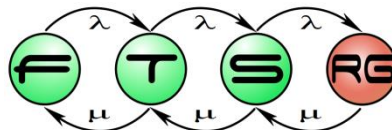


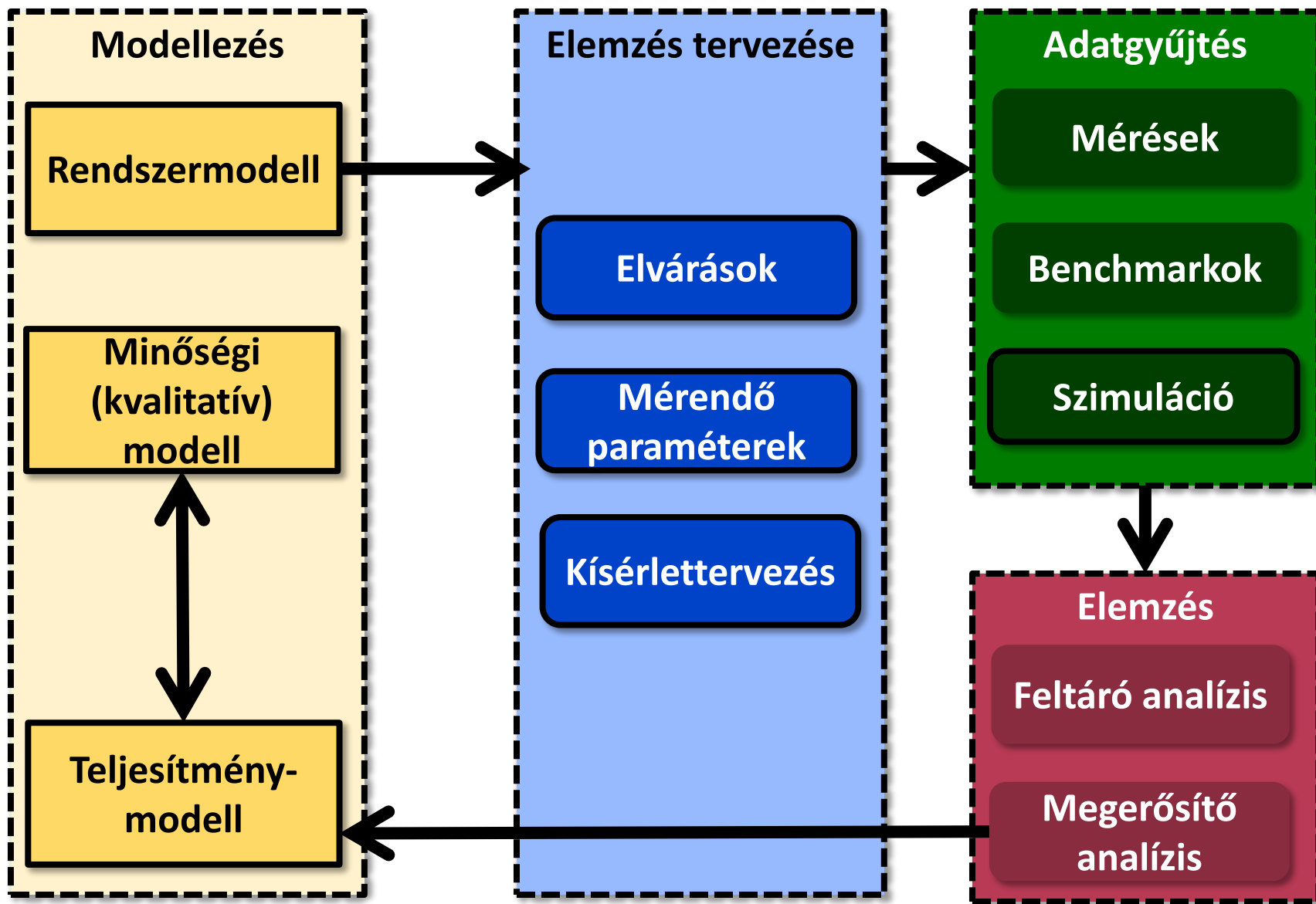
# Modellek paraméterezése: regresszió, benchmarkok

Rendszermodellezés 2017.

**Budapest University of Technology and Economics**  
**Fault Tolerant Systems Research Group**

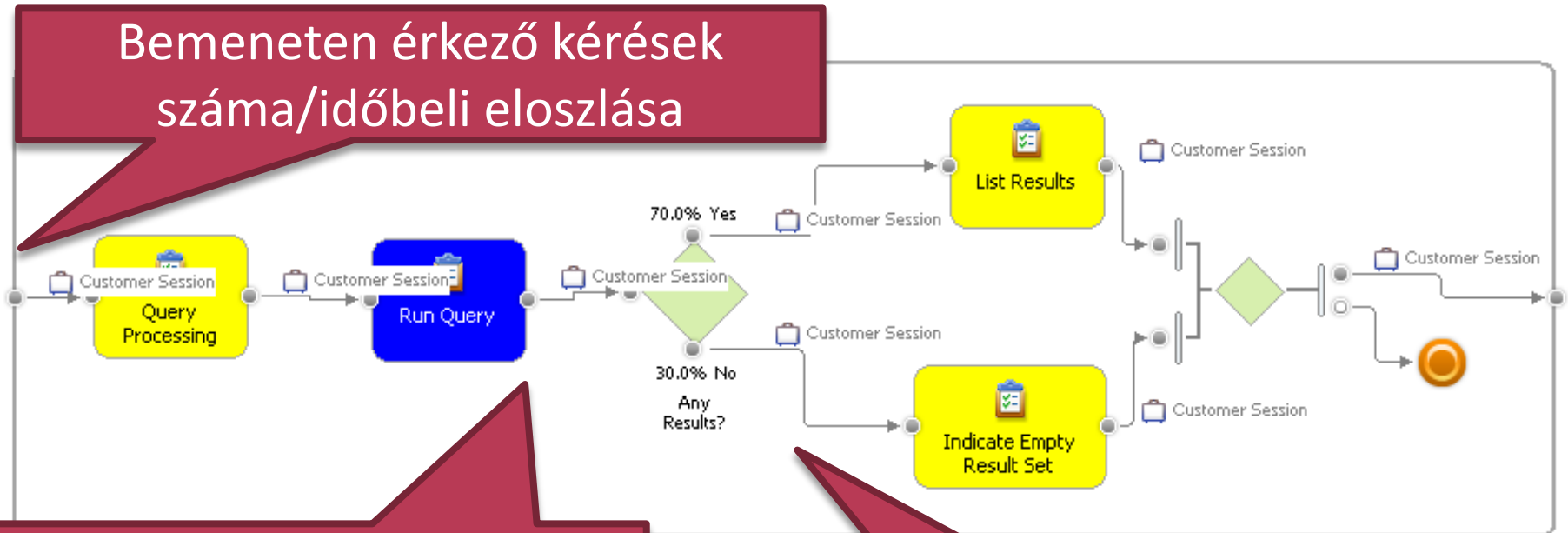


# Rendszermodelltől a teljesítménymodellig



# Alap kérdés

- Jól becsüljük meg a mennyiségi paramétereket?



Egy adott tevékenység végrehajtási ideje adott erőforráson

Várható értékkel közelített döntési valószínűség/gyakoriság

# Adatok hihetősége

## ■ Érzékenységvizsgálat

- A modell **kimeneti** paraméterei mennyire érzékenyek a modell **bemeneti** paramétereinek változására
- (erőforrások száma/kapacitása, felhasználók döntései)  
→ (folyamat válaszideje, átbocsátása)
- „parameter sweep”: egy paraméter vizsgálata adott tartományban  
→ Melyik paramétert mennyire kell tudnunk becsülni?

## ■ Ökölszabály: adatok hihetősége

- mérés\_bizonytalansága (szórás)  $\sim$  mérések\_száma<sup>2</sup>
- (kellő mennyiségű adat esetén) → Val.szám.

# MATEMATIKAI BECSLÉSEK: REGRESSZIÓS MÓDSZEREK

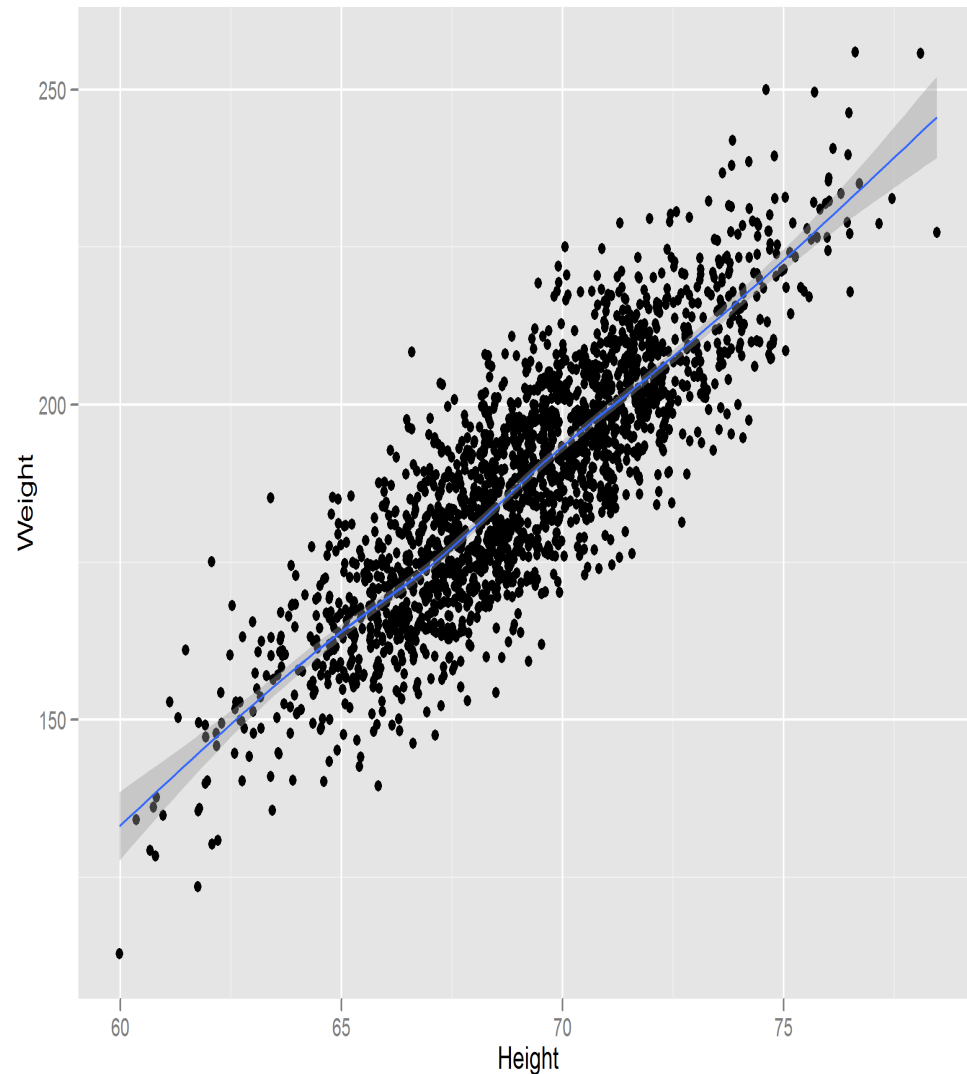
# Feladat

- Adott sok változó, hosszú időre
- Szeretnénk bizonyos változók értékét közelíteni
  - Nehéz mérni
  - Nem áll rendelkezésre
- Szeretnénk becsülni/előrejelezni
  - Még nem történt meg, az idő függvényében közelítjük
  - Nem volt adott bemenet (pl. adott számú felhasználó)
  - Csak később lesz látható egy hatás (pl. válaszidő csak a kérések kiszolgálása közben nő meg a várakozás miatt)
- Mennyire bízhatunk meg a feltételezésekben/következtetéseinkben?

# Regresszió

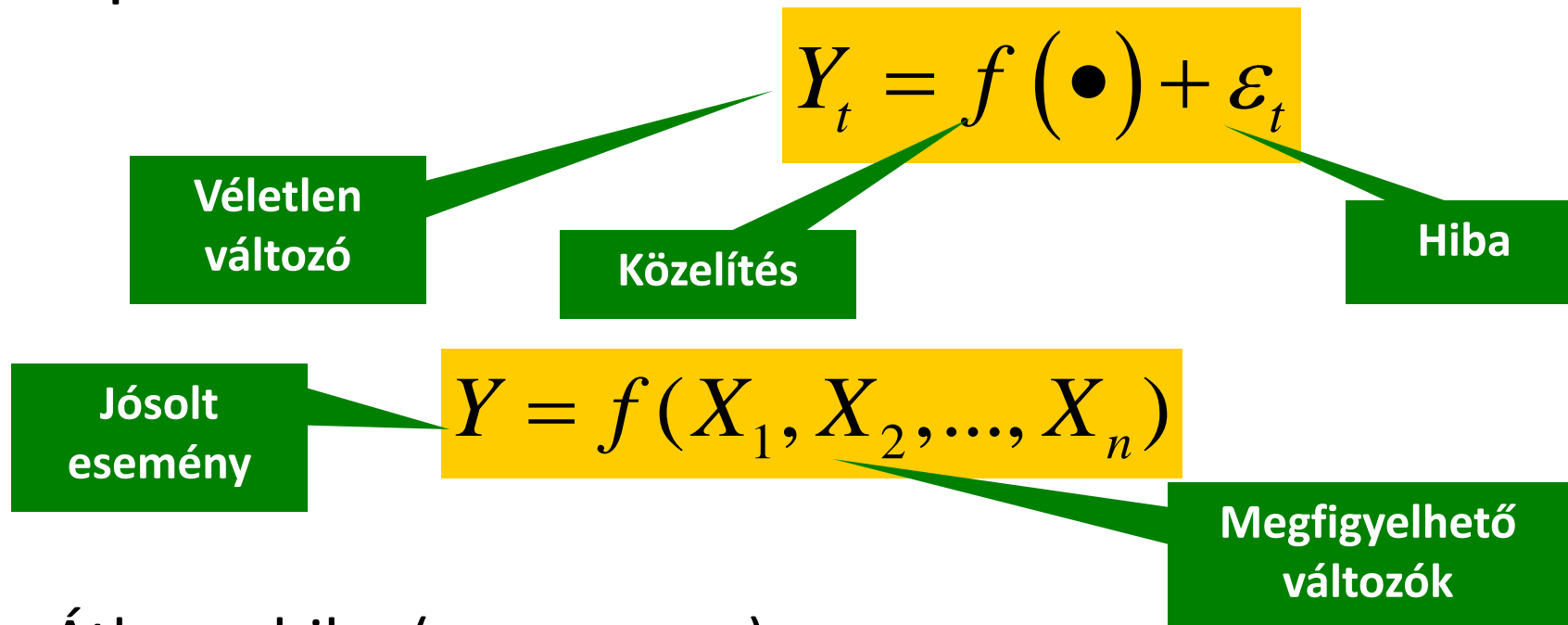
f függvény,

- bemenet:  
az attribútumok értéke,
- kimenet: megfigyelések  
legjobb közelítése
- „ökölszabály”
- Példa:  
testtömeg/magasság  
együttes eloszlás  
valójában egyenesre  
illeszthető

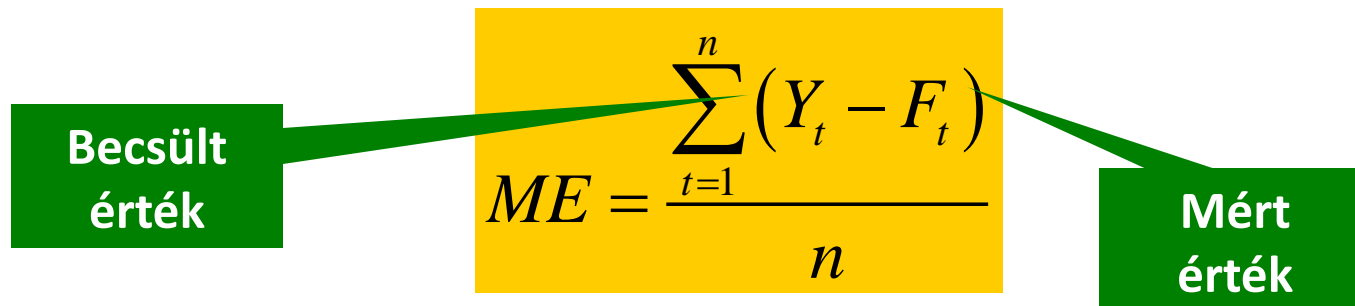


# Regressziós módszerek

## ■ Alapelv:



## • Átlagos hiba (mean error)





# Lineáris regresszió

- Egyszerű lin. függvény illesztése az adatokra
  - nem vár alapvető változást a rendszer viselkedésében

$$Y = a + bX$$

- Legkisebb négyzetek módszere
  - keressük azokat az  $a, b$  paramétereket (itt:  $a$ : eltolás,  $b$ : meredekség), amelyekre

$$SSE = \sum_{t=1}^n \varepsilon_t^2 = \sum_{t=1}^n (Y_t - F_t)^2 \quad \text{minimális (Sum of Squared Errors)}$$

- cél: 
$$\sum_{t=1}^n (Y_t - F_t)^2 = \sum_{t=1}^n [Y_t - (a + bX_t)]^2$$

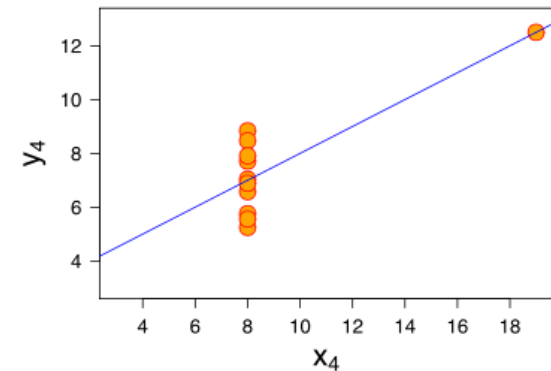
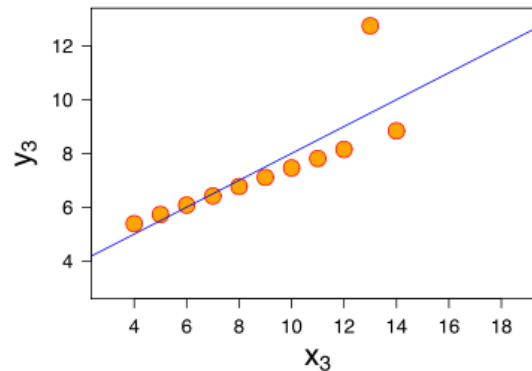
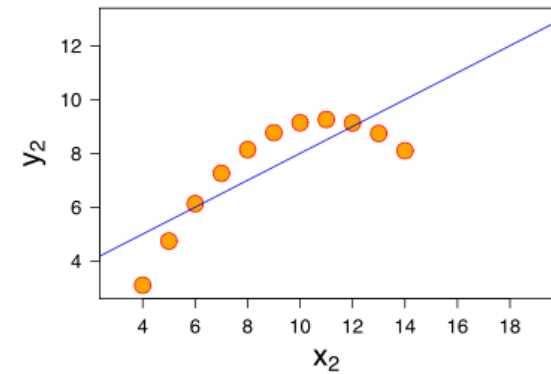
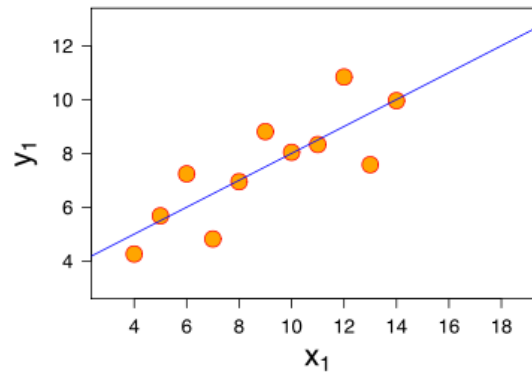
# Lineáris regresszió

- Legjobban illeszkedő egyenes

■ **DE:**

Anscombe's quartet

- Minőségileg különböző adatok
- Azonos regressziós egyenes



# Lineáris regresszió (folyt.)

## ■ Korrelációs együttható (négyzete)

- a változó becsült és tényleges értékének kapcsolata
- 0 és 1 közti érték
- 0: nincs kapcsolat
- 1: függvényszerű kapcsolat
- R maga: -1 .. 1 (kapcsolat iránya)

$$R^2 = \frac{\sum_{t=1}^n (F_t - \bar{Y})^2}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

## ■ Példa: E-mail szolgáltatás, 8 hétig mérjük a csúcsterhelést.

Hét	1	2	3	4	5	6	7	8
Max. terhelés (email/perc)	420	410	437	467	448	460	507	514

Hogyan közelíthető a terhelés változása?

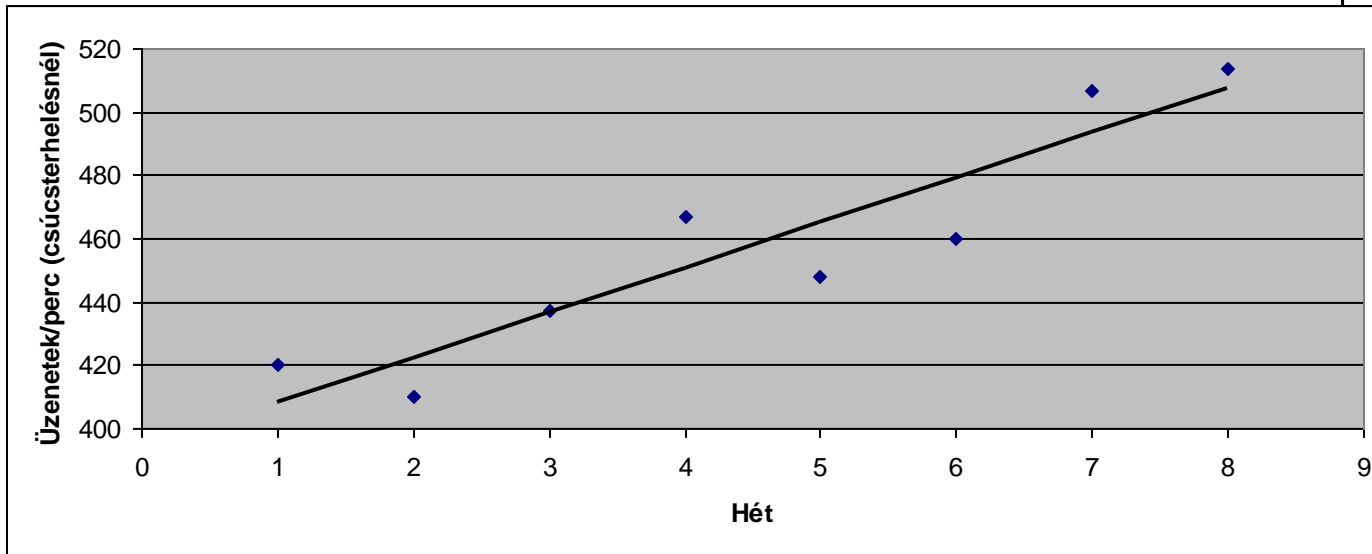
Mekkora a korrelációs együttható?

# Lin. regresszió példa

A legkisebb négyzetek módszerével  
 $Y=393.98+14.20X$

Korrelációs együttható:  
 $R^2=0.855$

Mért	Jósolt terhelés
420	408.18
410	422.38
437	436.58
467	450.78
448	464.98
460	479.18
507	493.38
514	507.58
	521.78



# Két változó kapcsolatának vizsgálata

- Tfh. lineáris kapcsolat van az egyszerre bejelentkezett felhasználók száma és az elküldött emailek közt. (pl logok alapján)

Bejelentkezett felh. átlagos száma (1 óra alatt)	2450	2765	2241	2860	3011	2907	3209
Átl. terhelés (kimenő+bejövő emailek/óra)	19257	20488	18152	21450	21077	20639	22142

- Lineáris regressziós közelítés a legkisebb négyzetek módszerével:

*ÜzenetekSzám* =  $f(\text{BejelentkezettFelhasználók})$

$Y=9480.48 + 3.95X$ ,  $R^2=0.937$  erős kapcsolat

# Nemlineáris módszerek

- Exponenciális közelítés:  $Y_t = a \times b^t$ 
  - jól illik a Web forgalom növekedéséhez
- Átalakítjuk a függvényt:
$$\log Y_t = \log a + t \log b$$
$$\log Y_t = Y', \log a = a', \log b = b'$$
$$Y' = a' + b't$$
- Legkisebb négyzetek módszere használható
- Pl. adottak a legnagyobb mért terhelés értékei

Mennyi a várható legnagyobb terhelés az év végén?

Hónap	1	2	3	4	5	6	7	8	9	10
Max. kérések/sec ( $Y_t$ )	1035	1100	1160	1250	1350	1555	1770	1950	2210	2630
$\ln(Y_t)$	6.942	7.003	7.056	7.13	7.207	7.349	7.478	7.575	7.7	7.874

# Exp. terhelés példa

- Becslőfüggvény:  $Y_t = a \times e^{bt}$
- Legkisebb négyzetek módszere a lineáris függvényre

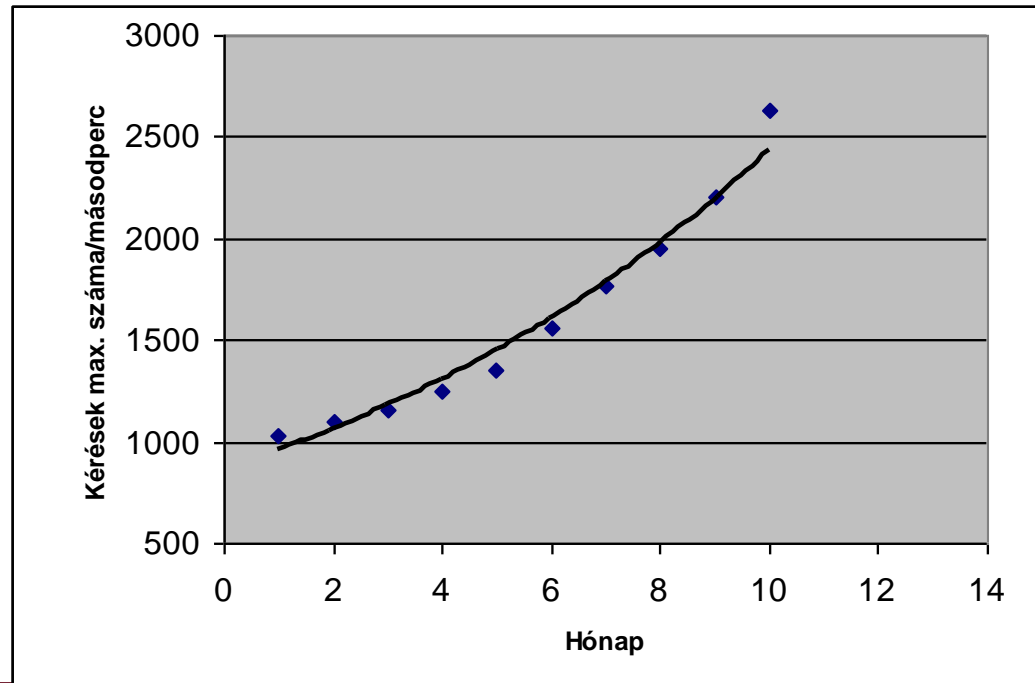
$$Y' = a' + b't, a' = 6.717, b' = 0.110, a = e^{a'}$$

- Eredmény:

$$Y_t = 826.33 \times e^{0.11t}$$

- 12. hónap:

$$Y_t = 3093.3$$



# Mozgó átlagok módszere

- Rövid távú előrejelzésre jó
- Egyszerre egy értéket ad meg
- A becsült érték az utolsó  $n$  érték átlaga

$$F_{t+1} = \frac{\sum_{i=t-n+1}^{t-n+1} Y_i}{n}$$

ahol  $Y_t$  a  $t$ . időpontban mért érték

$F_{t+1}$  a becsült érték

$n$  tipikusan 3 és 10 között van

(becslés hibája ne legyen túl nagy)



# Exponenciális csúszóablak

- Egy értéket ad meg, az előző méréseket átlagolva
- Későbbi mérés (és mérési hiba) nagyobb súllyal
- Rövid távú előrejelzésre alkalmas
  - (miért exponenciális?)

$$F_{t+1} = F_t + \alpha(Y_t - F_t)$$

ahol  $F_t$  : a t. időpontra becsült érték

$Y_t$  : t. időpontban mért érték

$Y_t - F_t$  : mérési hiba a t. periódusban

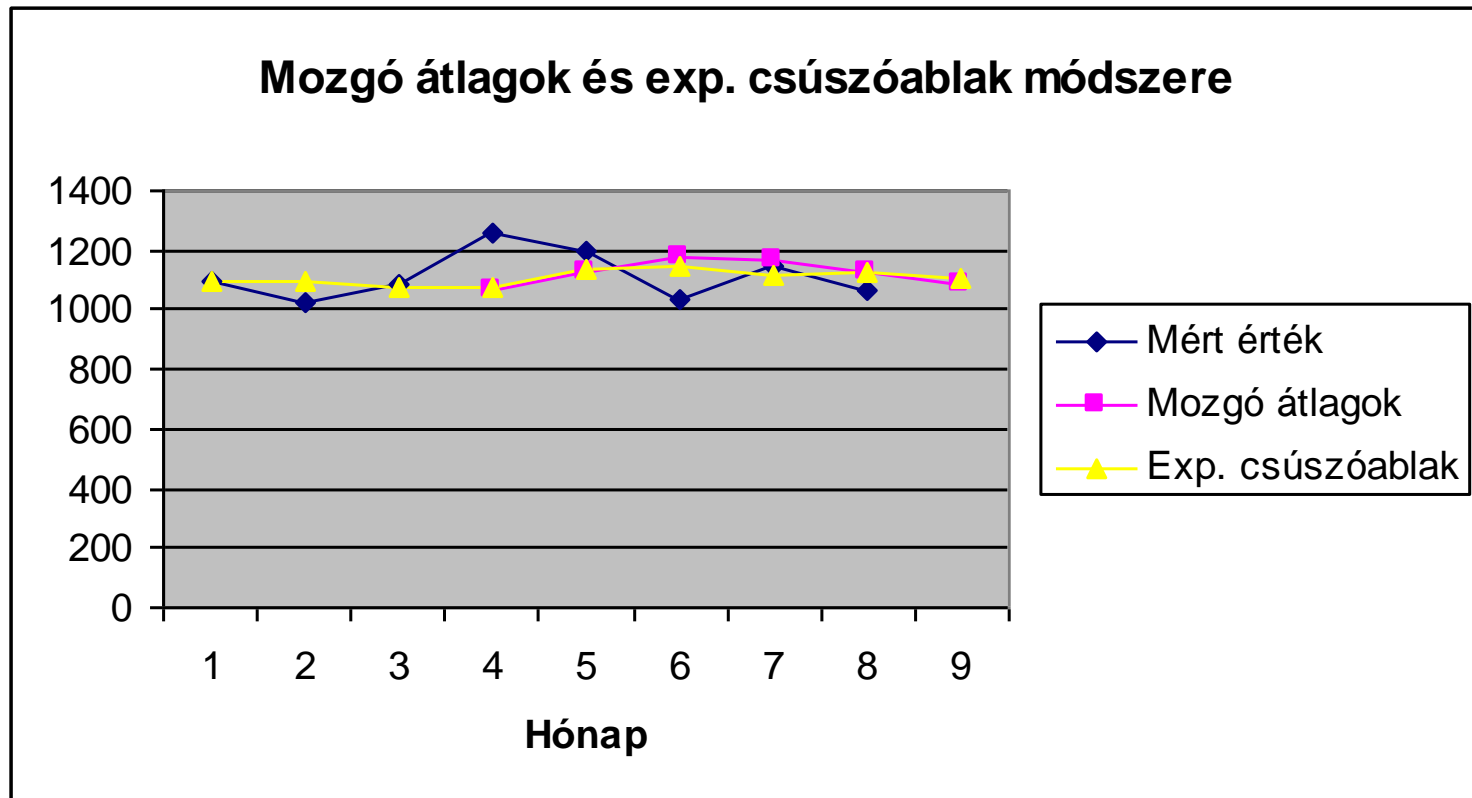
$\alpha$  : súlyozás ( $0 \leq \alpha \leq 1$ , gyakorlatban  $0.05 \leq \alpha \leq 0.3$ )

# A két módszer összehasonlítása

- Adott sávzélesség igények, a két módszerrel becsüljük a következő értéket

Hónap	Sávzélesség igény	Mozgó átlagok módszere (n=3)	Exp. csúszóablak ( $\alpha = 0.3$ )
1	1100		1100.00
2	1020		1100.00
3	1090		1076.00
4	1255	1070.0000	1080.20
5	1195	1121.6667	1132.64
6	1039	1180.0000	1151.35
7	1145	1163.0000	1117.64
8	1066	1126.3333	1125.85
9		1083.3333	1107.90

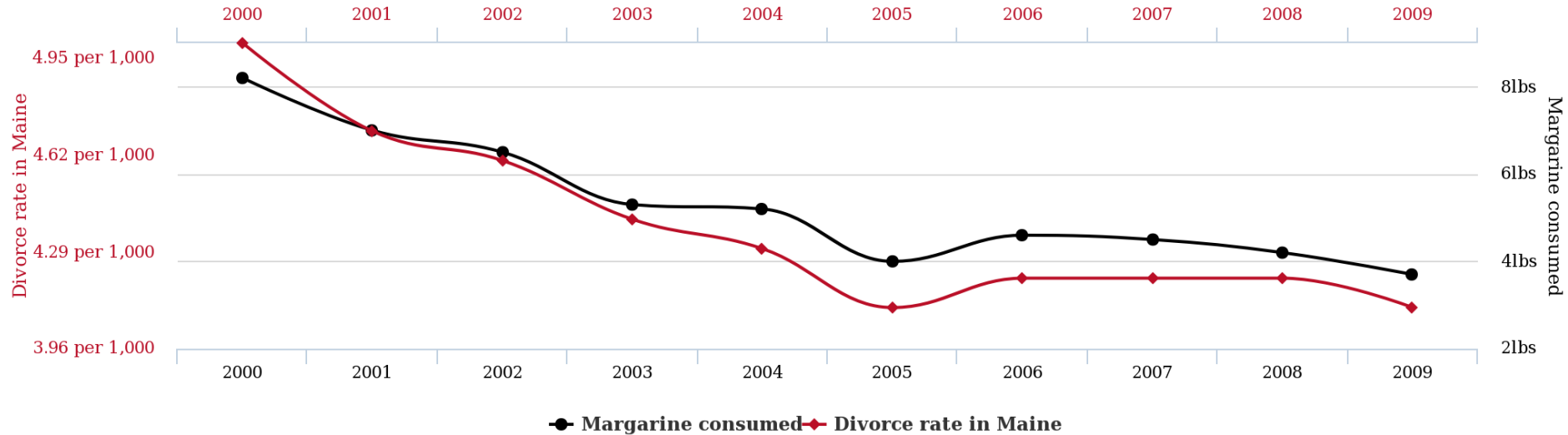
# A két módszer összehasonlítása



# Mire figyeljünk?

Kauzalitás (oksági kapcsolat) != Korreláció (együttes előfordulás)

## Divorce rate in Maine correlates with Per capita consumption of margarine



tylervigen.com

Informatikai példa: sok felhasználó → nagyobb kihasználtság ÉS magas válaszidő

# DEMÓ

Előrejelzés alapú kapacitástervezés, érzékenységvizsgálat

# BENCHMARKING

# Why benchmark?



# Benchmarking - Definíció

## ■ Wikipedia

*„In **computing**, a benchmark is the **act of running** a computer program, a set of programs, or other operations, in order to **assess the relative performance** of an object, normally by running a number of **standard tests** and trials against it.”*

## A benchmarkolás

- egy **program** (programok, vagy más műveletek) **futtatása**
- **szabványos tesztekkel** vagy bemenetekkel
- egy objektum **relatív teljesítményének felmérése** érdekében



# Benchmarking motiváció

- Fejlesztés alatt lévő rendszer
- Teljesítmény felmérése
  - Modell alapján becsült értékek
  - Megvalósított rendszer értékei „éles” helyzetben
- Tervezői és menedzsment döntések
  - Melyik részébe fektessünk több energiát ( $\approx$  pénzt)?
- Mi alapján döntsünk?
  - Jelenlegi erősségek és gyengeségek
- Mi számít erősségnek vagy gyengeségnek?
  - Mire képesek a hasonló rendszerek, versenytársak?

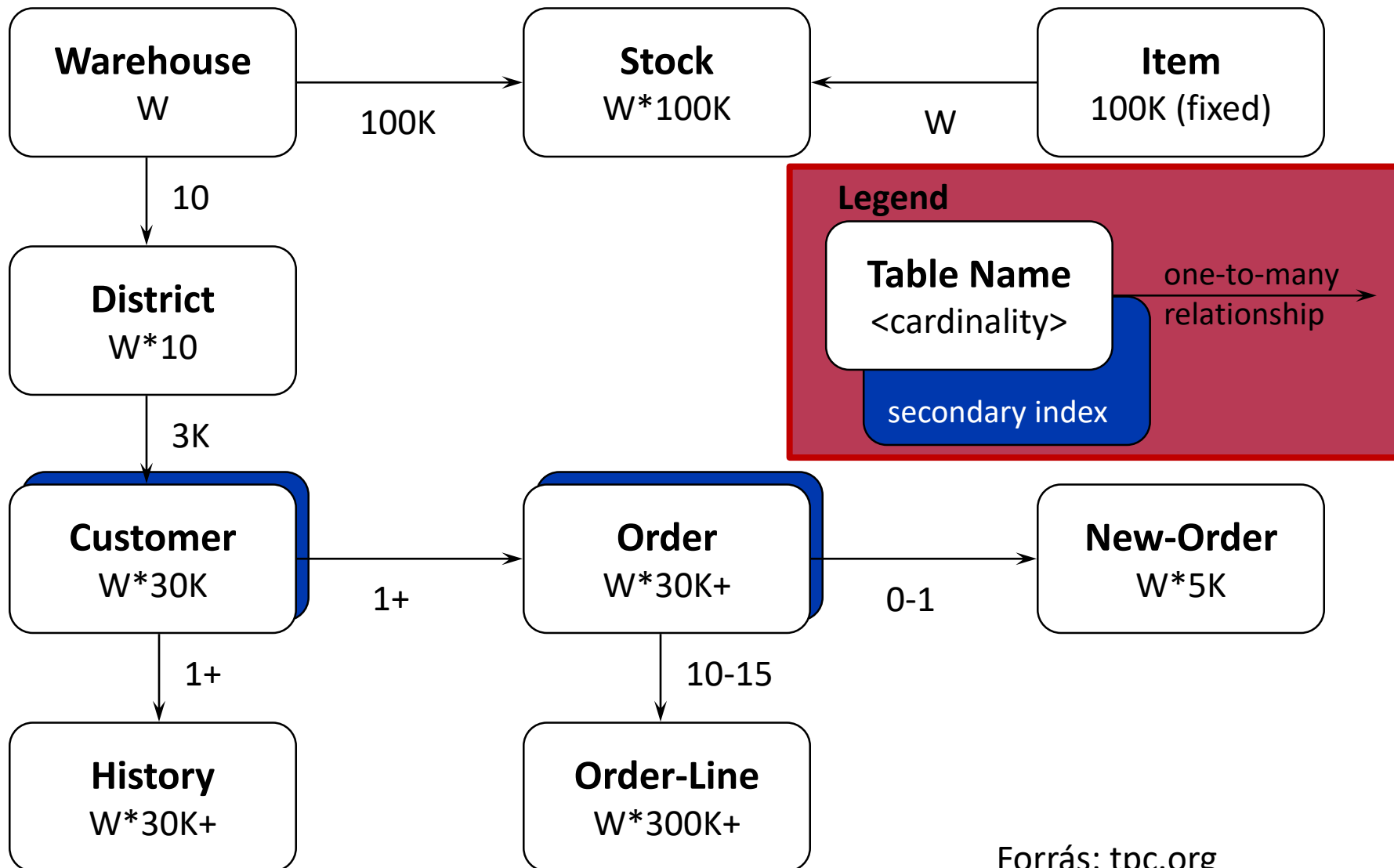
# Elvárások

- **Ismételhetőség (Repeatability)**
  - Egy elemen többször megismételt mérés/művelet eredményeinek változékonysága (szóródási faktor)
- **Reprodukálhatóság (Reproducibility)**
  - A mérési rendszer változékonysága, amelyet a műveletek viselkedéseinek különbsége okoz
- **Általánosított felhasználói eset**
  - Átlag felhasználó számára értelmezhető legyen az eredmény

- OLTP benchmark
  - Online Transaction Processing
  - Szoftverek és hardverek OLTP teljesítményét méri
- Komplex benchmark
  - Több tranzakció típus
  - Összetett adatbázis séma
  - Széles skálán mozgó adathalmaz méret
- Mért metrikák
  - Áteresztőképesség: transactions per minute (*tpmC*)
  - Ár-teljesítmény arány ( $\approx$  hatékonyság):  $\$/\text{tpmC}$

- **Mintaadatbázis**
  - Nagykereskedelmi beszállító cég rendelései
  - 9 különböző tábla
- **5 féle tranzakció típus**
  - Rendelés, fizetés, szállítás, rendelés státusza, raktár állapota
  - Frissítés, beszúrás, törlés, megszakítás
- **ACID tranzakciók**
  - Atomicity, Consistency, Isolation, Durability
- **Változatos felhasználói kérések szimulációja**
  - Tranzakció típusok kezdeményezése adott valószínűséggel
- **Időkorlát az egyes tranzakció típusokra**

# TPC-C adatbázis séma






Forrás: tpc.org

# KIÉRTÉKELÉS

# Adattisztítás

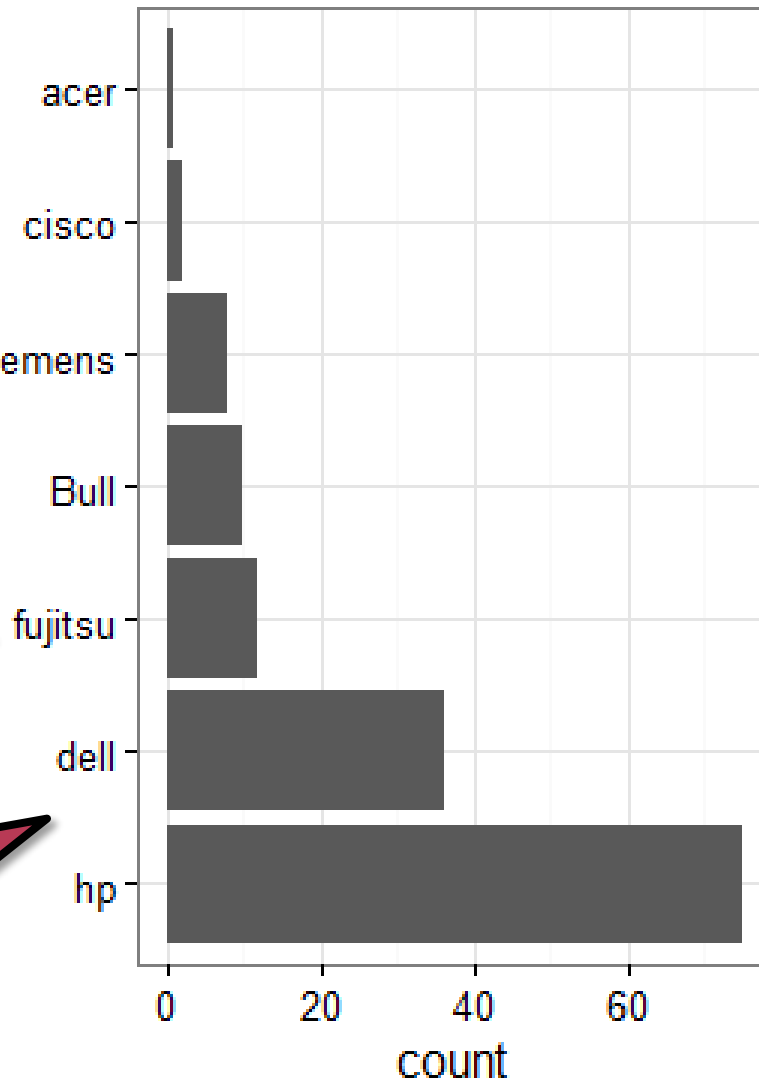
- Pénznem egységesítés (EUR, USD, AUD, JPY)
- Tizedesjegy, tízes elválasztó ellenőrzés

Hardware Vendor	System	tpmC	Price/tpmC	Watts/KtpmC	System Availability	Database	Operating System	TP Monitor	Date Submitted
	<a href="#">Altos R710 12GB/3.6GHz/2P C/S with 4 Altos R710</a>	66,543	12.42 AUD	NR	12/01/05	Microsoft SQL Server 2000 Enterprise Edition SP3	Microsoft Windows Server 2003 Enterprise Edition	Microsoft COM+	01/12/06
	<a href="#">Bull Escala PL6460R</a>	6,085,166	2.81 USD	NR	12/15/08	IBM DB2 9.5	IBM AIX 5L V5.3	Microsoft COM+	06/15/08
	<a href="#">Bull Escala PL860R</a>	629,159	2.49 USD	NR	06/11/08	IBM DB2 9.5 Enterprise Edition	IBM AIX 5L V5.3	Microsoft COM+	06/11/08

# Milyen rendszereket vizsgálunk?

Egységesítsünk?

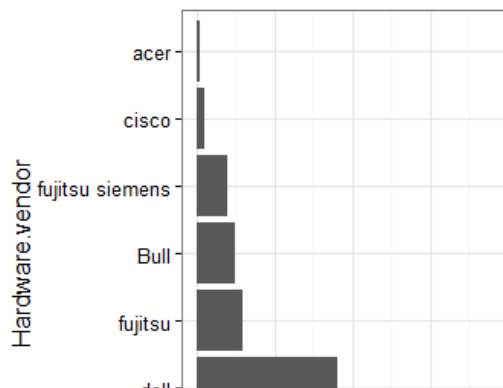
Hardware vendor



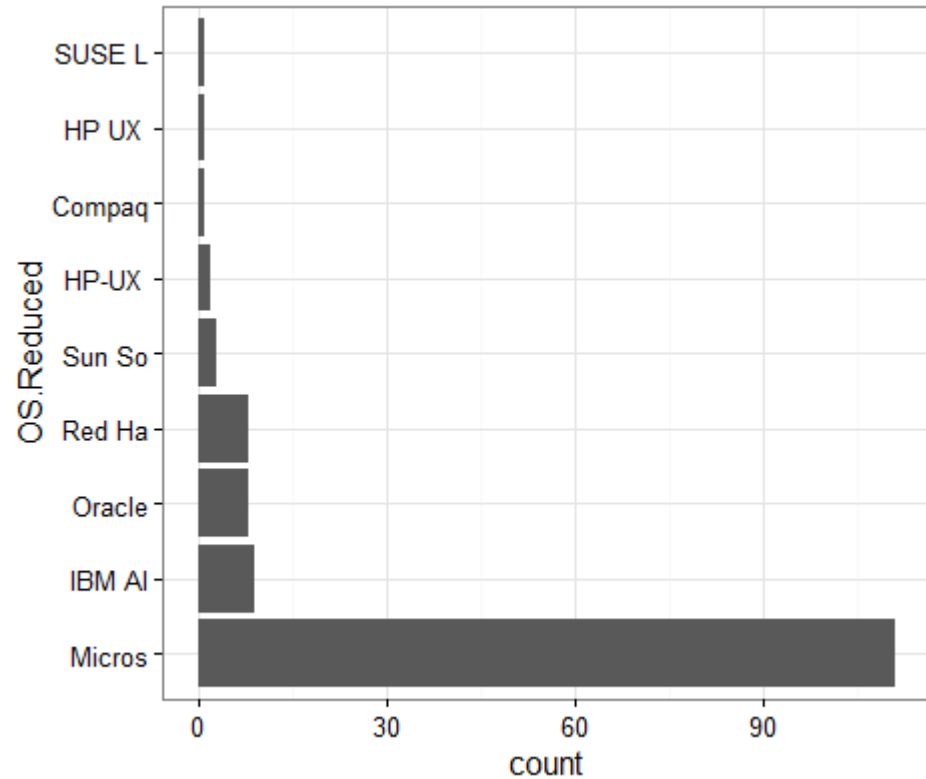
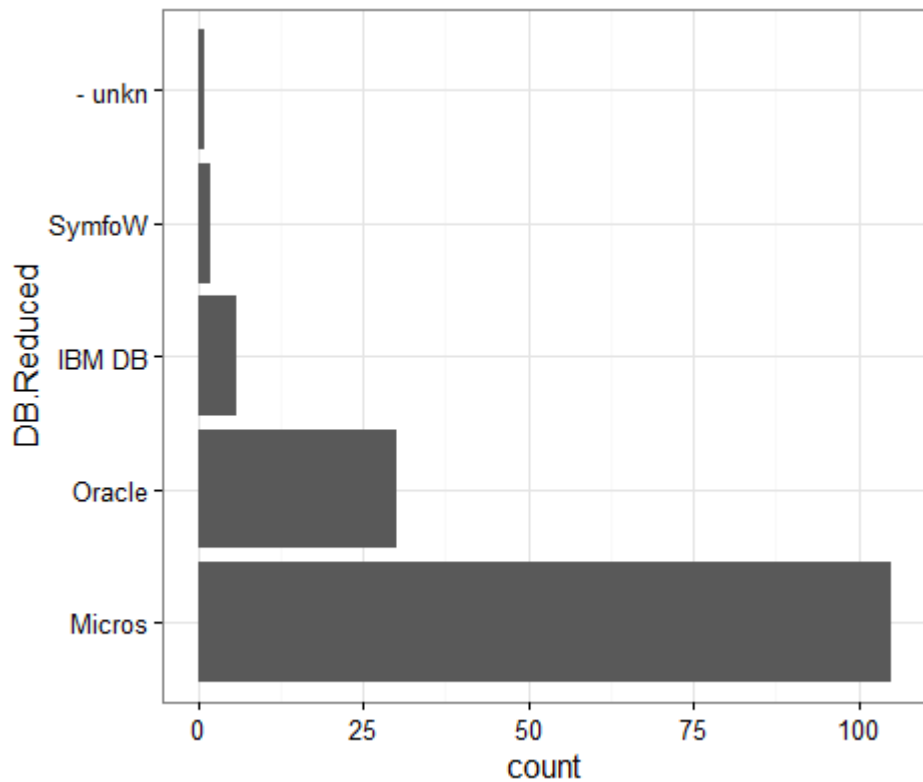
A két leggyakoribb gyártó az esetek 77%-ában szerepel



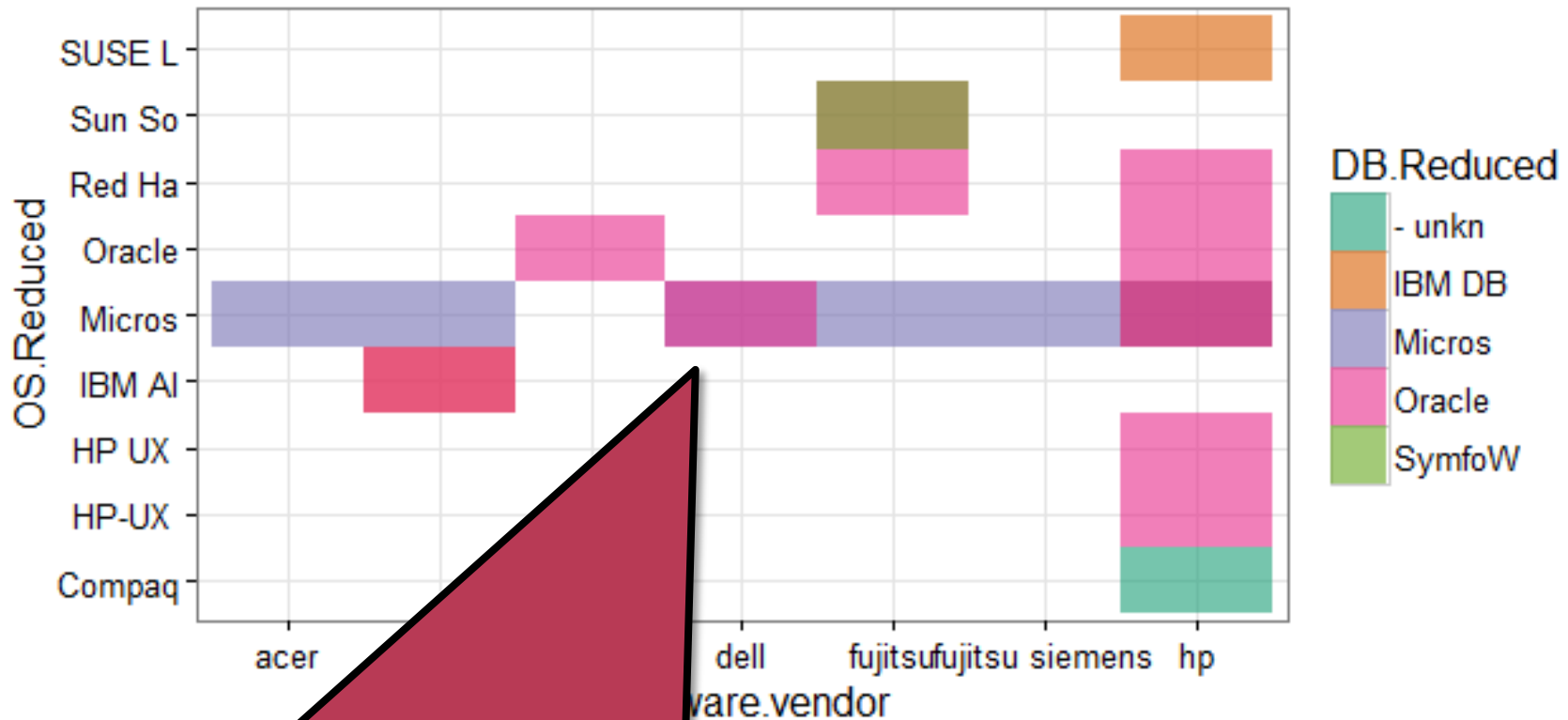
# Milyen rendszereket vizsgálunk?



Típus szerinti  
összevonás?

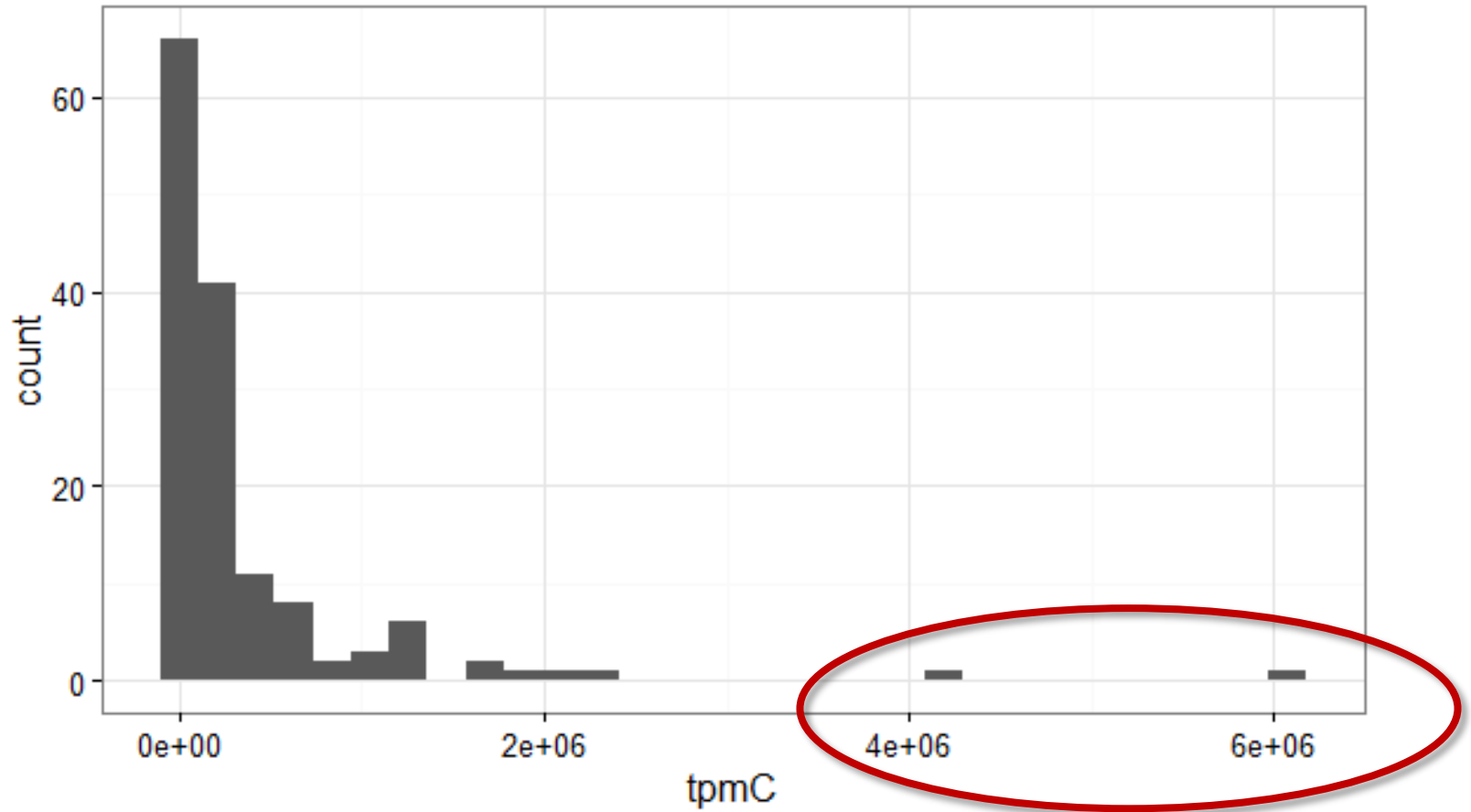


# Milyen rendszereket vizsgálunk?



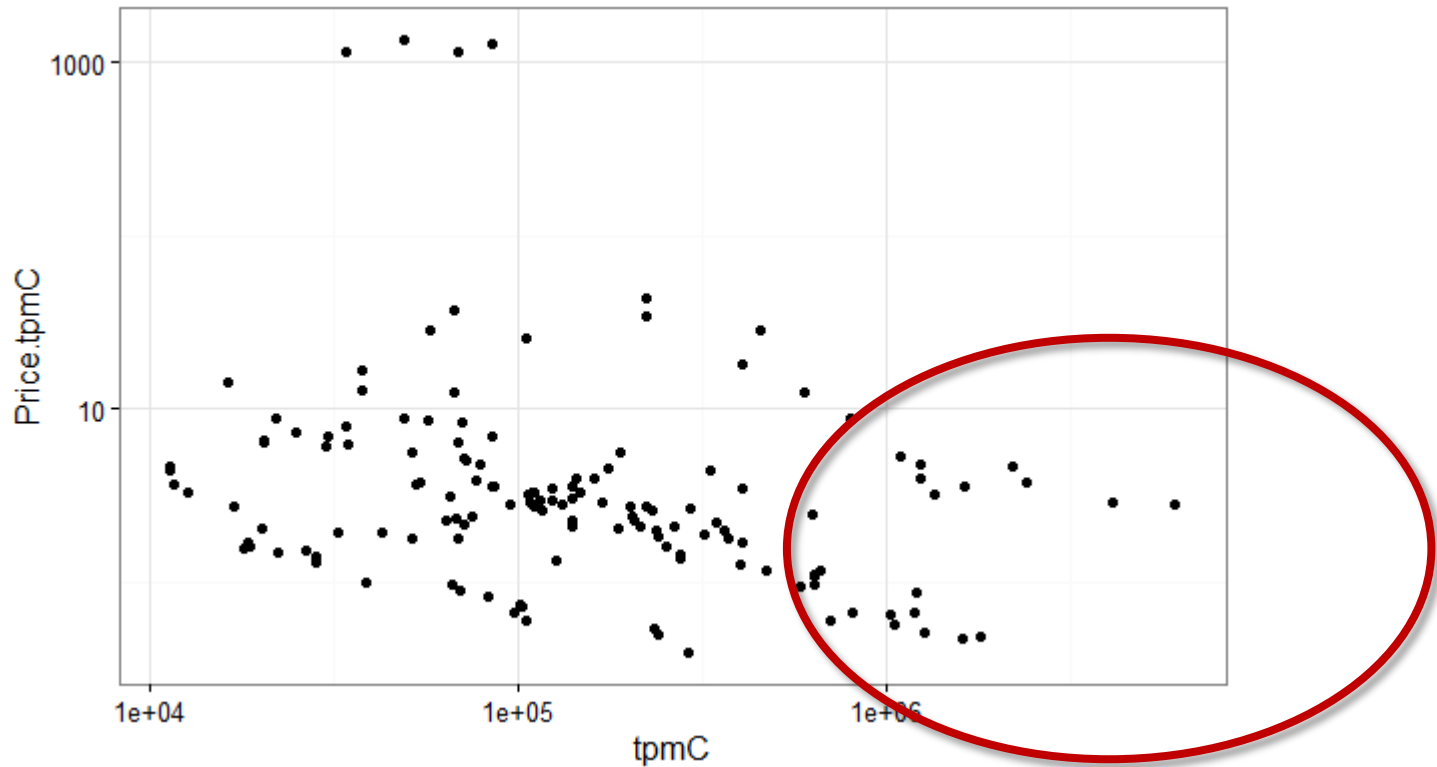
Általában többféle adatbázis többféle operációs rendszeren fut

# Benchmark eredmények



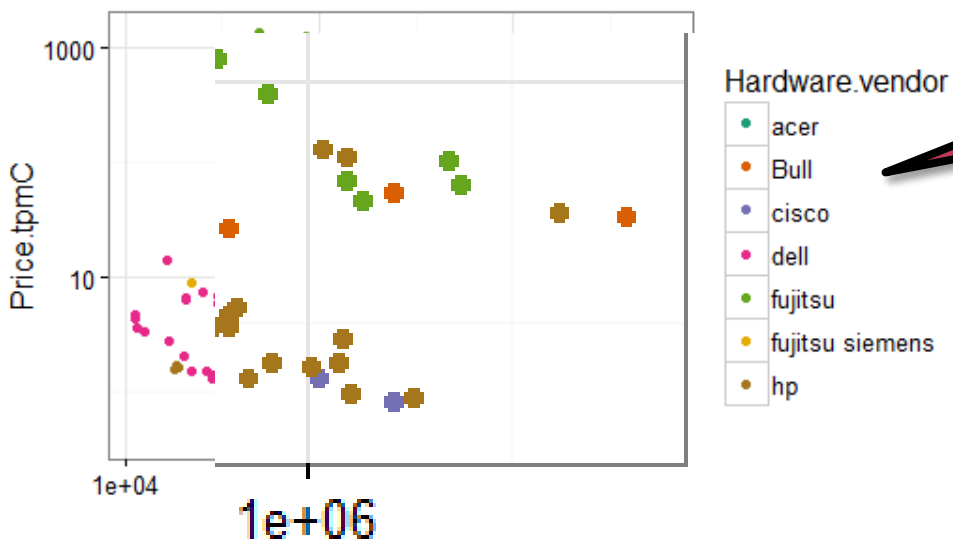
Milyen az ár/érték arányuk?

# Benchmark eredmények

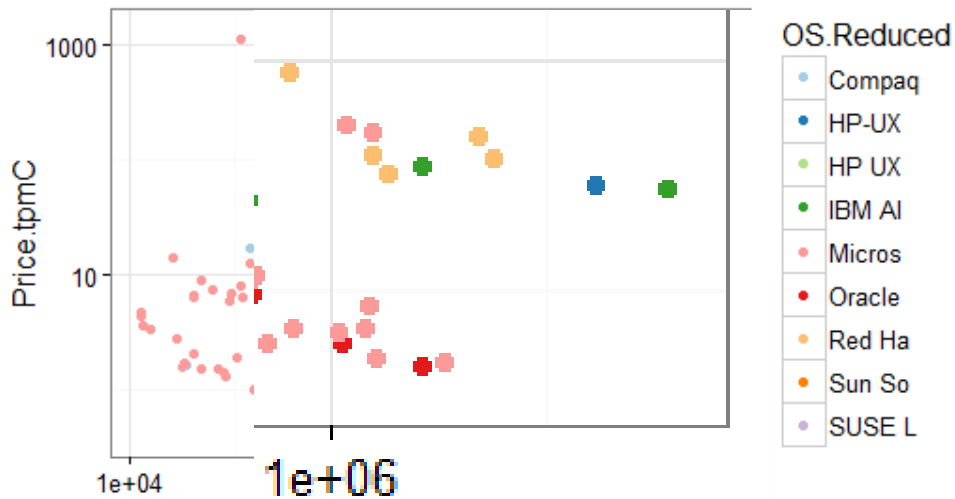


Logaritmus skála?

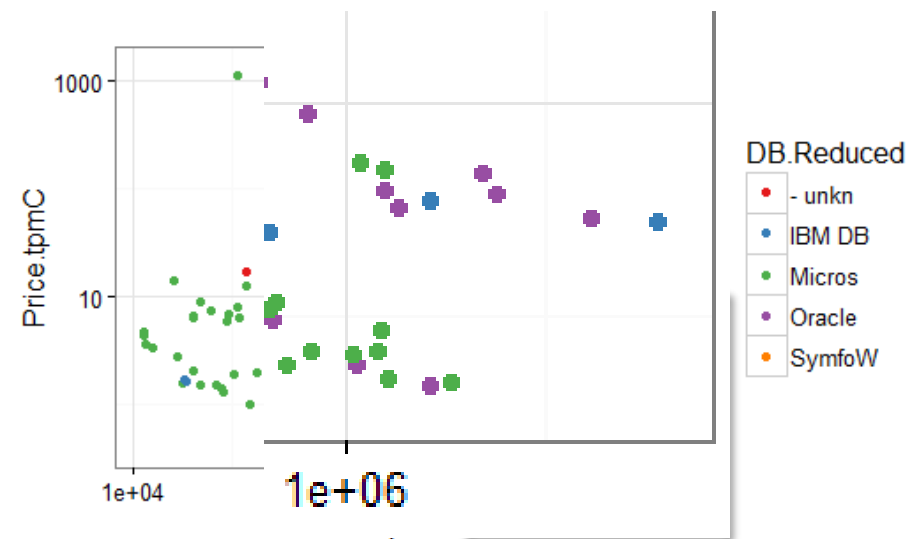
# Benchmark eredmények



Az élmezőny elég változatos



Nincs legjobb OS és DB konfiguráció sem



# Időbeli változások?

A legtöbb gyártónál  
10 év alatt  
egy nagyságrenddel csökkent a  
fajlagos ár

