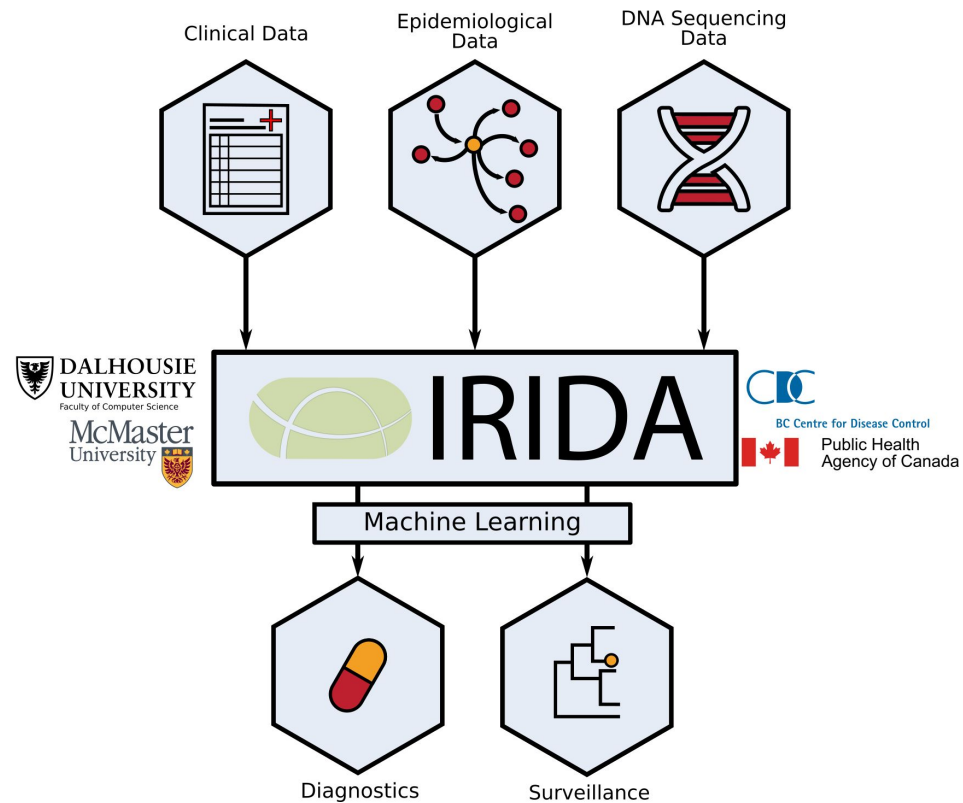
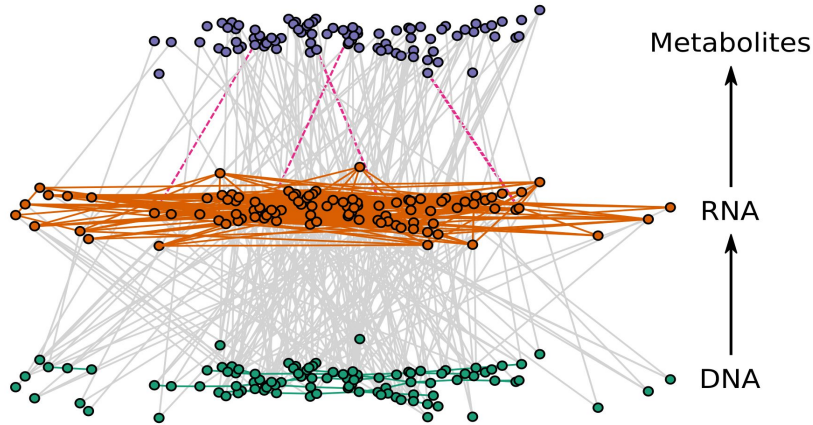


Introduction to Applied Research in Health Data Science

CSCI6XXX/CHE6XXX/CSCI4XXX
(CSCI6093)

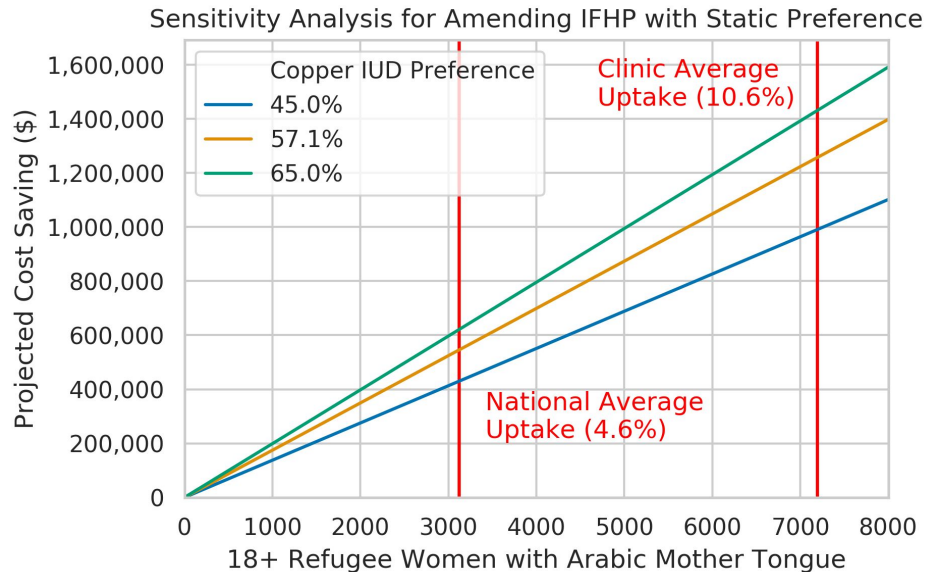
Finlay Maguire (finlay.maguire@dal.ca)

Why am I teaching this course?

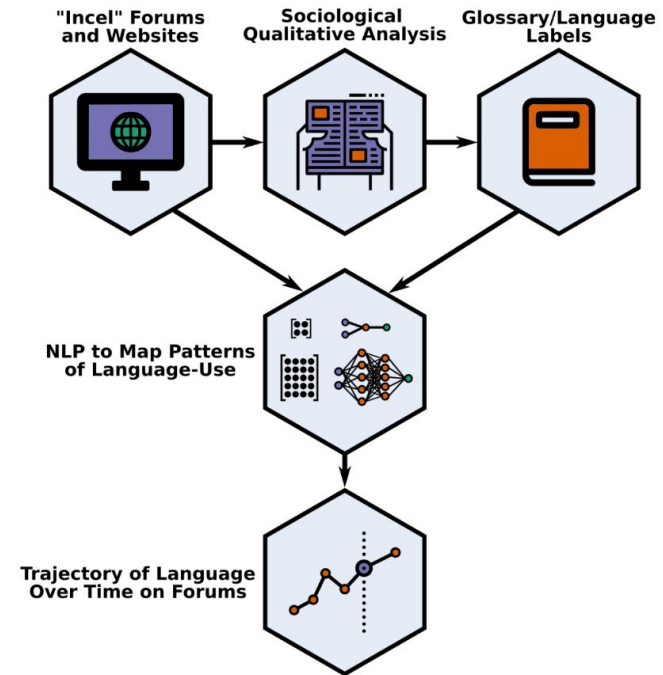


- **PhD (Bioinformatics)**: using large noisy datasets to understand how microbial systems and mechanisms evolve.
- **Postdoc (Genomic Epidemiology)**: using large noisy datasets to better diagnose, track and predict infectious diseases.

Why am I teaching this course?



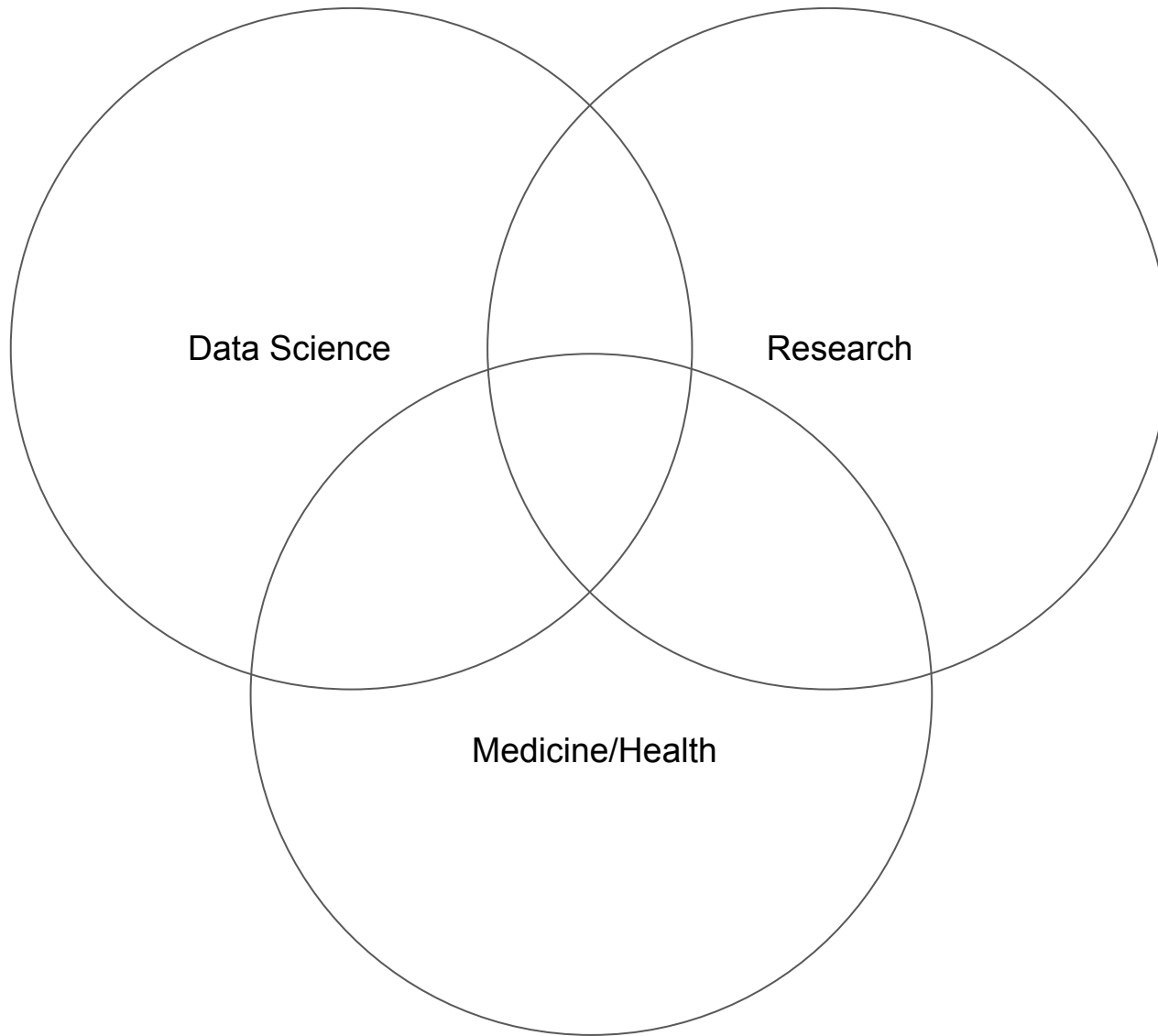
Modelling "Incel" Online Radicalisation via NLP



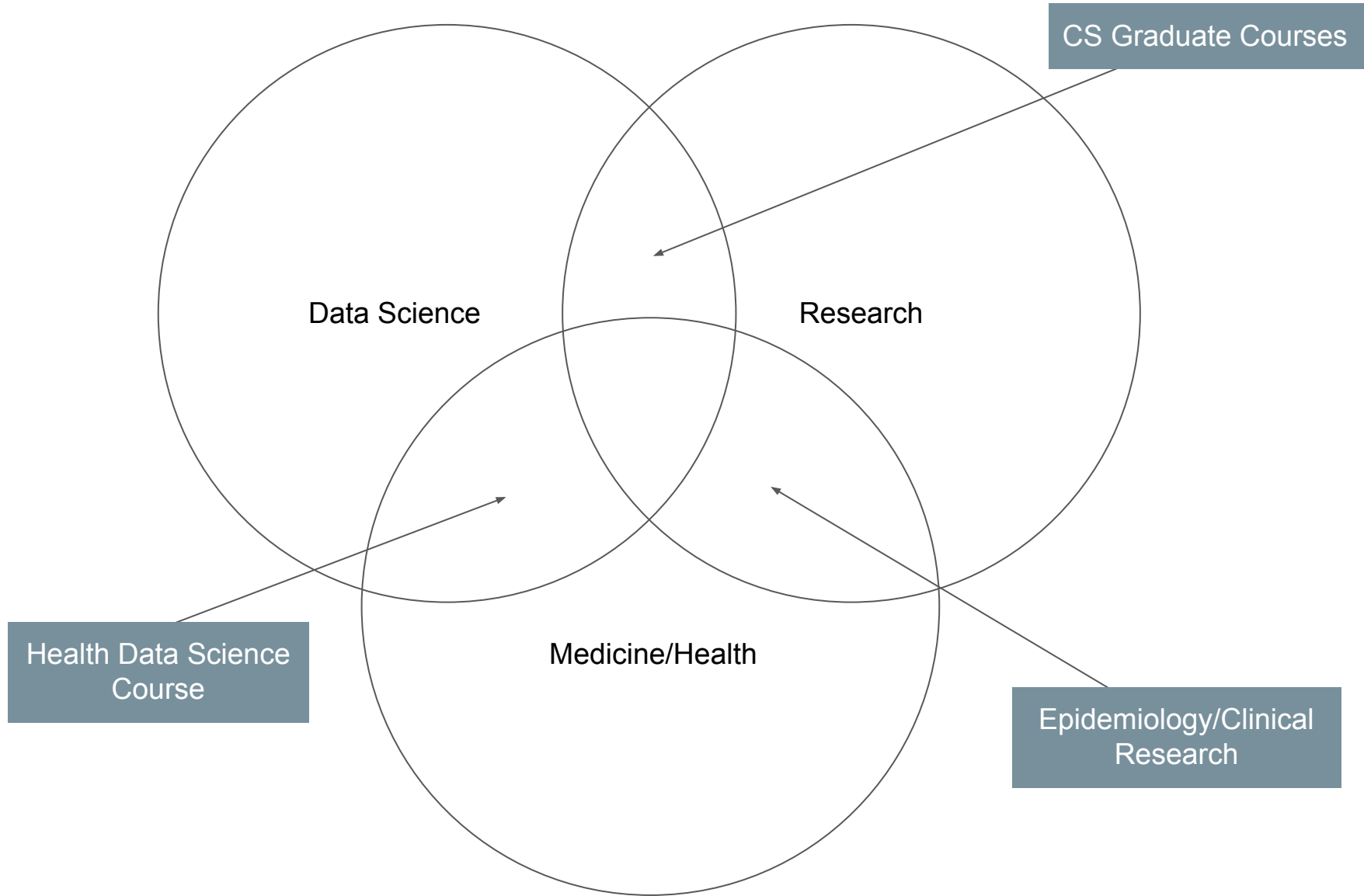
- **Research group:** using large noisy datasets:
 - Genomic epidemiology of infectious disease: **SARS-CoV-2, AMR**
 - Collaborations on socially/health focused problems: **refugee health, incel radicalisation, health inequality**

Overview of course

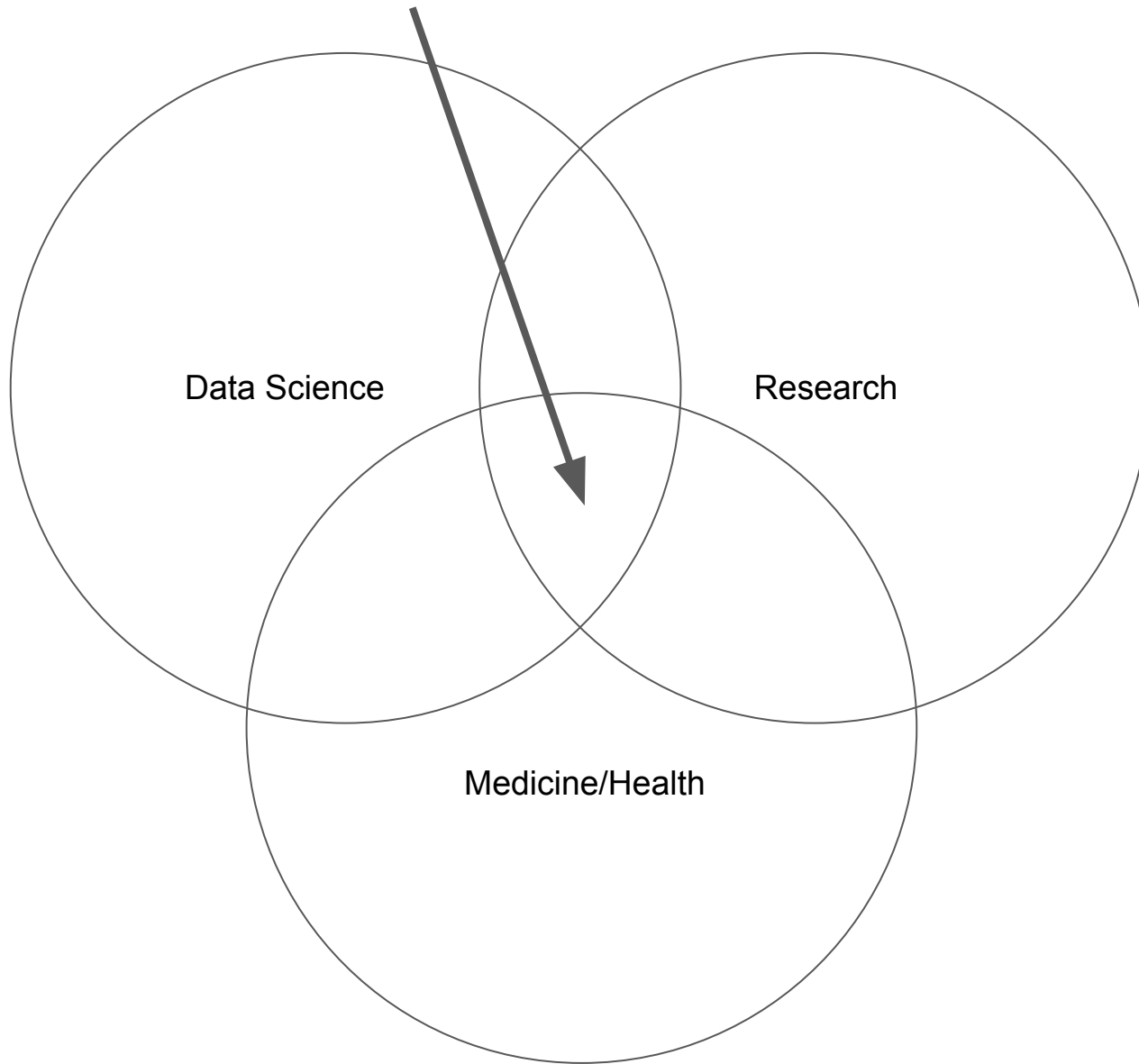
Applied Research in Health Data Science



Applied Research in Health Data Science



Applied Research in Health Data Science



Learning Outcomes

1. Understand the **4 principal sources and data types** of medical data:
 - a. longitudinal databases (tabular)
 - b. electronic medical records (structured, semi-structured, and unstructured text)
 - c. radiological imaging (image)
 - d. physiological (signal and time-series).
2. Identify and apply **appropriate type of method** to the analysis of each data type
3. Gain the technical skills necessary for effective health data science research including **data management, reproducibility**, and version control.
4. Understand the key **collaborative, legal, ethical, and knowledge translation** concepts required in interdisciplinary health data science research.
5. Critically **appraise research literature** in health data science.
6. Combine these skills to develop high-quality collaborative health data science **research proposals**

What is not covered in this course

- **Breadth/depth** of each data science method: *each could be multiple graduate CS courses*
- **Breadth/depth** of medical research: *again could be a whole PhD program*
- True **messiness** of real data: *provide tools but experience is invaluable*
- Some important forms of medical data (e.g., genomics): *see next year's **genomic medicine** course if interested.*

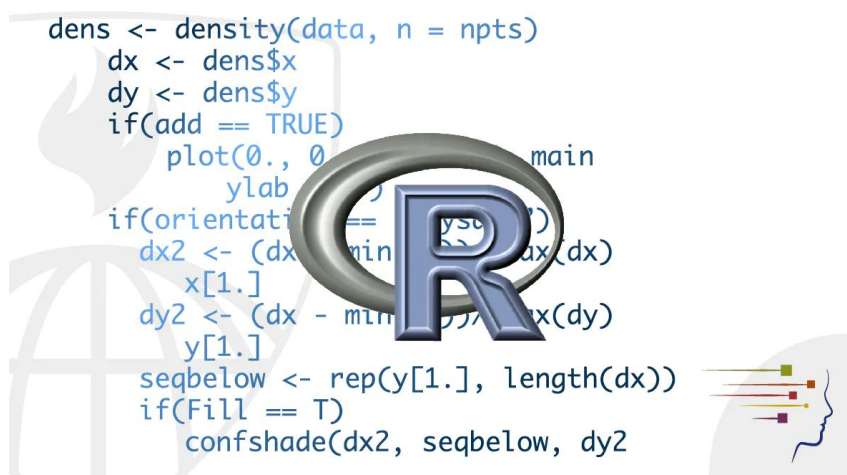

Course Structure

Survey of key methods for principal data types:

- **Lectures** (Mondays)
- **Practical Exercises** in R (Fridays)

Assessment: Submission of Practical Exercise
Answers by the following Monday

```
dens <- density(data, n = npts)
dx <- dens$x
dy <- dens$y
if(add == TRUE)
  plot(0., 0., main = "Density Plot",
       xlab = "x", ylab = "Density",
       if(orientation == "y")
         dx2 <- (dx - min(dx)) / (max(dx) - min(dx))
         x[1.]
       dy2 <- (dy - min(dy)) / (max(dy) - min(dy))
       y[1.]
  seqbelow <- rep(y[1.], length(dx))
  if(Fill == T)
    confshade(dx2, seqbelow, dy2)
```



<https://www.coursera.org/learn/r-programming>

Effective research in health data science:

- **Journal Club** (Wednesday A)

2 papers per week, rota for leading discussion of paper with rest of class.

Assessment: participation in discussion

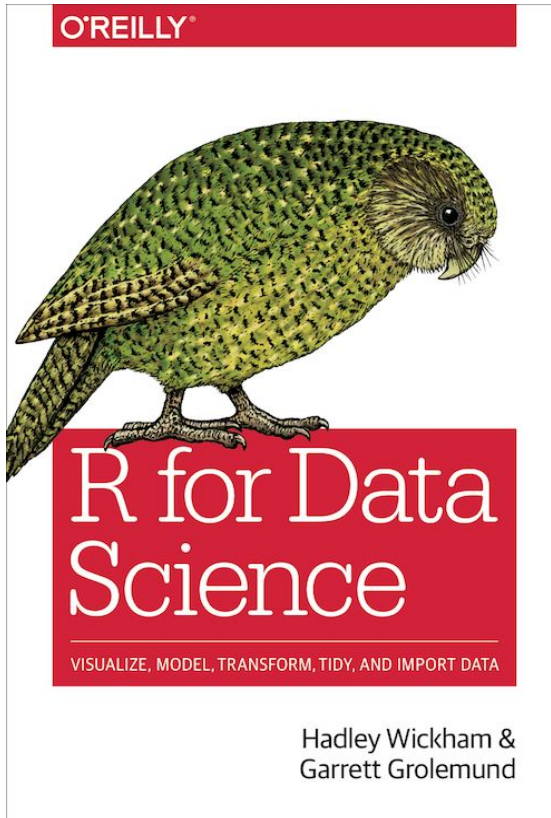
Development of a research proposal:

- **Class** (Wednesday B)

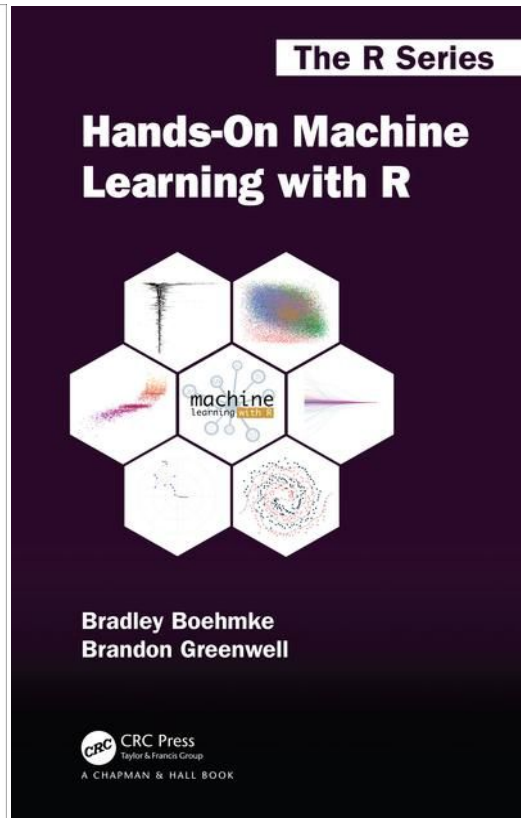
Assessment: Presentation in last week of class

Submitted final day of class

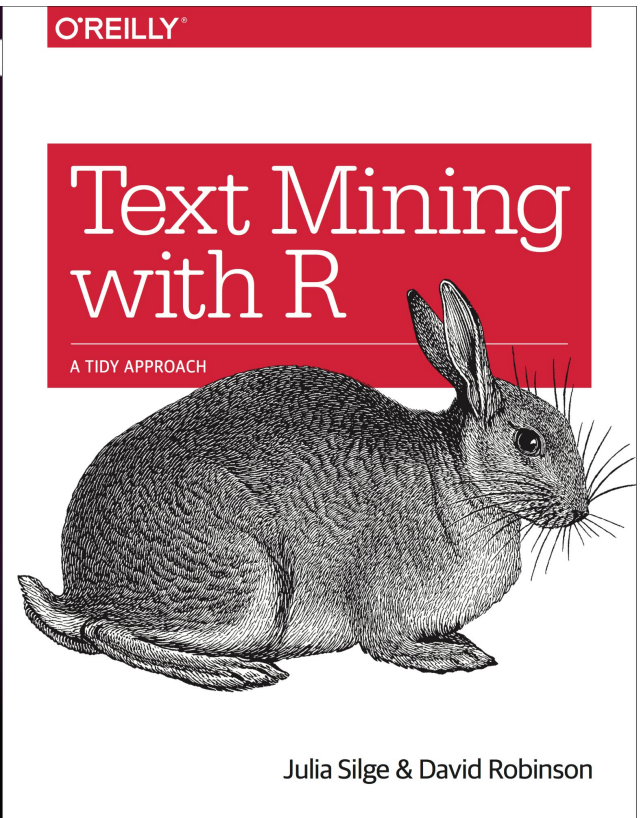
Course Materials



<https://r4ds.had.co.nz/>



<https://bradleyboehmke.github.io/HOML/>

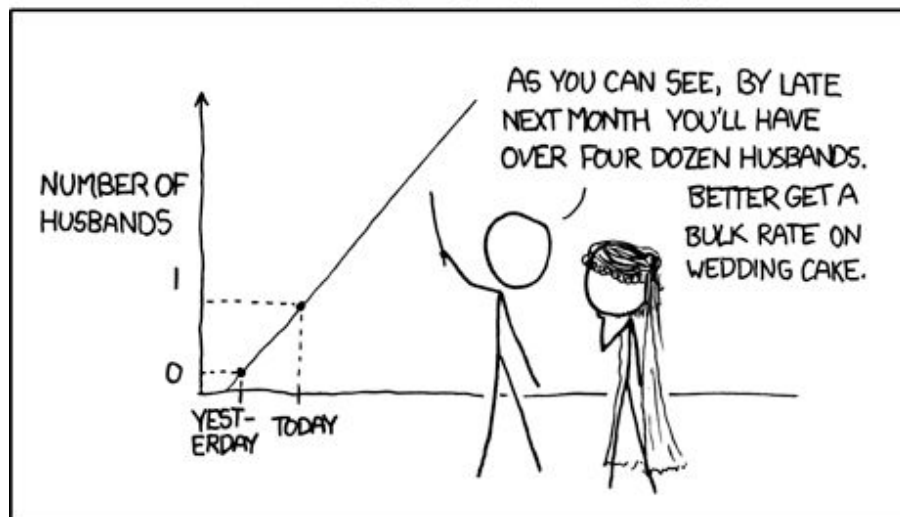


<https://www.tidytextmining.com/>

What is health data science?

Partially re-branded statistics

MY HOBBY: EXTRAPOLATING



Pitfalls:

- Less rigorous/principled
- Prone to reinventing the wheel

Benefits:

- More flexible
- Less prescriptive/intimidating

THIS IS YOUR MACHINE LEARNING SYSTEM?

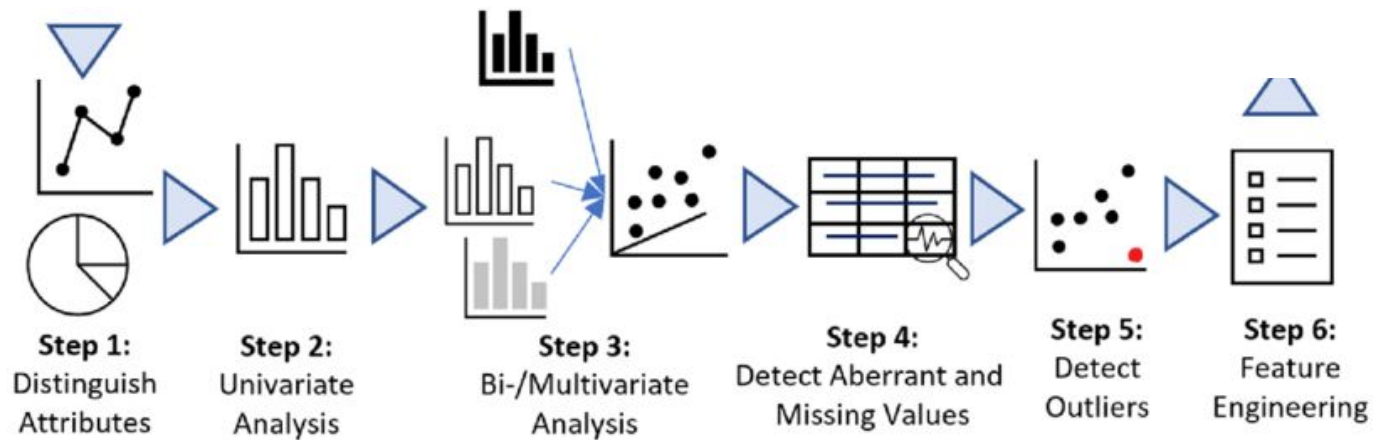
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.



Promotes exploratory data analysis



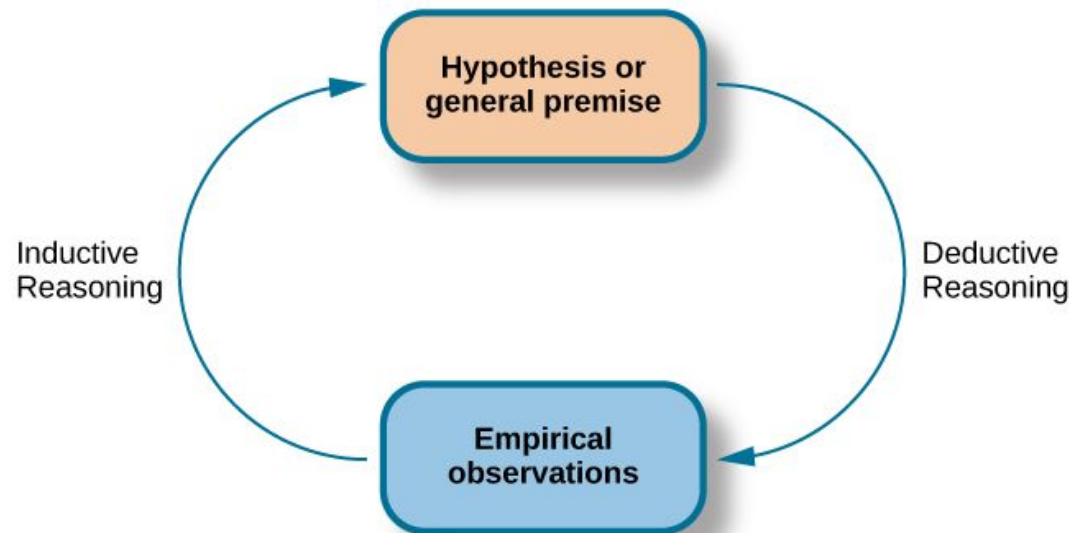
Supports inductive approaches

Deductive:

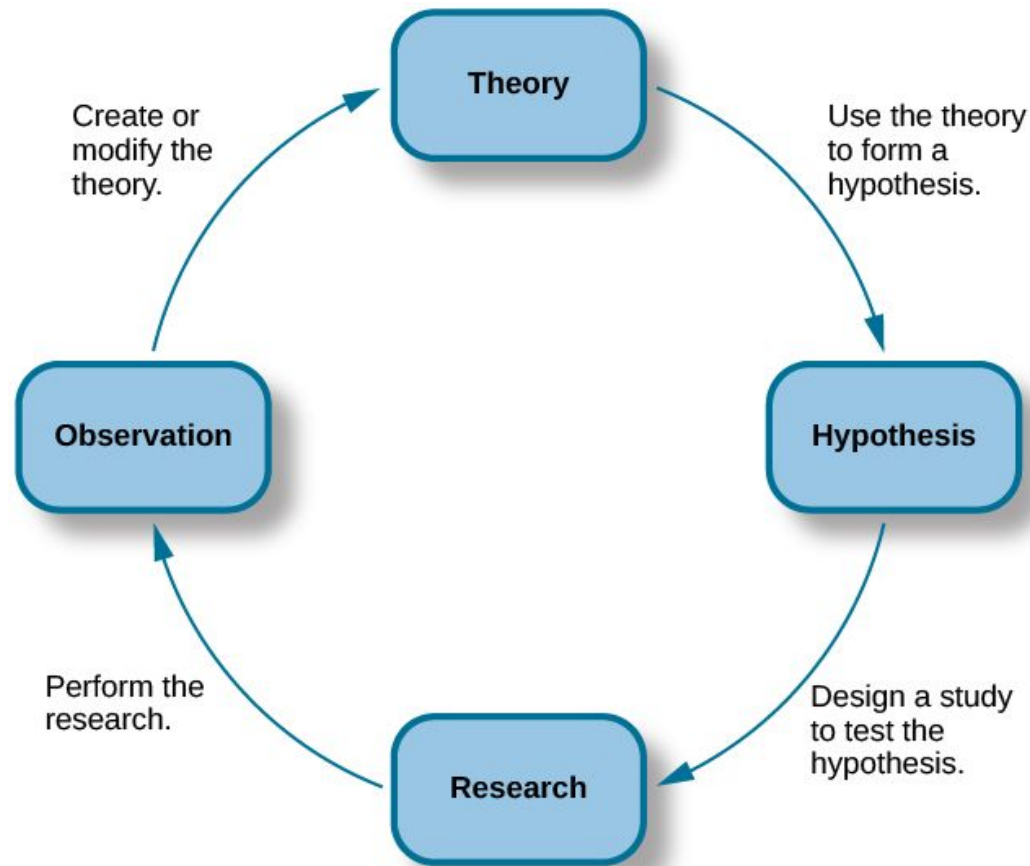
- “Condition X, causes Y”
- Collect data
- Perform frequentist statistical test
- Reject or confirm null hypothesis

Inductive:

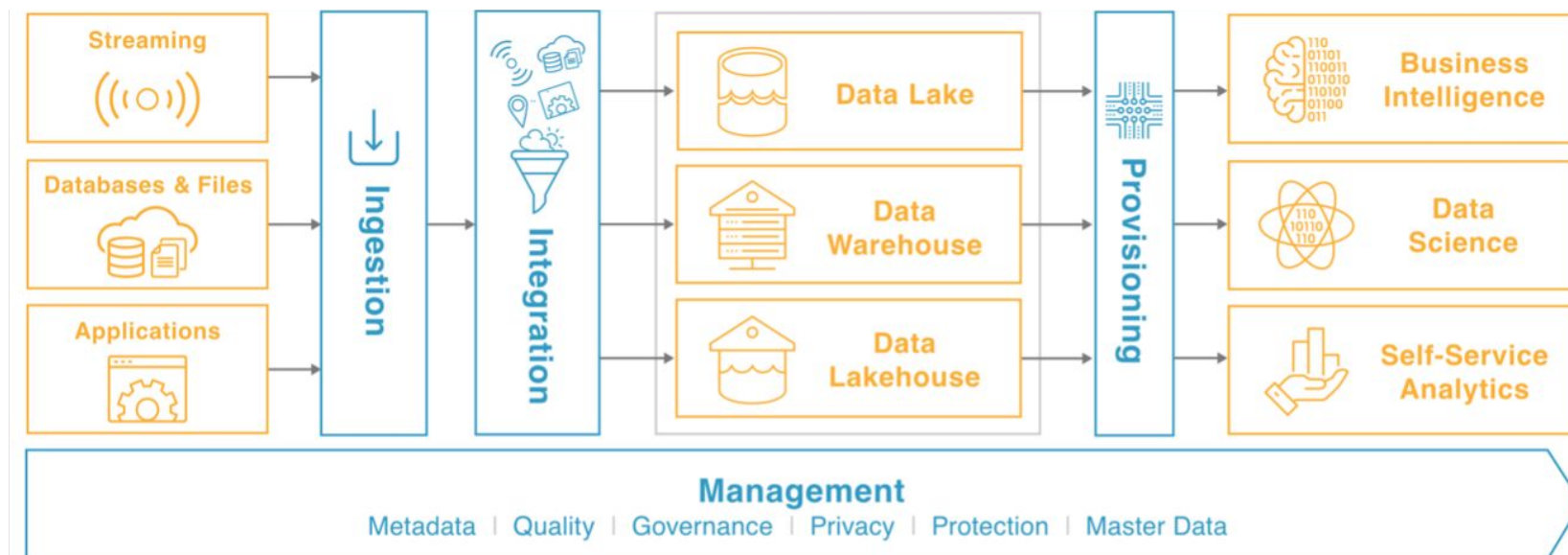
- Collect data
- Identify patterns in the data
- Observe X and Y seem connected somehow
- Quantify strength of association



Supports inductive approaches



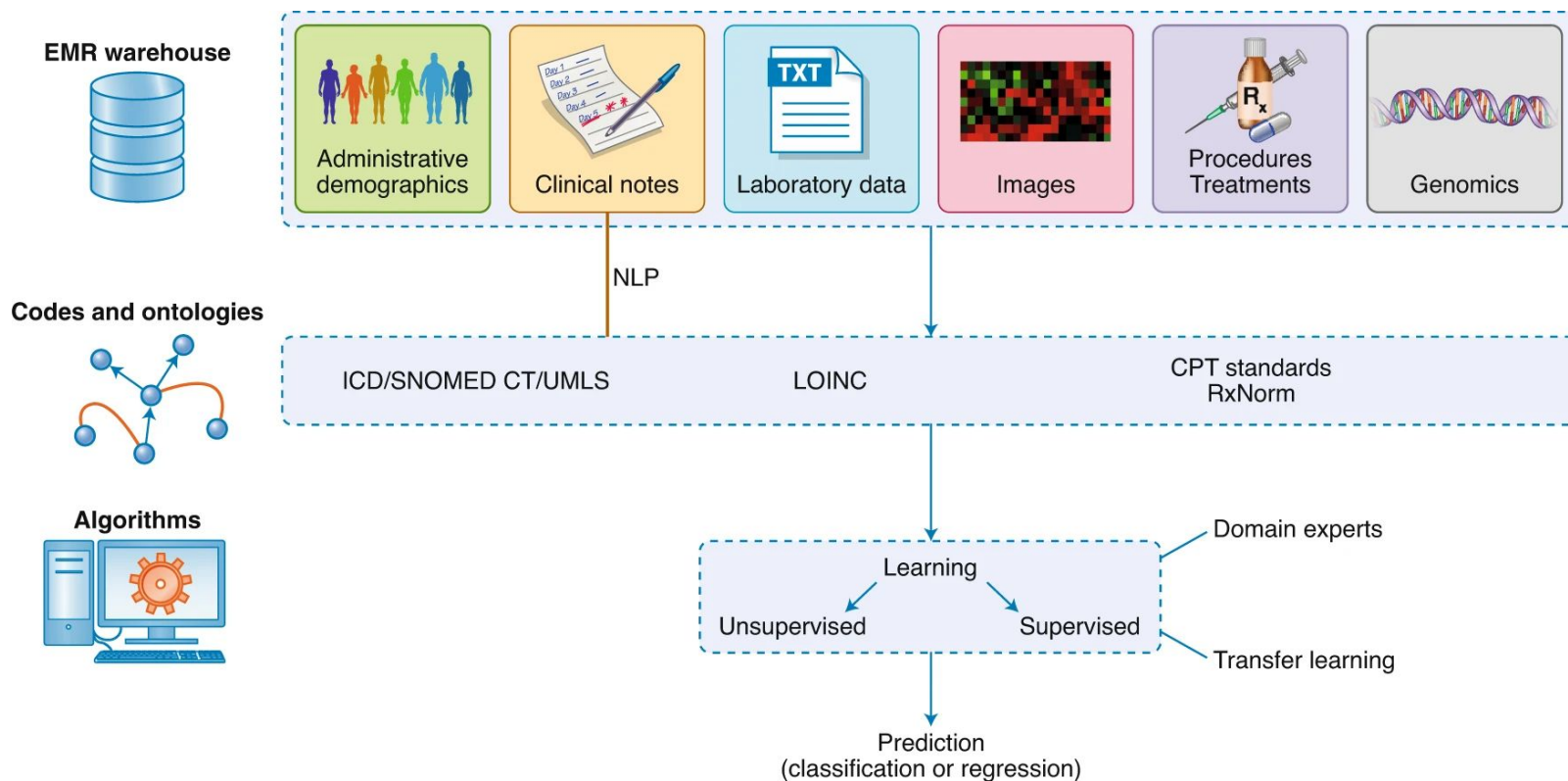
Data-intensive analytical design/tooling



<https://www.2ndwatch.com/blog/what-is-a-data-pipeline-and-how-to-build-one/>

So what about “Health Data Science”

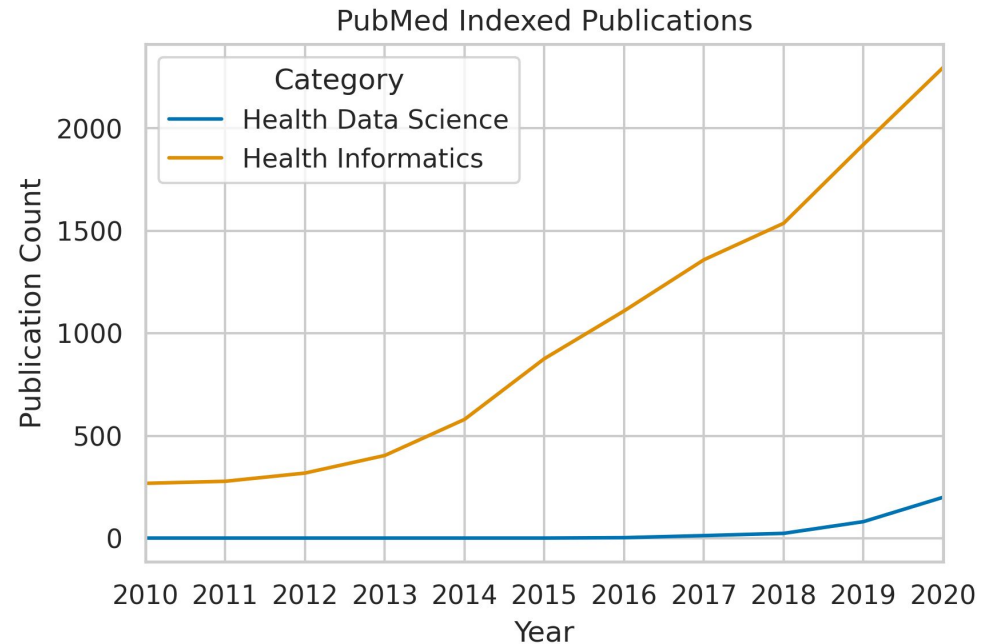
Data Science applied to Health Data



<https://www.nature.com/articles/s41588-020-0698-y/figures/2>

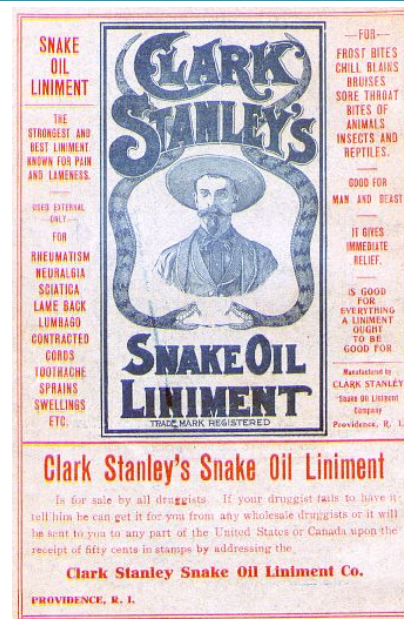
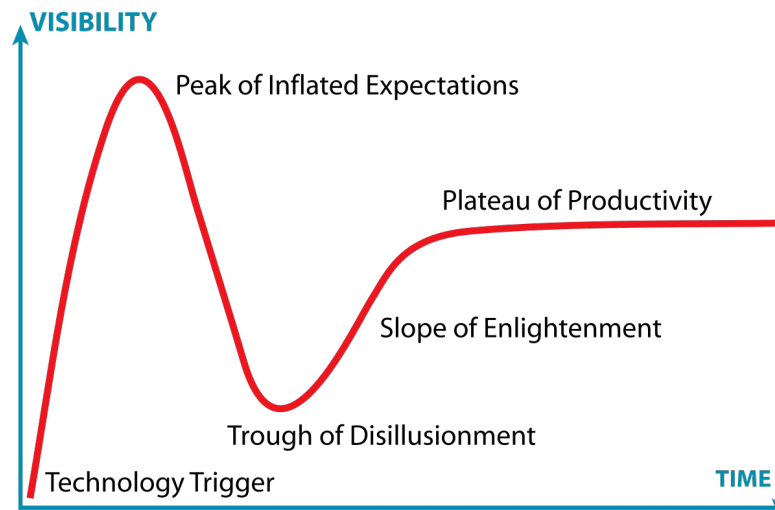
Benefits of Health Data Science

- Huge amounts of medical data
- Many open questions
- Domain experts are enthusiastic and willing
- People with broad skills are relatively rare



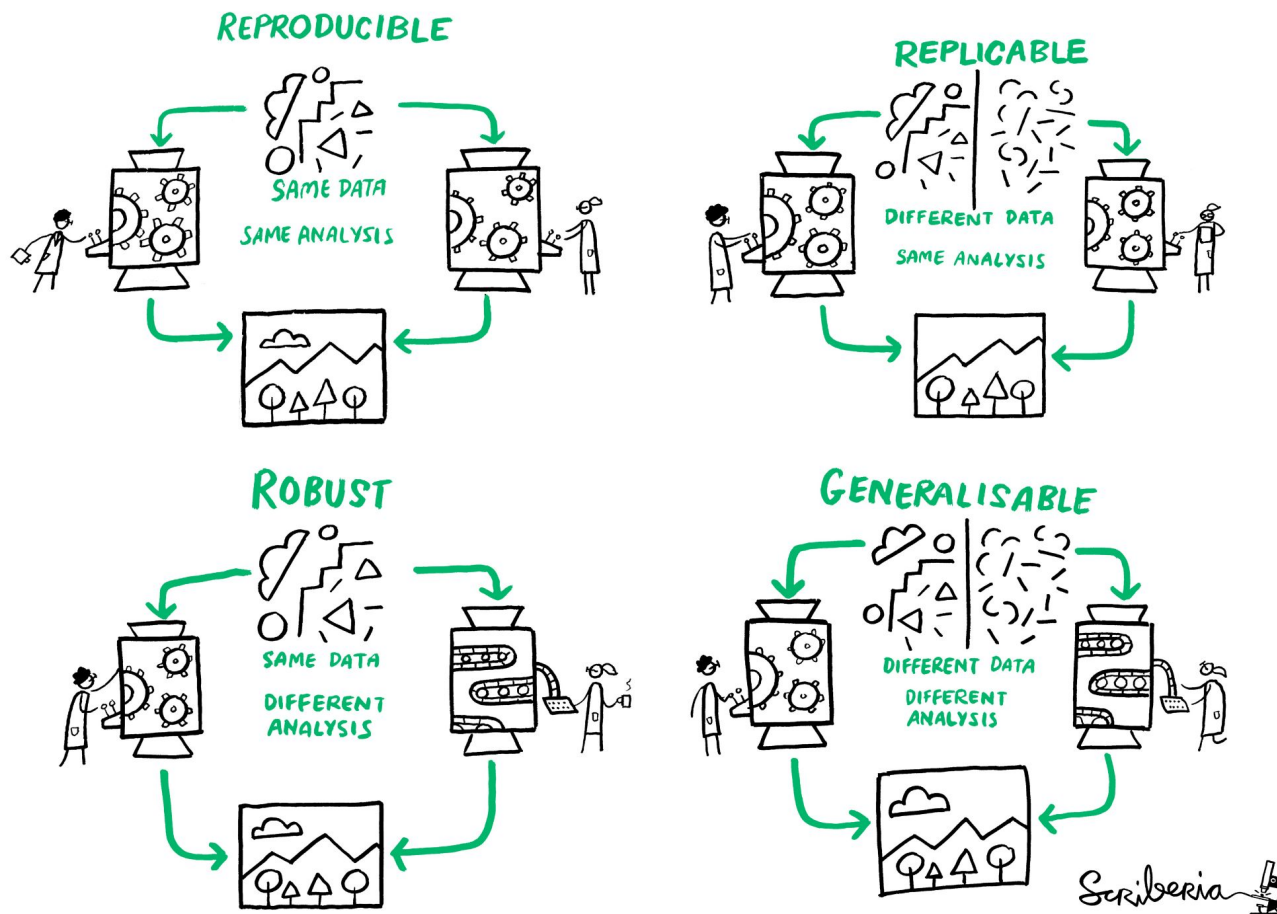
Challenges of Health Data Science

- Lots of hype
- Data quality issues
- Contextual/Metadata quality issues
- Influence of US health system
- Ethical pitfalls
- Treatment to the mean

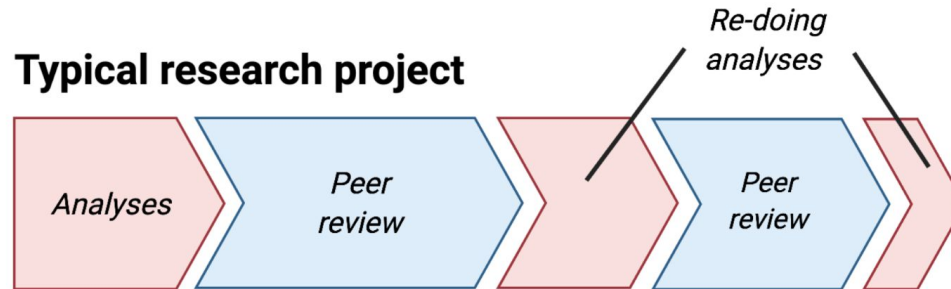


Reproducibility

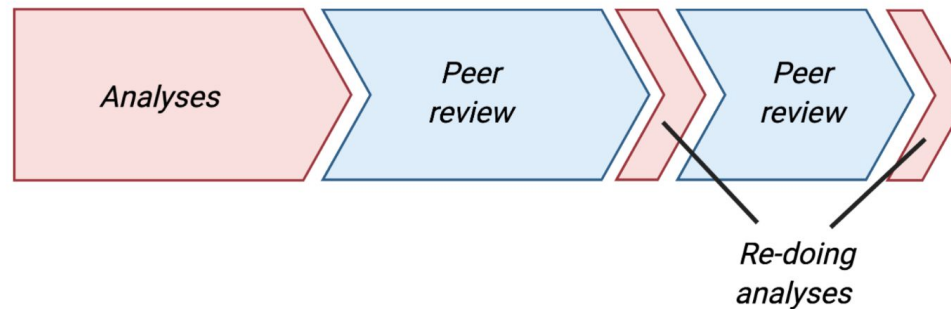
Reproducibility should be the bare minimum



Makes your own life easier



Research project using reproducible practices

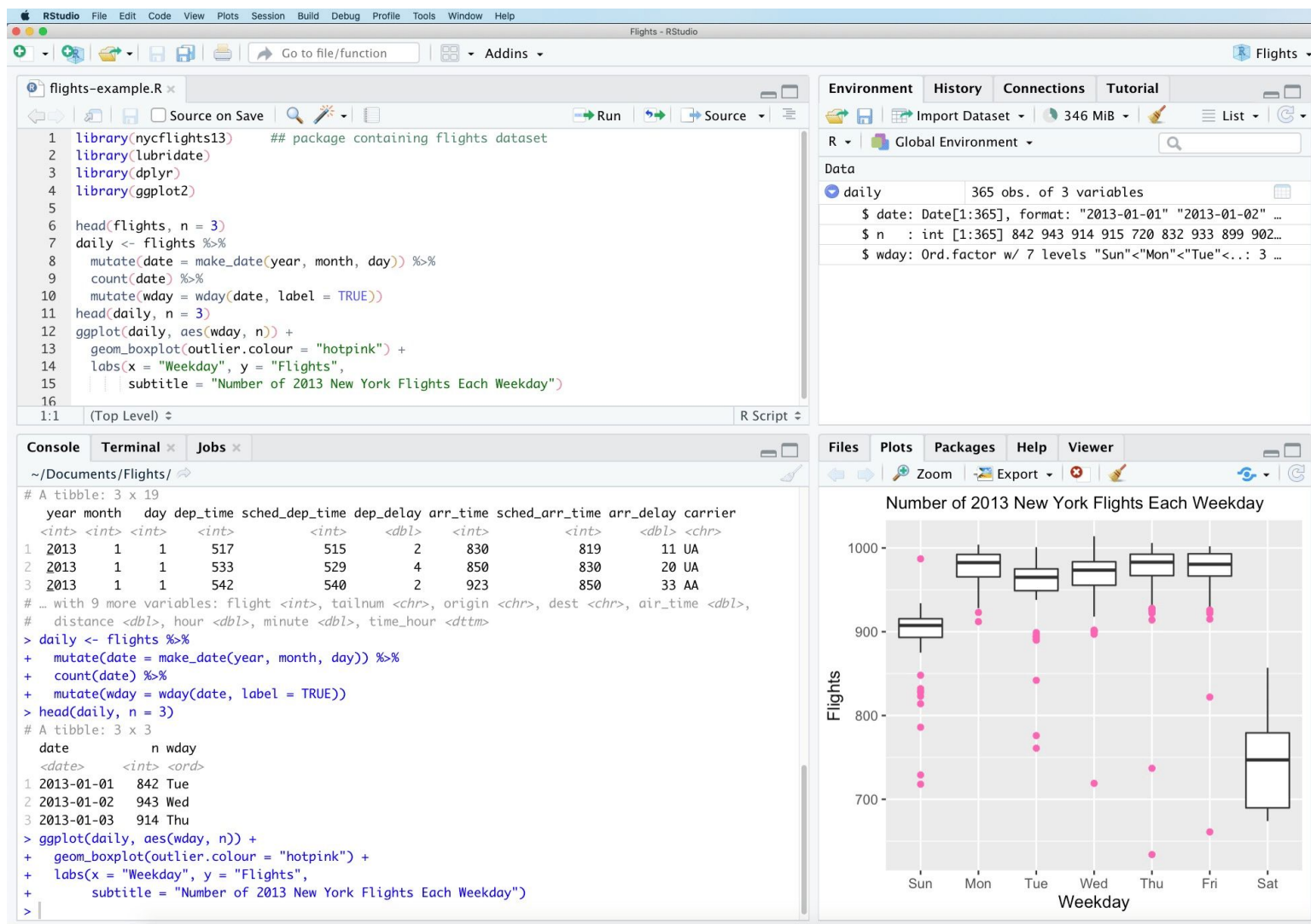


@dsquintana

oliviergimenez.github.io/reproducible-science-workshop

So, how do we actually do reproducible research?

Rstudio:



Rmarkdown Notebooks

settings). Therefore, from this time onward, case counts are likely underestimated and the sequenced virus diversity is not necessarily representative of the virus circulating in the overall population.

BC AB SK MB ON QC NS NB NL

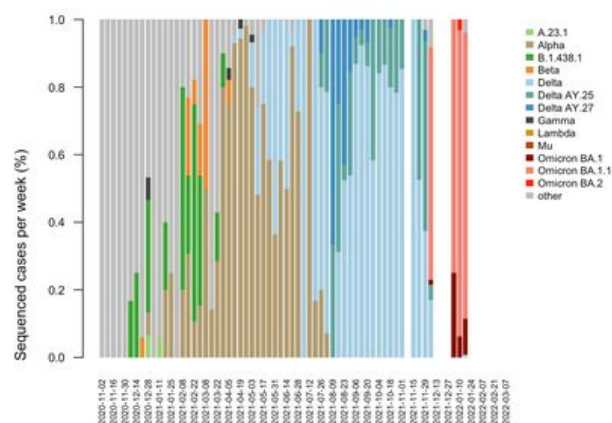
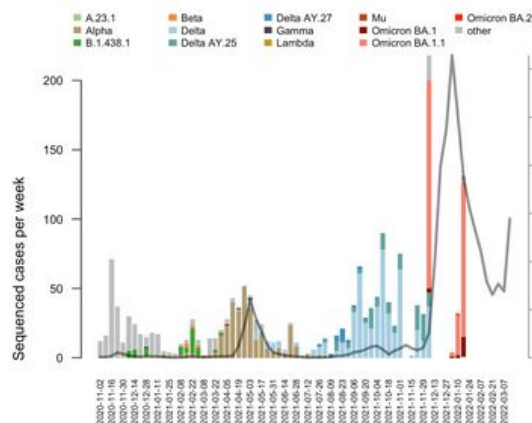
Nova Scotia

Additional up-to-date COVID data for this province can be found here:

<https://experience.arcgis.com/experience/204d6ed723244dfbb763ca3f913c5cad>

Hide

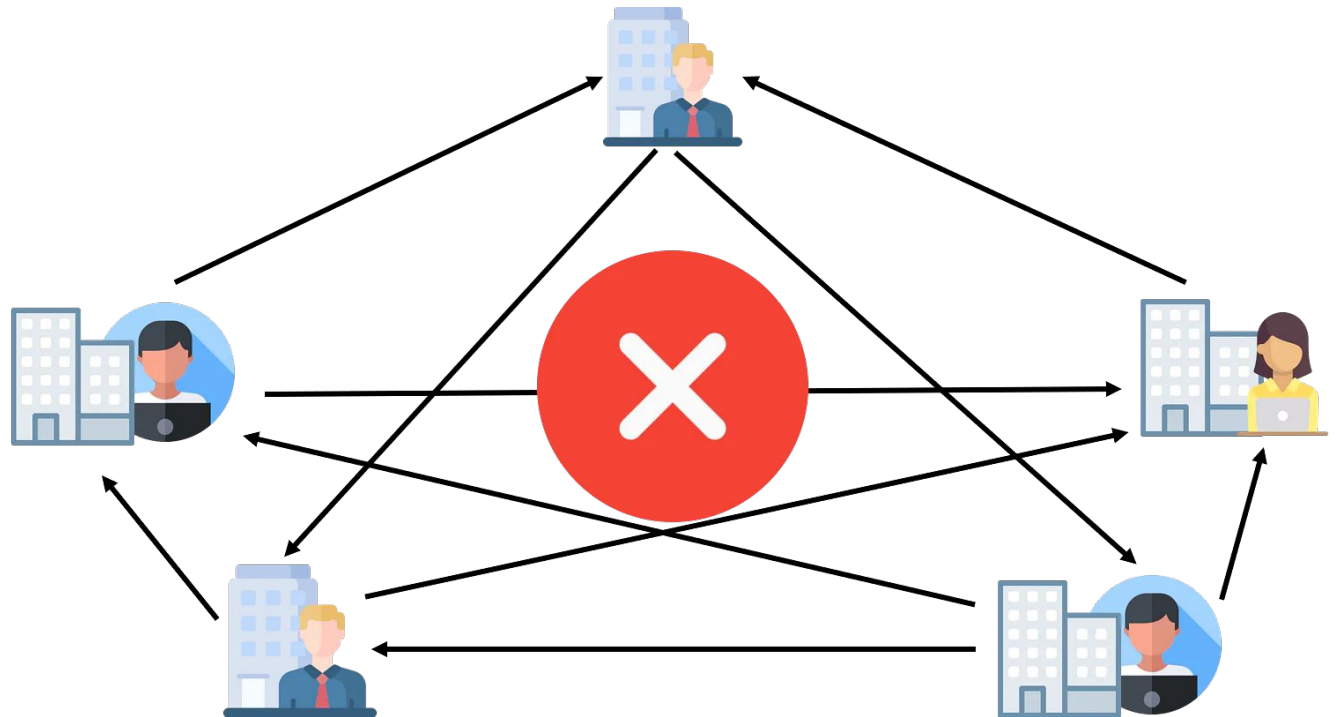
```
plot.variants(region='Nova Scotia')
plot.variants(region='Nova Scotia', scaled=T)
```



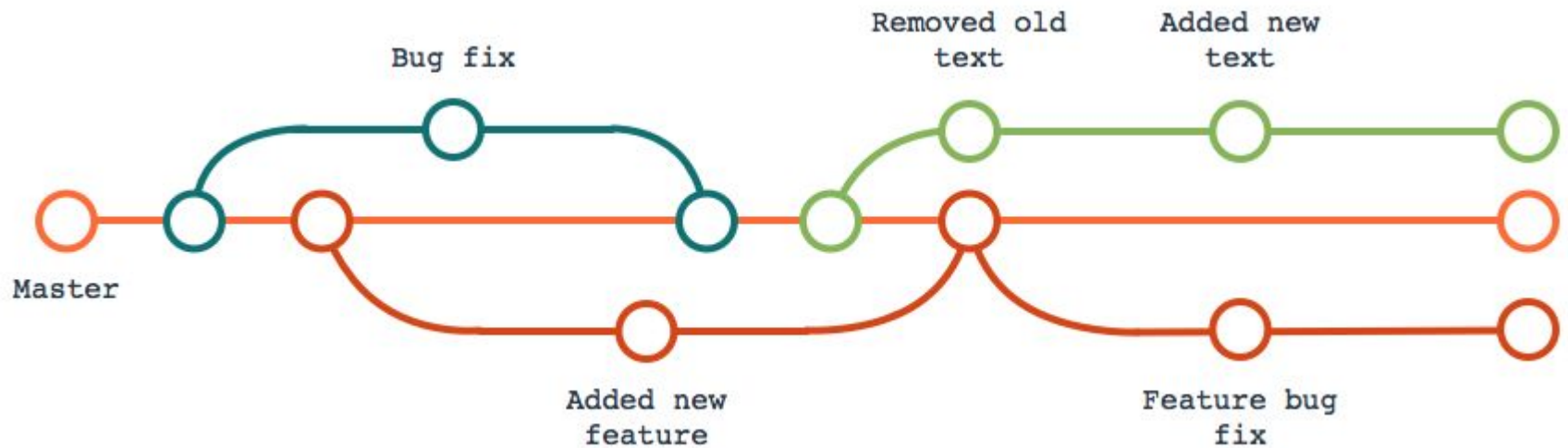
<https://covarr-net.github.io/duotang/duotang.html#>

Version Control

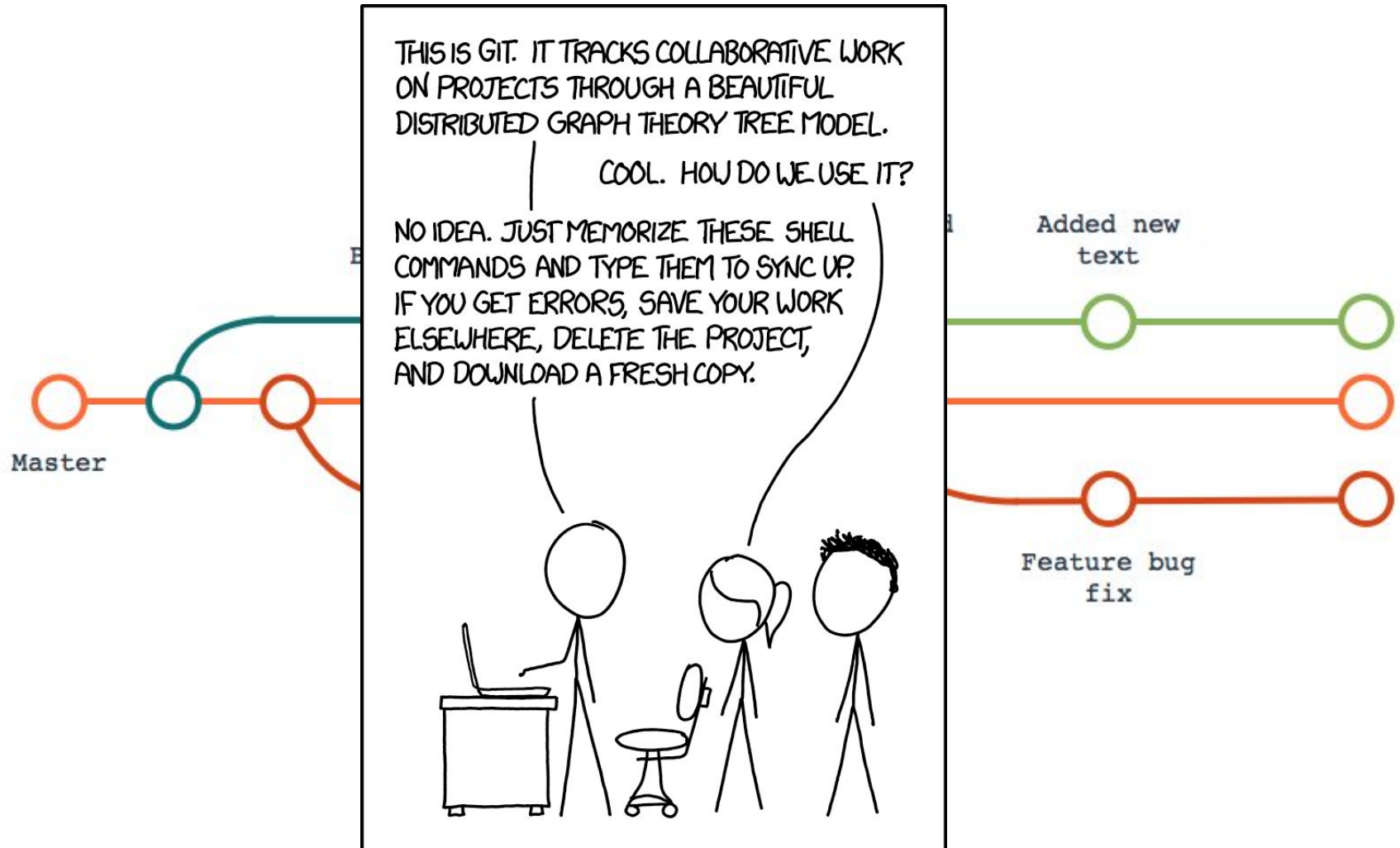
- Backup
- Collaboration
- Organisation



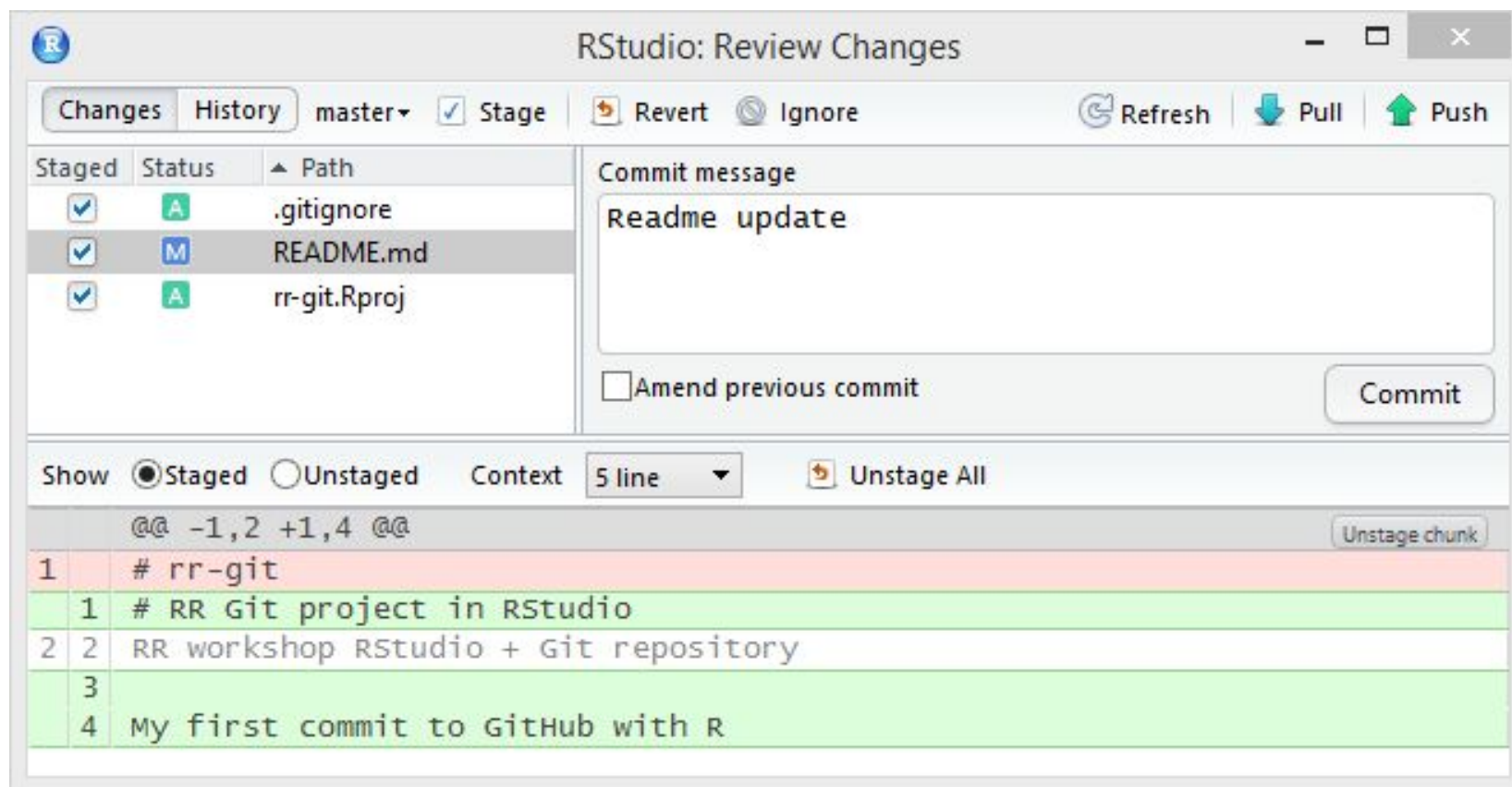
Git Version Control



Git Version Control

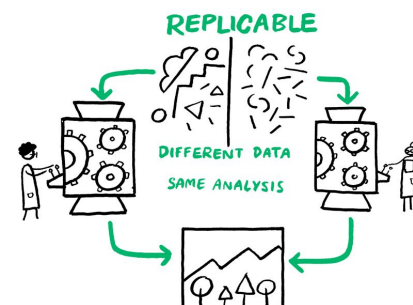
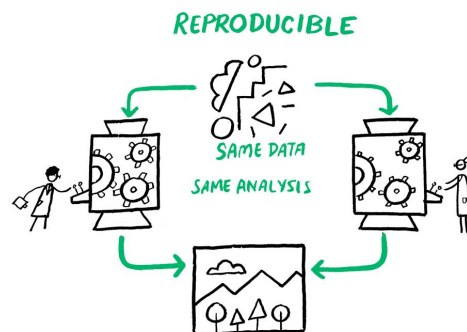


Git integrated into Rstudio!

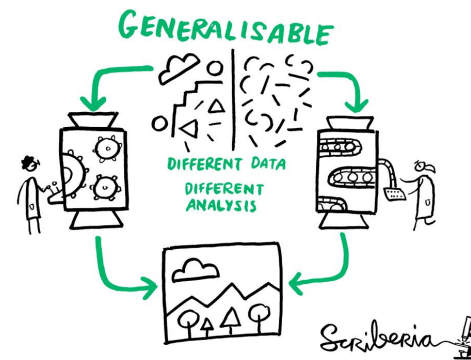
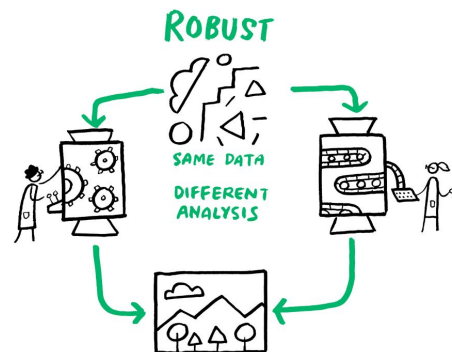


Combine Git+Rmd Notebooks for Reproducibility

1. Add analysis to notebook
2. Add changes to git
3. Find out you made a mistake
4. Revert changes



1. Share notebook with collaborator
2. They make changes
3. You make changes
4. Merge changes into single analysis



Friday's Practical

- Will go over the practical use of R, Rstudio, Rmd Notebooks, Git
- Try and install rstudio, git, and rmarkdown beforehand.
- 1st practical will not contribute to your course grade

Wednesday

- **Reproducibility in machine learning for health research:
Still a ways to go**

[Matthew B. A. McDermott](#) [Shirly Wang](#) [Nikki Marinsek](#) [Rajesh Ranganath](#) [Luca Foschini](#) [Marzyeh Ghassemi](#)

Science Translational Medicine • 24 Mar 2021 • Vol 13, Issue 586 • [DOI: 10.1126/scitranslmed.abb1655](https://doi.org/10.1126/scitranslmed.abb1655)

- **A Beginner's Guide to Conducting Reproducible
Research**

[Jesse M. Alston](#), [Jessica A. Rick](#) First published: 15 January 2021 <https://doi.org/10.1002/bes2.1801>