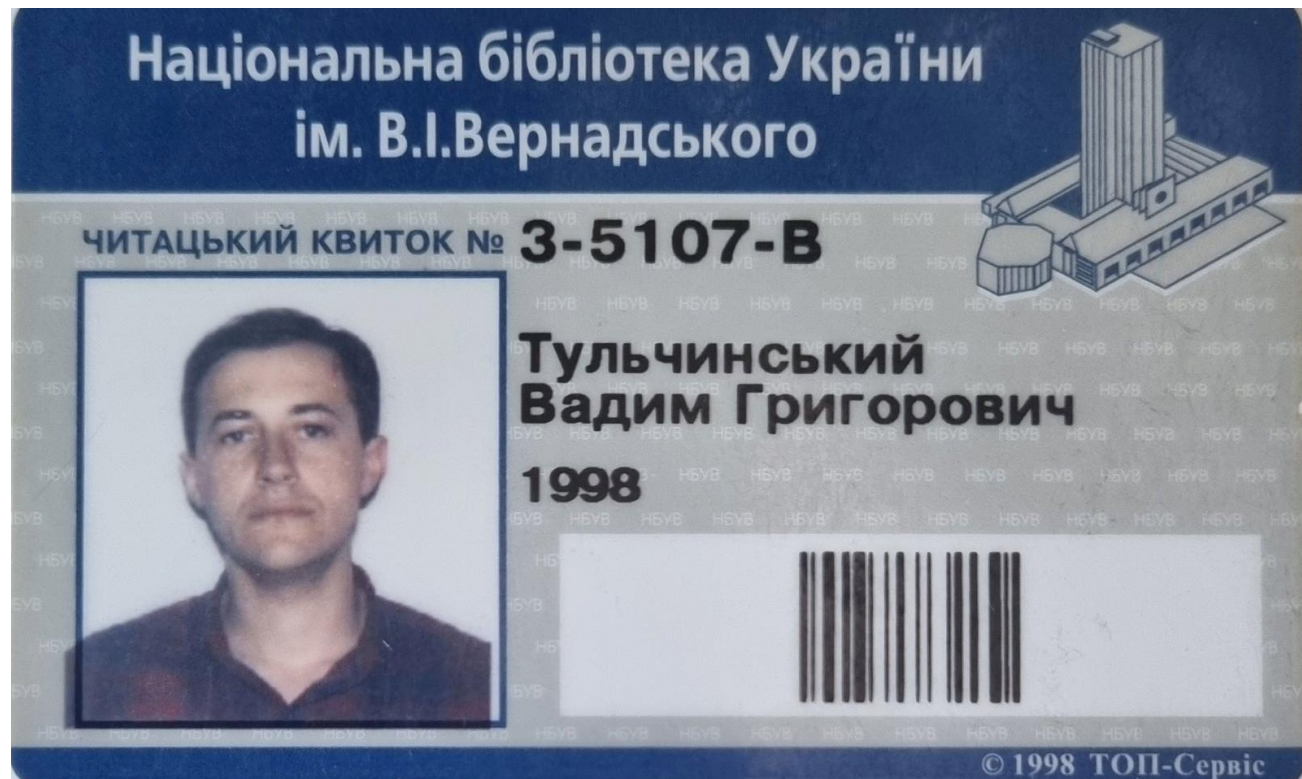
Although many medical tests have established norms for physiological measures, they are often based on data that are outdated or derived from few individuals^{12,13}. Moreover, because many of the measurements obtained in the HPP so far were not measured in large-scale cohorts of healthy people, reference values are lacking. Analyzing the distribution of these measurements across thousands of healthy individuals establishes norms tailored to specific clinical characteristics, such as age and sex. Examples include the establishment of reference values for CGM metrics¹⁴, sleep measures¹⁵ and the fundus microvasculature¹⁶. Data from the HPP were also utilized to develop an expanded reference set for the human gut microbiome, identifying hundreds of previously unknown species¹⁷. These studies provide a helpful resource for future research using these data types.

One example is the establishment of reference values for sleep-derived measures and sleep biomarkers, collected from 6,202,346 minutes of monitoring during 16,921 nights in 6,410 individuals. We examined the progression of sleep-derived features with age and obtained age–gender reference plots for peripheral apnea-hypopnea index (pAHI) (Fig. 2a). The percentage of the population in our cohort with moderate to severe obstructive sleep apnea (OSA)¹⁸ varies from 6.6% to 36.4%, depending on age and sex. Another example is the progression of fasting glucose (FG) levels with age, a key diagnostic criterion for prediabetes and diabetes¹⁹, which increases every 10 years by 2.6 mg dl⁻¹ in males and 2.5 mg dl⁻¹ in females (Fig. 2a). This approach can be applied to other complex data sets that lack established norms, such as metabolite profiles.

Deep clinical and molecular phenotyping of participants in the cohort could also challenge traditional diagnostic approaches. For example, in a previous study²⁰, CGM captured substantial day-to-day FG variability. Forty percent of individuals with FG levels in the normal range at study initiation had FG in the prediabetic range in sequential days, highlighting the deeper insights offered by continuous monitoring.

The study of aging in the HPP has provided insights into the distinction between chronological age and biological age. To explore this, we developed an ML model (LightGBM¹⁰; full details in ref. 23) to predict chronological age solely on the basis of cardiovascular measures. The model shows that individuals with a predicted cardiovascular age higher than their chronological age consistently exhibit worse clinical metrics. For example, when comparing individuals in the top versus bottom 10th percentiles of biological-age predictions, representing

МІЙ ЧИТАЦЬКИЙ КВИТОК



У 2000-му я останній раз скористався цим квитком: пред'явив як другий документ з фото (разом з закордонним паспортом), щоб отримати американський ID



МОЯ БІБЛІОТЕКА

Коли мені потрібна наукова публікація, якої нема у відкритому доступі, я типово звертаюсь в цю бібліотеку. Мій друг Андрій, що працює в Університеті Маастріхту, скачує файл.

А якщо він поїхав на конференцію, чи з ще кудись, чи хворий? Мій друг Олексій працює в Норвезькому університеті науки і технології у Тронгеймі.





ВІДКРИТТЯ №1: СТАТТІ ДЛЯ НАУКИ, ЧИ ДЛЯ КАР'ЄРИ?

Пол Гінспарґ – [arXiv.org](https://arxiv.org)

ПРЕПРИНТИ

У відкритому доступі

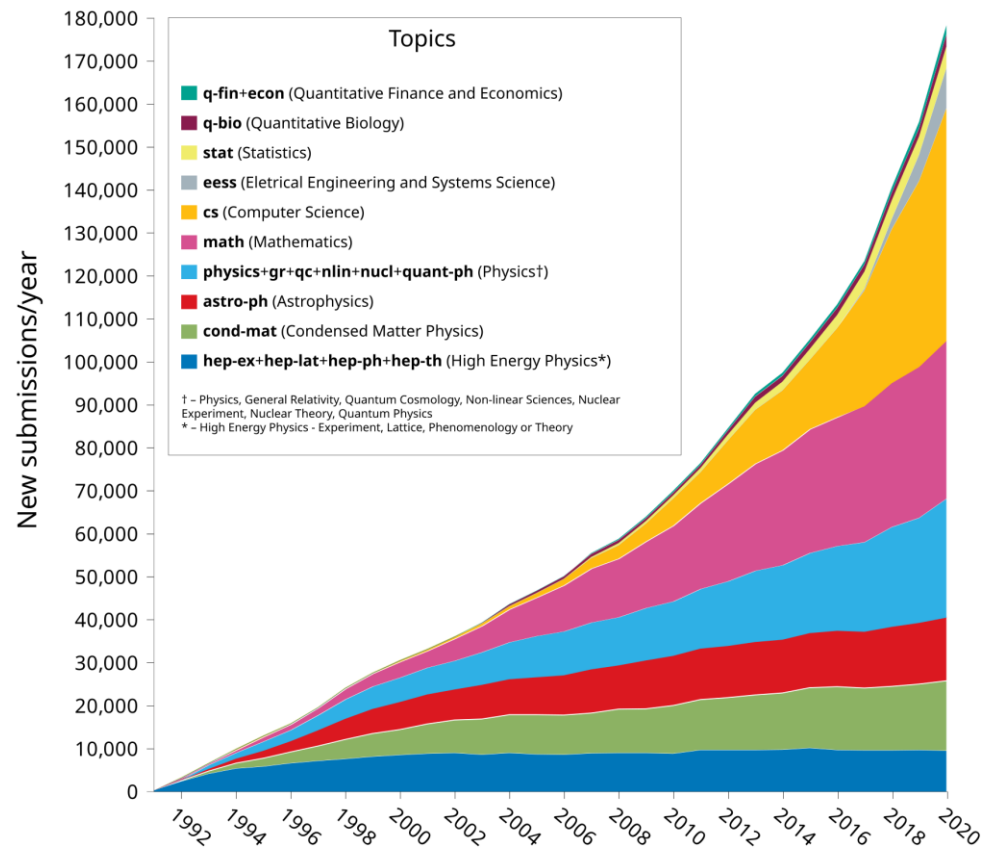
Доступні швидко

Безкоштовні для всіх

Не дають практичних вигід (не враховуються при захисті дисертацій)

Захищають авторський пріоритет при оприлюдненні разом з відкритими даними

Articles submitted to arxiv.org (1991–2020)



- **Висока якість контенту arXiv.org:** Carneiro, C.F.D., Queiroz, V.G.S., Moulin, T.C. et al. Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature. Res Integr Peer Rev, 2020, 5, 16. <https://doi.org/10.1186/s41073-020-00101-3>



ВІДКРИТТЯ №2: РІВНІСТЬ ЧИ ВЛАСНІСТЬ?

Олександра Елбакян – Sci-Hub

ВІДКРИТІ (ТІНЬОВІ) БІБЛІОТЕКИ НАУКОВИХ ПРАЦЬ

У 2011 р. був створений Sci-Hub, що надає вільний доступ до наукових статей, включно з тими, за які беруть плату видавництва.

Що це: боротьба з цифровою нерівністю, боротьба за відкриту науку, чи просто піратство?

Бібліотека	Особливості	Колекція
https://sci-hub.pub (Sci-Hub)	Багато дзеркал, 400 тис запитів на день	88 млн. статей
https://annas-archive.org/ (Anna's Archive)	Об'єднує багато колекцій (Sci-Hub, Libgen, Zlib,...)	54 млн. книг 22 млн. статей
https://librarygenesis.site/ (Library Genesis [Libgen])	Багато книг, багато непотрібного, реклама	960 млн. статей, книг, настанов...
https://z-lib.id/ (Z-Library [z-lib, b-ok.org, bookfi])	Вважався китайським, але США арештувало 2 росіян	13 млн. книг 85 млн. статей
https://unpaywall.org/ (Unpaywall database)	Легальні матеріали з 50 тис. видавництв і сайтів	54 млн. статей і препринтів
https://search.library.wisc.edu/database/UWI46808	Духи з науковими працями Китаю (КНР)	2 млн. книг і журналів

ВІДКРИТТЯ №3: НАУКОВА РОБОТА, ЧИ НАУКА?

Георгій Перельман – громадянська
наука



ПЕРЕХІД ДО ВІДКРИТОЇ НАУКИ

Головні фактори:

- Науковцям потрібна актуальна інформація + подолання цифрової нерівності + громадянська наука поза межами установ (Г. Перельман, МАН)
- Надавачам грантів не потрібно фінансування подібних і повторних досліджень, але важливі порівняння і перевірка результатів
- Видавництва вже не можуть заробляти на підписці/купівлі (на читачах), тож не борються проти відкриття текстів

Головні актори:

- S-коаліція, План S: <https://www.coalition-s.org/>
- Єврокомісія: <https://open-research-europe.ec.europa.eu/for-authors/> (Настанова журналу "Open Research Europe", де мають публікуватись результати проєктів Horizon Europe)
- Суспільство (наприклад, Хартія відкритих даних): <https://opendatacharter.org/>

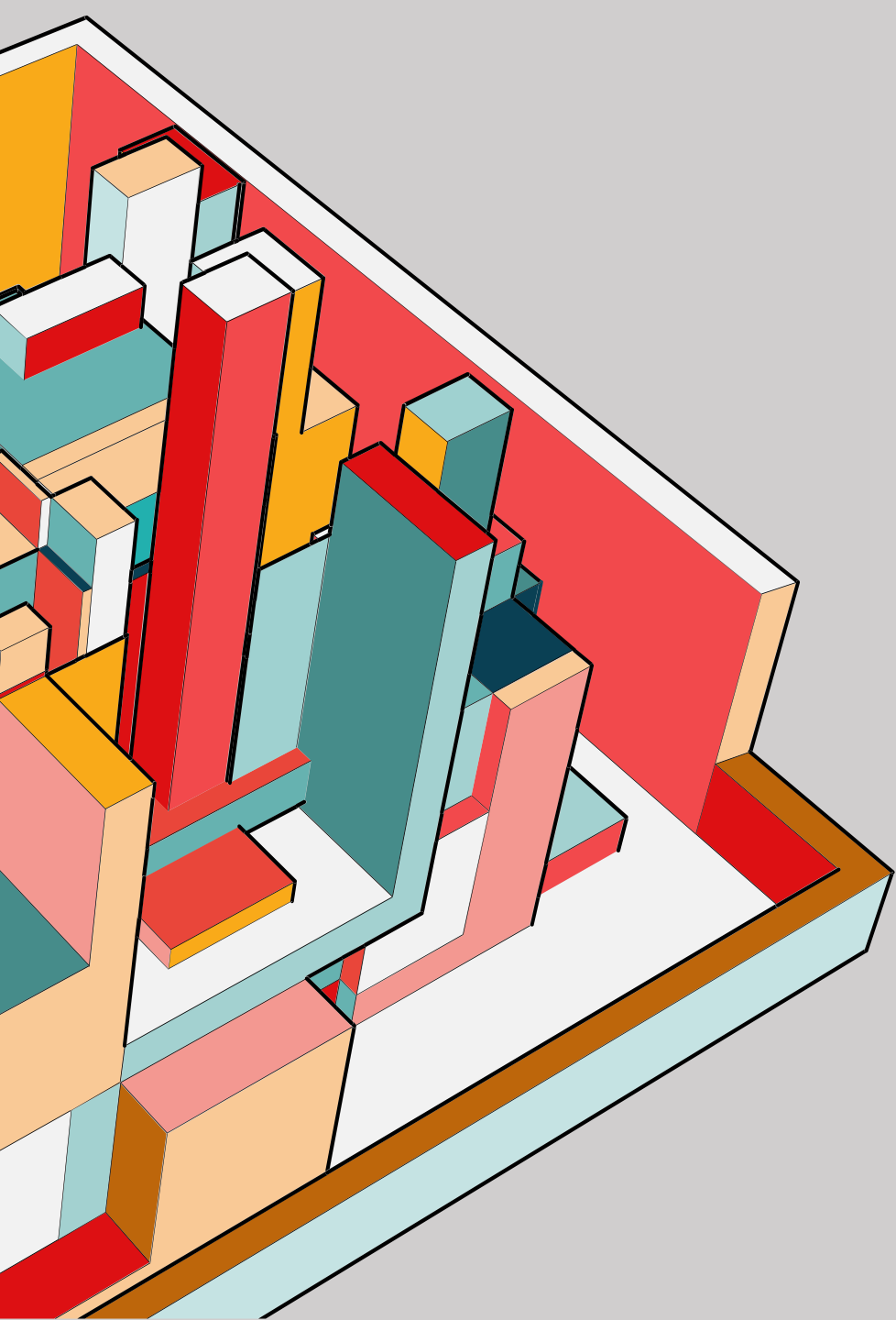


ВІДКРИТТЯ №4: ДЛЯ ВСІХ І ДАРОМ, ЧИ НІ ДЛЯ КОГО?

Ілля Суцкевер - ChatGPT

ВЕЛИКІ МОВНІ МОДЕЛІ

	ШІ-інструмент для науки	Особливості
Стають все краще	https://notebooklm.google/ (Google NotebookLM)	Обробляє завантажені тексти, формує заготовки наукових робіт і презентацій
Розуміють зв'язну мову	https://www.paperpal.com (Paperpal Generative AI)	Вбудовується в Word і Google Docs, сам робить структуру, огляд цитування
Розуміють контекст	https://www.jenni.ai/ (Jenni AI Research Assistant)	Сайт генерує структуру, огляд, цитати, допомагає розуміти і шукати плагіат
Враховують весь текст, а не лише метадані	https://www.thesisai.io/ (ThesisAI Academic Writing)	Здатен згенерувати повний текст наукової праці від 8 до 80 сторінок
Вбудовані у пошук	https://www.avidnote.com/ (Avidnote AI for Research)	Найширший набір функцій ШІ для науковців включно з аналізом даних
Вже є спеціалізовані ШІ сервіси для наукової роботи	https://scispace.com/ (SciSpace AI Research Agent)	Огляд, пошук, аналіз наукових робіт і даних, генерація і поліпшення тексту



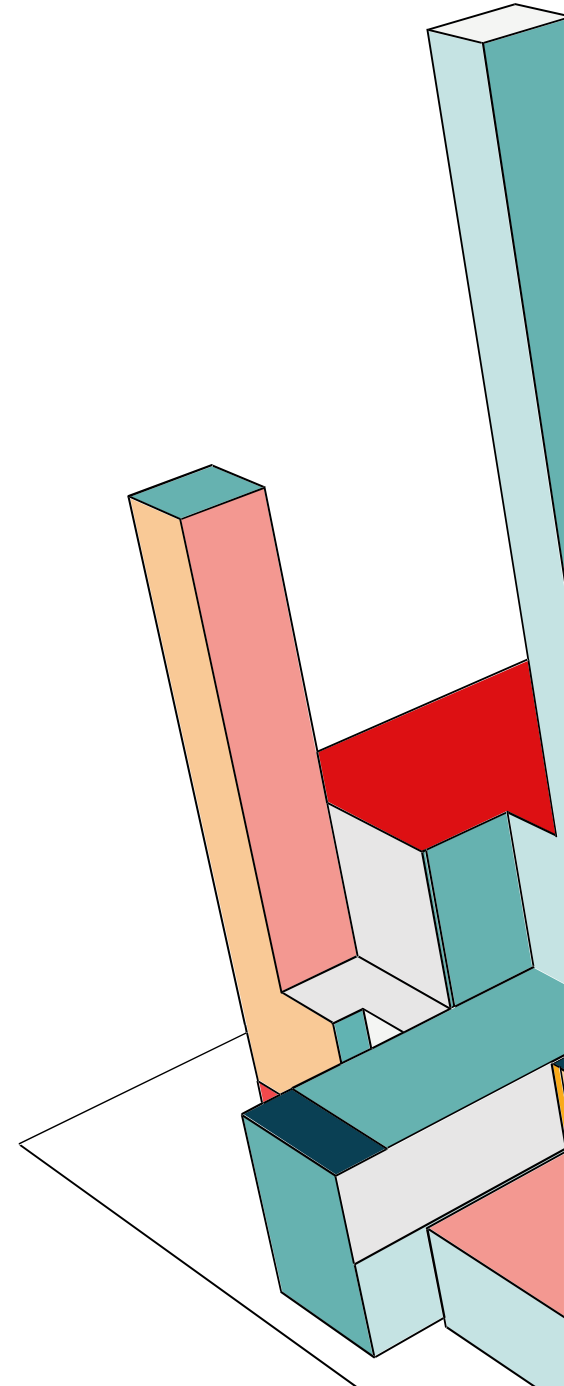
ЩО БУДЕ, ЩО РОБИТИ?

Тут мало б бути ваше фото

УМОВИ ЗА ЯКИМИ ТРЕБА ВИЖИТИ

- Обмеження на он-лайн доступ до нових наукових публікацій (з урахуванням препринтів) скасовані
- Метадані всіх відкритих публікацій доступні у міжнародних пошукових системах
- Автоматичний пошук публікацій працює на рівні мрії
- Людей, які не вміють всім цим користуватись, в науці та інженерії не залишилось (усі на пенсії)

Навіщо науково-технічні бібліотеки?





*Бібліотека монастиря Св. Катерини на Синаї (IV ст.)

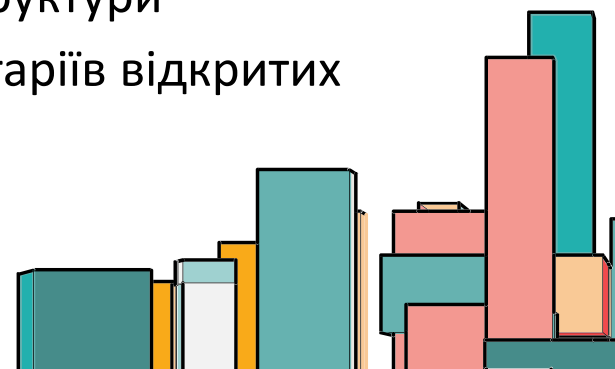
Головними функціями античної і середньовічної бібліотек було копіювання, реставрація і розмноження книг у скрипторіях.

Після розповсюдження друку скрипторії зникли.

Але бібліотеки, завдяки системі тимчасового користування книгами, стали ще актуальнішими.

ПРОПОЗИЦІЇ ЩОДО ТРАНСФОРМАЦІЇ

1. Орієнтація на автора і видавця, а не читача:
 - послуги з просування/реклами зокрема для ШІ
2. Надання в користування платних інструментів:
 - ШІ інструментів наукової роботи (з слайду 11)
 - комерційних програмних продуктів для науковців
 - тимчасових ліцензій та підписок
 - може обладнання і приладів?
3. Соціальні та інформаційні послуги:
 - збір на надання інформації про доступні поза інтернетом дослідницькі інфраструктури
 - допомога у використанні репозитаріїв відкритих даних, архівів препринтів тощо.



НАУКОВО-ТЕХНІЧНІ БІБЛІОТЕКИ ЗМІНЯТЬСЯ АБО ЗНИКНУТЬ

ДЯКУЮ!

Вадим Тульчинський

Інститут кібернетики імені В.М. Глушкова
НАН України

dep145@gmail.com

