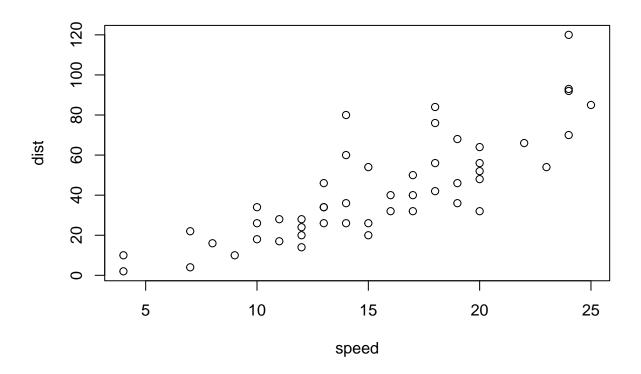
R Notebook og reproduserbarhet

Innlevering 1 i Data Science 2021 - Maren Sognefest og Daniel Karstad

plot(cars)



Minst et eksempel på bruk av følgende: italic bold italicbold

Reproduserbarhet

I senere tid har det oppstått en replikasjonskrise. Denne startet innenfor psykologien, og ble for alvor offentlig kjent i 2015 da 270 forskere samarbeidet om å forsøke å replikere 100 studier som alle var publisert i ledende tidsskrifter innenfor fagfeltet. De klarte kun å få samme resultat i under halvparten av studiene, og dette var selv med hjelp fra forskerne som

stod bak disse studiene Sætrevik (n.d.) (https://psykologtidsskriftet.no/fagessay/2017/07/replikasjonskrisen). Det har senere vist seg at disse replikasjonsproblemene finnes innenfor flere fagfelt, og i ettertid har det blitt større fokus på reproduserbarhet innenfor forskning. Reproduserbarhet er en forutsetning for replikerbarhet, så denne oppgaven skal vi ta for oss reproduserbarhet og hvorvidt bruk av R og R notebooks kan være en mulig løsning for å gjøre forskning reproduserbar, og dermed mer pålitelig.

Litteraturgjennomgang

Replikerbarhet/reproduserbarhet Det er enda ingen allmenn definisjon av "reproduserbarhet" og "replikerbarhet." Noen bruker disse begrepene om hverandre (Bollen et al., 2015), og andre er nøye med å skille dem fra hverandre (Leek and Peng, 2015: Goodman et al., 2016). I denne oppgaven vil vi skille begrepene tydelig fra hverandre og bruke Bollen et al. (2015) (et al., 2015) definisjoner av begrepene: "reproduserbarhet" oppstår dersom forskere klarer å komme frem til samme resultat ved å bruke samme prosedyre og samme datasett som gjort ved det opprinnelige studiet. "Replikerbarhet" oppstår dersom forskere klarer å komme frem til samme resultatet ved å bruke samme prosedyre og et nytt datasett. Hovedforskjellen er altså at ved "replikerbarhet" så skal det hentes inn nye data, men resultatet skal likevel bli det samme. De som prøver å replikere eller reprodusere studiet må altså ha tilgang til alt av data, kildekode og prosedyredetaljer. Man kan dermed si at reproduserbarhet er en betingelse for å kunne oppnå replikerbarhet.

Problemets omfang

I lys av replikasjonskrisen og det økte fokuset på replikasjon, har flere tidsskrifter begynt å publisere tilhørende datasett sammen med artiklene.

Vil dagens løsning med arkiv av data og eventuell programkode hos tidsskriftene kunne løse problemet?

Det er flere forskere o.l. som ikke ønsker å gi fra seg informasjon, dette gjelder data de besitter, koder, fremgangsmåte, dokumentasjon, resultat, feil, problemer de har møtt på, hypoteser osv. Dette gjør det svært vanskelig å reprodusere en tidligere studie, på et senere tidspunkt. Siden det samlet er flere problem, vil det naturligvis også være flere løsninger som må tas i bruk for at full reproduksjon skal være oppnåelig og man skal komme frem til lignende konklusjoner med ny data og ny sammensetning.

Vi kan definere og skille mellom tekniske og menneskelige løsninger. Det menneskelige aspektet i problemstillingen er ofte knyttet til det forskerne selv velger å dele av data, informasjon, koder, hypoteser, fremgangsmåte, programvare og så videre. Det har ikke vært praktisert, og standard retningslinjer for hva som bør ansees som god forskningsskikk og praksis er nødt å komme på plass for å møte kravene om tilfredsstillende reproduksjon og replikasjon. Det tekniske aspektet byr på mangel av data, koder, feil i programvare og feil som har oppstått underveis. Ved at man kan integrere og implementere programkode hos tidsskriftene, synlig

eller usynlig, så skal det være mulig for andre forskere å reprodusere studien og gjøre den replikerbar.

Oversikt over hva som bør sendes til tidsskriftene:

- En kode til å kune lese inn dataen med
- En kode til å kalkulere og analysere dataen
- En kode for å teste i henhold til hypoetese
- En kode for å generere en rapport av resultatet

Mulig løsning

I henhold til Gentleman (2005) er det nøye å integrere koder og beregninger som blir brukt i dataanalyser, metodebeskrivelser og simuleringer. Dette kan enkelt gjøres via et kompendium, hevdet av Gentleman (2005). Kompendium er en kortfattet oversikt over hovedinnholdet i f.eks. en studie gitt i dette tilfellet. Kompendiumet til da gi en oversikt over innholdet, slik som tekst, kode, data, metodikk, hypotese, problemstilling og så videre. Dette gjør at kompendiumet enkelt kan distribueres i ulike kanaler, enkelt kan håndteres og oppdateres i henhold.

Før var det RMarkdown som oftest ble brukt. Problemet der var at man ofte ikke fikk all tekst, data og koding i samme dokument, man måtte dele det opp i ulike tabs. RNotebook er den nyeste utgivelsen fra Rstudio.

Rstudio er en Integrated Developer Environment (IDE) for alt som er R relatert. Rstudio er gratis, både å laste ned og gratis å bruke. Man kan laste det ned lokalt på PC/Desktop eller jobbe online/remote. Alle vanlige operativ system (OS) skal være kompatible til å bruke Rstudio, bla. Mac, Windows, Linux. Rstudio må benyttes sammen med andre program og/eller extensions, f.eks GitAhead, GitHub Desktop, kommandolinje/terminal, for å kunne oppnå reproduserbarhet og replikerbarhet. Dette oppnås f.eks med kodeversjonskontroll, f.eks Git koblet med Github. En RNotebook er et RMarkdown dokument som inneholder kode+tekst-blokker, som henter inn koder og data, utfører beregninger og analyser i henhold til formler/kode som legges inn i Rmd filen. RNotebook vil da kunne vise oss et ferdig, vanlig tekstdokument, som inneholder tekst og koder for relevant innhold, istedenfor å ha dette i flere forskjellige filer. Dette gjør at du visuelt kan vurdere dataene mens du utvikler RMarkdown dokumentet uten å måtte «knyte» sammen hele dokumentet for å se resultatet.

Dette gjør at R Notebook kan brukes til å løse problemer knyttet til reproduserbarhet og replikerbarhet.

Vil dagens løsning med arkiv av data og eventuell programkode hos tidsskriftene kunne løse problemet?

En mulig løsning på problemet kan være å publisere forskningsartikler i kompendier, som også inneholder datasett og koder som er brukt i forskningen. I et slikt kompendium kan det

være dokumenter som kan oppdateres, også kjent som dynamiske dokumenter. I Rstudio kan man lage dynamiske dokumenter som blander tekst og R-kode. Et slikt dokument består av "text chunks" og "code chunks," altså bolker med både ren tekst og koding.

Analyse

Løser R Notebooks problemet med reproduserbarhet? Helt eller bare delvis? Med riktig bruk av R Notebook kan problemet med reproduserbarhet løses.

Dette dokumentet er skrevet i R studio, og det meste her er tekst, men som tidligere nevnt er fordelen med R Notebook at man kan blande bolker med tekst, sammen med bolker av koder. Når man laster ned pakker i Rstudio får man med noen dataset, som man kan bruke til å øve seg. Et av disse datasettene heter "cars" og code chunksene under henter data fra dette settet. Den først koden viser hvor langt den lengste bilen i datasettet kjørte.

max(cars\$dist)

[1] 120

Eksempler på "Code chuncks" ("R Code Block") og "text chunck i R notebook.

Har forskerne incentiver til å være "reproduserbare," eller må de tvinges?

I følge Florian Markowetz (2015) er det følgende fem egoistiske hovedgrunner til at forskerne selv burde ønske å publisere reproduserbar forskning:

- 1. Man unngår katastrofer
 - blabla
- 2. Det er lettere å skrive artikler
 - blabla
- 3. Lettere for fagfeller å forstå tankegangen
 - blabla
- 4. Det muliggjør kontinuitet i arbeidet
 - Det vil for eksempel ikke være noe stort problem dersom forskeren har glemt fremgangsmåten vedkommende brukte i forskningen sin i fjor. Det vil være muligheter for å kunne se hvordan man har tenkt og jobbet med studiet.
- 5. Hjelper deg å opparbeide et godt rykte

• Andre vil se på en forsker som publiserer reproduserbarforskning som en troverdig og grundig forsker, og dersom det noen gang blir problemer med noe av arbeidet, vil det være enkelt å vise og forklare hvordan man har tenkt og jobbet.

Er økt reproduserbarhet noe som vil tvinge seg frem eller er dagens økte interesse bare et blaff? Kan reproduserbarhet ha relevans i sektorer utenfor akademia?

Konklusjon

Vi kan konkludere med at RNotebook bidrar til å gjøre det mulig å reprodusere, replikere og generalisere et studie, dette ved hjelp av en dynamisk RMD fil som inneholder både data, koder, fremgangsmåte, resultat og referanser, som igjen produserer docx-, tex- og html-versjoner. Noe som kan by på hodebry og problemer er alle programmene, extensions og Git som skal kommunisere sammen. Dette bidrar til et uoversiktlig bilde i starten, og for de som skal ta det i bruk krever det en bratt læringskurve. Det man kan trekke frem som positivt for en forsker som skal ta dette i bruk er at man kan referere til logikk og utregninger direkte i dokumentet. Det viser hva som ligger bak og er ikke bare en visuell presentasjon. Om forskere da inkluderer alt av data, koder, fremgangsmåte og full utredelse for hva som har blitt gjort så vil dette kunne brukes av alle til å forstå og kunne brukes i en senere studie, eller bare brukes som en referanse i en ny studie/forskningsrapporter. Man bør ha flere retningslinjer og krav til hva man bør inkludere når man publiserer nye studier/rapporter/undersøkelser, dette vil bidra til økt standard for fremtidig bruk.

Litteraturliste

Sætrevik, B. (2017). Replikasjonskrisen. Psykologtidsskriftet. https://psykologtidsskriftet. no/fagessay/2017/07/replikasjonskrisen

Schmidt, M. L. 2015. Reproducible Research Using RMarkdown and Git through Rstudio. RPubs by Rstudio. https://rpubs.com/marschmi/105639

Markowitz, F., Five selfish reasons to work reproducibly. *Genome Biology*, 16(1) 274. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0850-7

Gentleman, R. & Lang, D. T. Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16(1) 1-23. https://www.tandfonline.com/doi/abs/10.1198/106186007X178663

Litteraturliste ved bruk av # References

Access to and preservation of scientific information in Europe: Report on the implementation of Commission Recommendation C(2012) 4890 final. (2015). Publications Office.

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., and Iannone, R. (2020). *Rmarkdown: Dynamic documents for r.*

American Economic Association. (n.d.). https://www.aeaweb.org/journals/data.

- Barnes, N. (2010). Publish your computer code: It is good enough. Nature, 467(7317), 753–753. https://doi.org/10.1038/467753a
- Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., and Goble, C. (2013). Why linked data is not enough for scientists. Future Generation Computer Systems, 29(2), 599–611. https://doi.org/10.1016/j.future. 2011.08.004
- Bollen, K., Cacioppo, J. T., Krosnick, J. A., Olds, J. L., and Kaplan, R. M. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science* (Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences). NSF.
- Brase, J. (2009). DataCite A Global Registration Agency for Research Data. 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 257–261. https://doi.org/10.1109/COINFO.2009.66
- Buckheit, J. B., and Donoho, D. L. (1995). WaveLab and Reproducible Research. In P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg, I. Olkin, N. Wermuth, S. Zeger, A. Antoniadis, and G. Oppenheim (Eds.), *Wavelets and Statistics* (Vol. 103, pp. 55–81). Springer New York. https://doi.org/10.1007/978-1-4612-2544-7_5
- Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). https://doi.org/10.1126/science.aac4716
- Conyers, R. (n.d.). Learn Git in 3 Hours. In O'Reilly | Safari. https://www.safaribooksonline.com/videos/lgit-in/9781789348231/9781789348231-video1 2.
- Dear Colleague Letter: Robust and Reliable Research in the Social, Behavioral, and Economic Sciences (Nsf16137) | NSF National Science Foundation. (n.d.-a). https://www.nsf.gov/pubs/2016/nsf16137/nsf16137.jsp.
- Dear Colleague Letter: Robust and Reliable Research in the Social, Behavioral, and Economic Sciences (Nsf16137) | NSF National Science Foundation. (n.d.-b). https://www.nsf.gov/pubs/2016/nsf16137/nsf16137.jsp.
- Dewald, William G., Thursby, J. G., and Anderson, R. G. (1986). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, 76(4), 587–603.
- Dewald, William G., Thursby, J. G., and Anderson, R. G. (2020). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. 18.
- Ezekiel, M. (1933). Some considerations on the analysis of the prices of competing or substitute commodities. Econometrica, 1(2), 172–180.
- Fomel, S., and Claerbout, J. F. (2009). Guest editors' introduction: Reproducible research. Computing in Science Engineering, 11(1), 5–7.
- Frisch, R. (1933). Editor's note. Econometrica, 1(1), 1–4.

- Gentleman, R. (2005). Reproducible Research: A Bioinformatics Case Study. *Statistical Applications in Genetics and Molecular Biology*, 4(1). https://doi.org/10.2202/1544-6115.1034
- Gentleman, R., and Lang, D. T. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16(1), 1–23. https://doi.org/10.1198/106186007X178663
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science (New York, N.Y.), 286 (5439), 531–537. https://doi.org/10.1126/science.286.5439.531
- Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341), 341ps12–341ps12. https://doi.org/10.1126/scitranslmed.aaf5027
- Grolemund, G., and Wickham, H. (n.d.). R for Data Science.
- Gruber, J., and Swartz, A. (n.d.). Daring Fireball: Markdown. https://daringfireball.net/projects/markdown
- Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124
- Iyengar, S., and Greenhouse, J. B. (1988). Selection models and the file drawer problem. Statistical Science, 3(1), 109–117.
- Jasny, B. R., Chin, G., Chong, L., and Vignieri, S. (2011). Again, and again, and again. Science, 334 (6060), 1225–1225. https://doi.org/10.1126/science.334.6060.1225
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., ... Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. Advances in Methods and Practices in Psychological Science, 1(4), 443–490. https://doi.org/10.1177/2515245918810225
- Knuth, D. E. (1984). Literate programming. The Computer Journal, 27(2), 97–111. https://doi.org/10.1093/comjnl/27.2.97
- Knuth, Donald E. (1992). Literate Programming. Cambridge University Press.
- Knuth, D. E. (1986). The TeXbook. Addison-Wesley.
- Lander, J. P. (2017). *R for Everyone: Advanced Analytics and Graphics* (2nd Edition). Addison-Wesley Professional.
- Leisch, F. (2002). Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis. In W. Härdle and B. Rönz (Eds.), *Compstat* (pp. 575–580). Physica-Verlag HD. https://doi.org/10.1007/978-3-642-57489-4_89
- Markowetz, F. (2015). Five selfish reasons to work reproducibly. Genome Biology, 16(1), 274. https://doi.org/10.1186/s13059-015-0850-7

- McCullough, B. D., McGeary, K. A., and Harrison, T. D. (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue Canadienne d'économique*, 41(4), 1406–1420. https://doi.org/10.1111/j.1540-5982.2008.00509.x
- McNutt, M. (2014). Reproducibility. *Science*, 343 (6168), 229–229. https://doi.org/10.1126/science.1250475
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348 (6242), 1422–1425. https://doi.org/10.1126/science.aab2374
- Nosek, Brian A., and et al. (2015). Estimating the reproducibility of psychological science. Science, 349(6251). https://doi.org/10.1126/science.aac4716
- Nüst, D., Konkol, M., Pebesma, E., Kray, C., Schutzeichel, M., Przibytzin, H., and Lorenz, J. (2017). Opening the Publication Process with Executable Research Compendia. *D-Lib Magazine*, 23(1/2). https://doi.org/10.1045/january2017-nuest
- Pandoc Pandoc User's Guide. (n.d.). https://pandoc.org/MANUAL.html.
- Peng, R. D. (2011). Reproducible Research in Computational Science. Science, 334 (6060), 1226–1227. https://doi.org/10.1126/science.1213847
- R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing.
- Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. Source Code for Biology and Medicine, 8(1), 7. https://doi.org/10.1186/1751-0473-8-7
- Ramsey, N. (n.d.). Noweb home page. https://www.cs.tufts.edu/~nr/noweb/.
- Riederer, E., Christophe Dervieux. (n.d.). R Markdown Cookbook.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. 86, 638–641.
- RStudio Team. (2020). RStudio: Integrated development environment for r. RStudio, PBC.
- Schwab, M., Karrenbach, M., and Claerbout, J. (1995). Reproducible electronic documents. 14.
- Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. https://doi.org/10.1177/0956797611417632
- Sætrevik, B. (n.d.). Replikasjonskrisen.
- Tierney, N. (n.d.). RMarkdown for Scientists.
- Wickham, H., and Grolemund, G. (2016). R for data science: Import, tidy, transform, visualize, and model data (pp. XXV, 492). O'Reilly.
- wikipedia. (2020). Meta-analysis. Wikipedia.

- Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, and R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman and Hall/CRC.
- Xie, Y. (2015). Dynamic documents with R and knitr (Second). Chapman and Hall/CRC.
- Xie, Y. (2020). Knitr: A general-purpose package for dynamic report generation in r [Manual].
- Xie, Y., Allaire, J. J., and Grolemund, G. (2018). R markdown: The definitive guide. Chapman and Hall/CRC.
- Young, N. S., Ioannidis, J. P. A., and Al-Ubaydli, O. (2008). Why Current Publication Practices May Distort Science. *PLoS Medicine*, 5(10), e201. https://doi.org/10.1371/journal.pmed.0050201

Appendiks

```
uthor: Maren <maren.sognefest@gmail.co
ate: Tue Sep 21 12:33:59 2021 +0200
mmit 5a4f146cbc3cc0b6a3244782df467ed699a48956
rge: 0557965 2ebc434
thor: Maren <maren.sognefest@gmail.com>
te: Tue Sep 21 10:54:59 2021 +0200
            0557965cd415bc445c152f042a556235986f33f6
Maren <maren.sognefest@gmail.com>
Tue Sep 21 10:49:15 2021 +0200
mmr. zemc43410fdfbcbed9958987=52ddd4b0a28d734
thor: Dankar03 <90779411+Dankar03@users.noreply.github.com>
te: Mon Sep 20 20:25:34 2021 +0200
              922c38ef9b11618810d567fe6c1e6cf667712578
Dankar03 <90779411+Dankar03@users.noreply.github.com>
Mon Sep 20 20:16:00 2021 +0200
             pp.HUG...
21794de79f1c3781439c87820ed1fc27as1067bb6 (HEAD -> master, origin/master)
Maren «maren.sognefest@gmail.com»
Tue Sep 21 12:33:59 2021 +0200
          5a4f146cbc3cc0b6a3244782df467ed699a48956

0557965 2ebc434

: Maren «maren sognefest@gmail.com>

Tue Sep 21 10:54:59 2021 +0200
 mmit 0557965cd415bc445c152f042a556235986f33f6
thor: Maren <maren.sognefest@gmail.com>
te: Tue Sep 21 10:49:15 2021 +0200
mmit Zebc43410fdfbcbed9958987e52ddd4b0a28d734
thor: Dankar03 <90779411+Dankar03@users.noreply.github.com>
te: Mon Sep 20 20:25:34 2021 +0200
            922c38ef9b11618810d567fe6c1e6cf667712578
Dankar03 <90779411+Dankar03@users.noreply.github.com>
Mon Sep 20 20:16:00 2021 +0200
ommit adeaal37fbe85741aeb7031623b9530ea3684541 (origin/litteraturliste, origin/konklusjon
tthor: DankarO3 <90779411-DankarO340users.noreply.github.com>
tte: Mon Sep 20 19:52:50 2021 +0200
    påbegynt konklusjon
           fc9849139e85.74758f71bc61e4e67a4e99853ce
056f1ee 0d5fbc?
: Dankar03 <090759411+Dankar03@users.noreply.github.com>
Mon Sep 20 19:28:21 2021 +0200
    Merge branch 'teori'
            Od5fbc22cb2c6779c668ebe8db31c568e0597921 (origin/teori
Dankar03 <90779411+Dankar03@users.noreply.github.com>
Mon Sep 20 19:17:38 2021 +0200
    mer tekst teori
ommit 5dbb5654e6ab999492d5ald7e83e2aea4d1b831e
ruge: 056flee ec7aece
thor: Maren <maren.sognefest@gmail.com>
te: Mon Sep 20 19:08:10 2021 +0200
```

Figure 1: git log/siste commits. Bildet viser siste commits, og at det er brukt flere branches (her: master, teori og konklusjon)