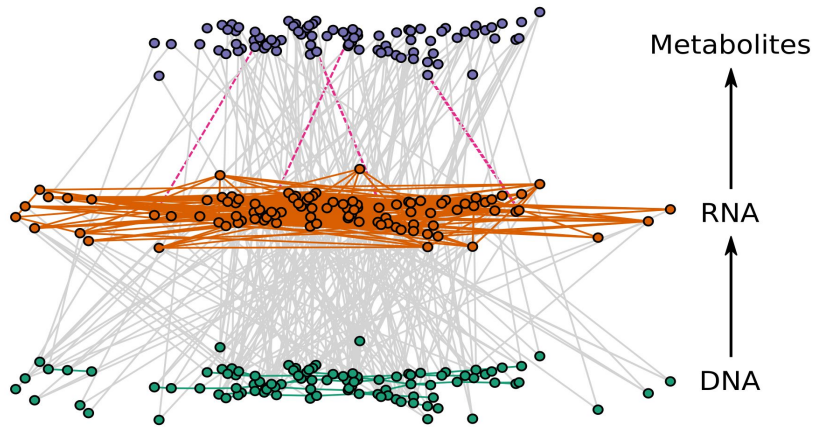


Lecture 0: Introduction to Applied Research in Health Data Science

CSCI6XXX/CHE6XXX/CSCI4148
(CSCI6093)

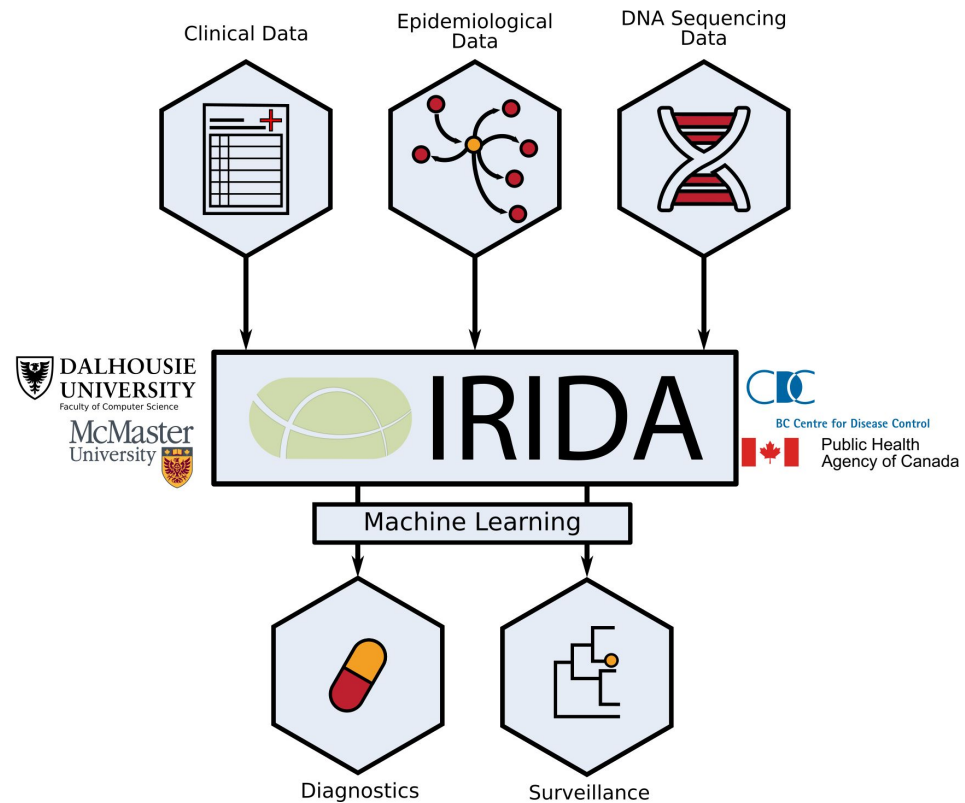
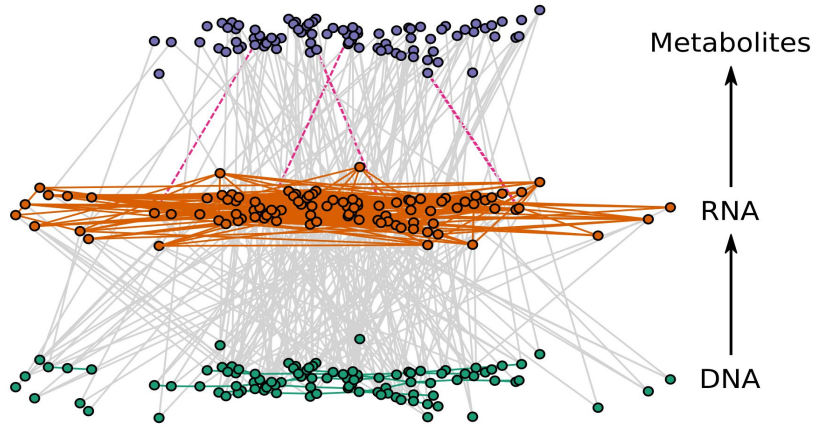
Finlay Maguire (finlay.maguire@dal.ca)

Why am I teaching this course?



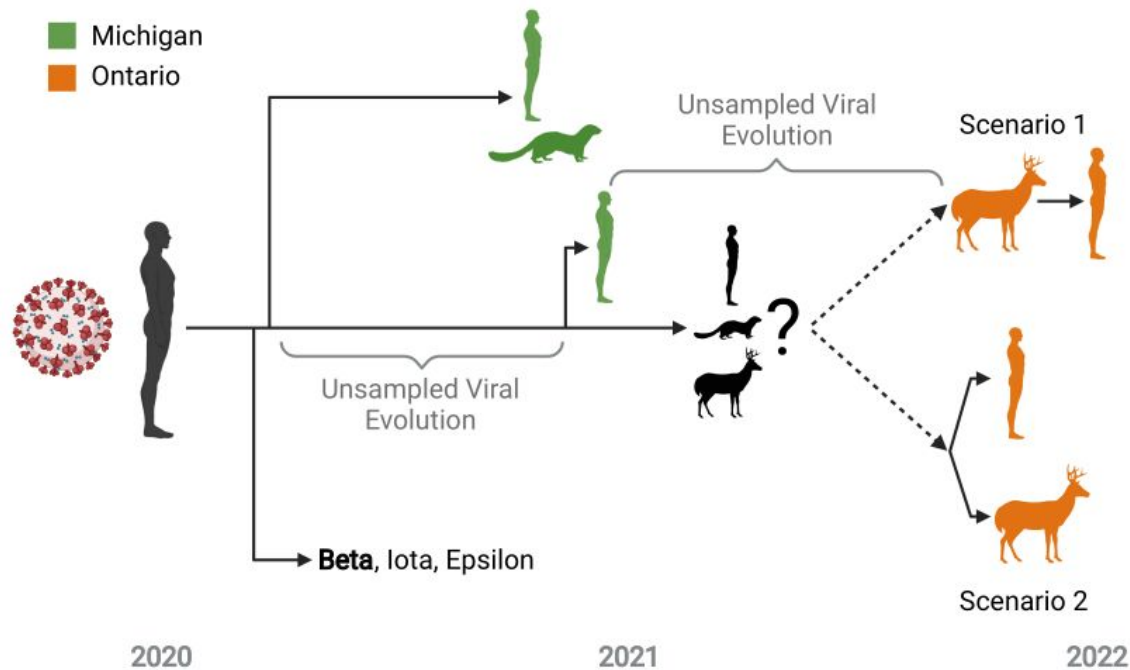
- **PhD (Bioinformatics):** using large noisy datasets to understand how microbial systems and mechanisms evolve.

Why am I teaching this course?



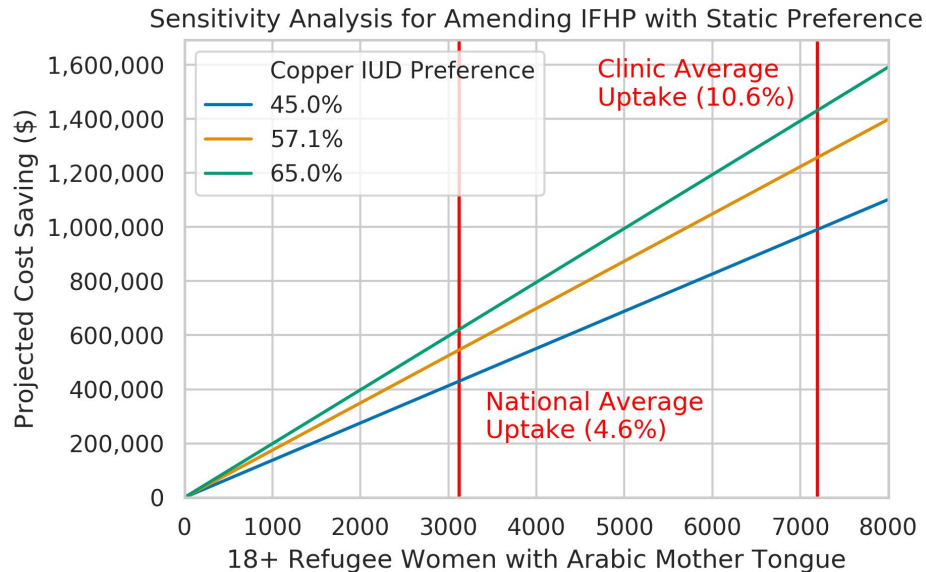
- **PhD (Bioinformatics)**: using large noisy datasets to understand how microbial systems and mechanisms evolve.
- **Postdoc (Genomic Epidemiology)**: using large noisy datasets to better diagnose, track and predict infectious diseases.

Why am I teaching this course?

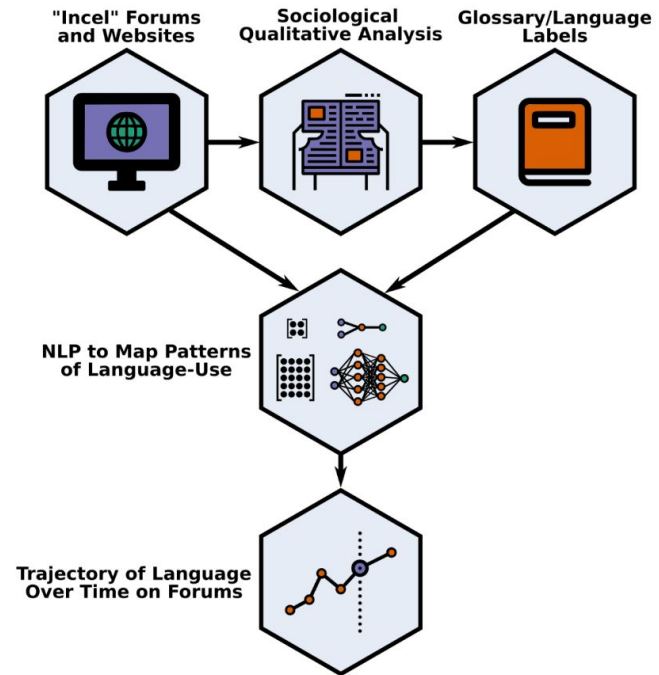


- **Research group:** using large noisy datasets:
 - Genomic epidemiology of infectious disease: **SARS-CoV-2, AMR**

Why am I teaching this course?



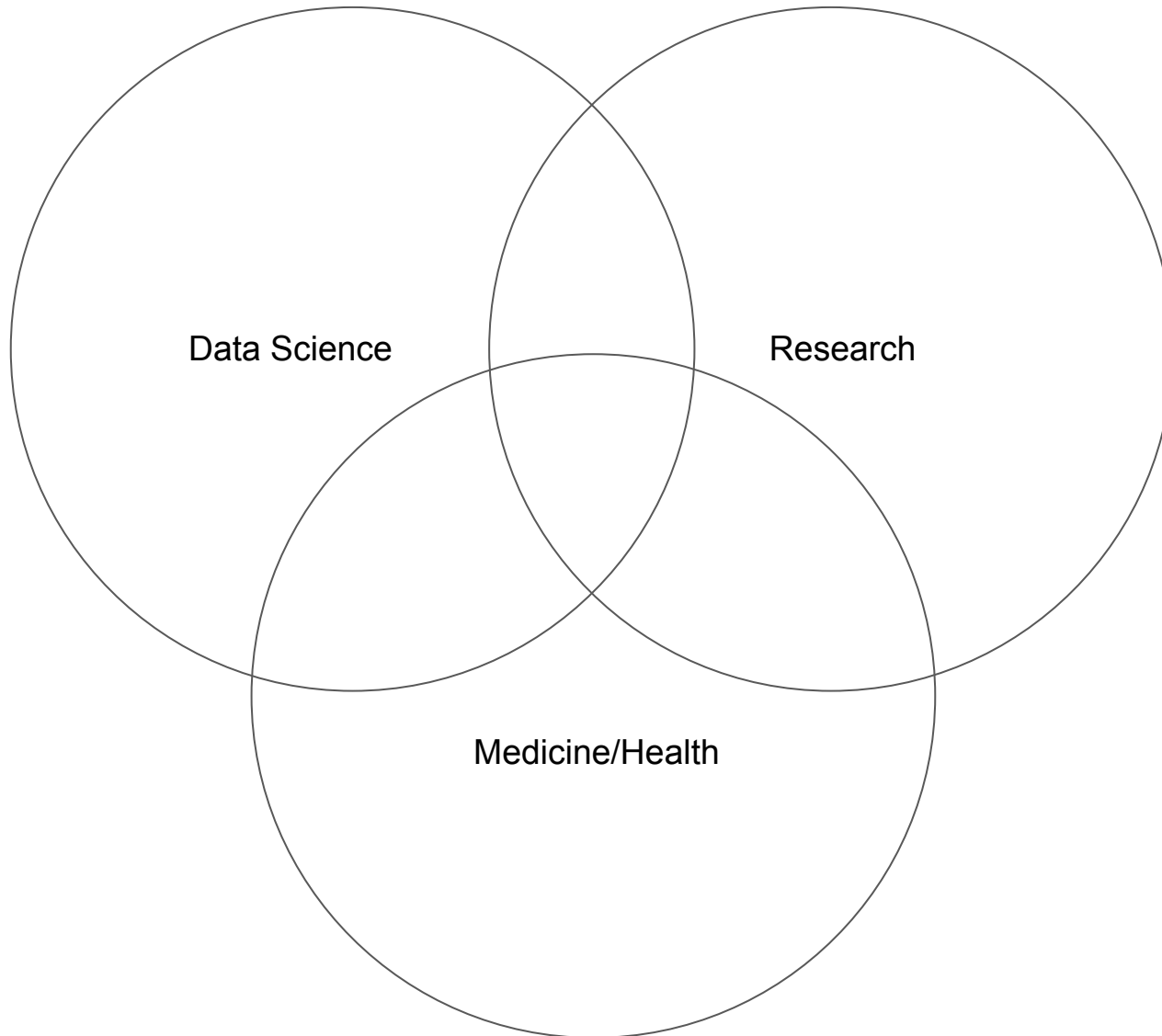
Modelling "Incel" Online Radicalisation via NLP



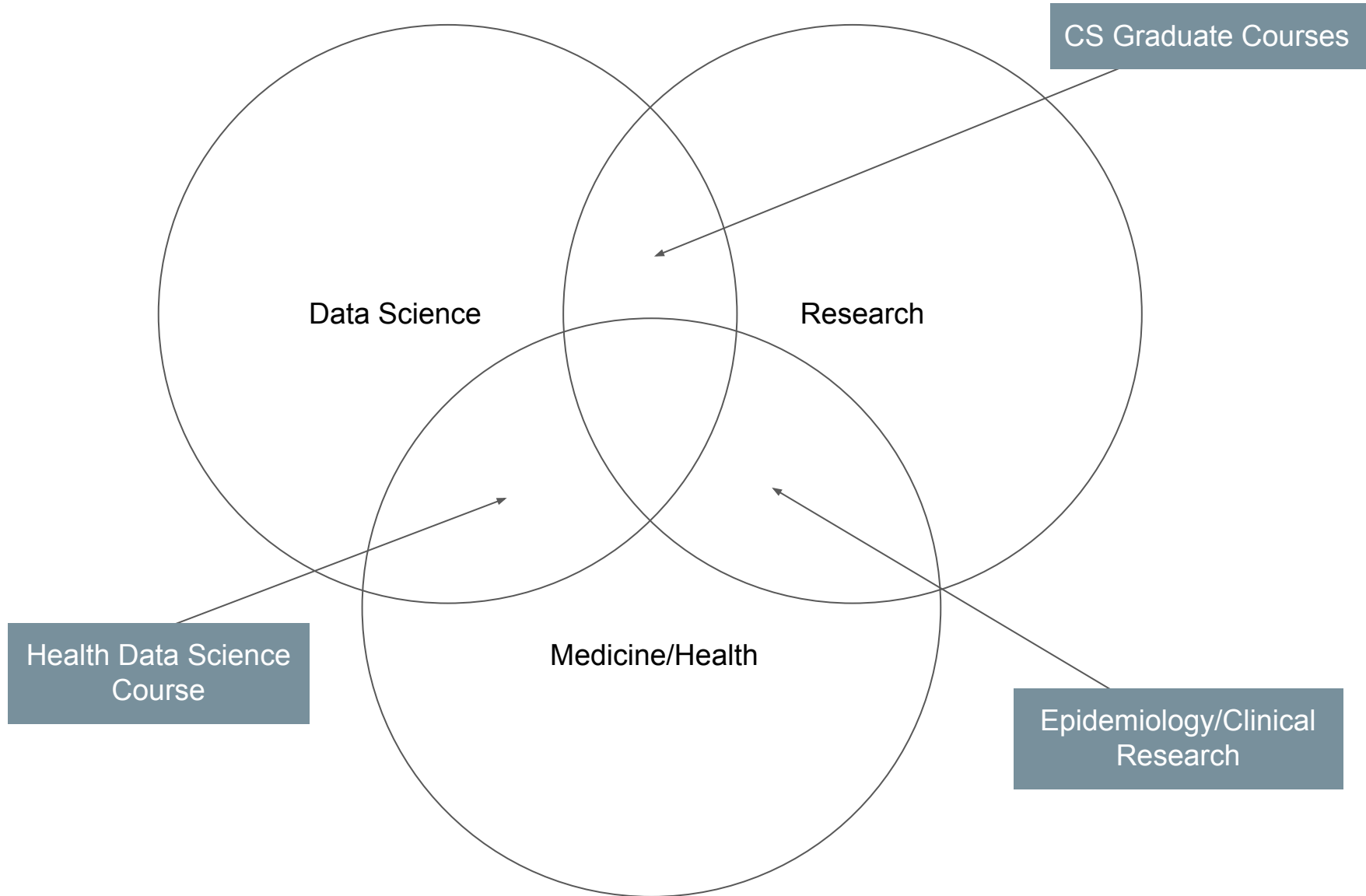
- **Research group:** using large noisy datasets:
 - Genomic epidemiology of infectious disease: **SARS-CoV-2, AMR**
 - Collaborations on socially/health focused problems: **refugee health, incel radicalisation, health inequality**

Overview of course

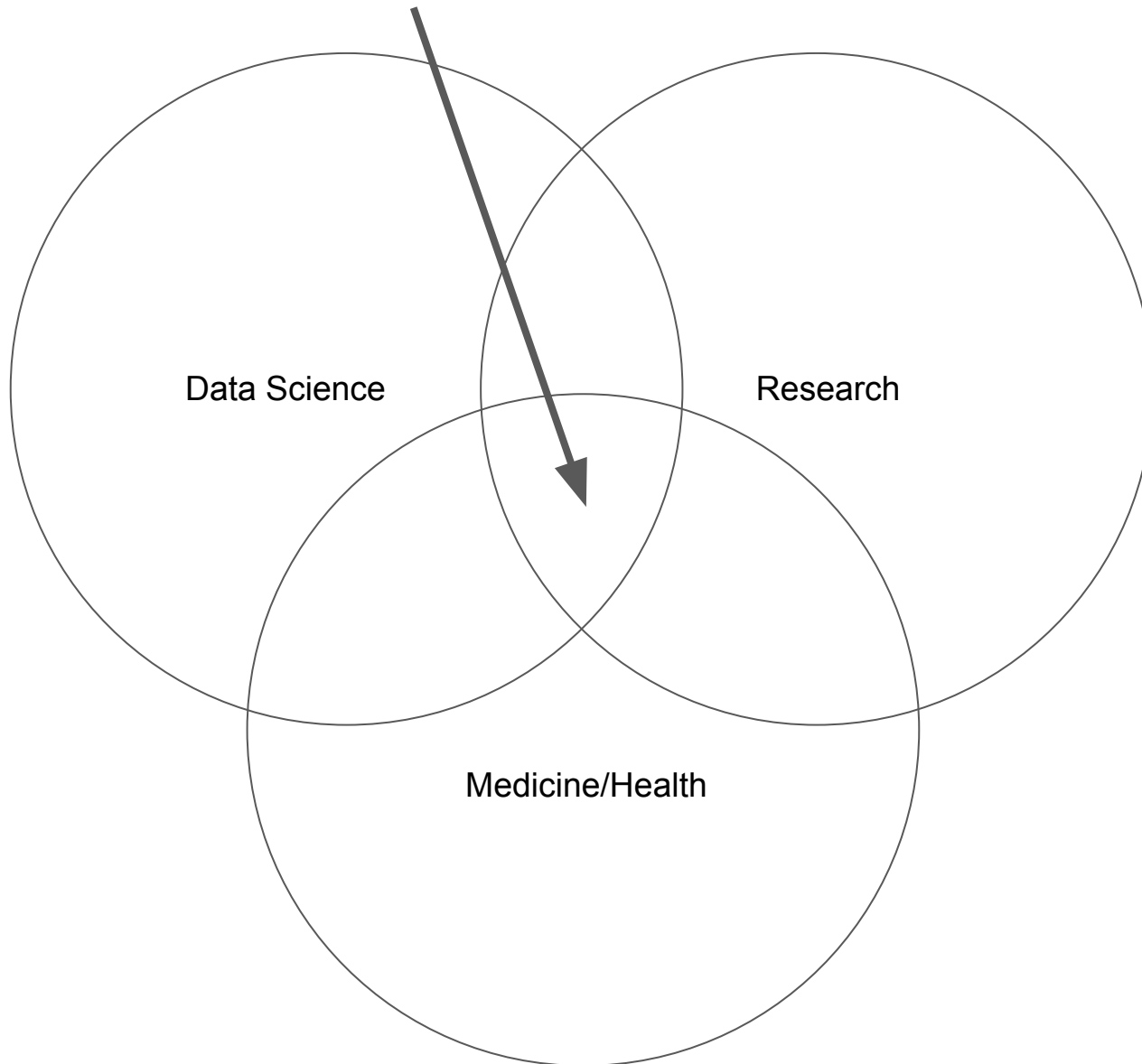
Applied Research in Health Data Science



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).

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

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.

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.

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.

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*

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*

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*

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

Overview of data types & analysis methods:

- **Lectures** (Monday/Wednesday)

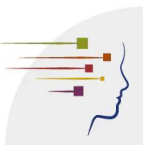
Course Structure

Overview of data types & analysis methods:

- **Lectures** (Monday/Wednesday)
- **Practical Exercises** (Friday/Monday)

Assessment: Submission of Practical Exercise Due the day before **following practical** (10% x 4)

```
dens <- density(data, n = npts)
dx <- dens$x
dy <- dens$y
if(add == TRUE)
  plot(0., 0., main,
       ylab)
if(orientati == yst)
  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)
```



Course Structure

Overview of data types & analysis methods:

- **Lectures** (Monday/Wednesday)
- **Practical Exercises** (Friday/Monday)

Assessment: Submission of Practical Exercise Due the day before **following practical** (10% x 4)

Research in health data science:

- **Journal Club** (Wednesday/Friday)

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

Assessment:

Paper presentation (10%)

Participation in discussion (10%)

```
dens <- density(data, n = npts)
dx <- dens$x
dy <- dens$y
if(add == TRUE)
  plot(0., 0., main,
       ylab,
       if(orientati == yst))
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)
```



Course Structure

Overview of data types & analysis methods:

- **Lectures** (Monday/Wednesday)
- **Practical Exercises** (Friday/Monday)

Assessment: Submission of Practical Exercise Due the day before **following practical** (10% x 4)

```
dens <- density(data, n = npts)
dx <- dens$x
dy <- dens$y
if(add == TRUE)
  plot(0., 0., main,
       ylab,
       if(orientati == "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>

Research in health data science:

- **Journal Club** (Wednesday/Friday)

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

Assessment:

Paper presentation (10%)

Participation in discussion (10%)

Development of a research proposal:

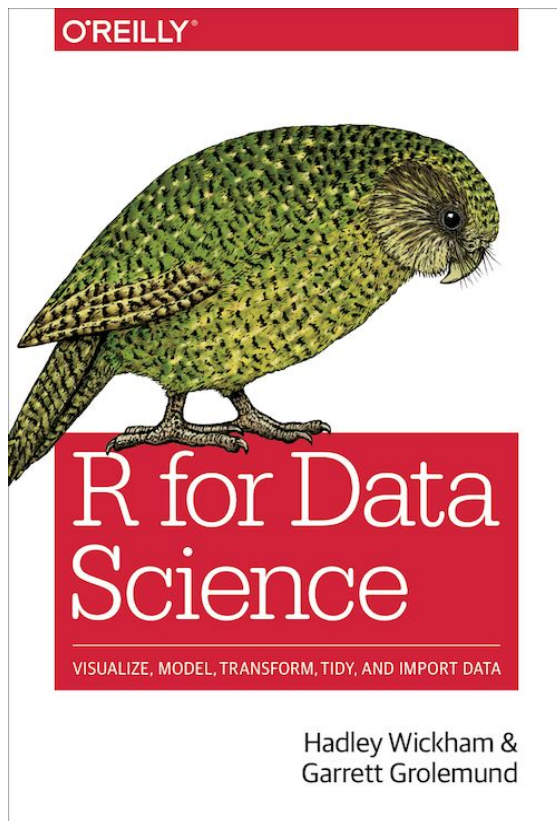
- **Class** (Wednesday/Friday)

Assessment:

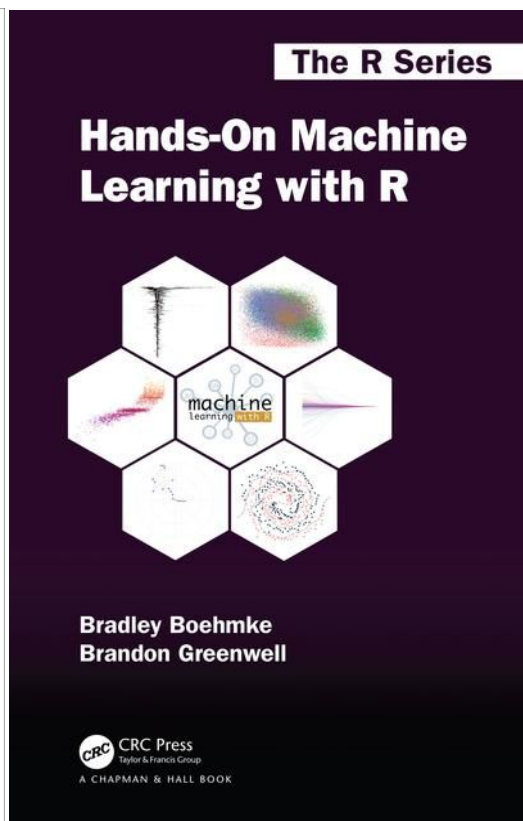
Presentation **last full week of class** (20%)

Submitted **final day of class** (20%)

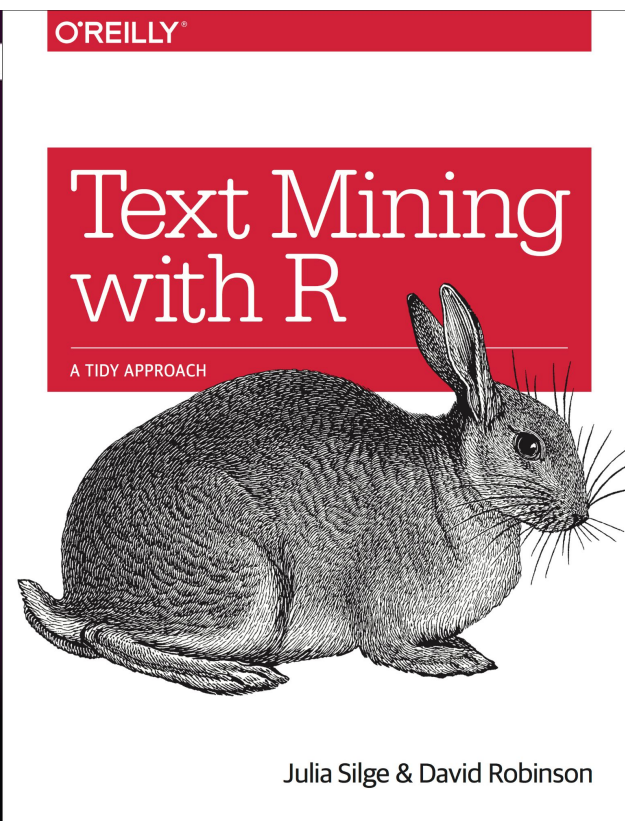
Course Materials



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



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



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

Course Website



The screenshot shows the top portion of a course website. The header is a dark blue banner with a white hexagonal logo on the left containing a calendar icon with a red cross. To the right of the logo, the text reads "Dalhousie University" in a small font, followed by "Applied Research in Health Data Science" in a larger, bold font, and "Summer 2021-2022" in a smaller font below it. A navigation menu is located at the bottom of the banner, featuring icons and labels for "HOME", "SCHEDULE", "LECTURES", "PRACTICALS", "PROPOSAL", and "LITERATURE". Below the banner, the page title "Applied Research in Health Data Science / Summer 2021-2022" is displayed in a purple font. A yellow box contains the heading "Updates" and a single bullet point: "New Lecture is up: Lecture 0 - Introduction to health data science [slides]".

Dalhousie University
Applied Research in Health Data Science
Summer 2021-2022

HOME SCHEDULE LECTURES PRACTICALS PROPOSAL LITERATURE

Applied Research in Health Data Science / Summer 2021-2022

Updates

- New Lecture is up: Lecture 0 - Introduction to health data science [slides]

https://maguire-lab.github.io/health_data_science_research/

Course Website

The screenshot shows the top navigation bar of the course website. It features the Dalhousie University logo on the left, followed by the text "Dalhousie University" and "Applied Research in Health Data Science Summer 2021-2022". To the right of the text are navigation icons for HOME, SCHEDULE, LECTURES, PRACTICALS, PROPOSAL, and LITERATURE. Below the navigation bar, the page title "Applied Research in Health Data Science / Summer 2021-2022" is displayed. Underneath, there is a yellow box titled "Updates" containing a single bullet point: "New Lecture is up: Lecture 0 - Introduction to health data science [slides]".

https://maguire-lab.github.io/health_data_science_research/

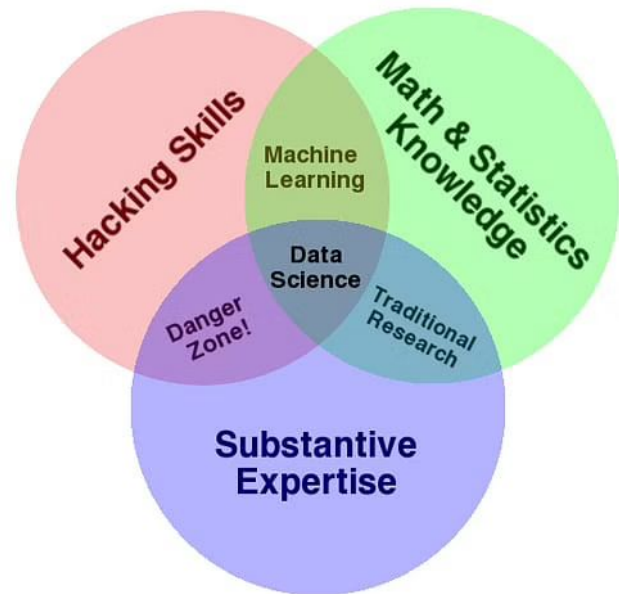
The screenshot shows a course page in Brightspace. At the top, the course title "CSCI6903 - Spec Grad Top Comp Sci (Sec 1..." is visible. Below the title is a navigation menu with options: Course Home, Content, Discussions, Assessments, My Tools, Help, and Course Admin. A large banner image of a vineyard is displayed with the text "CSCI6903 - Spec Grad Top Comp Sci (Sec 1) - 2022 Summer". Below the banner, there are two sections: "Announcements" and "Updates". The "Announcements" section contains a post titled "Course Website" posted on May 4, 2022, at 5:15 PM, with the text "I'm just putting together the main course website which will be hosted at: [https://maguire-lab.github.io](\"https://maguire-lab.github.io\")". The "Updates" section shows a message: "There are no current updates for CSCI6903 - Spec Grad Top Comp Sci (Sec 1) - 2022 Summer".

Grades/Submissions:

<https://dal.brightspace.com/d2l/home/221757>

What is ~~health~~ data science?

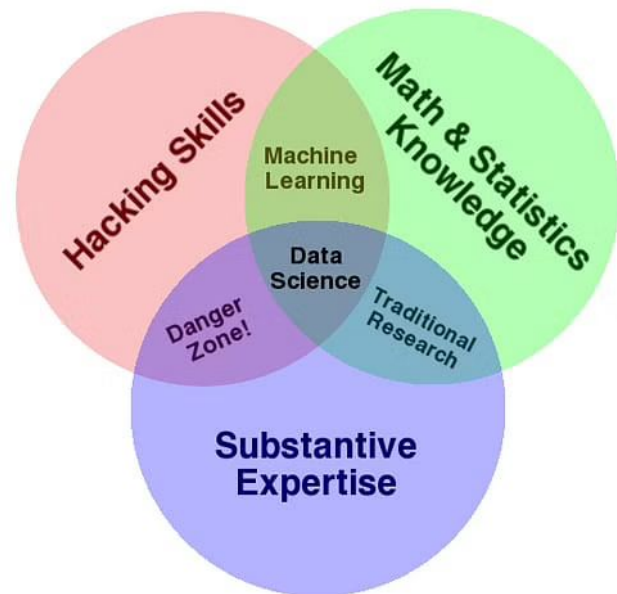
Data Science: *Using Data to Better Understand Things in the Real World*



<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

Data Science: *Using Data to Better Understand Things in the Real World*

A range of partial and totally overlapping terms:

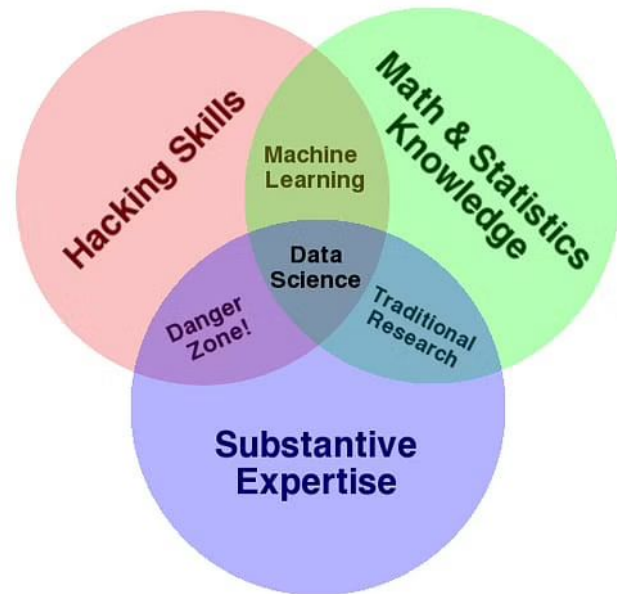


<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

Data Science: *Using Data to Better Understand Things in the Real World*

A range of partial and totally overlapping terms:

- Data Analytics
- Data Engineering
- Data Mining
- {Health,Bio,Medical}Informatics
- Database Analysis
- Business Intelligence
- Epidemiology
- Statistics
- Machine Learning
- Pattern Recognition
- Predictive Analytics
- Quantitative Researcher
- Scientist
- Analyst



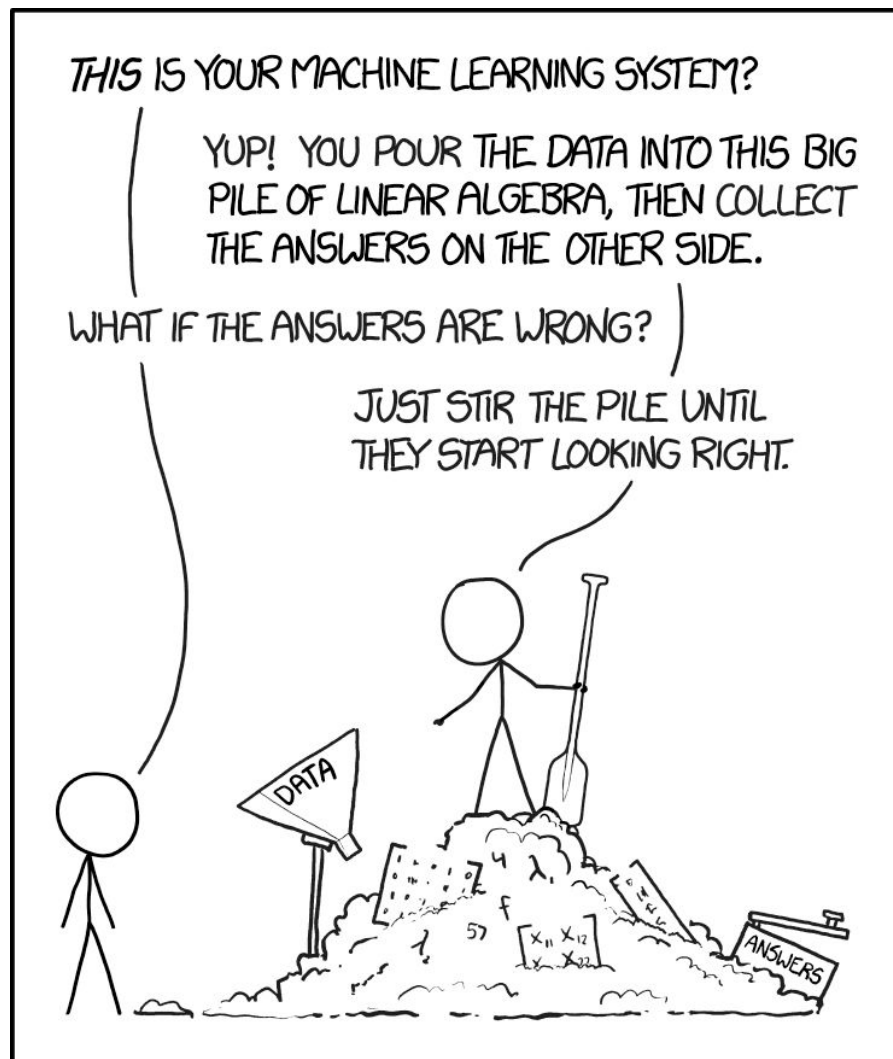
<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

So, it is just statistics?

Data Science (& Machine Learning): re-branded statistics

Pitfalls (can be):

- Less rigorous/principled
- Prone to reinventing the wheel



Data Science (& Machine Learning): re-branded statistics

Pitfalls (can be):

- Less rigorous/principled
- Prone to reinventing the wheel

Benefits (can be):

- More flexible
- Less prescriptive/intimidating



Data Science (& Machine Learning): re-branded statistics

Pitfalls (can be):

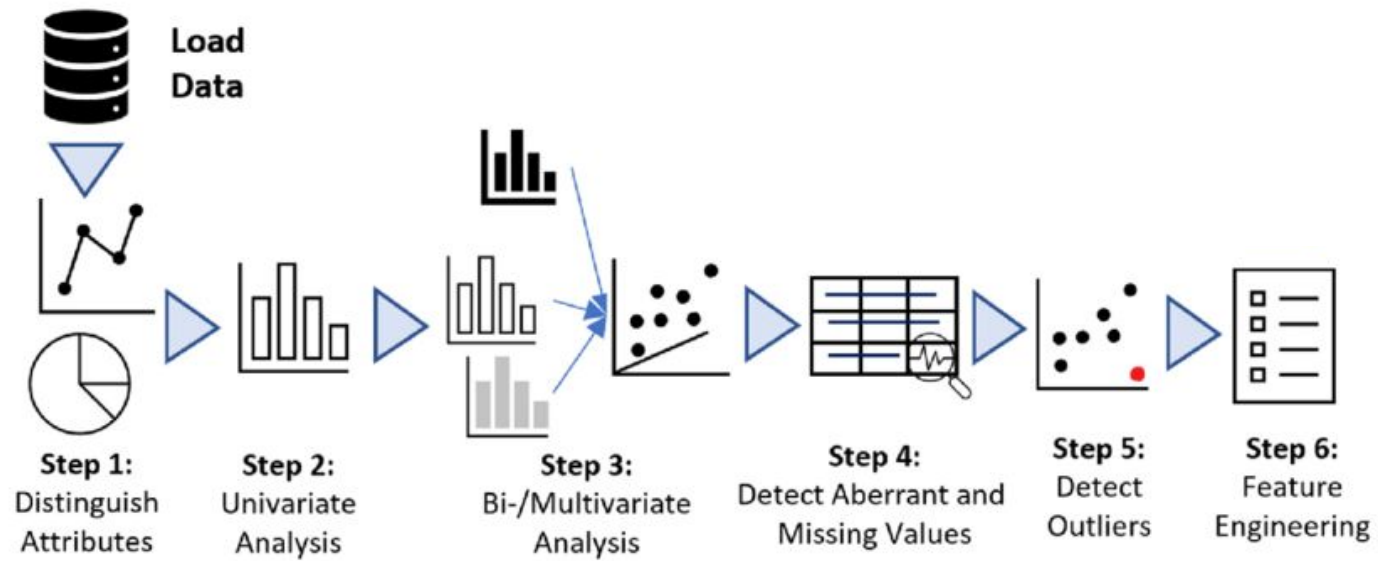
- Less rigorous/principled
- Prone to reinventing the wheel

Benefits (can be):

- More flexible
- Less prescriptive/intimidating



Data science centers exploratory data analysis

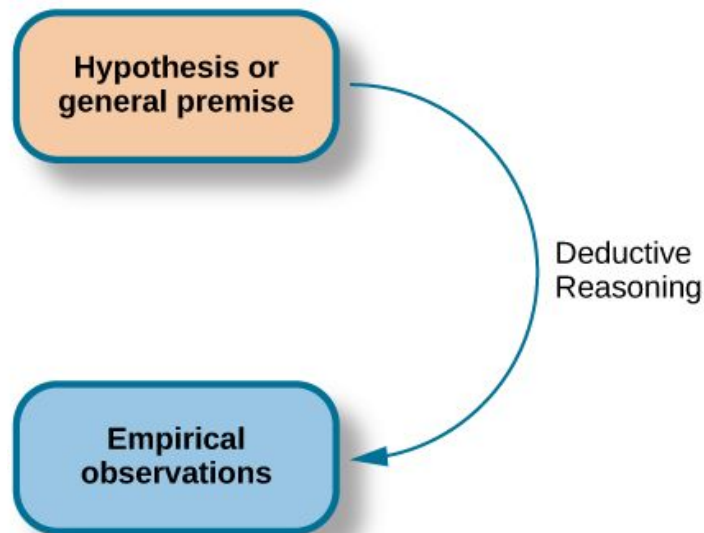


Data science supports inductive approaches

Data science supports inductive approaches

Deductive:

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



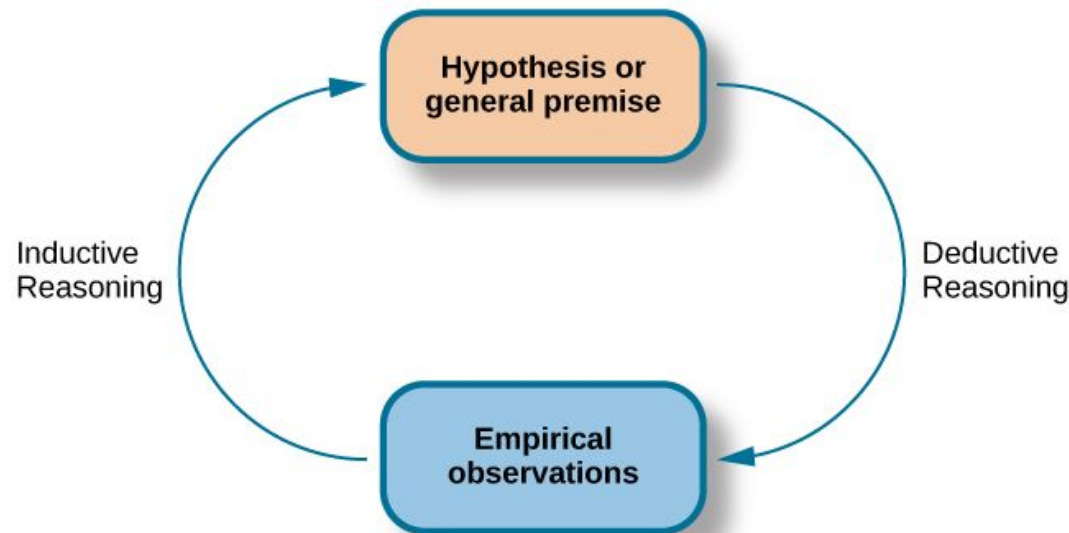
Data science supports inductive approaches

Deductive:

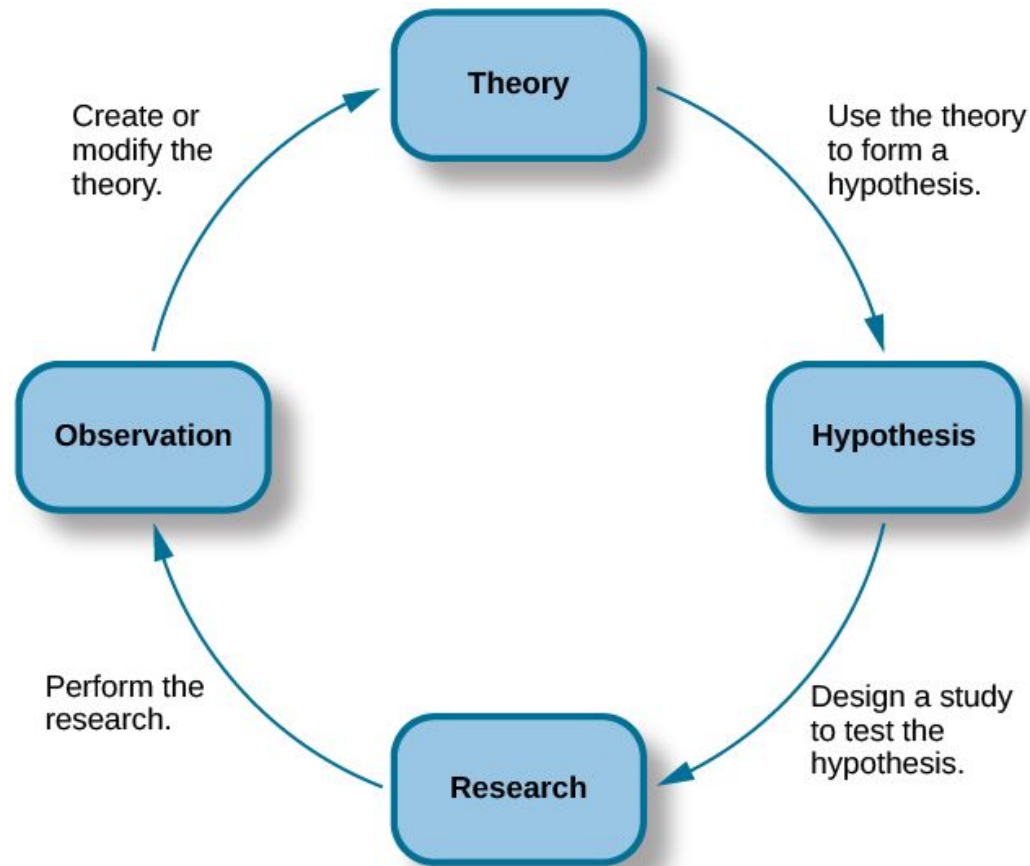
- “Condition X, causes Y”
- Collect data
- Perform frequentist statistical test
- Reject or confirm null hypothesis

Inductive:

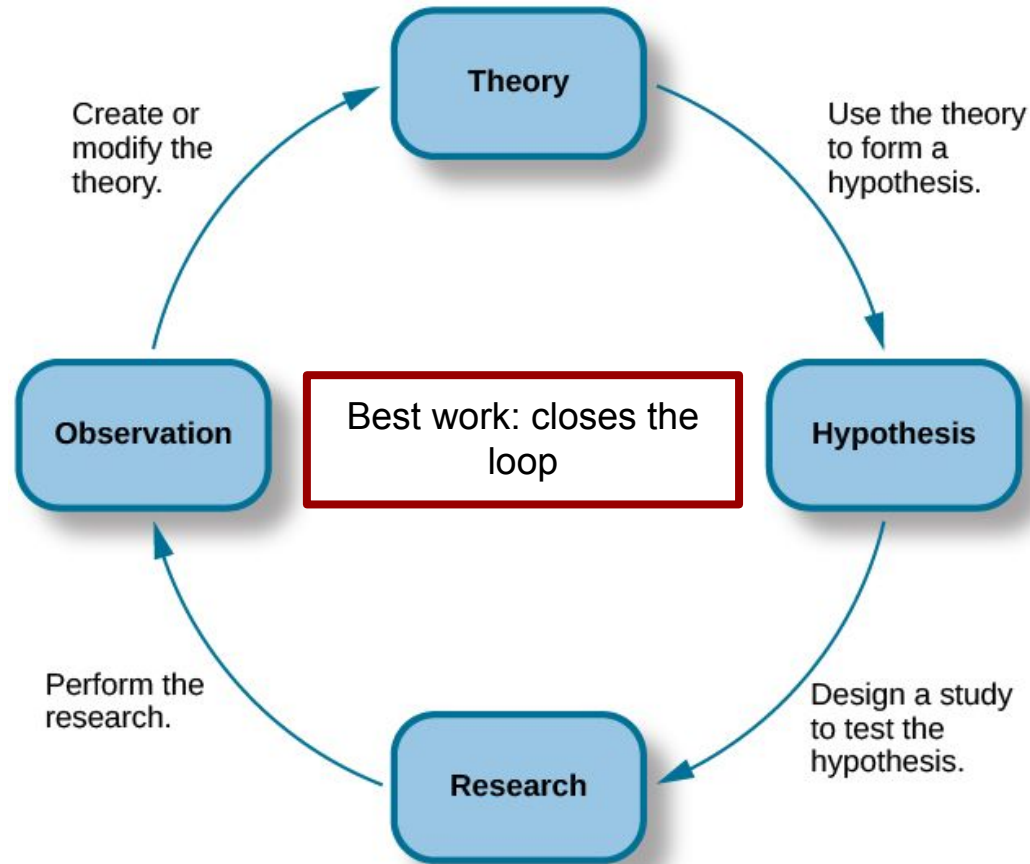
- Collect data
- Identify patterns in the data
- Observe X and Y seem connected somehow
- Quantify strength of association



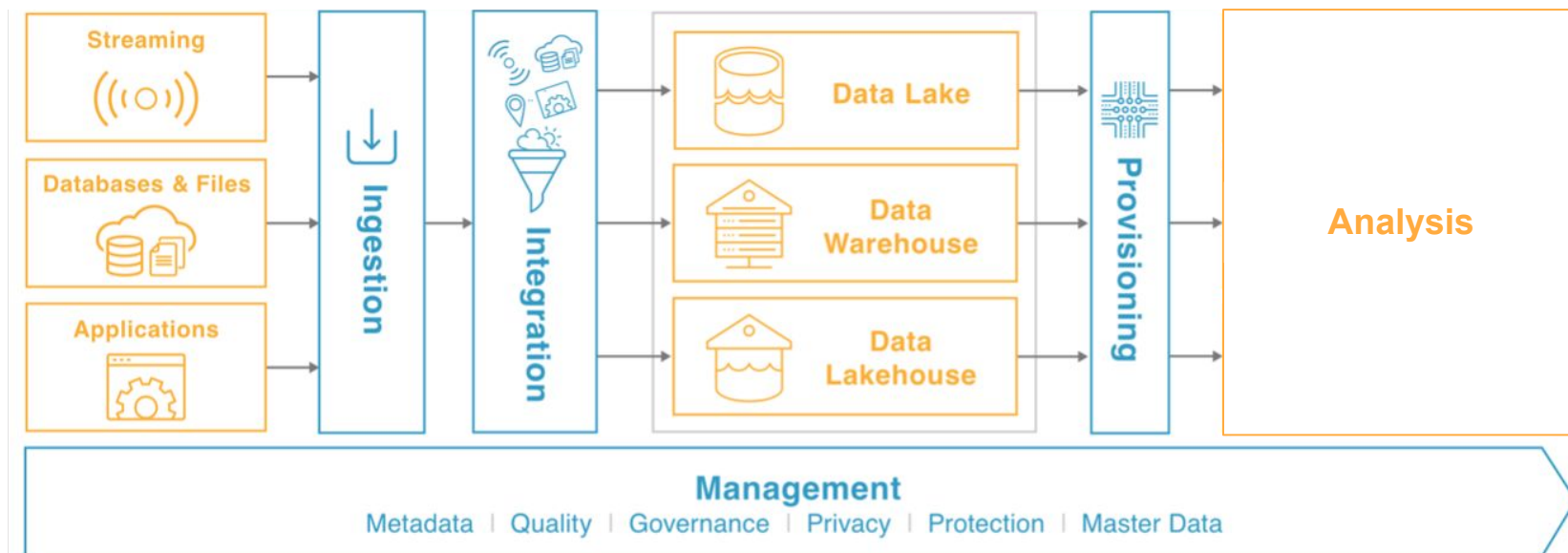
Data science is more realistic



Data science is more realistic

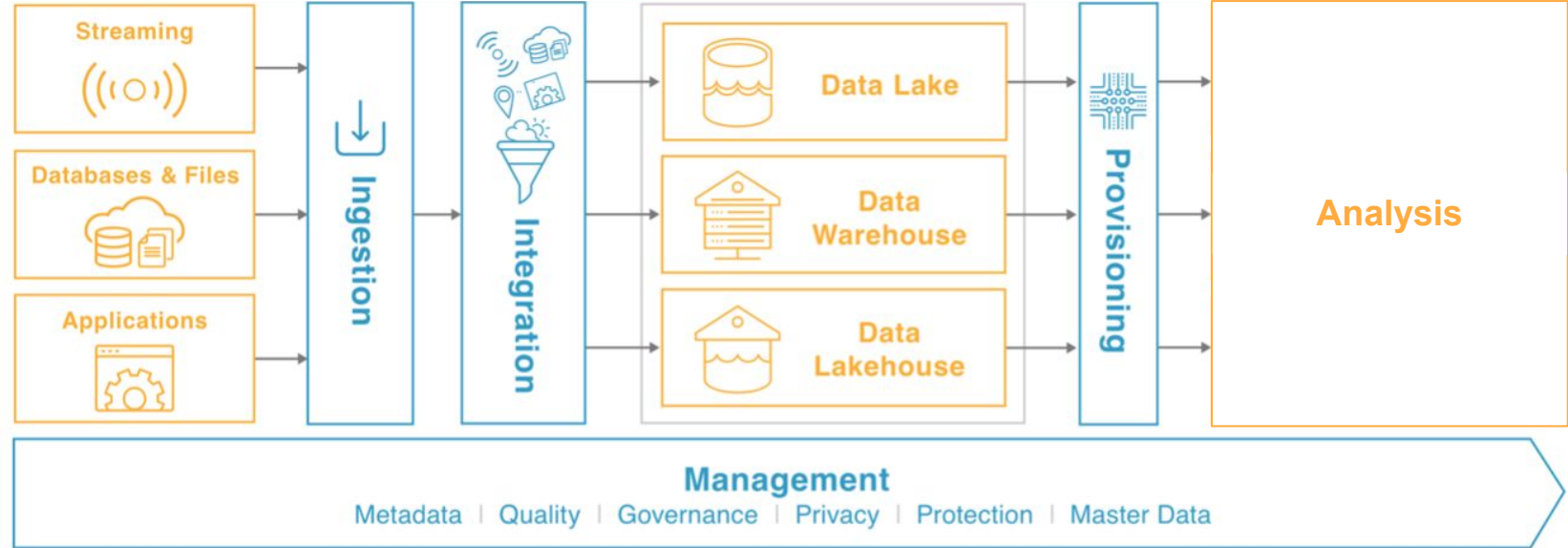


Data science is integrated into a data ecosystem



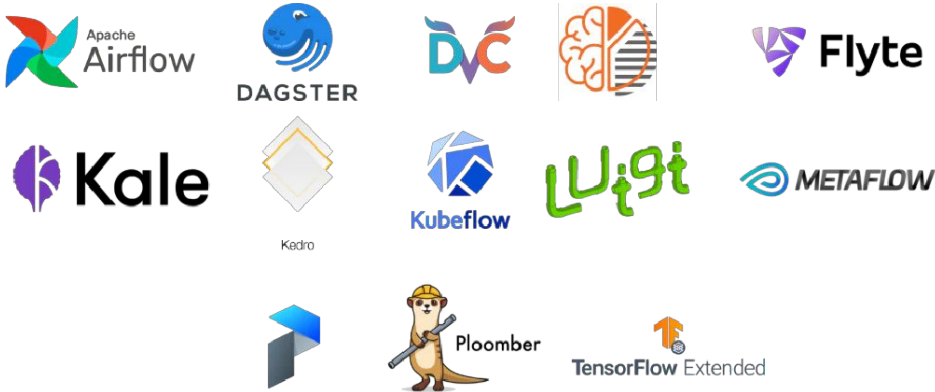
<https://www.2ndwatch.com/blog/what-is-a-data-pipeline-and-how-to-build-one/>

Data science is integrated into a data ecosystem



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

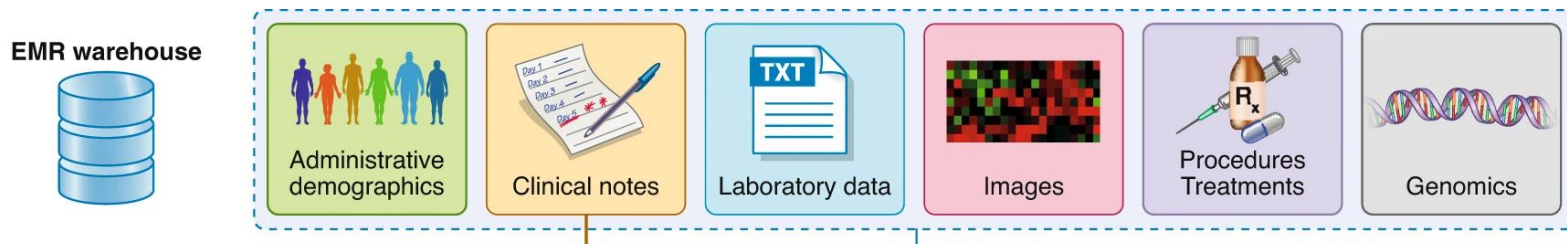
Some Open-Source Orchestration Tools:



<https://ploomber.io/blog/survey/>

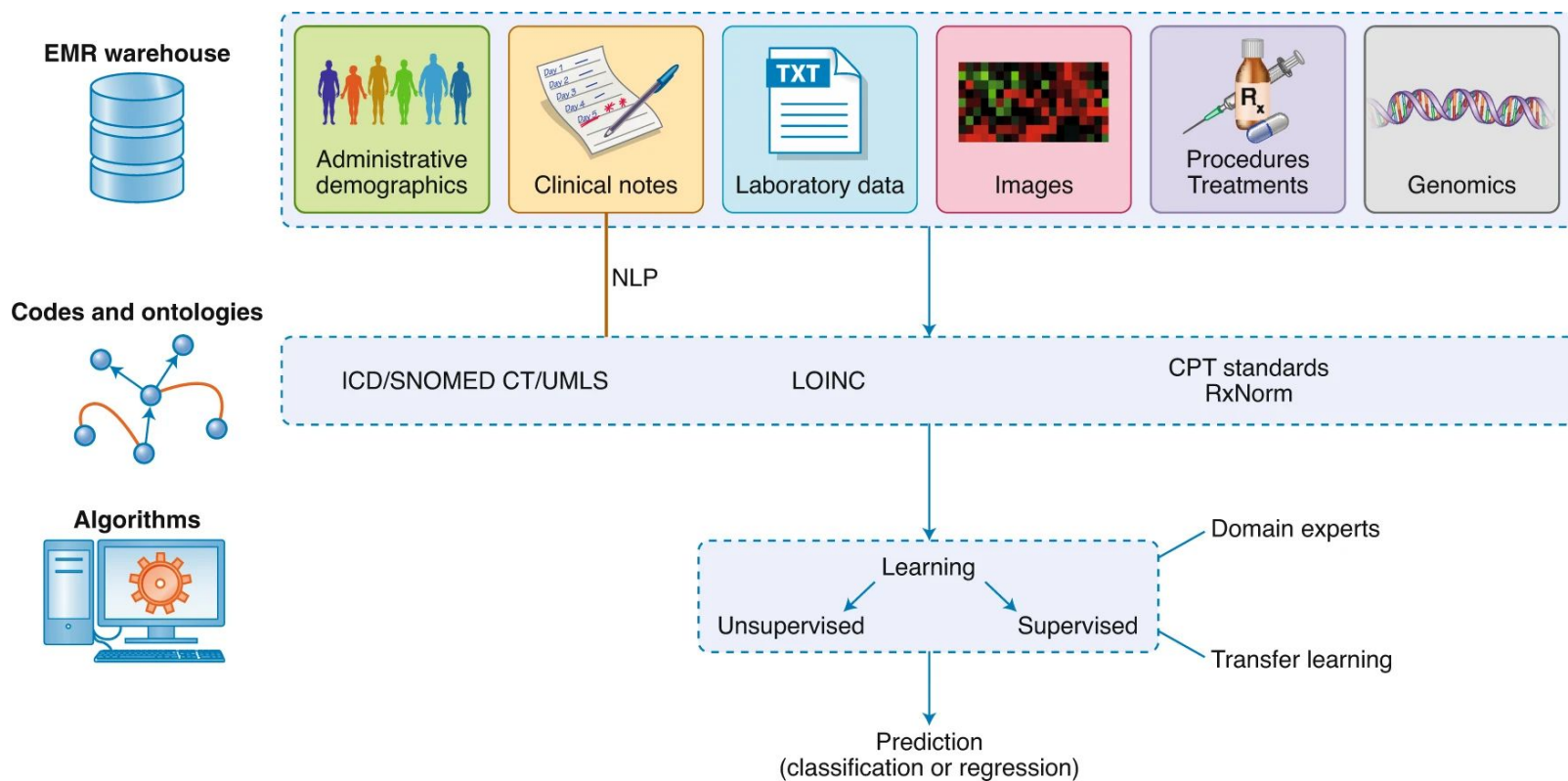
OK, what is **Health Data Science**?

Data Science applied to Health Data



Why “health data” instead of “medical data”: health encompasses medical (**contentious**)

Data Science applied to Health Data



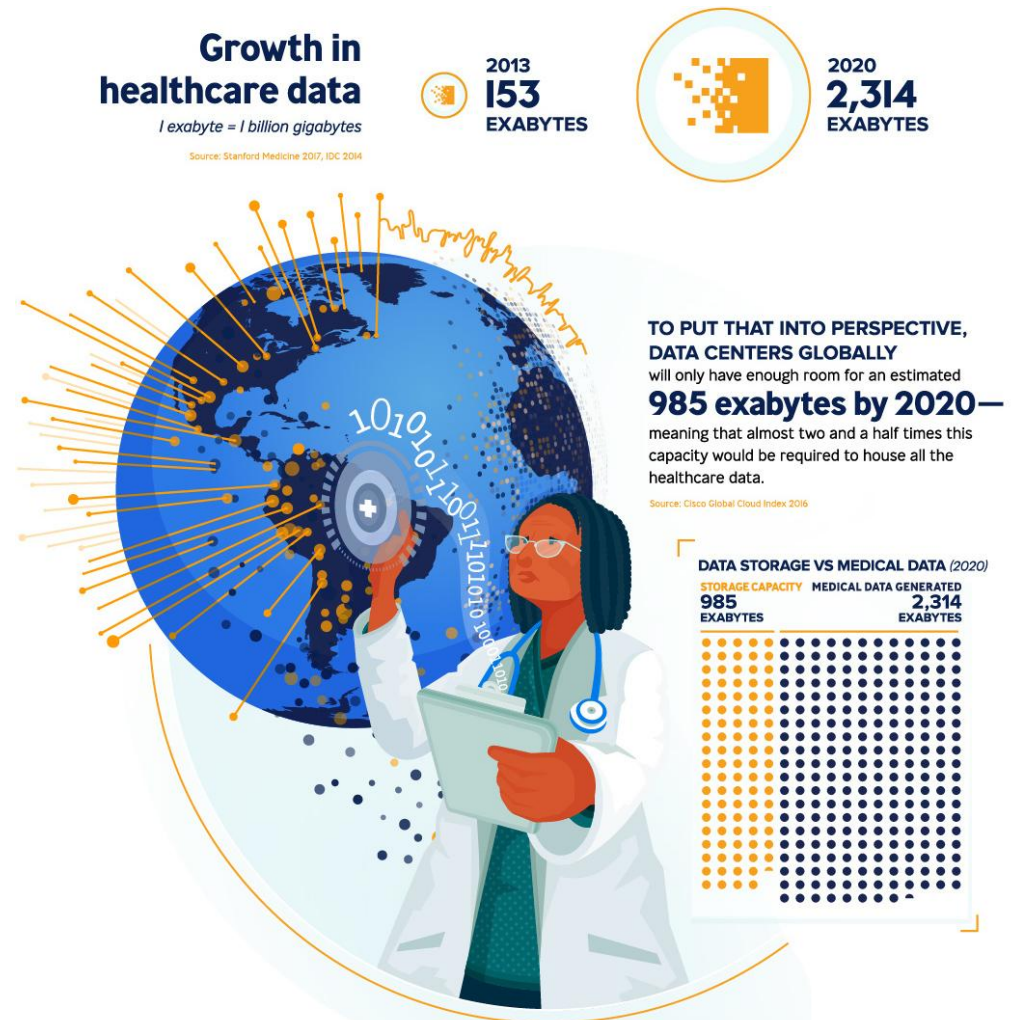
<https://www.nature.com/articles/s41588-020-0698-y/figures/2>

Why “health data” instead of “medical data”: health encompasses medical (**contentious**)

Opportunity of Health Data Science

Benefits (and pitfalls!) of data science in general combined with:

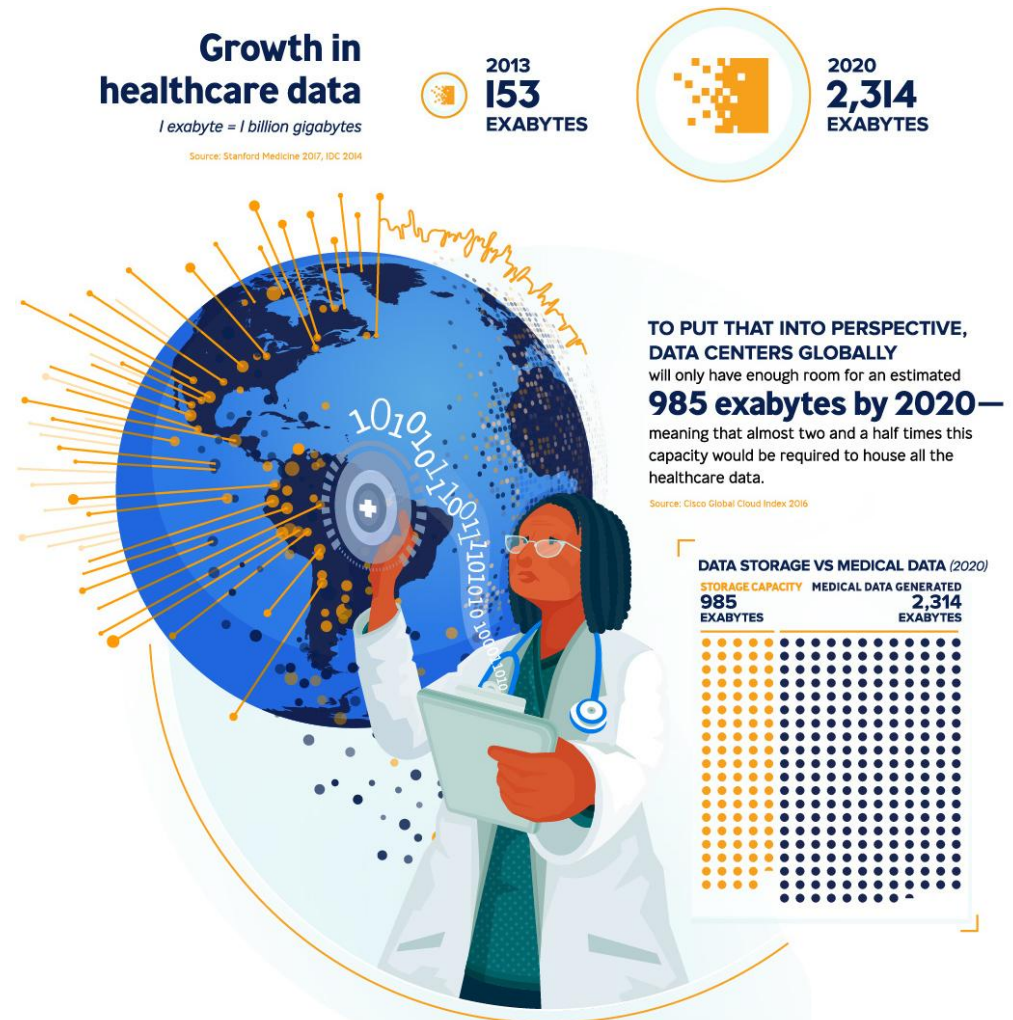
- Huge amounts of health data



Opportunity of Health Data Science

Benefits (and pitfalls!) of data science in general combined with:

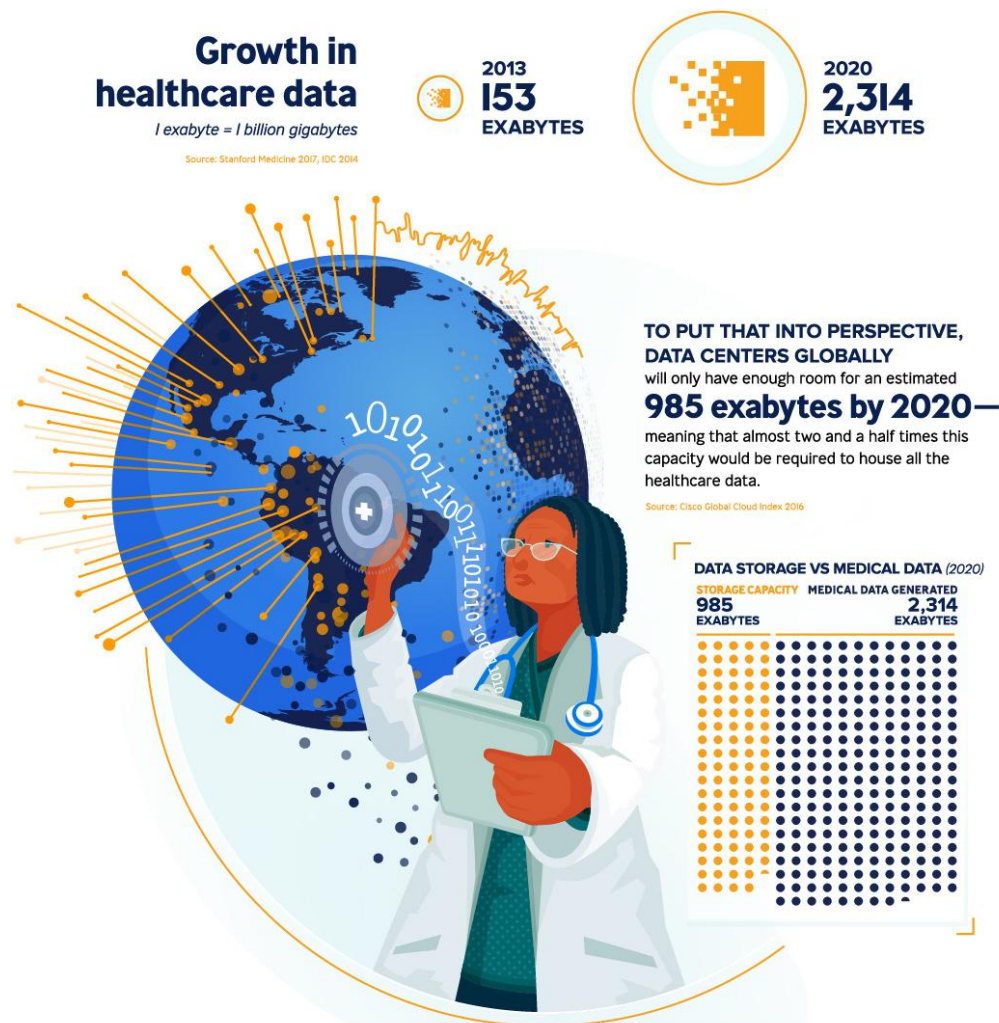
- Huge amounts of health data
- Many **interesting** and **important** problems



Opportunity of Health Data Science

Benefits (and pitfalls!) of data science in general combined with:

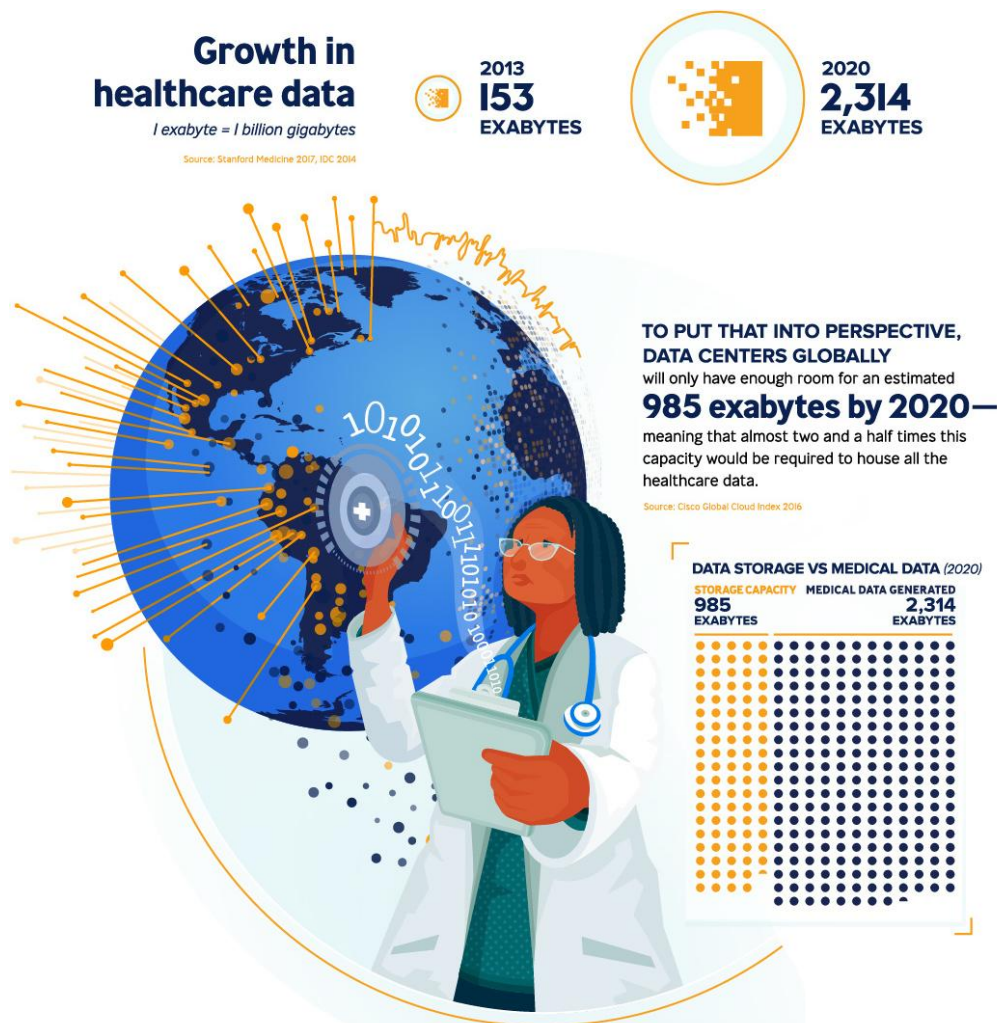
- Huge amounts of health data
- Many **interesting** and **important problems**
- Many domain experts desperate for data-related help with these problems



Opportunity of Health Data Science

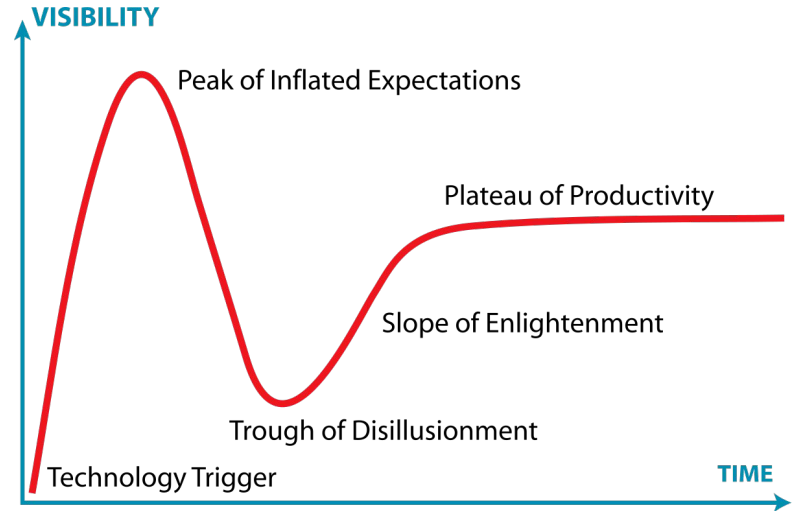
Benefits (and pitfalls!) of data science in general combined with:

- Huge amounts of health data
- Many **interesting** and **important problems**
- Many domain experts desperate for data-related help with these problems
- Relative few skilled data science practitioners



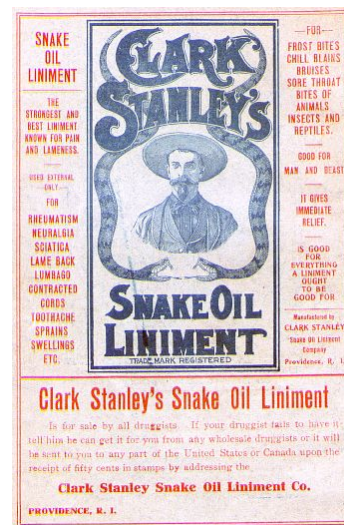
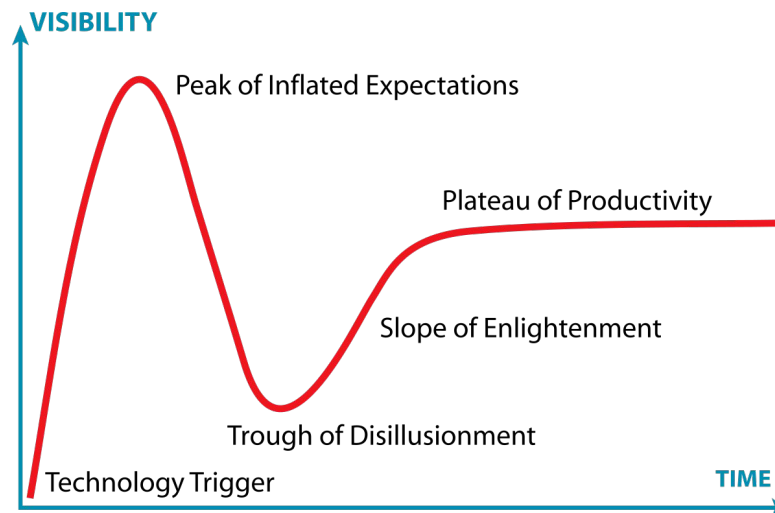
(Some) Challenges of Health Data Science

- Lots of hype



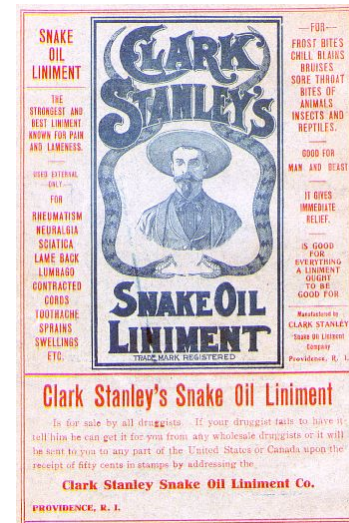
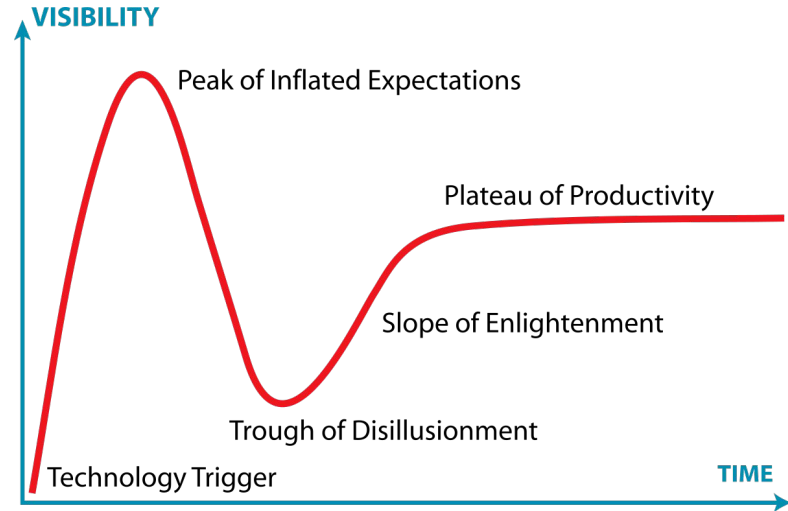
(Some) Challenges of Health Data Science

- Lots of hype
- Lots of grifters



(Some) Challenges of Health Data Science

- Lots of hype
- Lots of grifters
- Data quality issues
- Contextual/Metadata quality issues

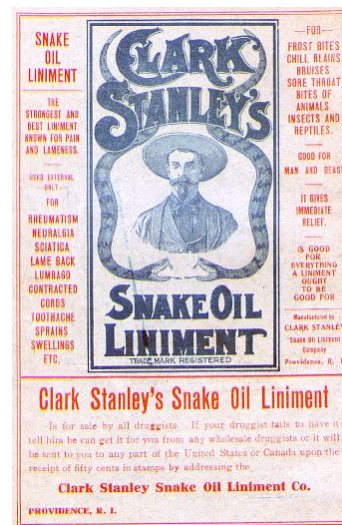
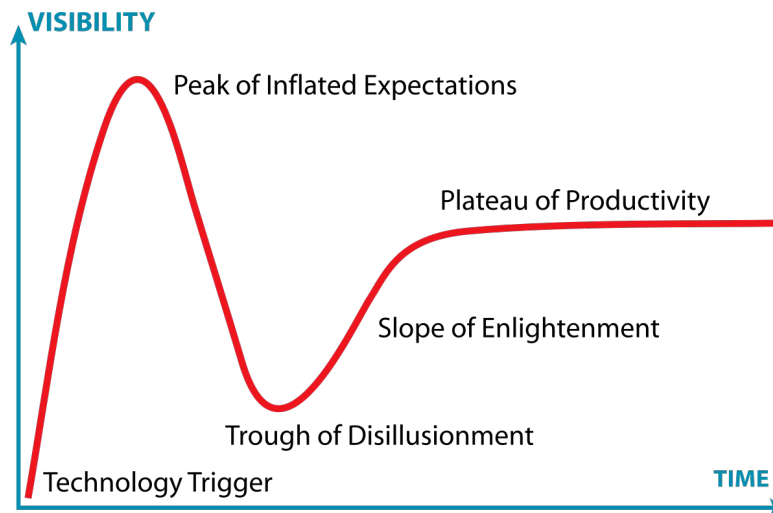


<https://www.r-bloggers.com/2019/08/new-course-learn-advanced-data-cleaning-in-r/>

https://upload.wikimedia.org/wikipedia/commons/9/94/Gartner_Hype_Cycle.svg
https://commons.wikimedia.org/wiki/File:Clark_Stanley%27s_Snake_Oil_Liniment.png

(Some) Challenges of Health Data Science

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

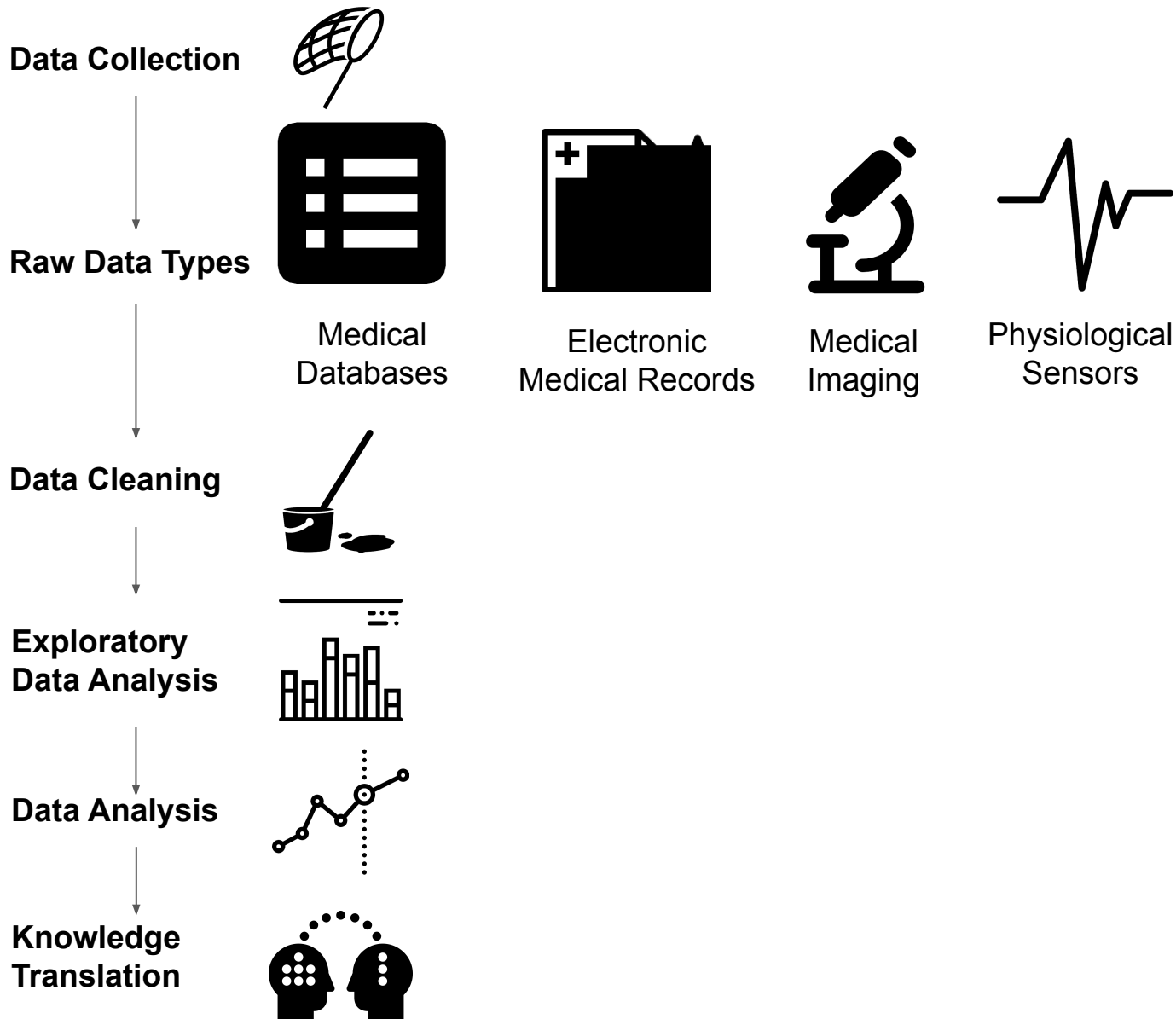


<https://www.r-bloggers.com/2019/08/new-course-learn-advanced-data-cleaning-in-r/>

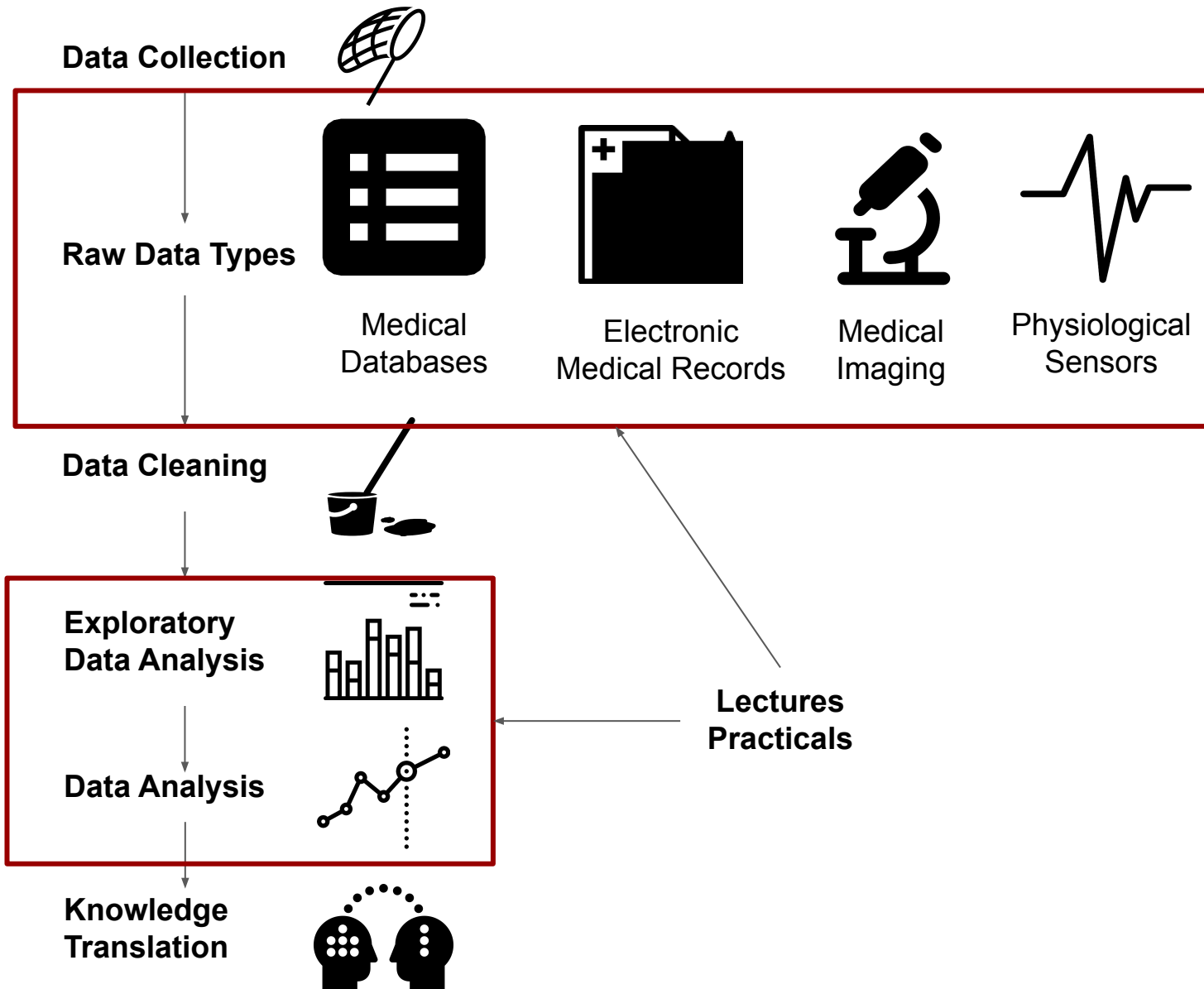
https://upload.wikimedia.org/wikipedia/commons/9/94/Gartner_Hype_Cycle.svg
https://commons.wikimedia.org/wiki/File:Clark_Stanley%27s_Snake_Oil_Liniment.png

What parts of health data science will this course cover?

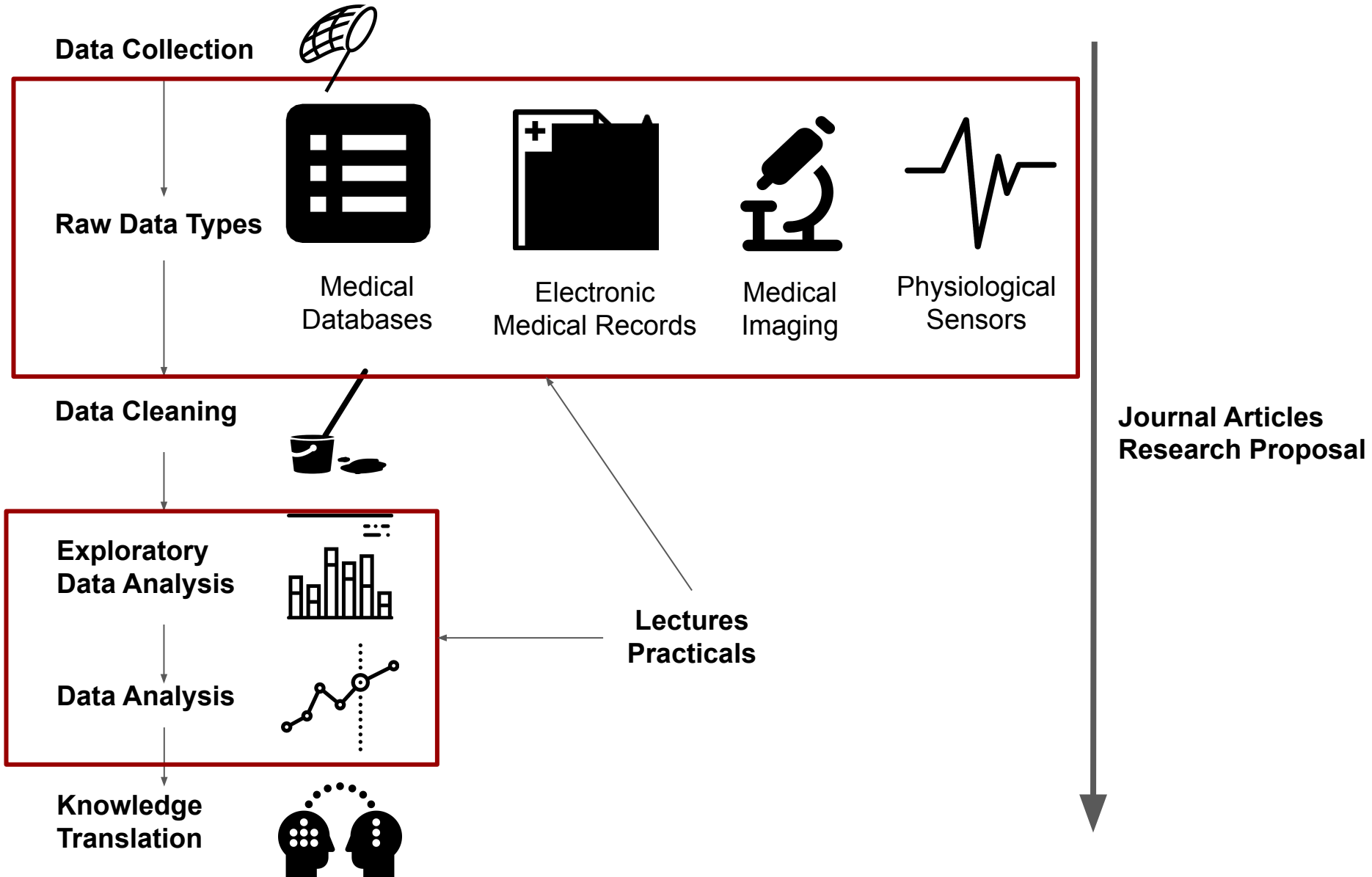
What parts of health data science will this course cover?



What parts of health data science will this course cover?



What parts of health data science will this course cover?



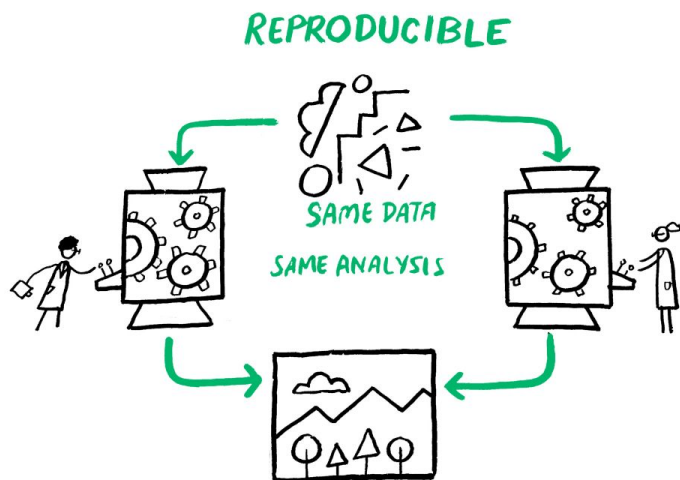
Let's take a 5 minute break!

Tools for Reproducible Health Data Science

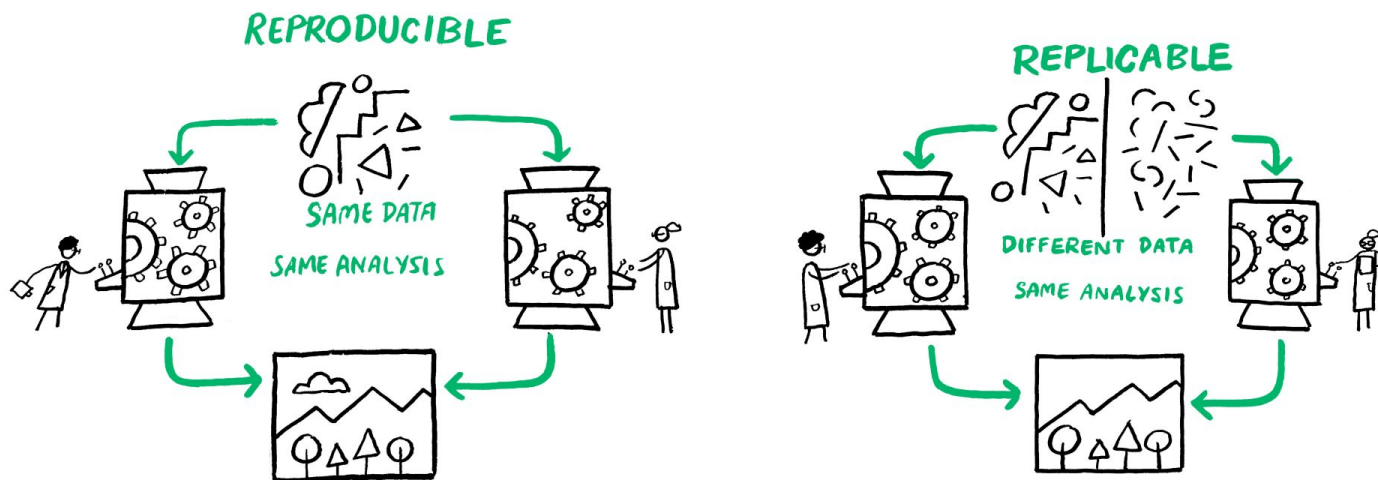
Rstudio, Rmarkdown, Git

Why do we care about reproducibility?

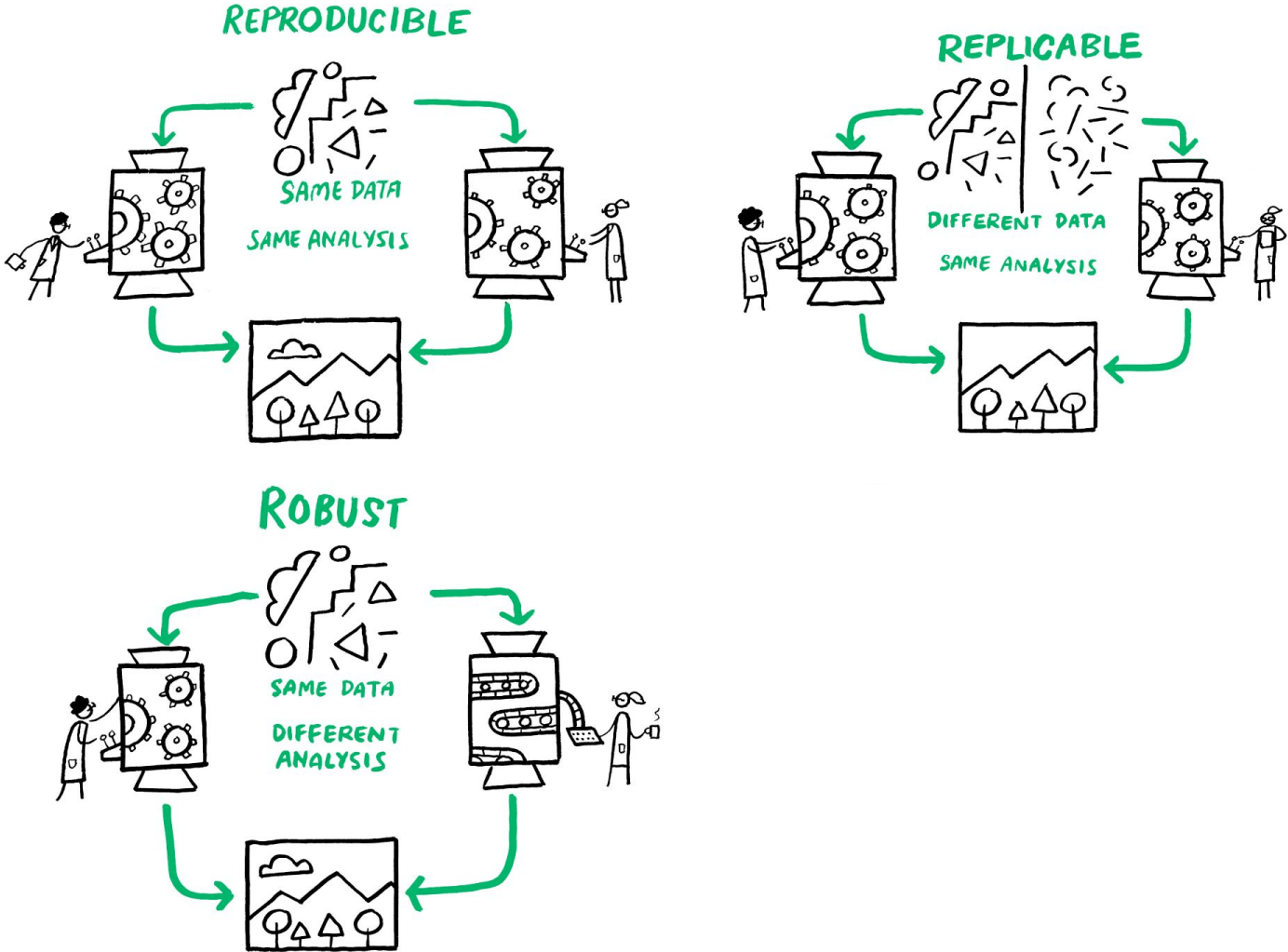
Reproducibility should be the bare minimum



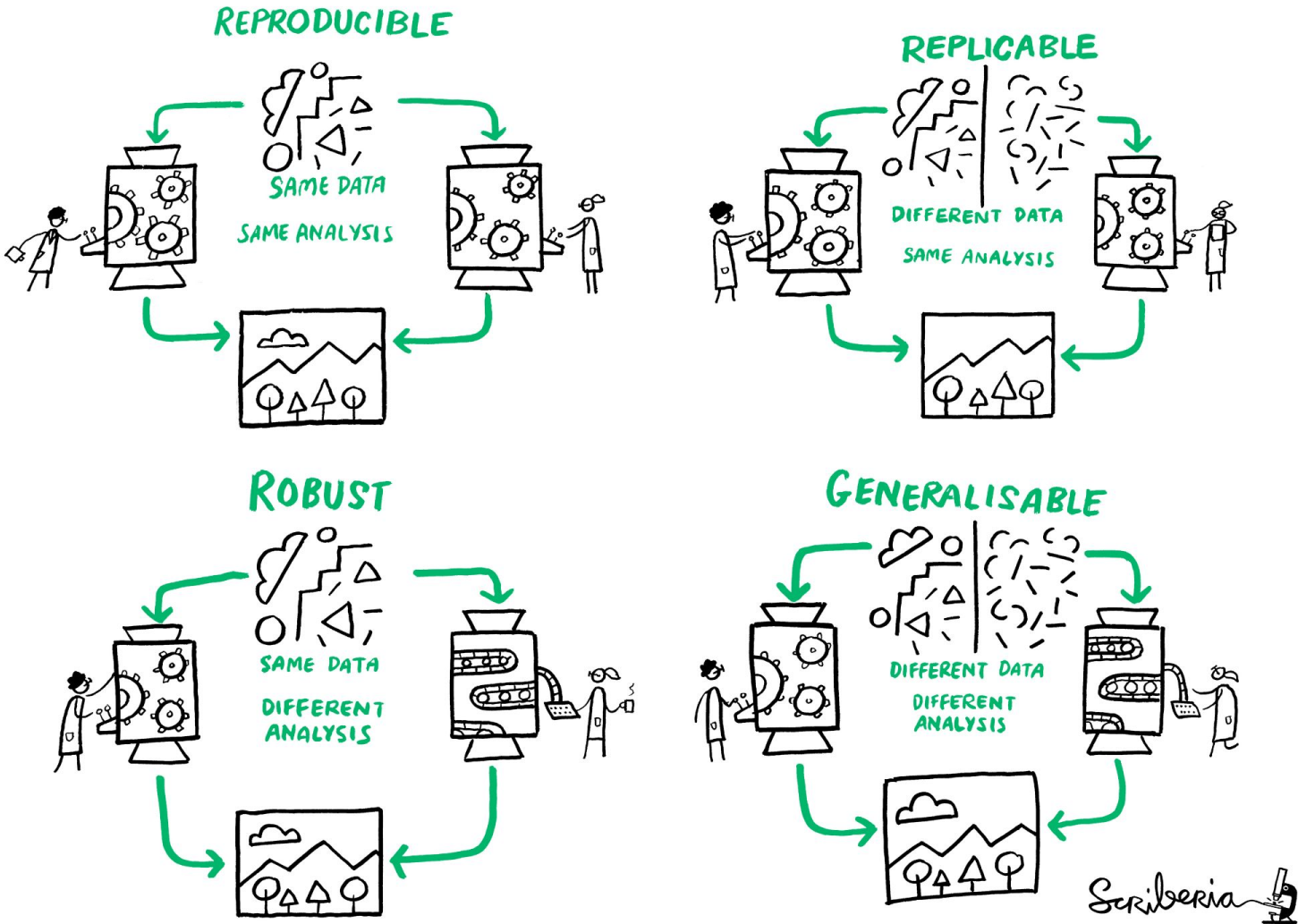
Reproducibility should be the bare minimum



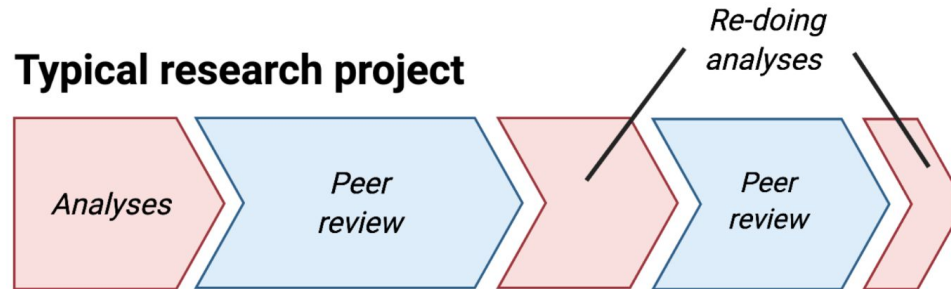
Reproducibility should be the bare minimum



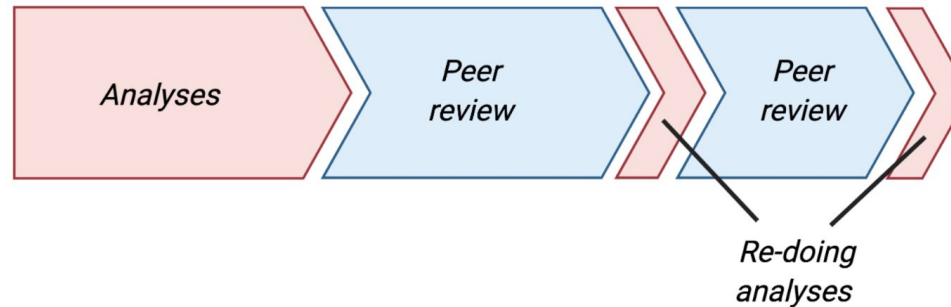
Reproducibility should be the bare minimum



Makes your own life easier



Research project using reproducible practices



 @dsquintana

oliviorgimenez.github.io/reproducible-science-workshop

What do we need to do to have reproducible research?

Reproducibility checklist

- Don't do anything by hand (even "one-off" tasks)

Reproducibility checklist

- Don't do anything by hand (even "one-off" tasks)
- Script every interaction with data:
 - Data collection
 - Moving data on your computer
 - Formatting datasets
 - Cleaning data
 - Exploratory data analysis
 - Main analyses
 - Report generation

Reproducibility checklist

- Don't do anything by hand (even "one-off" tasks)
- Script every interaction with data:
 - Data collection
 - Moving data on your computer
 - Formatting datasets
 - Cleaning data
 - Exploratory data analysis
 - Main analyses
 - Report generation
- Minimise interactivity/point and click interactions

Reproducibility checklist

- Don't do anything by hand (even "one-off" tasks)
- Script every interaction with data:
 - Data collection
 - Moving data on your computer
 - Formatting datasets
 - Cleaning data
 - Exploratory data analysis
 - Main analyses
 - Report generation
- Minimise interactivity/point and click interactions
- Version control all data, code, and documentation

Reproducibility checklist

- Don't do anything by hand (even "one-off" tasks)
- Script every interaction with data:
 - Data collection
 - Moving data on your computer
 - Formatting datasets
 - Cleaning data
 - Exploratory data analysis
 - Main analyses
 - Report generation
- Minimise interactivity/point and click interactions
- Version control all data, code, and documentation
- Use a random seed

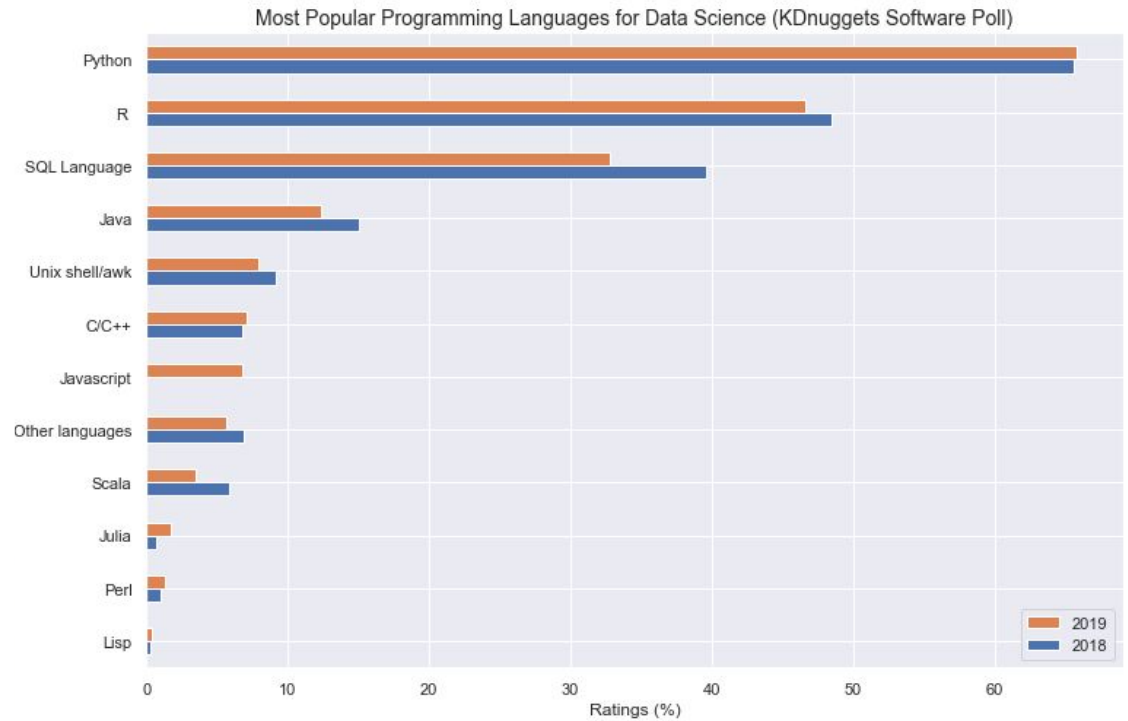
Reproducibility checklist

- Don't do anything by hand (even "one-off" tasks)
- Script every interaction with data:
 - Data collection
 - Moving data on your computer
 - Formatting datasets
 - Cleaning data
 - Exploratory data analysis
 - Main analyses
 - Report generation
- Minimise interactivity/point and click interactions
- Version control all data, code, and documentation
- Use a random seed
- Keep track of the exact version of every library/program you use

How do we actually do these things?

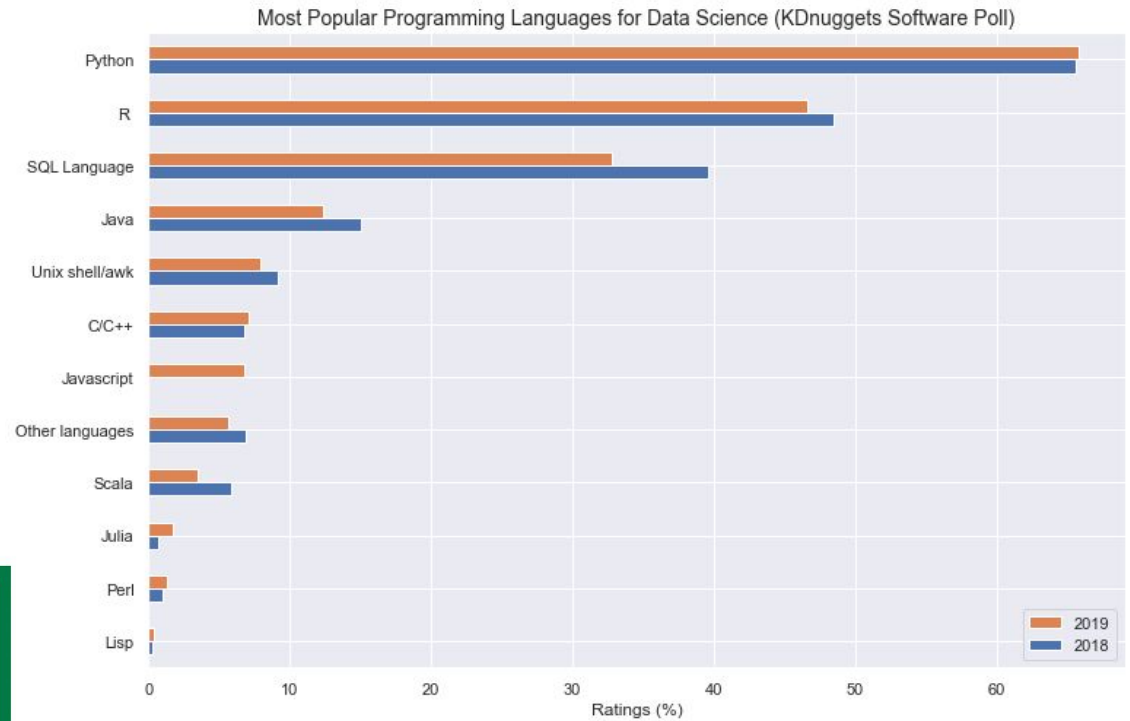
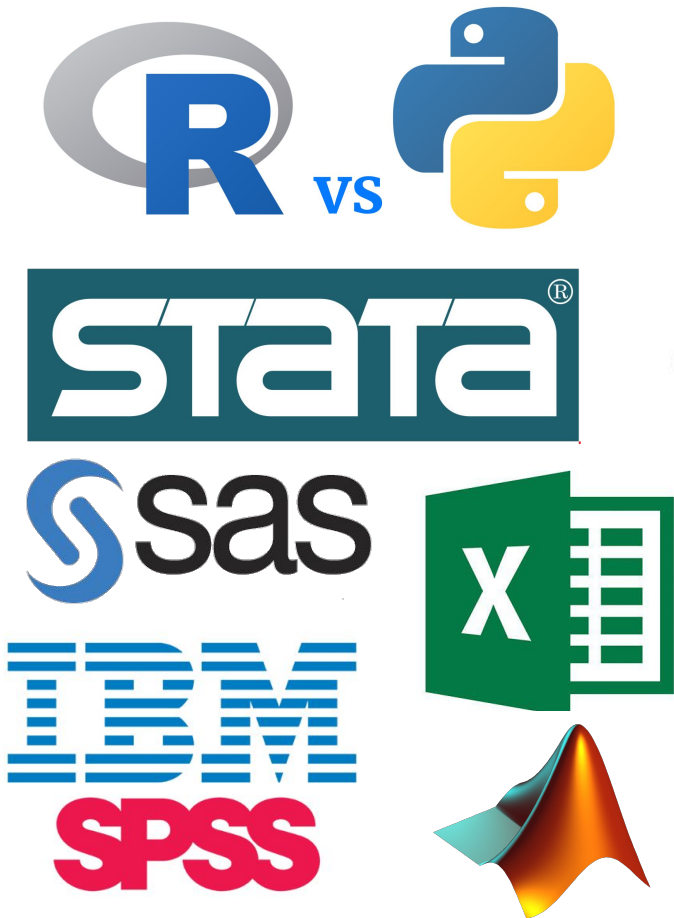
Choose a language that makes it easy to do most/all of your analysis

Choose a language that makes it easy to do most/all of your analysis



<https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html>

Choose a language that makes it easy to do most/all of your analysis



<https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html>

Use a data science focused IDE: Rstudio

The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains R code for loading packages, creating a 'daily' dataset, and plotting the number of flights per weekday.
- Environment:** Shows the 'daily' dataset with 365 observations and 3 variables.
- Console:** Shows the execution of the R code, including the creation of the 'daily' dataset and the execution of the plot command.
- Plots:** Displays a boxplot titled 'Number of 2013 New York Flights Each Weekday' showing the distribution of flights for each day of the week.

```
1 library(nycflights13) ## package containing flights dataset
2 library(lubridate)
3 library(dplyr)
4 library(ggplot2)
5
6 head(flights, n = 3)
7 daily <- flights %>%
8   mutate(date = make_date(year, month, day)) %>%
9   count(date) %>%
10  mutate(wday = wday(date, label = TRUE))
11 head(daily, n = 3)
12 ggplot(daily, aes(wday, n)) +
13   geom_boxplot(outlier.colour = "hotpink") +
14   labs(x = "Weekday", y = "Flights",
15        subtitle = "Number of 2013 New York Flights Each Weekday")
16
```

Console output:

```
# A tibble: 3 x 19
  year month day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier
  <int> <int> <int> <int> <int> <dbl> <int> <int> <dbl> <chr>
1 2013 1 1 517 515 2 830 819 11 UA
2 2013 1 1 533 529 4 850 830 20 UA
3 2013 1 1 542 540 2 923 850 33 AA
# ... with 9 more variables: flight <int>, tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
# distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dtm>
> daily <- flights %>%
+   mutate(date = make_date(year, month, day)) %>%
+   count(date) %>%
+   mutate(wday = wday(date, label = TRUE))
> head(daily, n = 3)
# A tibble: 3 x 3
  date           n wday
  <date> <int> <ord>
1 2013-01-01 842 Tue
2 2013-01-02 943 Wed
3 2013-01-03 914 Thu
> ggplot(daily, aes(wday, n)) +
+   geom_boxplot(outlier.colour = "hotpink") +
+   labs(x = "Weekday", y = "Flights",
+        subtitle = "Number of 2013 New York Flights Each Weekday")
>
```

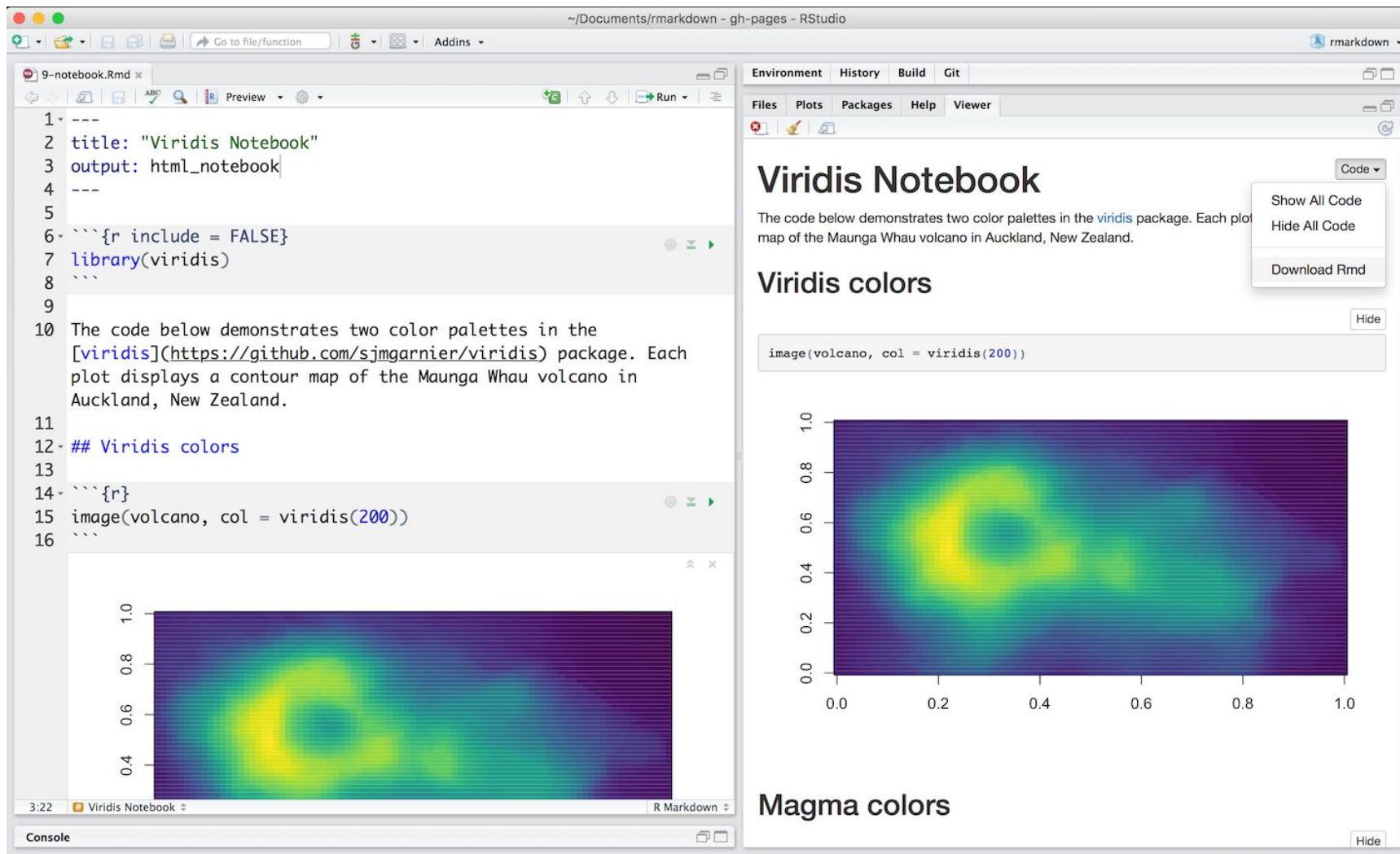
Boxplot visualization:

Number of 2013 New York Flights Each Weekday

| Weekday | Min | Q1 | Median | Q3 | Max |
|---------|-----|-----|--------|-----|-----|
| Sun | 720 | 890 | 900 | 910 | 990 |
| Mon | 910 | 960 | 970 | 980 | 990 |
| Tue | 760 | 940 | 950 | 960 | 990 |
| Wed | 720 | 940 | 950 | 960 | 990 |
| Thu | 740 | 940 | 950 | 960 | 990 |
| Fri | 760 | 940 | 950 | 960 | 990 |
| Sat | 680 | 730 | 750 | 770 | 860 |

set.seed()
sessionInfo()

Use notebooks to document analyses: Rmarkdown



The screenshot displays the RStudio interface with an R Markdown notebook open. The notebook title is "Viridis Notebook" and the output format is set to "html_notebook". The code in the notebook includes:

```
1 ---
2 title: "Viridis Notebook"
3 output: html_notebook
4 ---
5
6 ```{r include = FALSE}
7 library(viridis)
8 ```
9
10 The code below demonstrates two color palettes in the
11 [viridis](https://github.com/sjmgarnier/viridis) package. Each
12 plot displays a contour map of the Maunga Whau volcano in
13 Auckland, New Zealand.
14
15 ## Viridis colors
16
17 ```{r}
18 image(volcano, col = viridis(200))
19 ```
```

The notebook shows two contour plots of the Maunga Whau volcano. The first plot, titled "Viridis colors", uses the viridis color palette. The second plot, titled "Magma colors", uses the magma color palette. Both plots show a contour map of the volcano with axes ranging from 0.0 to 1.0. The RStudio interface also shows the Environment, History, Build, and Git panels, and the Files, Plots, Packages, Help, and Viewer panels.

Use notebooks to document analyses: Rmarkdown

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

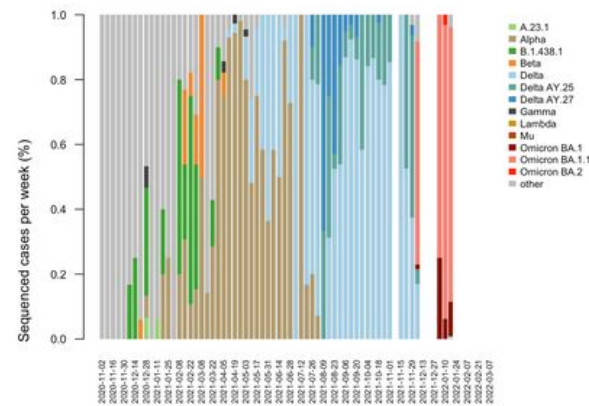
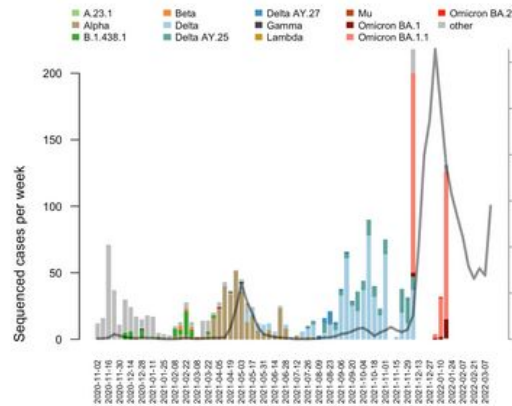
Nova Scotia

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

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

Hide

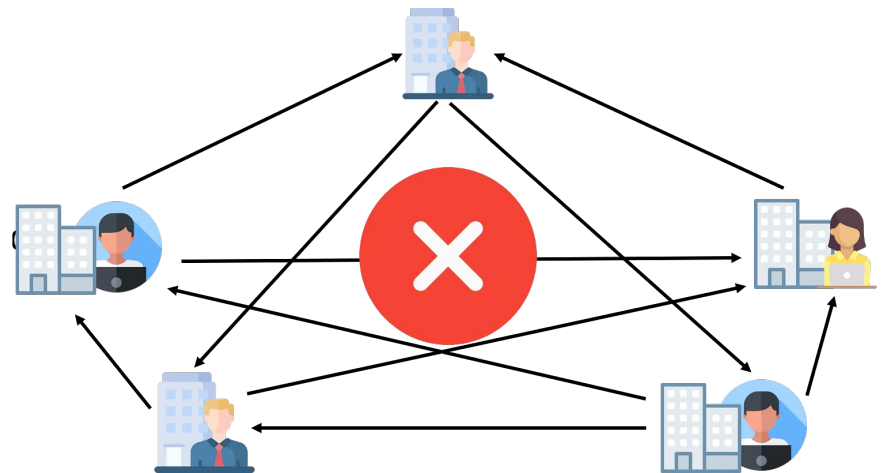
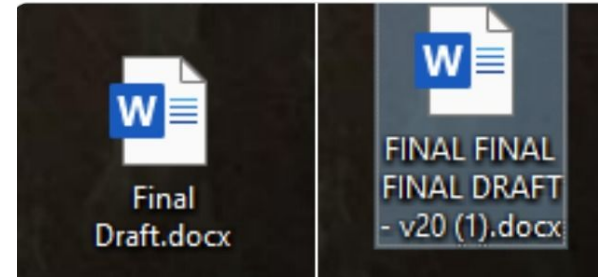
```
plot.variants(region='Nova Scotia')
plot.variants(region='Nova Scotia', scaled=T)
```



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

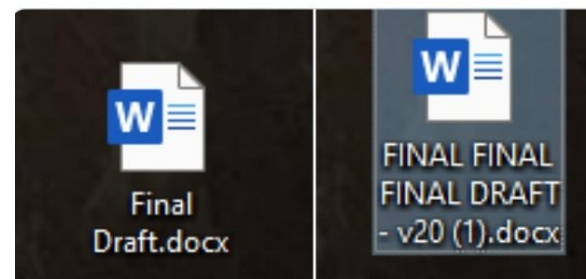
Use standard version control systems

- Ever had a nightmare of versioning even when just you?
- Add more people and the chaos grows exponentially!



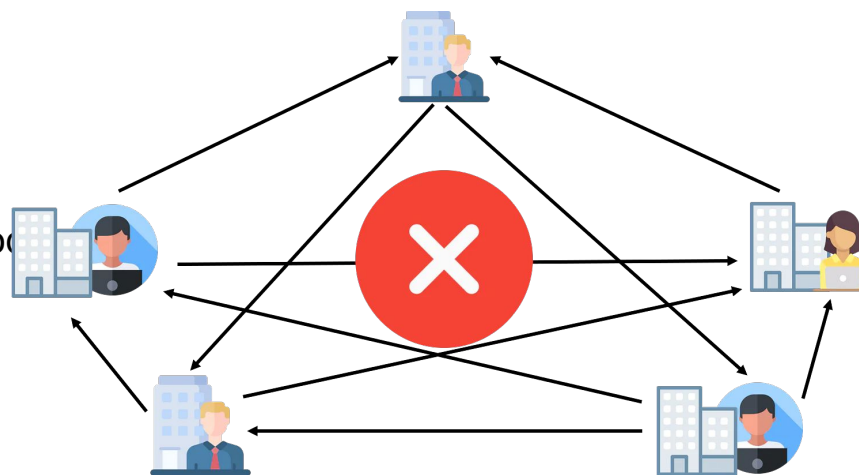
Use standard version control systems

- Ever had a nightmare of versioning even when just you?
- Add more people and the chaos grows exponentially!

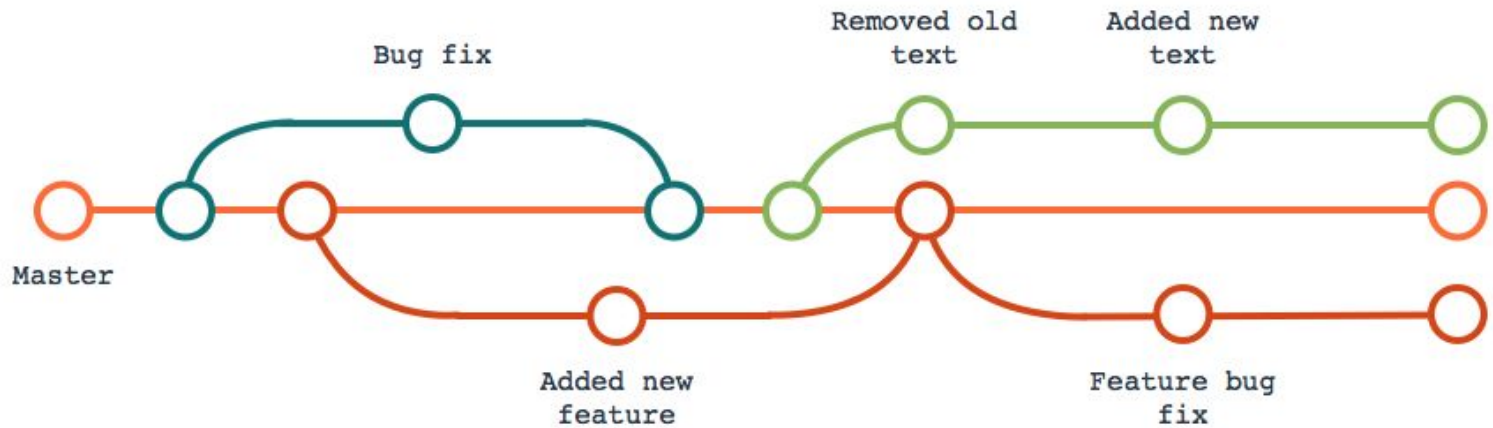


Version control let's you:

- Revert mistakes
- Acts as a comprehensive backup
- Let's you maintain multiple versions of your analysis
- Let's you compare different versions of your code
- Track down the who/what broke the analysis
- Work out why you did something in the past
- Build on someone else's work
- Share your own work
- Experiment without risk

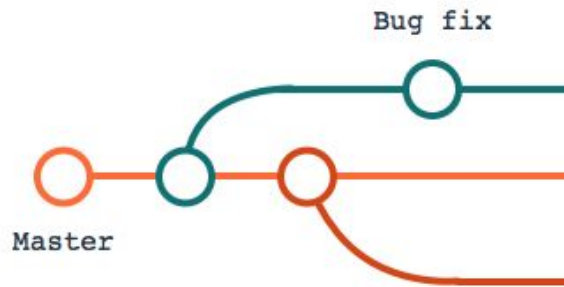


Git Version Control

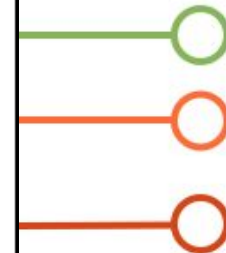


- Most popular
- Decentralised
- Designed for
- GitLab/GitHub Services

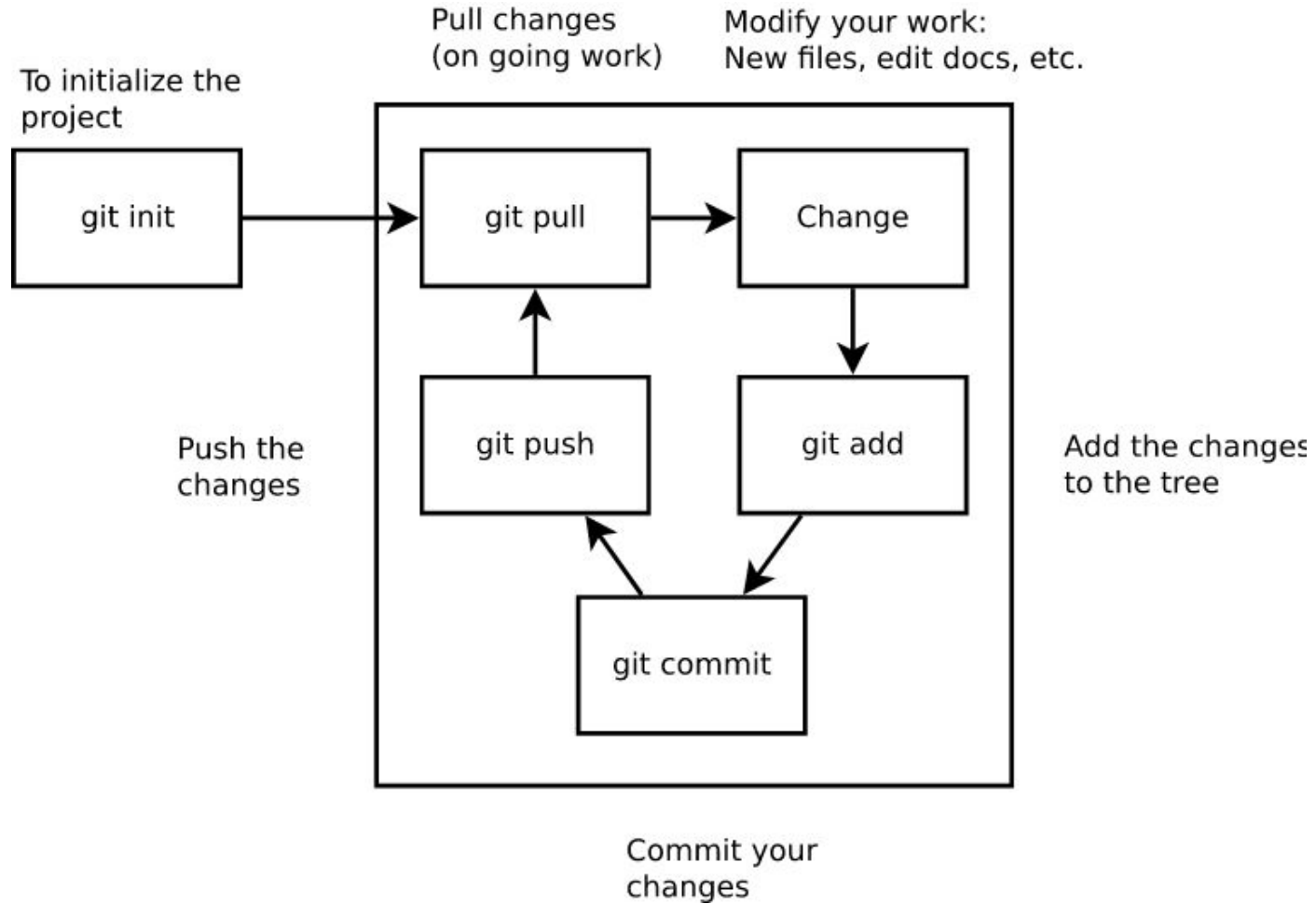
Git Version Control



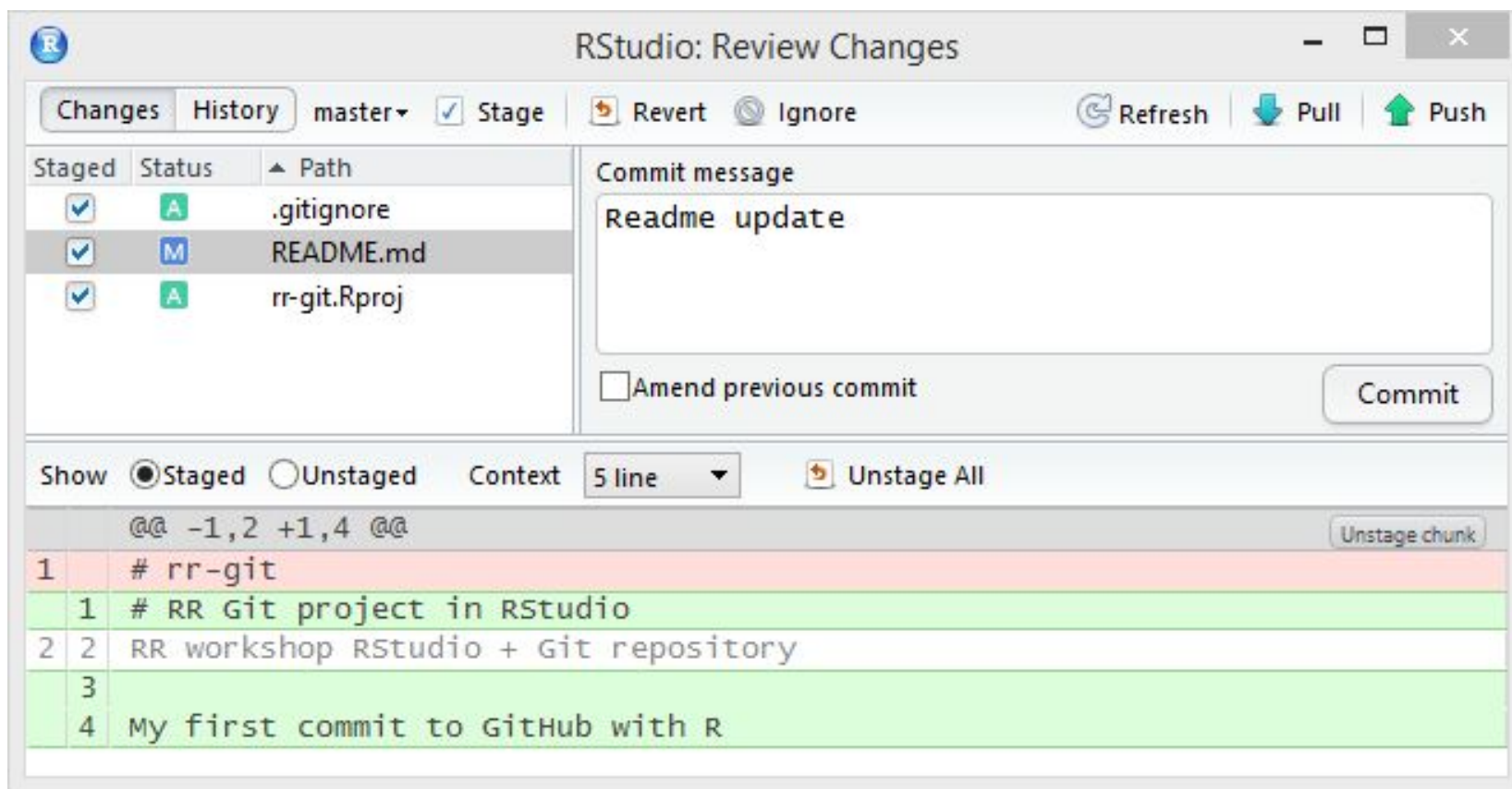
- Most popular
- Decentralised
- Designed for
- GitLab/GitHub Services



Git Workflow



Git is integrated into Rstudio!



The screenshot shows the RStudio 'Review Changes' window. At the top, there are tabs for 'Changes' and 'History', and a dropdown menu set to 'master'. Below this are buttons for 'Stage', 'Revert', and 'Ignore', along with 'Refresh', 'Pull', and 'Push' icons. The main area is divided into two panes. The left pane shows a table of staged files:

| Staged | Status | Path |
|-------------------------------------|--------|--------------|
| <input checked="" type="checkbox"/> | A | .gitignore |
| <input checked="" type="checkbox"/> | M | README.md |
| <input checked="" type="checkbox"/> | A | rr-git.Rproj |

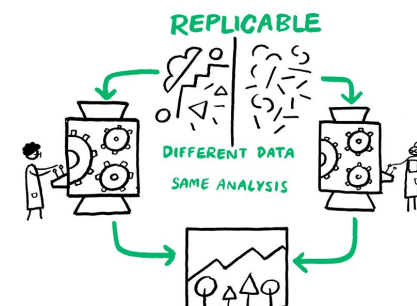
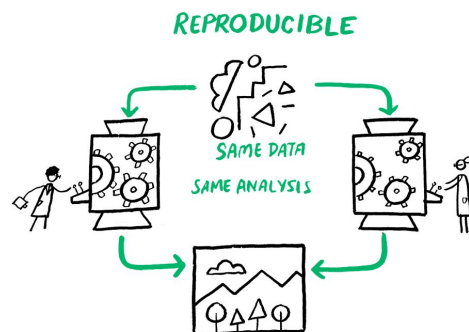
The right pane is for the commit message, with a text box containing 'Readme update' and an 'Amend previous commit' checkbox. A 'Commit' button is at the bottom right. Below the panes, there are controls for 'Show' (radio buttons for 'Staged' and 'Unstaged'), 'Context' (a dropdown set to '5 line'), and 'Unstage All'. A diff view shows the changes to README.md:

```
@@ -1,2 +1,4 @@
1 # rr-git
1 # RR Git project in RStudio
2 2 RR workshop RStudio + Git repository
3
4 My first commit to GitHub with R
```

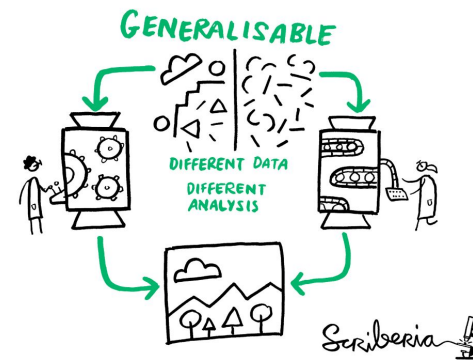
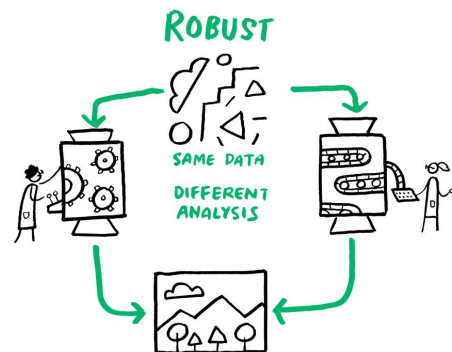
An 'Unstage chunk' button is visible on the right side of the diff view.

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



Summary

- Overview of course: Database/EMR/Imaging/Signal
- Main assessments: practicals, journal article presentations, research proposal
- Data science is statistics with an EDA/Inductive/Data-focused Spin
- Health Data Science is a massive and growing area with lots of opportunity and challenges
- R is a powerful and useful tool for health data science
- Reproducibility is vital to good ~~health data~~ science
- Rstudio, Rmarkdown notebooks and Git based version control facilitate that reproducibility

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's Journal Articles

- **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>