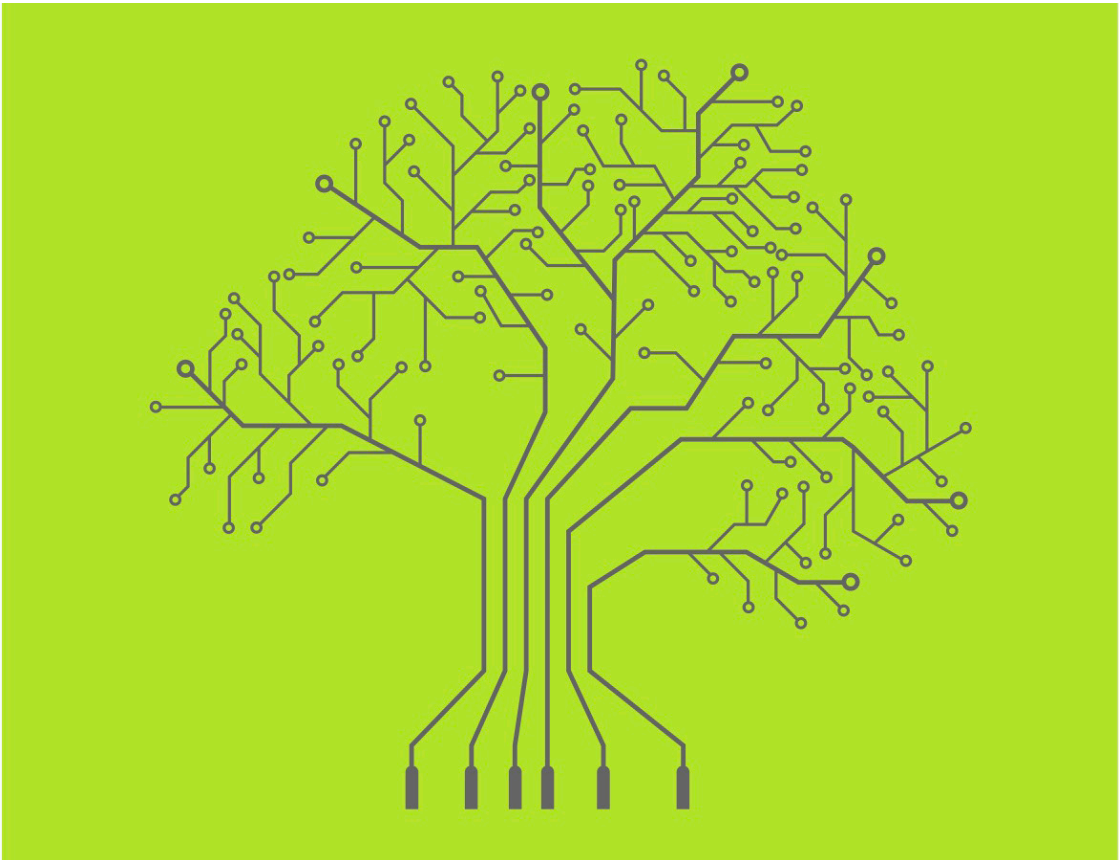
GovTech Challenges
rapportserie, del 4

Mittuniversitetet
FORUM FÖR DIGITALISERING

Datakvalitet – Bra data enligt vem? Beroende av vilka?

Eric Andersson



Inledning

Data är överallt, och i dagens samhälle krävs data för allt. Vi ska förbättra våra processer, vi ska rättfärdiga våra beslut, och vi ska upptäcka problem – allt med hjälp av data. Men för att detta löfte ska kunna uppfyllas så måste kvalitén hos data, eller mer allmänt hos vår information, vara god.

Vad händer om vi inte har god data- och informationskvalitet? Det finns många exempel där problem uppstått på grund av data och informationskvalitet. Exempelvis:

- 1988 – Krigsfartyget USS Vincennes skjuter på grund av feltolkad information och felanpassade data ner Iran Air Flight 655, ett civilt passagerarplan [1].
- 1999 – NASAs Mars-sond kraschar för att navigationssystemet förväntade sig metersystemet, och deras underleverantör använde sig av brittiska måttenheter [7].
- 2020 – Lågkvalitativa data förändrade vissa länders svar på COVID-19 genom att ge missvisande bilder, vilket resulterade i mer spridning av viruset [5].
- 2024 – Förenta nationerna presenterar sin rapport för de största hoten mot mänskligheten i nutid och snar framtid. Felaktig information och desinformation läggs fram med andra hot som klimatkatastrof och miljöförstöring högst på listan, över även resursbrist och ökad ojämlikhet [6].
- 2025 – Felanpassade data skapade överskattade siffror på brittisk inflation, vilket ledde till felaktiga beslut och förlust av offentlighetens tilltro för institutionen [4].

På en mer allmän nivå så påverkar data- och informationskvalitet verksamheter/organisationer. Det finns individuella rapporter som visar på mångmiljardförluster för att informationen organisationer har är felaktig och måste korrigeras, eller att fel beslut tagits: ett konkret exempel är hur IBM uppskattade USA:s årliga förlust på grund av dåliga data till 3100 miljarder dollar [2,3]. Exemplen målar en tydlig bild av konsekvenserna av låg datakvalitet. Men, trots detta, finns det inga enkla lösningar. Men vad menar vi när vi säger data- och informationskvalitet, mer specifikt? De är snarlika, men bör separeras för tydlighetens skull och ingen av dem kan uteslutas. Vi börjar med datakvalitet.

Vad är ”datakvalitet”?

Figur 1. Den output du får om du ber AI illustrera datakvalitet med en bild. Källa: OpenAI DALL-E 3

Figur 1 till vänster är en illustration över vad en språkmodell antar att datakvalitet är. Som man själv kanske hade gissat, så rör det ofta siffror och statistik - men det behöver inte göra det. I grunden så rör datakvalitet huruvida vår data speglar det fenomen det försöker avbilda väl. Vi skiljer därför på detta och den externa användningen av data, vilket är närmre besläktat med informationskvalitet.

Det går att studera denna spegling. Några av de vanligaste mätpunkterna att ha i åtanke, eller ”dimensionerna” som de oftast kallas, är [8,9]:

- Är data korrekt? (kallat Accuracy eller Correctness)
- Har vi all data som vi bör, eller saknas delar? (Completeness)
- Är datavärdet uppdaterat eller utdaterat? (Currency)
- Håller olika data med varandra eller är de i konflikt? (Consistency)
- Över tid, hur väl går det att förlita sig på att punkterna ovan håller måttet? (Reliability)

Alltså, datakvalitet rör huruvida data, i sig, är lik vad den ska motsvara. Praktiska problem som kan uppstå är exempelvis att vi har fel i våra siffror, att vi saknar siffror vi borde ha, eller siffror som var korrekta är nu gamla och därmed felaktiga. Med det sagt kan vi gå över till informationskvalitet.

Människan i fokus – ”informationskvalitet”

Som figur 2 illustrerar blir människan viktigare i informationskvalitet. Informationskvalitet fokuserar på användning av information: hur väl information kan hjälpa oss lösa våra mål. En datamängd kan vara av hög datakvalitet, men låg informationskvalitet för att den inte är anpassad för verksamheten och dess behov. Likaså kan man ha hög informationskvalitet men sakna grundläggande datakvalitet, men detta är mer ovanligt då det förutsätter en

Figur 2. Den output du får om du ber AI illustrera informations-kvalitet med en bild. Källa: OpenAI DALL-E 3

situation där det inte är viktigt vad som är sant, endast att vi har ett underlag att stödja oss på.

Vanliga dimensioner inom informationskvalitet är [8,9]:

- Hur viktig är just denna information för våra uppgifter? (Relevancy)
- Levereras informationen i tid för det den ska användas för? (Timeliness)
- Upplevs informationen som god av dess användare? (Believability)
- Är det rätt mängd data som hanteras? För mycket, för lite? (Volume)
- Är det lätt eller svårt att förstå vad informationen säger? (Interpretability)
- Är informationen lättillgänglig? (Accessibility)

Vad säger forskningen?

Under denna studieperiod har jag gjort klart 3 studier. **Studie 1** undersöker vad kvalitet faktiskt betyder, och hur människan är en faktor för att avgöra det. **Studie 2** ger en kritisk granskning av forskningsläget på hur kvalitet på information påverkar beslut. **Studie 3** fokuserade på hur beslutsfattare upplever datakvalitet i deras beslutsfattande. Härefter beskrivs deras huvudresultat.

Studie 1 – vad kvalitet faktiskt betyder

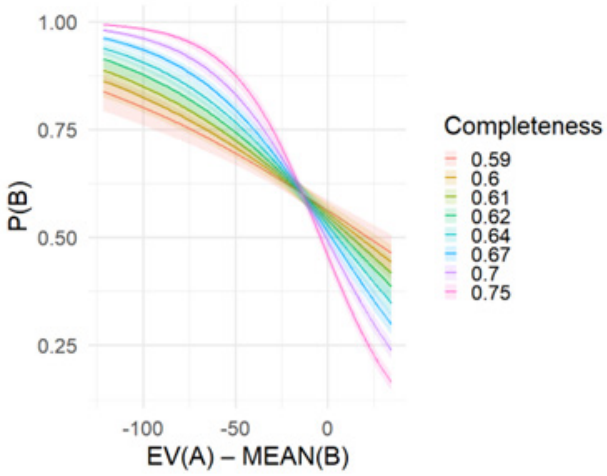
Det huvudsakliga resultatet av *studie 1* var att tidigare forskning har haft för lite fokus på människan när det kommer till kvalitet i information och data. Ett grundläggande problem är att vi ofta frågar experter om deras åsikter om sin data, vilket är ett subjektivt mått som inte tar hänsyn till deras begränsningar eller deras bias. Exempelvis kan en beslutsfattare oftast inte uttrycka sig om hur relevant eller korrekt deras data är, utan endast ge en åsikt som baseras på huruvida den information de hanterar har gett dem negativa konsekvenser. Studien kartlägger typiska beroendeförhållanden inom kvalitetsbedömning, som att vad som är ”complete” (fullständiga) data är beroende av vem som ska använda den och för vad.

Studie 2 – hur kvalitet på information påverkar beslut

Studie 2 ger en tydlig bild av kunskapsläget. I och med problemet i studie 1, att forskning ofta görs utan att ta hänsyn till mänskliga tendenser, så behövs mer rigorösa utvärderingar. Experimentella studier, som är den starkast formen av studie för att påvisa orsakssamband, har dock sällan använts för att studera data och informationskvalitet. Utöver att antalet studier är få, så använder många få deltagare och gör ofta misstag (likt Studie 1) i sina undersökningar över vad de faktiskt undersöker. Den sammantagna bilden kan beskrivas som:

- Ju mer korrekt vi blir informerade att informationen är, desto säkrare blir vi på våra beslut, desto godare bedömning gör vi av informationen. Experter (men inte noviser) tenderar att vara de som kan integrera information om datakvalitet i sin beslutsprocess.
- Det är oklart om fullständighet i materialet har en faktisk påverkan på våra beslut då studierna är få och relativt svaga ur ett designperspektiv.
- Information som är mer konsekvent tenderar inte att göra att vi känner att informationen är mer övertygande, men våra beslut blir trots det mer förutsägbara.
- Det är även otydligt huruvida att information anländer i tid (timeliness) påverkar våra beslut då ingen experimentell studie undersökt sambandet. Argumentativt bör försenad information kunna leda till förhastade beslut och oförsiktig användning av information som har andra brister.
- Några enstaka allmänna studier finns även – vilka antyder en påverkan av kvalitet på beslut – men deras analysnivå är för abstrakt för att kunna användas för planering

I en egen undersökning i samma studie undersöktes hur människor fattar beslut när de inte har fullständig information. Denna undersökning baserades på öppna data där deltagarna fick välja mellan alternativen A och B. Alternativen varierades systematiskt med avseende på värdeutfall, sannolikheter och mängden information som presenterades kontra doldes för deltagarna över ett stort antal beslutssituationer. Resultaten visar att när utfallschanser är ofullständiga blir det svårare att värdera alternativen:



Figur 3. Tendensen att välja alternativ B kontra A i relation till värdebedömning (Väntevärde (A) – känd risk, kontra förväntat Medelvärde B) över olika fullständighet i information. Källa: egenproducerad.

- Om B är det fördelaktiga valet i termer av utfall blir det lättare att välja B när informationen är mer fullständig. När informationen är ofullständig blir osäkerheten större, och beslutet svårare.
- På samma sätt blir det svårare att motivera ett val av A när data saknas i ökande grad – trots då A anses ha markant bättre värde.

Fenomenet liknar den så kallade **Ellsbergska paradoxen**, där människor undviker situationer med otydlig risk. Den här undersökningen visar att detta gäller när otydligheten är jämförbar, men att graden av ofullständighet i data också påverkar beslutet: ju fler möjliga utfall för B som är dolda, desto mer skiftar preferensen mot B. Dessa tendenser illustreras i Figur 3 på föregående sida, där X-axeln visar värdebedömning och Y-axeln visar sannolikheten att individen väljer alternativ B.

Sammantaget för studie 2 är att det finns flera möjligheter att fortsätta forska inom ämnet, som exemplifierats ovan, vilka är kritiska för att förstå hur vi påverkas av information av låg kvalitet. Dagens forskningsläge säger lite totalt, och än mindre om hur kvalitetsbristers påverkan på individnivå kan motverkas.

Studie 3 – datakvalitet i beslutsfattande

Studie 3 undersöker denna påverkan genom intervjuer med beslutsfattare. Det som framkom är hur viktig källans trovärdighet är för hur informationsmottagaren upplever informationen. Detta är särskilt framträdande när mottagaren av data saknar kunskapen att skilja mellan bra och dåliga data vilket får konsekvenser för vidare användning. I flera led av användning ökar avståndet till verksamheten som skapade informationen och minskar möjligheten att veta vad det handlar om och förmågan att kunna ifrågasätta den. Detta skapar ett beroende av tillit.

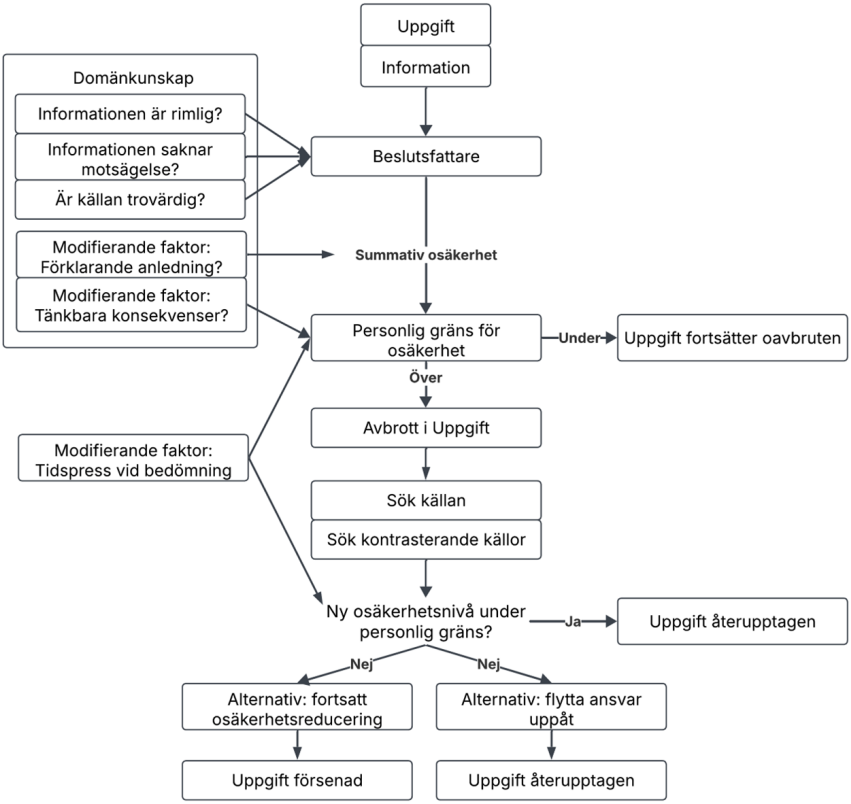
Tillit krävs för all verksamhet, men kan bli överdriven så att misstag inte ifrågasätts. Typiska konsekvenser är oftast tidsförluster i att försöka reda ut kvalitetsproblem, men även felaktiga beslut.

Lösningar kräver förstärkning av informationskedjan i alla led; som att lära ut vikten av korrekt rapportering, eller öka tydligheten i information genom att säkerställa definitioner, hur de beräknats, eller standardisera kommunikation.

En annan faktor av vikt är att vi inte förlorar den osäkerhet vi hade i ett visst skede. Informationens brister bör förmedlas så att denna vetskap inte missas när den senare ska användas i vidare beslut.

Vidare bygger studien en modell över hur beslutsfattare bemöter information som gör dem osäker. Det visar sig att frågor man oftast ställer sig relaterar

mycket till ens personliga kunskap och erfarenheter, så som: Är detta rimligt? Är källan trovärdig? Vet jag något annat som motsäger detta? Frågor som dessa väcker osäkerhet. I relation till konsekvenser och tidspress, tar individer handling för att reducera sin osäkerhet, bland annat genom att söka upp källan, eller jämföra källan mot andra perspektiv. Insikterna av modellen är: hur viktigt det är för informationsmottagaren att ha insikt i det informationen avser för att ens kunna agera kontrollör. Hur stark (eller svag) tillit till källan kan skapa problematiska situationer för verksamheten. Samt hur osäkerhet och låg kunskap om verksamheten tenderar att driva oss mot att acceptera information baserat på tillit; vilket inte garanterar kvalitet. Exempelvis kan tidspress ses som problematiskt, eftersom tidspress kan leda till att man behöver lita på data utan att faktiskt känna sig helt säker på det denna säger. Figur 4 nedan beskriver modellen, och läses ovanifrån och nedåt, detaljer kan förklaras på förfrågan.



Figur 4. Processmodell över bemötande av information för beslutsfattare. Källa: egenproducerad, Artikel 3.

Lärdomar i projektet

1. Tidigare forskning har lagt mycket energi på att visa att data och informationskvalitet har ett samband med utfall. Men förhållandevis lite energi på att visa hur. Forskningen som finns (Studie 2) är knaper och ofta bristande i att uppmärksamma de förhållanden som identifierats av Studie 1. Mitt arbete har bara påbörjat utforskningen kring den mänskliga faktorn, men många frågor återstår.
2. Informationsmottagaren som ska använda information bedömer den också. Om denna process fungerar väl så stärks kvaliteten, om den fungerar dåligt så får det konsekvenser som kanske inte är mätbart (e.g., tillitsbaserad acceptans av information).
3. Mottagarens upplevelse av informationens kvalitet påverkar hur denne kommer bedöma att deras processer fungerar, men det är oklart hur mycket det påverkar i mätbar effekt.
4. Tilltro till källan och verksamhetskunskap är båda viktiga fenomen i organisatoriska informationskedjor. Om dessa är för svaga eller för starka hos var informationsanvändare skapas problem för bedömning av information och vidare problem för korrekthet i beslut eller tidsåtgång.

I relation till ovan finns många balansgångar. Hur mycket bör man lita på den information som levereras av kollegor? Vad är en rimlig nivå av skepticism? Hur stor insikt i sina medarbetares eller avdelningspersonals arbetsdag och produktion kan man ha utan att vara påträngande? Hur mycket sådan information går att hålla koll på innan någon blir överbelastad? Vart går gränserna där individen börjar ignorera kvalitetsproblem för att bemöta organisatoriska kravbilder? Sådana frågor är i dagsläget obesvarade men av synnerlig vikt för utfall av kvalitetsproblematik.

Sammantaget är min bedömning att det finns få fullständiga kvalitetsmodeller och kvalitetsteorier som tar hänsyn till människan. Människan visas vara en faktor vi måste hantera om vi ska erhålla god kvalitet och vidare goda beslut. Studie 1 lyfter människan som en faktor i pusslet och studie 3 lyfter denna problematik till praktiken. Genom studie 1, 2, och 3 är det genomgående tydligt att mycket mer forskning behövs på området.

Mina resultat är bara början på det som behöver göras, framåt behöver energi riktas mot mer än traditionella kvalitetslösningar. Klassisk datarensning avser systemprocesser vilket automatiskt ska – genom regler och statistiska mått – höja kvalitén på materialet vi använder. Mitt arbete visar att det inte är så enkelt när vi har med människor att göra. Även om datarengöring hjälper, så måste vi börja fundera över hur information skapas, presenteras, och tas emot för att säkerställa att individens brister kan bemötas och styrkor lyftas.

Lästips

- Medium artikeln ”Data Quality: The Keystone of AI’s Arch to Success” ger en bra översikt och förklaringar till vikten av datakvalitet, speciellt inom implementering av AI.
- FN:s riskrapport: (https://unglobalriskreport.org)
- Kärnlitteratur: Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. För en mer teknisk beskrivning av vad som underbygger kvalitet. (kan förmedla PDF på förfrågan)
- Kärnlitteratur: Wang et al. (1996) Beyond accuracy: What data quality means to data consumers. Studie som la grunden för vikten av användaren av informations perspektiv. (kan förmedla PDF på förfrågan)
- Mina egna artiklar, om önskas fördjupning i det som nämns ovan, (kan förmedlas på förfrågan)

Referenser

[1] Fisher, Craig W., and Bruce R. Kingma. ”Criticality of data quality as exemplified in two disasters.” *Information & Management 39.2* (2001): 109-116.

[2] Nikiforova, A. (2020). Definition and Evaluation of Data Quality: User-Oriented Data Object-Driven Approach to Data Quality Assessment. *Baltic Journal of Modern Computing, 8*(3).

[3] Redman, T. C. (2016). Bad data costs the US $3 trillion per year. *Harvard Business Review, 22(*2016), 11-18.

[4] Romei, V., & Fleming, S. (2025). UK Inflation Overstated Due to Government Data Error, ONS Says. *Financial Times.*

[5] Setti, L., Passarini, F., De Gennaro, G., Barbieri, P., Perrone, M. G., Borelli, M., ... & Miani, A. (2020). Airborne transmission route of COVID-19: why 2 meters/6 feet of inter-personal distance could not be enough. *International journal of environmental research and public health, 17*(8), 2932.

[6] United Nations. (2024). UNITED NATIONS GLOBAL RISK REPORT 2024. https://unglobalriskreport.org/

[7] Young, T., Arnold, J., Brackey, T., Carr, M., Dwoyer, D., Fogleman, R., ... & Maguire, J. (2000). Mars program independent assessment team report.

[8] Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86-95.

[9] Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems, 12*(4), 5-33.

Mittuniversitetet
FORUM FÖR DIGITALISERING