

Szabad szoftveres OCR

a Videotorium keresőmotor támogatására



VIDEOTORIUM

Turcsányi Tamás, NIIF



Videotorium 2. workshop - Budapest, 2010. december 19.

Miért kell a prezentáció?



ISSC
International
Social
Sciences
Council

A new Action Agenda

- **New Focus:** Continued concern with the quality and growth of social sciences, but with a new emphasis on their
 - Capacity
 - Utility
 - Connectivity, also with the natural and human sciences
- **New Role:** Working as a catalyst, mobiliser, facilitator and coordinator, bringing together researchers, scholars, funders and decision makers from all parts of the world

Prezentációk és diák a Videotoriumon

Prezentációk: PDF, PPT, ODP

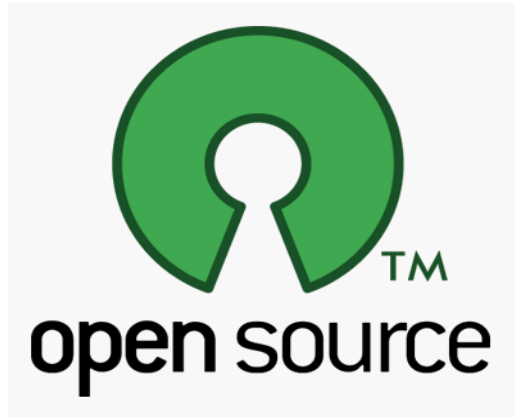
Diák: diaszerkesztő, migráció

Diák visszakereshetősége: a kezdetek

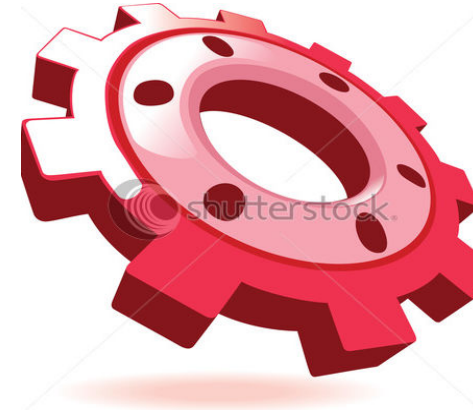


Célok

Ingyenes, OS megoldás



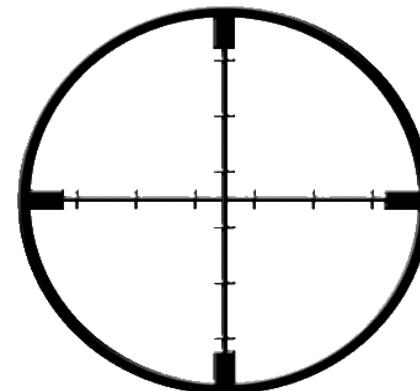
~46000 dia feldolgozása



Bitkép feltöltések



Precíz kereső

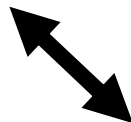
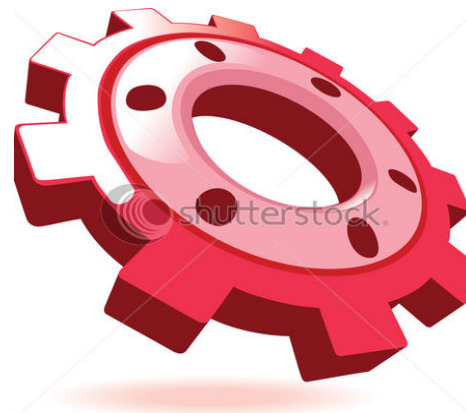


Az ideális OCR szoftver

OS



Parancssor



Stabil

Az ideális OCR szoftver



?

Recognita Omnipage

vietOCR

OCROpus

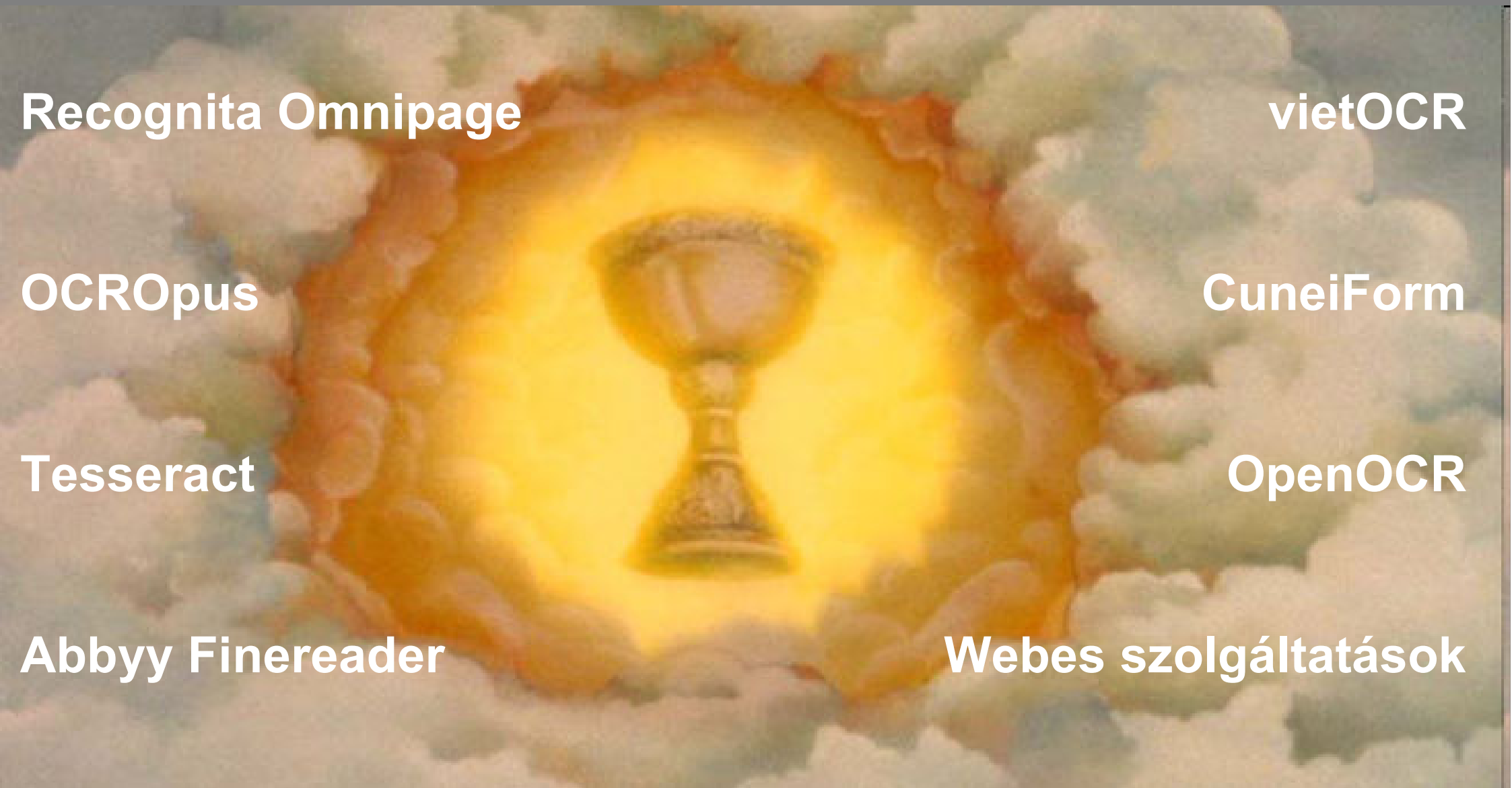
CuneiForm

Tesseract

OpenOCR

Abby Finereader

Webes szolgáltatások



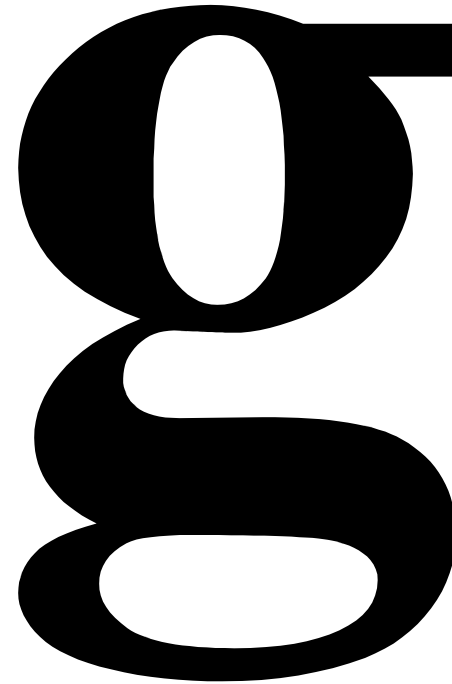
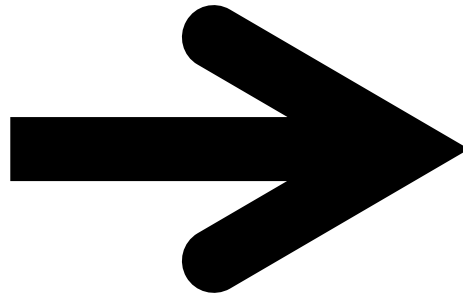
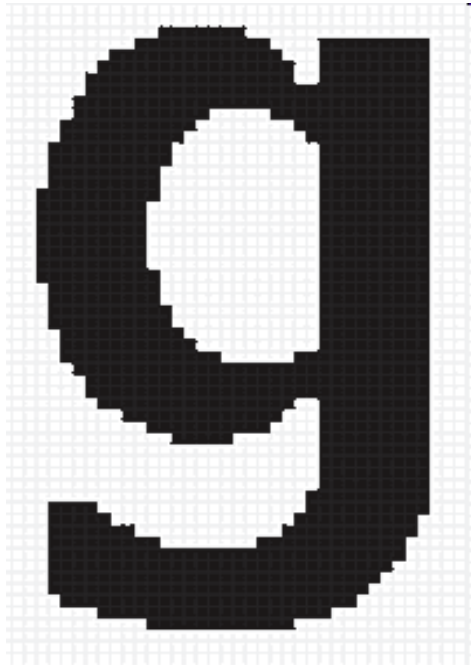
Tesseract tréning

Daniel Bell amerikai szociológus (szerint) 1788-ban vette kezdetét során civilizációnk egyre {inkább} elveszítette uralmát #információ véletlenül választotta ezt az évszámot, ugyanis ekkor jelent meg az 3. kiadása. Az előző kettő <anyagát> csupán 3-4% szerző állította hogy az „enciklopédiában” megjelenő [tudásanyagot] szerkesztői legalábbis át tudták látni. A következő kiadásoknál már egyre növe volt szükség.

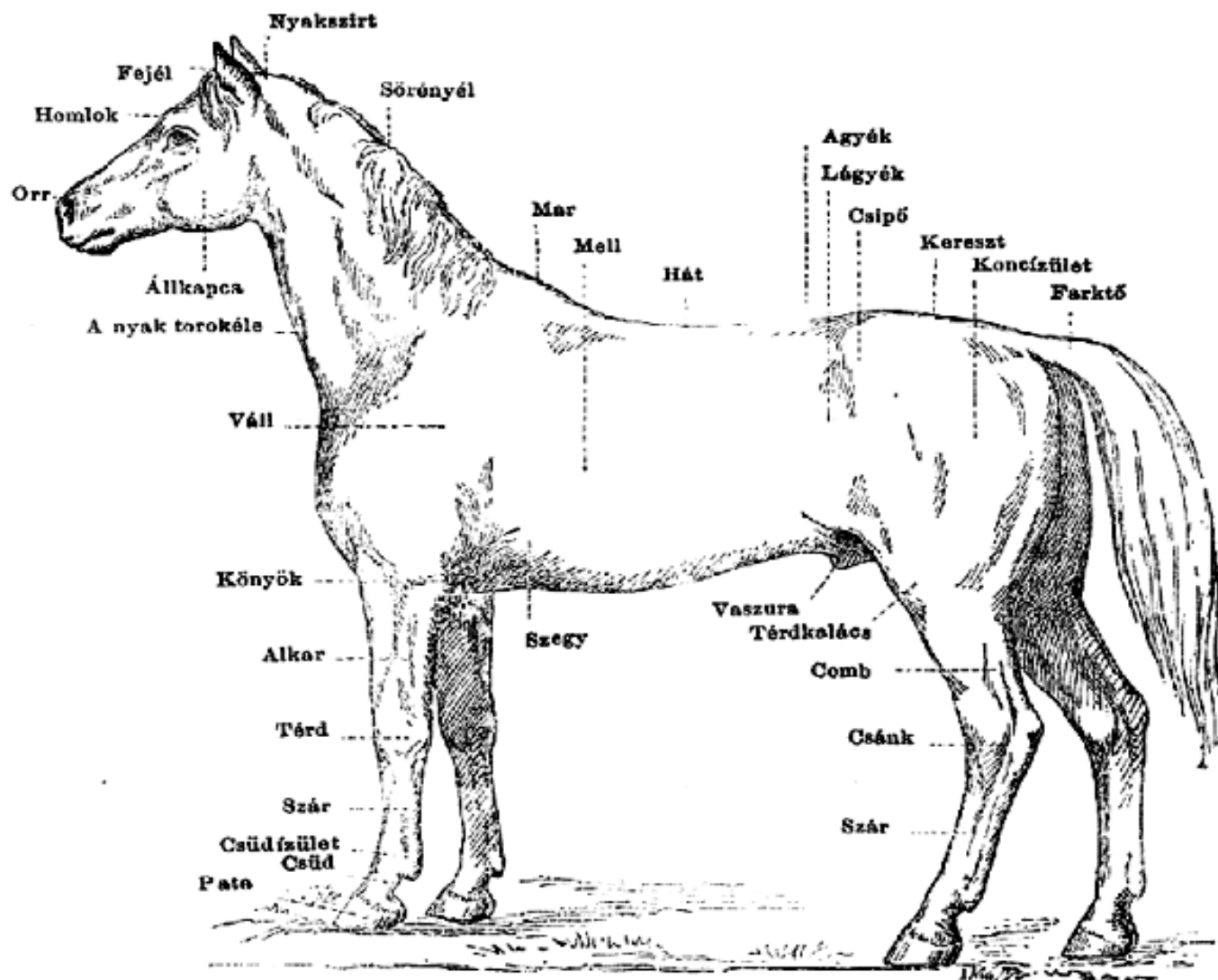
Ettől a kortól kezdve a felhalmozott ismeret modern kor egyik fő problémájává az hatalmas méretű ismeretanyagot az emberi tudás megdöbbentő felhalmozódásával

Daniel Bell a
során civilizá
véletlenül vá
3. kiadása. A

Amit az OCR szeret



Amink van



0 1 2 3 4 5 6 7

8 9 A B C D E F

G H I J K L M N

O P Q R S T U V

W X Y Z a b c d

e f g h i j k l

m n o p q r s t

u v w x y z



Hol tartunk ma?

Tétényi István

NIIF MT

Networkshop 2003

Pécs



ESZKÖZÖK:

INFORMÁCIÓS TÁRSADALOM: E-EUROPE;

INTEGRÁLT ÉS LIBERALIZÁLT KOMMUNIKÁCIÓ;

SZABÁLYOZOTT E-KERESKEDELEM.

ERA: KOORDINÁLT TUDOMÁNYPOLITIKA;

KIVÁLÓSÁGI KÖZPONTOK ÉS HÁLÓZATAIK;

INFRASTRUKTÚRA; 169§; MOBILITÁS;

INTEGRÁLT PROJEKTEK; FP6; NAGYSEBESSÉGŰ

ELEKTRONIKUS HÁLÓZAT. *

KÉRDŐJEL: SZAVAK ÉS TETTEK ARÁNYA.



Eszköz-tender I.

- **Közbeszerzési eljárás:**

- 2003 június – 2003 szeptember
- 1db nagytejesítményű Multipoint Control Unit (MCU)
- 40db professzionális H.323 végberendezés (15 helyett!)
- Szállító: LNX Hálózatintegrációs Rt.
- Gyártó: Polycom (USA)

- **MCU:**

- Accord MGC-100 (Polycom)
- 16 kártyahely
- Max. 24 @ 768 Kbps
- Max. 16 CP konferenciában

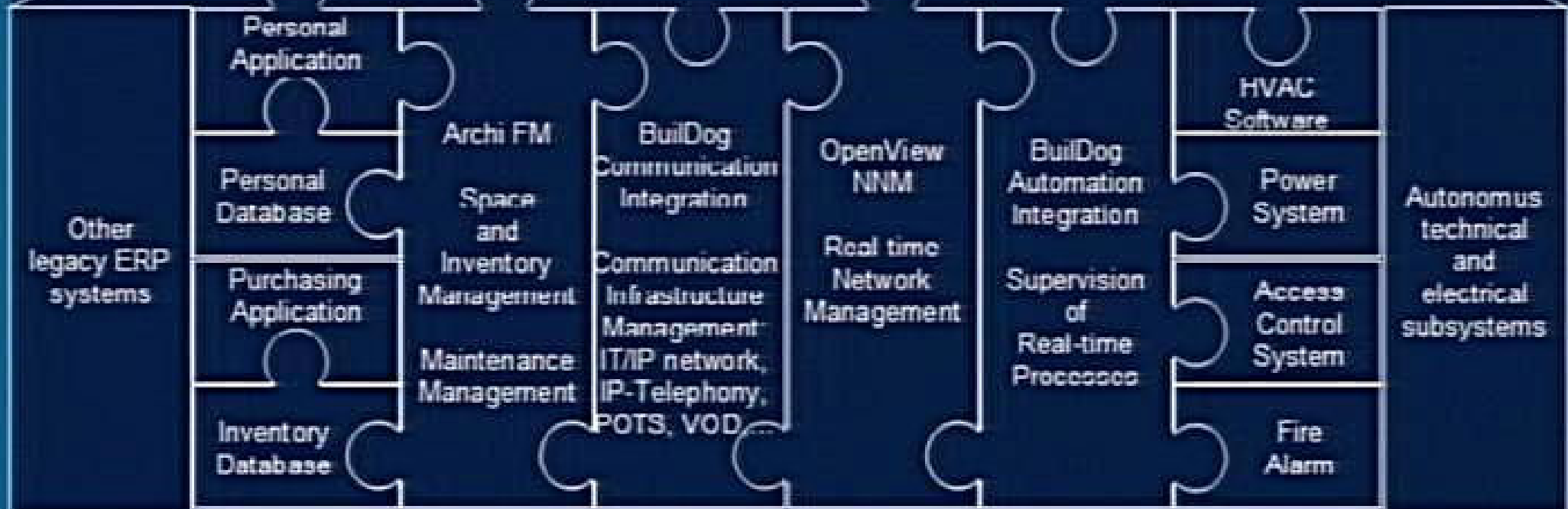


Az intelligens épület elemei és kapcsolatai



Intelligent Building
Management Integration Platform

OV Service Desk: Service Level Management



Előfeldolgozás



adaptív

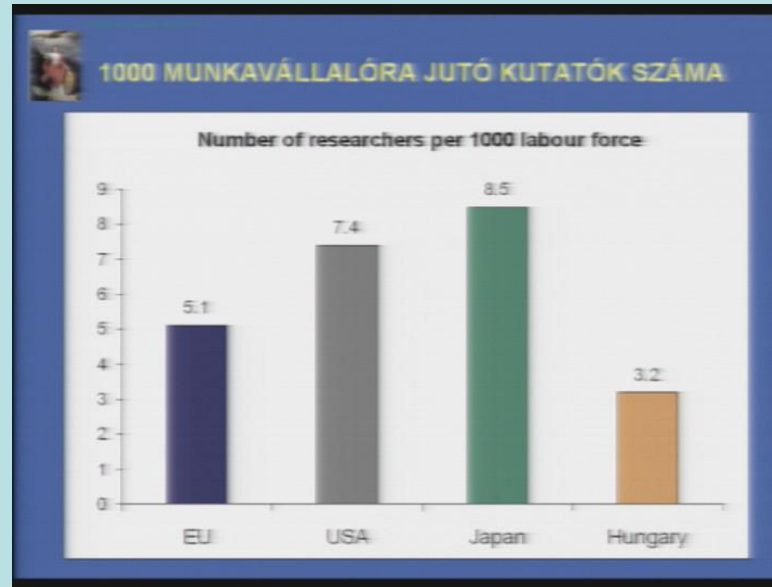


Lehetőségek



középutas „adaptív”

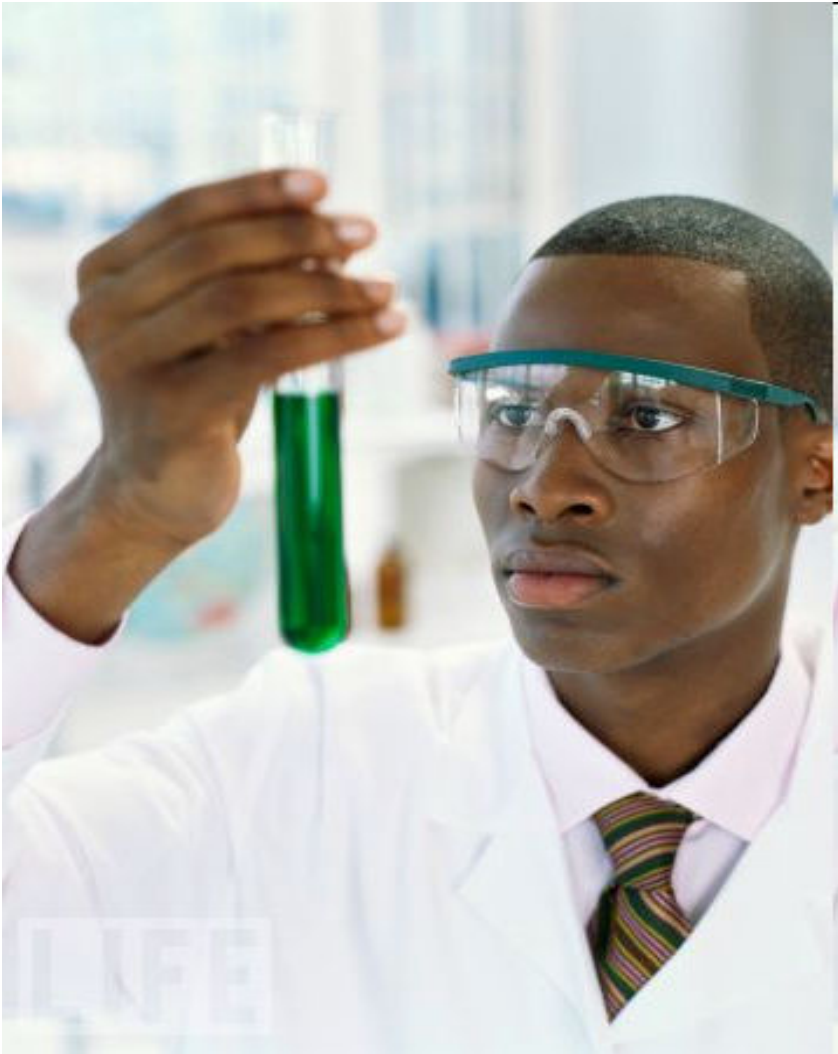
KERESSÜK!



**Aki bármit tud a fenti diáról,
kérjük hívja az alábbi számot:**

06-55/555-NIIF

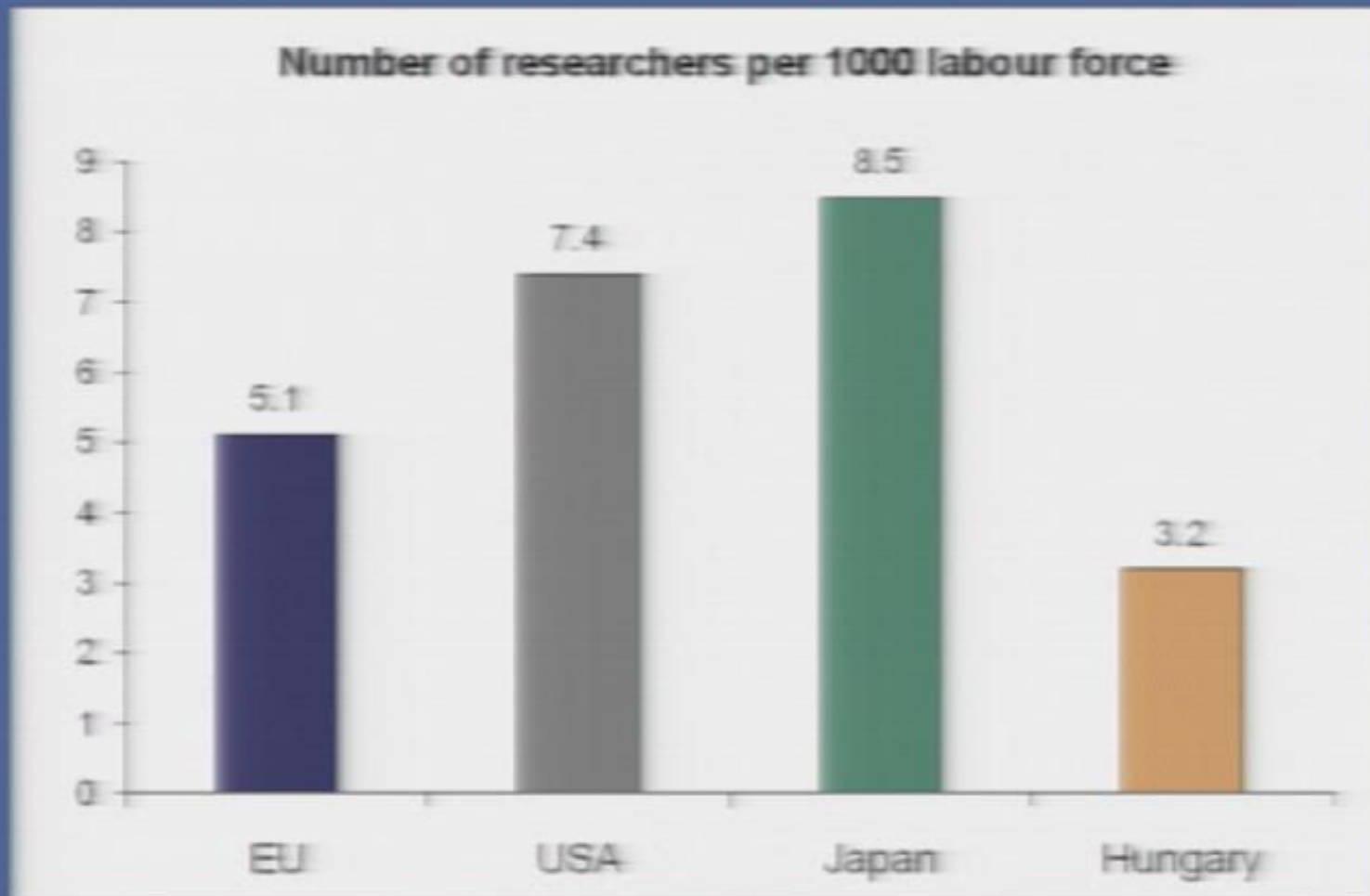
Year	Country	Value	Year	Country	Value
1990	Algeria	0.0	1990	Algeria	0.0
1991	Algeria	0.0	1991	Algeria	0.0
1992	Algeria	0.0	1992	Algeria	0.0
1993	Algeria	0.0	1993	Algeria	0.0
1994	Algeria	0.0	1994	Algeria	0.0
1995	Algeria	0.0	1995	Algeria	0.0
1996	Algeria	0.0	1996	Algeria	0.0
1997	Algeria	0.0	1997	Algeria	0.0
1998	Algeria	0.0	1998	Algeria	0.0
1999	Algeria	0.0	1999	Algeria	0.0
2000	Algeria	0.0	2000	Algeria	0.0
2001	Algeria	0.0	2001	Algeria	0.0
2002	Algeria	0.0	2002	Algeria	0.0
2003	Algeria	0.0	2003	Algeria	0.0
2004	Algeria	0.0	2004	Algeria	0.0
2005	Algeria	0.0	2005	Algeria	0.0
2006	Algeria	0.0	2006	Algeria	0.0
2007	Algeria	0.0	2007	Algeria	0.0
2008	Algeria	0.0	2008	Algeria	0.0
2009	Algeria	0.0	2009	Algeria	0.0
2010	Algeria	0.0	2010	Algeria	0.0
2011	Algeria	0.0	2011	Algeria	0.0
2012	Algeria	0.0	2012	Algeria	0.0
2013	Algeria	0.0	2013	Algeria	0.0
2014	Algeria	0.0	2014	Algeria	0.0
2015	Algeria	0.0	2015	Algeria	0.0
2016	Algeria	0.0	2016	Algeria	0.0
2017	Algeria	0.0	2017	Algeria	0.0
2018	Algeria	0.0	2018	Algeria	0.0
2019	Algeria	0.0	2019	Algeria	0.0
2020	Algeria	0.0	2020	Algeria	0.0
2021	Algeria	0.0	2021	Algeria	0.0
2022	Algeria	0.0	2022	Algeria	0.0
2023	Algeria	0.0	2023	Algeria	0.0
2024	Algeria	0.0	2024	Algeria	0.0
2025	Algeria	0.0	2025	Algeria	0.0
2026	Algeria	0.0	2026	Algeria	0.0
2027	Algeria	0.0	2027	Algeria	0.0
2028	Algeria	0.0	2028	Algeria	0.0
2029	Algeria	0.0	2029	Algeria	0.0
2030	Algeria	0.0	2030	Algeria	0.0
2031	Algeria	0.0	2031	Algeria	0.0
2032	Algeria	0.0	2032	Algeria	0.0
2033	Algeria	0.0	2033	Algeria	0.0
2034	Algeria	0.0	2034	Algeria	0.0
2035	Algeria	0.0	2035	Algeria	0.0
2036	Algeria	0.0	2036	Algeria	0.0
2037	Algeria	0.0	2037	Algeria	0.0
2038	Algeria	0.0	2038	Algeria	0.0
2039	Algeria	0.0	2039	Algeria	0.0
2040	Algeria	0.0	2040	Algeria	0.0
2041	Algeria	0.0	2041	Algeria	0.0
2042	Algeria	0.0	2042	Algeria	0.0
2043	Algeria	0.0	2043	Algeria	0.0
2044	Algeria	0.0	2044	Algeria	0.0
2045	Algeria	0.0	2045	Algeria	0.0
2046	Algeria	0.0	2046	Algeria	0.0
2047	Algeria	0.0	2047	Algeria	0.0
2048	Algeria	0.0	2048	Algeria	0.0
2049	Algeria	0.0	2049	Algeria	0.0
2050	Algeria	0.0	2050	Algeria	0.0
2051	Algeria	0.0	2051	Algeria	0.0
2052	Algeria	0.0	2052	Algeria	0.0
2053	Algeria	0.0	2053	Algeria	0.0
2054	Algeria	0.0	2054	Algeria	0.0
2055	Algeria	0.0	2055	Algeria	0.0
2056	Algeria	0.0	2056	Algeria	0.0
2057	Algeria	0.0	2057	Algeria	0.0
2058	Algeria	0.0	2058	Algeria	0.0
2059	Algeria	0.0	2059	Algeria	0.0
2060	Algeria	0.0	2060	Algeria	0.0
2061	Algeria	0.0	2061	Algeria	0.0
2062	Algeria	0.0	2062	Algeria	0.0
2063	Algeria	0.0	2063	Algeria	0.0
2064	Algeria	0.0	2064	Algeria	0.0
2065	Algeria	0.0	2065	Algeria	0.0
2066	Algeria	0.0	2066	Algeria	0.0
2067	Algeria	0.0	2067	Algeria	0.0
2068	Algeria	0.0	2068	Algeria	0.0
2069	Algeria	0.0	2069	Algeria	0.0
2070	Algeria	0.0	2070	Algeria	0.0
2071	Algeria	0.0	2071	Algeria	0.0
2072	Algeria	0.0	2072	Algeria	0.0
2073	Algeria	0.0	2073	Algeria	0.0
2074	Algeria	0.0	2074	Algeria	0.0
2075	Algeria	0.0	2075	Algeria	0.0
2076	Algeria	0.0	2076	Algeria	0.0
2077	Algeria	0.0	2077	Algeria	0.0
2078	Algeria	0.0	2078	Algeria	0.0
2079	Algeria	0.0	2079	Algeria	0.0
2080	Algeria	0.0	2080	Algeria	0.0
2081	Algeria	0.0	2081	Algeria	0.0
2082	Algeria	0.0	2082	Algeria	0.0
2083	Algeria	0.0	2083	Algeria	0.0
2084	Algeria	0.0	2084	Algeria	0.0
2085	Algeria	0.0	2085	Algeria	0.0
2086	Algeria	0.0	2086	Algeria	0.0
2087	Algeria	0.0	2087	Algeria	0.0
2088	Algeria	0.0	2088	Algeria	0.0
2089	Algeria	0.0	2089	Algeria	0.0
2090	Algeria	0.0	2090	Algeria	0.0
2091	Algeria	0.0	2091	Algeria	0.0
2092	Algeria	0.0	2092	Algeria	0.0
2093	Algeria	0.0	2093	Algeria	0.0
2094	Algeria	0.0	2094	Algeria	0.0
2095	Algeria	0.0	2095	Algeria	0.0
2096	Algeria	0.0	2096	Algeria	0.0
2097	Algeria	0.0	2097	Algeria	0.0
2098	Algeria	0.0	2098	Algeria	0.0
2099	Algeria	0.0	2099	Algeria	0.0
2100	Algeria	0.0	2100	Algeria	0.0



Tesztalany



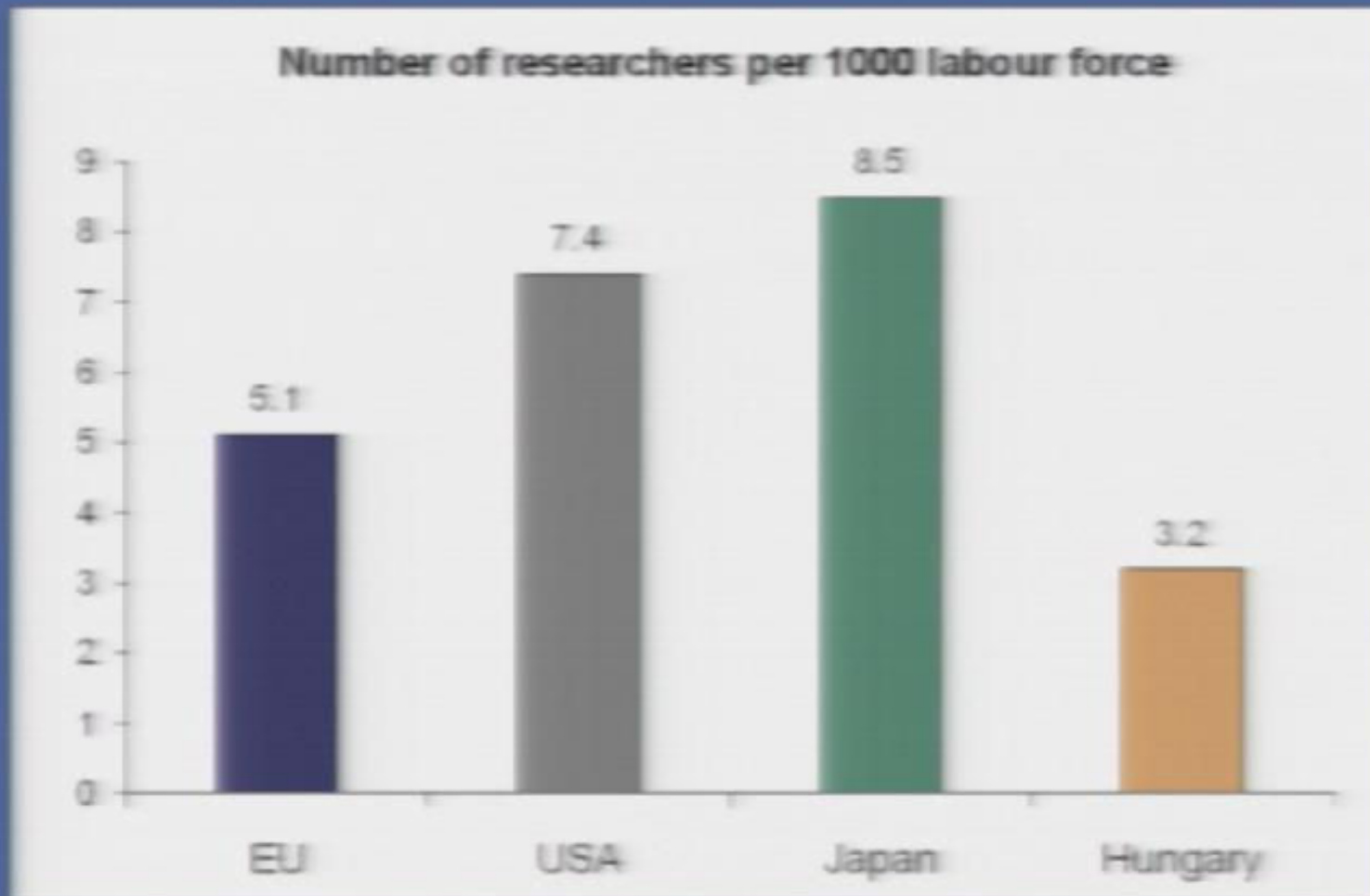
1000 MUNKAVÁLLALÓRA JUTÓ KUTATÓK SZÁMA



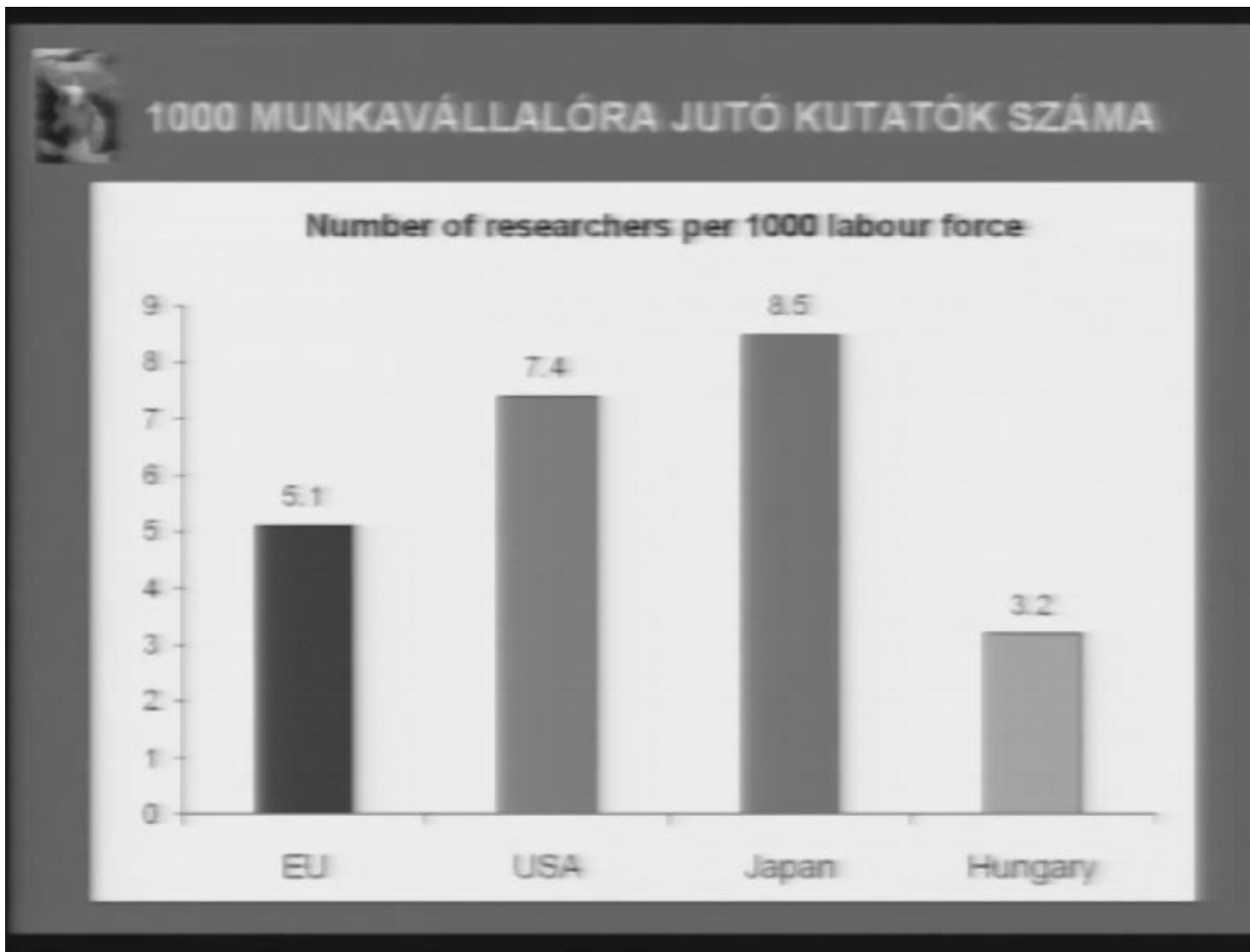
-despeckle



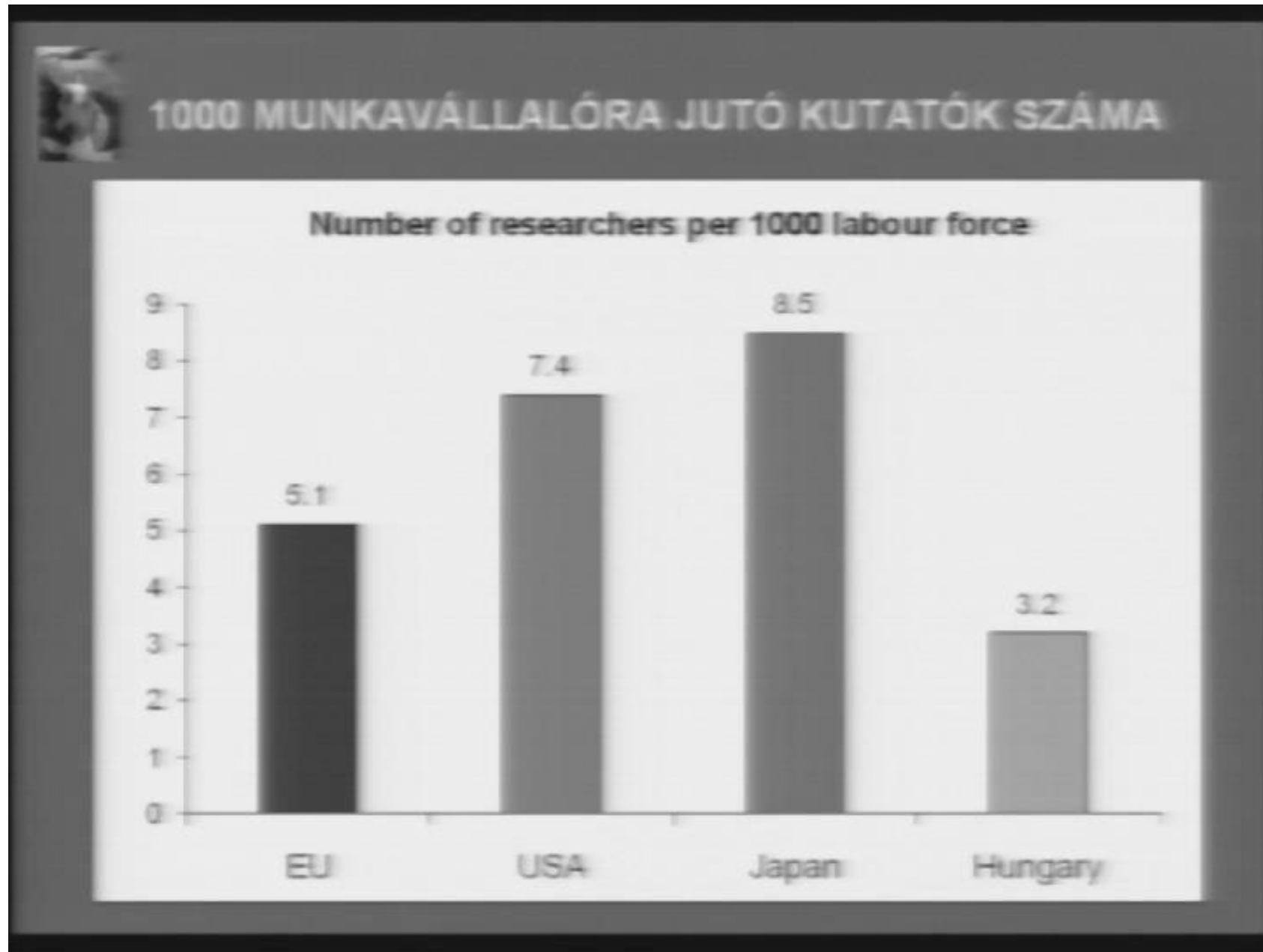
1000 MUNKAVÁLLALÓRA JUTÓ KUTATÓK SZÁMA



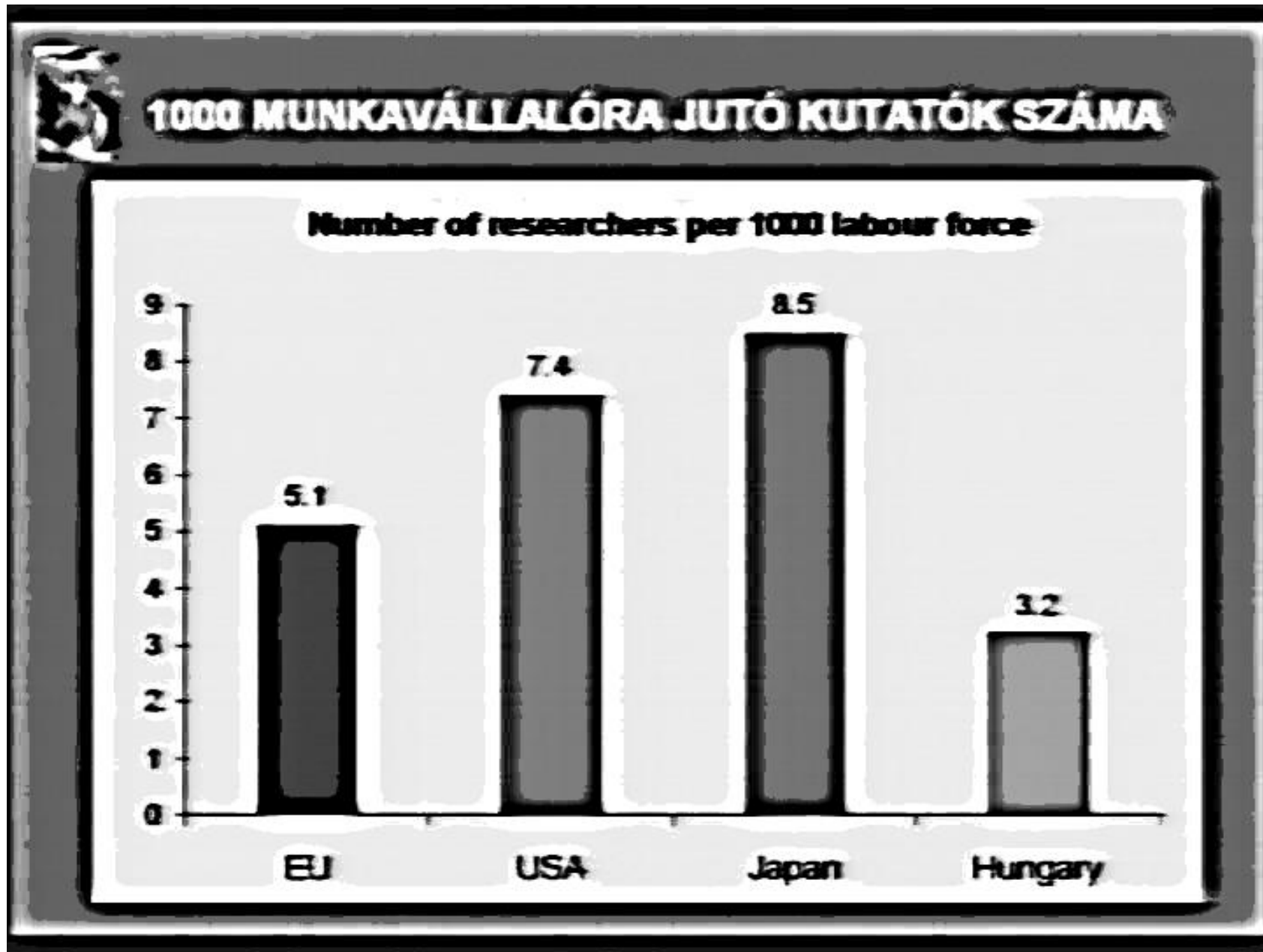
-colorspace gray



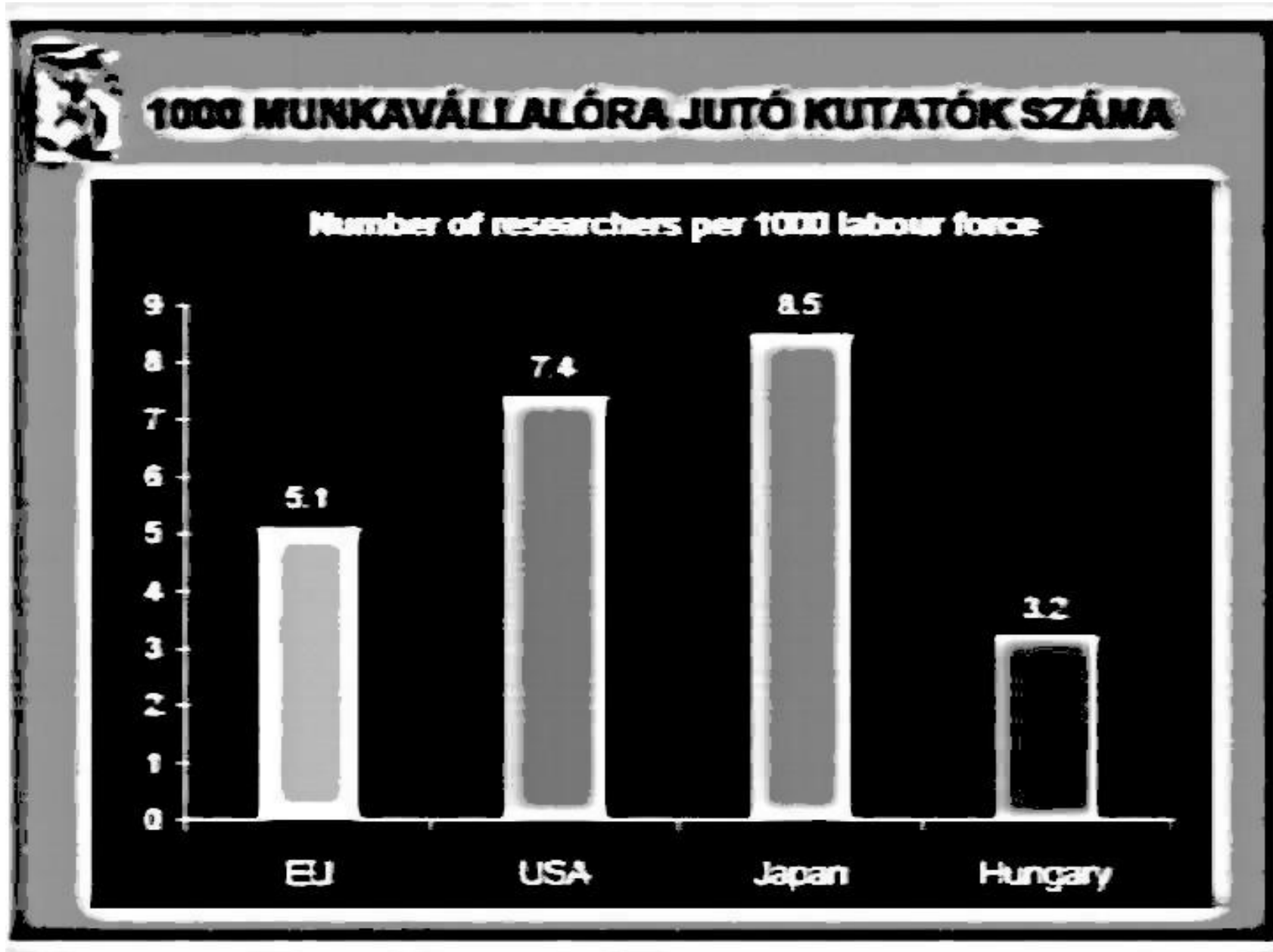
-resize 200%



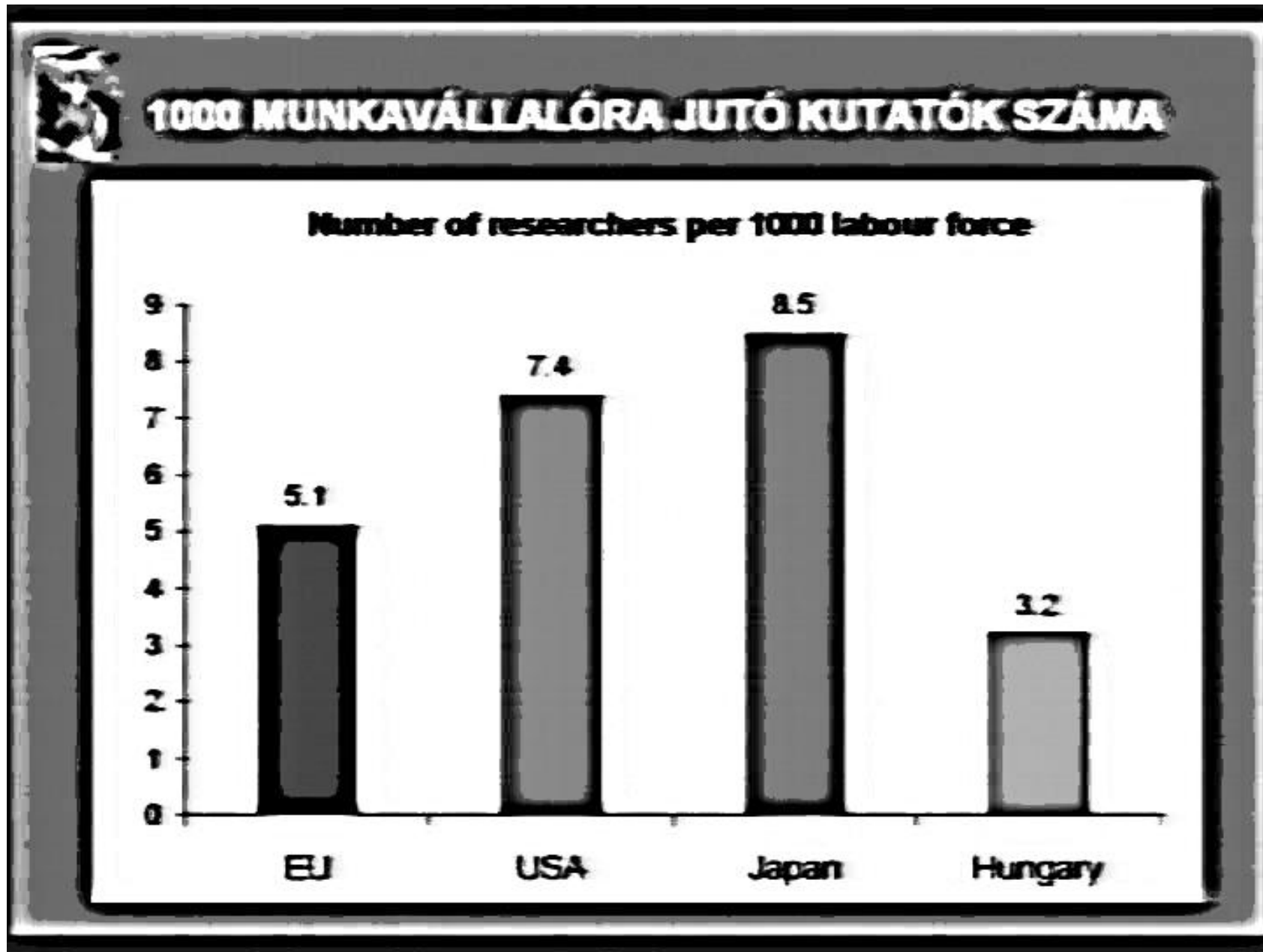
-unsharp 0x6+7%



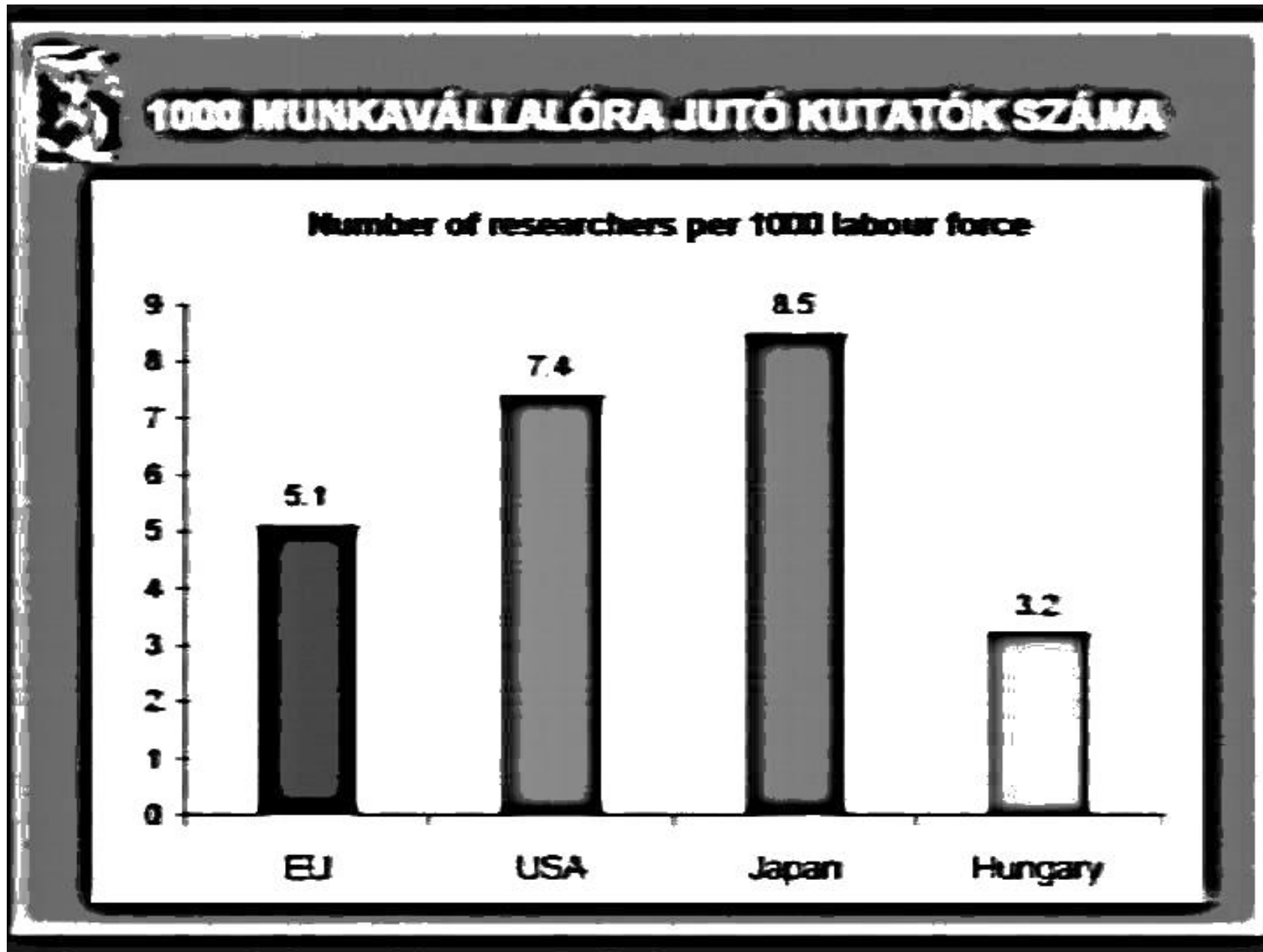
invertálás?



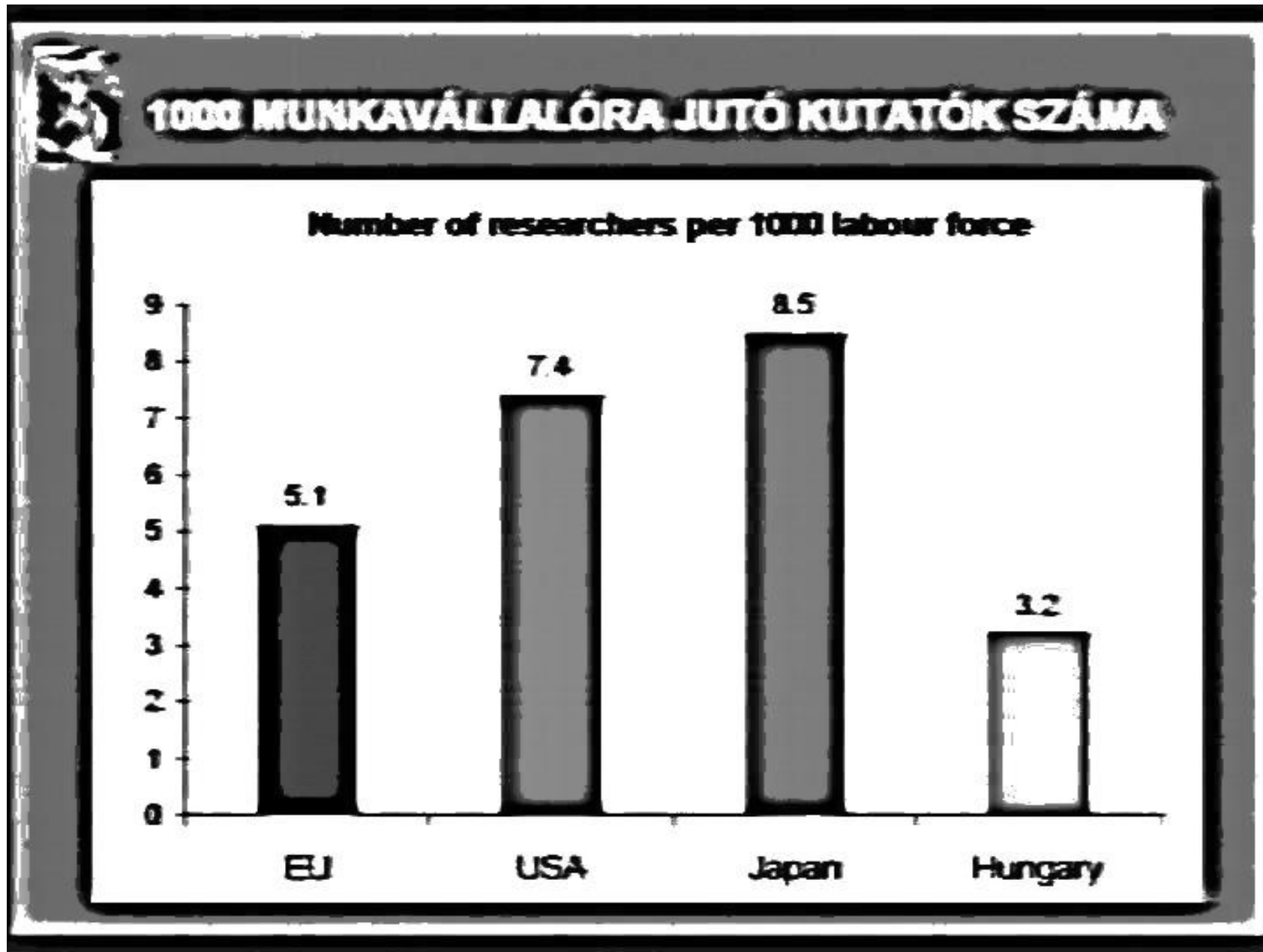
-contrast-stretch 0%x50%



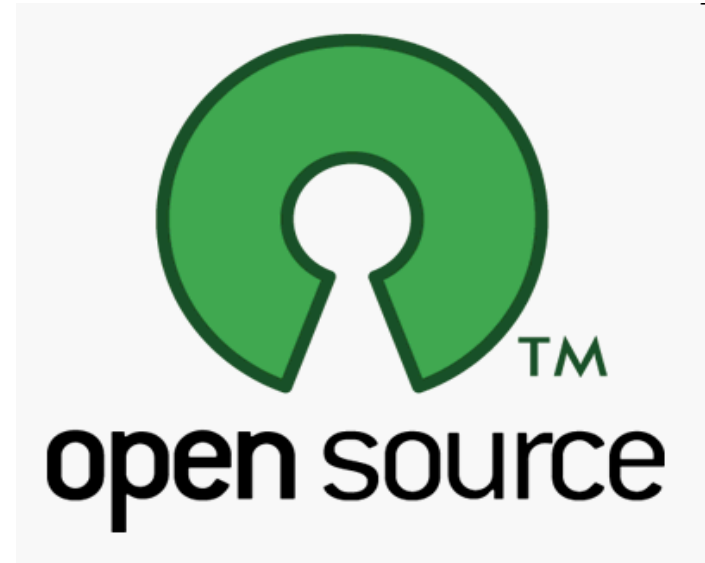
-white-threshold 68%



-despeckle

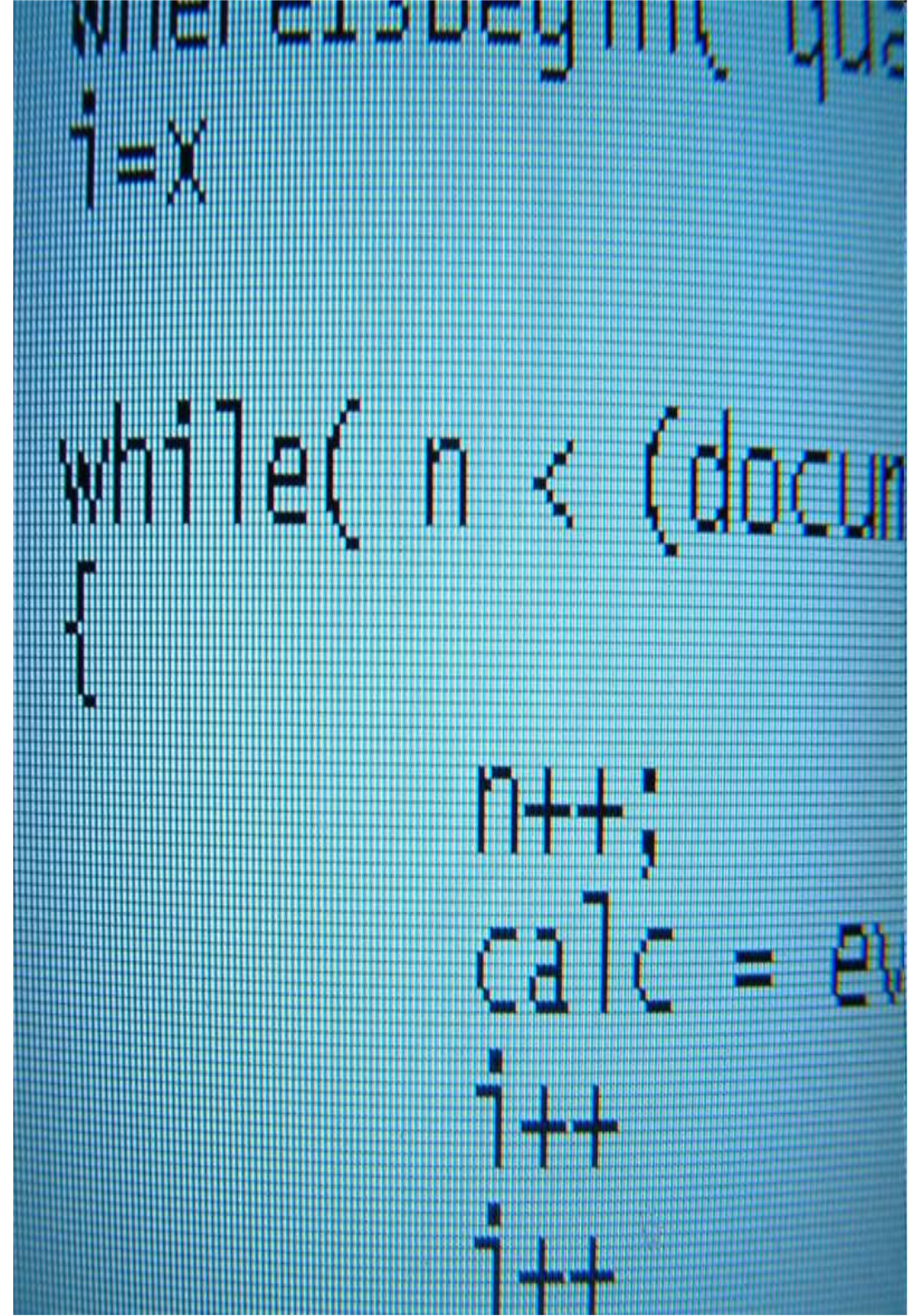


Cognitive OpenOCR (CuneiForm)



Fejlesztések

- előfeldolgozó és OCR konnektorok: cserélhető komponensek
- elérhetőek a migráció és az alkalmazás számára is
- adatbázis módosítások



Rendszerbe illesztés

- **feltöltött bitképek fogadása a szerkesztőfelületről**
- **előfeldolgozás**
- **OCR**
- **storage-re másolás**
- **részletes naplózás**

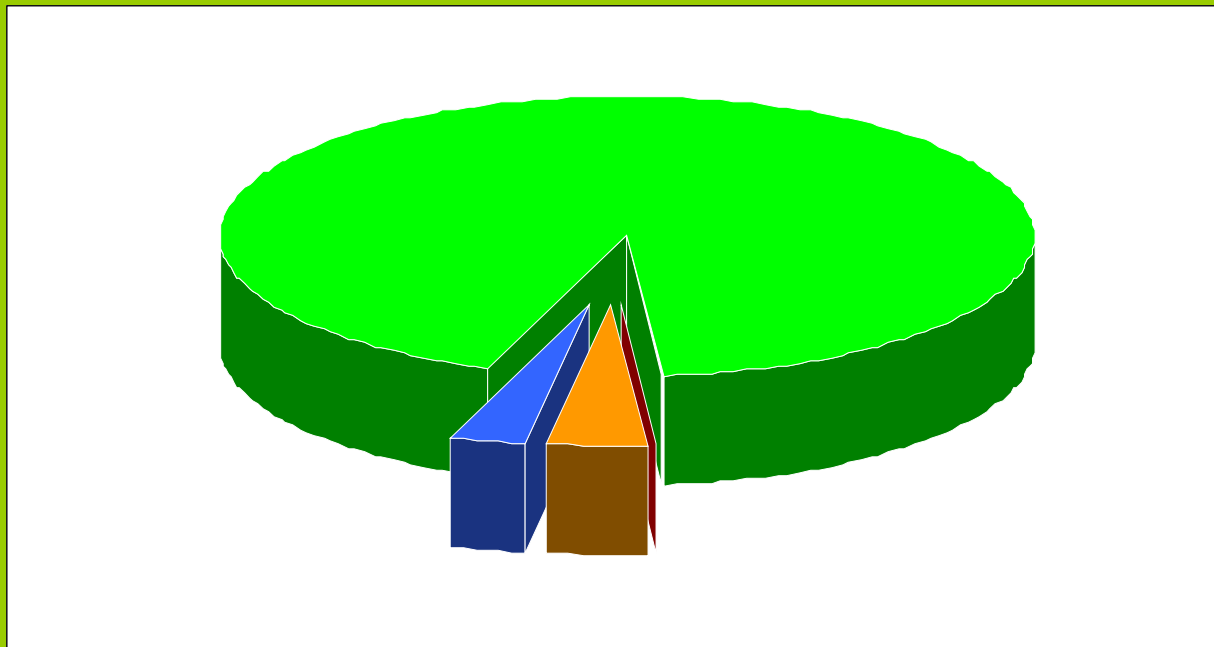


Feldolgozás

- **tesztkörnyezet kialakítása**
- **tesztfeldolgozások, javítások iterációval**
- **biztonsági mentés**
- **feldolgozás élesben**
- **kb. 1,9 GB**
- **10-12mp/dia**
- **6 nap**



A feldolgozás eredménye



- **93%** - sikeres feldolgozás
- **3%** - korábban feldolgozva (PDF, PPT, ODP)
- **3,9%** - üres kimenet
- **<0,1%** - szoftverhiba

A kereső továbbfejlesztése

Felhők átvonulása és naplemente

2 órás felhőátvonulás Budapestén

Felvétel ideje: 2009. július 7. Megtekintések: 68 Értékelés: ★★★★★

Találatok a fóliákon:



8mp



13mp



15mp



17mp



18mp



19mp



27mp



30mp

Tanulságok majdnem mindenkinek

- **Célhoz az eszközt**
- **Ingyenes OCR még mindig körülményes**
- **Ár és érték sokszor együtt jár**
- **Kompromisszumkészség, kitartás!**



**Köszönöm
a figyelmet!**



Turcsányi Tamás
tt@niif.hu

<http://videotorium.hu/>