A véletlen, kiszámítás , információ és Lem másodfajú démona

Benczúr András

ELTE IK

ELTE eScience RET

A véletlen, kiszámítás és információ viszonya

Benczúr András

ELTE IK

ELTE eScience RET

A Természet Világa alapításának 140. évfordulója alkalmából 2009. szeptember 18.

Staniszlav Lem másodfajú démona

Trurl és Klapanciusz készít egy másodfajú démont Mohónak, az okleveles zsiványnak, aki a tudás kincseire vágyik.

„ megadjuk neked az információt az összinformációról, vagyis saját kezűleg készítünk neked egy Másodfajú Démont, amely mágikus, termodinamikus, antiklasszikus és statisztikus. Ez egy öreg hordóból, vagy akár egy köhintésből kivonja és átviszi neked az információt mindenről, ami csak volt, van, lesz és lehet.”

LEM: TRURL ÉS KLAPANCIUSZ HÉT UTAZÁSA, HATODIK UTAZÁS:

avagy hogyan épített Trurl és Klapanciusz másodfajú démont az okleveles zsivány számára

Staniszlav Lem másodfajú démona

„ Adj egy dobozt, mindegy, hogy mekkorát, csak jól zárjon; fúrjunk belé tűheggyel egy parányi lyukacskát, és a lyukba ültetjük a Démont; ott üldögél majd, és kizárólag az értelmes információkat engedi ki a dobozból, semmi egyebet. Mihelyt egy molekularaj véletlenül úgy áll össze, hogy jelent valamit, a Démon nyomban fülön csípi, és az értelmes információt briliánstűvel feljegyzi a papírszalagra, amelyből rengeteget kell neki odakészíteni, mert éjjel-nappal működni fog, amíg csak fennáll a kozmosz, és százmilliárd bitet ír le másodpercenként. Hát így működik a Másodfajú Démon; hamarosan meglátod!”

Rényi Alfréd egy kérdése

„Lehet egy vizsga nehézségét azzal jellemezni, hogy hány bit-et kell a hallgatóknak tudni? Enciklopédikus jellegű tárgyakban ez nem is teljesen abszurdum, a matematikában, persze, ennek nincs értelme, hiszen a dolgok egymásból következnek, aki az alapokat tudja, elvben mindent tud, illetve tudhatna. Egy matematikai elmélet összes eredménye tulajdonképpen csírájában benne van az axiómákban – vagy mégsem? Erről egyszer még gondolkodni fogok.”

(Rényi „Az információ matematikai fogalmáról” (Egy egyetemi hallgató naplója) Ars Mathematica, Rényi Alfréd összegyűjtött írásai, TYPOTEX, 2005.)

Az információ mérőszámai
Kiindulás: mit mérnek?

Valaminek a leírását adjuk meg, helyettesítünk jelekből álló leírásokkal.

Lehetőleg rövid leírásokat keresünk.

A leírásból kívánt pontossággal visszanyerhető legyen, amit helyettesít.

Az információ mérőszámai leírások hosszára vonatkoznak, az adott feladat szempontjából optimális megoldás hosszát adják meg.

Az információ mérőszámai

Kolmogorov: Három megközelítés

1. Valószínűségi: Shannon-entrópia

$$H(p_1, p_2, \ldots p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

2. Algoritmusos: Kolmogorov-entrópia

$C(x) = C_U(x) = \min\{l(p) \mid U(p) = x\}$, és ∞, ha nincs ilyen p.

A definícióban U a rögzített referencia függvény,

tipikusan az univerzális Turing - gépet választják.

3. Kombinatórikus: azonos hosszú kód a halmaz minden elemére

A feltételes Komogorov-entrópia

Definíció:

$$C(x \mid y) = C_U(x \mid y) = \min\{l(p) \mid U(p, y) = x\}, \text{ és } \infty, \text{ ha nincs ilyen } p.$$

A definícióban U a rögzített kétváltozós referencia függvény.

Prefix-változat:

Az U által használt kódok prefixmentes rendszert alkotnak.

A megfelelő entrópiák jelölése:

$$K(x), \text{ és } K(x \mid y).$$

Az algoritmikus entrópiák konstans erejéig egyértelműek.

Játék a kombinatórikus entrópiával
n kocka felcímkézése

Azonos hosszú bináris kódot kell ragasztani minden kockára:

$n \log_2 n$ bit összesen.

(Ennél kisebb összhossz prefixmentes kódokkal nem érhető el.)

Legyenek színes kockáink, és színenként kell azonos hosszú kódot ragasztani:

$\sum_{i=1}^{k} n_i \log_2 n_i$ bitre van szükség,

ahol k a színek száma, és

az i - edik színnel n_i számú kocka rendelkezik.

Mennyi bitet nyertünk:

$$n \sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n}{n_i} = nH\left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots \frac{n_k}{n}\right)$$

Az egyenletesség szerepe

A halmaz szerinti prefix entrópia

A kombinatórikus entrópia – az egyenletes eloszlás optimális kódja - miatt a halmaz szerinti feltételes prefix entrópiára teljesül a következő egyenlőtlenség:

Legyen S egy m elemű, természetes számokból álló halmaz. Ekkor:

$$\sum_{x \in S} K(x \mid S) \geq m \log_2 m$$

Az egyenletes hosszúságú kódnál átlagosan nem jobb a feltételes optimális kód.

- Honnan következtethetünk a jövő eloszlására? Hogyan befolyásolhatjuk a jövő eloszlását?
- A múltra vonatkozó ismereteinket kell felhasználni. Ez az ismeret most a világháló adatbázisában gyűlik.
- A múlt leírásának tömöríthetősége: a Kolmogorov entrópia világa
- A Shannon entrópia: a jövő leírásának tömörségére ad mérőszámot: a jövő lehetséges kimeneteleinek eloszlása ismeretében mi a megkülönböztető leírások hosszának várható értéke? Ennek alsó határa a Shannon entrópia. (felkészülés a jövőre)
- Kétrészes kód: törvényszerűség leírása és a maradó véletlen, tipikus adat (jelentése a jövőre nézve)

- A feltételes Kolmogorov bonyolultság alkalmazása, a kétrészes kód:
- a) Adott elemet tartalmazó egyszerűen leírható halmaz, abban tipikus elem
- b) Kevés elemszámú, egyszerűen leírható halmaz
- A halmaz lehet például egy valószínűség-eloszlásból származó tipikus minták halmaza (Bernoulli eloszlás)
- Példa: 2n szögpontú, q,p paraméterű reguláris páros gráf

Az információ fogalma

Idegen szavak szótárából:

- Információ:
    1. felvilágosítás, tudósítás, tájékoztatás, hírközlés
    2. értesülés, adat
    3. tudósítási, tájékoztatási anyag, hír
- Informál: tájékoztat, felvilágosít, tudósít

Saját megközelítésem

- Visszatérve az információ fogalmára, a három matematikai mérőszám behatárol egy fogalmat: az információ az, aminek a mennyiségét mérik. Más kontextusban használt információra ezek a mérőszámok felelőtlenül nem használhatók.
- *Szerintem az információ fogalma történetileg érthető,*
- Kezdet: Az **információ** az élettelen és élő határvonalának egyik oldalán: az **élőnél** jelenik meg. Az élettelenben az információ értelmezhetetlen.

Saját megközelítés

- Az élő az információ reprezentálására nem csupán belső lehetőségeit használja, hanem az élettelent is felhasználja reprezentáció céljából. Hibás az a fogalomalkotás, ami ezt a reprezentációt az élettelen információjaként tünteti fel. Majd a kommunikációnál térünk vissza erre.
- Az információ a múltra vonatkozik, és a jövőre való felkészülésben hasznosul. Az eredeti latin szó egyik értelmezése sem vonatkozik a jövőre. A jövőről nincs tudósítás, a jövőre lehetnek előrejelzések, akár teljes biztosak is, de lehetnek csupán elképzelések is, akár minden esély nélkül.
- Valakinek a jövőre vonatkozó elképzelései is csak a múltról tudósítanak: egy múltban lezajlott gondolkodás eredményéről.

Információ

Az élet, az élő egyed gyűjt információt, átad, örökít információt.

Miben, hol jelenik meg **először** az információ? A sejten belül, majd az élő szövet szerveződéseiben és mozgásaiban.

Miből keletkezik az információ: a fizikai-kémiai hatásokból, az érzékelésekből. Hogyan hat az információ: érzékelődik. Ami érzékelődik, az mozgás (változás), és mozgást (változást) tud kiváltani. Az a mozgás megint érzékelődik is. És így tovább. **Van benne eleve rekurzió.**

Információ

Az állatvilág egyik jellemzője a célirányos mechanikai mozgás, helyváltoztatás. Ehhez a mozgás vezérlésében, koordinálásában igen nagy teljesítményű információ-feldolgozás, igen nagy ”kiszámítási” teljesítmény szükséges. Az idegrendszer méretezése ehhez a csúcsteljesítményhez igazodik. Amikor nincs mozgás, ez a kapacitás kihasználatlan, előfordulhat, hogy más információfeldolgozást is végezhet. Akaratlagossá is válhat ez a tevékenység, bizonyos mértékben irányíthatóvá, ami a gondolkodáshoz vezet.

Ennyi dióhéjban a biológiai individuumok belső „informatikája”.

Információ

A következő informatikai szint az élővilágban az egyedek közötti kommunikáció. Ennek kiinduló elve az **idegrendszerek függetlensége**. Ez alatt azt értem, hogy egyik élőlény a másik belső információreprezentációit és folyamatait közvetlenül nem ismerheti. Nincs az idegrendszereknek közvetlen együttműködése. Az egymásra hatás csak közbeiktatott közeg változásain keresztül történhet. Ez a kommunikáció.

Eddigi írásaim a Természet Világában innen indultak: a kommunikációból.

A kommunikáció fejlődése

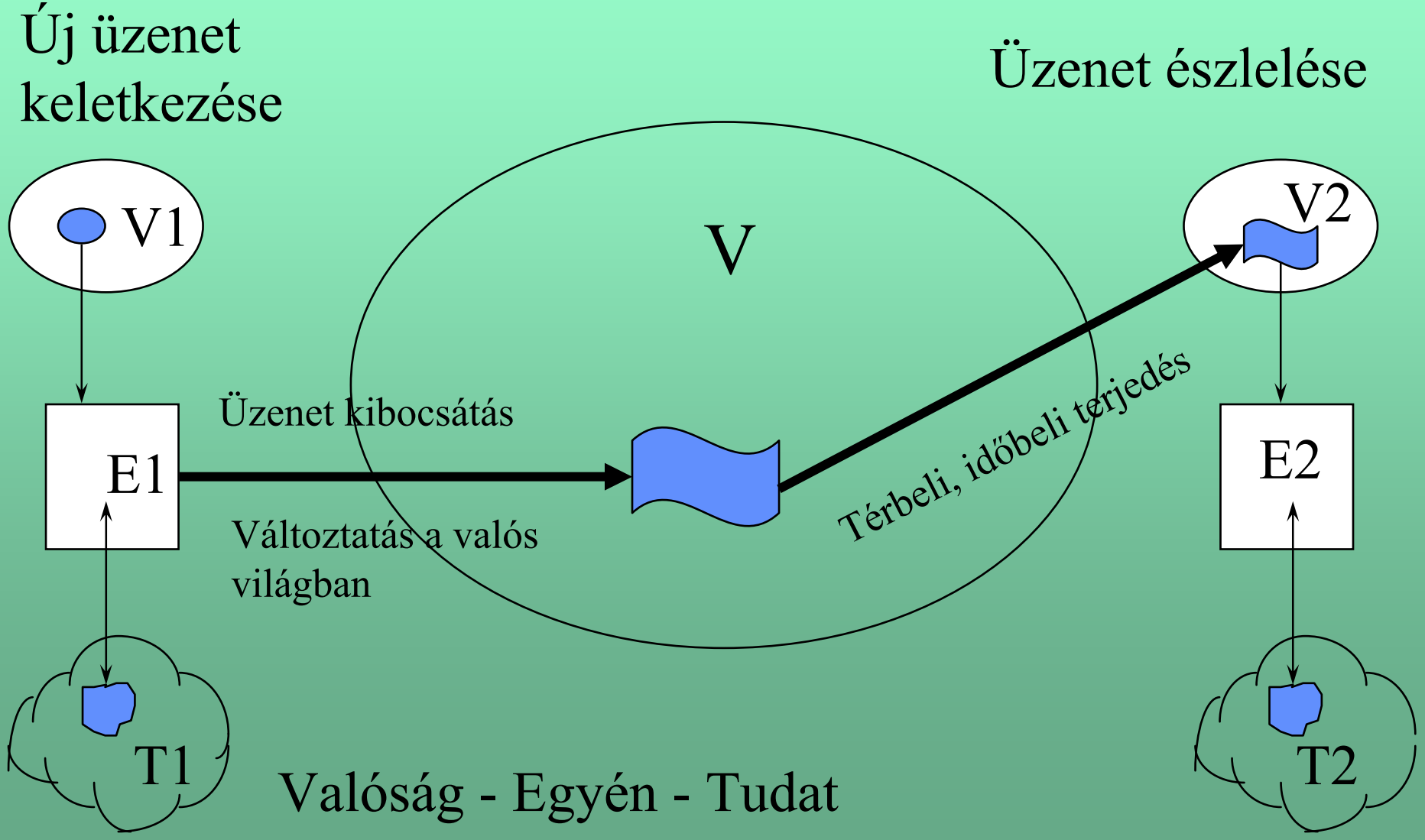
Fő állomások:

- beszéd
- írás, nyomtatás
- számítástechnika, telekommunikáció, multimédia, Internet, WWW, távjelenlét

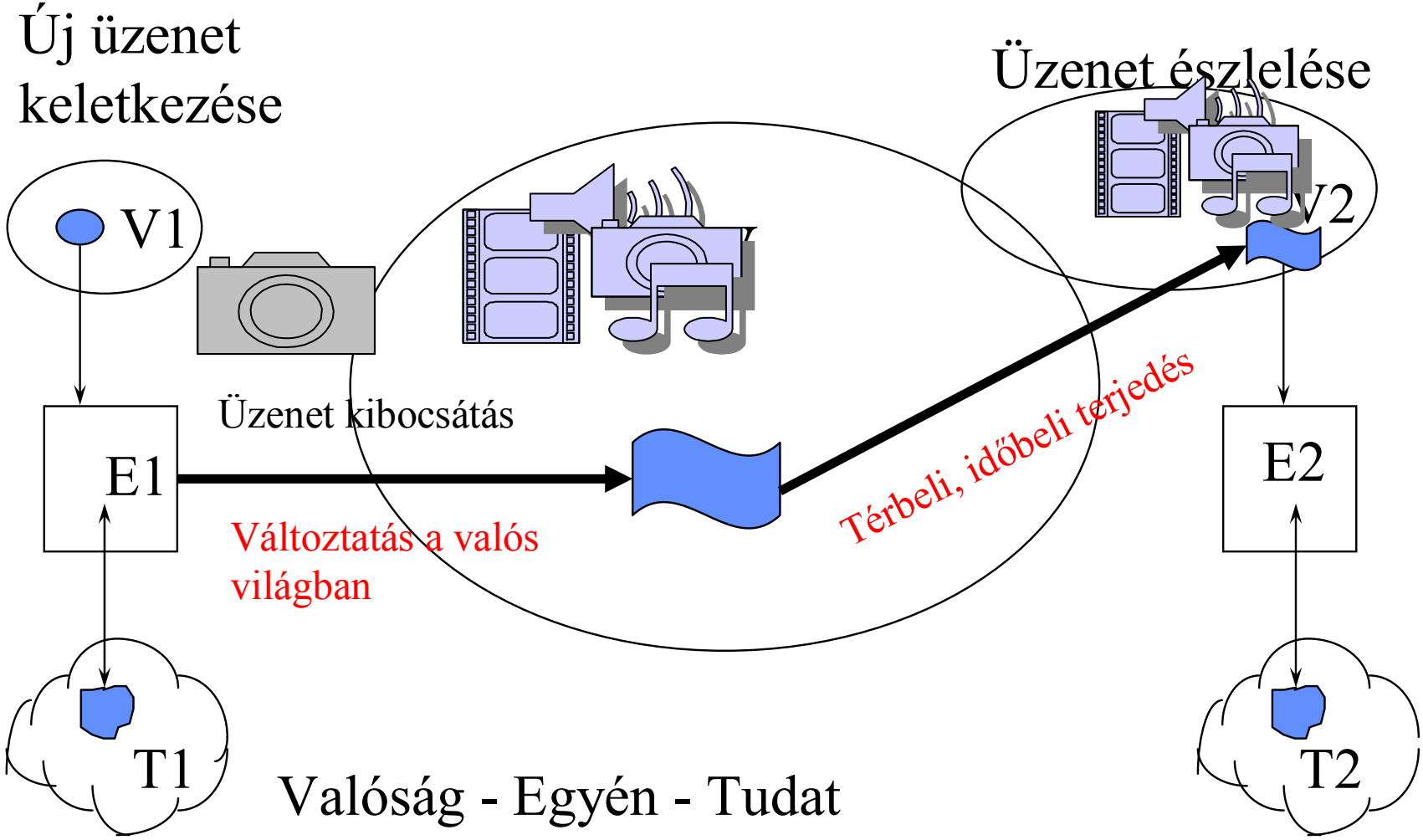
Hatásuk:

- nyelv, fogalomrendszer, egyeztetett alapismeretek
- a tudaton kívüli ismerettárolás, feljegyzés jövő számára, többszörözés
- tér és időkorlátok feloldása, tárolási, **elérési és feldolgozási** lehetőségek exponenciális ütemű növekedése, **intelligens közeg** épül az ember-ember és ember-természet interakció közé

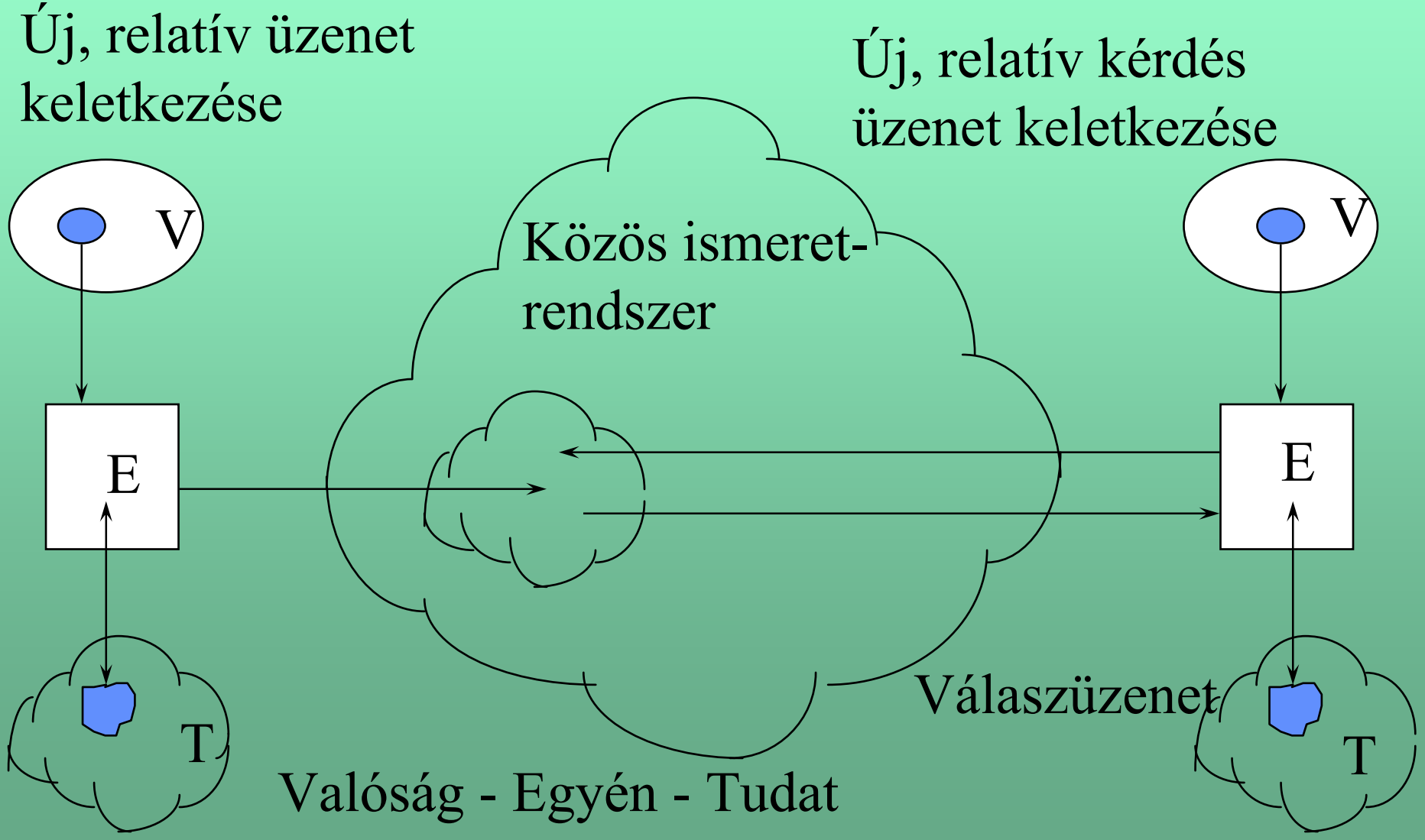
Kommunikáció
Információ átadás



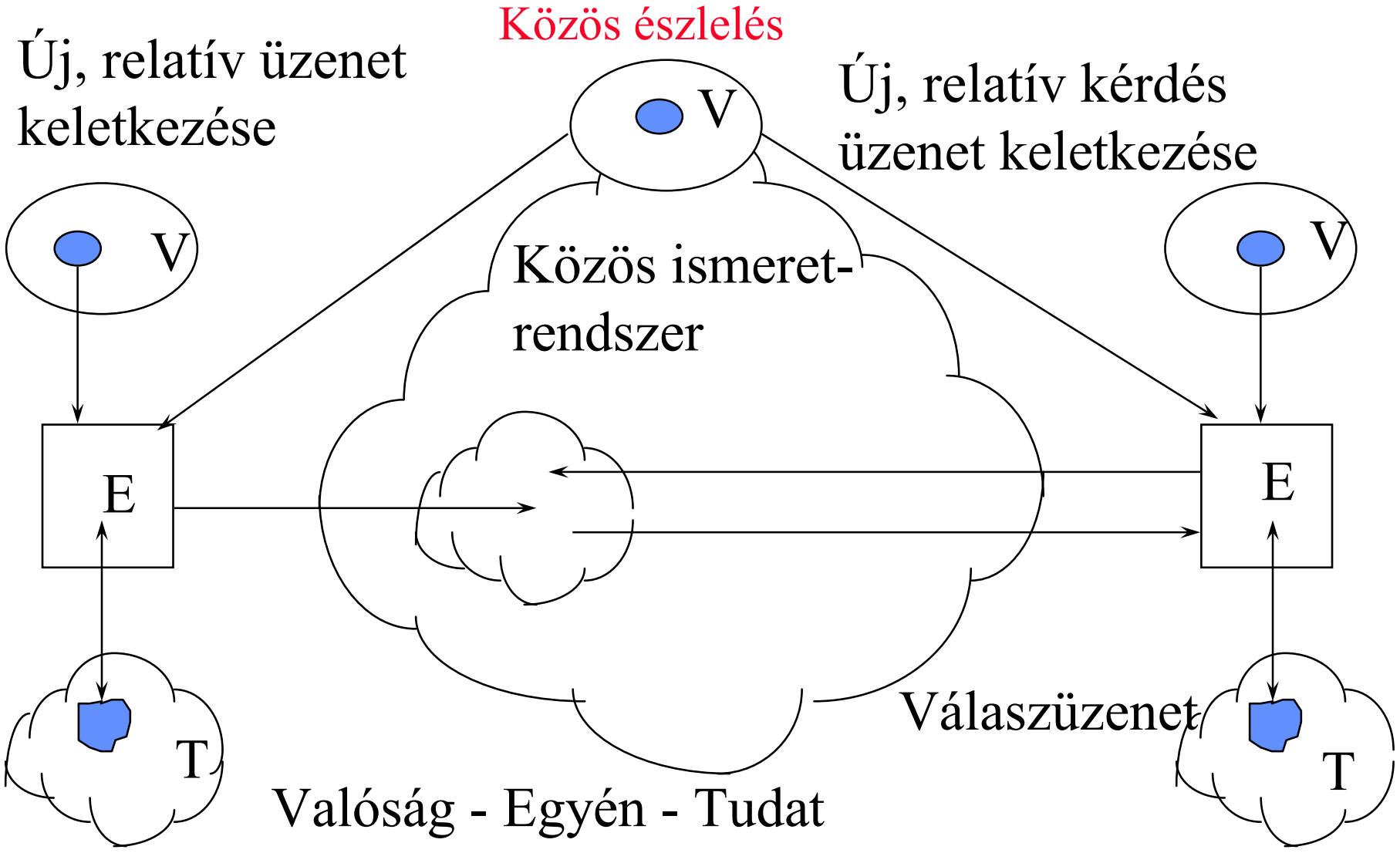
Új hangsúly : instrumentális elemi észlelés, élménymegosztás, vizuális üzenetek, tetszőleges multimédia forrás használható



Információs rendszer

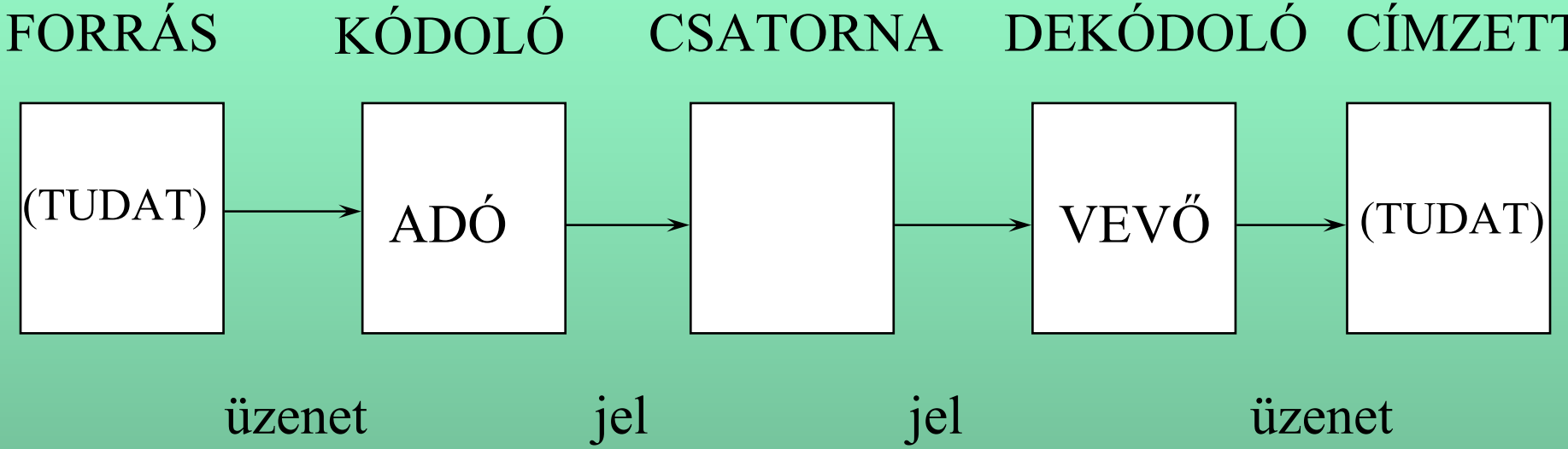


Információs rendszer



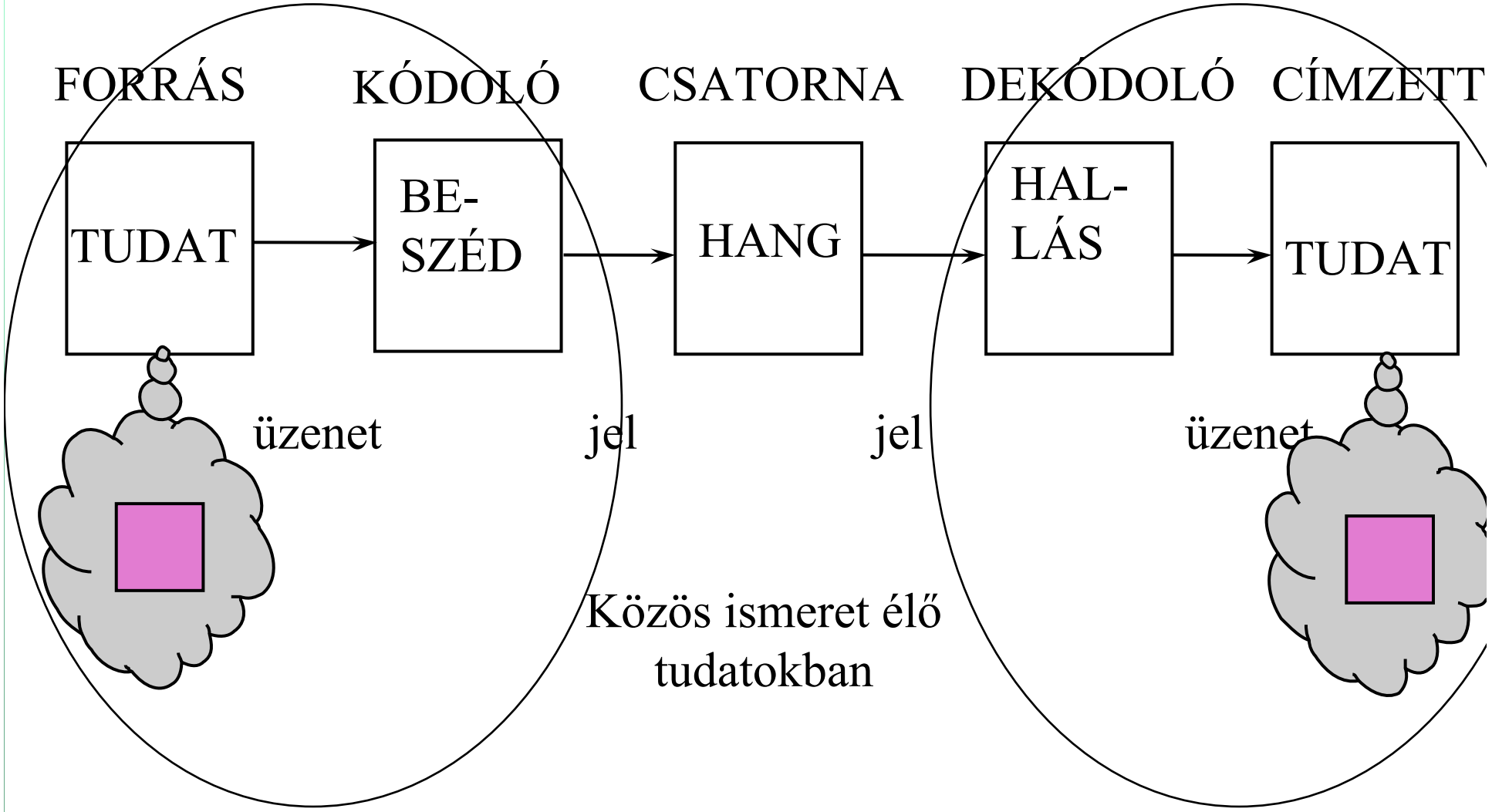
Shannon kommunikációs modellje

„A kommunikáció felöleli mindazokat az eljárásokat, amelyeken keresztül az egyik emberi elme a másikra hatni képes.” (W. Weaver)

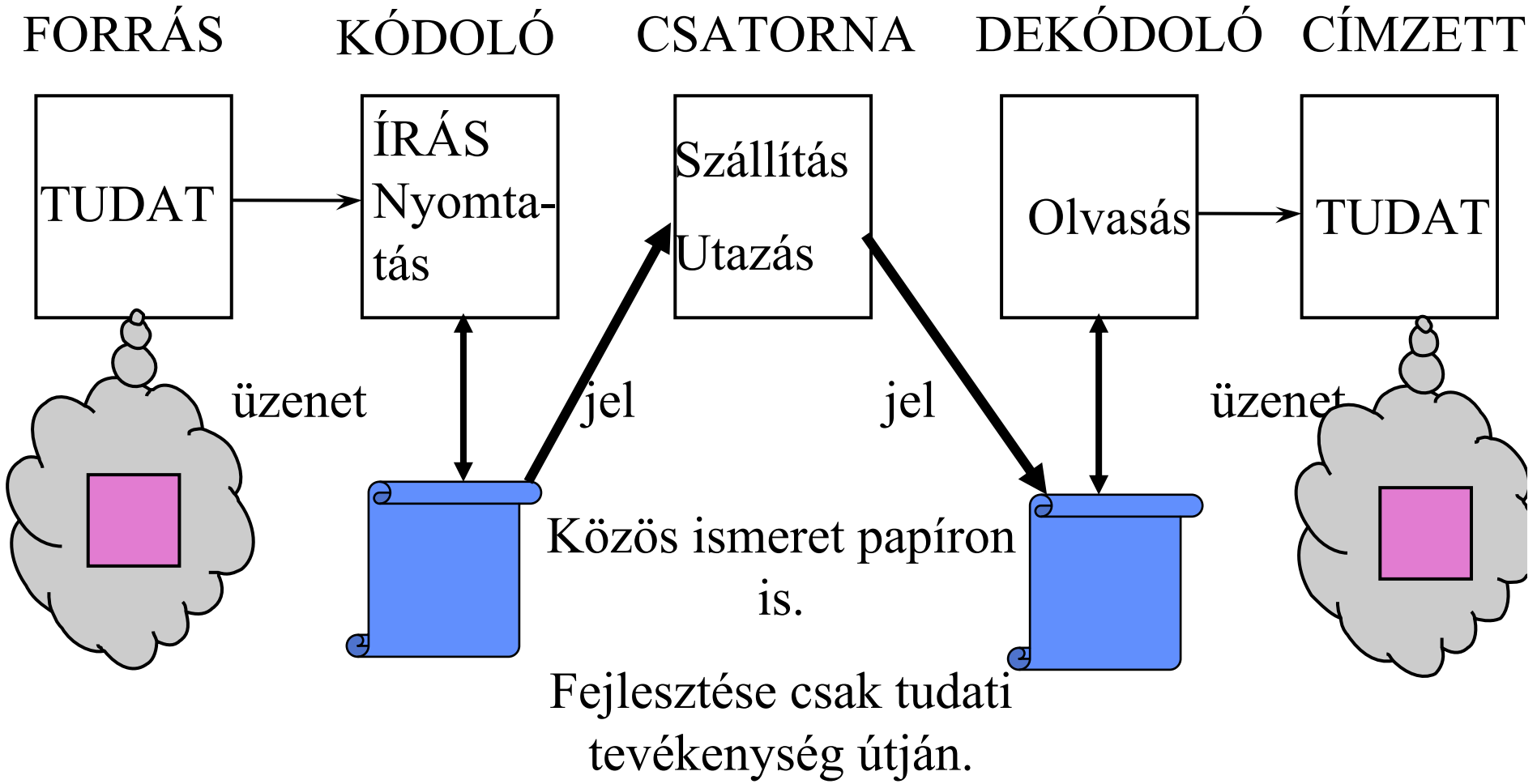


mennyiségi szint (entrópia és csatorna kapacitás)
megértési szint (jelentés, szemantika)
hatékonysági szint (a kívánt hatás elérése)

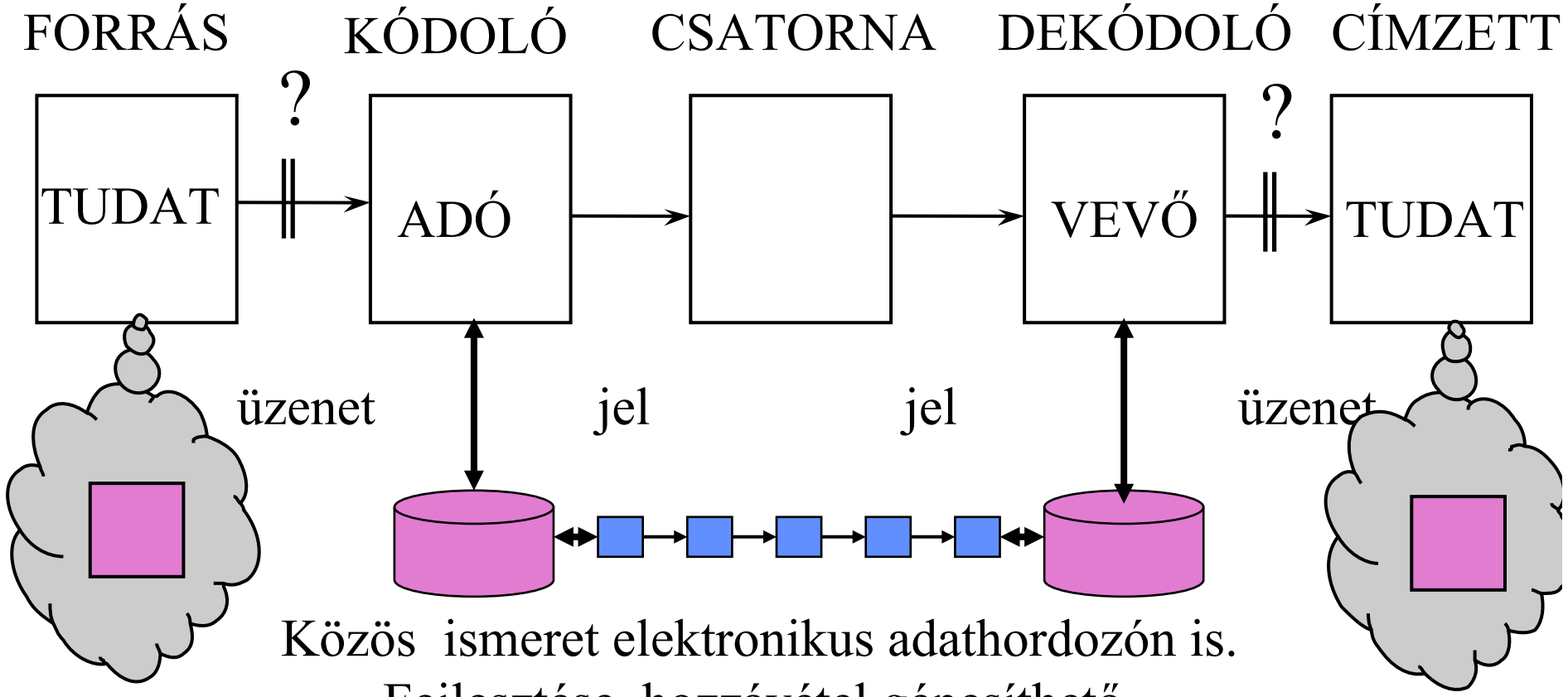
Beszéd


FORRÁS
KÓDOLÓ
CSATORNA
DEKÓDOLÓ
CÍMZETT
TUDAT
BE-
SZÉD
HANG
HAL-
LÁS
TUDAT
üzenet
jel
jel
üzenet
Közös ismeret élő
tudatokban

Írás


FORRÁS
KÓDOLÓ
CSATORNA
DEKÓDOLÓ
CÍMZETT
TUDAT
ÍRÁS
Nyomta-
tás
Szállítás
Utazás
Olvasás
TUDAT
üzenet
jel
jel
üzenet
Közös ismeret papíron is.
Fejlesztése csak tudati tevékenység útján.

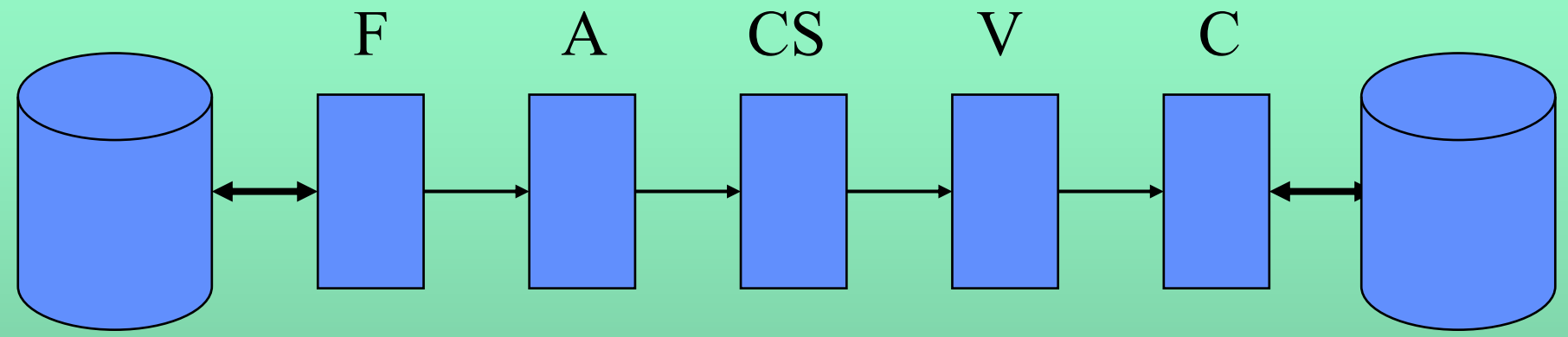
SZÁMÍTÁSTECHNIKA



Közös ismeret elektronikus adathordozón is.
Fejlesztése, hozzávétel gépesíthető.

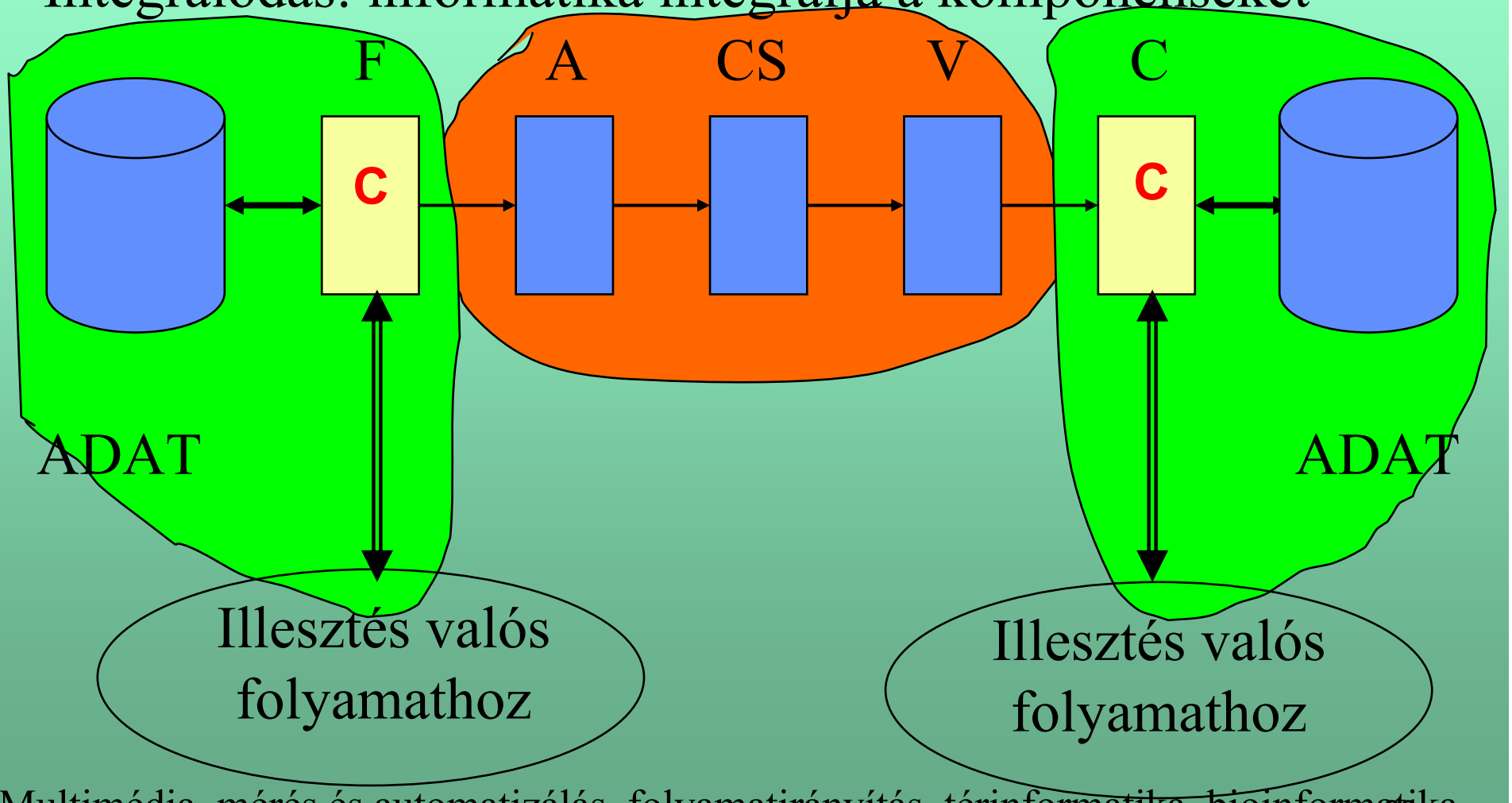
KISZÁMÍTÁS

ADATÁTVITEL



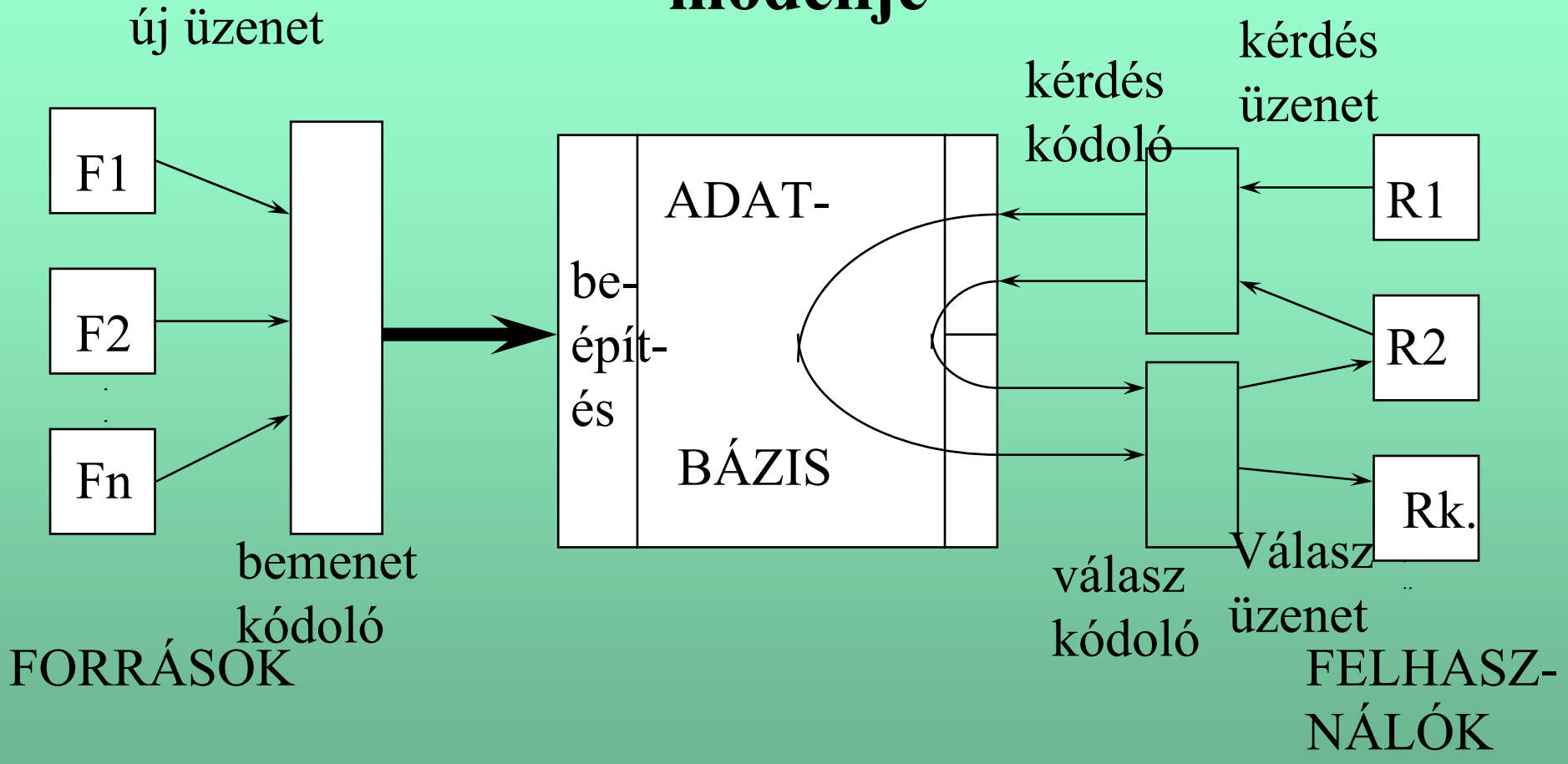
Konvergencia: informatika-távközlés-média

Integrálódás: informatika integrálja a komponenseket



Multimédia, mérés és automatizálás, folyamatirányítás, térinformatika, bioinformatika, e-világ, CAD, CAM, ...

Az automatizált információs rendszer modellje



A kódolók előre rögzített formájú, speciális nyelven írt üzeneteket fogadnak, ehhez intelligens interaktív felületet adnak a források és felhasználók számára.

Az emberi észlelés , tudatosulás mellett megjelent az instrumentális észlelés, ami nyomtatás, vizuális és audió- kijelzés, elektronikus rögzítés majd megjelenítés közvetítésével válik emberileg észlelhetővé, egészen a távkörnyezet megoldásokig. (Kollaboratív virtuális környezet, távjelenlét, teleimmersion.)

Az elektronikus formák átalakítása és terjedése sok fázison mehet keresztül emberi észlelés nélkül. Az irány fordítva is működik, a jelek felerősíthetők fizikai változásokká. A valós folyamatok bekapcsolhatók a kommunikáció folyamatába.

Ami észlelés, üzenet egyszer adattá, jellé válik, az utána önálló életet nyer. Új modellező nyelvek fejlődnek hozzá. A modell helyettesíti a valóságot, feldolgozhatóvá teszi az észleléseket.

A véletlen és kiszámítás

Információra nem lenne szükség véletlen nélkül, az információ nem lenne használható kiszámítás nélkül. Nem lenne semmi szerepe. A véletlen matematikai modelljét a jövő leírásának eszközeként használhatjuk, mint a jövő bizonytalanságának, kiszámíthatatlanságának leírási módszerét.

A múlt homályos leírása és a múlt véletlen gyökerei, valamint a jövő bizonytalanságát modellező valószínűség-eloszlásra való következtetéseink hogyan hozhatók össze?

A jövőre csak a múltból következtethetünk.

Beck Mihály a „Parajelenségek és paratudományok” c. könyv 67-oldalán: „ A különböző folyamatok időbeli lejátszódásának leírása valójában csak akkor lehetséges, ha a jövő nem más, mint *megismételt múlt*.”

Rényi Alfréd egy kérdése

„Lehet egy vizsga nehézségét azzal jellemezni, hogy hány bit-et kell a hallgatóknak tudni? Enciklopédikus jellegű tárgyakban ez nem is teljesen abszurdum, a matematikában, persze, ennek nincs értelme, hiszen a dolgok egymásból következnek, aki az alapokat tudja, elvben mindent tud, illetve tudhatna. Egy matematikai elmélet összes eredménye tulajdonképpen csírájában benne van az axiómákban – vagy mégsem? Erről egyszer még gondolkodni fogok.”

(Rényi „Az információ matematikai fogalmáról” (Egy egyetemi hallgató naplója) Ars Mathematica, Rényi Alfréd összegyűjtött írásai, TYPOTEX, 2005.)

A válasz az algoritmikus információelméletben van.

Egy matematikai elmélet összes (bizonyítható) eredménye felsorolható, csak győzzük kivárni, mikor érünk el az éppen kérdezett eredményig. A feltételes Kolmogorov-entrópia mutatja, hogy adott bitnyi ismeret matematikai tétel esetén kevesebb kiegészítéssel vezet el egy kérdés válaszáig, mint történelmi tétel esetén, másként mondva, kevesebb újonnan megtanult bit feldolgozása után érünk el a tételhez. Azonban a feldolgozás sebessége sem elhanyagolható. Sőt, a felhasználható tárméret sem. Fejben más a határ, mint papír-ceruzával.

Paradoxon

Paradox módon az algoritmusos információelmélet szerint a matematika egyszerűbb, mint a történelem. Intuitív magyarázata ennek az állításnak az, hogy ugyanannyi mennyiségű írásos válasz a vizsgán (matematikából a definíciók, tételek pontos megfogalmazását és a tételek bizonyításának leírását értve válasz alatt) kevesebb memorizálandót jelent matematikából, mint történelemből. Ugyanígy magyarázható, miért egyszerűbb a vers, mint a próza megtanulása szó pontossággal. A matematika kevésbé véletlenszerű, mint a történelem. (Még irritálóbb állítás, hogy az élő anyag egyszerűbb, mint az élettelen. Ennek kifejtése a genetika-genomika világába vezet már. A genetikai kódnak a bioszféra gépezetébe helyezésével történő kiszámításként elképzelve az egyed kifejlődését.)

enciklopédikus tudás

Mi az **enciklopédikus tudás jellemzője**: Az M eléréséhez elsősorban indexelés kell, és elért szekvenciák visszaadása. (Az asszociatív emberi visszakeresés is sajátos indexelés. Ebben is nagy a gyakorlás szerepe, és a korábban, más céllal megtanultak szerepe.)

matematikai tudás

Az axiómarendszerre fogalmak és erős szerkezetek épülnek. A szerkezetek egy része definíció és tétel jellegű, vagyis az axiómákból következő igaz állítások megfogalmazását könnyítő definíciók és utána tételek megfogalmazása. A tételek igazolása algoritmusokkal történik. A kettő erős együttműködése tesz lehetővé kevés új bit megtanulásával nagyobb visszaadható tudást. A tételek generatív megfogalmazhatósága (szabályok alapján állíthatjuk elő tételek sorát) rávezet a tételek megfogalmazására pontatlan tudás esetén is, ugyanígy a bizonyítások is generatív-felsorolható módon rekonstruálhatók.

Azonban a rekonstruáláshoz igen gyors számítási teljesítményre van szükség. Gondolkodni és érteni kell! A gondolkodást gyakorolni kell. **(Konfuciustól idézet emlékezetből: Tanulni csak gondolkodva érdemes. Gondolkodni tanulás nélkül veszélyes.)** A tételek generatív előállítása az elméletek formális nyelvként való megadhatóságára épül. A tételekből egy elég sűrű részhalmaz tudása alapján könnyű kitölteni sok további tételt. Ehhez viszont jól kell ismerni a nyelvet. A tételek és bizonyítások szövedéke újabb tételekhez vezet. A matematikai vizsgán a „ritka” tételtudás esetén gyors generatív, „kiszámító” gondolkodás is elég lehet, a sűrűbb tételhalmaz tudása viszont gyengébb gondolkodási képesség mellett is elegendő lehet.

. A matematika múltja igen sok lehetséges ág bejárását jelenti. Ez is a megismert múlt része, a megismert gondolatok múltja. Ebből építkezünk tovább. A matematika maga – lehetséges axiómarendszerei, tételei – nem változik. Az változik, hogy ebből emberi tudatok mit jártak be. A matematikusok teremtett világai ebben a legnagyobb kirándulások. A teremtés új és új terminális és grammatikai elemeket hoz be. Új nyelveket kell tanulni. Egyre többet lehet bennük megfogalmazni. **A megfogalmazottak (megismert tételek) mennyisége eléri-e az emberöltő által továbbdolgozható volument? Osztott, párhuzamos és ellenőrizhető matematikai gondolkodás jön-e majd létre?**

Informatikai tudásanyag

Az informatika maga, és ezért ismeretanyaga változik, és igen gyorsan. (Szemben a természettudományokkal, ahol nem a természet változik, hanem ismereteink és észlelési lehetőségeink bővülnek.) A változást emberi tudatok bejárása eredményezi. Teremtett világ. Vannak benne matematikai jellegű részvilágok, és tele van heurisztikákkal is. Az eredmény: programok, technológiák, materializált gépek. Utána azonban gépek folytatják a lehetséges „tételek” bejárást. Ezekről a bejárásokról megint emberi tudatnak kell tételeket fogalmazni. Mire a tétel kész, már lehet, hogy más bejárások (programok, technológiák) leváltották a tétel tárgyát képező bejárásokat. (A bejárás itt tulajdonképpen kiszámításokat jelent, vagy emberi tudat által végzettet, vagy gép által végzettet.)

A forrás véletlensége - matematikai idealizált modell. A valós szituációt egy bizonytalanabb, nagyobb entrópiájú matematikai jelenséggel, sztochasztikus folyamattal helyettesítjük. A valós forrás kimenetei ezért szabályosabbak és kevesebben vannak, mint a tipikus véletlen halmaz.

A nagyon hosszú, vagy kiterjedt konkrétan előforduló folyamatot (realizációt) önmagában is nézhetjük, eloszlás nélkül, tömöríthetőség szempontjából.

Ez átvezet az algoritmikus jellemzés világába, a nagy adatállományok tömörítésének kérdéseihez.

Korábban voltak a könyvek, oda ömlött be minden tudásunk (helyesek és hibásak, jó szándékúak és rossz szándékúak.)
Vörösmarty: „Gondolatok a könyvtárban”

Az USA Kongresszusi Könyvtár: 28 millió könyv. Teljes digitalizálása: 10-100 Mbyte/kötet:összesen 280-2800 TeraByte.

Most a világháló adatbázisába ömlik minden. Becsült mérete Zetabyte tartományban van. Másfél évenként kétszereződik , gyorsul.

Exabyte: az 1999-ben keletkezett információ (adattömeg) fele.

IDC tanulmány szerint közel fél Zetabyte, pontosabban
3 892 179 868 480 350 000 000
bitnyi információ keletkezett 2008-ban,
2009-ben ezerszer annyi várható, mint 1999-ben.

Programkódok

Mennyi az elkészült programok bitmennyisége? Hogyan viszonylik ez a világháló 10^{21} (Zeta-bájt) méret közelében lévő adatmennyiségéhez?

Megbecsülhető, hány bájt keletkezhet 100 millió programozó napi 16 órás munkájával másodpercenként egy leütés (fél bájt) írási sebességgel évente:

Az eredmény: 1.051.200.000.000.000 bájt.

Egyszerűsítve, ez 1 Petabájt. Ez a szinte irreálisan magas felső becslés évente ekkora növekedési korlátot mutat. (Ha 100 millió digitális kamerát működtetnénk másodpercenként egy felvétellel, az évente kétmilliószor ekkora adattömeget jelentene, 2000 Zetabyte lenne az eredmény.)

Tudományos adatok

Milyen arányt képviselnek?

A lagnagyobbak Petabyte tartományban.

CERN: évente 10Pbyte

A világháló nagy része azonban a közvetlen emberi érzékelésnek megfelelő észlelések rögzítéséből áll.

Méretarányok

A háttértároló 100-szor olcsóbb, mint a gyors memória, igen nagy tárolókapacitás jött létre. (60% PC-ken van)

Amdahl törvények a kiegyensúlyozott rendszerről: a másodpercenkénti lebegőpontos műveletek száma, az operatív memória mérete byte-ban, a másodpercenkénti I/O mennyiség byte-ban, valamint a háttértároló mérete úgy aránylik egymáshoz, mint 1:1:0,1:100.

A petaművelet/sec tartományban 1 GByte/sec adatelérési sebességű merevlemezekből 100000 kell, 1 TeraByte kapacitású tárolókból 100000.

Visszafele: 1 Zetabyte adatbázishoz 10 Exaflops processzor telejsítmény, 10 Exabyte memória, 1 Exabyte beolvasási sávszélesség kell.

A véletlen és kiszámítás

Információra nem lenne szükség véletlen nélkül, az információ nem lenne használható kiszámítás nélkül. Nem lenne semmi szerepe. A véletlen matematikai modelljét a jövő leírásának eszközeként használhatjuk, mint a jövő bizonytalanságának, kiszámíthatatlanságának leírási módszerét.

A múlt homályos leírása és a múlt véletlen gyökerei, valamint a jövő bizonytalanságát modellező valószínűség-eloszlásra való következtetéseink hogyan hozhatók össze?

A jövőre csak a múltból következtethetünk.

Beck Mihály a „Parajelenségek és paratudományok” c. könyv 67-oldalán: „ A különböző folyamatok időbeli lejátszódásának leírása valójában csak akkor lehetséges, ha a jövő nem más, mint *megismételt múlt*.”

A Másodfajú Démon

A véletlen matematikai modelljét a jövő leírásának eszközeként használhatjuk, mint a jövő bizonytalanságának, kiszámíthatatlanságának leírási módszerét.

Mit gyűjt ki a démon? A sok-sok bonyolult konfigurációból, amit a molekulák tánca eredményez, az egyszerűket. A nem tipikus konfigurációkat. A konfiguráció vizuális észlelése alapján „értelmes(nek ható)” szövegeket ír ki.

Modellezzük a jelenséget egy 1000*1000 pixeles fekete-fehér képpel. Az egyes pixelek az órafrekvencia szerint mindentől függetlenül 0,5 valószínűséggel lesznek feketék, vagy fehérek.

Egy kép $2^{-1000000}$ valószínűséggel keletkezik.

$1000000 * 2^{1000000}$ órajelet véve minden kép majdnem biztosan meg fog jelenni. Például ez a szöveg is. Meg a tiszta fehér is.

A Másodfajú Démon

Hogyan választ a démon:

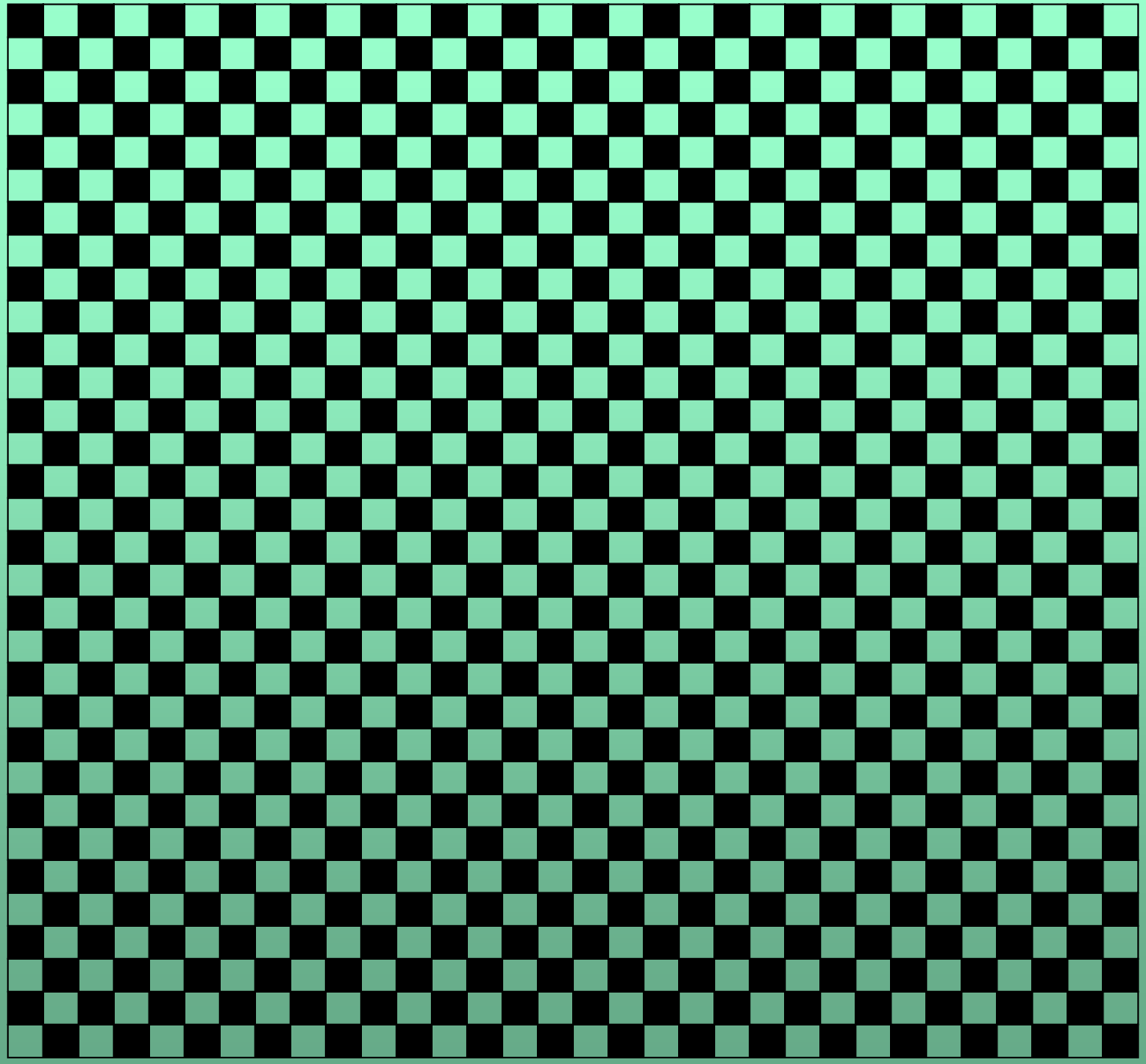
Kiszámolja a kép Kolmogorov bonyolultságát, és ha elég kicsi, kiírja a képet.

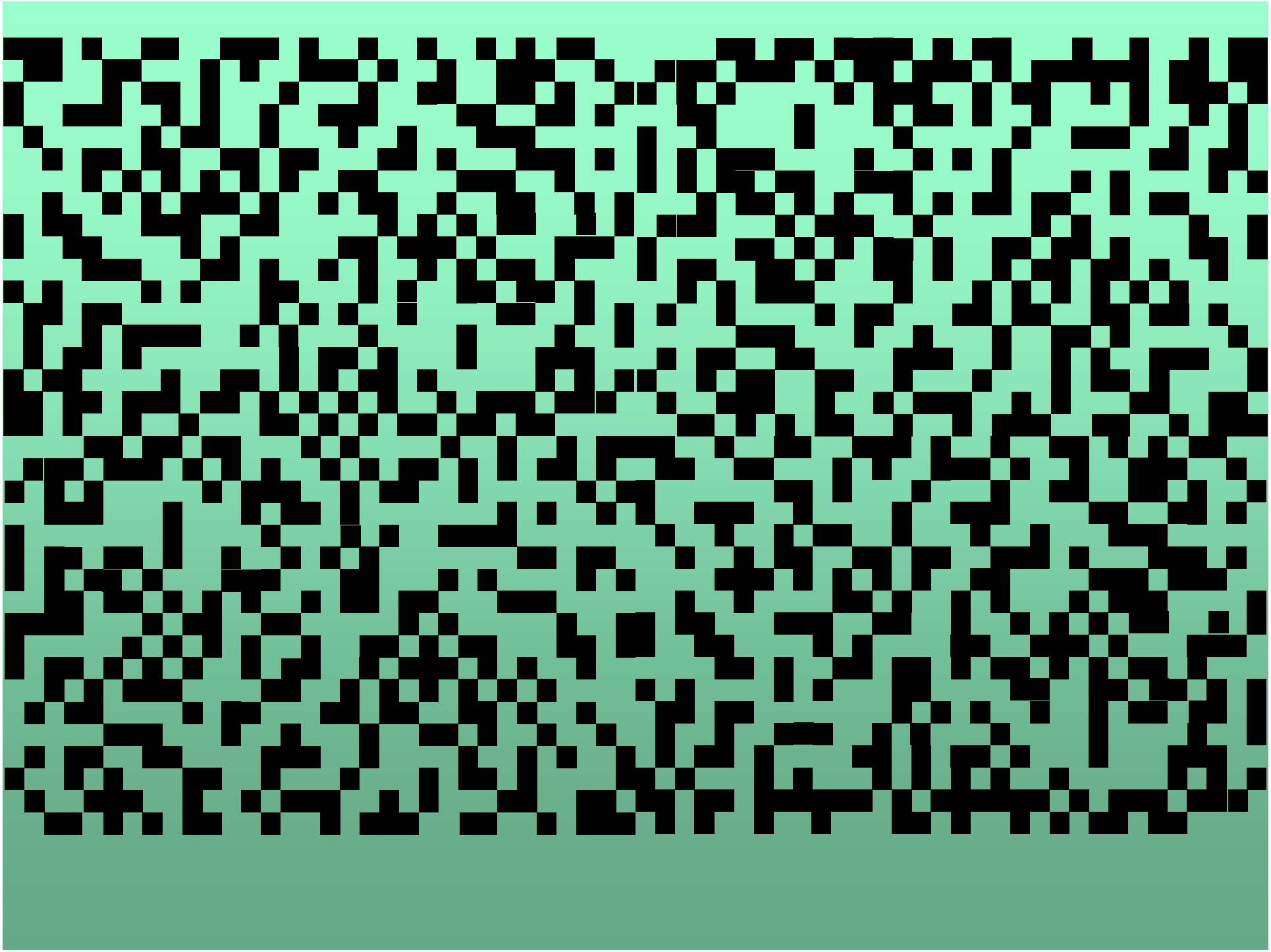
Mitől lesz a képnek jelentése?

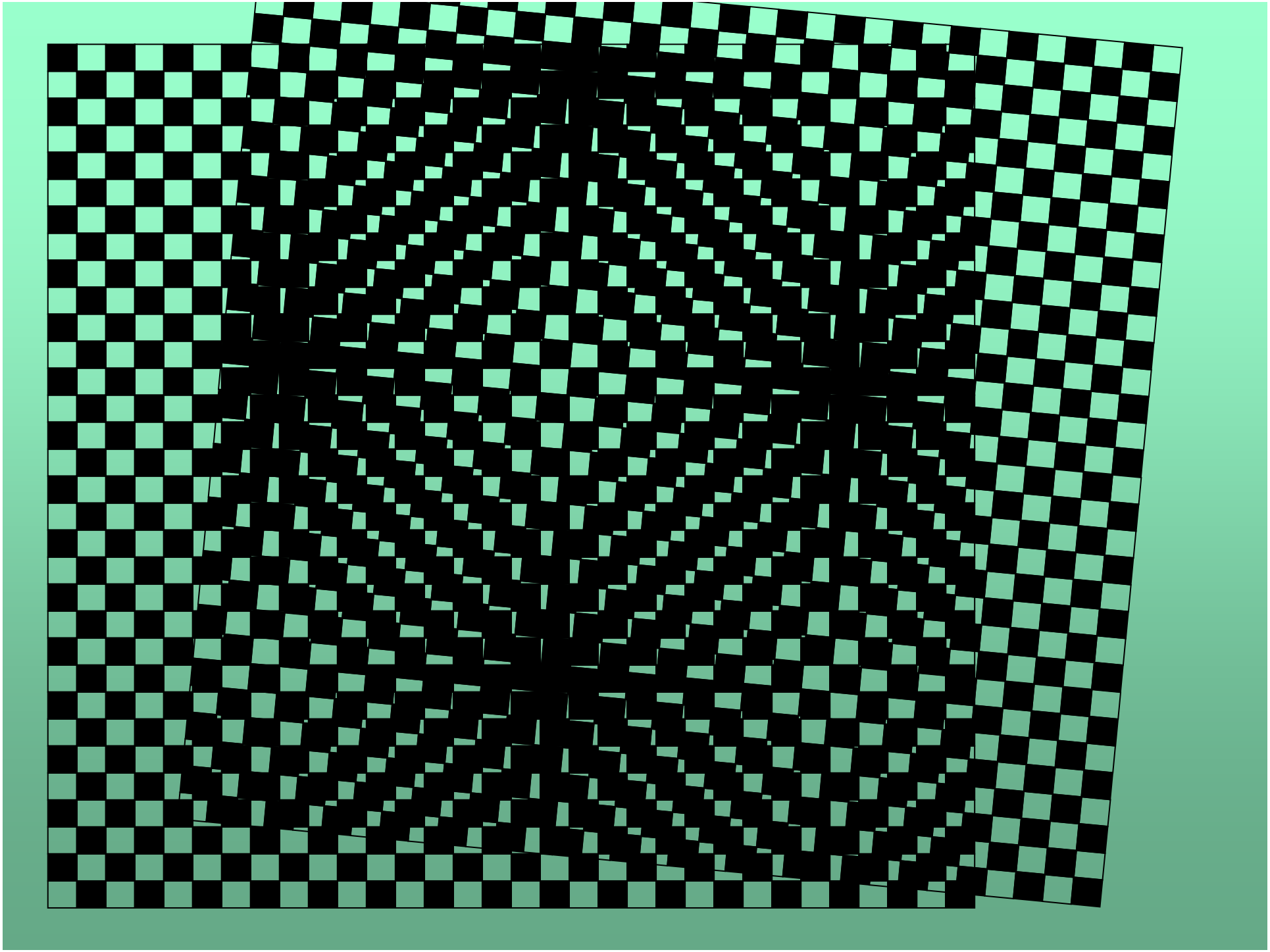
Lem démonja miket írt ki?

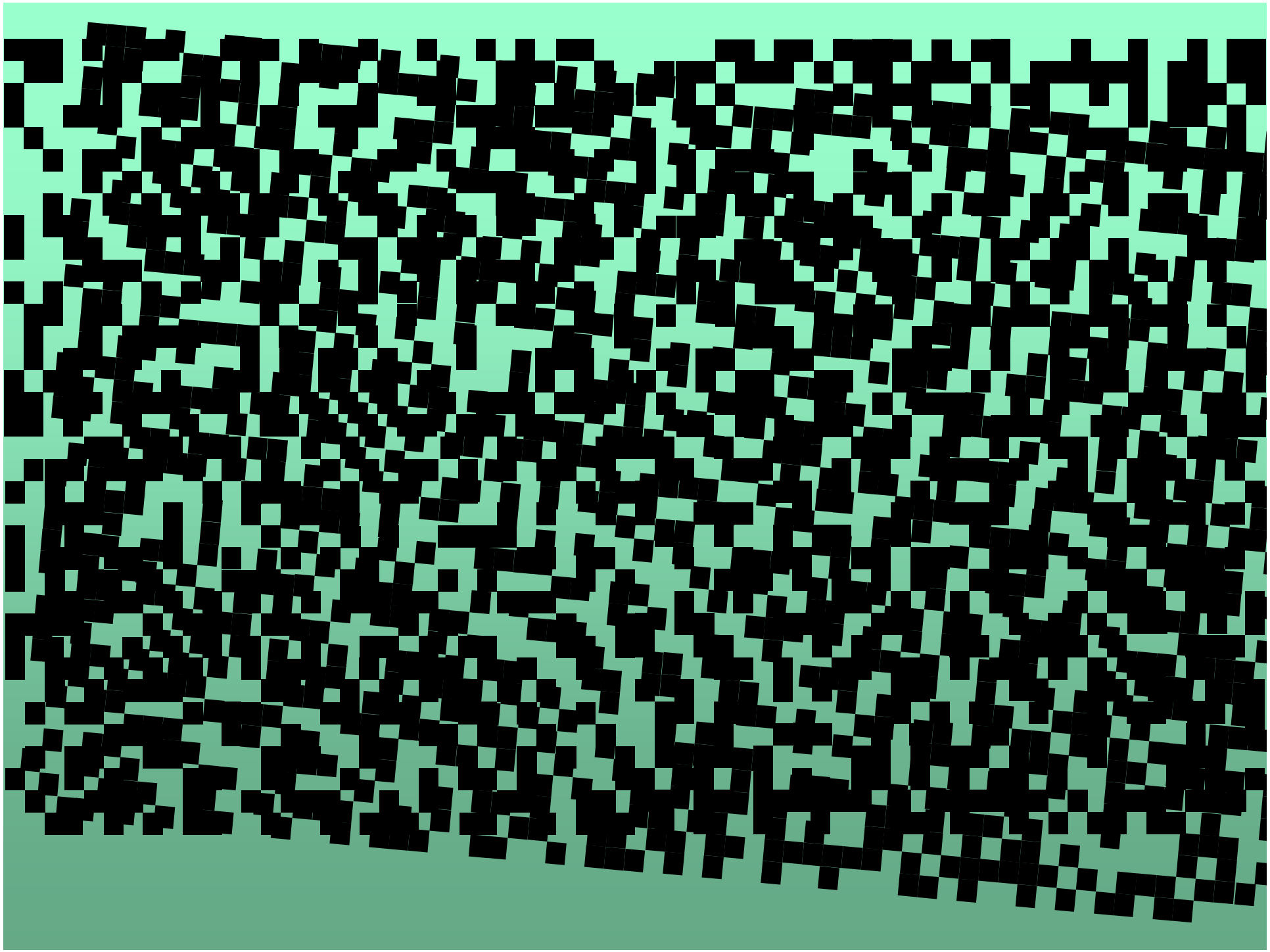
Információ, vagy észlelés?

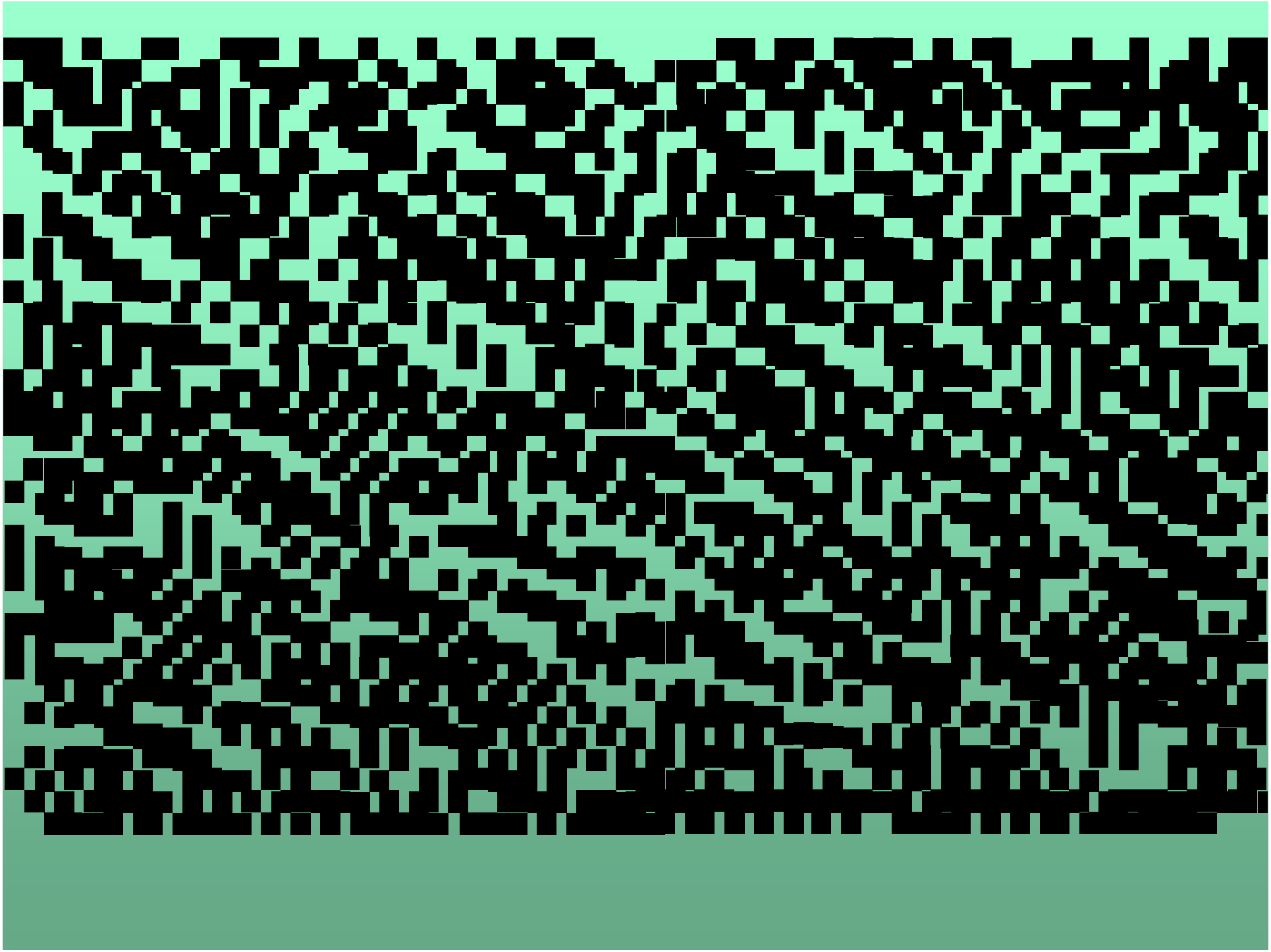
Az észlelt kép még nem kötődik máshoz, nem helyettesít mást.

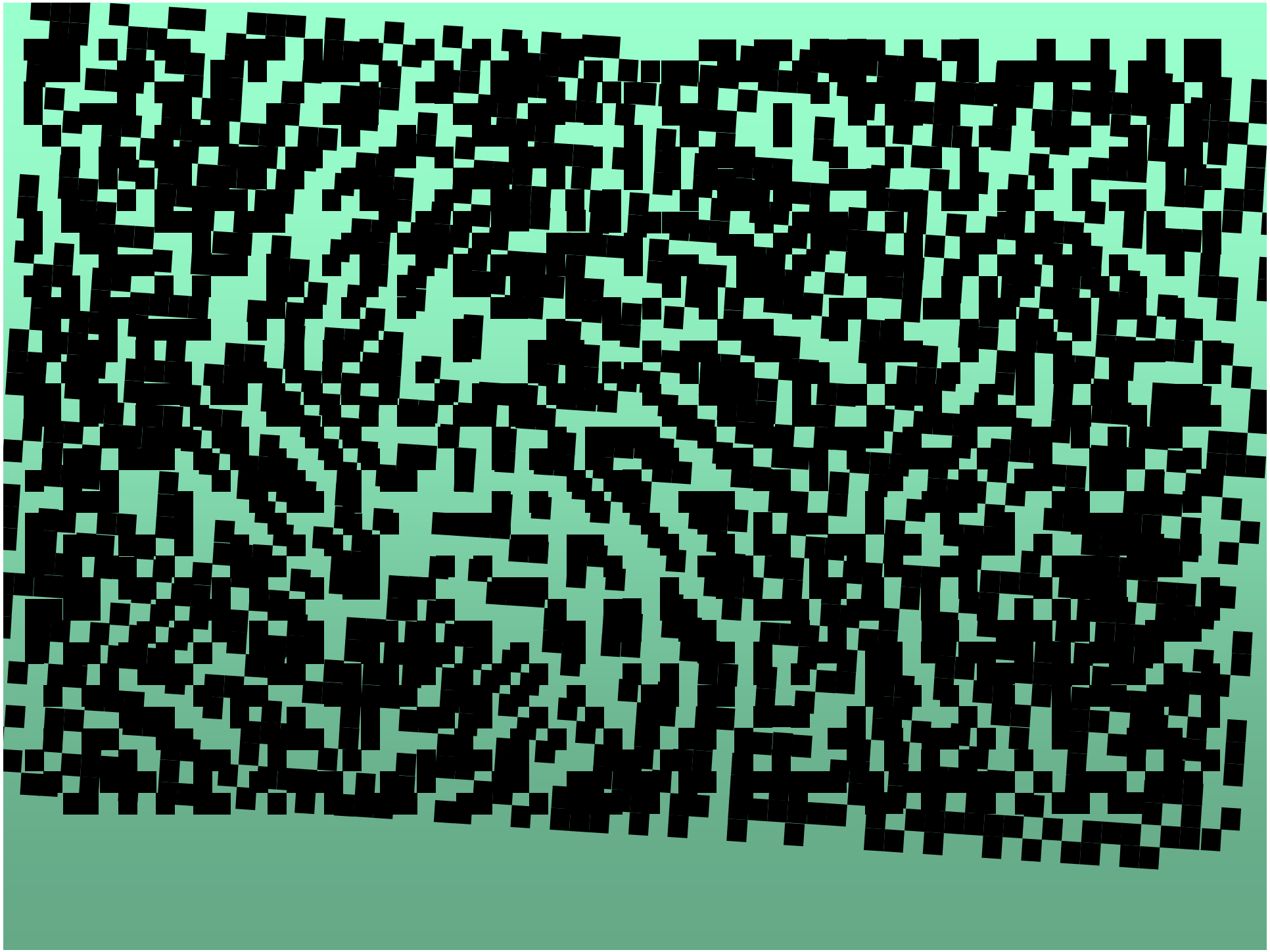












A jellemző szabályosságok írják le a tipikus halmazt, elemein a legbizonytalanabb, egyenletes eloszlást vesszük. Ha megtaláltuk a jó homogén halmazt, lehet, hogy a felhasználás, üzenet szempontjából a konkrét elem már érdektelen.
Példa: szürke színezés. Az 1/2 szürkeség - Fuzzy halmaz - bármelyik tipikus színezéssel szemléltethető.
Két színezés és sajátosságaik: a szórásnégyzet második tagja

$$np(2-p)(1-p)^2 + 2(n-k)p(1-p)^3$$

A véges világban - tehát a digitalizált világban - a matematika végtelen ideális konstrukciói - a folytonosság, végtelen, univerzális algoritmusok - csak segítenek, végtelennel közelítik a véges modellt, amit utána vissza kell véges közelítésbe hozni.

Az információs rendszerek új világa

Ami jellé alakult, utána veszteségmentesen rövidebb, hosszabb ideig, vagy véglegesen tárolódik a világháló címtartományában .

A kódolás feladata a módosítás és visszakeresés, lekérdezés feladatának megoldhatóságát is szem előtt viseli.

Információ-visszakereső rendszerek: kinyerő kérdés
Adatbázis rendszerek: transzformáló és kinyerő kérdés
Tudásbázisok: kérdés-adatbázis + adatbázis rendszer
e-kereskedelem: alkalmazás-adatbázis, interfész-adatbázis +...
Adatmodellek, sémák, előfordulási halmazok, félig strukturált adatok, ...
Módosító és lekérdező nyelvek. A lekérdező nyelvek világa
- a természetes nyelvek kérdezési lehetőségei

Építkezés folyik, a fontos feladatokra kifejlesztik a hatékony megoldást.

Feladatok:
A tárolható adatok világának jellemzése, adatmodellek, tudásreprezentáció, stb.
Változtatások lehetőségei
Kérdések, nyelvek, kifejező erő, bonyolultság, kiszámíthatóság, ekvivalencia,...
A címtartomány szerepe - a hely egyben adat is lehet - a címtér a világhálón
Hogyan programozzuk az egészet?

Hogyan értjük meg ami készül?

A közös ismeret elérése

- Beszéd: élő személyek tudata
- Írás: jegyzetek, levelek, könyvek, könyvtárak, levéltárak, irattárak, nyilvántartások, posta
- Számítástechnika, info-kommunikációs technikák: adatbázisok, Web szerverek, digitális könyvtárak, böngészők, ágensek, indexek, adatbányászat, e-mail

Épül az emberiség átfogó információs rendszere.

A megőrzött tudás elektronikus, digitalizált felvételre kerül kiegészülve automatizálható, algoritmizált feldolgozási lehetőségekkel.

A hozzáférés: kérdezés.

Csak annyi információt nyerhetünk ki, amennyit bevittünk. (A válasz információmennyisége nem nagyobb, mint a kérdésé.)

Mi kérdezünk, vagy helyettünk kérdeznek?

Új kérdést csak az élő tudat képes feltenni.

Az tud kérdezni, aki sokat tud.

A tudás évente megkétszereződik

- egyéni szakértelem kérészéletű
- minden tanító tanuló
- a tanulás nagy kihívása csak a világot behálózó hálózaton keresztül valósítható meg, amely minden tudatot és tudást összeköt

Észlelés – megismerés – információ

Fizikai jelenségek észlelése: egy másik, már ismertebb jelenségre való hatáson keresztül történik. A kezdeti észlelés, mint felismerhető jelenség, az emberi érzékszervekre, érzékekésre alapul, a telereceptoraink, azaz a látó és halló receptoraink informálnak bennünket..

helyettesítés

Minden észlelésből helyettesítő jelek maradnak. Tudatunkban is. Utána már csak ezekkel tudunk manipulálni.

Ez viszont csak algoritmusokkal történik – absztrakt értelemben. Az emberi gondolkodás, következtetés, stb, minden eddigi lehetősége a kiszámítási világon belül van. Ami ezen túl van, az nem ellenőrizhető. (Bár a kiszámítás, gondolkodás se mindig ellenőrizhető.)

megismerés

Miben áll a természettudományos megismerés legfontosabb lehetősége: Addig kell provokálnia természetet, míg olyan új múltat nem eredményez, amilyet eddig még nem észleltünk. A jövőre csak a múltból következtethetünk.

Ezt fogalmazza meg másként Beck Mihály a „Parajelenségek és paratudományok” c. könyv 67- oldalán: „ A különböző folyamatok időbeli lejátszódásának leírása valójában csak akkor lehetséges, ha a jövő nem más, mint *megismételt múlt*.”

Zárszó

Juris Hartmanis: Zárszó:

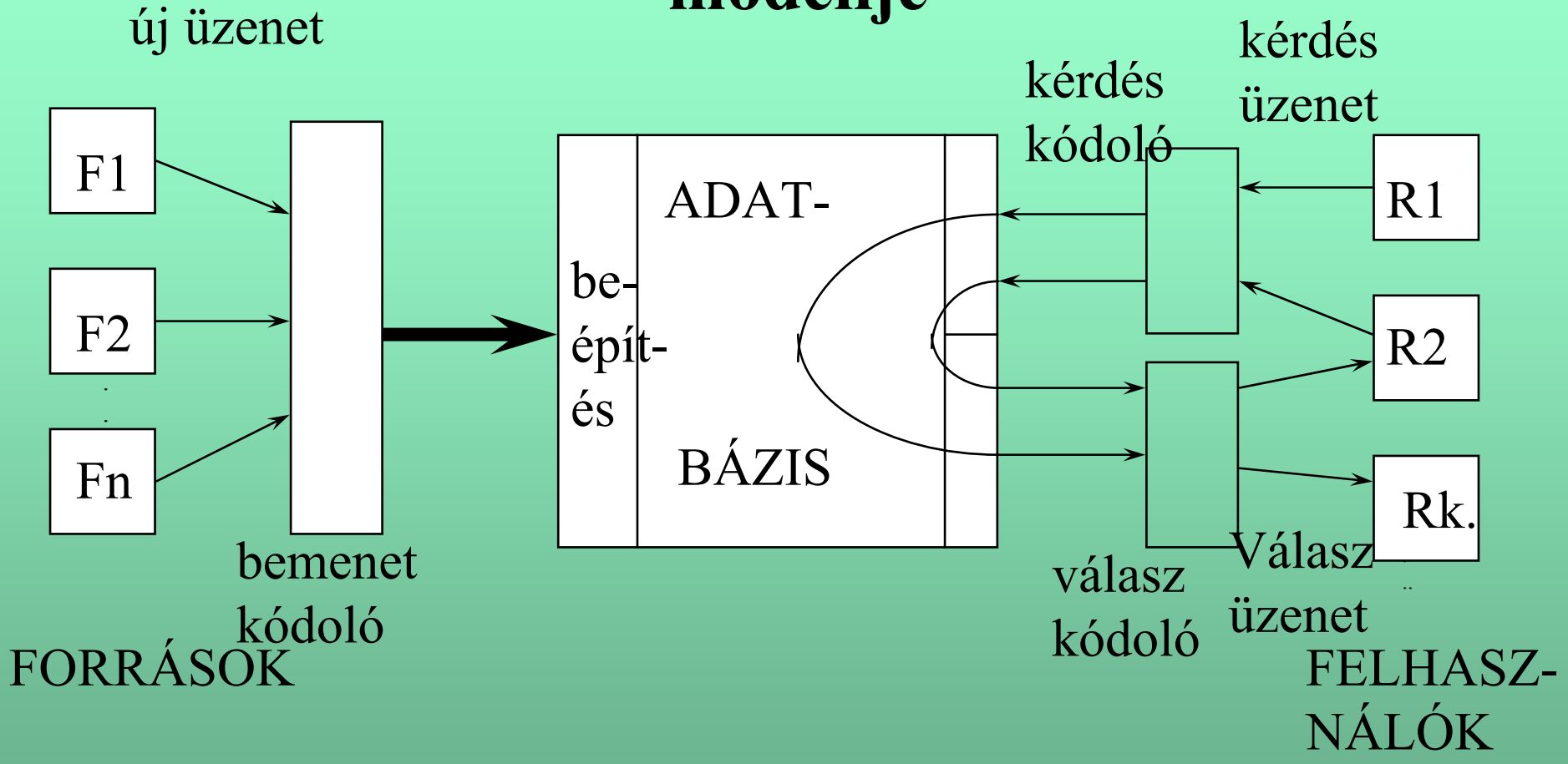
„Hiszek abban, hogy a számítógép-tudomány rendelkezik olyan potenciális erővel, amely által mélyebben be tudunk tekinteni a kiszámítás paradigmájába, valamint saját intellektuális folyamatainkba, kvantitatív megértésüket kaphatjuk, és így, esetleg talán, egy lehetőséget nyerünk a tudható határának átlépésében.”

Zárszó

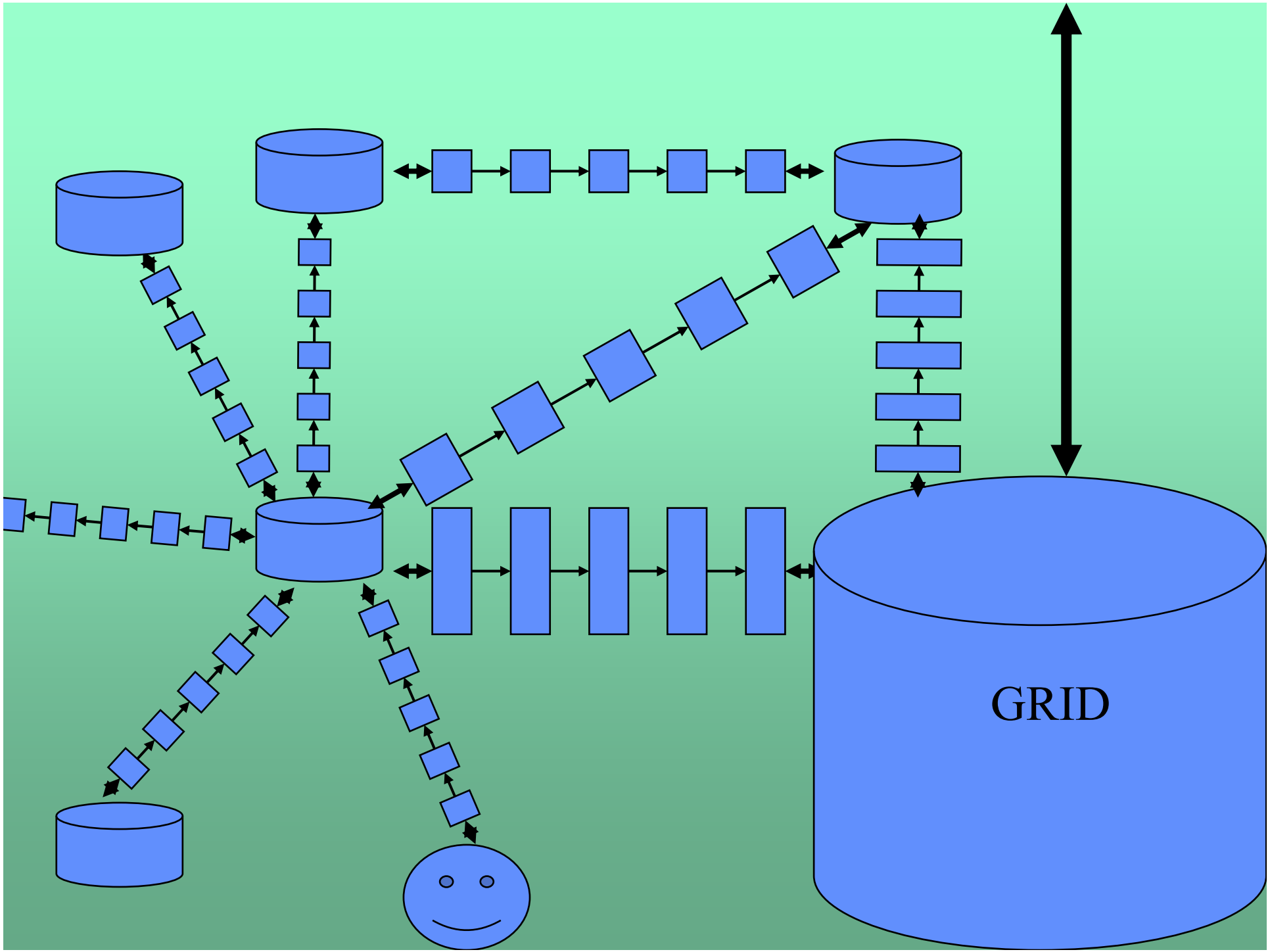
Juris Hartmanis: Zárszó:

„Hiszek abban, hogy a számítógép-tudomány rendelkezik olyan potenciális erővel, amely által mélyebben be tudunk tekinteni a kiszámítás paradigmájába, valamint saját intellektuális folyamatainkba, kvantitatív megértésüket kaphatjuk, és így, esetleg talán, egy lehetőséget nyerünk a tudható határának átlépésében.”

Az automatizált információs rendszer modellje



A kódolók előre rögzített formájú, speciális nyelven írt üzeneteket fogadnak, ehhez intelligens interaktív felületet adnak a források és felhasználók számára.


GRID

Matematikai feladatok

A lehetséges üzenethalmaz és a lehetséges csatorna jelsorozat halmaza közötti megfeleltetés:”skatulya elv”: ha több golyó van, mint skatulya, akkor van olyan skatulya, amelybe egynél több golyó jut.

Shannon modell: T idő, V szimbólum/sec sebesség, H bit/szimbólum átlagos entrópia, az választás bizonytalansága

Valószínűség-eloszlás a matematikai modell, $H = -\sum_{i=1}^{n} p_i \log_2 p_i$

Csatorna kapacitás: C bit/sec

skatulyák - csatorna jelsorozatok száma $2^{T(C-\delta)} \le N(T) \le 2^{T(C+\delta)}$

golyók - választható üzenetek T idő alatt,

a meghatározó többség számossága: $2^{TV(H+\lambda)}$

Csatorna kódolás alaptétele: $V = CH^{-1} - \varepsilon, \quad \varepsilon > 0$

Beszéd, írás esetében nem a csatorna adja a korlátot, hanem beépített kódoló-adó és vevő-dekódoló „berendezésünk”. A megértéshez és megjegyzéshez nagy redundancia kell!
A matematika a mesterséges csatornák megjelenésekor vált fontossá. (Adó, átviteli terjedés sebessége, vevő együtt adják N(T)-t, zaj, híradástechnika és információelmélet.)

Az alaptétel a kódolást nem adja meg, érték és permutáció szerint invariáns az entrópia.

Kódoláshoz algoritmus és adat kell, számítás nélkül csak analóg átvitel lenne lehetséges.

Csatorna kihasználása: időbeli átlag (blokkosítás)helyett térbeli átlag (multiplexálás)

2008-ra 1Tbit/sec kapacitású ethernet hálózat lesz!

Az objektív mérőszám: a Kolmogorov entrópia.

Def. Az f(p) kiszámítható parciális függvényhez

$$C_f(x) = \min(l(p) | f(p) = x),$$

végtelen, ha nem létezik p kód. A p kód hossza $l(p)$.

Tétel. Létezik optimális kódoló, $f_0(p)$, hogy

$C_{f_0}(x) \le C_f(x) + n_f$ tetszőleges f, x esetén, ahol n_f

csak f-től függ. (Univerzális függvény $U(n_f, p) = f(p) = x$)

Az optimális kódolók additív konstansban térnek el egymástól.

Rögzítve az f_0 optimális függvényt kapjuk az

$I(x) = f_0(x)$ Kolmogorov bonyolultságot.

Prefixmentes : K(x)

Feltételes: I(x|y) , illetve K(x|y)

$$I(x \mid y) = C_{f_0}(x \mid y) = \min(l(p) \mid f_0(p, y) = x)$$

Ahol f_0 két változós optimális függvény.

Egyik sem kiszámítható.

Skatulya elv: A(t) legyen a t paraméter szerint rekurzívan felsorolható véges elemű halmaz, m(t) legyen elemeinek száma. Akkor $x \in A(t)$ esetén

$K(x \mid t) \leq \log_2 m(t) + c$, és fordítva, A(t) elemei döntő többségének legalább ekkora a t szerinti feltételes bonyolultsága. (a $\log_2 m(t)$ hosszúságú kódok száma!)

Információ megmaradás – nem növekedés - törvénye:

A q kérdésre adott y válasz információmennyisége az addig bevitt x adat ismeretében:

$K(y|x) \leq K(q)+c$.

Mert a v(q,x)=y válaszfüggvényre

$$K(x \mid y) \leq C_v(x \mid y) + n_v \leq l(q) + n_v$$

A két entrópia kapcsolata:

- Hosszú üzenetsorozatokra a tipikus halmaz megadása

- A kódolandó elemek halmazának jó megválasztása

Mindkét esetben a megtalált halmazokon az egyenletes kódhossz választásánál nincs jobb lehetőség.

A Shannon entrópiával a tipikus halmaz elemszáma $2^{TV(H+\lambda)}$

A tipikus halmaz elemeinek feltételes Kolmogorov entrópiája az elemszám logaritmusa, $TV(H+\varepsilon)$, tehát az egy szimbólumra jutó bonyolultság $H\left(1+\frac{\varepsilon}{TV}\right)$,

feltéve hogy algoritmikusan megadható a tipikus halmaz.

A feladat: a lehető legszűkebb halmaz keresése. (Mindkét esetben.)
Véges esetben az algoritmikus világ törvényei érvényesülnek.

A Kolmogorov entrópiára univerzális tanuló eljárások épülnek, amelyek nem kiszámíthatóak, csak közelíteni tudjuk. Például az egész számok univerzális apriori eloszlása a Bayes módszer kiindulásához: $\pi(x) = 2^{-K(x)}$, és $\sum_x \pi(x) < 1$.

A Bayes elv: a legnagyobb apriori bizonytalanságból indulunk ki, véges esetben az egyenletes eloszlásból. Az egész számok felett az algoritmikusan legbizonytalanabb eloszlás a fenti univerzális Levin-féle eloszlás. Csak közelíthető!

Az eloszlásból származó végtelen minta algoritmikusan 1 valószínűséggel felismerhető. Adott eloszlás szerinti végtelen véletlen sorozatok halmazának komplementere algoritmikusan jellemezhető, és u.n. effektíven null-mértékű halmazt jelent az adott eloszlás szerint.

Véges esetben minden más. A nagyon hosszú előfordulások egyre nagyobb hányada viselkedik a fenti módon, vagyis relatíve egyszerű algoritmikus tesztekre a határértékben nulla-egy törvényűség teljesül.

Tanulás: kiszűrjük a jellemző szabályosságokat, s a megmaradó egyedi tulajdonságokat adatként adjuk meg.

Cyber-infrastruktúra

Z. Karvalics László

A cyber-infrastruktúra mint aktuális kihívás és mint tudományszociológiai probléma I. Magyar Tudomány 2007.

A cyberkörnyezet egyaránt támogatja a tudomány-gyár működésének átfogó újratervezését (*re-engineering*) és a kutatási folyamatok jobb programozását - evvel **a tudomány új korszaka** születik meg (*next generation science*), amit bátran nevezhetünk **adat-intenzív tudománynak** *(data intensive science)*

(Jim Grey, Alex Szalay et.al. 2005).

Jeltömeg és kontrollválság

Z. Karvalics László

Az adatsilóktól a tudomány kontrollforradaImáig.
Magyar Tudomány 2008.

„A tudósok lényegesen gyorsabban hozzák létre az új adatokat, mint ahogy azokat elemezni tudnák. Az eredmény leginkább az optikai csalódásra hasonlít”.

(Hugh Kiefferttől származó idézet)

Mi az e-science?

Az ELTE eScience RET pályázatból

Új feltörekvő technológia, melynek révén nagyléptékű, komplex tudományos tevékenység fejthető ki a modern információs technológia felhasználásával. Legfőbb jellemzője a rendkívül sok, gyakran különböző helyekről elérhető adaton operáló kiértékelő munka, melynek eredményes véghezvitelére az adatok automatikus gyűjtése, optimális adatbázisba rendezésére, rendkívül nagy számítástechnikai kapacitást igénylő feldolgozására, és a lényeget megragadó vizualizációjára van szükség.

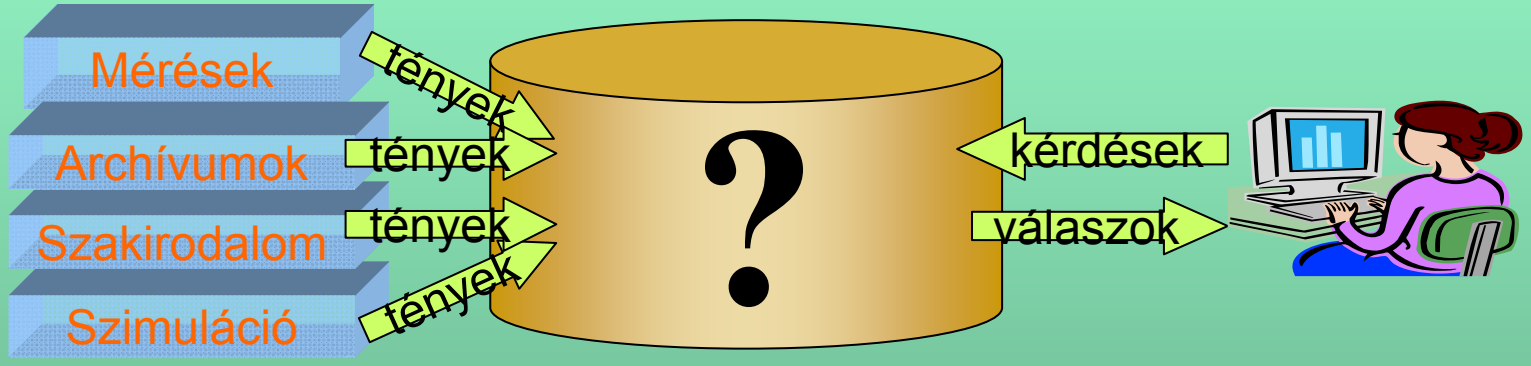
CyberinfrastruCture Vision for 21st Century DisCoVery

National Science Foundation

Cyberinfrastructure Council

March 2007

Az átfogó kép a tudományban



- Adatfeltárás
- Petabyte kezelése
- Közös séma
- Hogyan szervezzük?

- Hogyan szervezzük át?
- Hogyan működjünk együtt?
- Adatlekérdezés és vizualizáció
- Hatékonyság

Zárszó

Juris Hartmanis: Zárszó:

„Hiszek abban, hogy a számítógép-tudomány rendelkezik olyan potenciális erővel, amely által mélyebben be tudunk tekinteni a kiszámítás paradigmájába, valamint saját intellektuális folyamatainkba, kvantitatív megértésüket kaphatjuk, és így, esetleg talán, egy lehetőséget nyerünk a tudható határának átlépésében.”