Adatelemzés és adatbányászat MSc

1. téma

IR és adatelemzési alapok

Dr. Kovács László
ME GEIAL

Információ szintjei

Kell-e targonca?

Mennyi targonca kell?

Mennyi kezelők kellenek?

Mit szállítsanak ma?

Mikorra vigyék át?

?

Hova kell menni most?

Mit kell vinni?

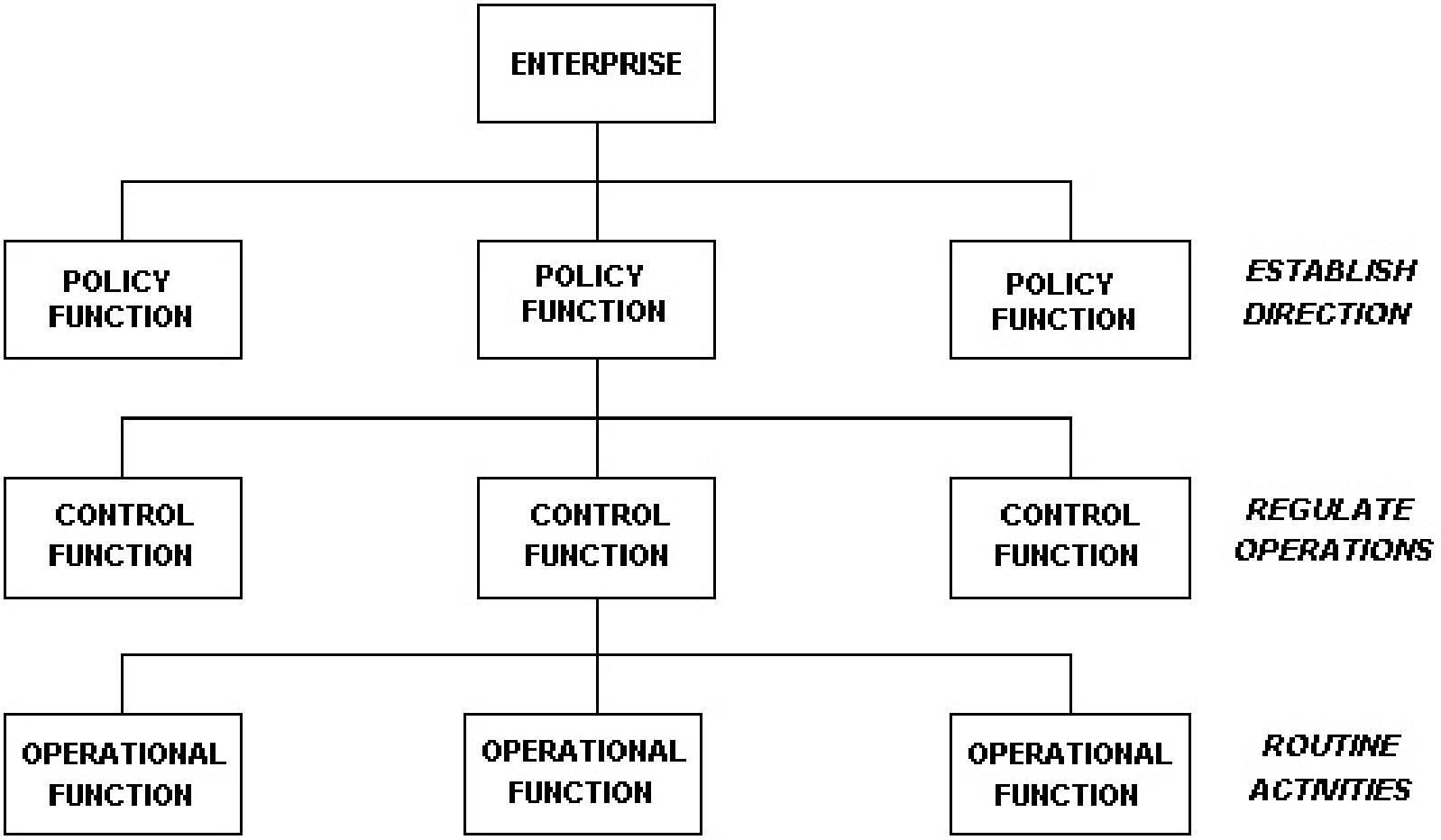
Merre kell menni?

Hogyan kell beindítani?

Hogyan kell fékezni?

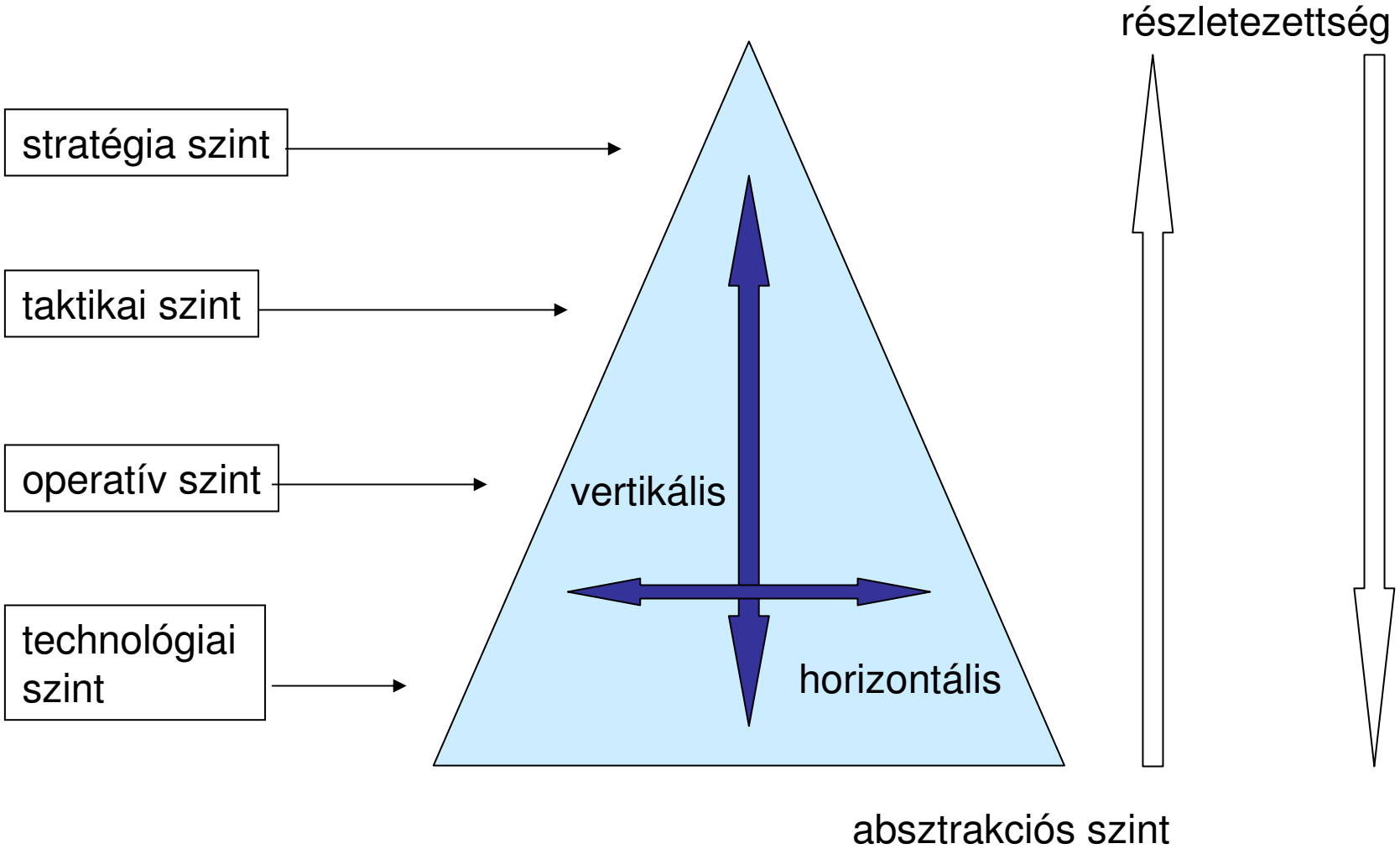
Dr. Kovács László
ME GEIAL

Vállalati irányítási szint modulok



VIR információ szintjei

Információ áramlások funkció nézetben



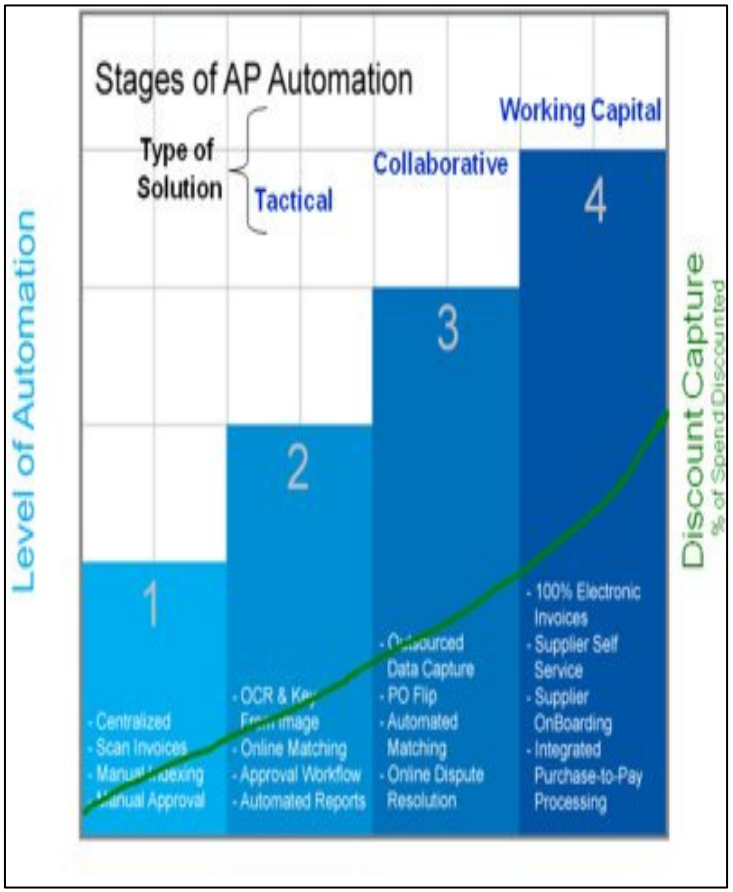
Dr. Kovács László
ME GEIAL

Vállalati információs rendszer (VIR)

VIR: Olyan információs (informatikai) rendszer, melynek célja, a vállalati információkezelés hatékonyságjavítása, az információk integrált kezelése.

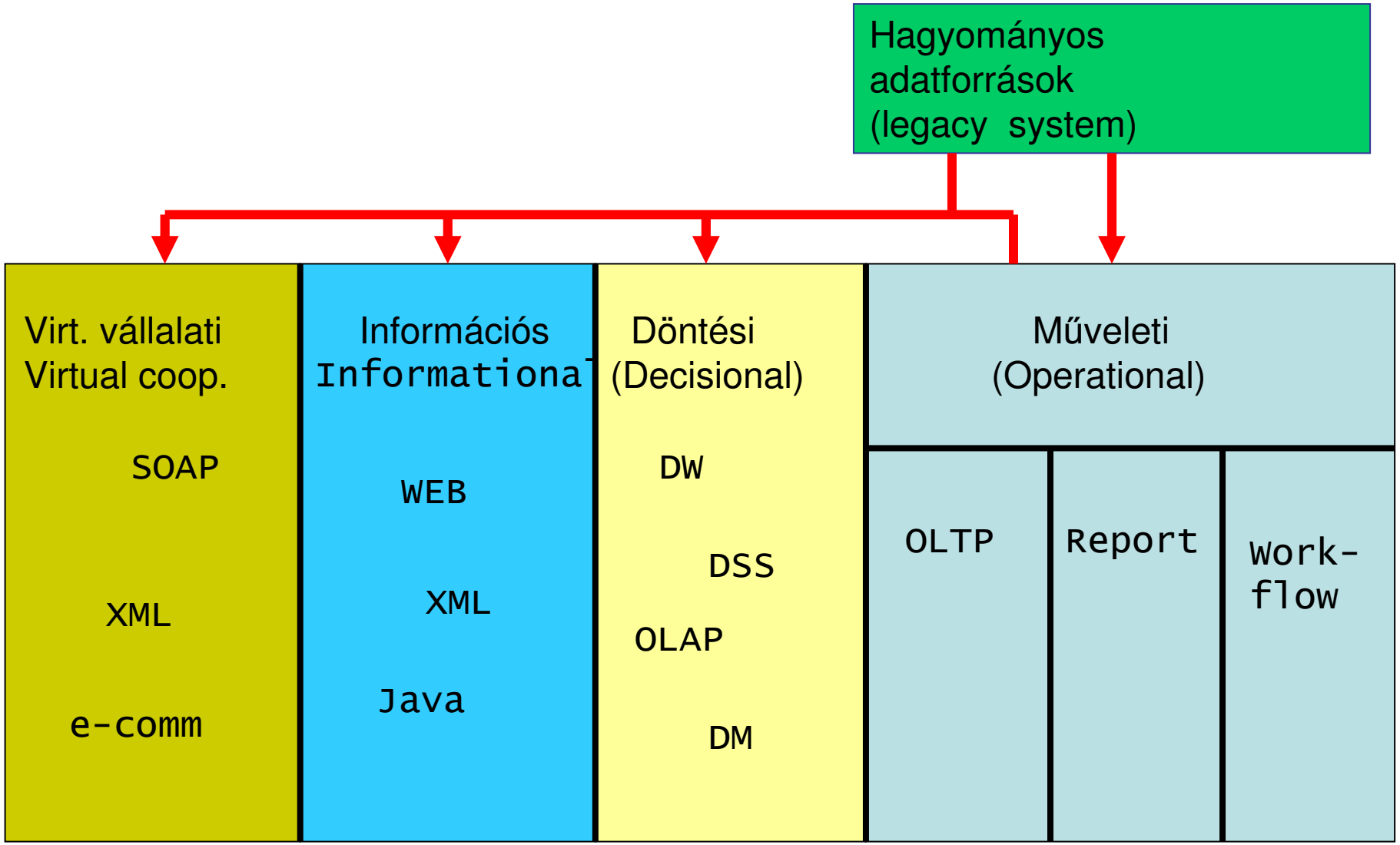
A VIR előnyei:

- információ gyors elérése
- információ pontossága
- piaci versenyhelyzet javítása
- termelékenység növelése
- ügyvitel javítása
- technológia fejlesztése
- hatékonyabb kapcsolat a partnerekkel
- hatékonyabb döntéshozatal , tervezés



Dr. Kovács László
ME GEIAL

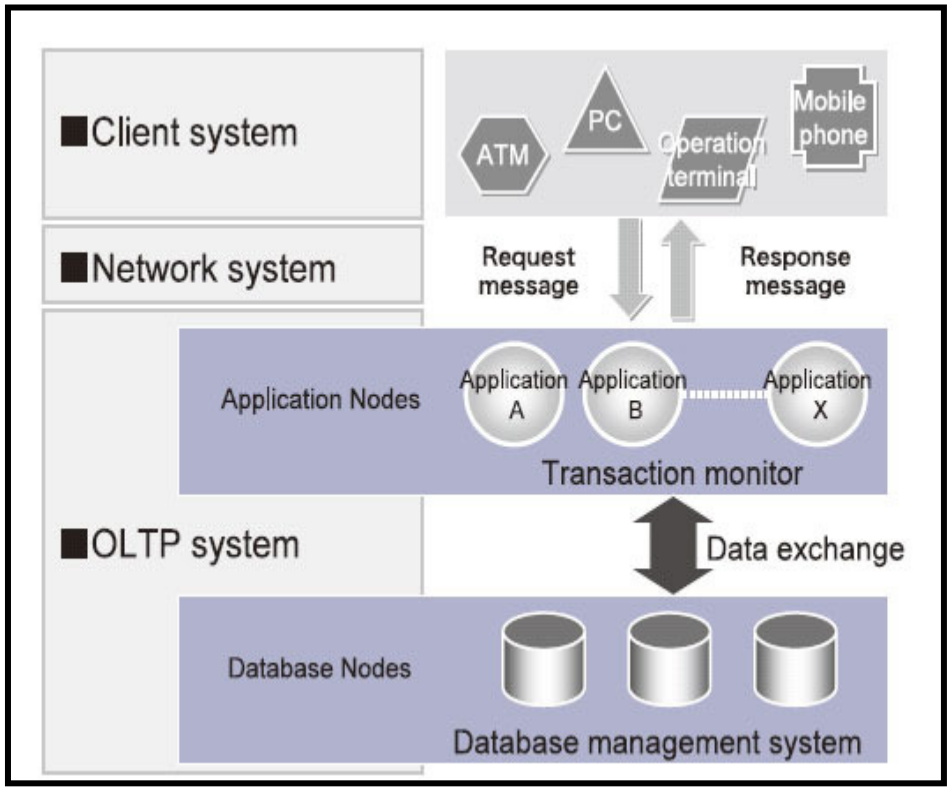
Információs rendszerek típusai



Dr. Kovács László
ME GEIAL

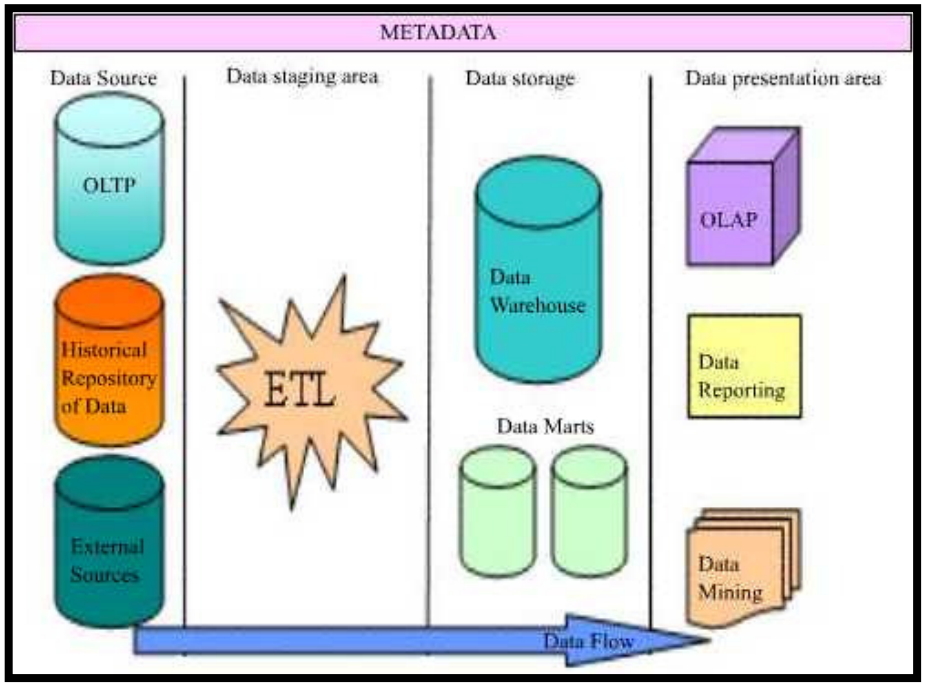
OLTP jellemzői

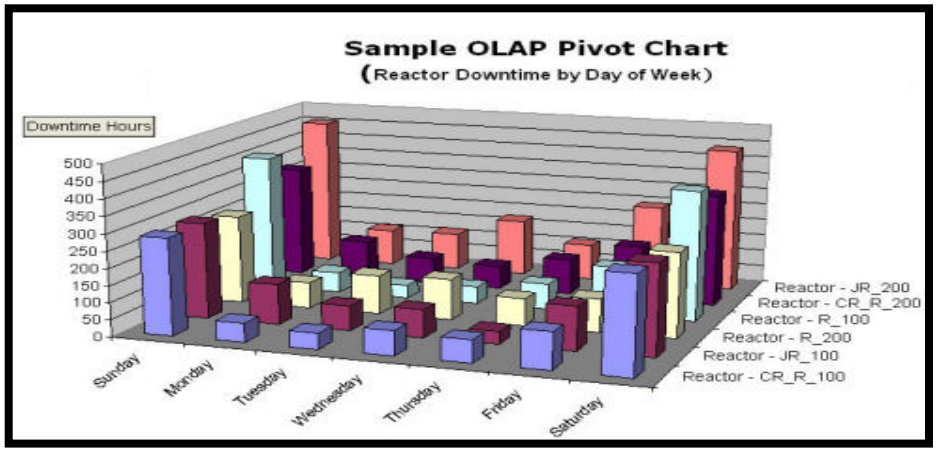
- adatmódosítás
- aktuális állapot
- nagy konkurencia
- konzisztencia
- rövid tranzakciók
- homogenitás
- normalizált
- relációs és OO
- SQL felület
- ACID elvek
- adatvesztés elleni védelem



OLAP jellemzői

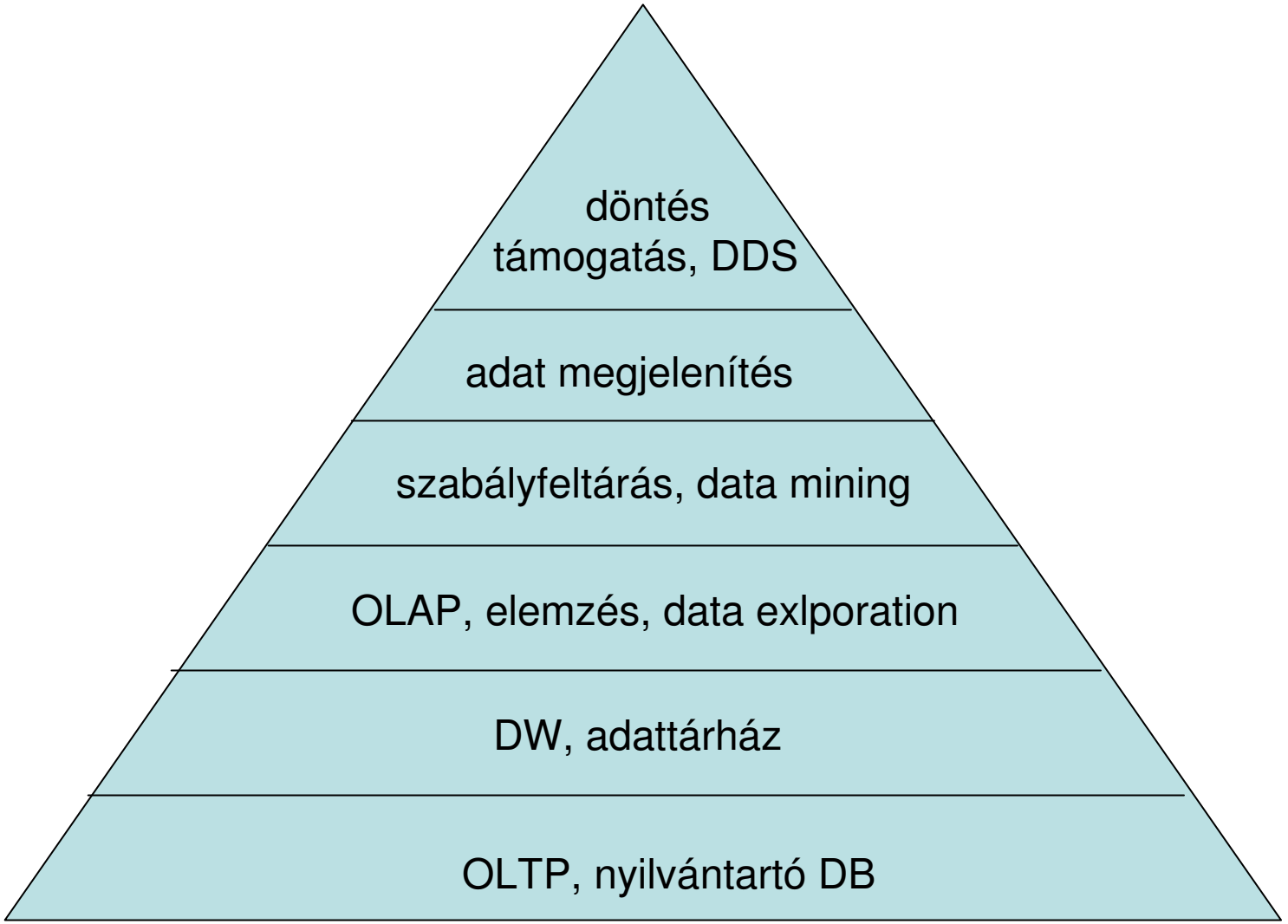
- adatlekérdezés
- korábbi állapotok
- kis konkurencia
- betöltés
- hosszú tranzakciók
- heterogenitás
- adatkocka
- modulokból áll
- nincs szabvány
- nem normalizált
- adatvesztés elleni védelem





Dr. Kovács László
ME GEIAL

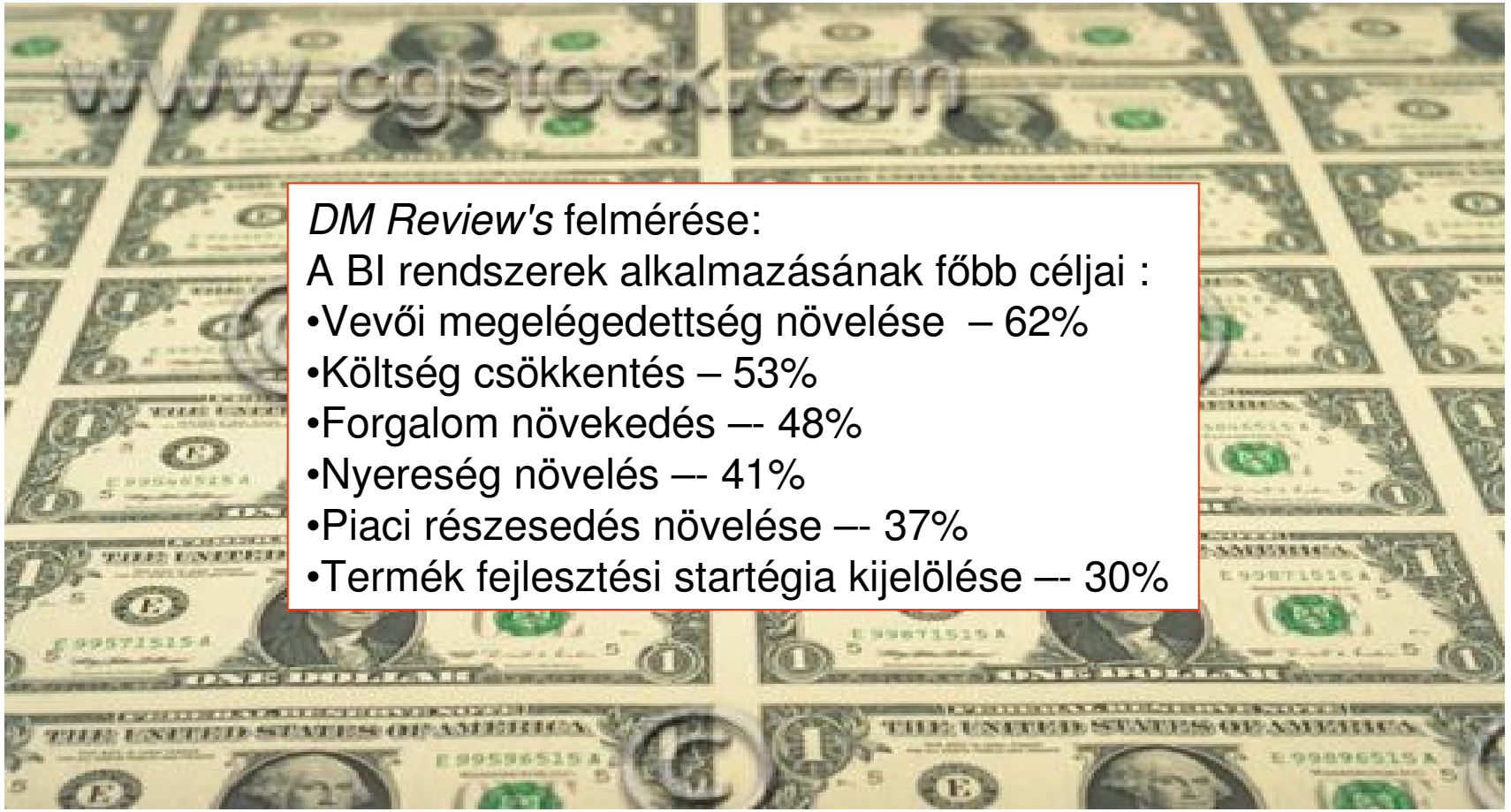
DSS, OLAP rendszerek szerepe



Információ feldolgozási szintek

Dr. Kovács László
ME GEIAL

OLAP rendszerek célja

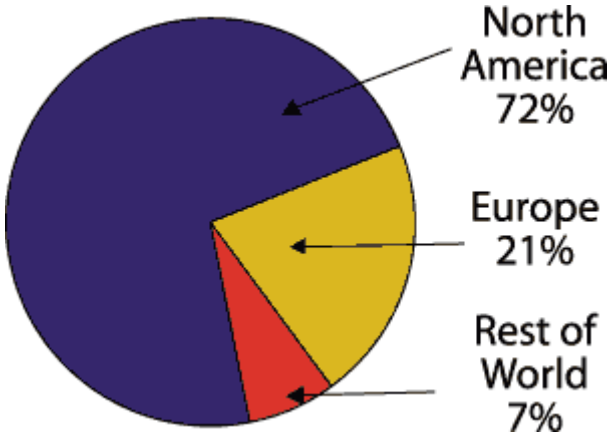


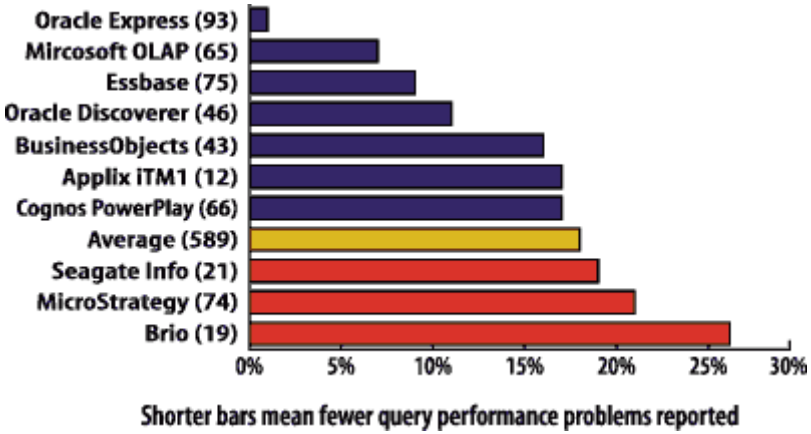
DM Review's felmérése:
A BI rendszerek alkalmazásának főbb céljai :
•Vevői megelégedettség növelése – 62%
•Költség csökkentés – 53%
•Forgalom növekedés –- 48%
•Nyereség növelés –- 41%
•Piaci részesedés növelése –- 37%
•Termék fejlesztési startégia kijelölése –- 30%

OLAP/DW rendszerek piaca

- évi 3 milliárd dolláros piac
- több szereplős
- főbb területek:
  - ERP
  - CRM
  - Sales analysis
  - planning
- 85% sikeres projekt
  - MicroStrategy, MS, Oracle
- megbizhatóság
- kb 2/3 részben kihasználtak
- nem dominál a Web-OLAP
- piaci részesedés:
  - MS
  - Oracle
  - Essbase
  - MicroStrategy

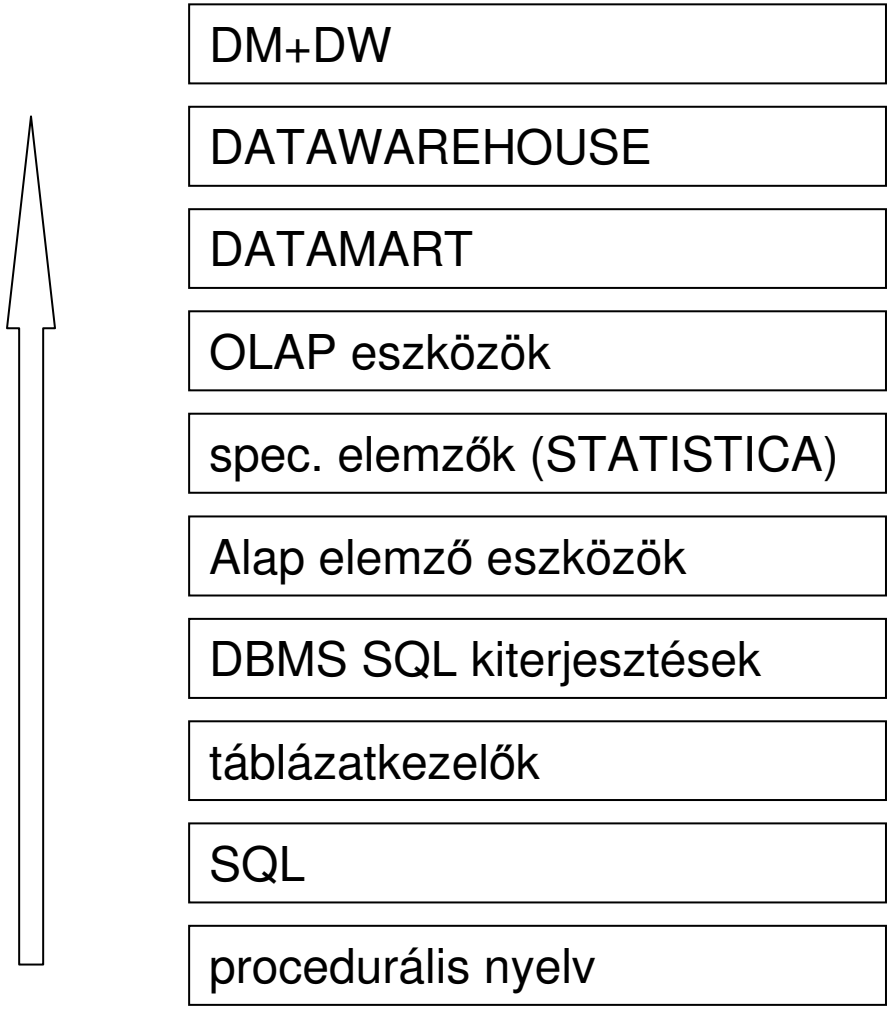
OLAP elterjedtsége





teljesítmény viszonyok

Adatelemzési eszközök



Dr. Kovács László
ME GEIAL

Adatelemzés SQL-ben

szórás

```
SELECT
SQRT(AVG(POWER(x.ar – (SELECT AVG(y.ar) FROM auto_eladas),2)))
FROM auto_eladas;
```

Gyakorisági tábla
(évenként a Fiat autók relatív aránya)

```
SELECT
ev, SUM(CASE WHEN tipus='FIAT' THEN 1 ELSE 0 END) / SUM(1)
FROM auto_eladas
GROUP BY ev;
```

Adatelemzés SQL-ben

korreláció
(a Fiat és Opel autók havi darabszámai között)

```
SELECT CORR (dbF,dbO) FROM
(  SELECT
ev, SUM(CASE WHEN tipus='FIAT' THEN 1 ELSE 0 END) / SUM(1)  dbF,
SUM(CASE WHEN tipus='OPEL' THEN 1 ELSE 0 END) / SUM(1)  dbO
FROM auto_eladas
GROUP BY ev )
WHERE dbF * dbO IS NOT NULL;
```

differencia képzés
Havi összforgalom változás

```
SELECT T1.ho, T1.db – T2.db dbdiff FROM
(  SELECT  ho, SUM(ertek) db  FROM auto_eladas  GROUP BY ho ) T1
INNER JOIN
(  SELECT  ho, SUM(ertek) db  FROM auto_eladas  GROUP BY ho ) T2
ON (T1.ho = T2.ho  + 1 and  T1.ev = T2.ev)  OR
(T1.ho =1 and T2.ho =1 and  T1.ev = T2.ev + 1)
```

Adatelemzés SQL-ben

keresztreferencia tábla
(autótípusok éves darabszámai)

```
SELECT
   SUM (CASE WHEN ev = 2006 THEN db ELSE 0 END) e06,
   SUM (CASE WHEN ev = 2007 THEN db ELSE 0 END) e07,
   SUM (CASE WHEN ev = 2008 THEN db ELSE 0 END) e08,
   SUM (CASE WHEN ev = 2009 THEN db ELSE 0 END) e09,
   ---
FROM auto_eladas
GROUP BY tipus
```

differencia maximuma
Legnagyobb havi összforgalom változás

```
SELECT MAX(dbdiff)  FROM
(SELECT T1.ho, T1.db – T2.db dbdiff FROM
(  SELECT  ho, SUM(ertek) db  FROM auto_eladas  GROUP BY ho ) T1
INNER JOIN
(  SELECT  ho, SUM(ertek) db  FROM auto_eladas  GROUP BY ho ) T2
ON (T1.ho = T2.ho  + 1 and  T1.ev = T2.ev)  OR
(T1.ho =1 and T2.ho =1 and  T1.ev = T2.ev + 1))
```

Adatelemzési módszerek

Regresszió

Feladat: adott mérési pontra legjobban illeszkedő görbe megkeresése

Adottak: mérési pontok, függvényosztály (paraméteresen)

Feladat: a mérési pontokra legjobban illeszkedő paraméterek meghatározása

Optimalizálási feladat:

Célfüggvény: illeszkedési hiba: eltérések négyzetösszege

Optimalizálási módszerek:

Derivált zérushelye
Gradiens módszer
Sztohasztikus keresés

$$\{(x_i, y_i)\}$$

$$\{f(\overline{p}_i, x)\}$$

$$E(\overline{p}_i) = \sum_i (f(\overline{p}_i, x_i) - y_i)^2$$

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Lineáris regresszió

többváltozós lineáris regresszió: a mérési pontokat legjobban közelítő függvény meghatározása

egy függő változó feltételes várható érték becslésére szolgál

$$E(y|x_1\ x_2..) = F(x_1,x_2,..,\alpha_1,\alpha_2,..)$$
$$y = F(x_1,x_2,..,\alpha_1,\alpha_2,..) + \varepsilon$$

lineáris regresszió : a paraméterekben lineáris az F függvény

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 + \varepsilon$$

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 + \alpha_4 x_1^2 + \alpha_5 x_2^2 + \varepsilon$$

a feltétel szerint ε egy 0 várható értékű, azonos paraméterű normál eloszlású

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Lineáris regresszió

a paraméterek várható értékének meghatározása a legkisebb négyzetek elvével történik
elemi esetre:

$$y_i = \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 + \varepsilon_i$$

$$\varepsilon_i = y_i - (a_1 x_{i1} + a_2 x_{i2} + a_3)$$

$$E(\varepsilon_i) = 0$$

$$\Sigma \varepsilon_i^2 \rightarrow \text{ minimális}$$

a szélsőérték szükséges feltétele a deriváltak zérus értéke

$$\partial_{a1} \Sigma_i (y_i - (a_1 x_{i1} + a_2 x_{i2} + a_3))^2 = 0$$

$$\partial_{a2} \Sigma_i (y_i - (a_1 x_{i1} + a_2 x_{i2} + a_3))^2 = 0$$

$$\partial_{a3} \Sigma_i (y_i - (a_1 x_{i1} + a_2 x_{i2} + a_3))^2 = 0$$

Adatelemzési módszerek

Lineáris regresszió

egy változós esetre:

$$\partial_{a1} \Sigma_i (y_i - (a_1 x_i + a_2))^2 = 0$$

$$\partial_{a2} \Sigma_i (y_i - (a_1 x_i + a_2))^2 = 0$$

$$\partial_{a1} \Sigma_i (y_i^2 + a_1^2 x_i^2 + a_2^2 + 2 a_1 a_2 x_i - 2 y_i a_1 x_i - 2 y_i a_2) = 0$$

$$\partial_{a1} \Sigma_i (a_1^2 x_i^2 + 2 a_1 a_2 x_i - 2 y_i a_1 x_i + a_2^2 - 2 y_i a_2 + y_i^2) = 0$$

$$\partial_{a2} \Sigma_i (a_2^2 + 2 a_1 a_2 x_i - 2 y_i a_2 + a_1^2 x_i^2 - 2 y_i a_1 x_i + y_i^2) = 0$$

$$a_1 \Sigma_i x_i^2 + a_2 \Sigma_i x_i - \Sigma_i y_i x_i = 0$$

$$a_2 n + a_1 \Sigma_i x_i - \Sigma_i y_i = 0$$

$$a_1 = (n \Sigma_i x_i y_i - \Sigma_i x_i \Sigma_i y_i) / (n \Sigma_i x_i^2 - \Sigma_i x_i \Sigma_i x_i)$$

$$a_2 = (\Sigma_i y_i - a_1 \Sigma_i x_i) / n$$

Adatelemzési módszerek

Lineáris regresszió

Minta {(2.,4.),(4.,6.2), (6.,4.5)} és f(a,b,x) = ax+b

Excel megvalósítás:

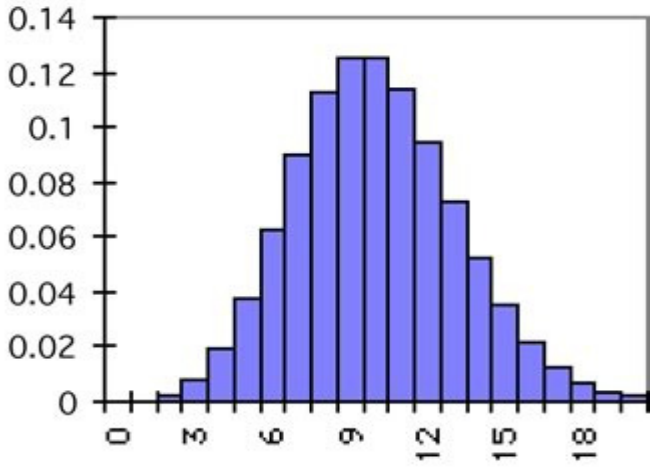
- elemi lépésekben

- regressziós görbével

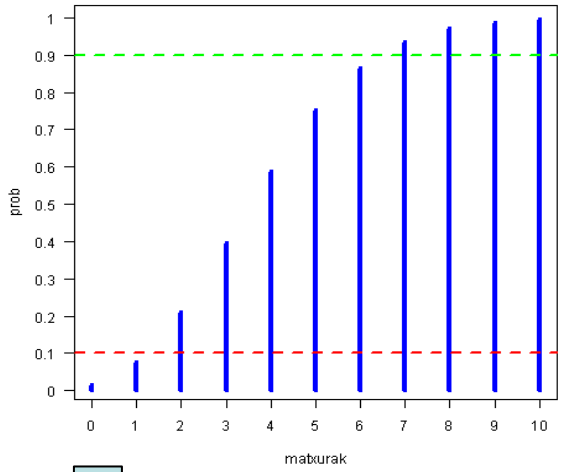
Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

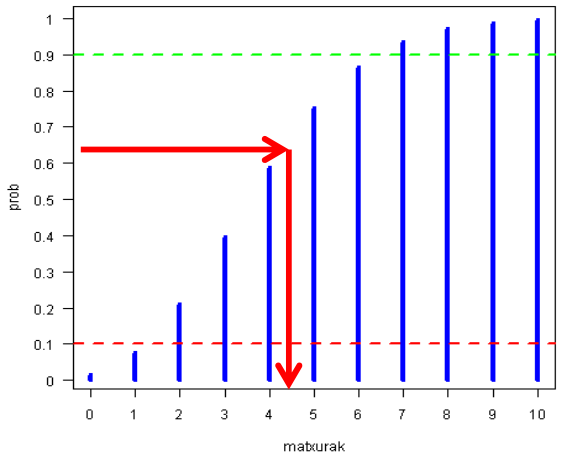
Minta generálása adott eloszlás függvényhez



sűrűség fv.



eloszlás fv.



érték

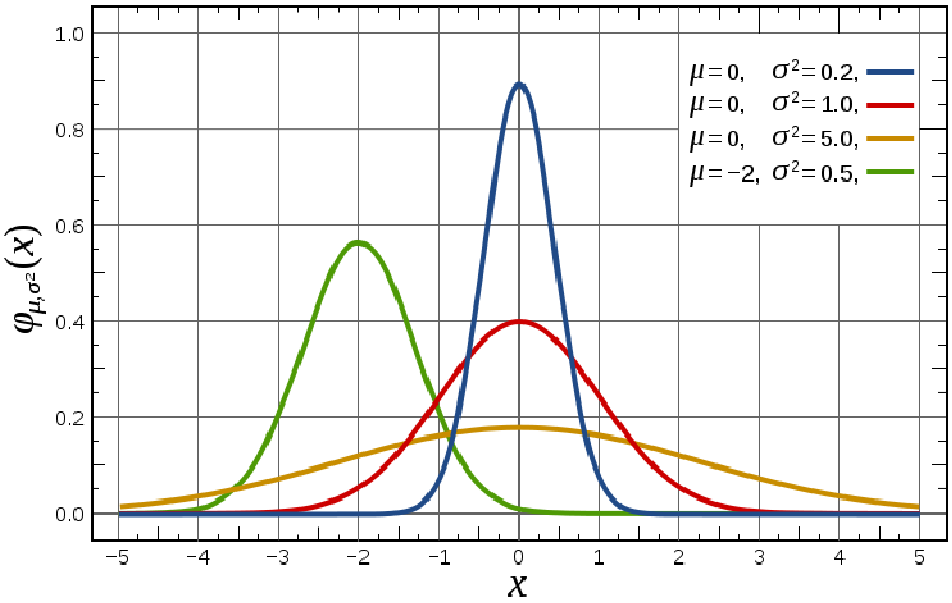
Adatelemzési módszerek

Normál eloszlás

Centrális határeloszlás tétele: Nagy n-re a mért empirikus átlagok normális eloszlást mutatnak

A x = (a – a')/ σ változó
N(0,1) normál eloszlású lesz

A N(0,1) eloszlás esetén az
|x| >≈ 2.8 pontok „lehetetlen" eseményeknek tekintetők



várható érték

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

szórásnégyzet

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Zajok szűrése

Zajok kiszűrésének módszere: az elosztást N(0,1)-re hozva a 2.5-nél nagyobb abszolút értékek zajoknak tekinthetők

Példa: 46, 48, 38, 45, 47, 58, 44, 45, 43, 44

46	0,2	458	0,04	25,73333	0,039426
48	2,2	45,8	4,84	5,072803	0,433685
38	-7,8		60,84		-1,53761
45	-0,8		0,64		-0,1577
47	1,2		1,44		0,236556
58	12,2		148,84		2,404982
44	-1,8		3,24		-0,35483
45	-0,8		0,64		-0,1577
43	-2,8		7,84		-0,55196
44	-1,8		3,24		-0,35483
a	a-a’	a’	$(a-a')^2$	σ	x

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Véletlenszerűség ellenőrzése

Wald-Wolfowitz teszt: figyeli a sorozatok (runs) eloszlását (ne legyen se túl kevés, se túl sok sorozat)

Induló adatsor: mérési értékek

Lépések:

- a' átlag kiszámítása
- sig(a-a') –val helyettesítjük a-kat
- n^+, n^- (elemek db), R (sorozatok száma) meghatározása
- a" =1 + $2n^+n^-/(n^+ + n^-)$
- $\sigma^2 = (a"-1)(a"-2)/(n^+ + n^- - 1)$.
- $z = (R-a") / \sigma$
- ha $|z| > z_0$ akkor nem véletlen a sorozat (~2.5)

Dr. Kovács László
ME GEIAL

Adatelemzési módszerek

Véletlenszerűség ellenőrzése

Példa: 3, 5, 12, 7, 9, 8, 21, 17, 87, 22, 18, 24

Excel megvalósítás:

- elemi lépésekben

- regressziós görbével

Dr. Kovács László
ME GEIAL

Adatelemzések statisztikai háttere

A statisztikusok is ideális világból indulnak ki.

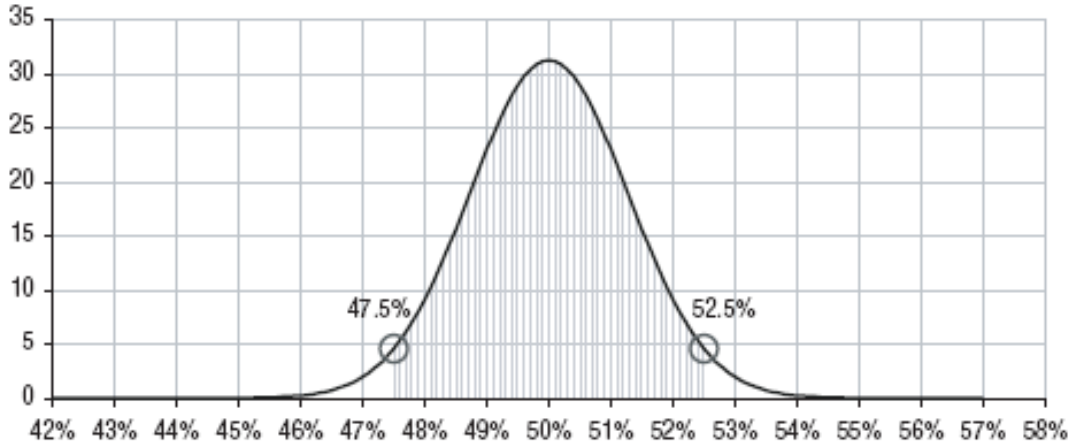
Null-hipotézis elve: a mérési eltérés csak a véletlen műve

A vizsgálat tárgya:

- milyen paraméterű az ideális eloszlás?
- mennyire teljesül a null-hipotézis?

A mérési adatokon próbákat lehet végrehajtani a hipotézis ellenőrzésére, a hipotézis konfidencia szintjének megállapítására

Az elemzés megadja, hogy milyen konfidencia értékkel fog a paraméter egy megadott konfidencia intervallumba esni.



Dr. Kovács László
ME GEIAL

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-eloszlás

Feltétel: legyenek X_i független normál eloszlású változók, (μ,σ) paraméterekkel.

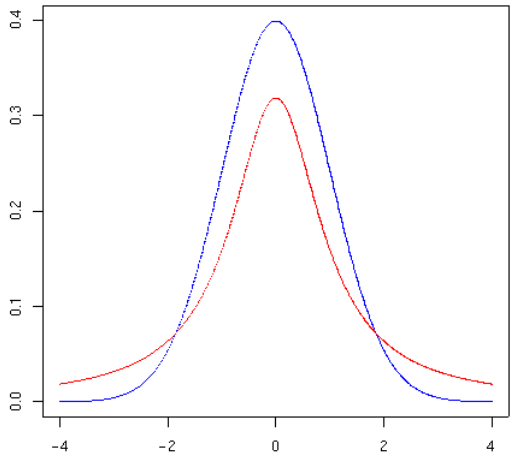
Ekkor a minta átlaga (n: mintaszám): $\overline{X} = \frac{\sum X_i}{n}$

minta szórásnégyzete: $S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$

Ekkor normál(0,1) eloszlású: $\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$

Illetve Chi-négyzet eloszlású: $\frac{(n-1)S^2}{\sigma^2}$

Emiatt T Student eloszlású lesz: $\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$



Dr. Kovács László
ME GEIAL

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-teszt

Egyváltozós eset.

- null hipotézis: az eloszlás várható értéke: μ

- feladat: a tapasztalati eloszlás illeszkedik-e?

- vizsgált eloszlás: $$t = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

- függetlenségi tényező: $n - 1$

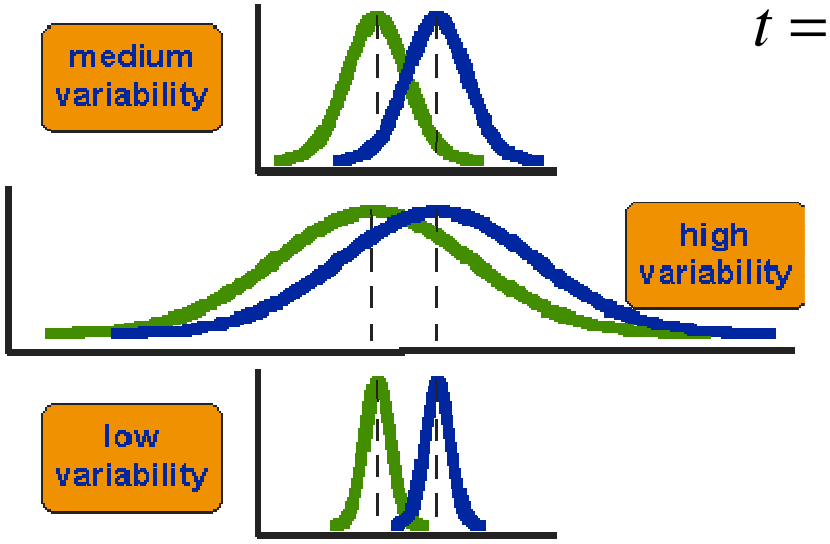
Dr. Kovács László
ME GEIAL

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-teszt

Kétváltozós eset.

A feladat adott kontroll és mérési eloszlás mellett eldönteni, hogy a mérési eloszlás mennyire illeszkedik a kontrollra



$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{(n_1 - 1)S^2{}_1 + (n_2 - 1)S^2{}_2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 n_2}}}$$

Függetlenségi tényező: $n_1 + n_2 - 2$

Dr. Kovács László
ME GEIAL

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-teszt

T-tábla használata:

- az oszlop jelöli a konfidenciát
- a sor jelöli a függetlenségi értéket
- a cella adja meg az előírt maximum t értékek

(ha a tábla érték nagyobb mint a számított, akkor megtartjuk a hipotézist)

FD	0.1	0.05	0.01
5	2.02	2.57	4.03
6	1.94	2.45	3.71
7	1.89	2.37	3.50
9	1.83	2.26	2.68
20	1.72	2.09	2.85
30	1.70	2.04	2.75

5% a kockázat, hogy úgy vetjük el a hipotézist, hogy mégis igaz

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-teszt

Adott az alábbi adatsor: 483, 502, 498, 496, 502, 483, 494, 491, 505, 486.

Kérdés: tekintető-e 5%-os kockázat mellett a eloszlás m=500 várható értékűnek?

$$t = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

X=494, S = 8.05, μ=500, n=10, df=9

t=2.36, tablazat:2.26

megoldás: nem fogadható el a hipotézis

Dr. Kovács László
ME GEIAL

Adatelemzések statisztikai háttere

Hipotézis vizsgálat, T-teszt

Adott az alábbi adatsor, két eltérő helyen élő egyedhalmaz súlyértékei:

X1: 52; 57; 62; 55; 64; 57; 56; 55

X2: 41; 34; 33; 36; 40; 25; 31; 37; 34; 30; 38.

Kérdés: tekintető-e azonosnak a két eloszlás 5%-os kockázat mellett?

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{(n_1 - 1)S^2{}_1 + (n_2 - 1)S^2{}_2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 n_2}}}$$

X1=57.25, X2=34.45

n1=8, n2=11

$S1^2$=15.36, $S2^2$=21.87

t=11.12, táblázat=2.11

megoldás: nem fogadható el a hipotézis

Dr. Kovács László
ME GEIAL