1 Adatelemzési és statisztikai alapok

Milyen típusú változótípusokat különböztetünk meg?

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

Hol van ezeknek szerepe?

Adatelemzésben

Milyen típusú változók fordulhatnak elő egy olyan adatsorban, amely egy magyarországi lakosok vásárlási szokásait felmérő, alábbi pontokat tartalmazó kérdőívből született:

- nyilatkozó neme, életkora, lakóhelye, legmagasabb iskolai végzettsége;
- vásárlási gyakorisága, hetente hányszor vásárol X terméket;
- a standard vagy a prémium alterméket szereti?

Mondjon példát mindhárom típusra!

Strukturált

- Rögzített formátum
- Általában $n \times p$-s táblázat
- „Tidy”
  - Sor: pontosan egy megfigyelés
  - Oszlop: pontosan egy változó

Nemstrukturált

- Nincs előre rögzített tárolási/értelmezési modell
- Csak metaadat
- Pl. e-mail, audio anyagok

Szemistrukturált adat:

- $\subset$ Strukturált
- Nem ábrázolható hatékonyan adattáblában
  - Azonos típusú objektumok más attribútumokkal is
  - Az attribútumok sorrendje nem számít
- Pl. XML, JSON

Mi a felderítő és mi a megerősítő statisztikai elemzés? Mondjon példát mindkét megközelítésre!

Felderítő analízis

- **Cél:** hipotézisek megfogalmazása
- Ismerkedés az adatokkal/doménnel
- Erősen ad-hoc

- **Fő eszköz**: leíró statisztika + adatbányászat, sok vizualizáció
- **Példa**: Sejtés: az x változó normális eloszlású

Megerősítő analízis

- **Cél:** hipotézisek tesztelése
- Előre megsejtett összefüggések ellenőrzése
- **Fő eszköz**: statisztikai tesztek + következtető módszerek
- **Példa**: Az x változó hihetően $N(12,4)$ eloszlást követ

2 Vizuális analízis

Mik a fő különbségek az EDA és a CDA között a statisztikai elemzés során?

EDA: *Exploratory Data Analysis*: statisztikai tradíció,

- mely koncepcionális
- és számítási eszközökkel segíti
- minták felismerését és ezen keresztül
- hipotézisek felállítását és finomítását.

Cél: adatok „megértése”

- „detektívmunka”
- erősen ad-hoc

Fő eszköz: adatok „bejárása” grafikus reprezentációkkal

Hipotézisek: iteratív folyamat

Flexibilitás és pragmatizmus

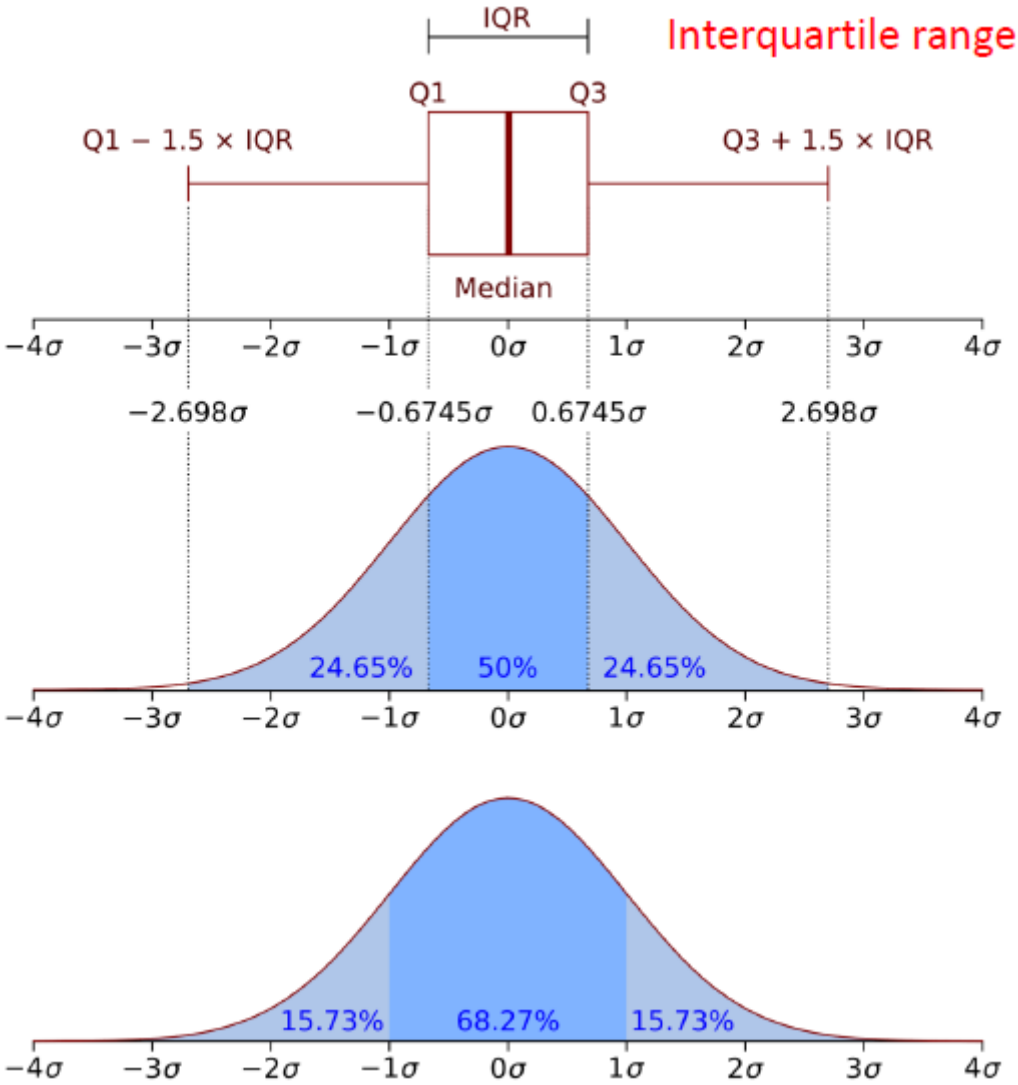
CDA: *Confirmatory Data Analysis*

- Hipotézistesztelés
- modellválasztás
- paraméterillesztés, …

Mi a dobozdiagram (*boxplot*)? Minek a szemléltetésére használjuk? Ábrán szemléltesse, hogy a dobozdiagram hogyan reprezentálja egy megfigyelés-halmaz alapvető leíró statisztikáit!

- Megjelenített dim.k: 1
- 5 értékkel jellemzésként
- Ábrázolt összefügg:
    - folytonos változó fontos percentilisei
- **Adategység**:
    - Doboz – szélei jelzik az alsó és felső kvartiliseket,
    - Középen a medián.
    - A minimum és a maximum általában még pontosan jelezve,
    - Outlierek már csak pöttyökkel

Mi a dobozdiagram mediánjának, „bajszainak“ és „sarokpontjainak“ (*whiskers and hinges*) kapcsolata a normális eloszlás paramétereivel? Diszkutálja, hogy alkalmas-e a dobozdiagram más eloszlások szemléltetésére is, és ha igen, milyen korlátokkal!



Mi a SPLOM? Miért használjuk a vizuális EDA során? Mik alkalmazásának legfőbb korlátai?

- Megjelenített dim.k: *n*
- Ábrázolt összefügg:
  - A változópárok együttes eloszlása (scatterplotok mátrixban)
- Adategység:
  - Scatterplot – minden diagram a neki megfelelő változók együttes eloszlását mutatja be

Mozaik-diagram: szöveges definíció, szemléltetés ábrával, jellemző alkalmazási esetei és alkalmazásának korlátai.

- Megjelenített dim.k: 2
- Ábrázolt összefügg:
  - két diszkrét változó együttes eloszlása
- Adategység:
  - Téglalap – a téglalap *területe* arányos az ($X = x_i$, $Y = y_i$) értékpárok gyakoriságával
- Korlát:
  - Sorfolytonos olvasása nehézkes

Párhuzamos koordináták: szöveges definíció, szemléltetés ábrával, jellemző alkalmazási esetei és alkalmazásának korlátai

- Megjelenített dim.k: *n*
- Ábrázolt összefügg:
  - Rekordok/attribútumok hasonlósága
- Adategység:
  - Törött vonal – az egyes attribútumtengelyeken felvett értékek rendezett sorozata
- Korlátok:
  - Tengelyek (attribútumok) más mértékegysége/nagyságrendje stb. torzíthat

Mit jelent, ha egy párhozamos koordináta diagram két szomszédos tengelye között futó szakaszokra illesztett egyenesek egy pontban metszik egymást?

Sok/két rekord egy attribútumának azonos az értéke.

Mit jelent, ha egy párhozamos koordináta diagram két szomszédos tengelye között futó szakaszokra illesztett egyenesek két pontban metszik egymást? Milyen hipotézist állítana fel ebből a megfigyelésből?

Van egy elválasztó változó érték.

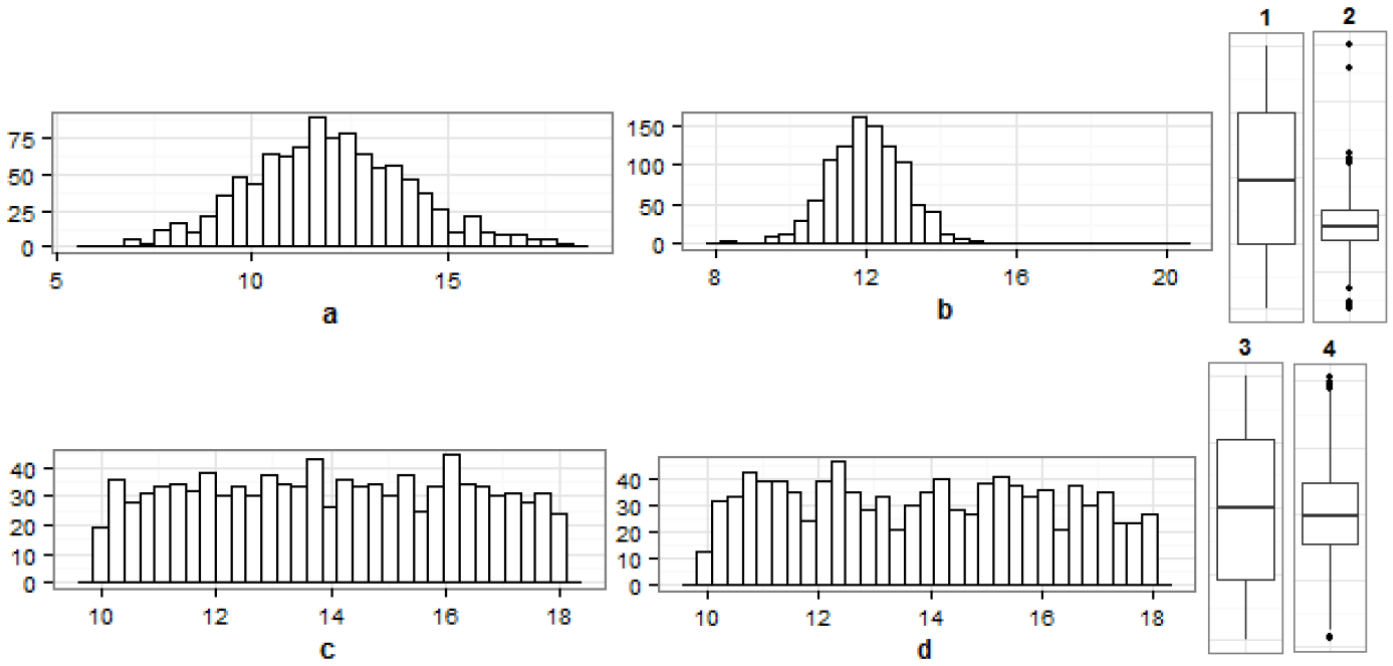
Kiszámítjuk egy folytonos változó értékeit tartalmazó adatsor mediánját, móduszát és átlagát. Válassza ki az igaz állításokat!

- A medián **biztosan** kisebb az átlagnál. **Hamis**
- Az átlag **legfeljebb** kétszerese lehet a mediánnak. **Hamis**
- A medián és a módusz az adatsor egy-egy kitüntetett értékét jelölik. **Hamis** (a módusz lehet több is)
- A módusz megegyezhet a mediánnal. **Igaz**
- Találunk olyan reguláris kategorikus változót, amelynek mediánja megegyezik az általunk kiszámolt mediánnal.
- Találunk olyan reguláris kategorikus változót, amelynek módusza megegyezik az általunk kiszámolt mediánnal.

Egy folytonos változó jellemző értékeit doboz diagrammal (boxplottal) és hisztogrammal is ábrázoljuk. Válassza ki az igaz állításokat!

- A doboz diagramról mindig könnyedén leolvasható az első kvartilis. **Igaz**
- A hisztogramról mindig könnyedén leolvasható az első kvartilis. **Hamis**
- A doboz diagramról mindig könnyedén leolvasható a 40. percentilis **Hamis**
- A hisztogramról mindig könnyedén leolvasható a 40. percentilis **Hamis**
- A doboz diagramról mindig könnyedén leolvasható a módusz. **Hamis**
- A hisztogramról mindig könnyedén leolvasható a módusz. **Igaz**
- Minden információ ami a doboz diagramról könnyen leolvasható a hisztogramról is könnyen leolvasható, emiatt tekintjük a doboz diagramot a hisztogram egyfajta absztrakciójának. **Hamis**

Egy adatsor *a*, *b*, *c* és *d* változóját ábrázoltuk hisztogramon és boxploton is, de sajnos a boxplotok címkéi elvesztek, így nem tudjuk, mely ábrák tartoznak ugyanazokhoz a változókhoz. Válassza ki az igaz állításokat!



- Az 1-es boxplot biztosan a *c* hisztogramhoz tartozik **Hamis**
- A 2-es boxplot biztosan a *b* hisztogramhoz tartozik **Hamis**
- A 3-as boxplot biztosan az *a* hisztogramhoz tartozik. **Igaz**
- A 4-es boxplot biztosan a *d* hisztogramhoz tartozik. **Hamis**

3 Nagyméretű adatok vizualizációja

Mik a disztributív, algebrai, holisztikus típusú statisztikai aggregátorok? Hová tartozik a szórás, az IQR és a percentilis?

Disztributív

- egyetlen, adott méretű köztestár
- eredmények kombinálhatóak
- pl. count, sum

Algebrai

- disztributív statisztikák fix száma kell hozzá
- Pl. átlag: count + sum, szórás

Holisztikus

- bemenettel növekvő köztestár kell
- Pl. medián: legalább az elemek fele, percentilis, IQR

Mi a *small multiplies* elv a vizualizációban? Miért különösen fontos ez a nagyméretű adatsorok vizualizációja témakörben?

Small multiples vizualizáció: rendező attribútumlista többelemes is lehet (a *tabplotban* egy változó szerint rendezünk csak)

Általános elv: *small multiples*

- Ugyanazt rajzoljuk ki kategóriák szerint
- Pl., matematika-olvasás pontszám *scatterplotja* országonként

- R specifikusan: *facet in ggplot2, trellis in lattice*

Mi az alapvető különbség a Map&Reduce és a Divide&Recombine minták között?

Mondjon néhány alapvető megközelítést/tippet/trükköt nagyméretű adatsorok vizualizációjára!

- **Scagnostic measures**: Megpróbáljuk kitalálni a 2D scatterplotból az összefüggést, amit ábrázol
- **Bigvis**: simítás
- **Tableplot**: zoom

Tároljunk HDFS-ben egy folytonos megfigyelt változó feletti, időbélyeggel ellátott megfigyeléseket (pl. "timestamp, value" szerkezetű CSV). Hogyan állítaná elő a megfigyelések hisztogramját MapReduce algoritmusszervezéssel? (Pszeudokódot is kérünk.)

Tároljunk HDFS-ben két folytonos megfigyelt változó feletti, időbélyeggel ellátott megfigyeléseket (pl. "timestamp, var1, var2" szerkezetű CSV). Hogyan állítaná elő a megfigyelések hőtérképét (*heatmap*) MapReduce algoritmusszervezéssel? (Pszeudokódot is kérünk.)

4 A MapReduce algoritmusszervezési minta

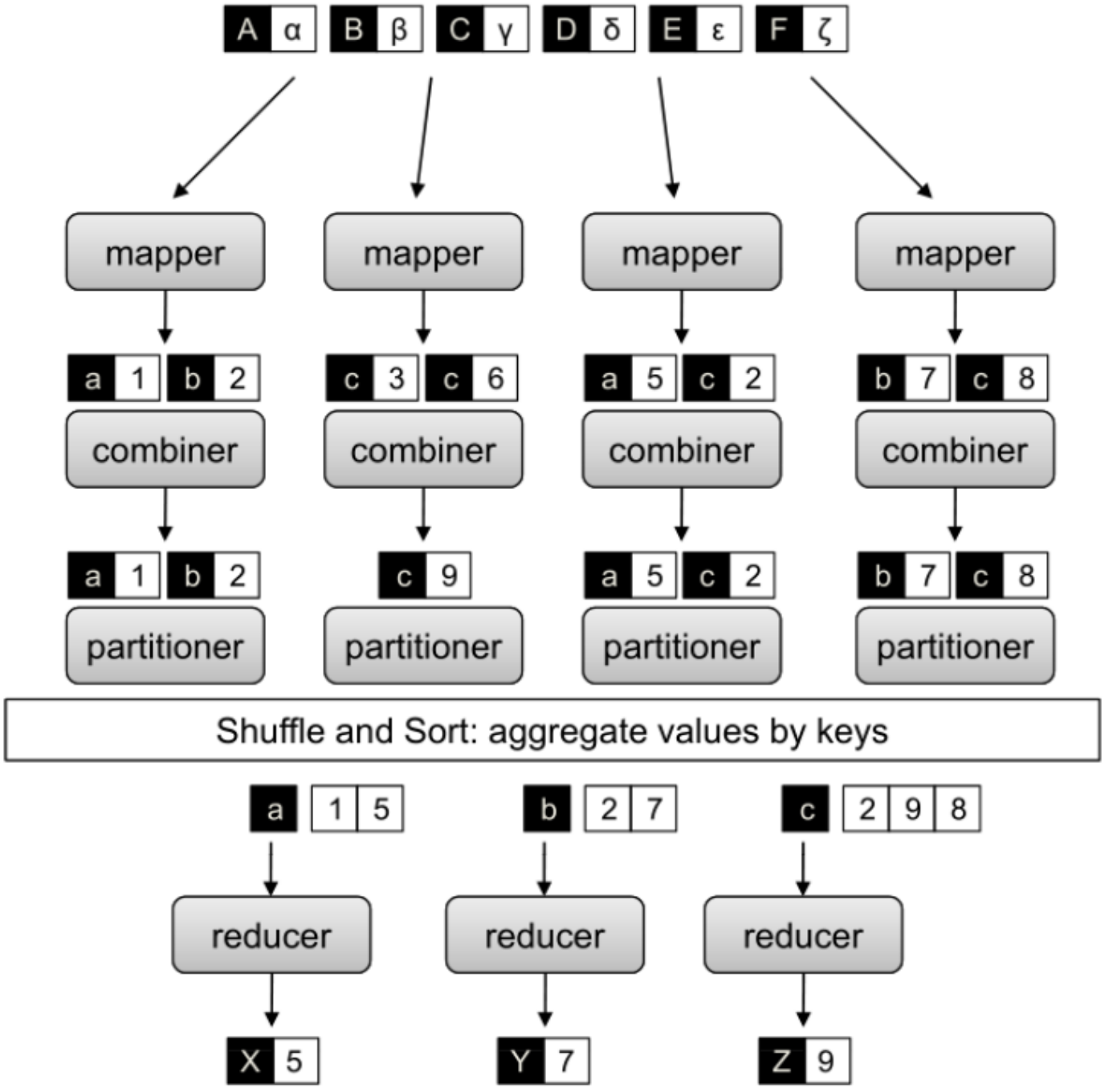
Mi a horizontális és mi a vertikális skálázási megközelítés? Az, hogy a BigData házi feladatok háromfős csapatokban oldhatóak meg, az melyik típusú erőforrás-bővítési mechanizmust követi?

A skálázhatóság két fő típusa

- **vertikálisan** skálázható: nagyobb hardver kapacitás (gyorsabb vagy több processzor, nagyobb vagy gyorsabb memória és merevlemez) nagyobb számítási teljesítményt ad
- **horizontálisan** skálázható: újabb szerverek hozzáadásával növelhető a teljesítmény

A házi feladat horizontális.

Legyen adott a következő input: 5 p dimenziós A, B, C, D, E középpontról szeretnénk eldönteni, hogy az általuk definiált Voronoi cellákban a szintén bemenetként érkező *n* darab *p* dimenziós adatpontból hányat tartalmaznak. Adjon MapReduce stílusú megoldást a problémára! A megoldási mód tetszőleges lehet, írjon például pszeudokódot, vagy töltse ki az alábbi ábrán a mapper doboz kimenetét és a reducer ki- és bemenetét! Ügyeljen arra, hogy megoldása kellően konkrét legyen, a megoldási elgondolást tartalmazó szöveges megoldást nem fogadjuk el.



Mi a “shuffle and sort” fázis feladata a MapReduce végrehajtás során?

Aggregate values by keys, a kulcsok mentén összesíti az értékeket.

A kiterjesztett MapReduce sémában mi a “combiner” feladata? Miért érdemes alkalmazni?

Csökkenteni lehet vele a lemezre írt és a hálózaton átküldött adat mennyiségét.

Tároljunk a HDFS-ben fix formátumú CSV állományokat, melyek *n* folytonos változó feletti megfigyeléseket írnak le egy időbélyeggel kiegészítve. Adjon Mapper és Reducer pszeudokódot az egyes megfigyelt változók időbeli maximum-helyének meghatározására!

K-means klaszterezés megvalósítása MapReduce segítségével: algoritmusszervezés szöveges ismertetése, map és reduce pszeudokódok.

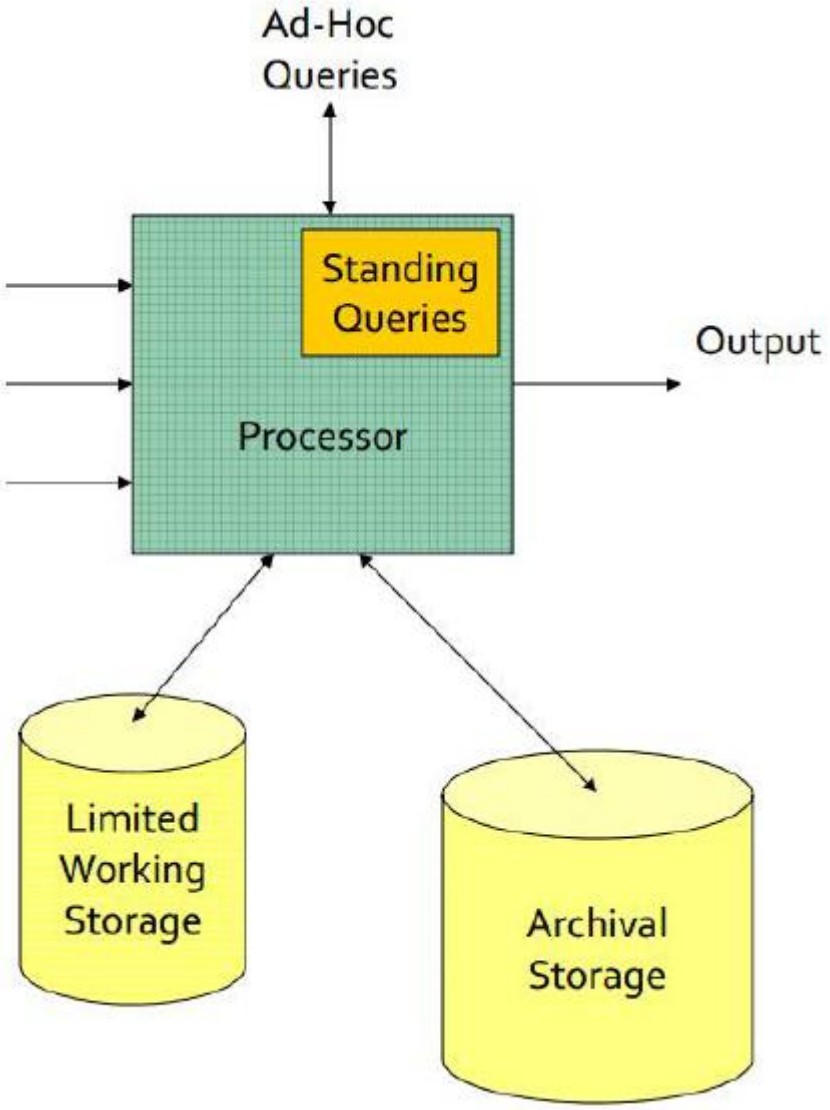
```
kmeans.map =
      /*Map kap néhány pontot*/
      function(., P) {
            nearest = {
                  if(is.null(C))
                        /*Első sor inicializálása*/
                        sample(1:num.clusters,nrow(P),
                              replace = T)
                  else {
                        D = dist.fun(C, P)
                        /*Legközelebbi klaszter*/
                        nearest = max.col(-D)}}

            if(!(combine || in.memory.combine))
                  keyval(nearest, P)
            else
                  keyval(nearest, cbind(1, P))}
/*P pont Ci klasztertől vett távolsága*/
dist.fun = function(C, P){
      apply(C,
            /*Klaszter középpontok mátrixának minde sora*/
            1,
            function(x)
                  colSums((t(P) - x)^2))}

kmeans.reduce = {
if (!(combine||in.memory.combine))
      function(., P)
            t(as.matrix(apply(P, 2, mean)))
else
      /*k klaszerközépponthoz lekérjük az összes P pontot*/
      function(k, P)
            keyval(
                  k,
                  /*Előbb csak összeget számolunk*/
                  t(as.matrix(apply(P,2,sum))))}
```

5 Adatfolyam-feldolgozás

Ismertesse az adatfolyam-feldolgozás elemi blokkjának tekintett “stream processor” mintát! Hogyan történik ezekkel a bejövő adatfolyamok feldolgozása? Mit értünk adatfolyam-feldolgozás esetén az alábbi fogalmakon: ismeretlen mintavételezési gyakoriság (unknown sampling frequency), korlátozott számítási/tárolási kapacitás (limited computational/storing resources), ad-hoc és állandó lekérdezések (ad-hoc and standing queries)?



Ismeretlen mintavételezési gyakoriság

Nem ismert mintavételezési időköz, vagy akár nem ismert nem szabályos a mintavételezési időköz.

Milyen problémák merülhetnek fel adatfolyamok mintavételezésénél? Vázoljon egy lehetséges megoldást kulcs-érték párok kulcsok fölötti mintavételezésének problémájára?

Probléma

Egy kulcsnak vagy minden értéke megjelenjen, vagy egy sem.

Megoldás

- *a/b*méretű mintához *a* (kulcstér) méretű folyamon a kulcsot *b* vödörbe hasheljük
- A hash-függvény valójában „konzisztens random-generátor”: $a < b$ esetén tárolunk
- Nem véges minta – kisebb módosítás

Példa

A felhasználók mekkora része ismétel meg lekérdezéseket” a felhasználók 1/10 mintáján

Mik a Bloom filterek? Térjen ki a bitvektor és a hash függvények szerepére a megközelítésben! Hogyan alkalmazzuk őket halmazba tartozás közelítő ellenőrzésére adatfolyam-feldolgozásban?

Bloom filter

- *n* bites vektor, kezdetben azonosan 0
- Hashfüggvények kollekciója: $h_1,h_2,\dots h_k$. Mindegyik kulcsokat rendel *n* vödörhöz (a vektor elemeinek felelnek meg).
- S: kulcshalmaz ($S = m$)

Cél

Minden $K \in S$ átengedése, a ***legtöbb*** $K \notin S$ kiszűrése –tárhely-hatékonyan

Példa

Spam email-cím alapján

Indulás

Minden j bit-et 1-re állítunk, amire van h_i és $K \in S$, hogy $h_i(K) = j$

Kulcs tesztelése: minden függvény eredménye 1 értékű bitbe visz-e

- Igen: továbbengedés (lehet hogy S-ben)
- Nem: dobás (nem lehet S-ben)

Web crawlerünkben Bloom-filtereket alkalmazunk a már látogatott URL-ek felismerésére. Bloom filterünk két hash függvényt használ, a következő paraméterekkel működik:

- N = 11
- Input: egész számok (az URL-eket reprezentálandó)
- h1(x): a páros bitekből képezett y mod N (tehát h1(585) = h1($1\mathbf{0}0\mathbf{1}0\mathbf{0}1\mathbf{0}0\mathbf{1}_2$) = 01001_2 mod 11 = 9 mod 11 = 9)
- h2(x): a páratlan bitekből képzett y mod N (tehát h2(585) = h2($\mathbf{1}0\mathbf{0}1\mathbf{0}0\mathbf{1}\mathbf{0}0\mathbf{1}_2$) = 10010_2 = 18 mod 11 = 7)

A rendszerünkben eddig a következő műveleteket hajtottuk végre: Beszúr(25), Beszúr(159). Mit ad vissza a KERES(118) művelet? Interpretálja a végeredményt!

6 Mintavételezés és anomáliadetektálás

Mutassa be a 3 tanult mintavételezési technikát és illusztrálja példával ezek működését pl. egy közvélemény-kutató cég esetén, ahol a bemeneti populáció Magyarország teljes lakossága!

SRS

- Simple Random Sample
- random mintavétel

Stratified Sample

- Homogén „réteg”
- Mindegyikből random *m*.

Cluster sample

- ~azonos méretű klaszterek
- Azokból random *m*.

Mit nevezünk kollektív anomáliának? Mi a viselkedési és kontextus anomáliák (outlierek) közötti különbség? Szemléltesse a különbséget példával!

Milyen alapvető módszerei vannak az offline outlier detektáló algoritmusok adatfolyamokra történő közvetlen adaptálásának?

Periodikus

- Minden n. adat pont után futtassuk le az *X* algoritmust
- Probléma: $x_n{-}t$ nem tudjuk jelezni

Iterált

- Minden lépésben újrafuttatjuk az *X* algoritmust
- Probléma: lassú

"Felügyelt"

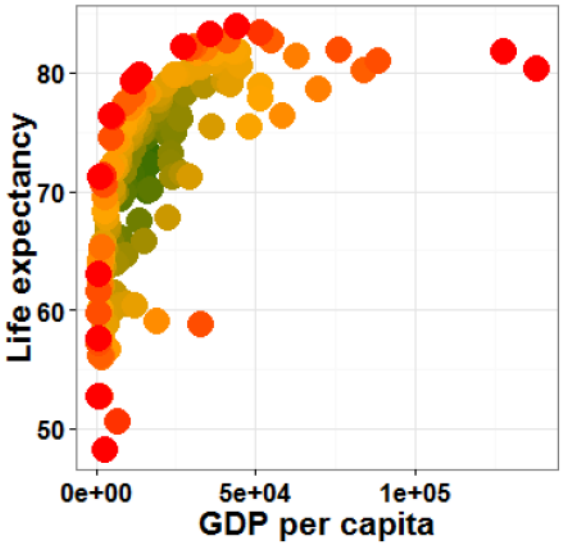
- Az elején kiszámítjuk a "normál" működést, aztán mindent ahhoz viszonyítunk
- Probléma: azx_{n+3} is outlier lesz, hiszen a normál működést nem frissítjük.

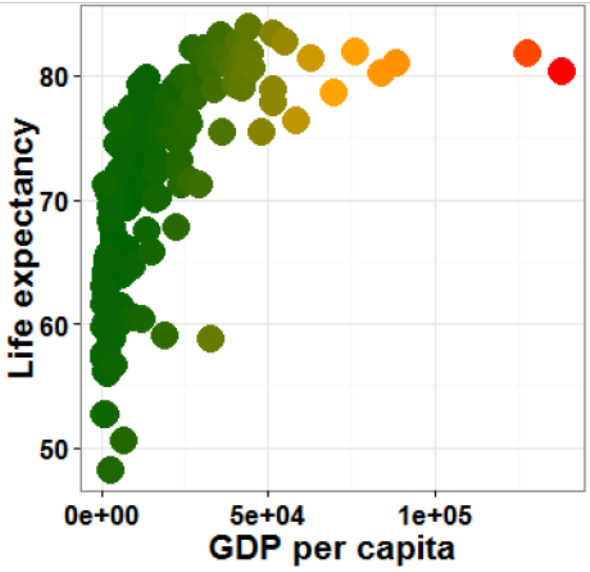
Hibák lehetnek benne!

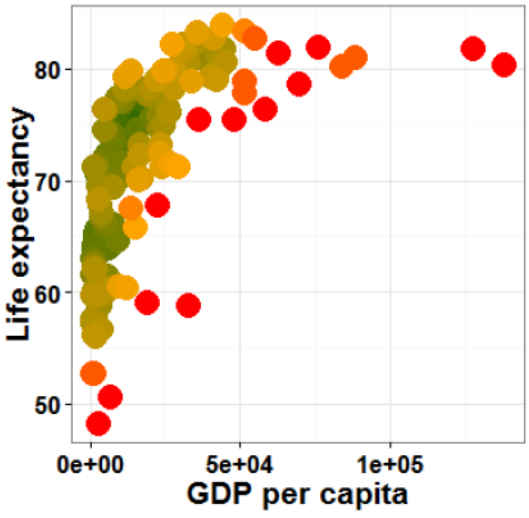
Felelősséget nem vállalok!

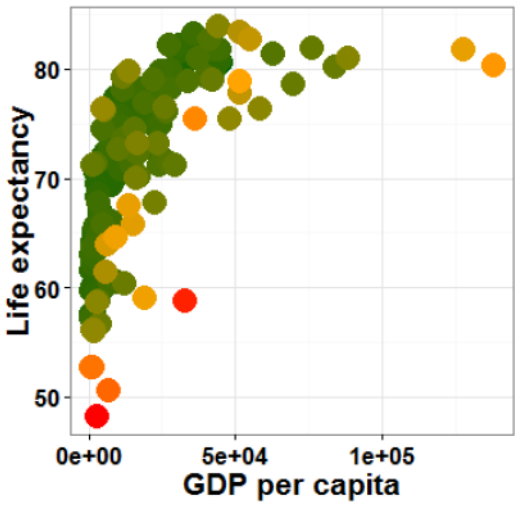
Készítette: Horváth Gábor

Az ábrán szemléltetett kétdimenziós adatsoron lefuttattuk az alábbi outlier detektáló algoritmusokat: DB, LOF, féltér-mélység keresési és BACON. A végeredményt vizualizáltuk, minden esetben a nagyobb outlier score-ral rendelkező adatpontokat színeztük pirosra. Sajnos elfelejtettük, melyik ábra melyik algoritmushoz tartozik. Párosítsuk össze az algoritmust az ábrákkal!









féltér-mélység	BACON
DB – Distance Based: szomszédok száma alacsony	LOF: a lokális sűrűség jóval kisebb, mint a szomszédjaimnak átlagosan