

Pályázati azonosító	TKP2021-NKTA-34
Pályázat címe	BigData-technológiával támogatott UDBD-Health adattárház fejlesztése és üzemeltetése

A Debreceni Egyetem TKP2021-NKTA-34 azonosító számú „BigData-technológiával támogatott UDBD-Health adattárház fejlesztése és üzemeltetése” elnevezésű projektje az Innovációs és Technológiai Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a TKP2021-NKTA pályázati program finanszírozásában valósul meg 1 400 000 000 Ft összegben.

A támogatás intenzitása 100 %.

A Kutatás megvalósítása 2021. december 01. napjától 2025. november 30. napjáig tart.

A Debreceni Egyetem, mint a magyar felsőoktatás kiemelkedő kutató, fejlesztő intézménye, küldetésnyilatkozatában, tevékenységének fókuszát az egészségiparban és annak innovációs lehetőségeiben határozta meg. Ezért jelen Big Data projektben célul tűztük ki a Debreceni Egyetemen képződő nagy tömegű adatok megfelelő tárolásának és feldolgozásának optimalizálását, az Egyetem kutató-fejlesztő-innovatív tevékenységének legkorszerűbb támogatása céljából. Ugyancsak célul tűztük ki, hogy a strukturált adatbázisok az Egyetem szolgáltatási portfóliójában is megjelenjenek.

A Debreceni Egyetem adatvagyonának jelentős részét teszik ki a Klinikai Központban folyó betegellátás során keletkezett medikai és képi adatok, melyek kutatási célú elérhetősége erősen korlátozott. A BigData tématerület keretében 2018 óta folytatott K+F tevékenység egyik legfontosabb eredménye az, hogy ezt az adatvagyonot megfelelő adattranszformációs eljárások kidolgozásával a kutatók számára elérhetővé tettük. A Microsoft Azure környezetben kialakított UDBD Health adattárház és a hozzá kapcsolt skálázható számítógéprendszer biztosítja a pseudo-anonim adatok folyamatos gyűjtését, valamint a kutatók számára az adattudomány korszerű elemző módszereinek (pl. a gépi tanuláson alapuló eljárások) használatát.

A megfelelő módon strukturált és anonimizált adattárház, a kórházi informatikai rendszerből származó, 20 évet lefedő szöveges adatokból, képi és genetikai diagnosztikai információkból áll össze. Ezzel lehetővé válik olyan rejtett összefüggések megtalálása és elemzése is, melyek nem csak egyes népbetegségek (pl. cukorbetegség, magas vérnyomás), hanem akár a ritka betegségek megértéséhez is információval szolgálhatnak.

A célok elérése érdekében négy kutatócsoport fogott össze:



1. Klinikai BigData (KBD) - a klinikai orvosi adatokat előkészítő és azok elemzésének fő szempontjait megfogalmazó kutatócsoport,
2. GEN - a genomikai feldolgozást előkészítő kutatócsoport,
3. MI - a mesterséges intelligencia alkalmazásának optimális lehetőségeit kidolgozó kutatócsoport, valamint a
4. HPC - nagy számításigényű elemzések hardveres feltételrendszerét optimalizáló egyetemi szuperszámítógépet működtető kutatócsoportból áll.

A fejlesztési projekt első két évben az adatgyűjtés, az adattisztítás és az adatintegrálás IT háttérét dolgoztuk ki a négy munkacsoport összefogásával. A KBD munkacsoport a medikai adatok feldolgozásával, valamint a PACS rendszerben tárolt CT, MRI és PET képekből származó adatok integrálásával ért el eredményeket. A HPC munkacsoport a nagy mennyiségű adatfeldolgozásához szükséges számítási kapacitás támogatására fejlesztett kommunikációs alkalmazást, valamint kidolgozott egy párhuzamos feldolgozási algoritmus gyűjteményt. Az MI munkacsoport feladatai közé tartozott a PET/CT felvételeken légúti és tüdő daganatok, kóros halmozások automatizált detektálása és szegmentálása, citológiai minták automatizált kiértékelése, valamint bőrpigmentációk osztályozása gépi tanuláson és képfeldolgozáson alapuló módszerek kidolgozásával. A GEN munkacsoport több olyan fontos eljárást fejlesztett ki, amivel a nagymennyiségű genomikai adatok elemzése megoldható.

A négy munkacsoport összehangolt K+F tevékenysége révén egy olyan modell értékű adattárházat sikerül kialakítani a projekt 3. évének végére, amely hazai viszonylatban egyedülálló, nyitottsága révén megoldható bármely szakterület speciális adatainak integrálása is (pl. proteomika)

2021-2025 között a fejlesztés fókuszai a következők lesznek:

1. az adatvagyon kezelésének szabályozása, kiemelkedő figyelmet fordítva a humán genetikai adatok adattárházba való integrálásával kapcsolatos adatbiztonsági kérdéseknek,
2. véglegesíteni az Egyetemen belüli kutatásokat támogató adattárház alkalmazásokat, és annak szabályrendszerét létrehozni
3. kialakítani az Egyetemen kívüli, az adattárház kutatásával kapcsolatos megkeresések kielégítésére egy az adatbiztonságnak, a titkosságnak és a GDPR-nak egyaránt megfelelő környezetet,
4. a külső megkeresések (gyógyszergyár, egyéb egészségipari fejlesztés, EU-n belüli kutatások stb.) kielégítésére megrendelési és szolgáltatási szabályok kialakítása, pilot projektek lefuttatása.

A Big Data HPC Kutatócsoport az egyetlen nagyméretű fájl gyors továbbítását megvalósító applikációt fejleszti tovább sok nagyméretű fájl egyidejű továbbítása céljából.

A kutatás célja mesterséges intelligencián - elsősorban mélytanuláson - alapuló természetes nyelvi feldolgozó és képfeldolgozó módszerek fejlesztése klinikai adatok elemzéséhez. A kifejlesztendő



természetes nyelvi feldolgozó módszerek a klinikai információs rendszerekben tárolt, változatos formátumú szöveges adatok feldolgozásának előkészítését, és elemzését teszik lehetővé.

2023-ra tervezzük, hogy az adattárház az egyetem kutatói megfelelő szabályrendszer mellett használhassák (1. mérföldkő), majd a szerzett tapasztalatok alapján egy új egyetemi szolgáltatásként ezt az adatvagyonot piaci résztvevők számára is elérhetővé kívánjuk tenni (2. mérföldkő).



NEMZETI KUTATÁSI, FEJLESZTÉSI
ÉS INNOVÁCIÓS HIVATAL

AZ NKFI ALAPBÓL
MEGVALÓSULÓ
PROJEKT