



I. A felfedési kockázat mérése és a mikroadatokhoz való hozzáférés jövője

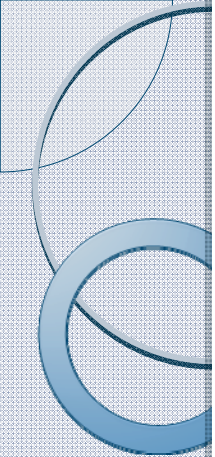
II. Paraadatok használata a rugalmas (responsive) mintavétel során

Vereczkei Zoltán

Központi Statisztikai Hivatal

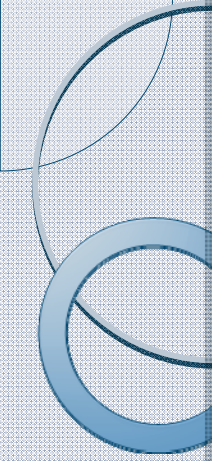
Módszertani főosztály

2012. január 24.



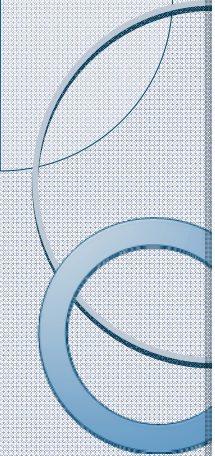
A felfedési kockázat mérése

- Felfedési kockázat mérése rendkívül fontos, mégsem könnyű feladat
- Elmúlt évtized kutatásai: eredmények általánosságban felhasználhatók vagy csak adott felvételre, adott körülmények között alkalmazhatók?
- Táblázatos, de különösképp mikroadatok kiadása előtt a lehetőségekhez mérten biztosítani kell, hogy az adott statisztikai egység(ek) felfedési kockázata rendkívül alacsony



Felfedési kockázat – A „nagy” kérdések

- Mi a felfedési kockázat elfogadható alacsony szintje?
- Lehetséges mérni?
- Mikroadatok esetén az azonosíthatóság megakadályozásával biztosan kiadhatók az adatok, különösen, ha „érzékenynek” számító változók is szerepelnek benne?
- Milyen külső adatfájlok érhetőek el, melyek a hozzáférhetővé tett adatokkal összekapcsolhatók?

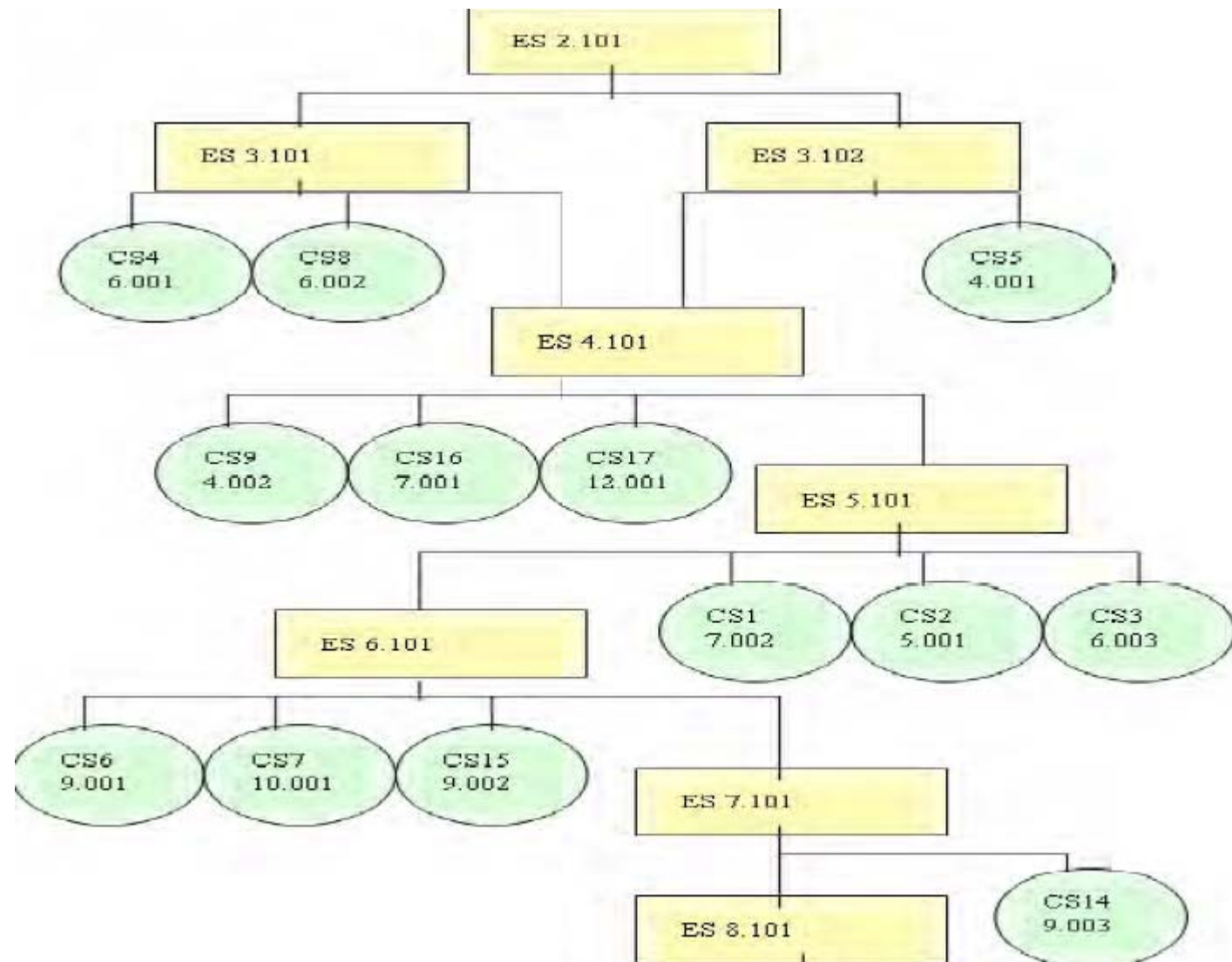


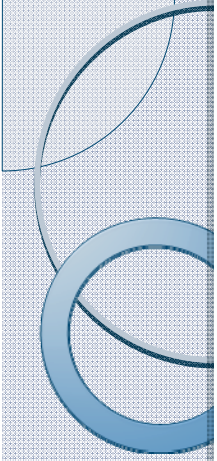
Angol példa (University of Manchester)

DEA: Data Environment Analysis

- Cél: kapcsolható adatbázisok feltérképezése, katalogizálása, kategorizálása és dokumentálása meta rendszerben
- Felfedési forgatókönyvek:
 - Fennáll a más állományokkal való kapcsolat veszélye?
 - Lehetséges egyáltalán más állományokkal kapcsolni (van potenciális állomány) és ha igen, pontosan hogyan (milyen változókat érint)?
- Feltételezések:
 - Egyedi adat felvétele esetén az adat egyedi rekordok formájában adatbázisban tárolásra kerül
 - Az adatokat olyan bontásban tárolják, ahogyan azokat összegyűjtötték

DEA: metaadat struktúra



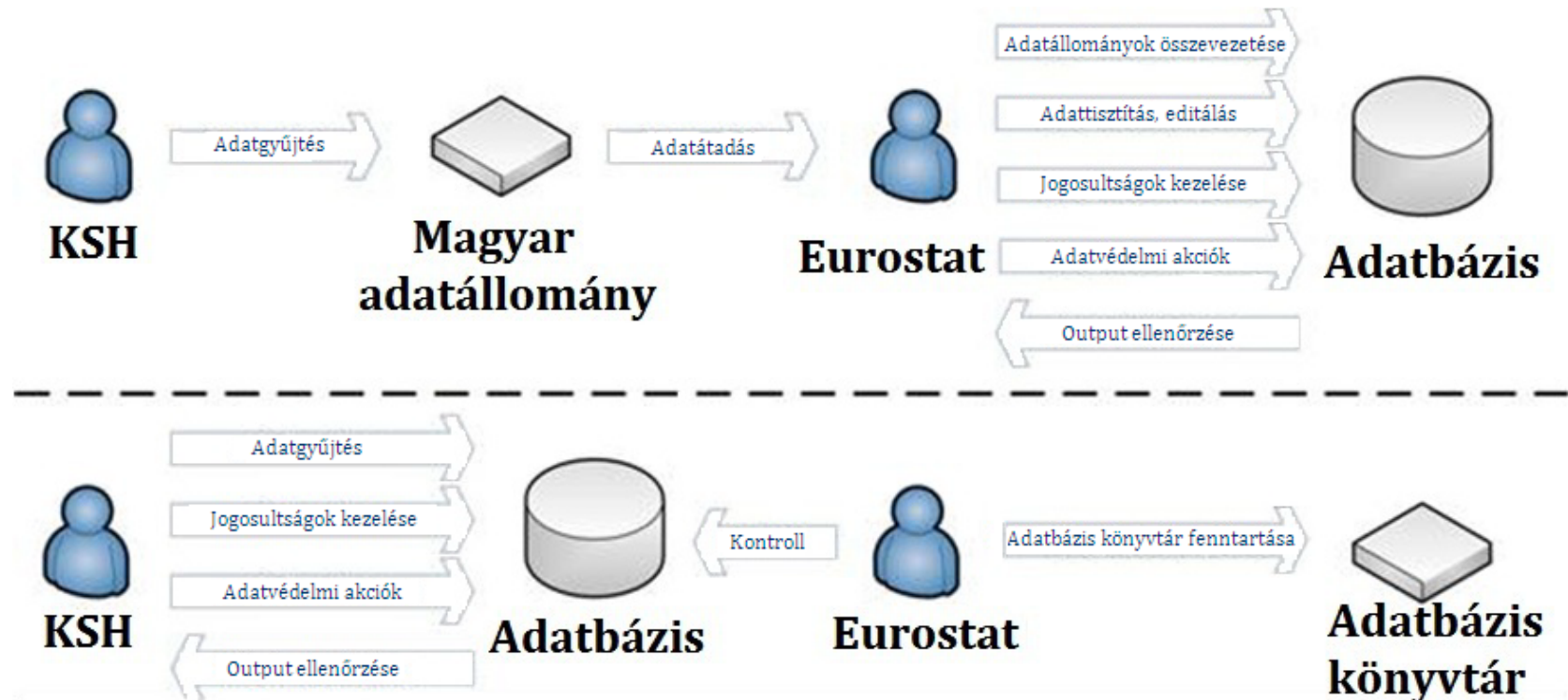


Mikroadatokhoz való hozzáférés jövője

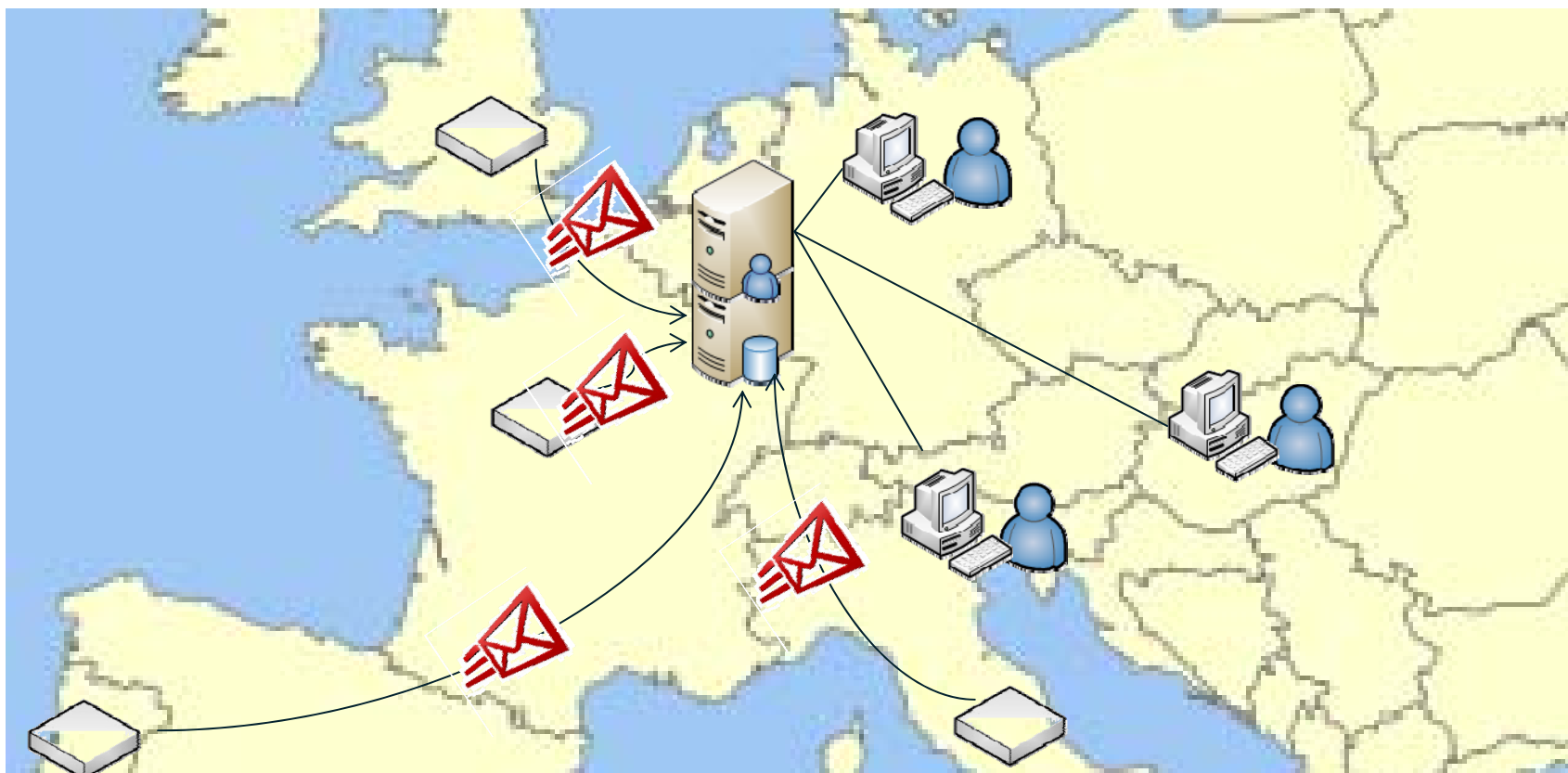
Jövő: **távoli hozzáférés**, nemzeti és európai mikroadatok széles köréhez

- Európai mikroadatokhoz való tudományos célú hozzáférés fejlődése
- Data without Boundaries
- Morpheus

Európai mikroadatokhoz való tudományos célú hozzáférés fejlődése – Eurostat szerepének változása



Európai mikroadatokhoz való tudományos célú hozzáférés fejlődése – Hálózatépítés (ESSnet DARA)



Data without Boundaries – Hozzáférés téma és mód szerint az európai országokban (2008)

	Bárki számára hozzáférhető mikroadatfájl (public use file)	Védett mikroadat-állományok kiadásra	Kivonatok (részminták)	Kötött szerkezetű táblázatos adatok	Táblázatos adatok egyedi kérésre	Távoli hozzáférés/ távoli végrehajtás	Kutatósobai hozzáférés
Népszámlálás							
Társadalom-statisztikai felvételek							
Természetes személyek regiszter adatai							
Gazdaság-statisztikai felvételek							
GSZR (business register)							
Egyéb							

□ Nincs

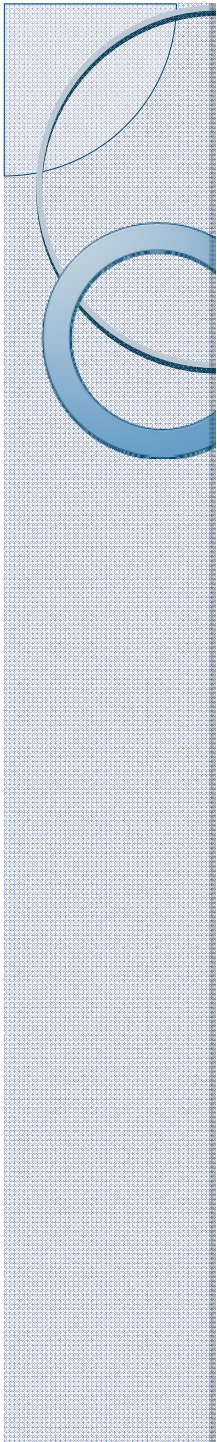
□ 1-3 ország

□ 4-6 ország

□ 7-9 ország

□ 10-12 ország

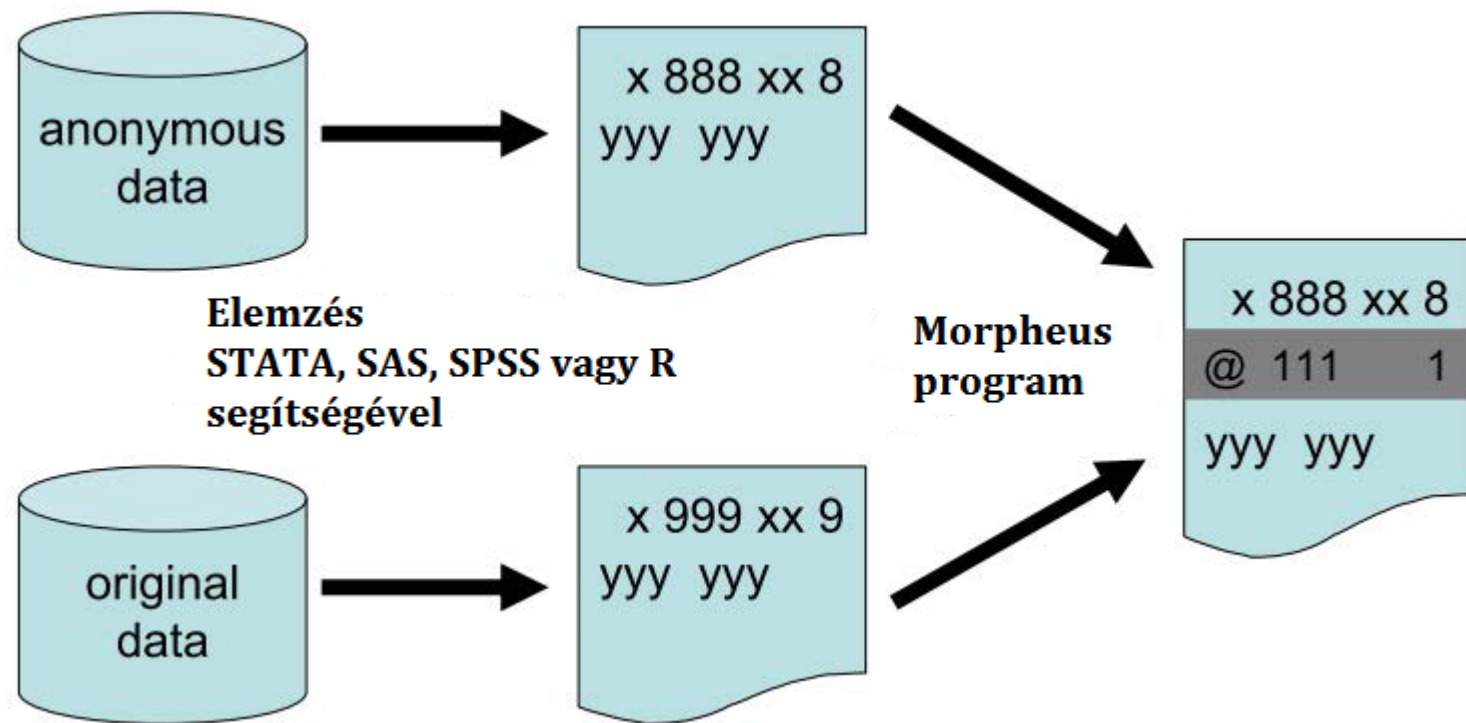
□ 13 ország vagy több

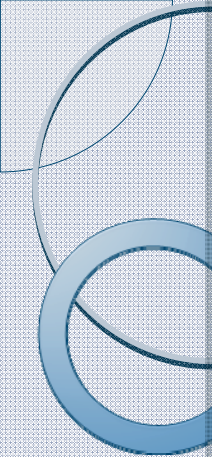


Data without Boundaries – Célok

- 1/3 nemzeti statisztikai hivatalok, 1/3 CESSDA adatarchívumok, 1/3 kutatóintézetek és egyetemek
- Pilot projektek indítása kutatószobai és távoli hozzáférés módokkal a nemzeti adatok határokon túli hozzáférhetőségének előremozdításához
- Technikai, jogi, gyakorlati megoldások tesztelése, melyeket tagországi és remélhetőleg EU-s szintre is kiterjesztenek majd
- Keretet ad az érintettek közötti stabil és rendszeres kommunikációra

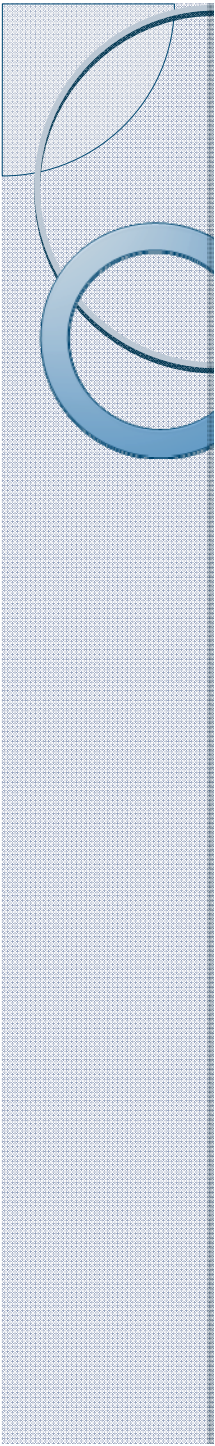
Morpheus – Működési modell (State Statistical Institute: Berlin-Brandenburg)





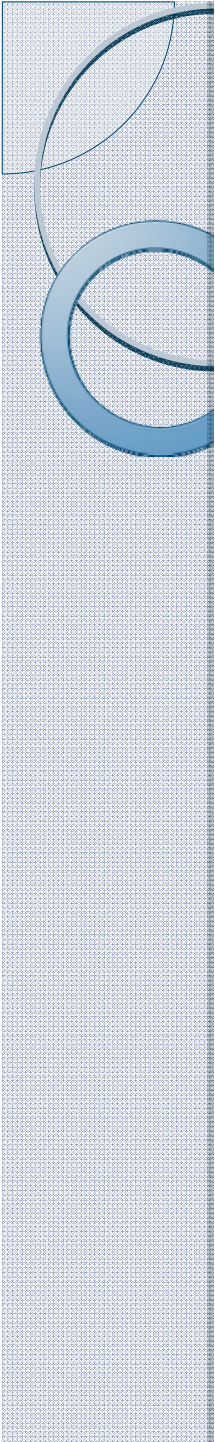
Morpheus – Konceptió

- Távoli hozzáféréseken alapul
- Felhasználók védett állományon végzik számításait
- A háttérben a számítások az eredeti mikroadatfájlon is futnak, amit a felhasználó nem lát
- Automatikus adatvédelmi szempontú ellenőrzés a védett állományon
- A kapott eredmények mellett megjelenik egy minőségmutató, mely az illeszkedés jóságát mutatja
- Az eredményeket (majdnem) valós időben kapja a kutató
- Az éles állományon végzendő akciók eredményét manuálisan kell adatvédelmi szempontból ellenőrizni



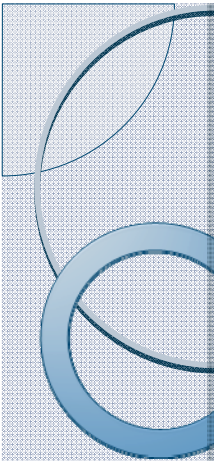
Paraadatok és responsive design

- Paraadat: kiegészítő információk, melyek az adatgyűjtési folyamat során állnak elő.
- Például:
 - Hívásadatok (automatikus rögzítéssel, összeíró által)
 - Összeíró megfigyelései (szomszédtól, kapcsolatfelvétel adatai, az összeírás körülményei)
 - Összeíró teljesítményadatai (ledolgozott órák száma, utazási távolság, stb.)
 - Rögzítési adatok (leütött billentyűk, adatjavítás, validáló üzenetek hatása, stb.)



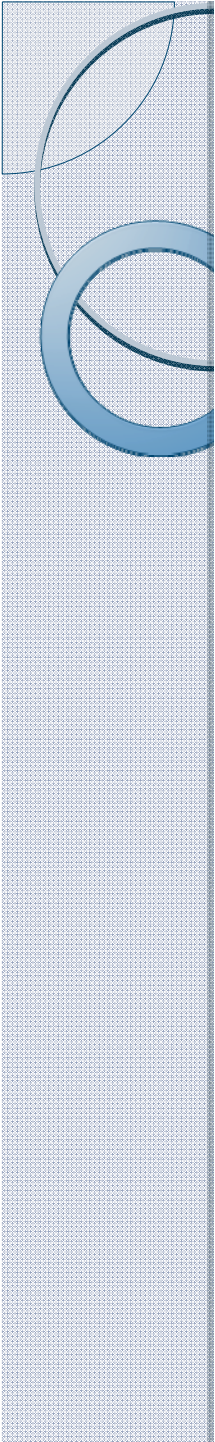
Responsive design fő elemei

- Kulcsváltozók figyelése paraadatok segítségével:
 - statisztikai folyamatirányítási módszerek alkalmazása a tevékenységek figyelésére
- Statisztikai szempontú megfontolások a design változtatásához:
 - koncentráltan egy vagy több mérhető eredményváltozóra
- Célzott beavatkozás a minta egyes részeire:
 - nem a teljes mintánál követendő stratégiát változtatja
- A döntési folyamat dokumentálása
- A beavatkozás sikerességének értékelése



NSFG: National Survey of Family Growth – Felvétel jellemzői (University of Michigan)

- Interjú két fő része:
 - *Előszűrő interjú* a lehetséges személyek beazonosítására (15-44 év közöttiek)
 - *Adatfelvétel személyes interjúval:* a kiválasztott személlyel
- Adatgyűjtés két fő része:
 - 1. fázis: adatgyűjtés 10 héten keresztül
 - 2. fázis: fennmaradó esetek részmintája



NSFG: Responsive design beavatkozások

Beavatkozás az 1. fázisban:

- A negyedév során az összeíró munkaidejének változtatása az előszűrő interjúkra, illetve adatfelvételre szánt idő módosításával
- „Kiemelt fontosságú” esetek beazonosítása és előre sorolása

Beavatkozás a 2. fázisban:

- „Kiemelt fontosságú” esetek kiválasztása a válaszadási arány növelésére és a torzítás csökkentésére

Mindkét beavatkozás során:

- Napi futású válaszadási modellek eredményei
- Kulcsváltozók napi szintű követése, felvétel céljaival, valamint előző negyedévi eredményekkel összevetve

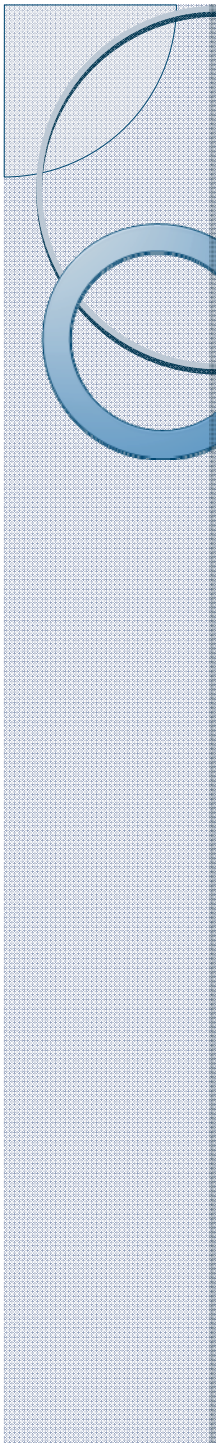
SCA: Survey of Consumer Attitudes – Felvétel jellemzői (University of Michigan)

- Telefonos felvétel
- Egyes hívások alapján a kapcsolatfelvétel valószínűségének mérése
- Kapcsolatfelvétel stratégiájának (hívás időzítésének) változtatása a tapasztalatok alapján
- 4 csoport kialakítása:

Csoport	Tartalma
1	SZO-V-H: 16.00-21.00
2	K-P: 17.00-21.00
3	SZO-V: 09.00-16.00
4	H: 09.00-16.00, K-P: 09.00-17.00

Table 1. Calls, Contacts, and Contact Rates by Experimental Group and Month

Month	Control			Experiment			Pr> χ^2
	Calls	Contacts	Contact Rate	Calls	Contacts	Contact Rate	
August	4,467	470	0.105	4,238	517	0.122	0.179
September	5,418	507	0.094	5,025	596	0.119	0.016
Combined	9,885	977	0.099	9,263	1,113	0.120	0.008



Köszönöm a figyelmet!