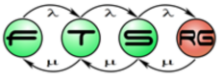
Intelligens rendszerfelügyelet (VIMIA370)

IT adatok vizuális elemzése

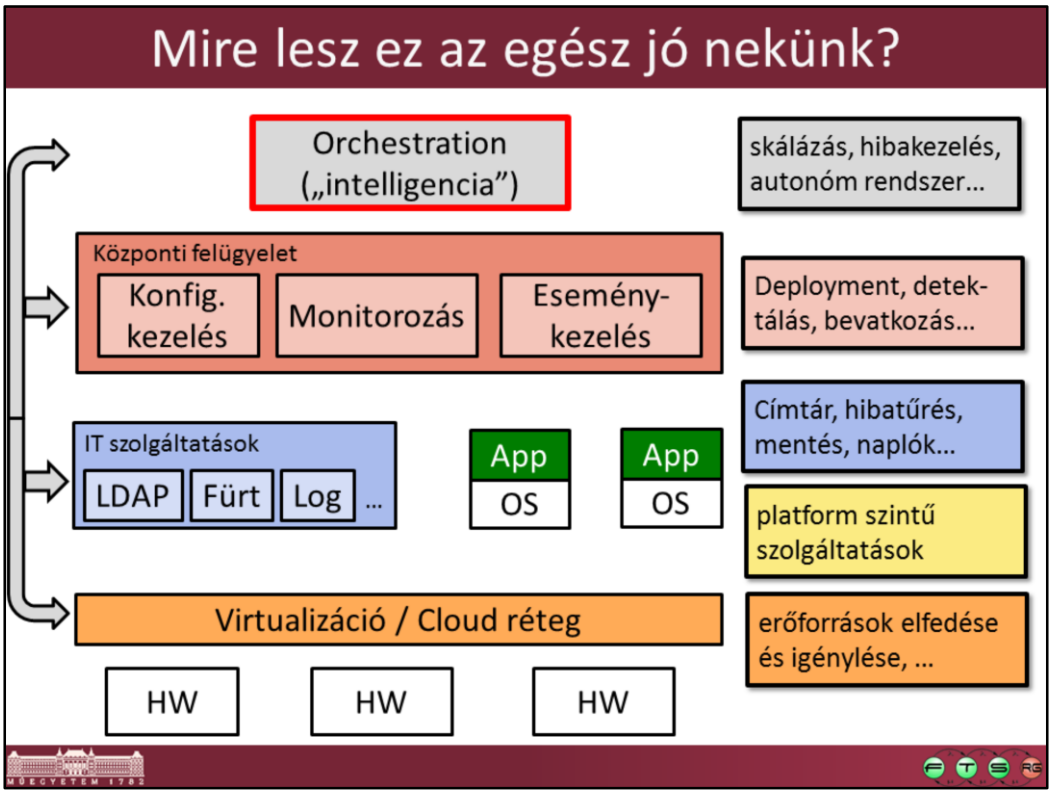
Salánki Ágnes



Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

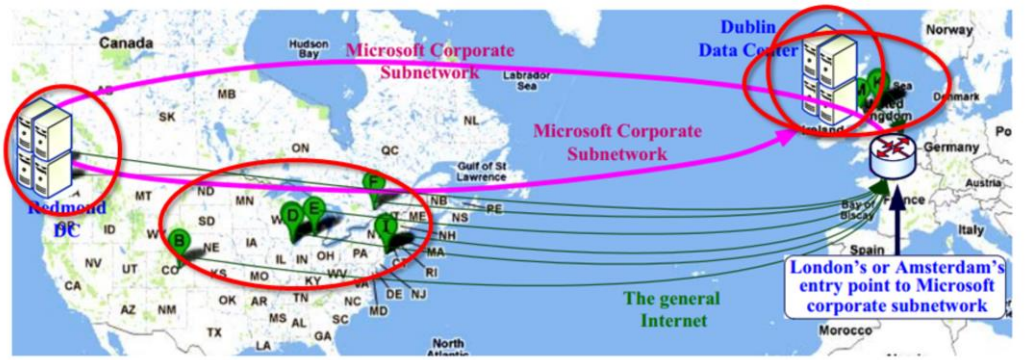
Utolsó módosítás: 2014.05.08.

Dr. Pataricza András (Rendszermodellezés) és Kocsis Imre (Big Data elemzési módszerek) idevágó fóliáit felhasználva.



Az alsó három szinttel foglalkoztak már a korábbi előadások. A vizuális analízis a monitorozott adatokat használja bemenetként, a rendszerbe általában a Központi felügyelet blokkba kapcsolódik vissza.

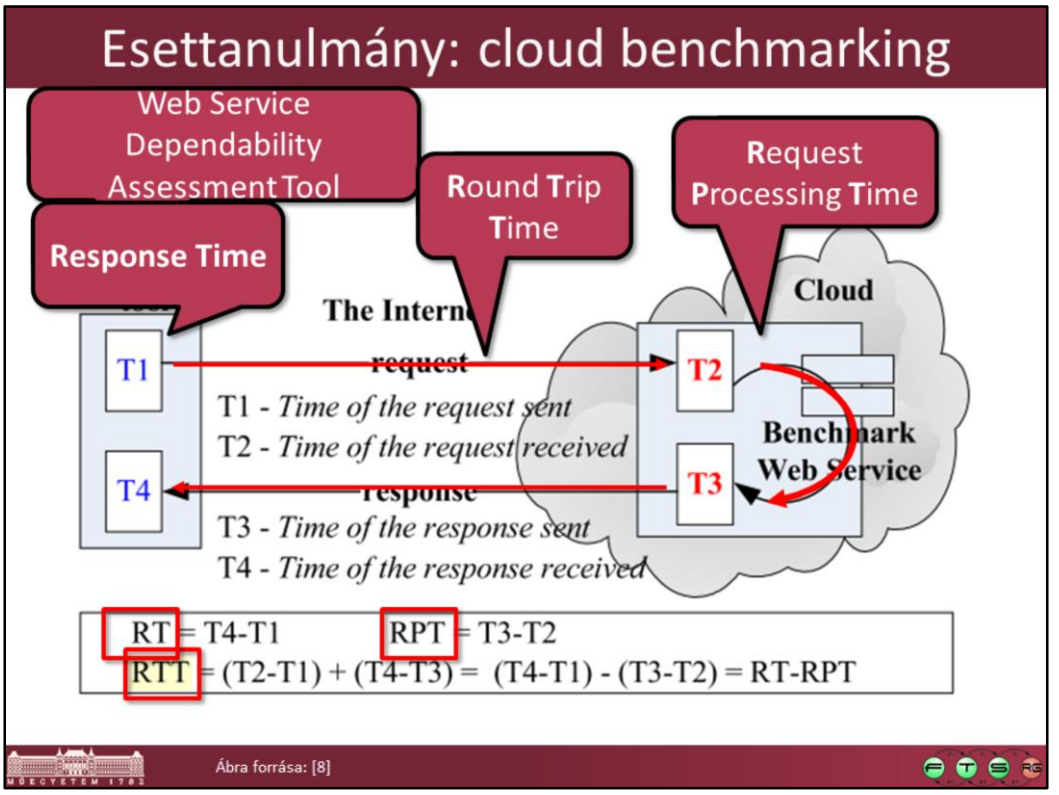
Esettanulmány: cloud benchmarking



Ábra forrása: [6], [7]

Kísérlet eredeti célja: hogyan befolyásolják a kliens/szerver implementációk/helyszínek egy MS Azure-ban futó webszolgáltatás teljesítményét? Felhasználói szemszög: a QoS metrikánk ezúttal a válaszidő.

Az eredeti esettanulmányát lásd a [6] cikkben.



WSsDAT – percenként kiküld egy ilyen kérést a megfelelő szerverhez
A vizsgált webszolgáltatás: kérésre egy 50x50x50-es mátrixot rendezzünk ellenkező irányba, majd ennek az első 100KByte-nyi adatával térjünk vissza

Elemzési megközelítés 1: leíró statisztika

Client's location		Client's IP address	Response Time (RT), ms				
			Min	Avg	Max	Std Dev	CV, %
Canada	Ottawa	216.151.172.x	1484	2007	3917	255	12.7
	Burnaby	64.151.226.x	1841	2329	287448	7648	328.4
UK	Newcastle	10.8.146.x	859	1577	4016	480	30.4
	Newcastle	10.8.151.x	890	1498	3438	423	28.2
	Durham	213.175.197.x	814	1922	47740	2132	110.9
USA	New York	69.72.183.x	1331	2018	7145	575	28.5
	Chicago	209.188.85.x	1133	2104	11729	875	41.6
	Peyton	64.64.0.x	1289	1961	4637	350	17.9
	Lansing	67.225.254.x	902	2051	4483	396	19.3
		67.225.254.x	1421	2024	3992	365	18.0
		67.227.193.x	1441	2068	19365	1004	48.5
		67.227.216.x	6123	6911	11520	366	5.3
		67.227.216.x	1420	2016	3770	344	17.1
	Secaucus	204.14.93.x	1262	1983	47056	1499	75.6
		208.87.24.x	1216	1991	46457	1271	63.8
		208.87.25.x	1212	2030	59475	2232	110.0
		64.20.37.x	1216	2275	132498	5135	225.7

CV -- coefficient of variation: szórás / átlag. Miért fontos: mert pl. a 2 ms-es szórás mást jelent 4 ms-os átlag válaszidőnél és egy 4 s-os átlagos válaszidőnél.

Elemzési megközelítés 2: felderítő adatanalízis

- *Exploratory Data Analysis (EDA)*
  - statisztikai tradíció,
  - mely koncepcionális
  - és számítási eszközökkel segíti
  - minták felismerését és ezen keresztül
  - hipotézisek felállítását és finomítását.

[1] és [2] alapján

- Statisztikai tradíció: ugyanolyan mint a matematikai statisztikából leadott Confirmatory Analysis (hipotézistesztelés, modellválasztás, paraméterillesztés – amit valószínűségszámításból tanultunk).
- Koncepcionális és számítási eszközök: nem használunk mély statisztikát, elemi számításokat és ábrákat annál inkább + az interpretáció sikeressége nagyban függ a szakértői tudástól.
- Minták felismerését és hipotézisek felállítását segíti: valahol tehát az adatbányászat és a statisztika között van.

Exploratory Data Analysis

- Cél: adatok „megértése”
  - „detektívmunka”
  - erősen ad-hoc
- Fő eszköz: adatok „bejárása” grafikus reprezentációkkal
- Hipotézistesztelléssel: iteratív folyamat

Erősen ad-hoc: az adatok szisztematikus átvizsgálása, de nincs jól bevált recept, leginkább mélységi keresést végzünk, ha találunk valami érdekeset, akkor aztán megfogalmazunk egy hipotézist („Statisztikailag szignifikáns-e a válaszidő várható értékének különbsége Java és .Net kliens esetén?” vagy „Normál eloszlást követ-e az RPT?”), aztán a megfelelő teszttel leellenőrizzük. Valamit mond a teszt, aztán görgetünk vissza és folytatjuk a felderítést.

Mi nem kell hozzá? Mély statisztikai ismeret: se centrális határeloszlás tétel, se nagy számok törvénye, az majd a 2. fázisban
Mi kell hozzá? Szakértői tudás nélkül nem megy – ezért beszélünk róla egyáltalán IRF-ből.

Miről lesz szó?

- Adatelemzési alapfogalmak
- Alapvető diagramtípusok
- Interaktív EDA eszközök – elvárt funkcionalitás

- Esettanulmány: cloud benchmarking

Miről nem lesz szó?

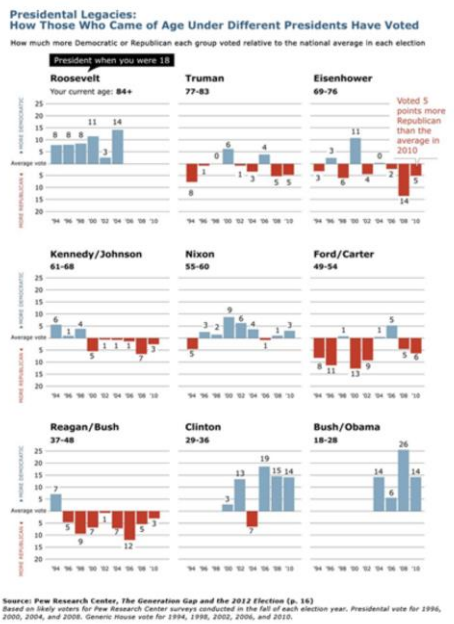
- Adatbányászat
- Hipotézistesztelés
- Kísérlettervezés
  - Pl. Rendszermodellezés tárgyunk
- Számítógépes grafika
- Információvizualizáció
  - Pl. blogok: Junk charts [8], Flowing data [9]

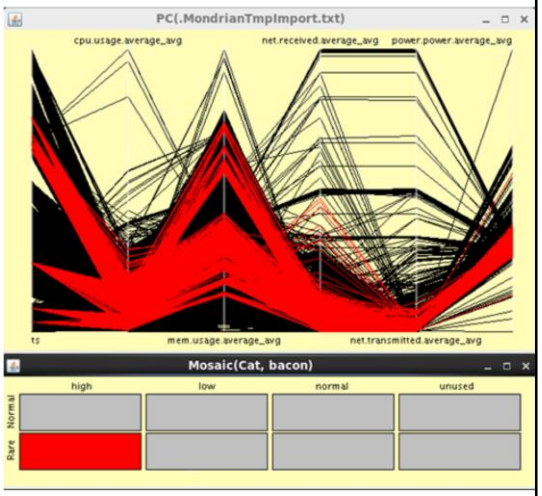
Prezentáció vs. felderítés

- Prezentáció
  - Statikus
  - Jó minőségű
  - Tömör
  - Sok annotáció: nagy közönség
- ~ bizonyítás
- ggplot2 csomag (R)
  Adobe Illustrator, Inkscape

- Felderítő ábrázolás
  - Interaktív
  - Gyors
  - Több különálló ábrát kapcsol össze
  - Néha tengelyfeliratok sem: az elemző az interpreter
- ~ matematikatörténet
- Pl. Mondrian, iplots (R)
  Many Eyes, Tableau

Prezentáció vs. felderítés





Adatmennyiség?

- Instrumentáció: pl. HF3
  - 8 metrika, egyperces mintavételezéssel egy hónapig
  - $8 \times 60 \times 24 \times 30 \approx 350\,000$ adatpont
- Hipotézismentes adatgyűjtés
  - 1 Windows 7 OS: perfmon
    - ≈ 100 körüli metrikaszám, egy másodperces mintavétel
    - Egy nap alatt $100 \times 3600 \times 24 \approx 860\,000$
  - Tanszéki VCL
    - ≈ 70 metrika, hosztonként 20 mp-es mintavételezéssel
    - Egy hónap alatt $70 \times 10 \times 180 \times 24 \times 30 \approx 90M$

Mi a baj ekkora mennyiségnél? A táblázatokat

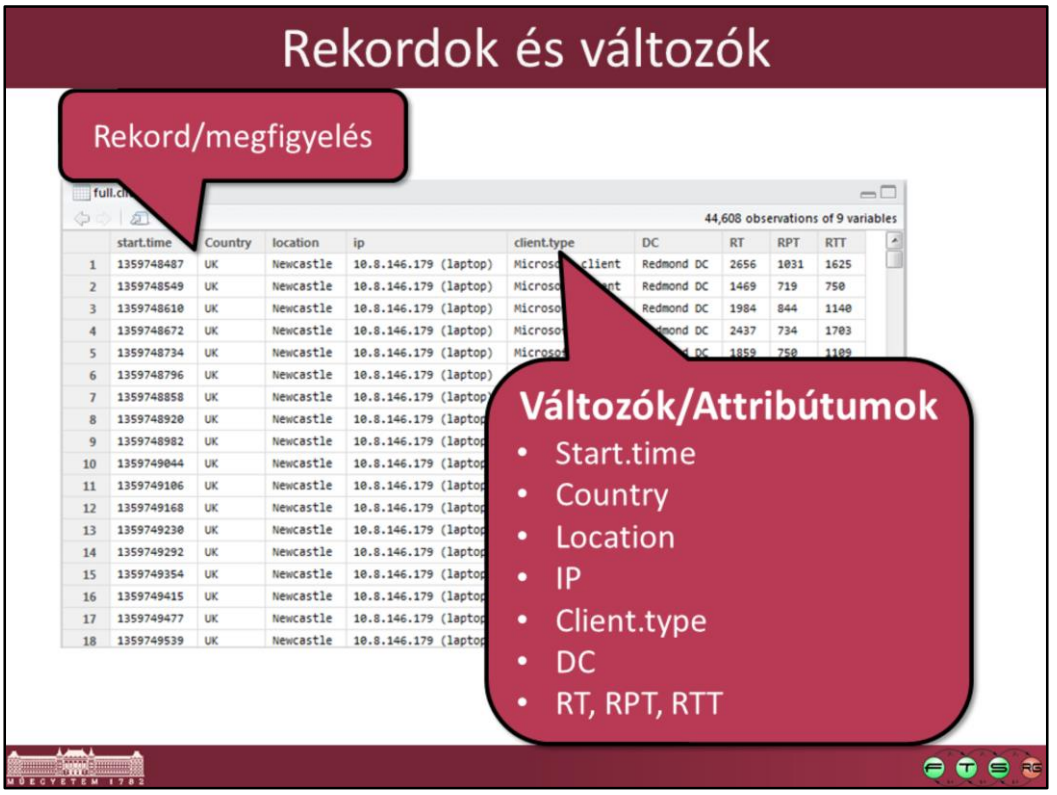
1. vagy nagyon sokáig kell böngészni, vagy
2. az aggregálás a világon mindent kisimít ekkora távon, a trend éppen látszik, a tranziensek pont nem és ez baj, mert minket általában az érdekel, hol és mikor történt valami hiba.

Miről lesz szó?

- **Adatelemzési alapfogalmak**
- Alapvető diagramtípusok
- Interaktív EDA eszközök – elvárt funkcionalitás

- Esettanulmány: cloud benchmarking

VÁLTOZÓK

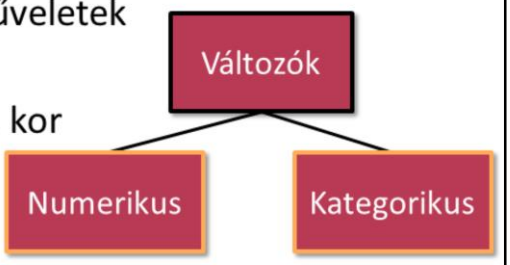


Rekord: általában egy egyértelműen azonosítható mérési regisztrátum
Változó: minden, amit az adott pillanatban tudunk a mérésről és a mért adatról

Mindig a kísérletből derül ki, hogy egy adott változó az kontextus vagy viselkedési. Például:
a) Hőmérsékletet mérünk MO-n: a hőmérséklet viselkedési
b) Adott hőmérsékletre felmelegítjük a szerver szobát és számoljuk, hány gép adja meg magát: a hőmérséklet kontextus

Numerikus és kategorikus változók

- Numerikus (numerical)
  - az alapvető aritmetikai műveletek értelmesek
  - Pl. napi átlaghőmérséklet, kor

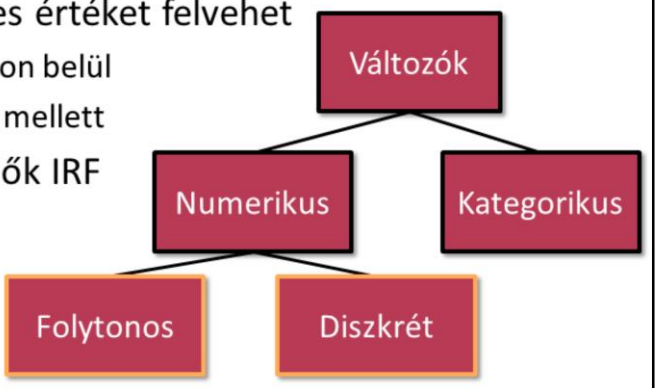


- Kategorikus (categorical)
  - Csak a megkülönböztetés miatt
  - Pl. telefonszám, nem

Nem a reprezentáció a lényeg, hanem az interpretáció!
Ha a nem az {1, 2} halmazból veszi fel az értékét (és nem a {nő, férfi}-ból), akkor is kategorikus lesz, mert minden művelet értelmetlen rá.

Numerikus változók

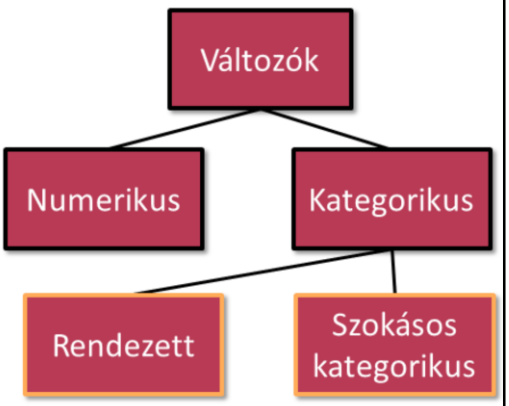
- Folytonos
  - Mért – tetszőleges értéket felvehet
    - adott tartományon belül
    - adott pontosság mellett
  - Pl. a teremben ülők IRF jegyének átlaga
- Diszkrét
  - Számolt – véges sok értéket vehet fel adott tartományban
  - Pl. IRF előadáson ülők száma



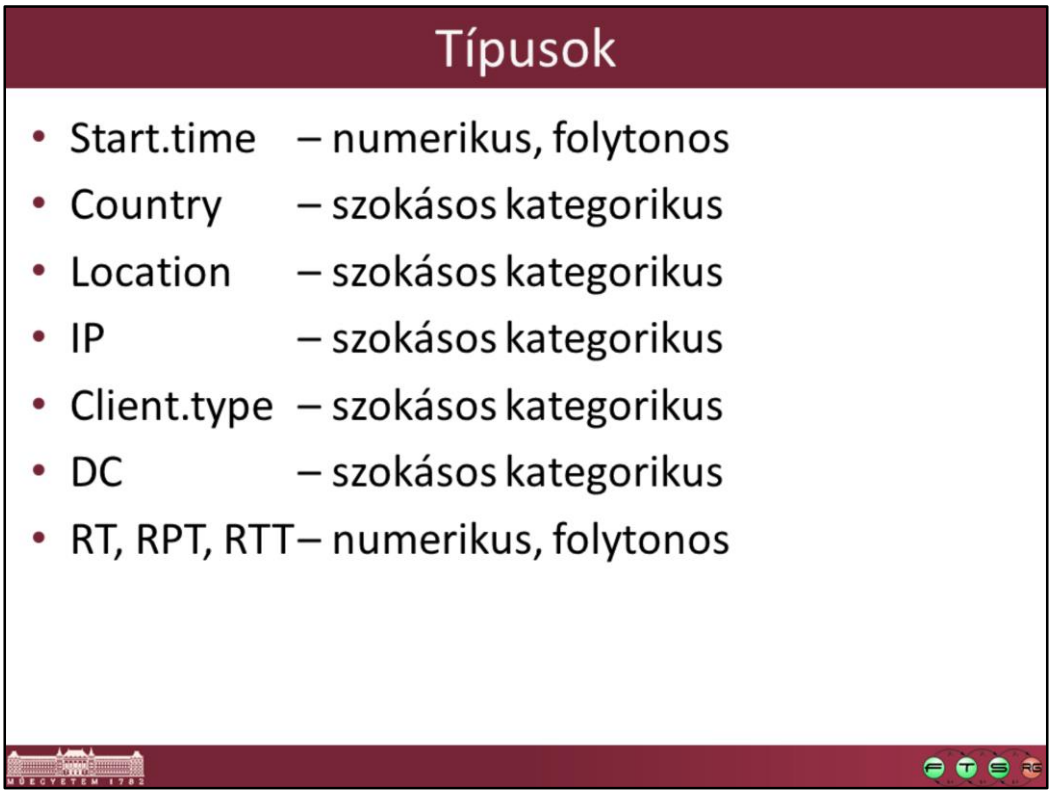
Kategorikus változók

- Szokásos kategorikus (regular)
- Rendezett
  - szintek között hierarchia

9. Ajánlanád-e a tárgyat másoknak?
- Mindenkit rábeszélnék
- Nyugodtan ajánlanám
- Esetleg ajánlanám
- Inkább lebeszélném róla
- Feltétlenül lebeszélném
- Nem kívánok válaszolni



Miért fontos, hogy tudjuk a típust? Mert tudnunk kell, milyen ábrázolásmód passzol hozzá.



Ha tudjuk a megfelelő ábrázolásmódot, onnan akár vissza is következtethetünk a típusra: pl. az RT vs. Start.time inkább scatterplotra menjen és nem mozaikplotra → inkább legyen numerikus, mint kategorikus.
A folytonos-diszkrét általában értelmezés kérdése.

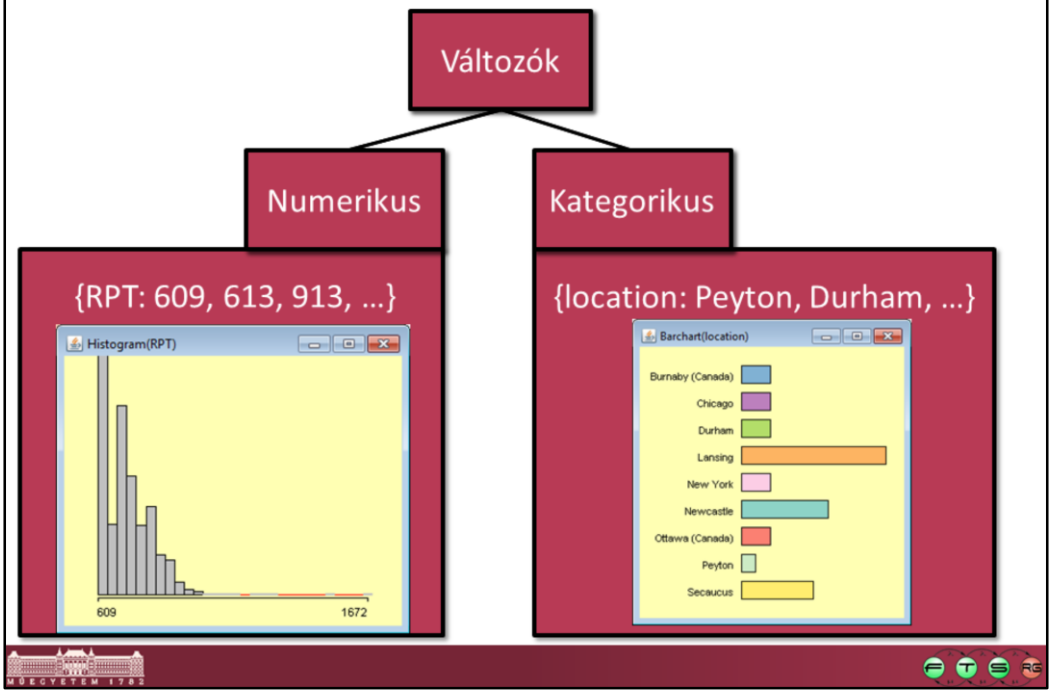
Miről lesz szó?

- Adatelemzési alapfogalmak
- **Alapvető diagramtípusok**
- Interaktív EDA eszközök – elvárt funkcionalitás

- Esettanulmány: cloud benchmarking

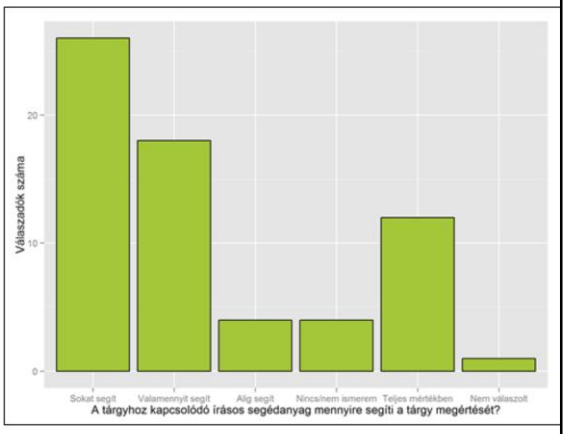
ALAPVETŐ DIAGRAMTÍPUSOK

1 változó


Változók
Numerikus
Kategorikus
{RPT: 609, 613, 913, …}
Histogram(RPT)
609
1672
{location: Peyton, Durham, …}
Barchart(location)
Burnaby (Canada)
Chicago
Durham
Lansing
New York
Newcastle
Ottawa (Canada)
Peyton
Secaucus

Oszlopdiagram (bar chart)

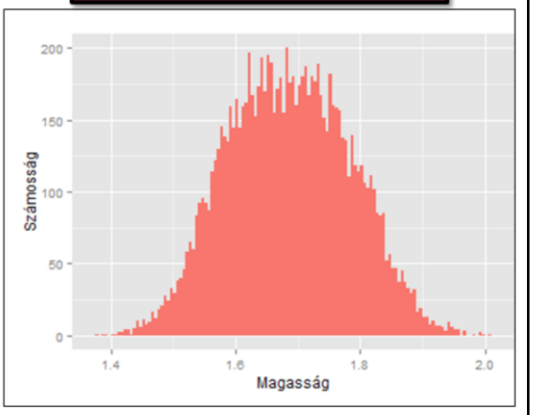
- Megjelenített dimenziók száma: 1
- Ábrázolt összefüggés:
  - Kategorikus változó egyes értékeinek abszolút gyakorisága
- Adategység:
  - Oszlop – magassága: adott érték gyakorisága
- Tervezői döntés:
  - Értékkészlet darabolása?



Hisztogram

- Megjelenített dimenziók száma: 1
- Ábrázolt összefüggés:
  - Folytonos változó egyes értékeinek abszolút gyakorisága
- Adategység:
  - Oszlop – magassága: adott érték gyakorisága
- Tervezői döntés:
  - Oszlopszélesség/kezdőpont?

Fontos percentilisek?



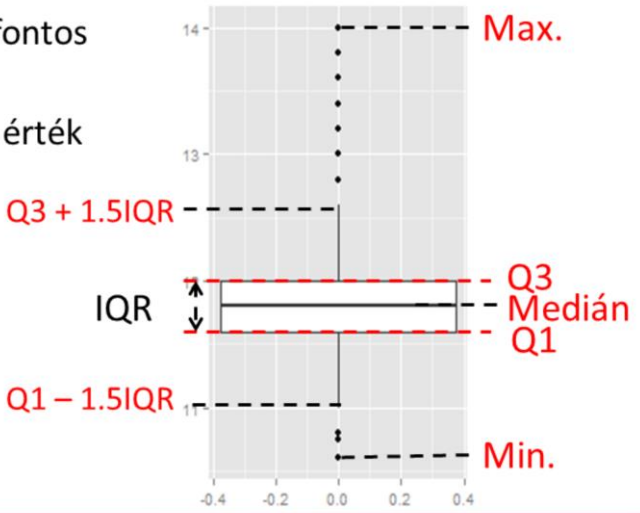
Kezdőpont függés:
S := {1.0, 2.0, 3.0}, az oszlopszélesség legyen 1.5.
Két különböző hisztogram is ér: ([0, 1.5), [1.5, 3.0]) és ([1.0, 2.5), [2.5, 4.0]) → ezek mindegyike érvényes, csak máshonnan indul!

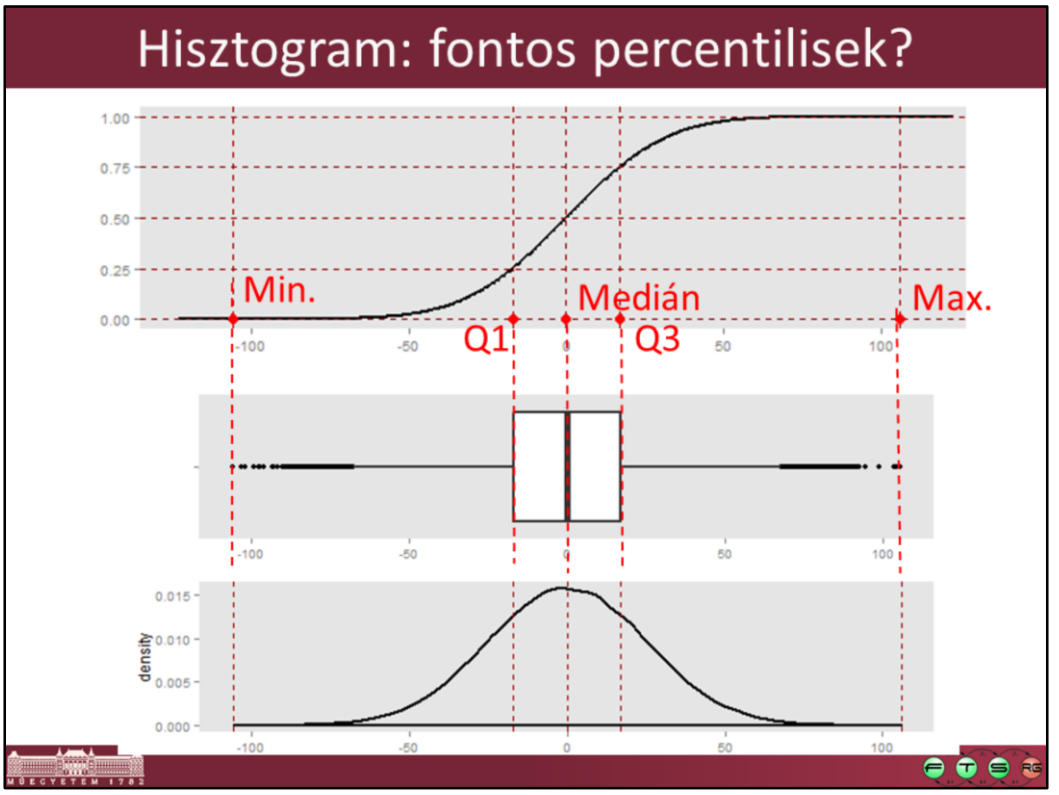
Egy kis leíró statisztika…

- Átlag, medián, módusz
- Percentilis
  - Az n-edik percentilisnél az adatok $n\%$-a kisebb
- Kvartilis
  - Q1, Q3: 25. és 75. percentilis
  - Q2: medián
- Inter-quartile range (IQR)
  - Q3 – Q1

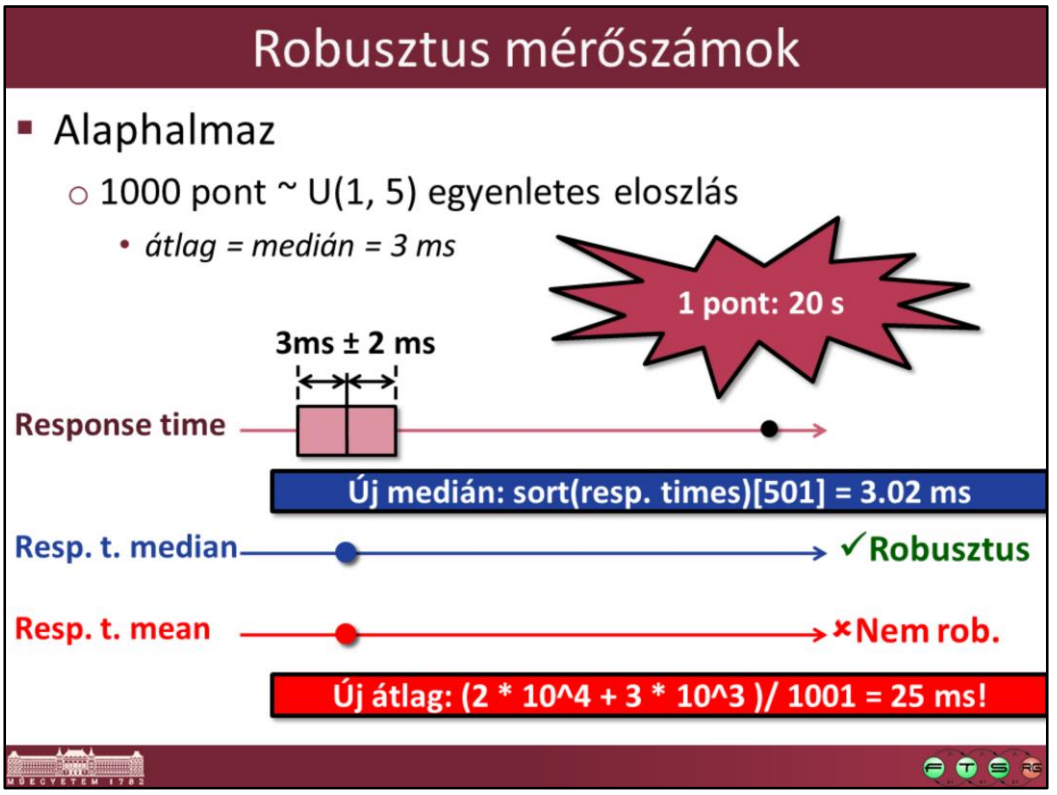
Doboz diagram (boxplot)

- Megjelenített dimenziók száma: 1
- Ábrázolt összefüggés:
  - Folytonos változó fontos percentilisei
  - Általában 5 fontos érték
- Adategység:
  - Doboz
- Tervezői döntés:
  - Outlierek?



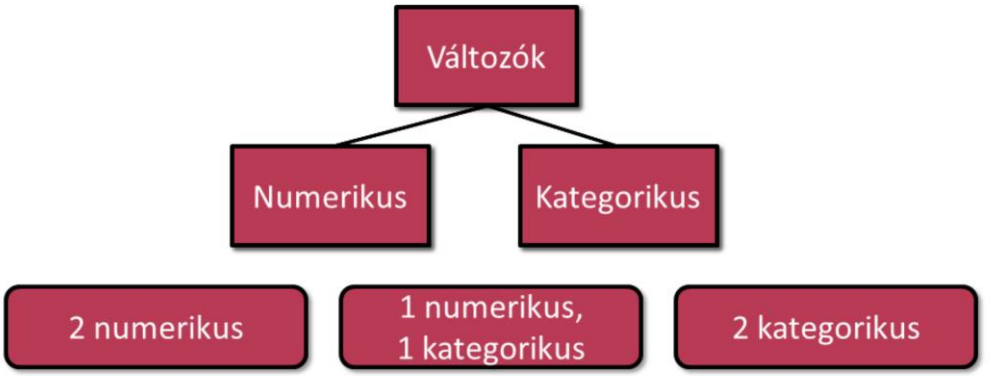


Felső: eloszlásfüggvény, alsó: sűrűségfüggvény

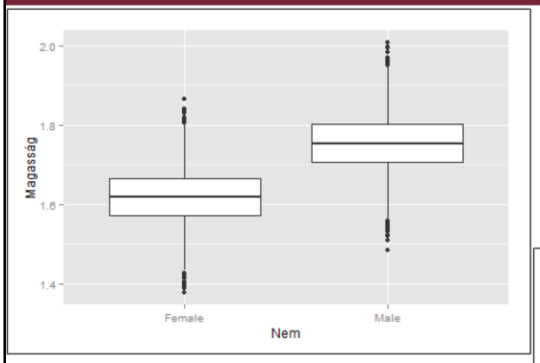


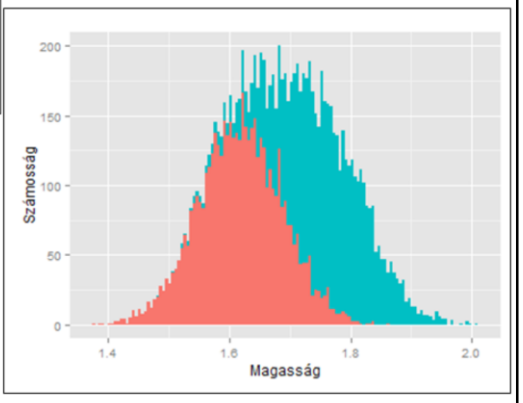
Ha az outlierekre vagyunk kíváncsiak, akkor persze számoljunk átlagot, de ha az adatsort jellemezni szeretnénk, akkor jobb robusztus statisztikákat használni.

2 változó kapcsolata

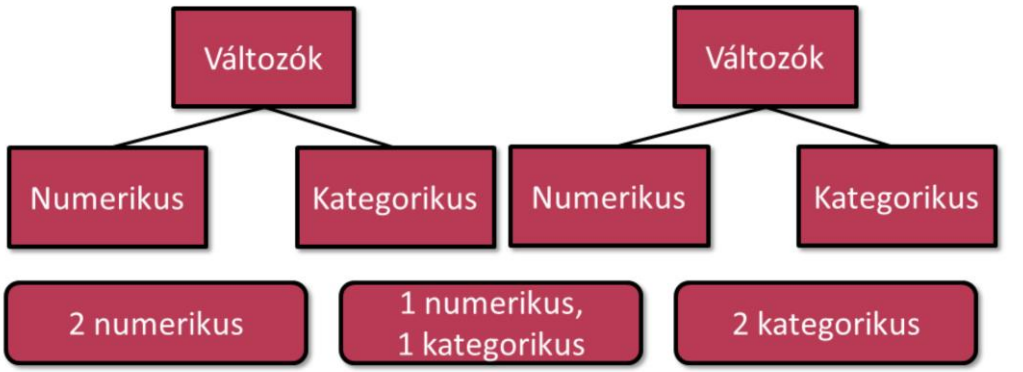


Numerikus kategóriánként





2 változó kapcsolata



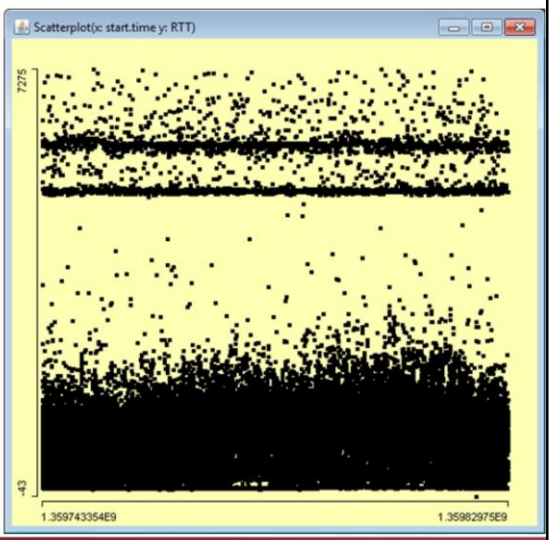
2 numerikus

1 numerikus,
1 kategorikus

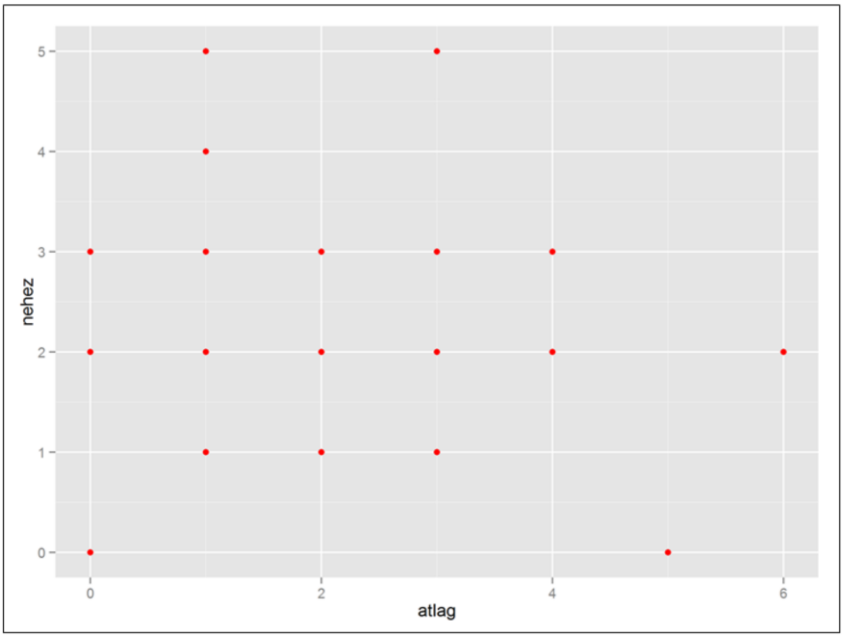
2 kategorikus

Pont – pont diagram (scatterplot)

- Megjelenített dimenziók: 2
- Ábrázolt összefüggés:
  - folytonos változók együttes eloszlása
- Adategység:
  - pont – $X = x_i, Y = Y_i$ előfordulás
- Korlát:
  - ha az egyik változó értéke hiányzik, nem tudjuk felrajzolni
- Tervezői döntés:
  - Overplotting?

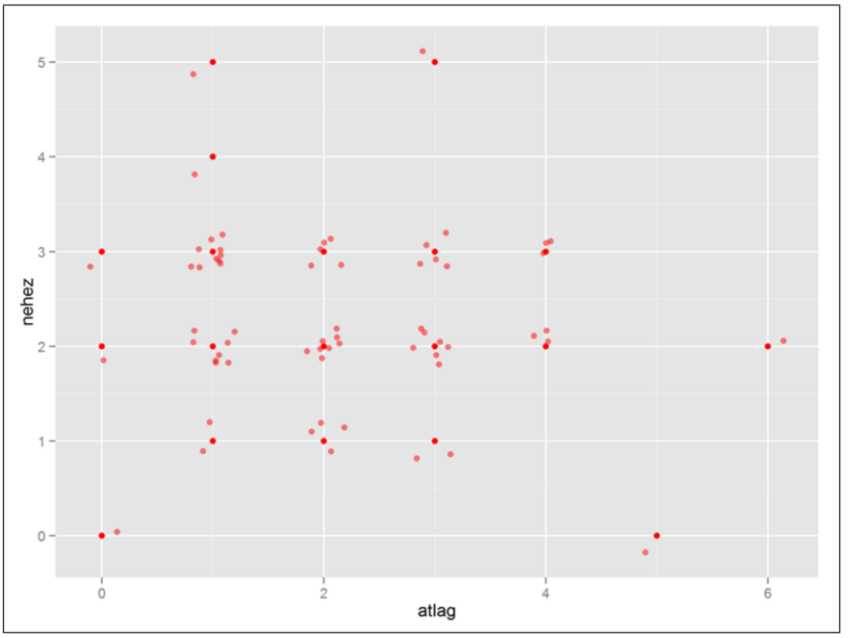


Overplotting megoldások 1: jitter

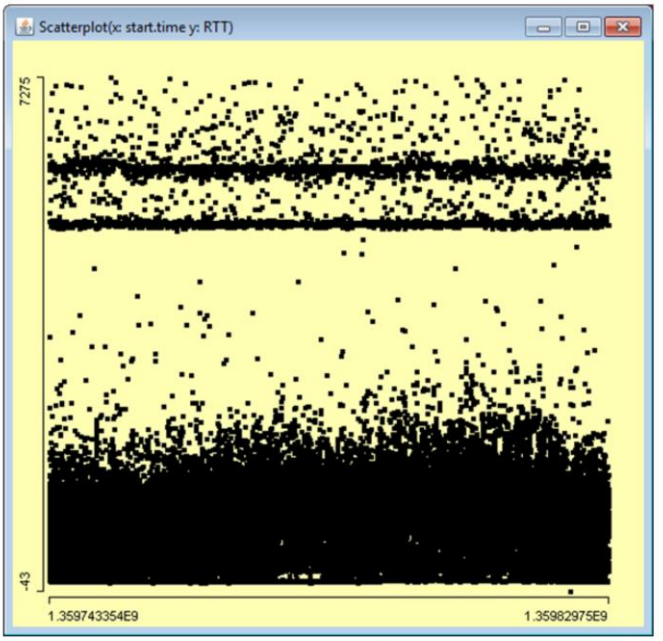


Két diszkrét numerikusnál – ritka, szét kell szedni

Overplotting megoldások 1: jitter

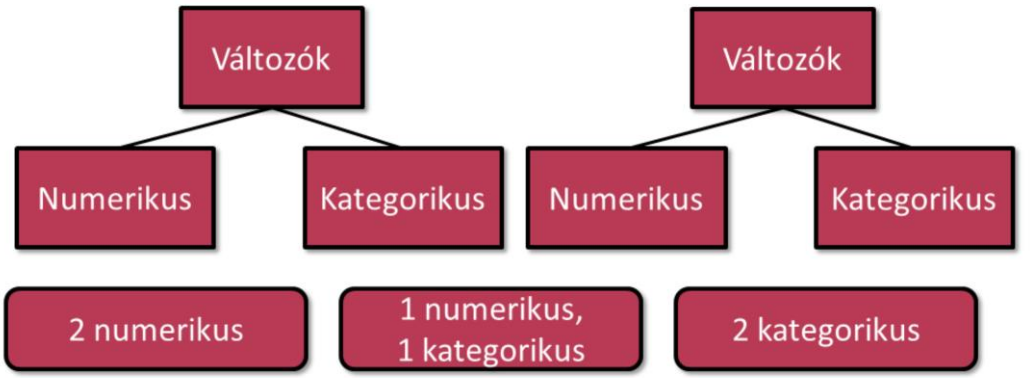


Overplotting megoldások 2: átlátszóság



Két folytonosnál: túl sűrű, szedjük szét!

2 változó kapcsolata

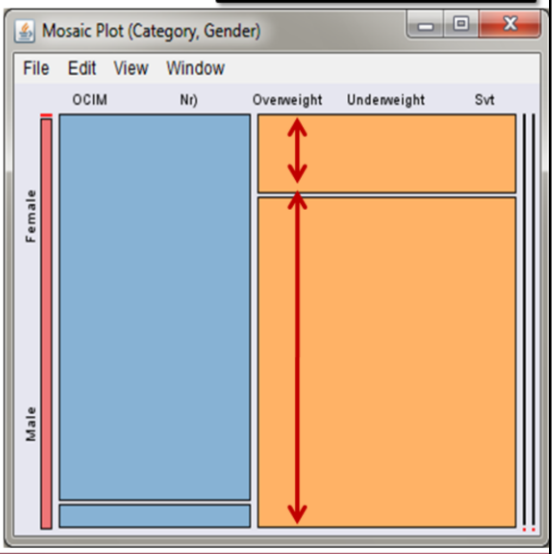


2 numerikus	1 numerikus, 1 kategorikus	2 kategorikus

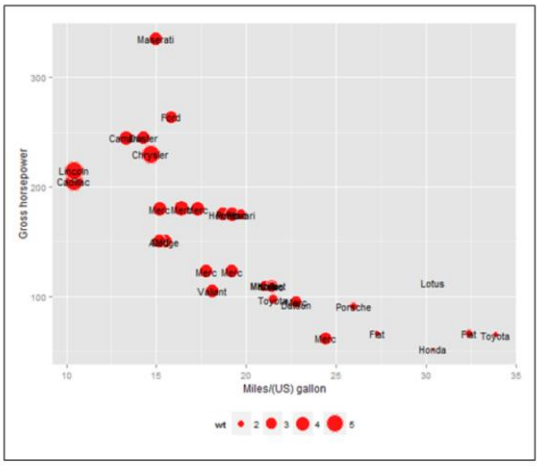
Mozaik diagram (mosaic plot)

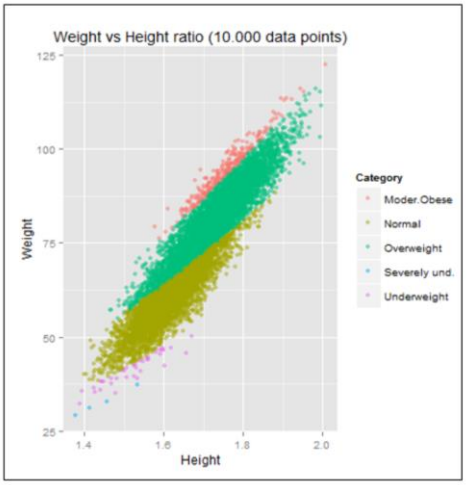
- Megjelenített dimenziók száma: 2
- Ábrázolt összefüggés:
  - 2 diszkrét változó e. e.
- Adategység:
  - Téglalap – *területe* arányos az $(X = x_i, Y = y_i)$ értékpárok gyakoriságával
- Korlát:
  - Sorfolytonos olvasás?

A túlsúlyosak nagy része férfi!

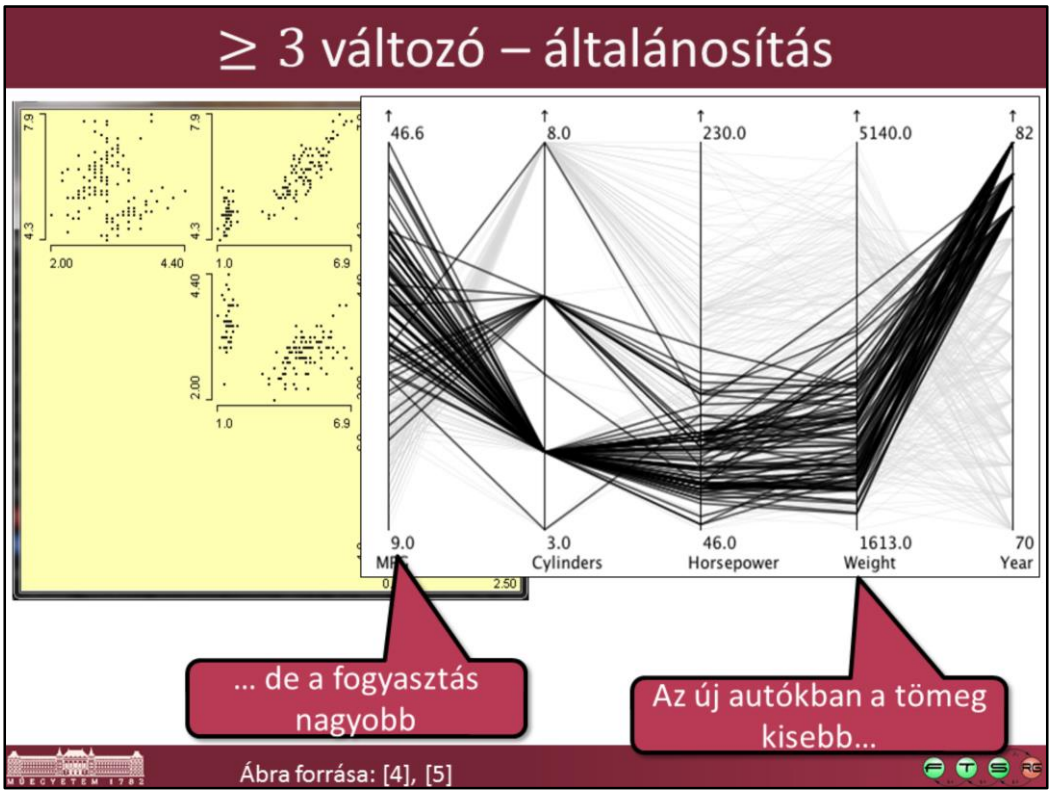


≥ 3 változó – mesterséges dimenziók





Alak (telített pötty vagy sima, csillag, négyzet stb.) + szín + méret – 3 extra dimenzió, szerencsés esetben akár 5 dimenziót is ábrázolhatunk egyszerre



Scatterplot matrix: rettenetesen rosszul skálázódik ☹, gyorsan kiszűrhetők az egymással asszociáló változópárok ☺

Párhuzamos koordináták: a tengelyek különböző nagyságrendje torzíthat az összefüggésen ☹, aki érti a geometriáját, az könnyen/gyorsan következtet (Pl. egy pontban metszik a tengelyek egymást → 1-es vagy -1-es korreláció látszik) ☺ → Optimális változósorrendet találni? ☹

Miről lesz szó?

- Adatelemzési alapfogalmak
- Alapvető diagramtípusok
- **Interaktív EDA eszközök – elvárt funkcionalitás**

- Esettanulmány: cloud benchmarking

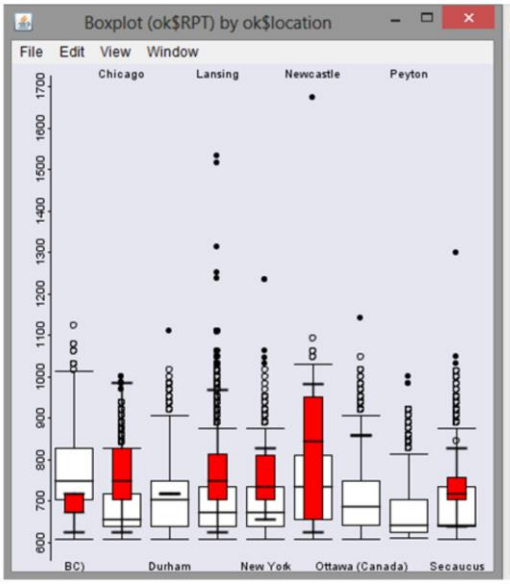
FUNKCIONALITÁS

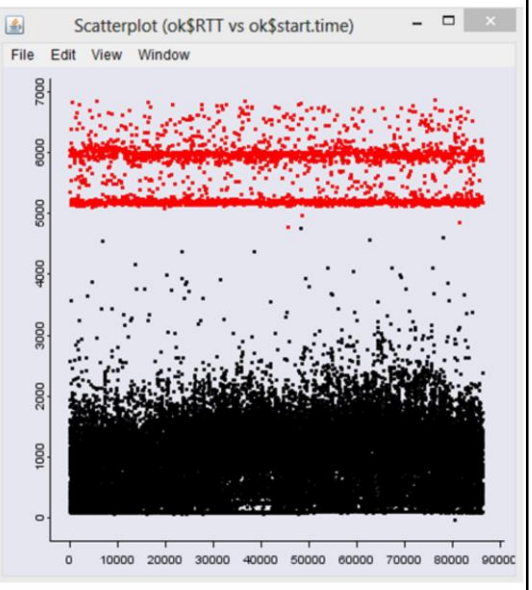
Prezentáció vs. felderítés

- Prezentáció
  - Statikus
  - Jó minőségű
  - Tömör
  - Sok annotáció: nagy közönség
- ~ bizonyítás a matematikában

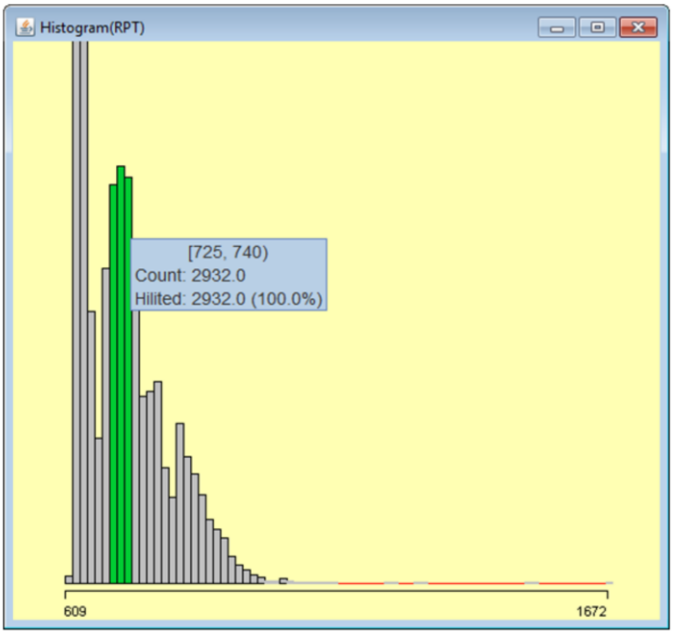
- Felderítő ábrázolás
  - Interaktív
  - Gyors
  - Több különálló ábrát kapcsol össze
  - Néha még tengelyfeliratok sem: egyedül az elemző kell hogy megértse
- ~ matematikatörténet

Adatkötés

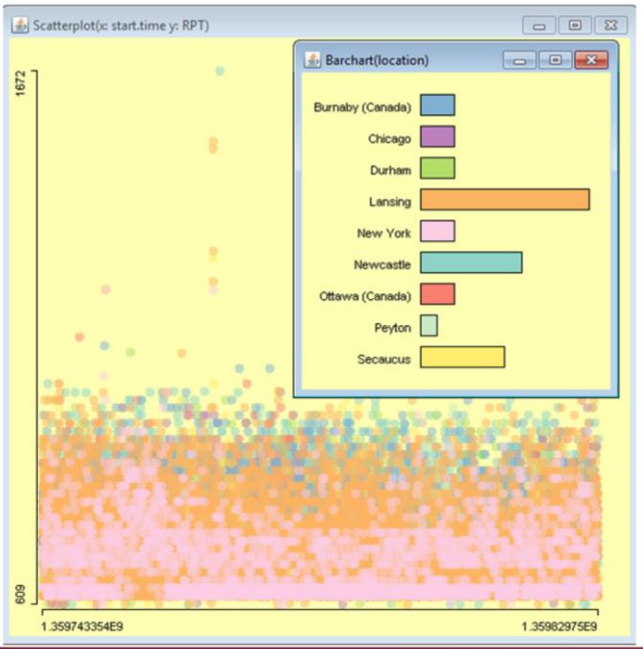

Boxplot (ok$RPT) by ok$location
File Edit View Window
Chicago
Lansing
Newcastle
Peyton
BC)
Durham
New York
Ottawa (Canada)
Secaucus


Scatterplot (ok$RTT vs ok$start.time)
File Edit View Window

Lekérdezések



Színezés/átlátszóság



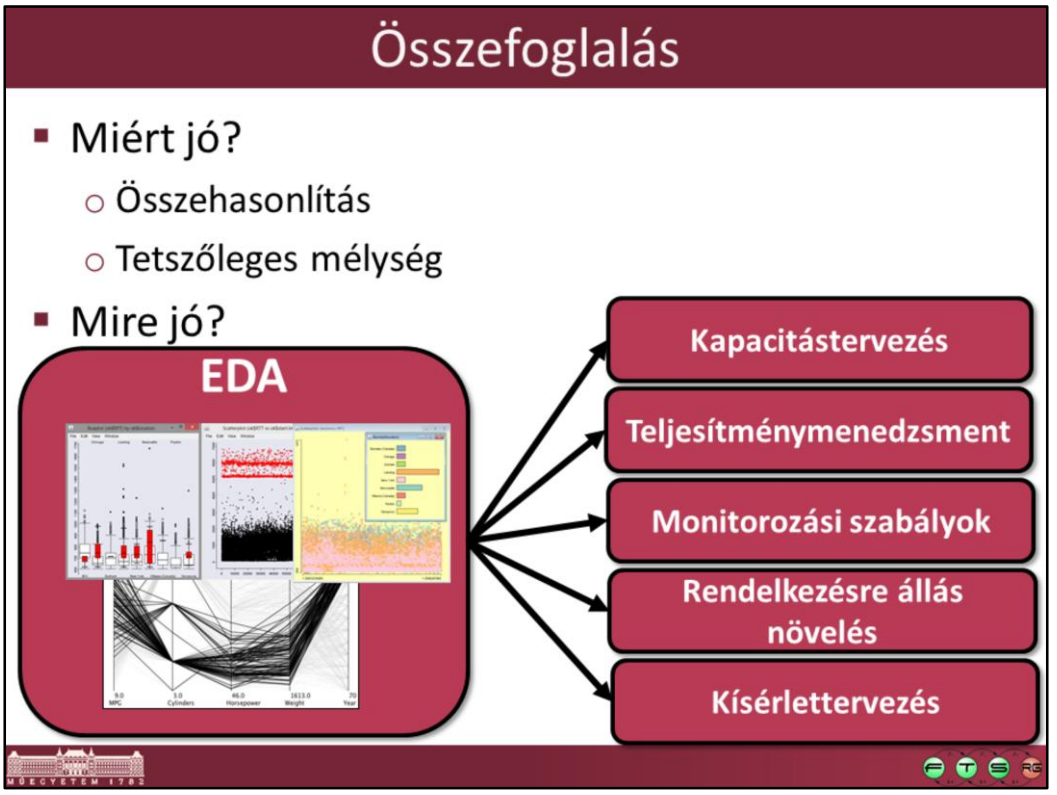
Miről lesz szó?

- Adatelemzési alapfogalmak
- Alapvető diagramtípusok
- Interaktív EDA eszközök – elvárt funkcionalitás

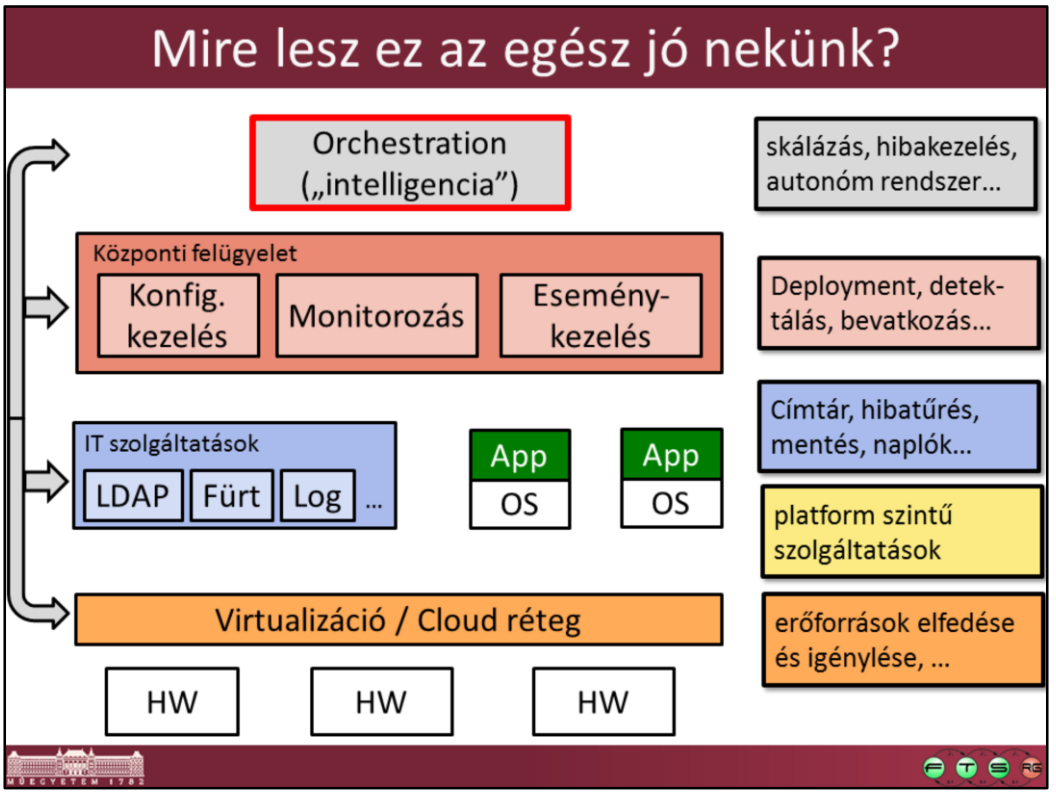
- **Esettanulmány: cloud benchmarking**

Cloud benchmarking

- Alapvető RT-RTT összefüggések
- Kísérlettervezési hiányosságok
- Konfiguráció hibák
- Térbeli/időbeli/kliensbeli függőségek



Szűk keresztmetszetek keresése, legyen az szolgáltatásbiztonság vagy teljesítménymenedzsment : időbeli/térbeli függés fontos lehet.
Kapacitástervezés, monitorozás: RTT befolyásol inkább, mint az RPT, akkor monitorozzuk/javítsuk ezt a részét a dolgoknak, felesleges jobb processzort venni, inkább a hálózatunk legyen normálisan bekonfigurálva



Nagyrészt a központi felügyelet dobozkába vezetjük vissza az eredményeket, legfeljebb az IT szolgáltatásokhoz.
Az adatelemző nem tudja, mi van az adatok mögött, a rendszermérnök kompetenciája viszont jól használható.

Hivatkozások

[1] Behrens, J.T.: Principles and procedures of exploratory data analysis. Psychological Methods 2, 131–160 (1997)

[2] Tukey, J.: We need both exploratory and confirmatory. The American Statistician 34, 23–25 (1980)

[3] Yau, Nathan. *Visualize this: the FlowingData guide to design, visualization, and statistics*. John Wiley & Sons, 2011.

[4] Inselberg, A.: Parallel Coordinates: Visual Multidimensional Geometry and its Applications. Springer Science+Business Media, New York (2009)

[5] Theus, M., Urbanek, S.: Interactive graphics for data analysis: principles and examples. CRC Press (2011)

[6] Gorbenko, A., Kharchenko, V., Mamutov, S., Tarasyuk, O., Romanovsky, A.: Exploring Uncertainty of Delays as a Factor in End-to-End Cloud Response Time. In: 2012 Ninth European Dependable Computing Conference, pp. 185–190. IEEE (2012)

[7] Pataricza, András, et al.: Empirical Assessment of Resilience. Software Engineering for Resilient Systems. 1-16. (2013)

[8] Funk, Kaiser: Junk Charts blog, URL: http://junkcharts.typepad.com/

[9] Yau, Nathan: FlowingData blog, URL: http://flowingdata.com/