R intensiivisesti

Erkki Räsänen
Ecitec Oy

Päivän tavoitteet

Yleinen perehdytys R:ään; miten sitä käytetään ja mitä sillä voi tehdä

Ymmärrämme yleisimpiä analyysimenetelmiä ja osaamme tulkita tuloksia

Madallamme oppimiskynnystä ja helpotamme itseopiskelua

Sopisiko R sinun tai yrityksenne käyttöön

Datan ominaisuudet ja analyysimenetelmän valinta

tärkeimmät huomioitavat seikat

Minkälaista dataa on olemassa?

Kategorista; mies, nainen, punatukkainen, omakotitalo, koira, kissa… Data, johon liittyy kategoriat, joita ei voida verrata toisiinsa “arvon” perusteella.

Ordinaalidataa; ikäryhmät, tulotaso, erilaiset arvoluokitukset… Data, joka on luokiteltu toisiinsa verrannollisiin arvoluokkiin.

Kategorisen ja ordinaalidatan välimuoto; Likertin skaalan mukainen data

Jatkuvaa dataa; lämpötila, ilmankosteus, nopeus… Erilaiset jatkuvasti mitattavat suureet

Käytettävät analyysimenetelmät on valittava datan tyypin mukaan. Harva menetelmä sopii kaikille datatyypeille.

Miten datasta tehdään havaintoja?

Tilastollisten ominaisuuksien perusteella

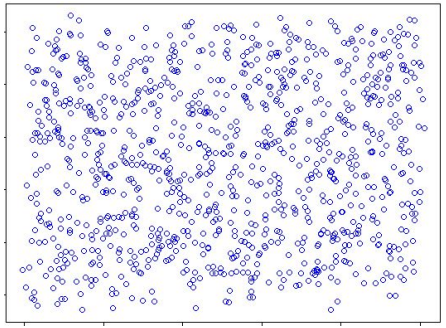
Korrelaatioiden perusteella

- korrelaatio- ja regressioanalyysi

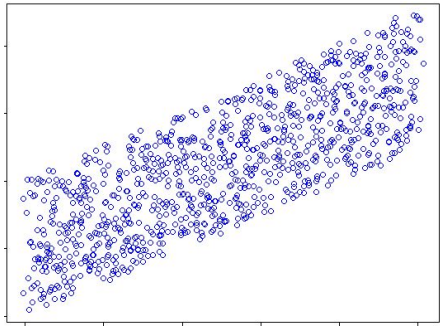
Datajoukkojen eroavaisuuksien/samankaltaisuuksien perusteella

- luokittelu, klusterointi
- pääkomponenttianalyysi
- neuroverkot...

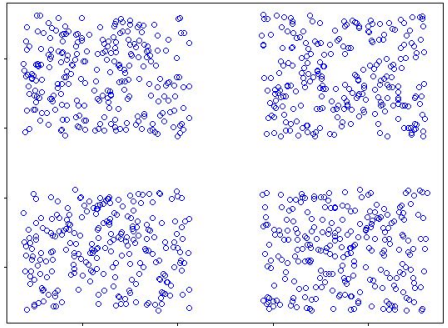
Miten datasta tehdään havaintoja?



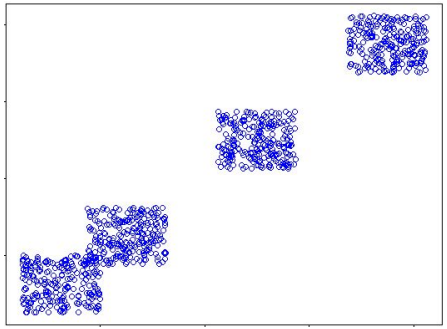
ei klustereita,
ei korrelaatiota



korrelaatio,
ei klustereita

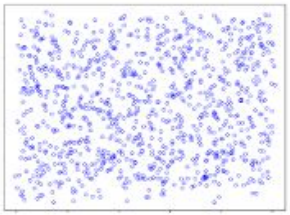


klustereita,
ei korrelaatiota

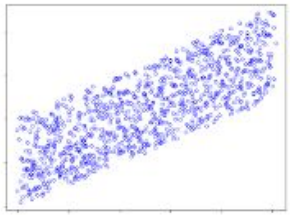


klustereita ja
korrelaatio

Datan ominaisuudet → käyttökelpoiset menetelmät

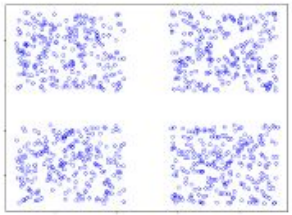


ei klustereita,
ei korrelaatiota

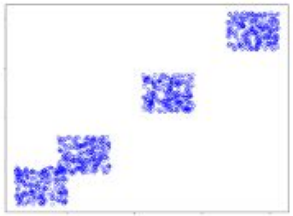


korrelaatio,
ei klustereita

Korrelatiiviset menetelmät,
riippuvuussuhteiden määrittely



klustereita,
ei korrelaatiota



klustereita ja
korrelaatio

Klusterianalyysi,
samankaltaisuuksien luokittelu

Klusterianalyysi ja
korrelatiiviset menetelmät

Miten saan analyysin onnistumaan?

Tiedosta, minkälaista dataa ja/tai muuttujia sinulla on; kategorista, ordinaalidataa, jatkuvaa… usein näitä kaikkia.

Tiedosta, mitkä menetelmät sopivat datallesi. Jos yksi menetelmä sopii joillekin muuttujille, se ei välttämättä sovi kaikille muuttujille.

Tunnista datasi erityispiirteet, jotka ilmeisimmin vaikuttavat analyysiin. Esim. aikasarjadatan korrelaatio/regressioanalyysi; kuinka leveällä aikaikkunalla analyysin voi luotettavasti tehdä.

Varmista tuloksesi useilla menetelmillä.

Hyödynnä visuaalisia menetelmiä. “Silmä näkee parhaiten”.

Kurssilla käytetyt esimerkkidatat

Markkinatutkimusdata “kuntoilutottumukset”; ainoastaan kategorisia ja ordinaalimuuttujia, *Likertin* 5-tasoinen mielipideasteikko

Prosessidataa paperikoneelta; jatkuvasti mitattuja suureita, myös muutama kategorinen muuttuja. Data on otettu tilanteesta, jolloin koneella on ollut toimintahäiriö.

IoT-dataa kiinteistön vedenkulutuksesta

R:n demonstroinnissa usein käytettyjä demodatoja; NHL & NBA Statistics, Cars, Diamonds

R on visuaalisen analytiikan työkalu

johdantoa harjoituksiin - yleisimmät analyysimenetelmät ja niiden soveltaminen

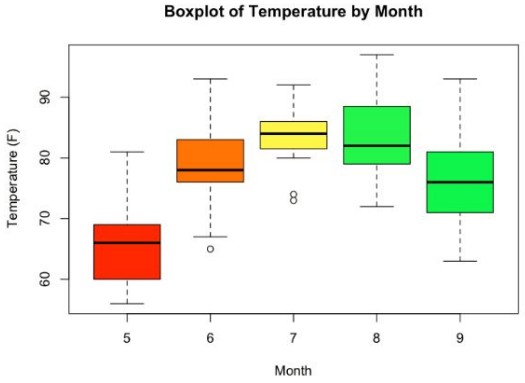
datan tilastolliset ominaisuudet ja laatu - korrelaatiot - klusterianalyysi - itseorganisoituvat kartat - muut visuaaliset analyysimenetelmät - ennustaminen aikasarjadatoilla

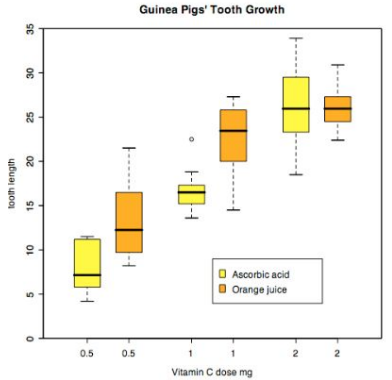
Datan tilastolliset ominaisuudet ja laatu

Hajonta, kvartiilit, keskiarvo, mediaani, poikkeamat… Box Plot on havainnollinen ja tehokas työkalu datan laadun arviointiin

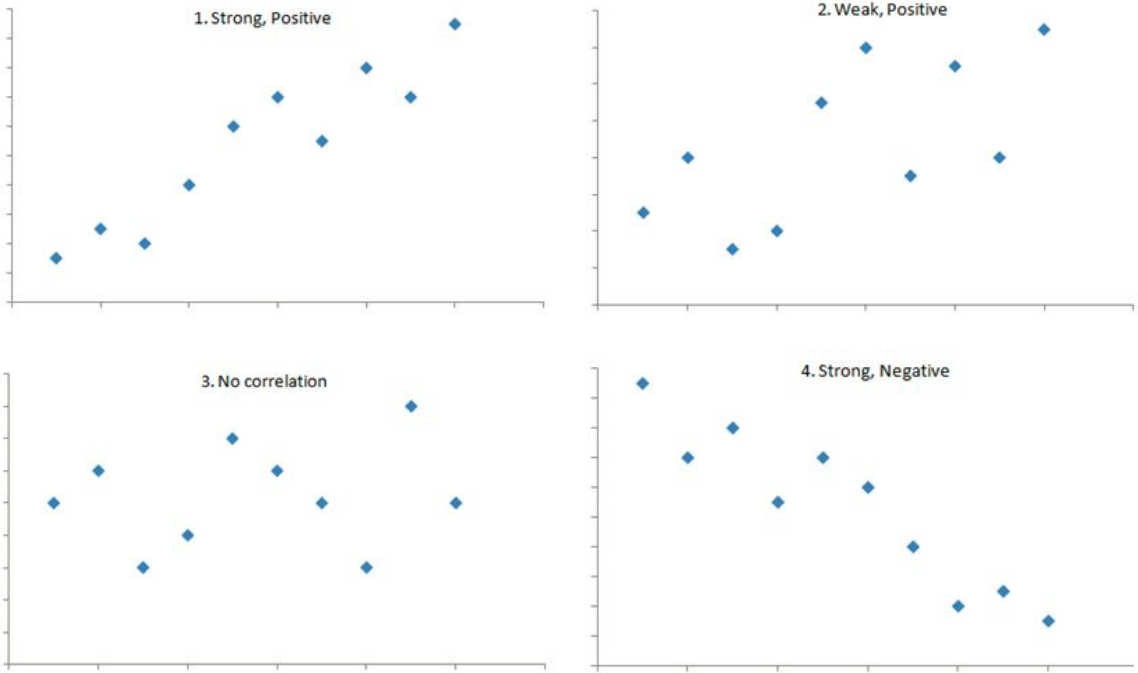
Box Plot on erittäin käyttökelpoinen analyysimenetelmä, kun vertaillaan jatkuvia datamuuttujia kategorioittain

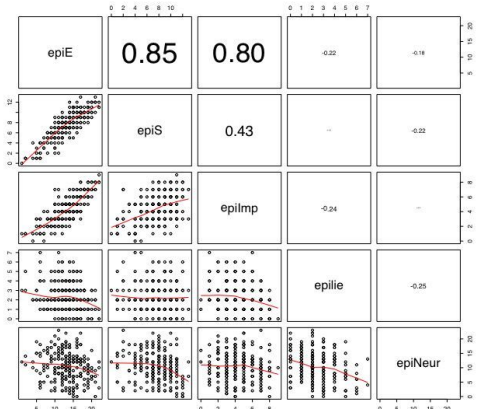
Visualisoitavana muuttujana ei kuitenkaan saa olla kategorinen tai ordinaalidata

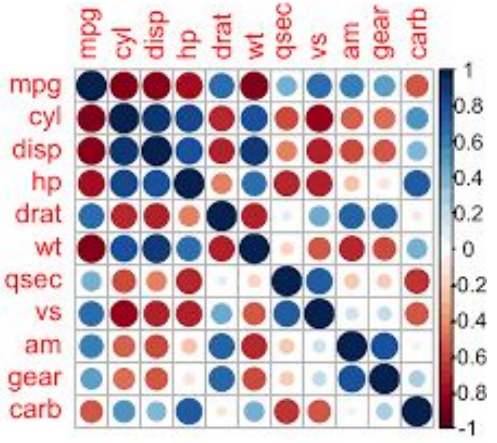




Korrelaatioanalyysi





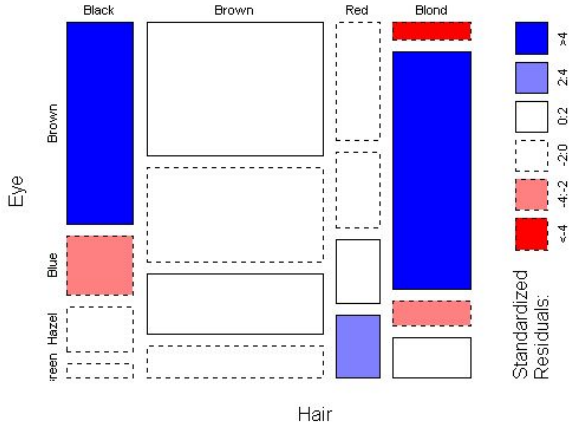


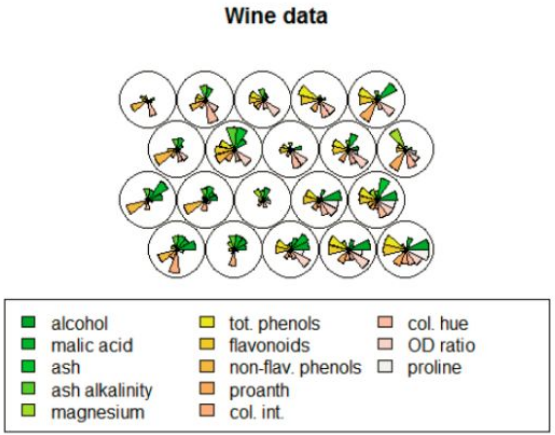
Datajoukkojen samankaltaisuus

Mosaic Plot (yläkuva); sopiva kategoriselle ja ordinaalidatalle

Itseorganisoituvat kartat (SOM, alakuva) ovat klusterianalyysin ilmentymiä

- SOM sopii datalle, jossa tarkasteltavaan asiaan liittyy paljon laadullisia muuttujia
- SOM on hyvin yleiskäyttöinen monilla sovellusalueilla

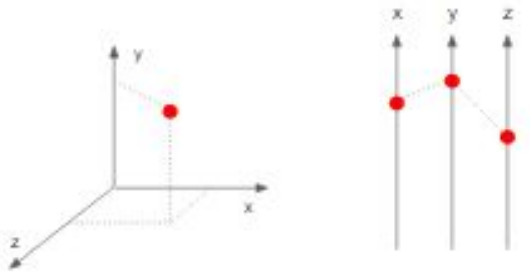


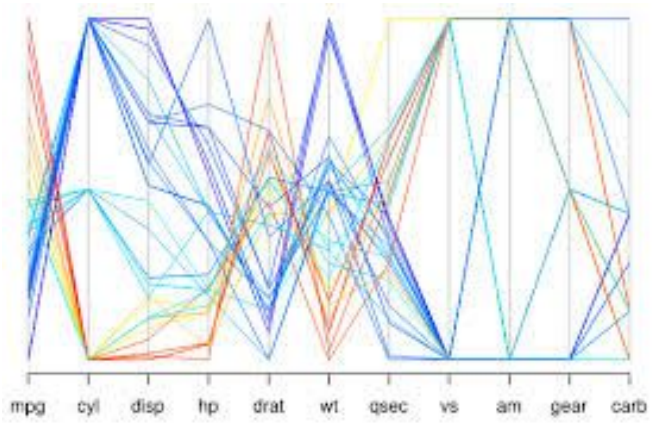


Muita käyttökelpoisia visuaalisia analyysimenetelmiä

Rinnakkaiskoordinaatisto (Parallel Plot)

- perustuu moniulotteisen datan esittämiseen rinnakkaisilla akseleilla
- erinomainen menetelmä esim. laadullisten ominaisuuksien vertailuun ja normaalista poikkeavien asioiden tai tilanteiden löytämiseen
- käytetään paljon teollisten prosessien toimintaongelmien ratkaisuun ja optimointiin





Muita käyttökelpoisia visualisointeja

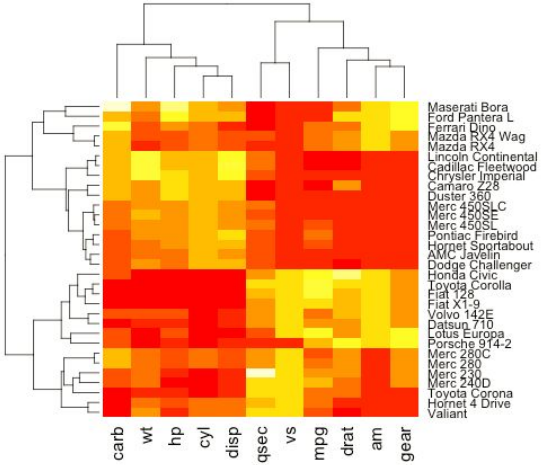
Heatmap

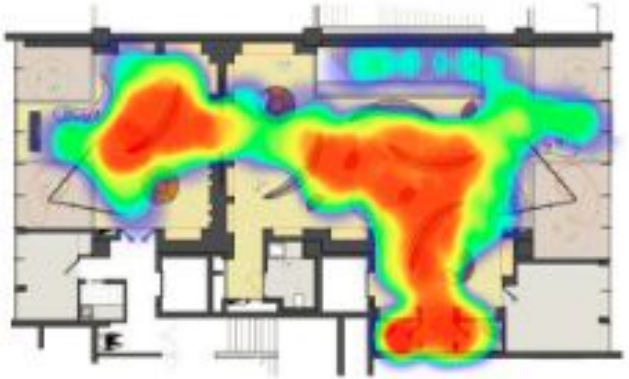
vertailutaulukko (tai koordinaatisto), jossa numeroarvot on korvattu väreillä

menetelmä helpottaa arvotasojen (min-max) löytämistä

datassa esiintyvät säännöllisyydet tulevat näkyviin (jos niitä on)

Big Data -analytiikassa paljon käytetty menetelmä



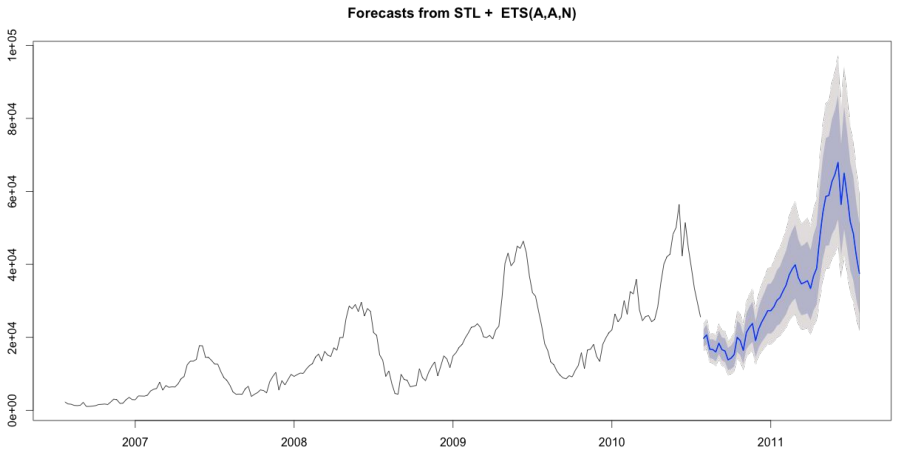


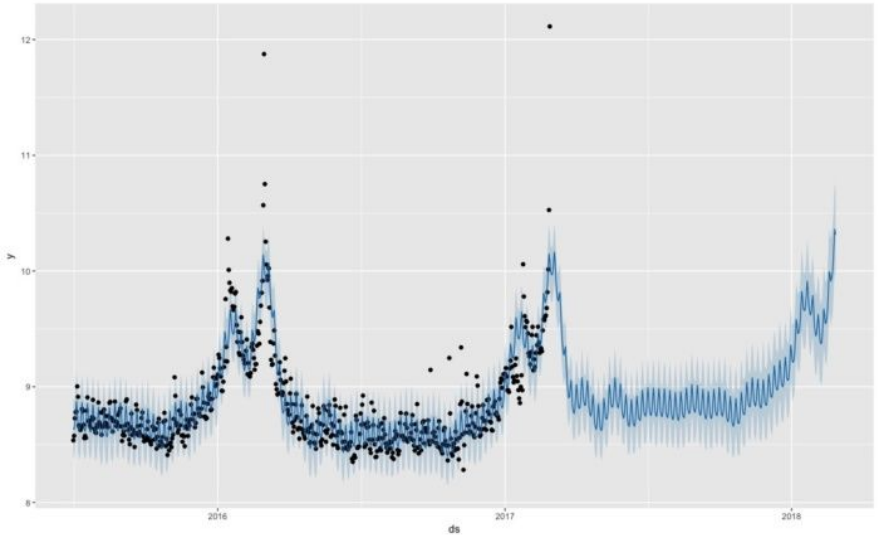
Ennustaminen (Forecast)

Ennustavaa (prediktiivistä) analytiikkaa voi tehdä monella tavalla:

- Datojen luokitteluun perustuen (SOM, klusterointi, neuroverkot) → jos luokittelun perustana on riittävän monta laatumuuttujaa, hyvä aineisto voi antaa luotettavan ennusteen
- Korrelatiiviset menetelmät → kohdemuuttujan/muuttujien käyttäytymisen ennustaminen korreloivien muuttujien perusteella
- Aikasarjamenetelmät;
  - Luokitteluun perustuvat ja korrelatiiviset menetelmät aikadynaamisesti sovellettuina → Machine Learning, Deep Learning
  - Yhden muuttujan ennustaminen sen aikaisemman käyttäytymisen perusteella (varianssi, kohina, transientit)

Ennustaminen (Forecast)





NEWS

Facebook releases 'Prophet' -- its free forecasting tools -- for Python and R

The code is available on GitHub

Interaktiivisten visualisointien ja appien toteutus

R Studioon saa lisäosat, joilla voi toteuttaa interaktiivisia web-visualisointeja ja sovelluksia;

Plotly

Shiny

Kirjastot saa ilmaiseksi rajoitetuin ominaisuuksin.

R, Big Data ja IoT

R Big Data - analytiikan työkaluna

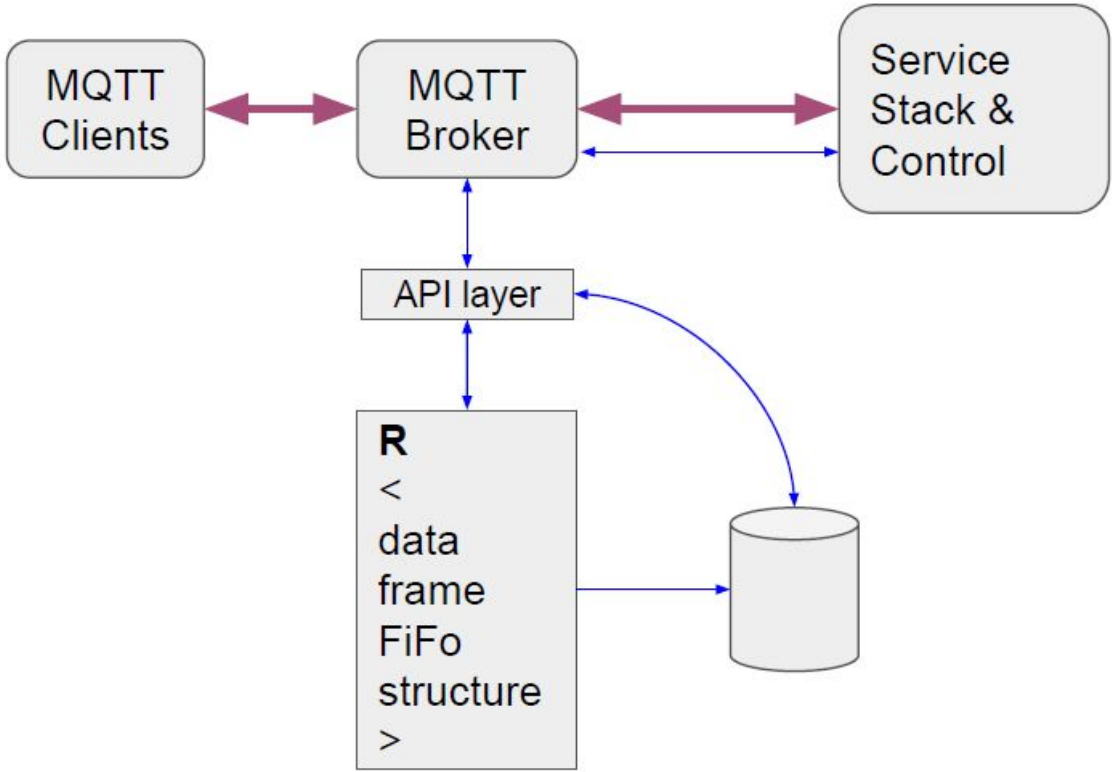
R toimii tehokkaasti klusterilaskennassa:

- HIVE - Hadoop InteractiVE
- R Hadoop Streaming API
- SparkR

Miten pääset alkuun; tarvitset klusterin, jonka voi edullisesti toteuttaa Raspberry PI -tietokoneista

Klusterin toteutus ja Big Data-työkalujen käyttö opastetaan toisessa Ecitec Oy:n kurssissa

Data processing architecture



Harjoitukset ja demot

Harjoitus 1 - datan käsittelyn perusteet

R Studio työympäristönä

Datan import / export → data frame (tiedostot ja tietokantadata)

Sarakkeisiin ja riveihin viittaaminen ja niiden käsittely

Dataobjektit ja -muuttujat

Ensimmäiset visualisoinnit plot-funktiolla

Datojen esikäsittely; datasettien yhdistäminen, sarakkeiden ja rivien poisto, poikkeamien poisto

Harjoitus 2 - datan tilastolliset ominaisuudet

Box plot -visualisointien toteutus ja tulkinta

Tarkastellaan tässä harjoituksessa myös, miten R:llä tuotettujen visualisointien yhdistäminen tapahtuu

Harjoitus 3 - klusterit ja korrelaatiot

Miten tuodaan esiin datan klusteroituminen

- jitter-funktio ja kohinan lisääminen ordinaalidataan
- hexbin-visualisointi (density plot)

Korrelaatiokartat (correlation map, correllogram)

- tavanomainen korrelaatiokartta tuo esiin korrelaatiot
- klusterointia voi hyödyntää keskenään korreloivien muuttujien ryhmittelyyn (datajoukkojen samanlaisuuksien / erilaisuuksien perusteella)

Harjoitus 4 - Self Organizing Maps

Tarkastellaan SOM-visualisointien generointia ja tulkintaa erilaisilla datoilla

- markkinatutkimusdata
- demografinen data
- NHL & NBA-statistiikka

Harjoitus 5 - muut visualisointimenetelmät

Parallell plot - prosessidata ja Cars-data

Heatmap - NHL ja NBA Stats

Big Data -esimerkki

Harjoitus 6

R:n Forecast -kirjaston käyttö aikasarjaennusteisiin

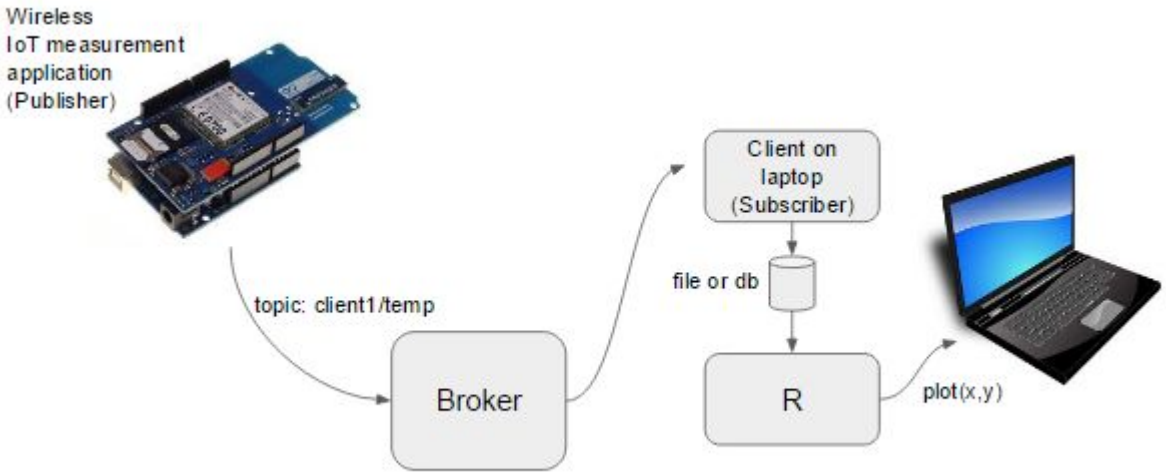
Harjoitus (demo) 7 - edistyneemmät visualisoinnit

Ggplot2 -kirjasto; R:ään sisäänrakennettu visualisoinnin “makrokieli”

Plotly; Interaktiivisten visualisointien toteutus

Shiny; Interaktiivisten appien toteutus, miten tuoda omaa dataa esimerkki-appeihin

Harjoitus (demo) 8 - reaaliaikaisen IoT-datan tuonti R:ään



Miten tästä eteenpäin?

Analytiikka R:llä vaatii koodaustaitoa. Koodaamista oppii vain, kun sitä itse tekee. Onneksi R-kieli on verrattain helposti omaksuttavissa. Kurssilla käytetyt esimerkit ovat vapaasti käytettävissä, myös netistä löytyy paljon esimerkkejä ja oppimateriaalia.

Järjestämme räätälöityjä yrityskohtaisia kursseja.

Erkki Räsänen

050 371 6229

erkki.rasanen@ecitec.fi