analystyper.md

#KvaffKap1

Deskriptiv analys

- syftar **beskriva** ett fenomen mha data

Prediktiv analys

- Syftar förutsäga framtiden (mha data?)
- Om man kan säga något om framtiden från dåtida data beror ex på om något fundamentalt ändrats med systemet - GÖR SÅDANA HÄR GRUNDLÄGGANDE FRÅGESTÄLLNINGAR

Preskriptiv analys

- syftar t algoritmiskt välja bästa beslutsalternativet

--

- Ofta vilseledande att förknippa vissa verktyg med olika analystyper
-

Använda data.md

#KvaffKap1

Data gör att man kan göra mer objecktiva slutstaser, undvika confirmation bias

- kan dock också, om man inte förstår verkligheten datan uppstår ur, och gör okristiska slutsatser, göra misstag
- data inte heller fri fr confirmation bias, men kan vara lättar att undvika än i kvalitativ analys

Data utan kritiskt tänkadande kan ge felaktiga slutsatser, det är inte datan som gör att vi förstår omvärlden

Beskrivning av mätserie.md

#KvaffKap2

- Analytikers jobb är analysera relevanta mönster från data som avslöjar något om verkligheten
- Måste dock kunna beskriva en mätseria och få översikt
- God vana att alltid skaffa översikt av dataset
  - bra översikt: förenkla så mycke som möjligt utan förvanska
- Mest översiktligt: Sammanfatta med ett antal *mått*
  - Lägesmått:
    - Medelvärde
    - median
    - mode
  - Spridningsmått:
    - Varianse
    - Standardavvikelse
- Kombination av dessa ger bild av intervall för var "typiskt" värde befinner sig
- Ovannämnda kan förvrängas av extremvärden, etc
  - Kan säga mer om fördelnignens utseende med kvantiler
    - vanligt med kvantiler (4)
- Låddiagram kan visualisera kvantiler och medelvärde
- Histogram delar i i delintervall

    - Kan ge information om outliers, extremvärden
- Täthetskurva = en utjämning av histogram
    - Om liknar en täthetsfunktion kan vi använda den för att modellera

Jämförelse mellan mätserier

- Ofta interesant för analytiker
- Kan jämföra lägesmått, spridningsmått, eller grafiskt genom histogram
- Analytiker föredrar i många fall överskådlighet ist för nyanser - gillar arbeta med jämförelser medelvärden
    - Jämförelse av medelvärden kan dock ge oriktiga bilder - kan ha varit väldigt varmt en dag och ungefär lika varmt resten av dagarna, medelvärde säger ändå att ena är varmare
- --> Ofta vanligt vilja ha mer kvalificerad jämförelse, formellt hypotesttest
    - Kan testa nollhypotes medelvärden skiljer sig ej åt
    - gör genom t-test, ok enl CGVS om lagom stora mätserier

Om har flera mätserier behövs generaliserat test för jämföra varians i dem

- Standardtest variansanalysis, ANOVA #läsmerkvaff
    - Om inte ger signifikant resultat förkastar man INTE att serierna har samma medelvärde
    - om ger signifikant res vill veta vilka serier som skiljer sig
        - Kan vara problematiskt jämföra 2 och 2: risk för felaktigt signifikant res ökar med antal mätserier
        - kan köra post-hoc test #läsmerkvaff efter anova som korrigerar för detta, bonferroni-korrigering för göra svårare för resultat att vara signifikanta #läsmerkvaff

Beskrivning av samvariation

- korrelation och kovarians bra mått
- Kan vara så att mätserier korrelerar väl för vissa värden, dåligt för andra, måste använda scatter plots för att undersöka detta

felaktig data och rensning.md

#KvaffKap3
dålig reliabilitet och validitet kan uppstå genom att man helt enkelt gjort fel

- Kan ex ha mätfel vid uppsamling, man tar observationer som inte är från målpopulation, blir fel
- Kan kolla på extrema mätvärden o relatera de till den datagenererande processen, är det rimligt att de kommer därifrån?
    - Kan "trimma data"
        - = Ersätta höga med låga värden (bra om misstänkar att värde är för högt, men ändå ska vara högt)
    - Om misstänker data är helt fel, kan radera helt
    - Måste dock vara uppmärksam att inte radera utstickande värden: De kan ha verkliga implikationer / "En sannolikhet som inte kan avfärdas"
        - Ex ovanlig börskrasch, vulkanutbrott, blodpropp
    - Måste ha förståelse för datagenererande processen för att kunna avgöra allt detta
    - kan vara frestande göra justeringar för att ex få serie likna någon fördelning, men man bör bara ändra med grund i förståelse för hur datan kom till
        - Utstickande värden kan vara väldigt interessanta/påverka analys, bör inte ta bort de *enbart* för att de sticker ut
        -

reliabilitet och validitet.md

#KvaffKap3

Reliabilitet:

- Att mätning av samma underliggande storhet ger samma utfall till stor grad oberoende av när, hur, var mätningar görs
- Mätningar konsistenta?

Validitet:

- Hur väl våra mätvärden faktiskt mäter den underliggande storheten;
- Ex våg med för tunga motvikter, låg validitet
- ofta svårare att utgöra än reliabilitet

Om den underliggande storheten kanske inte är förknippat med ett mått och en mätmetod (ex som mätning av vikt med våg är), ex pris kund villig att betala, kan mycket mätdata, många mätserier krävas för att få god validitet

Reliabilitet och validitet avgör tillsammans förutsättningar för att data ska kunna vara användbar i analys

- validitet - väntesvärdesriktigt?
- Måste ha hög reliabilitet för att ha hög validitet - för stort konfidensintervall gör att vi inte mäter rätt underliggande storhet lagom väl (?)
- kan vara filosofisk fråga att säga om dålig data beror på dålig reliabilitet/validitet

---

- väldigt vag skattning kan inte sägas ha hög validitet
  - ex om gjort för få observationer för skattning
  - stora talens lag

För bedömma om problem med reliabilitet och validitet måste bedömma hur data förhåller sig till den verklighet man undersöker

- = "*Att förstå den datagenererande processen*"
  - Denna ofta komplicerad och omöjlig att helt förstå, behöver modellera matematikst genom ex normalfördelning
- **förstå den datagerenerande processen kräver två typer förståelse:**
  - *god insikt i verkligheten interesserade av*
  - *insikt i metoder som används för att samla data*
  - exempel:
    - Mäta trafiken utanför skola morgonar under 30 min
      - dessa måste vara representativa av morgonar i framtiden
      - hög spridning; högre trafik fredagar
        - om inte inser detta skattar man för högt, måste ha insikt i verkligheten
        - Betingat väntevärde kan ofta mätas med högre validitet / reliabilitet
      - mer insikt gör att man kan välja bättre mått, bätre mätmetod
        - <-- man kan förbättra sätt att mäta, eller mäta faktorer som påverkar utfallen man är interesserad av

Slutsatser från data.md

#KvaffKap3

efter har gjort Beskrivning av mätserie eller dataset, dags att ställa frågan vad data kan användas till

- Nyckel t god analys: *Data måste kunna ge god representation av verkligheten vi är interesserad av*

Kan vara lätt att för djuptgående dyka in i data, god analytiker håller alltid anlytiska problemet och verkligheten inom synhåll

- avancerade analysmetoder kompenserar inte för brister i data

- Begreppen reliabilitet och validitet ger översikt, används ofta i kvalitativt arbete, användbara i diskussion av vad man kan och inte kan göra med data

Vanliga misstag vid tolkning/insamling av data

- ställ enkla frågor om komplicerade förteelser
  - ex svårtolkad eller för öppen
- kan få olika svar från samma person vid olika tillfällen
  - Vanligt om man ställer frågor om ex värderingar, eller om frågor m ord och begrepp som är onaturliga för den utfrågade
- Måste se till att inte leda den utfrågade mot ett visst svar, genom sammanhang eller frågeställning (ger lägre validitet)
  - ändrar ofta beteende om vet blir bedömd
- Man kan övertolka data, analysera data som bara "råkar finnas"
  - Ex patentdata används för modellera innovation, men de är inte direkt korrelerad
  - Ekonomer kan bli frestade att göra det ändå
  - kan undvika kritik genom formulera sin studie som en studie av det man mäter (ex patent) istället för underliggande storhet (ex innovation)
    - Kan ske glidning: över tid tar måttet platset som storheten man är intereserad av
    - ex KPI:er man har blivit mer interesserad av att öka dessa, de blir mål, än att öka performancen de är avsedda mäta (man har löst ett problem med validitet (bristande koppling verklighet - data) genom att ändra verkligheten istället för metodiken kring datan)

se validitet och urval

validitet och urval.md

#KvaffKap3

i reliabilitet och validitet fokus *begreppsvaliditet:* representera teoretisk storhet med mätbar storhet på ett riktigt sätt

- I många andra fall är urvalet av data främst utmanning för uppnå god validitet
  - hjälper inte att mått perfekt om urval inte representativa
- De man vill studera = *Målpopulation* (företag, individer, mängd)
- Rampopulation = de man kan nå för undersökningen
- Urval/stickprov: den grupp/mängd man har lyckats samla in användbar data från
- Om har tillg´ång till data om hela rampopulationen behöver inte dra stickprov: Totalundersökning
- Ex målpopulation: alla medlemmar, rampopulation: alla medlemmar man har kontaktuppgifter till, stickprov: alla man väljer att skicka till och som svarar

Olika validitet i olika fall

- ideal: totalundersökning
- Näst bäst: urval som inte avviker på viktigt sätt från målpopulationen
  - uppnås genom slumpmässigt urval/stickprov fårn rampopulation som inte avviker från målpopulation
  - Annars: Tänka på om olika undergrupper som är representerade på olika sätt finns i målpopulationen? hur speglas dessa i rampopulation respektive stickprovet?
  - *Selektionseffekter* när de ex systematiskt väljs objekt från en viss undergrupp, eller skevhet uppstår, vilket påverkar datan (och sänker validitet?)
    - ex: likgiltiga kunder underrepresenterade i stickprov (de svarar sällan), ändrar resultat
    - rampopulation kan representera målpopulation dåligt, stickprov kan representera rampopulation dåligt, ex rädsla terror-undersökning (alla människor, folk på gatan, de som går med på utfrågning)

modeller och samband.md

#KvaffKap4

stokastisk variabel är grundläggande modell, beskriver en datagenererande process

- i enklaste form beror utfall på fasta parametrar, definierar fördelningen
- i andra fall fokus på samband melllan variabler

För beskriva process kan diagram med lådor och pilar mellan komponenter av process användas

- lådor = variabler, pilar = samband ("kausala diagram")

Hur kan modeller se ut?

- Modeller kan bestå av ekvationer, ekvationsystem, ex i klassisk fysik
- mer generellt ett antal storheter, regler, beskrivningar för hur storheterna hänger ihop
    - regler representerar orsakssamband
        - kvantitativ analys interesserad av hur starka orssakksamband är

Två typer av orsakksamband:

1. Deterministiska kausala samband:
    1. Sådant som (empiriskt) alltid gäller
    2. regelbundna samband
    3. ex newtons lagar
2. Statisktiskt kausala samband:
    1. Mindre regelbundna
    2. ex inom medecin, ekonomi
    3. man kan inte automatiskt hävda kausalitet från data
        1. ex bara för X mer "statisktiskt signinfikant" korrelerad med Y ***innebär inte*** X och Y kausalt kopplade
        2. Logik funkar ist åt andra hållet: om X o Y kausalt kopplade finns nog korrelation
    4. Kan inte lita på att går utforska alla kausala samband genom undersöka korrelation, kan finnas gömd 3e variabel som ingår i kausala sambandet men inte undersöks, => ingen korrelation uppmäts

Modeller som analytiska verktyg.md

#KvaffKap4
Modeller förändrar komplex verklighet/värld till något greppbart / översiktligt

modeller som representation

- modell = representation komplexa/mångtydliga verkligheter
- Människans vardagilga tänkande bygger på tumregler/heurestiker, slags modeller
- modeller samlar erfarenheter som tolkas som generellt mönster
    - ex antiker som värderar möbel
        - är en informell modell, ofta bättre i analys med kvantitativa data med formell modell

Vad är en bra modell??

- teori = systematisk förståelse av fenomen
- Modell = konkret manifestation av teori
    - Behöver förhålla sig till teori för göra bra modell, för få hög reliabilitet och validitet i föreänderlig värld
- Bra model != modell som beskriver världen så exakt som möjligt
    - Istället en som lätt hjälper en få syn på några interessanta aspekter av fenomenet
    - modellera bara det avgränsade problemet!
- Hur en modell värderas beror på hur den ska användas
    - Finns skillnad mellan att modellera systemet man hämtar data från vs ett mer allmän princip
    - val av analysmetod beror på förhållandet mellan det man vill förutsäga och historiken man använder för leta efter mönster
    - ***Finns tre viktiga typfall:***
    - 1. **Modellen som Prediktionsmaskin**

        - om stabilt system kan använda algoritmiska metoder för finna bästa beskrivning (både i struktur och parameterval)
        - stabilt = systemet fungerar på samma sätt i framtiden som under tiden data genererades
        - Bra modell = en som skapar riktig output med indata som inte använts under träning (*out of sample prediction*)
            - denna förmåga beskrivs av roc-kurva elr precision recall-kurva #läsmerkvaff
        - Maskininlärning vanligt för dessa tillämpningar
            - Ex stödvektormaskin (support vector machine), klassificeierar/skiljer datapunkter #läsmerkvaff
            - Random forest: aggregera flera beslutsträd t genomsnittsmodell
                - #clarifyKvaff s 41 hur random forest funkar, fatta inte riktigt #läsmerkvaff
    - 2. **Modellen som systemrepresentation/beskrivning av systemdynamik**
        - om modellerar system som är förändligt måste förändringen modelleras utifrån teoretisk förståelse
        - =>Måste ha minne, bero på *tillståndsvariabler* som ändras över tid och som har påverkanssamband med andra tillståndsvariabler. Påverkanssamband kan beskrivas med parametrar/kunskap från tidigare undersökningar eller skattas
        - system är ofta system av differentialekvationer eller differensekvationer. matematiskt Komplexa, modellera ist med simuleringar => möjligt utforska utfall som går bortom observerade data
        - Dessa typer av modeller används ex för logistik, naturliga ekosystem
        - Bra modell i denna klass leder till insikt om hur underliggande system kommer bete sig i framtiden
        - kan kanske göra presktiptiv analys baserat på utfallen'
        - Enklaste form av denna typ av modell: länkade variabler med återkopplingsstruktur
    - 3. **Modellen som inferensverktyg**
        - Ibland inte datagenererande systemet som interessant utan underliggande principer som tar uttryck i systemet
        - ofta interesserad orsakssamaband/samvariation mellan två/flera faktorer
        - naturligt utgå från teoiretisk förståelse för modellera strukturen på system, skatta parametrar
            - I detta fall är det istället *parameterskattningarna* själva som är interessanta istället för prediktioner / dynamiken från modellen
            - ex skattning av koeffecienter i linjär regression (?) för att påvisa sambandsstyrka

modellera relationer mellan två variabler.md

#KvaffKap4

Tre slags grundläggande relationer mellan två variabler X o Y:

- Direkt samband: X o Y beroende
- inderekta medierade samband: Y beror av Z, Z beror av X
- Indirekta modererade samband: Y beror av Z, styrkan av sambandet beror av X

Enklast anta linjärt samband mellan två variabler: $Y = \alpha + \beta X$

om linjär i parametrar man skattar kan man göra (linjär) regression med vilken formel för y som helst

- kan man ex göra att logaritmen av y beror på X, $ln(Y) = \alpha + \beta X$
- eller godtyckligt polynom $Y = \alpha + \beta_1 X + \beta_2 X^2$
    - Om signifikant att $\beta_2 = 0$ så visat sambandet linjärt

olika sorters variabler.md

#KvaffKap4

Viktigt avgöra vilket utfallsrum en variabel kan tänkas ha
viktigt avgöra om den är kvantitativ eller kvallitativa värden

- Kvalitativa värden

    - Kan vara rent nominella (ex färg, genus), kan användas för kbalificiering men inte skala/rangordning
    - Kvalitativa variabler kan motsvara ordinalskala, inbördes rangordning finns (ex betyg, ordnade men går inte automatiskt att översätta t siffror)
- Ofta omvandlas nominella variabler till indikatorvariabler (kvantitativa) för att kunna använda statistiska verktyg
    - Ex omvandla "färg" (nominellt) till "röd", "blå", "grön" som alla är antingen 0 eller 1 (kvantitativa)
- Om översätter ordinala värden till siffror måste man se upp vilka antaganden man gör, annars blir inte analys rimlig (låg valididtet?)
    - ex E,D,C,B,A -> 0,1,2,3,4 är väldigt starkt antagande att betygen är ekvidistanta
- Finns observerade/latenta variabler
    - Observerade: ganska direkt koppling observerbara världen och modellen
    - Latent: abstraktion. Tänkt fånga konceptuell viktig faktor, till natur svår/omöjlig fånga upp i praktisk data (ex beteende/personlighet)
        - exemplet kan modelleras som oobserverad förklaringsfaktor till observerbara variabler som hämtas ex från personlighetstest

Analysalternativ.md

#KvaffKap5

Analysalternativ

finns många verktyg tillgängliga, ex maskininlärning etc, lätt komma in i

- dock ofta svårtolkade/komplicerade resultat
    - => viktigt överväga om man kan göra tolkningar man interesserad av med enklare deskriptiv analys

För exmpel att mäta om kvinnor äter mer frukt än män:

Enkla medelvärden

- beräkna medelvärde för fruktkonsumption för de datapunkter med kategoriskt (nominellt) värde män respektive kvinnor
    - Måste se om statisktiskt signifikant bör använda något test:
        - t-test (jämförelse av medelvärde mellan två dataserier med kontinuerliga värden)
        - ANOVA-test (jämförelse av medelvärde mellan två dataserier med kontinuerliga värden)
        - χ^2-test (jämförelse mellan två eller felera dataserier med kategoriska (nominala eller ordinala) dataserier)

Normering av data

- Man kanske fått signifikant svar på frågan man ställde tidigare, och god validitet, men vet inte exakta orsaker bakom observerade resultat
    - Kan av denna anldening studera normerade data: ex *kvoten* mellan fruktkonsumption och total matkonsumption mellan datapunkter med olika kategoriska/nominella värden på kön för undersöka preferenser

Trimming av data

- Ibland kan finnas faktorer som är korrelerade med variabeln man är interesserad av, och som systematiskt är korrelarad/annorlunda mellan olika kategorier (män/kvinnor)
- Ex kanske fruktkonsumption beror av hur många ggr/vecka man tränar, och denna systematiskt olika för män/kvinnot
- -> svårt att tolka vad resultaten faktiskt beror på, kanske vill undersöka skillnaden mellan två kategorier *allt annat lika*
    - Ett primitivt sätt åtgärda allt detta: Skära bort mätpunkter där den störande faktorn är över en viss gräns etc
        - Nackdelar: Studien får bara validitet för mindre del av stickprovet, och gör sig av med observationer, sämre precision i punktskattningar

Regressionsanalys

Kan applicera multipel regressionsanalys, ex av normerade värden:
$Konsumtion = \alpha + \beta_1 \cdot kön + \beta_2 \cdot träningsfrekvens$

och ta β_1 (m konfidensintervall) som en indikator på styrkan av ev. samband mellan kön och konsumtion

- Mycket mer svårtolkat, kräver tekniskt förståelse hos mottagare av analys

utforska mönster i data.md

#KvaffKap5
Minns många val när modell av problem ska göras
enkel modell kan vara mer användbar när ska kommunicera och identifiera mönster i data

se Analysalternativ

Studera orsak och samband.md

#KvaffKap6
Analytiker uppgift utoforska möjliga framtider, måste etablera ide om kausala samband orsak och verkan
--> Hur undersöka orssakssammanbad mellan två storheter?

Hypotetisk-deduktiv vetenskap

klokt när studera orsak verkan amtt använda *the scientific method*

- analys formas kring hypoteser som ankras i bästa möjliga befintliga uppfattning av fenomenet
- Interessant hypotes kan tydligt falsifieras, bestäm innan vilka observationer som leder t förkastning
- hypotes testas genom empirisk undersökning - analys, som ger underlag för slutsatser
  - för att komma ifrån ex confirmation bias
  - strukturerat arbete -> kan tydliggöra antaganden och slutsatsers validitet för sig själv/andra
  - hypotesprövande arbetssätt -> uppfattning endast kombination teoretiska idéer och empirisk undersökning som ger använbar kunskap
- arbete genom *scientific method* vanligt i naturvetenskap, lite mindre vanligt i samhällsvetenskap
- Om man tror på scientific method påverkar vilka analysmetoder man väljer
  - Ex om tror på det kan betrakta kvalitativa data som mindre tillförlitliga kvantitativa data
  - Andra menar enda kunskap är den man får genom djupt reflektera över process utan göra anspråk på orsakssamband
  - #clarifyKvaff allt detta var lite otydligt, s . 56
- svenska "vetenskapligt arbete" är bredare, "scientific method" = "hypotetisk-deduktiv ansats"
  - deduktiv - deduktiv logik grundad i teoretisk förståelse (eftersom x och y följer att...)**
    - Hypoteser följer från detta, **som sedan testas**
  - skillnad fr induktiv logik, som bygger på empiriska observationer (pga mina observationer av x, y, ser vi...)
- Utan hypoteser grundad i teoretisk hypoteser blir sannolikheten för bias och feltolkningar under undersökningar av orsakssamband, ex att man tolkar slumpvisa mönster som res av kausala samband, och confirmation bias

Ny teknik och gamla principer

- ai teknologi kan göra många analyser/slutsatser människor kan och kan göra
- Dessa bygger ofta på väldigt stora datamängder --> induktiv logik tar större plats
  - detta ersätter inte scientific method/hypotetisk-deduktiv kunskap, eftersom vi inte kan säga att de upptäcka mönstrena ger ny *vetenskaplig* kunskap innan man har en (teoretisk) förklaring till varför mönstrena uppstår

Kontrafaktisk analys

- etablerat att analytiker som vill testa orsakssamband bör ställa upp och testa hypoteser, men inte sagt vilken data som behövs
  - Grundprincip = slutsatser om detta kräver *kontrafaktisk analys*: Att vår data så långt som möjligt ska möjliggöra analys om hur ett visst utfall hade blivit med en annan historik
- **Krav på data för att möjliggöra analys om orsaksamband:**

- ***Data med variation***
  - ex om undersöker pris - efterfrågan
    - Omöjligt dra slutsatser om pris varit konstant
    - högre variation -> lättare att fånga upp effekter av pris på efterfrågan
    - extra stor variation om finns brus eller felkällor i mätning
      - om lite variation kan brus störa ut mönster
      - vill ha så mycket användbar variation som möjligt, högt signal-brusförhållande
      - mätfel och brus under kontroll -> mindre variation ok
      - Ex kommer osäkerhet i linjär regression genom att man vill dra slutsatser om underliggande population genom ett stickprov
      - standardfel #clarifyKvaff fatatde inte riktigt vad det var
    - kan ex inte dra slutsatser om värden utanför range av variation
- ***Data som möjliggör rekonstruktion av orsakssamband***
  - Detta innebär två tumregler:
    - Experimentella data är bättre än kvasi-experimentella data
      - experiementella data samlas in genom experiment
      - Idealfall: randomiserad kontrollstudie
        - = full tillgång till försöksobjekt. Väljer slumpmässigt ut ett stickprov, behandlar det, får data, och får resultat. Sen jämför resultat med kontrollgruppen (obehandlade gruppen), representerar kontrafaktiska utfallet
      - detta ideal ganska långt i från verkligheten utanför vetenskapliga studier, kan inte få den kontrollen
        - i detta fall får göra kvasi-experimentell ansats: samla in data från verklighet vi inte själva manipulerat och analysera den som att den vore experimentell
          - Kan inte isolera faktorer som är interesserad av, måste också samla data på alla faktorer som påverkar vår beroende variabel
          - kan då rensa bort trendar som irellevanta för studier (ex om pris ökar över tid och det inte är relevant för studien)
            - Om inte gör detta drar man felaktiga slutsatser om orsakssamband
      - mellanting mellan kvasi-experimentell och experimentell: studie som kan uttytja naturligt experiment
        - om system av intresse haft exogen (yttre) påverkan, som ändrat ex variabler vi intresserade av, kommer närmare "allt annat lika" analys man kan få i labb
        - Ex om studerar pris-efterfrågan elpriser: EU-lagändring som höjer elpriser är exogen påverkan, låter oss studera allt annat lika
          - Annan variation som ändrar priset kan eventuellt vara sammankopplad med priset också, mindre validitet
    - Paneldata är bättre än tvärsnittsdata
      - Tvärsnittsdata: Data från många observationsenheter, men bara vid ett tillfälle (tvärsektionellt)
      - Paneldata: Data från många observationsenheter, vid flera tillfällen (i tiden)
      - Paneldata bättre för att
        - Data över tid kan användas för skatta samband
        - Mer variation ger mer precisa skattningar
        - kan fokusera på skillnader över tid hos enskilda observationsenheter, hitta saker/eleminera vissa faktorer som är tidsinvarianta

Triangulering och komplementerande analyser

- Komplementerande delanalyser för en närmare målet att formulera pålitliga orsak-verkan-samband
- Kan göra placeboförsök:
  - Ge randomiserad kontrollgrupp en behandling som antas inte ha någon verkan, se vilka resultat det får (se så att det faktiskt inte har någon verkan)
  - bra exempel lottvinnare s. 62

reliabilitet och validitet för samvariation och kausalitet

- reliabilitet = spridningsmått (varians, standardavvikelse) på punktskattning av beroendemått (kovarians, korrelationskoeffecient, regressionskoeffecient) kan användas
- Validitetet behandlas som två separata frågor: Intern och Extern validitet:
  - Intern validitet
    - finna giltiga samband för den egenskap vi vill studera hos populationen vi studerar
  - Extern validitet
    - Hur representativ av målpopulationen är den populationen (stickprovet) vi har valt?
    - Blir ofta fråga om hur långt resultat kan generaliseras
      - ex kan inte generalisera resultat om sömns vikt för studier till sömns vikt generellt
    - Många studier har problem med detta, samband i testmiljö men inte utanför den (ex i samhällsvetenskap, etc etc)

Angående regressionsanalys.md

#KvaffKap7

Kontrollvariabler

= *Variabler som vi har med i regressionsanalys för att inte dra felaktiga slutsatser om samvariation, men som vi inte är analytiskt interesserade av (?)*
för studera samvariation mellan X och Y börjar man med fundera över vilka kausala länkar som (teoretiskt) kan sammankoppla de (se Studera orsak och samband)

- ex genom formulera ekvation vars parametrar kan skattas med regression
- Kan vara vanligt misstag att i modellen/ekvationen lägger till alla kända variabler + felterm:
  $Y = \alpha + \beta_1 \cdot X_1 + \ldots + \beta_n \cdot X_n + \epsilon$
- Kan då hålla alla variabler bi inte interesserade av konstanta och ändra den vi är interesserad av

Val av variabler i regressionsmodell

- många kontrollvariabler != bättre modell
- Ofta är enklaste möjliga modell den bästa
  - Mer pedagogiskt när uppvisa resultat, lättare att replikera tester
- Kan också bli problem med overfitting om man har för många variabler
  - Sådant som egentligen bara är brus (osystematiska fel) "bakas in" i parametrarna för (kontroll)variablerna
  - modellen blir väldigt specefikt anpassad till just de data den skattats med, sämre lämoad att beskriva underliggande processen
  - överanpassning försämrar möjlighet t prediktion
- Vanligt att experimentera med modellen; ex ta bort icke statistiskt signifikanta variabler
  - Kan göra detta om andra X-variablers relation till Y ej ändras, eller om modellens exmpelvis MSE eller R2-värde ej ändras/blir sämre
- grupper av variabler kan behöva testas tillsammans, eftersom en av de kan ha ingen påverkan på Y medan de tillsammans har påverkan
  - exempel med olika typer av test på s. 67

- På senare tid vanligt med algoritmer som väljer ut variabler i en modell och skattar modellen
  - exx lassomodell, säter parametrar som är nära 0 till 0
    - kan på så vis komma frammåt i analys, eleminera icke signifikanta variabler, speciellt om inte har teoretisk förståelse för vilka som är relevanta för att förklara Y
    - Kan bli väldigt olika modeller beroende på vilka variabler som finns till hand, används därför bäst som komplement

- #clarifyKvaff väldigt svårtolkad paragraf slutet av s. 68

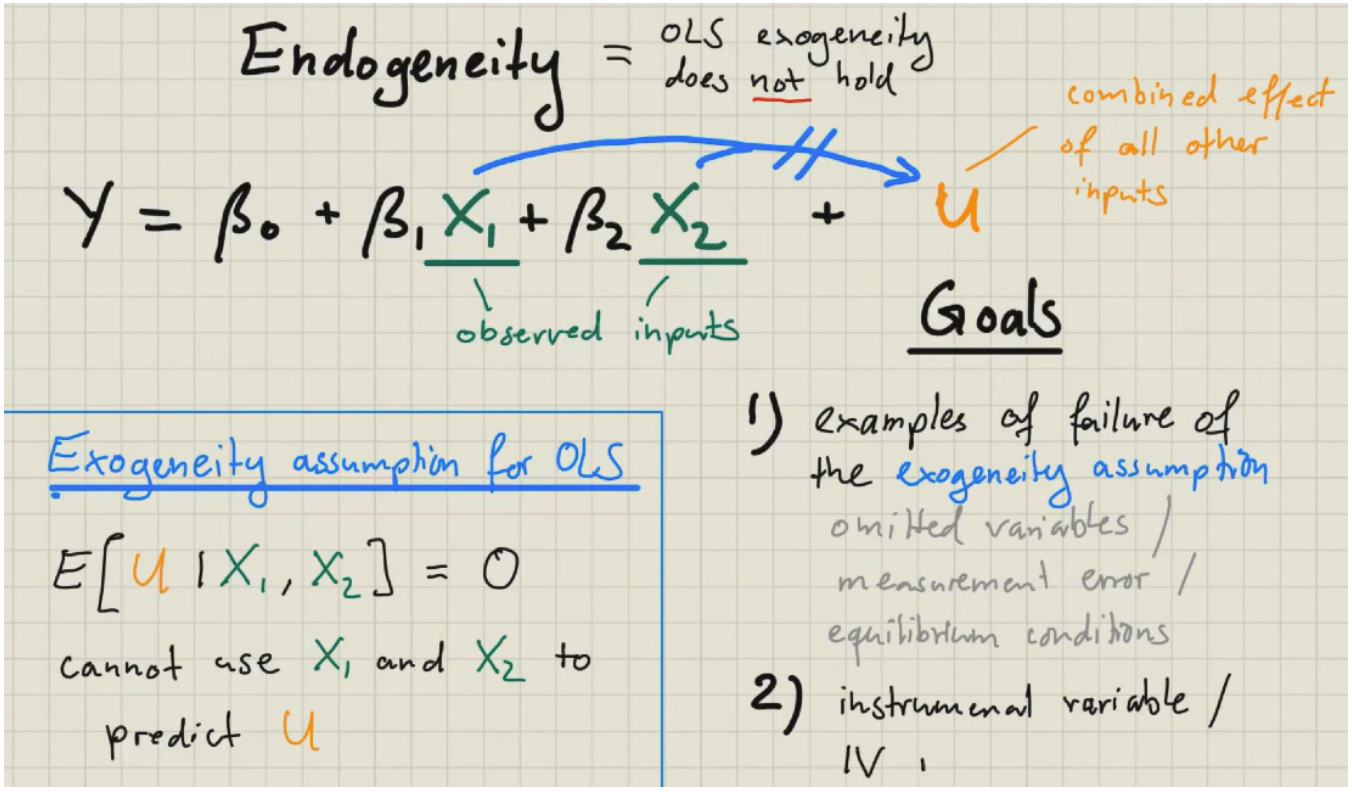
Endogenitetsproblem.md

#KvaffKap7

Om skattar sambandandet X -> Y med linjär regressionsmodell blir sambandet bara väntervärdesriktigt om påverkan är exogen i sammanhanget (att det skett variation utifrån som bara påverkar X?)

- om finns confounding factors, förväxlingsfaktorer, som modellen ej fångar upp blir skattning skev
  - => modellen lider av endogenitetsproblem ouppfångad kontrollvariabel
    - (Kan ockås kanske säga att endogenitetsproblem = ej uppfattad intern påverkan av olika variabler?)
    - svårt avgöra vilka variabler som påverkar varandra, vår modell inte lagom komplicerad för fånga upp det
  - God teoretisk förståelse -> kan åtgärda detta, samla in data om förstådda kontrollvariabler/förväxlingsfaktorer
    - Kan dock vara svårt add åtgärd engogenitetsproblem eftersom:
      - Kanske bristfällig förståelse för problemet eller osäker om fångat alla förväxlingsfaktorer
      - kan finnas omätbara förväxlingsfaktorer
      - X kan påverka Y, ömsesidigt snarare än ensidigt orsakssamband
        - Dessa endogenitetsproblem kallas simultanitet, ex studier tillgång efterfrågan
  - För att åtgärda endogenitetsproblem som inte kan lösas med mer/bättre datauppsamling:
    - Experimentell anstats
    - Använda naturligt experiment
    - Om anser de faktorer som orsakar endogenitetsproblem är tidsinvarianta kan man använda paneldata, många olika mätpunkter över tiden och använd endast tidsvariation för skatta koeficienter
      - vanligt i ex epidimeologi eller statsvetenskap där experimentella ansatser svåra att ådstakomma
      - #clarifyKvaff vad är experimentella anstatser nu igen?
    - Ett viktigt verktyg här är skattningsestimator med fasta effekter (fixed effects estimator)
      - $Y_{it} = \alpha_i + \beta x_{it}$
        - α tidsinvariant från t och unik till varje individ i
          - På så vis fångar den upp/absorberar alla faktorer som är unika till individen (och som inte ingår i x (!?), x fångar istället det vi interesserade av)
        - endast β skattas för hela stickprovet/målpopulationen
        - På detta vis undviker vi endogenitetsproblem som härstammar från tidsinvarianta faktorer som påverkar utfallet vi studerar
          - !!! Men inte andra källor t endogenitet
  - Om vill hantera ömsesidigt beroende variabler eller båda faktorer (X och Y?) påverkas av tidsinvariant oobserverad faktor, måste man ha mer avancerade regressionsmodeller för att undvika endogenitetsproblem i kvasi-experimentella data
    - ex instrumentalvariabelmodeller

Endogenitetsproblem är när någon förklaringsvariabel är korrelerad med felterm, dvs all påverkan av det som vår modell inte har fångat upp



Estimatorer för utfall som ej är normalföreldade.md

#KvaffKap7

Minstakvadratmetoden för regressionsanalys kan vara oträffsäker om datan inte ens är nära att uppfylla antaganden om att vara kontinuerlig och normalfördelad

- För dessa måste man använda andra moderatorer:
  - Tabell s. 75
- Dessa metoder kan ge bättre skattningar, men bra att jämföra med minstakvadratmetoden för regression
  - Vanligt att görea det vid dikotoma beroende variabler
  - När minsta-kvadrat-estimator används för detta kallas det linjär sannolikhetsmodell
    - Lättare att tolka än logit-koeeficienter, ger också ofta träffsäkra medelvärden
    - linear probability model kan dock vara olämplig om vid analys av hur sannolikhet varierar vid exremt höga/LÅGA värden, kan ge sannolikhet som inte $\epsilon\ [0, 1]$.

Från modell till slutsats genom regessionsanalys.md

#KvaffKap7

Vanligt använda regressionsmodell för undersöka styrkan av samvariation, under beaktande av vår teoretiska förståelse för systemet

- viktigt göra avvägning enkehlhet och fullständigheten i modellen

se Angående regressionsanalys

Kausala diagram.md

#KvaffKap7

ofta är analytikers arbete cykel av att modellera och utvärdera

- menar dock finns värde i börja modellering tidigare än så

Kan vara bra att visa orsakksammanband på grafiskt sätt med DAG (directed acyclic graph)

- visa förståelse för samband X o Y baserad på tillgänglig teoretisk förståelse/kunskap

- visa förståelsen av systemet och datagenererande och dess samband
- kan vara utgångspunkt i vilken strukturell form i regressionsmodell man ska använda
- Kan finnas förväxlingsfaktor (confounder) som påverkar både X och Y, om iinte kontrollerar för denna skattas sambandet mellan x och y systematiskt fel
  - Ex kan upplevt samband mellan x o y egentligen bero på att confoundern c gör att både x och y blir höga
- Om S linjärt beroende av X (kolinjär) kommer de samvarierar, en del av variationen i X kommer tillskrivas S, blir skevt
- Medierande variabel är en vilken är mellan beroendet av X och Y
  - om X endast påverkar Y genom M så är det full mediering
  - kan skatta M genom ekvationsystem, med strukturell ekvationsmodell där båda ekvationer skattas samtidigt:
    - $Y = \alpha + \beta_1 X + \beta_2 C + \beta_3 M + \epsilon$
    - $M = q + wX + \epsilon$
- Kan bortse från övriga variabler, om man inte anser att de kan påverka Y och *kanske* kan påverka X och således vara förväxlingsfaktorer

Finns vidare *Modererande Faktorer*

- Variabler/faktorer som påverkar styrkan av förhållandet mellan X och Y
- om = Z kan få följande ekv i regressionsanalys:
  - $Y = \alpha + \beta_1 X + \beta_2 XZ$
- kan också kallas interaktionsterm, påverkar hur X påverkar Y; interagerar med X
- Kan studera moderation genom dela upp datamängd efter variabel man vill undersöka (som man tror är modererande (?))
  - Ex dela in i olika kategorier, Ex olika utbildningsnivåer
  - skatta β_2 i exemplet ovan, visar "styrkan" på moderationen
  - bra exmpel på detta s. 73
  - Nackdel med denna indelning är man förlorar data, färre mätpunkter för varje skattning -> bredare konfidensintervall, större std

Kanske läsa mer om i kvaff.md

#läsmerkvaff

Modeller som bara nämns men inte förklaras i boken

- [ ] ANOVA (se Beskrivning av mätserie)
- [ ] post-hoc test (se Beskrivning av mätserie)
- [ ] bonferroni-korrigering (se Beskrivning av mätserie)
- [ ] roc-kurva, presicions-recall kurva (se Modeller som analytiska verktyg)
- [ ] stödvektormaskin (support vector machine) (se Modeller som analytiska verktyg)
- [ ] random forest (se Modeller som analytiska verktyg)
- [ ] stratifierade undersökningar (se planering av statistiska undersökningar)

- Falsk positiv och negativ och omkringliggande kunskap, grafer etc, se #KvaffKap8
- Mer om vilka matematiska prognosmodeller som finns att använda
- hypotestester för prognosmodeller
- Hur t-test går till
- R2-värde
- falsk positiv och negativ

Prediktion.md

#KvaffKap8

Analytikers uppgift ofta att göra förutsägningar om framtid snarare än endast blottlägga underliggande orsak-verkan samband

- risk göra felaktiga prognoser om bristfällig kunskap om kausala samband

Validering

- För god prediktion behöver en modell som ligger nära verkliga datagenererande processen
  - I process skapa denna modell hjälper förståelse för den processen
  - Kan också ta an prediktionsproblem genom helt enkelt träna ML algoritm
    - Under antagande att systemet inte förändras för mycket över tid, att träningsdatan är representativ för framtiden, kan man uttnyttja modell utan djupare insikt i hur den fungerar eller vad den säger om underliggande process
      - processen blir då "Svart låda", vi bryr oss bara om indata utdata
      - Om har ett ickeförändrande system finns här fördel att man tydligt kan mäta precision hos maskininlärningsalgoritm: Hur bra kan den göra precisa prediktioner på dataset det inte tränats på? = out of sample prediction
        - kan dela upp dataset i ex 70-30 procent, 70% för träna 30% för utvärdera
        - Validering enligt samma principer kan göra för optimera modellval genom att kolla deras presicion på samma sätt
          - kan undvika dataförlust kan man använda korsvalidering: dela upp dataset i olika delar som sätts ihop t flera träninigsset ch testset

Skillnader mellan deskriptiv, prediktiv och preskriptiv analys

- säger ofta komplexitet ökar fr deskriptiv prediktiv preskriptiv
- Ofta analytikers mål att säga något om framtiden
  - inte omöjligt att göra detta med deskriptiv analys: Kan hitta/beskriva mönster och argumentera varför de kommer upprepa sig i framtiden
- I andra frågeställningar krävs kartläggning av kausala samband för koppla handlingsalternativ till utfall
  - Detta kallas ibland (något slarvigt) preskriptiv analys
  - Kan dock också vara grund för prediktion, om modell har extern validitet och fångar upp ansenlig del av datagenererande processen
  - En skillnad när man försöker göra en enskild förutsägelse/prognos är att man kan göra starkare antaganden om stabilitet i datagenererande processen och ist för validitet maximera reliabilitet
    - Detta fokus reliabilitet kan ofta leda till enkla modeller, men de kan ofta ha högre precision av prediktion av framtiden jämfört med komplexa modeller som representerar underliggande kausala samband
  - Prognosmodeller är dock ofta väldigt komplicerade, kan söka efter optimal metod med algoritmer

Prediktionsförmåga

- Kan mätas i MSE, eller standardavvikelse (lätttolkad eftersom har samma enhet som storheten man undersöker)
- R2-måttet säger hur stor andel av den totala variationen i beroende variabeln som förklaras av modellen
  - Är bra för jämförelse/utvärdering mellan prediktionsmodeller
- Om dikotoma eller kategoriska klasificieringar (utifrån andra mätvärden från en datapunkt) är det man söker kan man mäta prediktionsförmåga som andel korrekta klassificieringar
- bör göra avvägningar mellan positiva och negativa klassificieringar: Om värre med falsk negativ/positiv bör man favoura den andra
- Kan använda ROC-kurva för att visa dessa klassificieringar, y axeln visar andel positiva fall som klassas som positiva, x-axeln andel negativa fall som klassas som positiva
- Kan istället ha presicion-recall-kurva: Horisontell axel har mått för hur stor andel av alla observationer som klassas som positiva som är korrekt klassificierade #clarifyKvaff
- Dessa kurvor kan simplifieras som AUC (area under the curve) fr 0 t 1, värde nära 1 visar att nästan alla mätpunkter som klassas som positiva verkligen är positiva

Prognosmodeller

- Kan använda många olika verktyg när man ska göra prognoser av framtiden

- regressionsmodeller m mkm ofta dåliga för tidsserier eftersom framtida utfall ofta beror på tidigare utfall
  - Dessa har så kallad autokorrelation, de beroendena bryter mot "antagandet om oberoende feltermer i gauss-markov-villkoren"
  - autokorrelation – utfall påverkas av ett eller flera tidigare utfall
  - riskerar underskatta standardfelen
  - måste modellera autokorrelation med prognosmodeller som ARCH och GARCH
    - Kan också använda enklare modeller
      - Ex exponential smoothing: Nästa värde i tidsserie är linjärinterpolation mellan prognosvärdet L_{t-1} och senast uppmätt värde Y_t:
        - $L_t = \alpha Y_t + (1-\alpha)L_{t-1}$
        - eftersom L_{t-1} beräknas från tidigare värden är vår modell ett viktat medelvärde av mätserien, mer vikt senare delar i mätserien om α är högt
  - Holt-winters metod är mer komplicerat ekv-syst som bygger på att säsongsvariationer (eller trender) har återkommande påverkan (se s. 87 i boken)
- Dessa modeller är mellanting mellan rent prediktiva modeller och statiska modeller för testa hypoteser om kausala samband
- Utgår dock ofta från teoretisk förståelse för processen: fördel kan göra ändringar i funktionen/modellen
  - Kan ex justera ner styrkan för parametrar som anger hur mycket försäljning som kommer ske om vi vet att det är lågkonjunktur
  -

Data collection and storage.md

#DataProcessing

en del av the data science process

två typer data:

- company data
  - collected och använd av företag för göra så de kan ta data-driven desicions
- open data
  - Alla har tillgång

vi genererar massa data genom att använda services

- företag som ger service samlar denna data i ovannämnda syfte

Company data

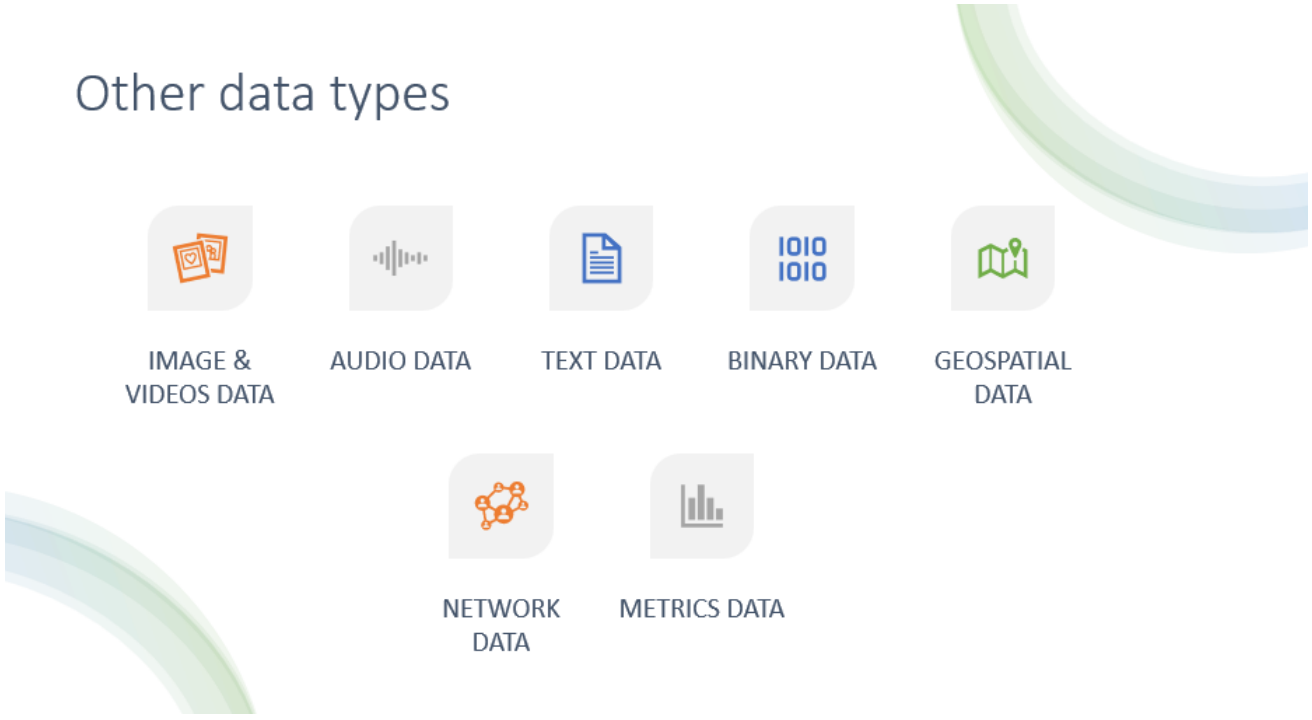
- kan vara
  - web events
  - survey data
  - customer data
  - logistics data
  - financial desicions

Open data

- kan vara
  - APIs
  - Public records
    - från olika institut, ex universitet
  - Finns internationella / nationella organisationer som ger data

DATA TYPES

- ***Två genrerella typer av data:***
  - Numerisk (quantitative)
    - Sånt som kan bli mätt, expressed using numbers
    - Kan vara diskret eller kontinuerlig
      - bara integer values om diskret?
  - Categorical (qualitative)
    - deskriptiv, conceptuell
    - Inte alltid en underliggande storhet som existerar?
    - Kan vara nominal eller ordinal
      - Nominal har ingen ordning eller numerisk värde
        - Hårfärg, religion
      - Ordinal kan ordnas
        - Ex betyg, education level



  - Dessa är essentiellt blandningar av kvalitativ/kvantitativ data

Data storing and retreiving

- hur ska man lagra och ta ut insamlade datan?
  - **Location**
    - Kan använda single/cluster av datorer/servers beroende på mängden data
    - Cloud storage
      - azure, AWS, google cloud
  - **Data type**
    - Olika data types lagras på olika sätt
      - *Ostrukturerad data* = email text image etc
        - Non relational databases, NoSQL
        - Länkade tables, joins, queries
      - *Strukturerad data* = Information som kan lagras i tabeller, ex xlsx eller csv-format
        - Relational databases, SQL
        - Mer komplex flexibel manner av storing, kan accessa snabbt o effektivt

Data preparation.md

#DataProcessing

Mycket tid läggs på detta

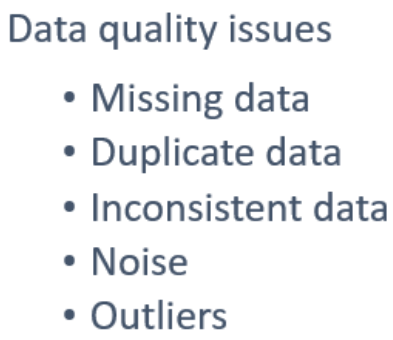
- essentiellt för få high quality results från data analysis and modelling

Varför data preparation?

- Förbättrar resultat, ta bort errors/inconsitencies, ökar reliabilitet
- Ökar effketivitet: minskar tid/computational resources det tar för data anlaysis/modelling
- Lättare organisera cleaned och organiserad data
- Minskar risk m använda felaktig data

Vad är data preparation?

- transformera data till något format suitable för analy/rapporter
- Steg:
    1. ***Data cleaning***
        - Ta bort inacuracies, inconsitencies, irrelevant information, öka usability/reliabilitet



        - Dessa saker försämrar datans o analysens kvalite
        - *Vanliga fel:*
            - kan ha missing data om folk inte vill uppge saker/ fältet appliar inte till dem/errors i insamlingen
                - Kan hantera genom droppa missing values, keep missing values, inputta vanliga/rimliga values där de saknas
            - dupllicate data
                - om flera kopior/när kopior av samma objekt finns
                - delete/merge records
            - Inconsistent/invalid data
                - Om omöjliga värden under någon kolumn
                - replace med rimlig value eller använd extern data source för få correct value
            - Noise
                - Allt som kan förstöra data
                - filtrera eller ta bort frequent noise components
                - ta bort del av data
            - Outliers
                - Mycket annorlunda värden från resten av datasettet
                - Kan vara failure/real anomaly
                - Undersök ytterligare innan tar bort, men endast om de är fokus av analysen
    2. ***Data formatting***
        - Ändra struktur för göra mer suitablre för analys, reporting, eller andra applications
        - ***Feature selection***
            - Välja set av features som är lämpliga för analys
            - Metoder:
                - adding features - data enrichment
                    - derived/extern data source
                - removing features - reduction

        - Väldigt korrelerad, saknade värden, irrelevanta värden
      - combining features
        - Om två features inte ger information ensamma kombinerar man de, ex BMI
      - Recoding features
        - Kan göra continous till kategorisk eller liknande
      - breaking up features
        - Göra om ex address till gata, län
    - Välj så få features som möjligt som relevanta för problemet, simplare analys
  - ***Feature transformation***
    - Ändra format av data på något sätt för att minska noise/variabilitet eller göra data enklare att analysera
    - Mappa om set av värden till nytt set för att göra representation av data enklare/mer suitable
    - ***Scaling***
      - Normalisera/standardisera värden som varierar väldigt mycket
      - kan göra ML enklare
      - Normalisering
        - shifta/scala så hamnar i rangen $[0, 1]$
      - Standardisering
        - Ändra data så att mean $\mu = 0$ och std $\sigma = 1$
      - Använd stanadrinseng om vet normalfördelning råder (men ingen set in stone regel)
    - ***Aggregation***
      - Summera/averega flera av samma features
        - Kan minska noise och ge tydligare representation
        -

Exploration & visualisation.md

#DataProcessing
del i data science process

- beräkna beskrivande mått för mätserien, mean median std etc
- Kör statistiska modeller
  - Linreg, logistic reg
- Plots
  - Scatter, bar, histogram

visualiseringsredskap: *(kanske inte superviktigt?)*
•Microsoft Excel and Google Sheets

Microsoft Excel and Google Sheets are spreadsheet software with built-in charting capabilities. They are user-friendly and suitable for creating basic charts and graphs.

•Tableau

Tableau is a powerful and widely used data visualization tool that offers a range of interactive and customizable charts and dashboards.

It supports various data sources and allows you to create complex visualizations without coding.

•Power BI

Microsoft Power BI is a business intelligence tool that enables you to create interactive reports and dashboards. It offers a wide range of data visualization options.

•Python Libraries

Python libraries like Matplotlib, Seaborn, Plotly, and Bokeh are popular among data scientists and analysts for creating customized visualizations. They provide fine-grained control over chart appearance and behavior.

•R and ggplot2

R is a programming language used for statistical analysis and data visualization.

The ggplot2 package is widely used for creating elegant and customizable data visualizations.

•R Shiny

R Shiny is an open-source R-based web application that allows you to create interactive, web visualizations and dashboards within the R environment.

•D3.js

D3.js is a JavaScript library for creating interactive and custom data visualizations for the web.

It's highly flexible and allows to create unique and dynamic visualizations.

data science.md

#DataProcessing

- iNterdisciplinary field som andvänder statistical and computational methods to extract insights and knowledge from data
  - Innehåller matte, programmering, mechine learning, stats, etc
- Använde allt detta för att göra informed desicions, få insigths o knowledge från data

data science process

- Iterativ, flera stages refinement improvement
- Hur process går til beror på data tjillgänglig, verktyg
- ***Generellt dock 4 steg:***
  1. Data collection and storage
     - Samla in data från olika källor och lagra på säkert och accesible sätt
  2. Data preparation
     - Har nu rå data, transformera och cleana så blir suitable för analyssis
  3. Exploration & visualisation
     - Hitta mönster o relationer genom ex visualisera, statistisk analys
  4. Discovery & Prediction
     - Använda resultat av förra steg för beskriva data/göra predictions om framtiden eller trender. Involverar analytiska metoder, ex statistiska, simulationer, experiment, etc

Data processing

- Subset av data science som fokuserar på tre första stegen
-

planering av statistiska undersökningar.md

Allmänt om planering

- planering viktig eftersom man i planeringsstadiet kan påverka valet av teoretisk modell för data som kommer fram
- kanvid god planering välja modell som möjiggör klara slutsatser
- mer än bara storlek som påverkar undersöknings kvalitet
- lägg upp plan som är så bra som möjlig med tillgängliga resurser
  - enkla övervägningar kan lösa sådana frågor
- behöver både statistisker och folk som är insatta i undersökta processen -> lagarbete gör gott
- **List med ungefärliga etapper för planeringsarbete:**

Minneslista för planering av statistiska undersökningar

Etapp 1. Inledande praktiska moment

Inventering av redan tillgänglig information
Inventering av resurser (tid, personal, ekonomi etc)
Formulering av praktiskt problem
Formulering av krav på lösningen
Val av undersökningstyp

Etapp 2. Teoretiska moment

Konstruktion av slumpmodell
Formulering av teoretiskt problem
Formulering av krav på lösningen
Utarbetande av undersökningsplan
Angivande av statistisk analysmetod

Etapp 3. Avslutande praktiska moment

Uppgörande av skriftlig detaljplan för
datainsamling
bearbetning
statistisk analys
presentation

-
- Finns betydande skillnader i planeringsarbete mellan jämförande och icke-jämförande undersökningar

Icke-jämförande undersökningar

- önskar information om någon okänd storhet inom någon disciplin
- måste ta ett stickprov med n element från målpopulationen
  - om fysiska element kallas det *enkelt urval*
- **För att kunna ställa upp användbar modell måste några krav ställas på stickprovstagningen**:
  - 1. ***måste vara slumpmässig***
    - om ändlig population som går att numrera gör slumptalsteknik
    - annars får göra på något annat sätt
    - ibland kanske inte går att ta slumpmässigt element - går då att använda slumpmodell?
      - avgörande för om kan betraka som slumpmässigt dragna är om experiment är reproducerbart
    - ofta svårt att avgöra vilken population observationerna/stickprovet kommer från (ex medecinstudier på ett enskilt sjukhus). Måste klargöra vad man menar med att datamängd är slumpmässigt stickprov från en population
    - om inte kan garantera slumpmässighet krävs stor erfarenhet för att avgöra vilka slutsatser som är giltiga att göra
  - 2. ***störande faktorer får inte förekomma***
    - inga systematiska fel får finnas
    - ex felkalibrerat instrument ger felaktigt resultat oavsett antal mätningar
      - -> är en *störande faktor*
    - bortfall (elementen i populationen som inte upträder i stickprovet) kan påverka ndersökning
      - Små bortfal % kan ibland vara devestating
    - egentligen ouppnåeligt att inte ha några störande faktorer, snarare ska de inte påverka slutsatserna

- **Flera stickprov**
- Ibland samlar man stickprov från k olika populationer
- mer omfattade planering eftersom måste välja hur insamlingen ska fördelas över olika stickprov
- Kan göra ***Stratifierad undersökning***
  - = dela in populationer i k delpopulationer efter någon förmåga, ta stickprov från varje
  - om väljer storlek på stickproven lämpligt kan ofta lättare få önskad information om hela popualtionen än om tar ett enda stickprov från hela populationen
- ***matematisk beskrivning av hur göra stratifierade undersökningar och varför de är bra kring sidan 382*** #läsmerkvaff

Jämförande undersökningar

- undersöka olika "behandlingar"
- två typer: *jämförande experimentell undersökning*: kontroll över hur behandlningarna fördelas på de i undersökningen ingående experimenten, om inte är det *jämförande icke-experimentell undersökning*
  - förstnämnda ger säkrare slutsatser, kan eleminera störande faktorer
- **Funnständigt randomiserat experiment**
  - behandlingarna fördelas slumpmässigt på ingående elementen
  - -> Jämförelsen påverkas inte systematiskt av störande faktorer
  - när planerar för längd på konfidensintervall genom antal observationer kan man ex optimera för kostnad
- **Randomiserat blockexperiment**
- om stora skillnader mellan elementen, kan dela in de parvis efter vilka objekt som liknar varandra, och sedan ge varje par antingen A- eller B-behandlingen
- -> Kan jämföra skillnader inom block utan att skillnader mellan blocken påvekar
  - kan sedan använda stickprov i par (normalfördelningsantagande)
  -

statistiska undersökningar.md

statistisk undersökning har vanligtvis fyra delar:

1. Planering - alla förberedelser
2. datainsamling - ex läsa mätinstrument, intervjuer
3. bearbetning - olika former. i enkla fall ex diagram/tabell, annars mer komplicerade mått
4. presentation - ex graafisk framställning, sammanfatta resultat/prediktioner/rekomendationer

Statistiska undersökningar förekommer inom nästan alla vetenskaper

- även onom ex samhäll-ämnen, ekonmi, handel, industri

Population och element

- Vid undersökning studerar man population
- population avser mängd element,
- ett element abstrakteras ofta till en viss egenskap hos ett objekt
- en population abstrakteras då till en mängd data eller observartioner som man studerar
- måste noggrant definiera vad element är och vilka element som ingår i populationen
  - Om inte gör det risk för feltolkningar, felaktiga slutsatser o rekomendationer

Exempel 9.1 *Läkemedelskontroll*
En statlig myndighet vill undersöka om halten av det verksamma ämnet i ett visst läkemedel är den avsedda. Som element kan man välja t.ex. den förpackning som läkemedlet tillhandahålls i. Som population kan man välja t.ex. alla förpackningar tillverkade i Sverige under ett år eller alla förpackningar tillhandahållna på apoteken under ett år. I det förra fallet bör undersökningen utföras hos tillverkarna, i det senare fallet på apoteken. □

-

Variabler och skalor

- element ser olika ut beroende på vad man studerar
- element kan vara tal eller en kategori eller mycket mer subjektiv
- element kan vara talpar/vektorer
- element kan ha diskret variation : de kan bara anta ändligt eller uppräknerligt oändligt många states
    - respektive kontinuerlig variation
- Skalor:
    - nominalskala : ingen ordning, ex ögonfärg
    - ordinalskala : ordning, ex betyg
        - fortfarande inte alltid meningsfullt med differenser, även om omvandlar till tal (ex smiling faces skalan)

Ändlig/oändlig population

- ändlig antal element: ex alla träd i en skog
- oändlig: Alla tänkbara mätningar av fysikalisk storhet, alla 20g prov av 1 ton järnmalm

Totalundersökning/stickprovsundersökning

- Om väldigt allvarliga konsekvenser kan föredra totalundersökning
    - i de studeras hela populationen
- Kan dock vara väldigt kostsamma, opraktiska, tidskrävande
- totalundersökning kräver ej statistikteori
- *stickprovsundersökning* -> del av populationen undersöks
    - stor praktisk användning
    - Ibland kan även få observationer ge väldigt mycket information
    - klokt avvägd stickprovsundersökning kan vara mer reliable än totalundersökning

Jämförande och icke-jämförande undersökningar

- jämför man populationer i något avseende?
- bara interesserad av en population?
- ex: A snabbare än B? vs hur snabb är A?

Hur skatta parametrar/dra slutsatser?

- konfidensintervall, hypotesprövning, punktskattning

Huvudproblem inom statistiken

- parametern $\theta \epsilon \Omega_{\Theta}$ har ofta någon motsvarighet i sinnevärlden
- Tänk på vad som är modell och vad som är verklighet

Ett schema för problemlösning genom dataanalys.md

TRE FRÅGOR FÖR FÖRDJUPAD ANALYS (BORTOM DESKRIPTION):
• Vilket är det fokala samband som jag vill undersöka?
(validitet gentemot den analysfråga jag vill besvara)
• Vilka andra samband är viktiga att ta hänsyn till?
(modellera ditt problem)
• Hur stämmer detta med data?
(testa / skatta / kalibrera modellen)

Vilka variabler bör ingå i en multivariat regressionsmodell?

- Grundprincip: variabler som kausalt påverkar den beroende variabeln
- Men: se upp med 'overfitting' i alltför komplicerade modeller
  (LASSO är ett ML-verktyg som reducerar antalet variabler algoritmiskt - användbart om svag teoretisk grund att stå på för modellering och väldigt många tänkbara förklaringsvariabler (står till hands)
- Framför allt: se till att inkludera förväxlingsfaktorer (confounders) som påverkar den beroende variabeln (Y) OCH är korrelerade med den oberoende variabel (x) som står i fokus
- Om förväxlingsfaktorer inte kan observeras direkt – finns det en relevant proxyvariabel som går att använda för att fånga den faktorn (ex: IQ-mätning)?
- *Medierande samband* bör modelleras om den medierande variabeln är av eget intresse
- *Modererande samband* är ofta viktiga komponenter i en modell
- DAG – directed acyclical graphs – som ett verktyg för att
  - åskådliggöra DGP
  - identifiera 'colliders' / gemensamma effekter som inte bör vara med i regressionsmodellen

KTH
TENSKAP
CH KONST

Åter till vårt exempel

Nu är det dags att testa mönster om könsskillnader i attityder, karriärval och entreprenörskap i en regressionsanalys

- Regressionsanalysen är **mer pålitlig** än beskrivande analys och korstabuleringar om det finns viktiga förväxlingsvariabler som vi kan mäta och inkludera i modellen
- … men regressionsanalysen kan också vara **mindre pålitlig** än enklare analyser om det finns viktiga förväxlingsvariabler som vi **inte** kan mäta eller inkludera i modellen
- I avsaknad av väl utvecklad teori som gör att vi kan uttala oss om ovanstående med säkerhet (med andra ord: teori som säger hur DGP:n verkligen ser ut) är vi allra säkrast ute om vi får liknande resultat från såväl enklare korstabuleringar som från regressionsanalys

Vad är korstabulering? Korstabulering (korstabell) är **ett användbart analysverktyg som ofta används när man vill jämföra resultat för en eller flera variabler med resultaten för en annan variabel**. Det används med data på en nominell skala där variabler namnges eller etiketteras utan någon särskild ordning.

Bias och skevheter

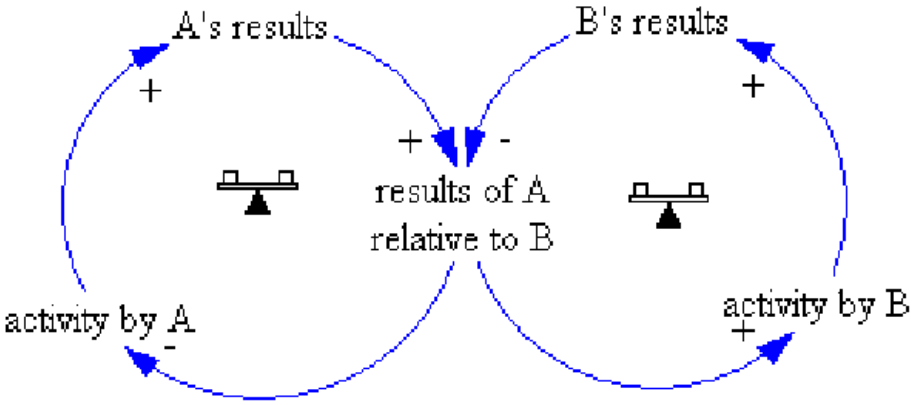
(m. a.o.: icke-väntevärdesriktiga skattningar)
Validitetsproblem
• Bekräftelsebias (confirmation bias)
• Systematiska mätfel
slumpmässiga mätfel påverkar reliabilitet och precision – men inte validitet
• Urvalsfel (sample selection bias)
Endogenitetsbias
• En statistisk definition: förklaringsvariabeln samvarierar med feltermen
• Konskvens av endogenitet på statistikspråk: "skattningar blir inte väntevärdesriktiga p.g.a. modellfel"
Två viktiga former / orsaker / problem:
• Simultanitet - ömsesidigt samband (simultaneity)
• Utelämnade viktiga variabler (omitted variable bias)

Från statistisk analys till systemförståelse

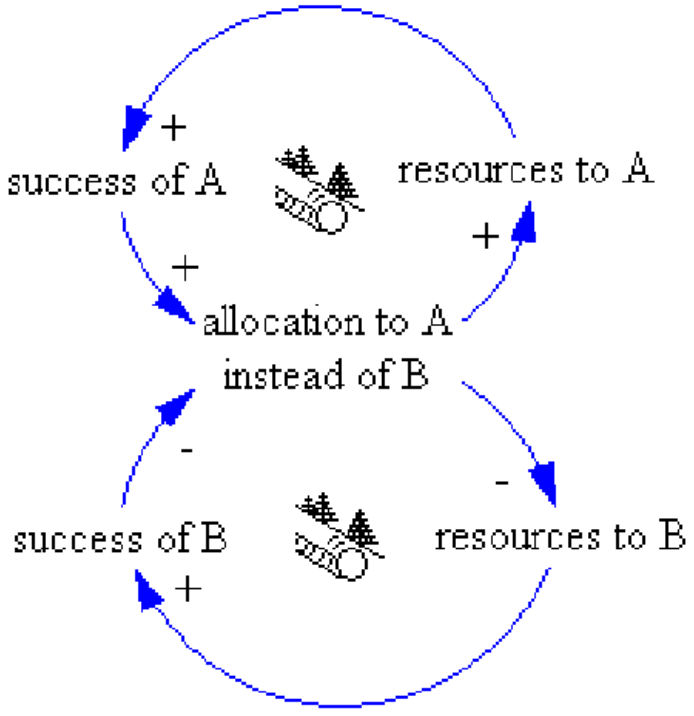
Statistisk inferensanalys
• Utveckla modell som representerar hypoteser om kausala samband
• Skatta / träna / kalibrera modellens parametrar på bas av data
• Typisk målsättning: förstå det datagenererande systemet
Analys med hjälp av simulering
(reglerteknik, styrteori, men även inom ekonomisk analys och prognos/scenario-arbete)
• Utveckla modell som representerar kända kausala samband
• Simulera utfall
• Typisk målsättning: undersöka möjliga framtider

typer av samband

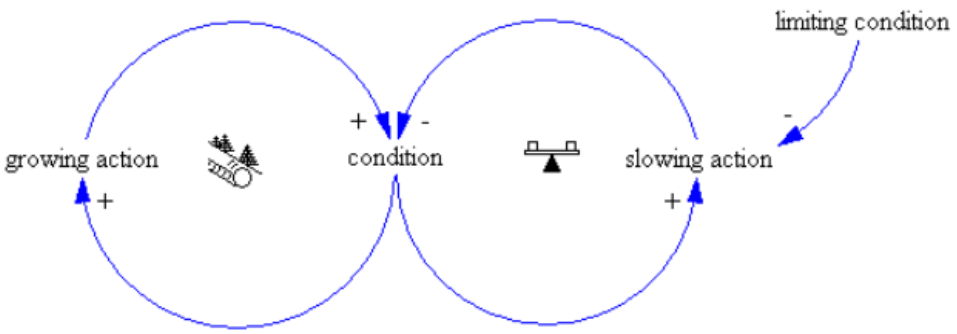
Eskalation



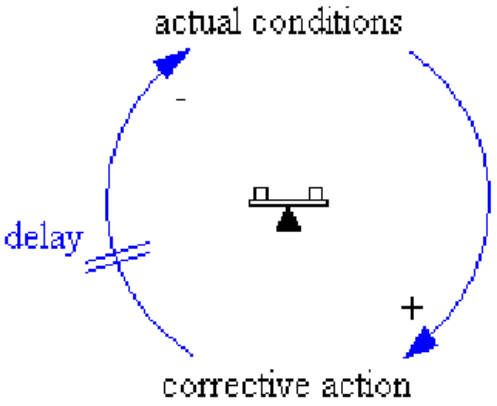
MatteusEffekt



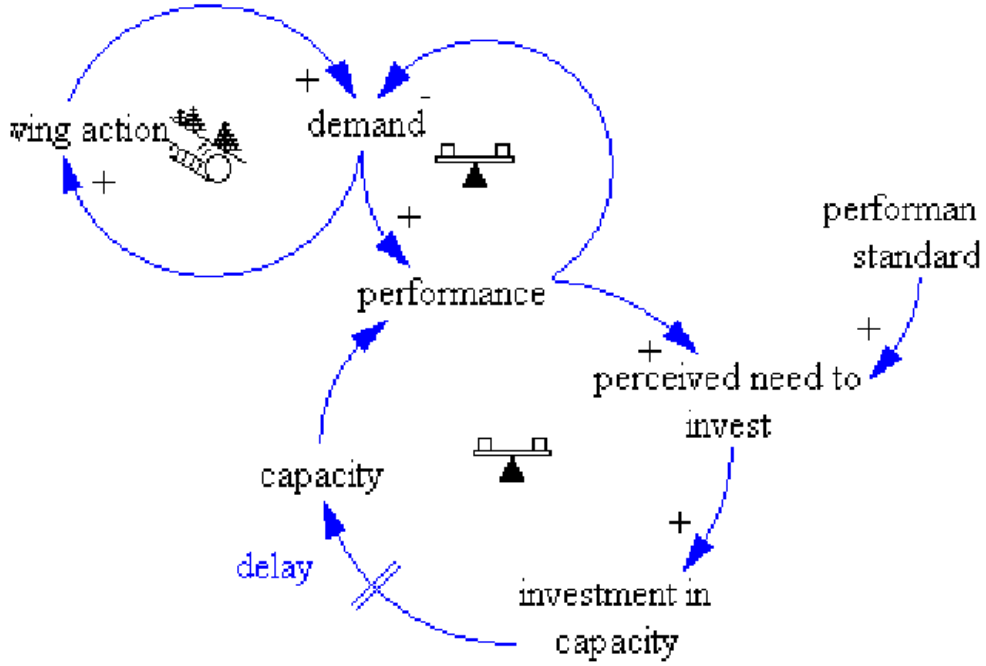
Limits to growth



Balancing process with delay

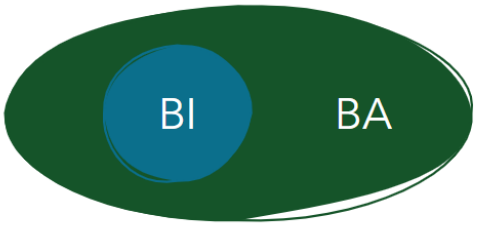


Growth and underinvestment



business intelligence vs business analytics.md

Skillnaden mellan Business Intelligence och Business Analytics



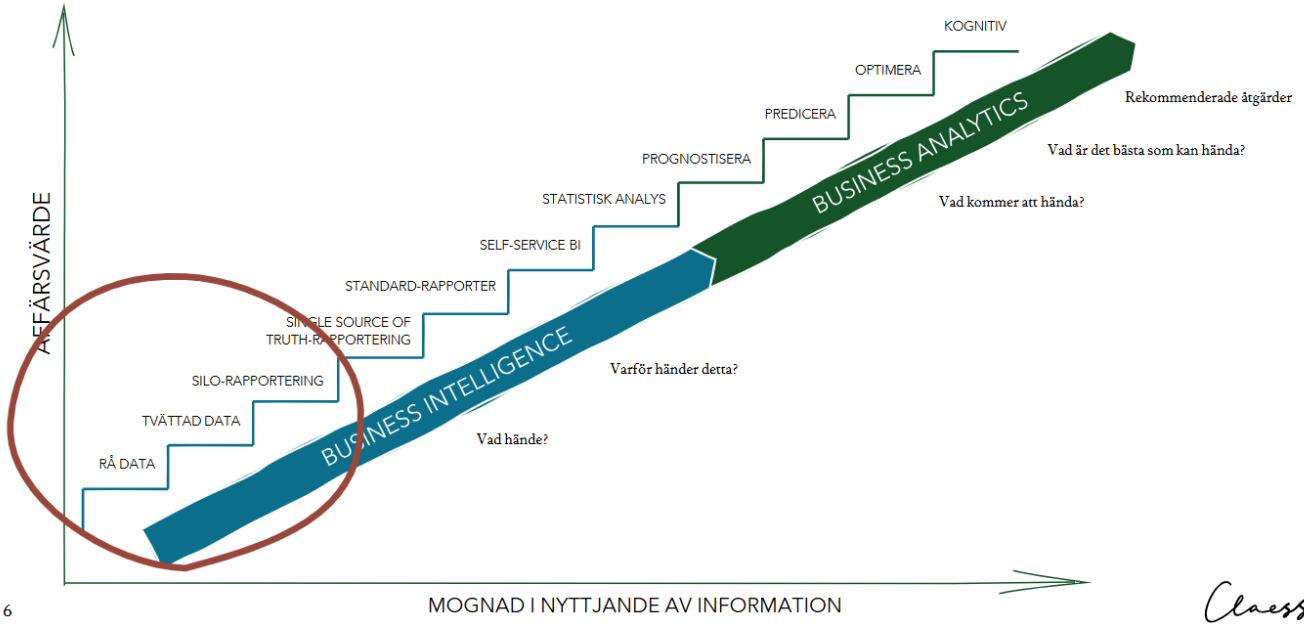
BUSINESS INTELLIGENCE
Business intelligence handlar om att analysera historisk data för att förstå vad som har hänt i verksamheten.

Det inkluderar skapande av rapporter, dashboards och visualiseringar för att stödja beslut baserade på tidigare prestationer.

BUSINESS ANALYTICS
Business analytics omfattar också analys av historisk data men går längre genom att använda avancerade metoder som prediktiv och preskriptiv analys.

Målet är att identifiera mönster, förutse framtida trender och rekommendera åtgärder för att proaktivt förbättra verksamheten.

Analysmognadstrappa – Var är vi och var skulle vi kunna vara?



Övrigt om prognoser.md

Viktiga "lagar" för prognoser

1. Prognoser är alltid fel.
   Bör därför inkludera både förväntat värde och en felskattning.

2. Långsiktiga prognoser är alltid mindre tillförlitliga än kortsiktiga (prognoshorisont är viktig)

3. Aggregerade prognoser är, relativt sett, mer tillförlitliga. Möjligt att aggregera på t.e.x.
   - Tid
   - Produkter/produktgrupper
   - Marknader

Sex faser för prognoser

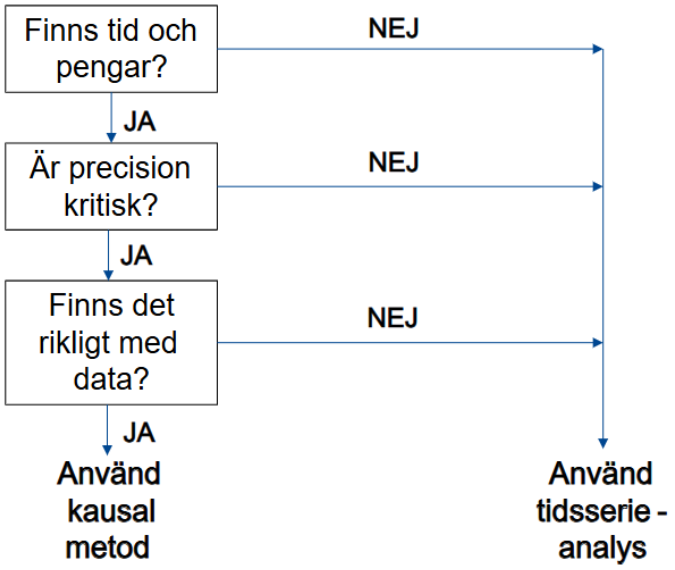
1. Tvätta /rensa data
2. Visualisera data
3. Metodval och Modellbygge
   4. Kontrollera modellen mot data
4. Generera prognos
5. Kontrollera prognos mot utfall

Metoder för prognosticering

- Kvalitativ: baseras på omdöme och åsikt.
- Tidsserieanalys: baseras på historisk efterfrågan
  - Statisk
  - Adaptiv
- Kausal: Använd kända korrelationer mellan efterfrågan och andra faktorer för att ta fram en prognos.
- Simulering
  - Imitera konsumentbeteende och val som leder till ökad efterfrågan.
  - Kan kombinera både kausala metoder och tidsserieanalys
  - M.m

Mer avancerade prognoser använder flera metoder

Val av prognosmetod - tumregel



Källa: Numbers guide

Alla databaserade prognoser bygger på att vi identifierar historiska samband som vi tror kommer att hända igen .

• Tidsse rie analys använde r t id som obe roe nde variabe l. Korre lat ion snarare än kausalit e t . Me n väldigt användbart !

• Många sake r korre le rar m e d t ide n m e n få sake r har e t t kausalt sam band.

• Som m aräpple n falle r från t räde n varje se pt e m be r. Diskut e ra kausalit e t och korre lat ion.

Prognoser med tidsserieanalys

- Statisk (metod)
  - Linjär regression med trend och säsongsvariationer (modell)
- Adaptiv (metod)
  - Glidande medelvärde, med eller utan viktning (modell)
  - Simple exponential smoothing (modell)
  - Holt's modell (**med trend**) (modell)
  - Winter's modell (**med trend och säsongsvariation**) (modell)

Statiska prognoser viktar alla datapunkter jämnt
medan adaptiva prognoser lägger mer vikt på
senare datapunkter.

Problem med överanpassning

- Modell passar data, men genererar en värdelös prognos
  - Vart är modellen på väg när den "lämnar" data?
- Måste finnas en hypotes för modellens utseende.
- Tumregel: Använd enkla modeller om det inte finns starka skäl att välja en mer komplex modell.