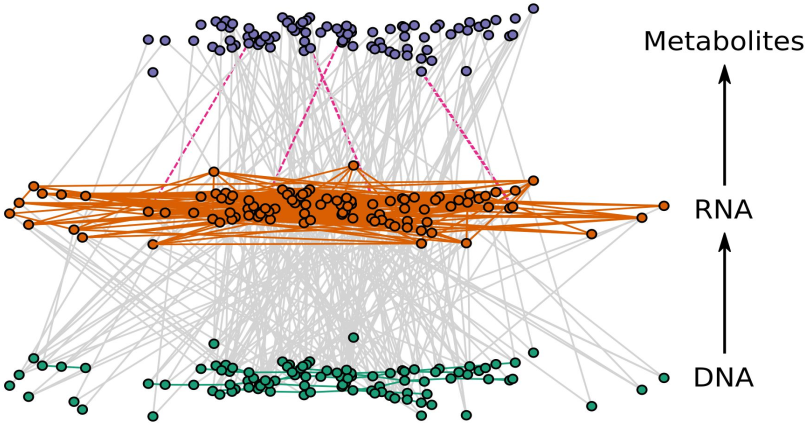
Lecture 0: Introduction to Applied Research in Health Data Science

CSCI6410/4148 & EPAH6410

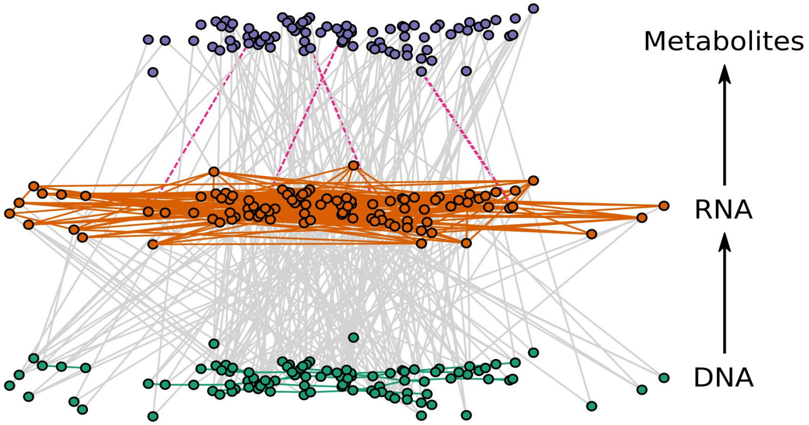
Finlay Maguire (finlay.maguire@dal.ca)

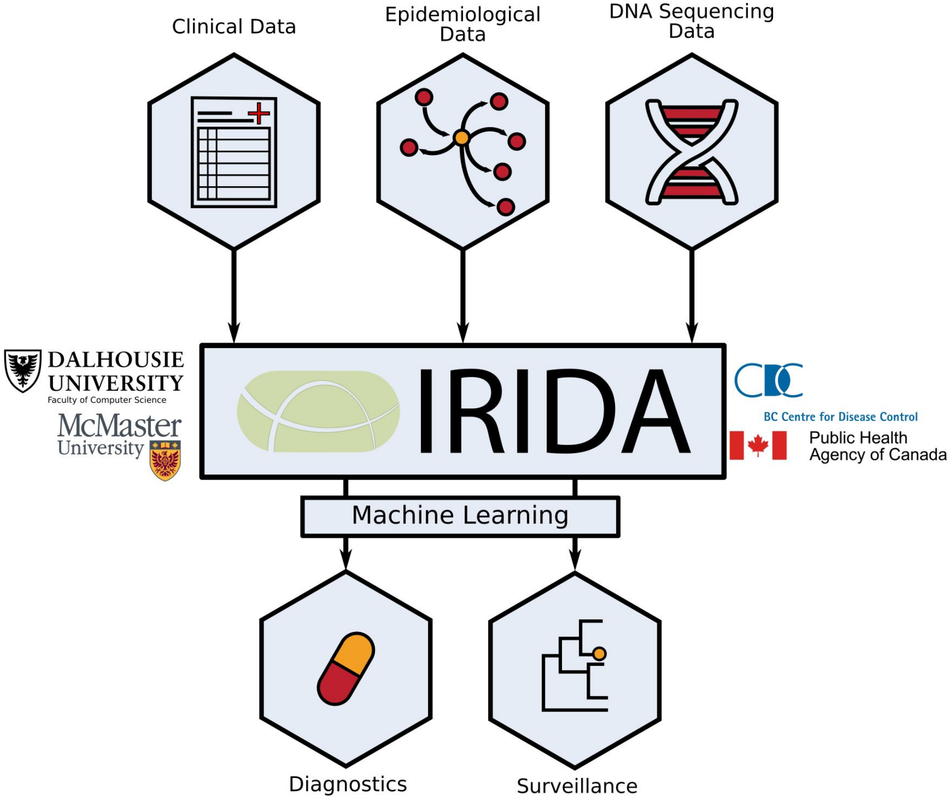
Why am I teaching this course?



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

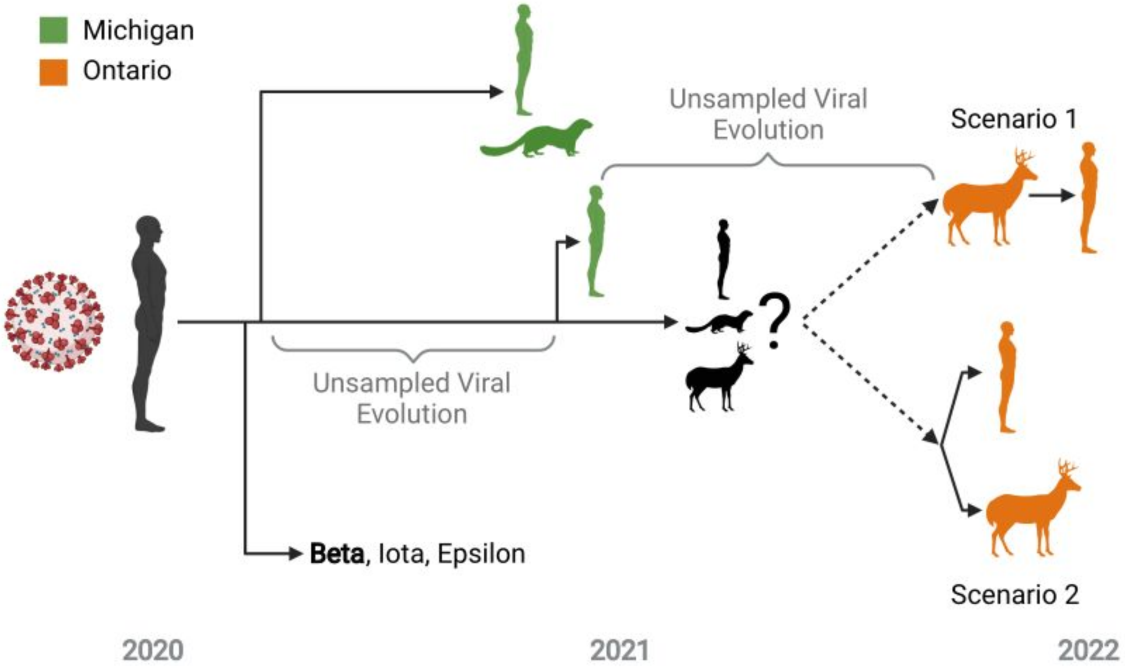
Why am I teaching this course?





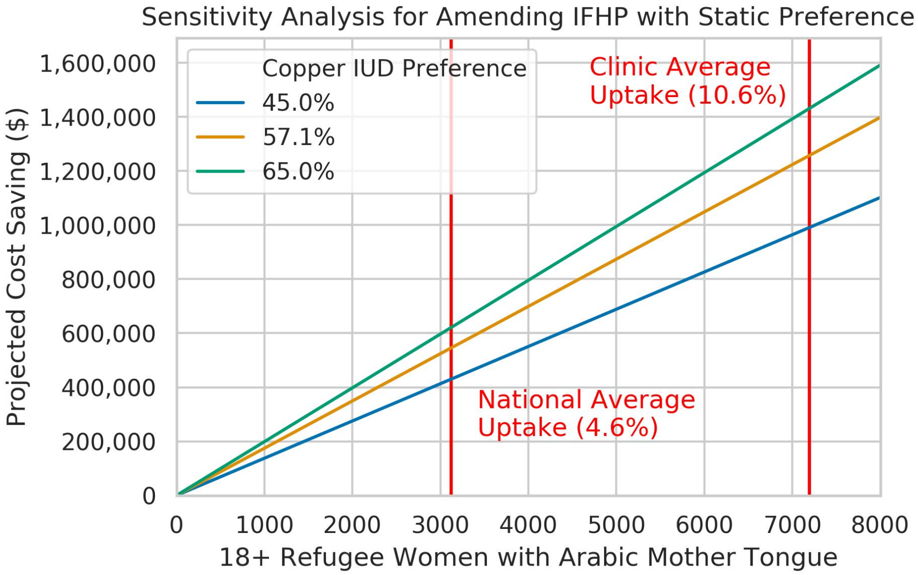
- **PhD (Bioinformatics)**: using large noisy datasets to understand how microbial systems and mechanisms evolve.
- **Postdoc (Genomic Epidemiology)**: using large noisy datasets to better diagnose, track and predict infectious diseases.

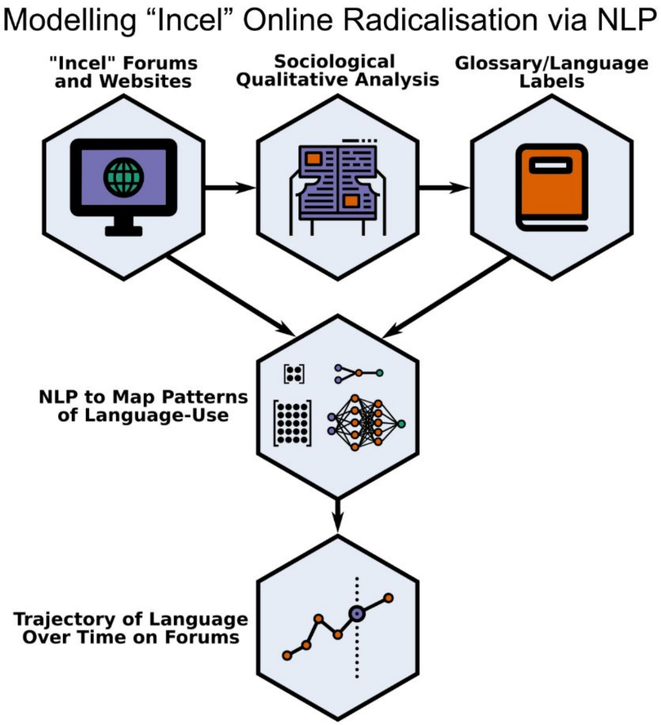
Why am I teaching this course?



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

Why am I teaching this course?

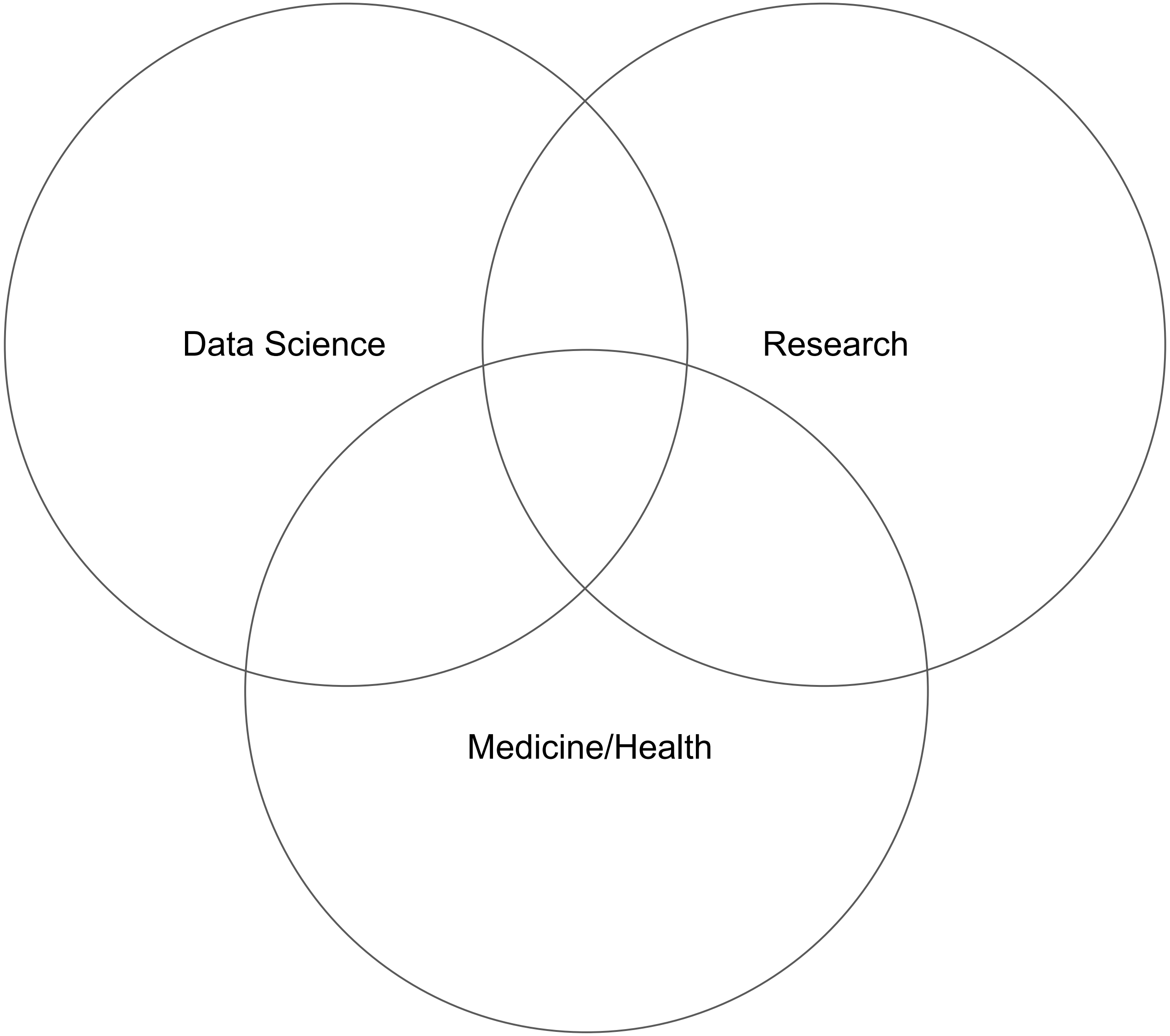



Modelling "Incel" Online Radicalisation via NLP
"Incel" Forums and Websites
Sociological Qualitative Analysis
Glossary/Language Labels
NLP to Map Patterns of Language-Use
Trajectory of Language Over Time on Forums

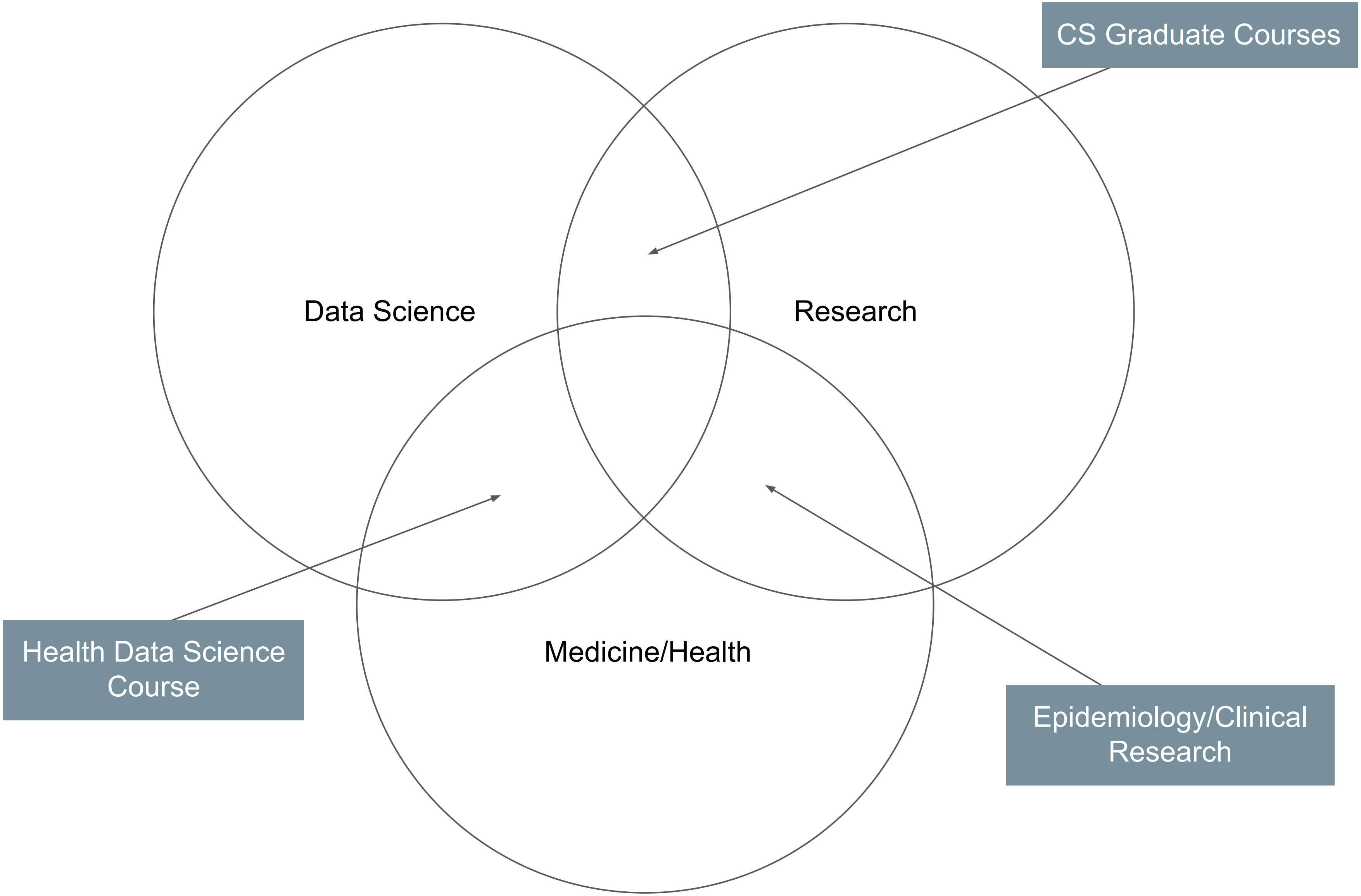
- **Research group**: using large noisy datasets:
  - Genomic epidemiology of infectious disease: **SARS-CoV-2**, **AMR**
  - Collaborations on socially/health focused problems: **refugee health**, **incel radicalisation, health inequality**

Overview of course

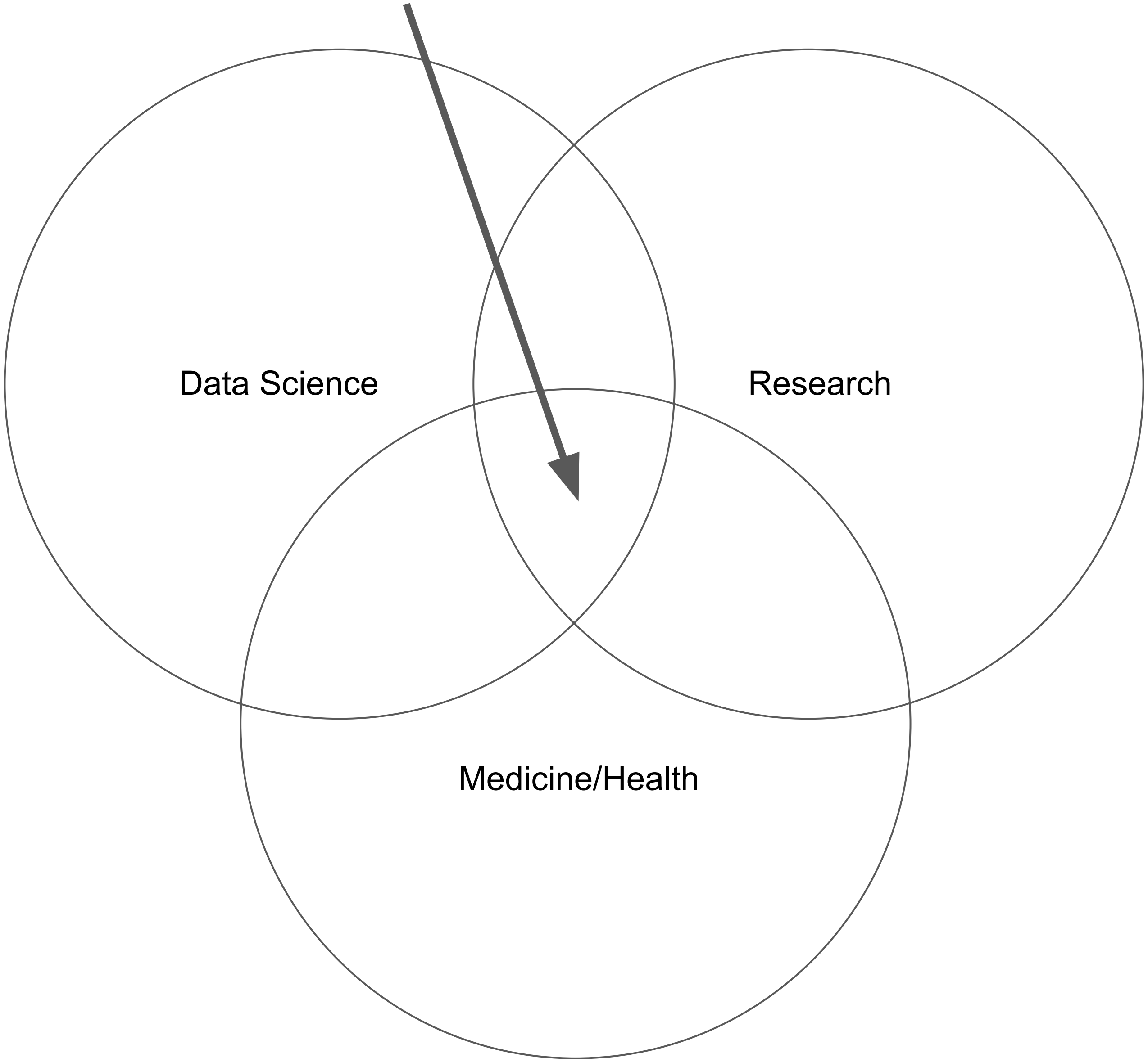
Applied Research in Health Data Science



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:

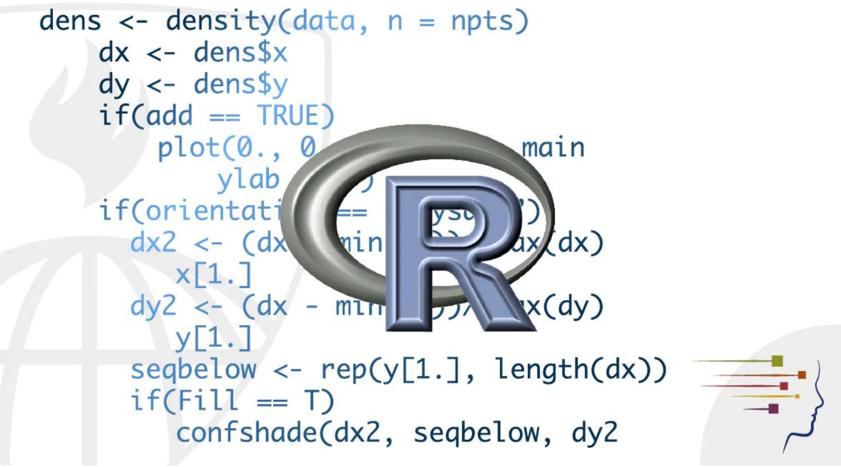
- **Lectures** (Monday/Wednesday)

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   ==   ys     ")
      dx2 <- (dx   min   ))   ax(dx)
         x[1.]
      dy2 <- (dx - min   ))/   x(dy)
         y[1.]
      seqbelow <- rep(y[1.], length(dx))
      if(Fill == T)
         confshade(dx2, seqbelow, dy2
```

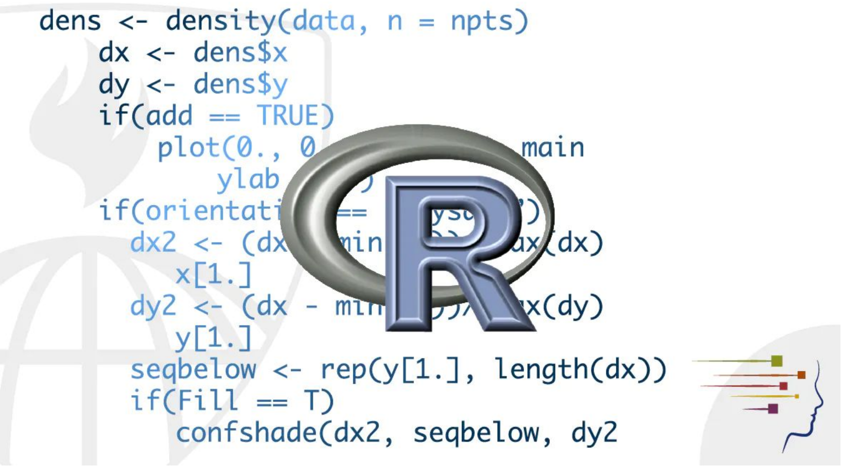
https://www.coursera.org/learn/r-programming

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    ==     ysc    ")
      dx2 <- (dx   min           ax(dx)
         x[1.]
      dy2 <- (dx - min          x(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%)

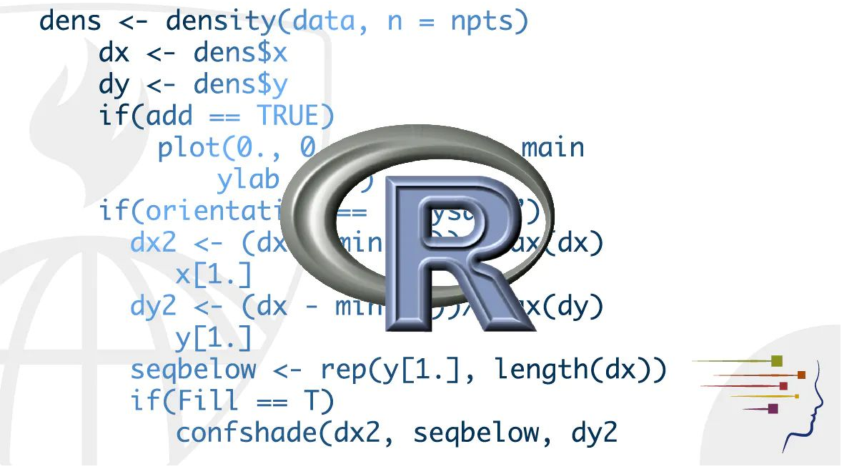
Participation in discussion (10%)

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     ==     ys      ")
      dx2 <- (dx     min           x(dx)
        x[1.]
      dy2 <- (dx - min      ))/    x(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:

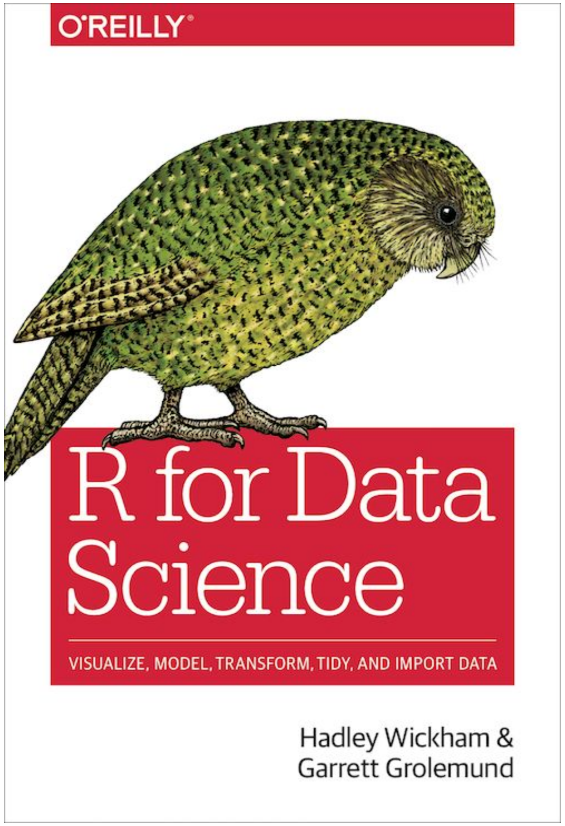
- **Class** (Wednesday/Friday)

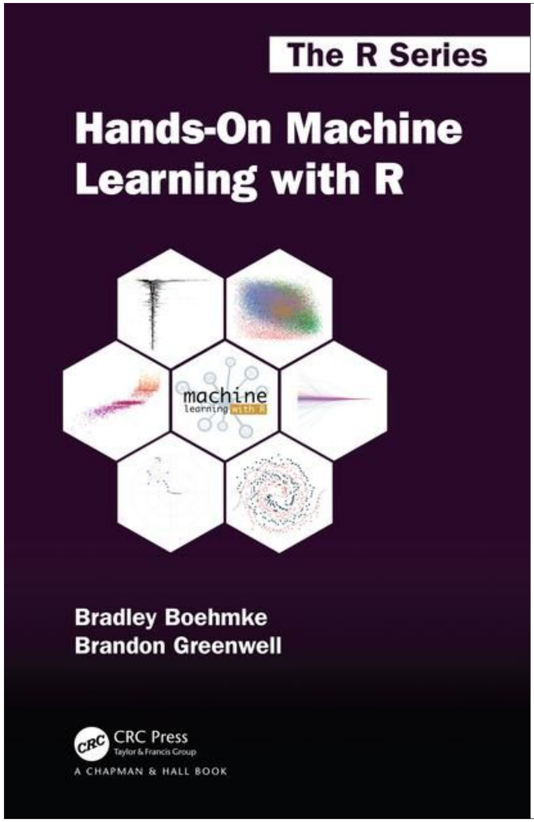
Assessment:

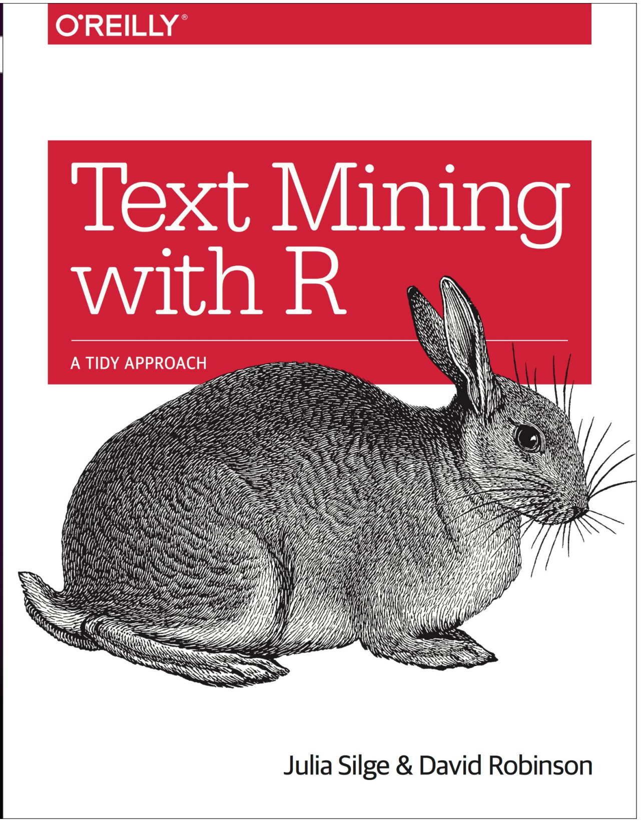
Presentation **last full week of class** (20%)

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

Course Materials





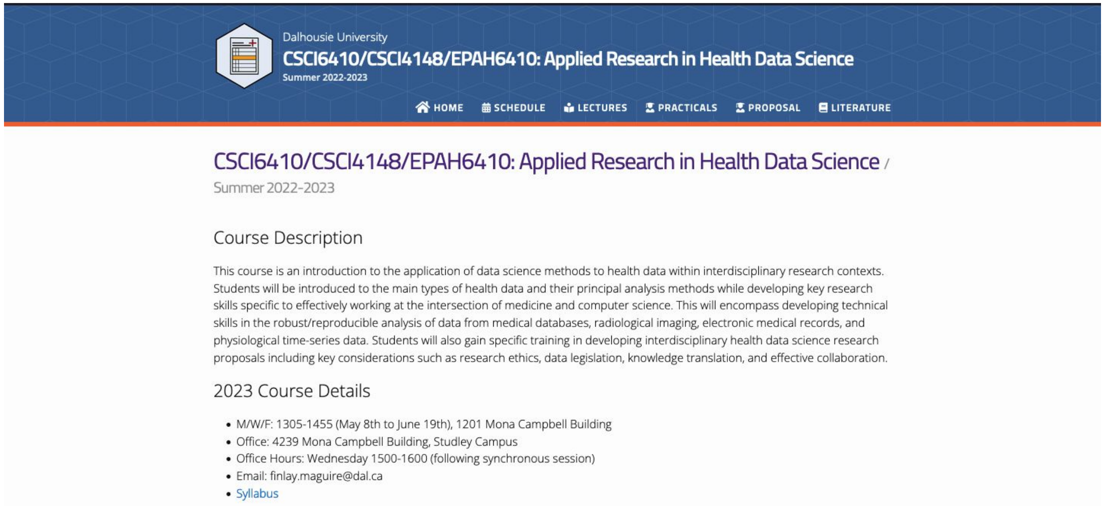


https://r4ds.had.co.nz/

https://bradleyboehmke.githu
b.io/HOML/

https://www.tidytextmining.com/

Course Website



Dalhousie University

CSCI6410/CSCI4148/EPAH6410: Applied Research in Health Data Science

Summer 2022-2023

CSCI6410/CSCI4148/EPAH6410: Applied Research in Health Data Science / Summer 2022-2023

Course Description

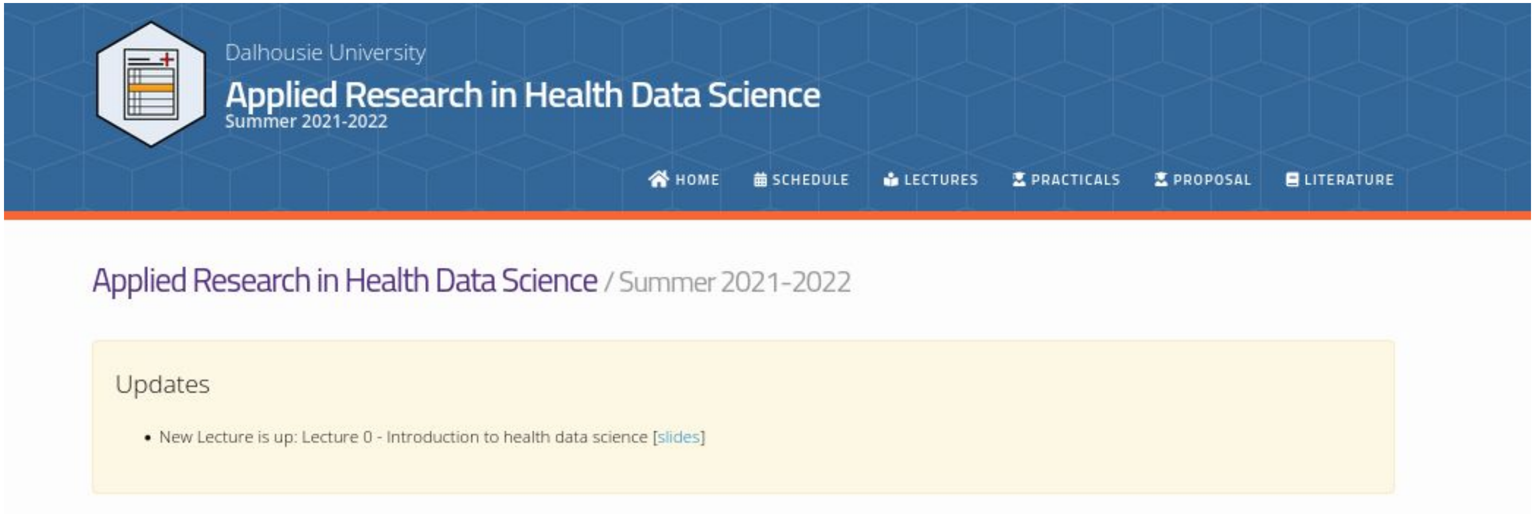
This course is an introduction to the application of data science methods to health data within interdisciplinary research contexts. Students will be introduced to the main types of health data and their principal analysis methods while developing key research skills specific to effectively working at the intersection of medicine and computer science. This will encompass developing technical skills in the robust/reproducible analysis of data from medical databases, radiological imaging, electronic medical records, and physiological time-series data. Students will also gain specific training in developing interdisciplinary health data science research proposals including key considerations such as research ethics, data legislation, knowledge translation, and effective collaboration.

2023 Course Details

- M/W/F: 1305-1455 (May 8th to June 19th), 1201 Mona Campbell Building
- Office: 4239 Mona Campbell Building, Studley Campus
- Office Hours: Wednesday 1500-1600 (following synchronous session)
- Email: finlay.maguire@dal.ca
- Syllabus

https://maguire-lab.github.io/health_data_science_research_2023/

Course Website



Dalhousie University

Applied Research in Health Data Science

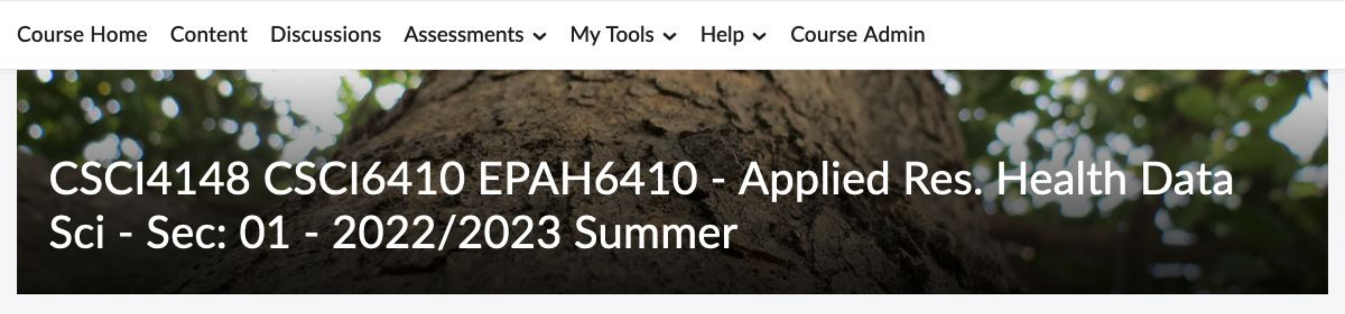
Summer 2021-2022

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/

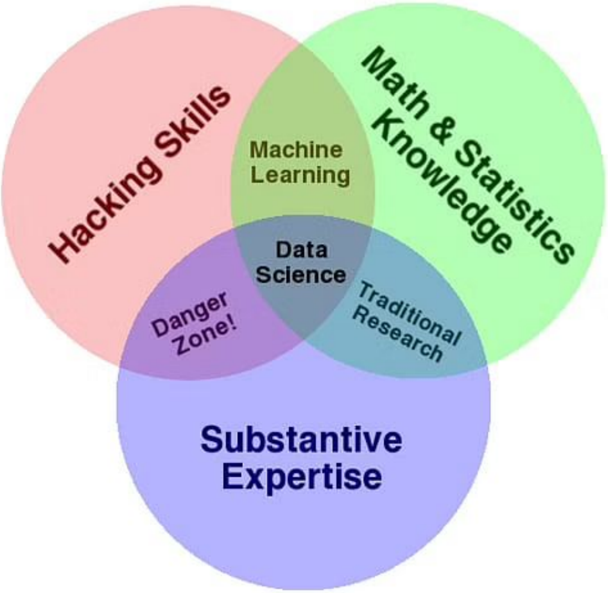


CSCI4148 CSCI6410 EPAH6410 - Applied Res. Health Data Sci - Sec: 01 - 2022/2023 Summer

Grades/Submissions:
https://dal.brightspace.com/d2l/home/221757

What is ~~health~~ data science?

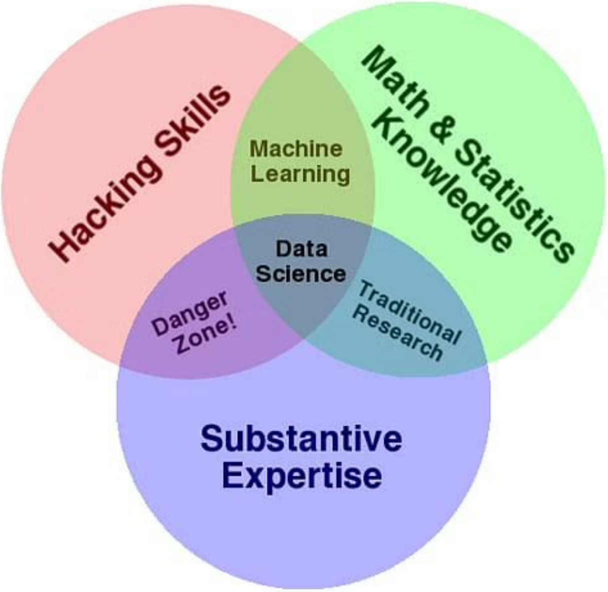
Data Science: *Using Data to Better Understand Things in the Real World*



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:

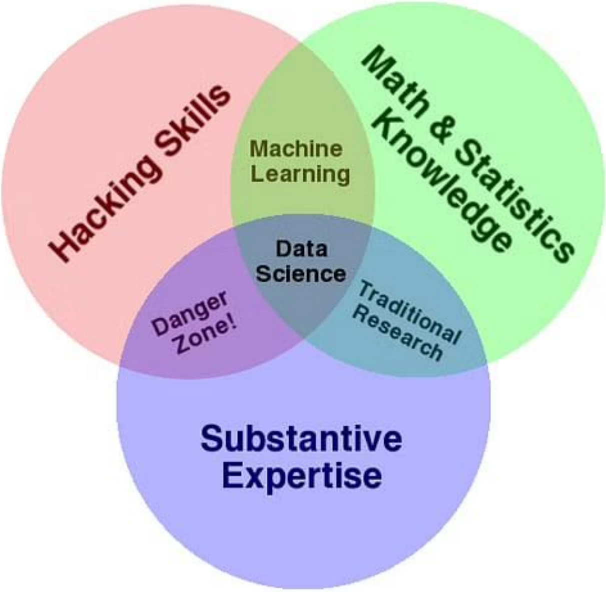


http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

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



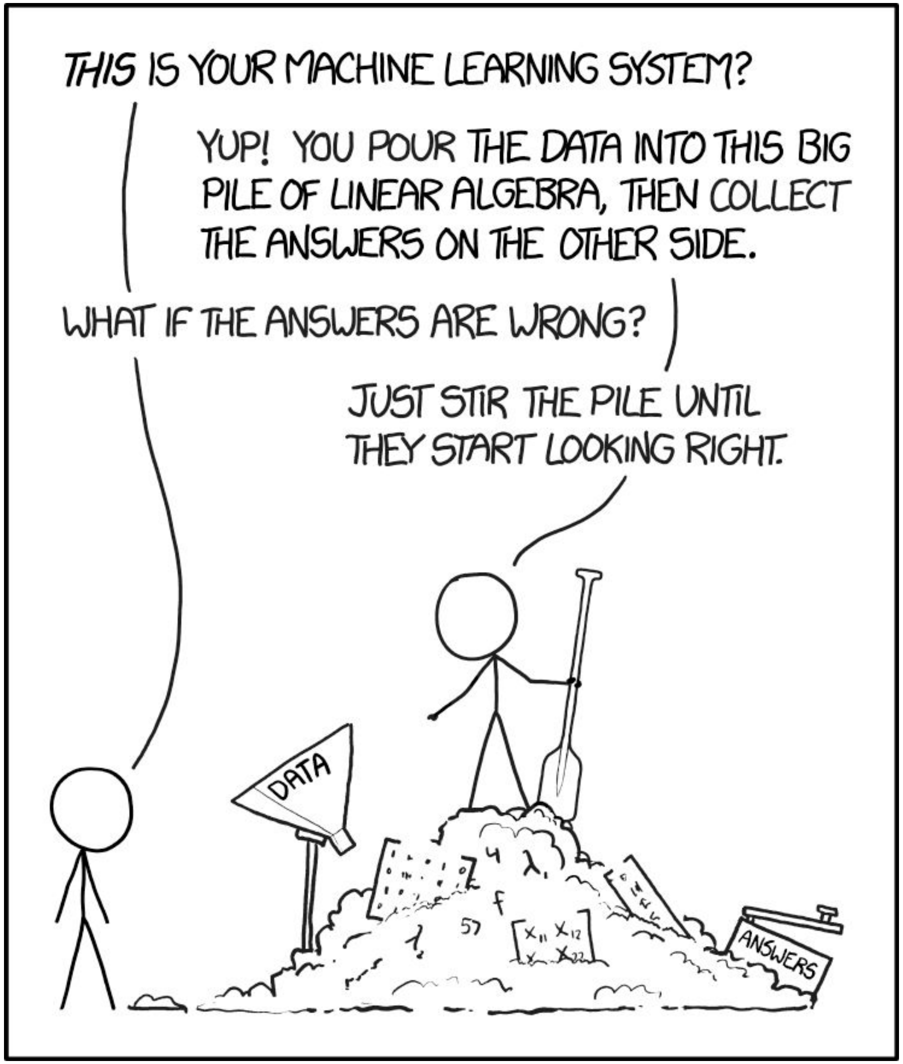
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

So, it is just statistics?

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

Pitfalls (can be):

- Less rigorous/principled
- Prone to reinventing the wheel


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.
DATA
ANSWERS

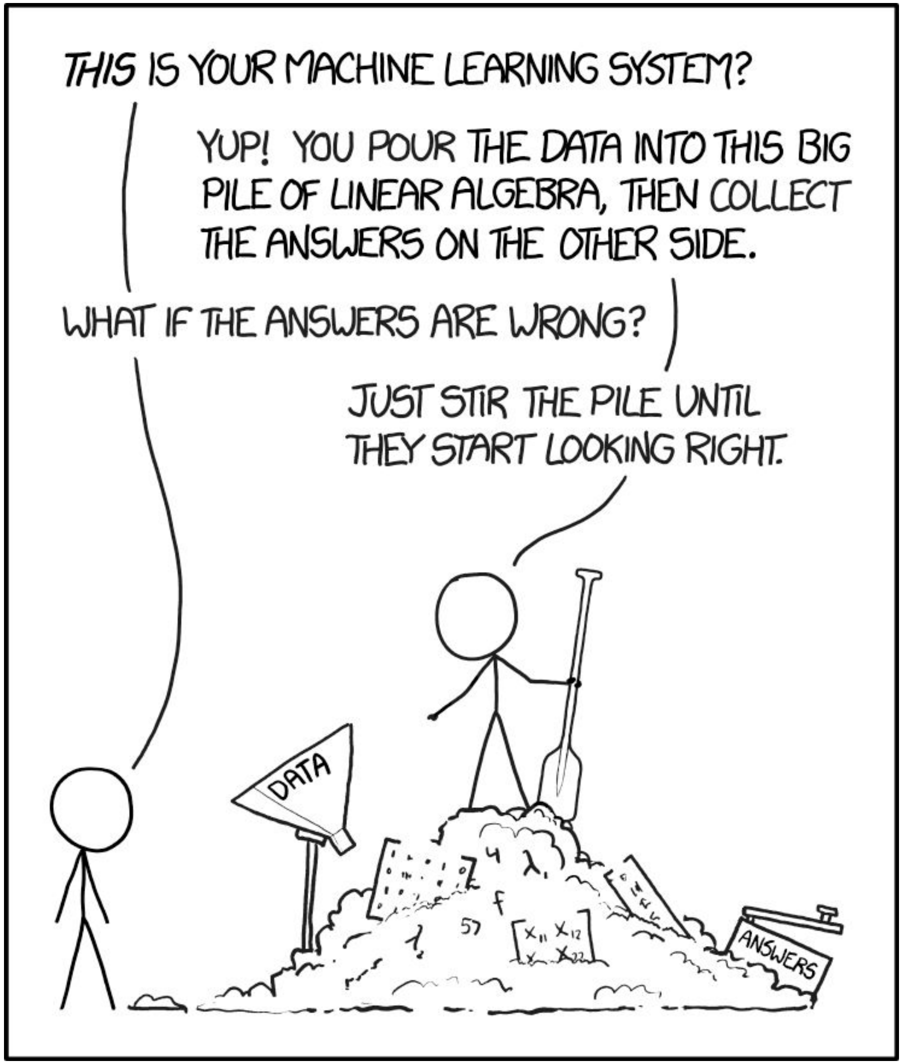
Data Science (& Machine Learning): re-branded statistics

Pitfalls (can be):

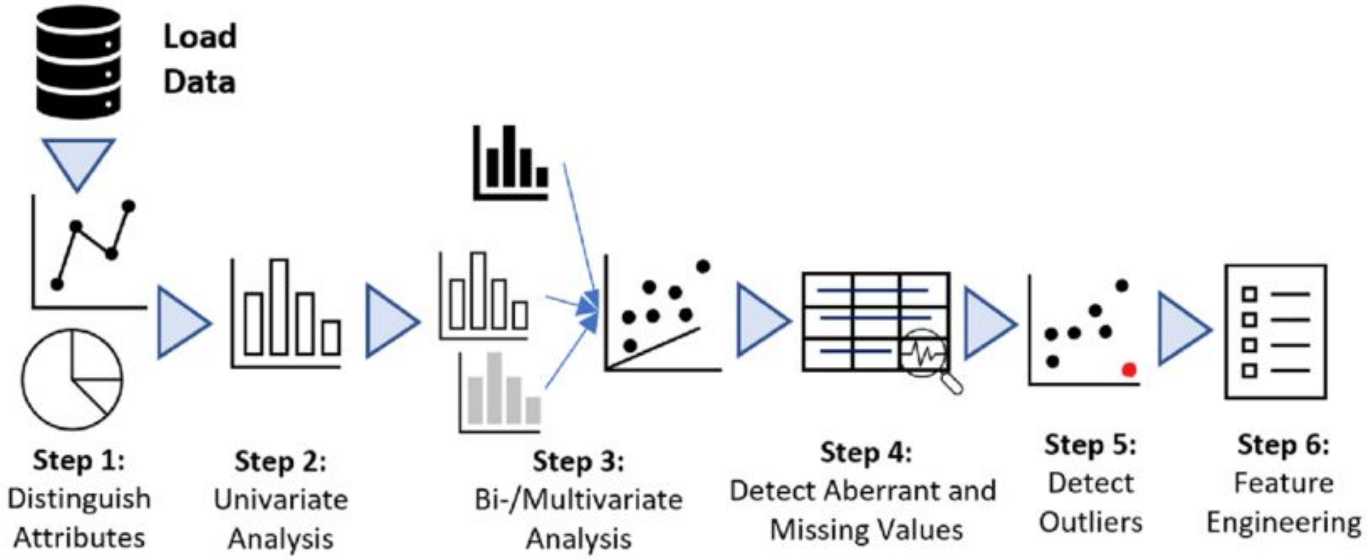
- Less rigorous/principled
- Prone to reinventing the wheel

Benefits (can be):

- More flexible
- Less prescriptive/intimidating


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.
DATA
ANSWERS

Data science centers exploratory data analysis


Load Data
Step 1: Distinguish Attributes
Step 2: Univariate Analysis
Step 3: Bi-/Multivariate Analysis
Step 4: Detect Aberrant and Missing Values
Step 5: Detect Outliers
Step 6: Feature Engineering

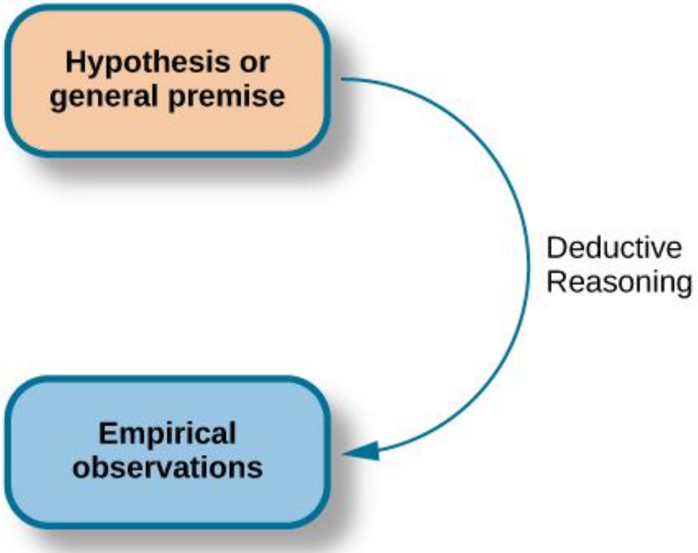
10.3390/su12124995

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



https://opened.cuny.edu/courseware/lesson/14/student/?task=3

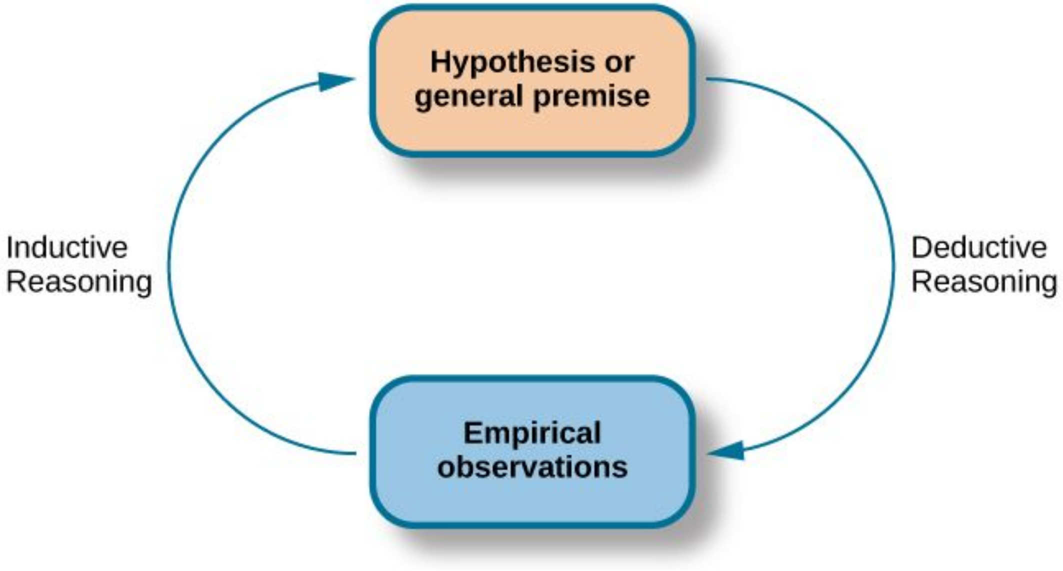
Data science supports inductive approaches

Deductive:

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

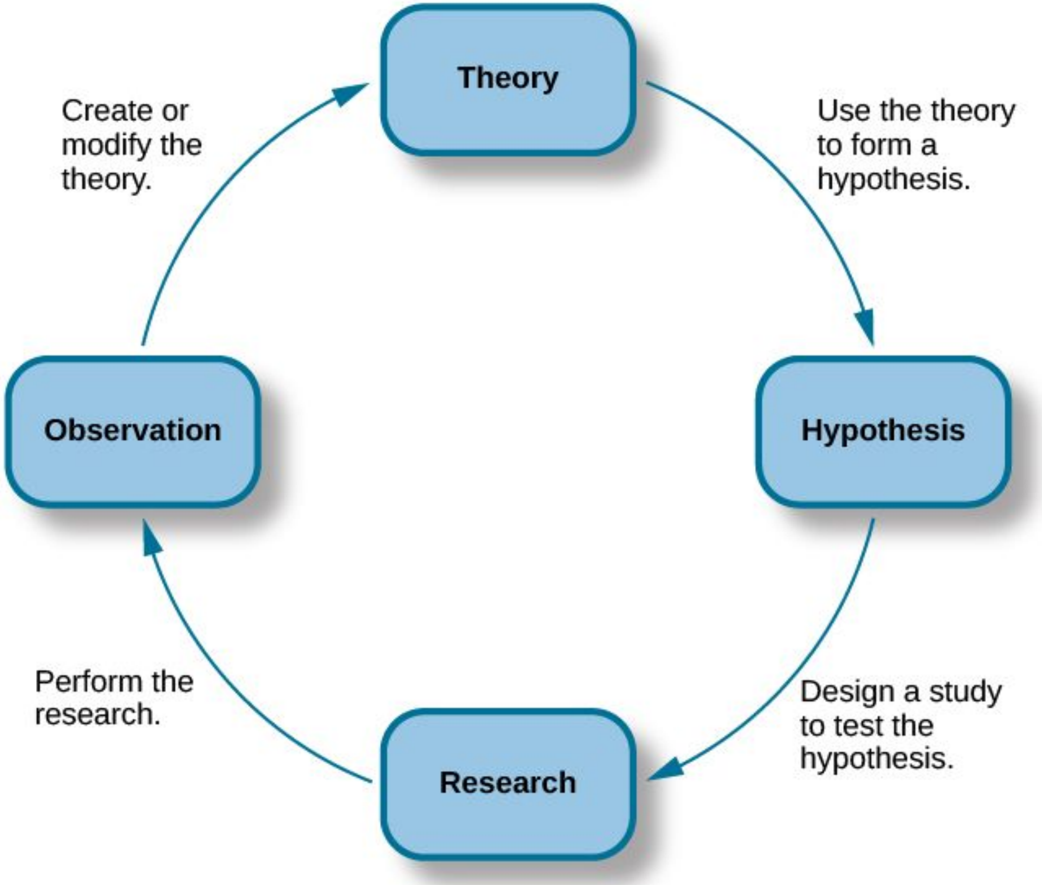
Inductive:

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



https://opened.cuny.edu/courseware/lesson/14/student/?task=3

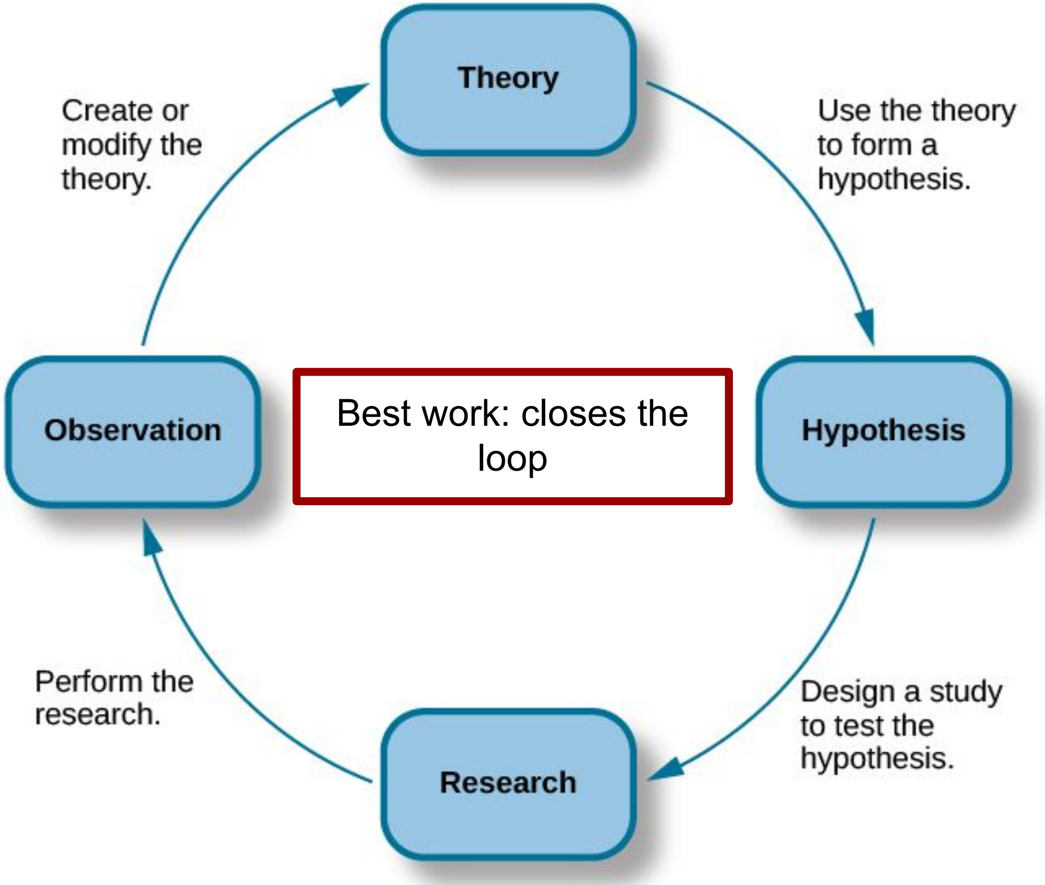
Data science is more realistic



Theory
Use the theory to form a hypothesis.
Hypothesis
Design a study to test the hypothesis.
Research
Perform the research.
Observation
Create or modify the theory.

https://opened.cuny.edu/courseware/lesson/14/student/?task=3

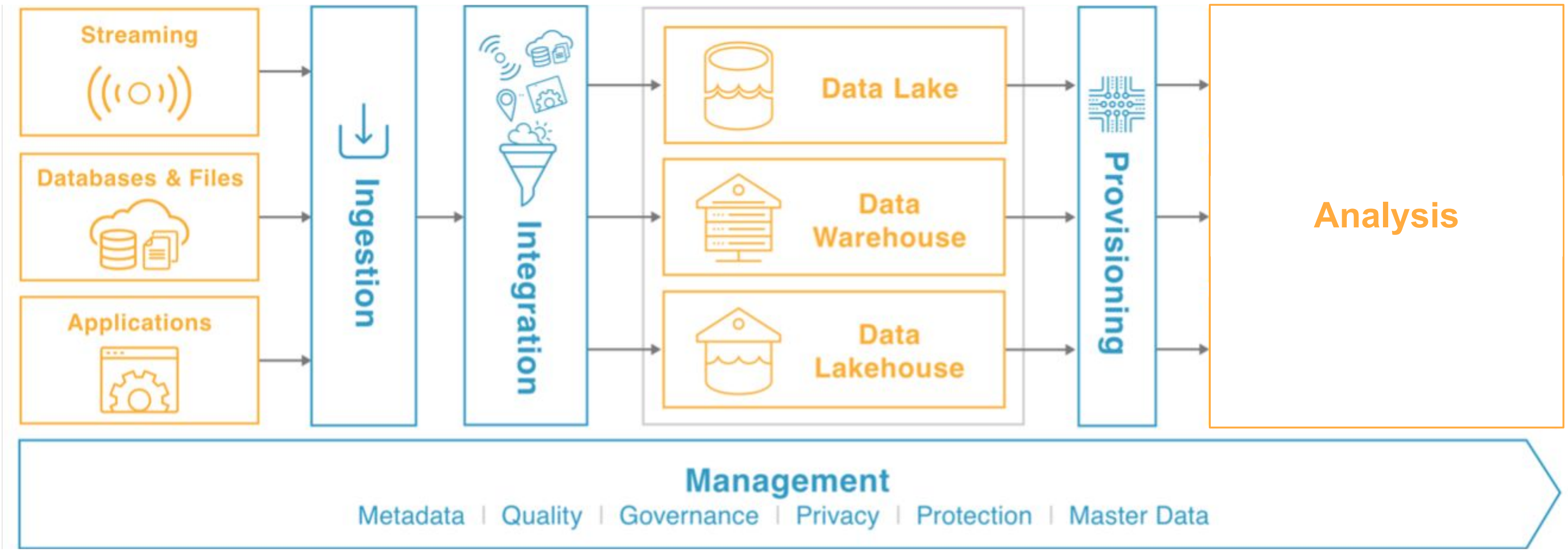
Data science is more realistic


Theory
Use the theory to form a hypothesis.
Hypothesis
Design a study to test the hypothesis.
Research
Perform the research.
Observation
Create or modify the theory.

Best work: closes the loop

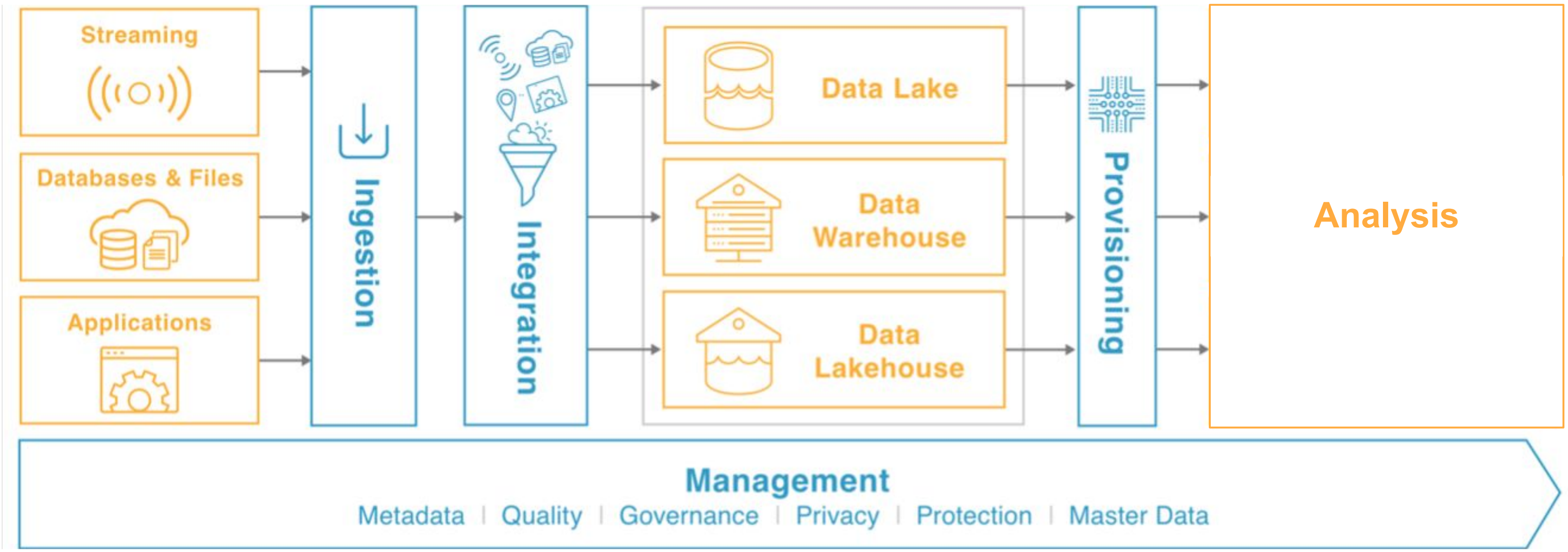
https://opened.cuny.edu/courseware/lesson/14/student/?task=3

Data science is integrated into a data ecosystem



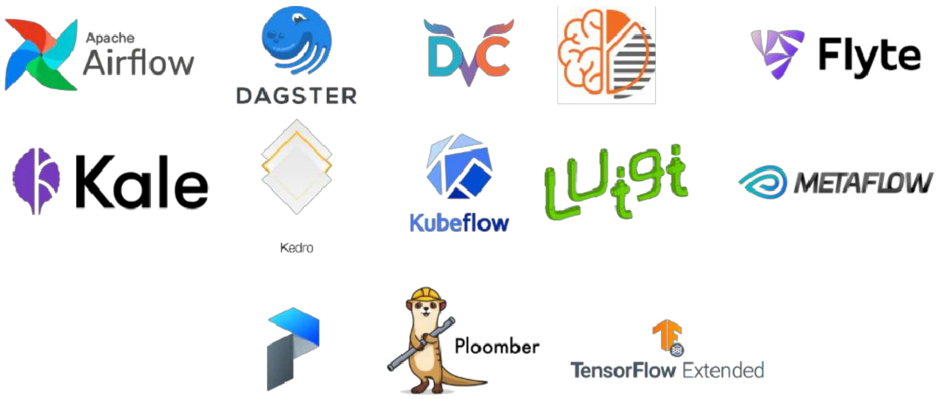
https://www.2ndwatch.com/blog/what-is-a-data-pipeline-and-how-to-build-one/

Data science is integrated into a data ecosystem



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

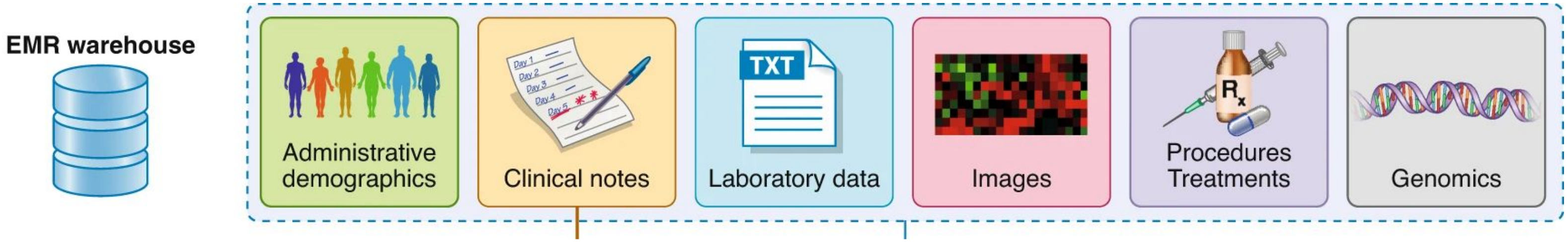
Some Open-Source Orchestration Tools:



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

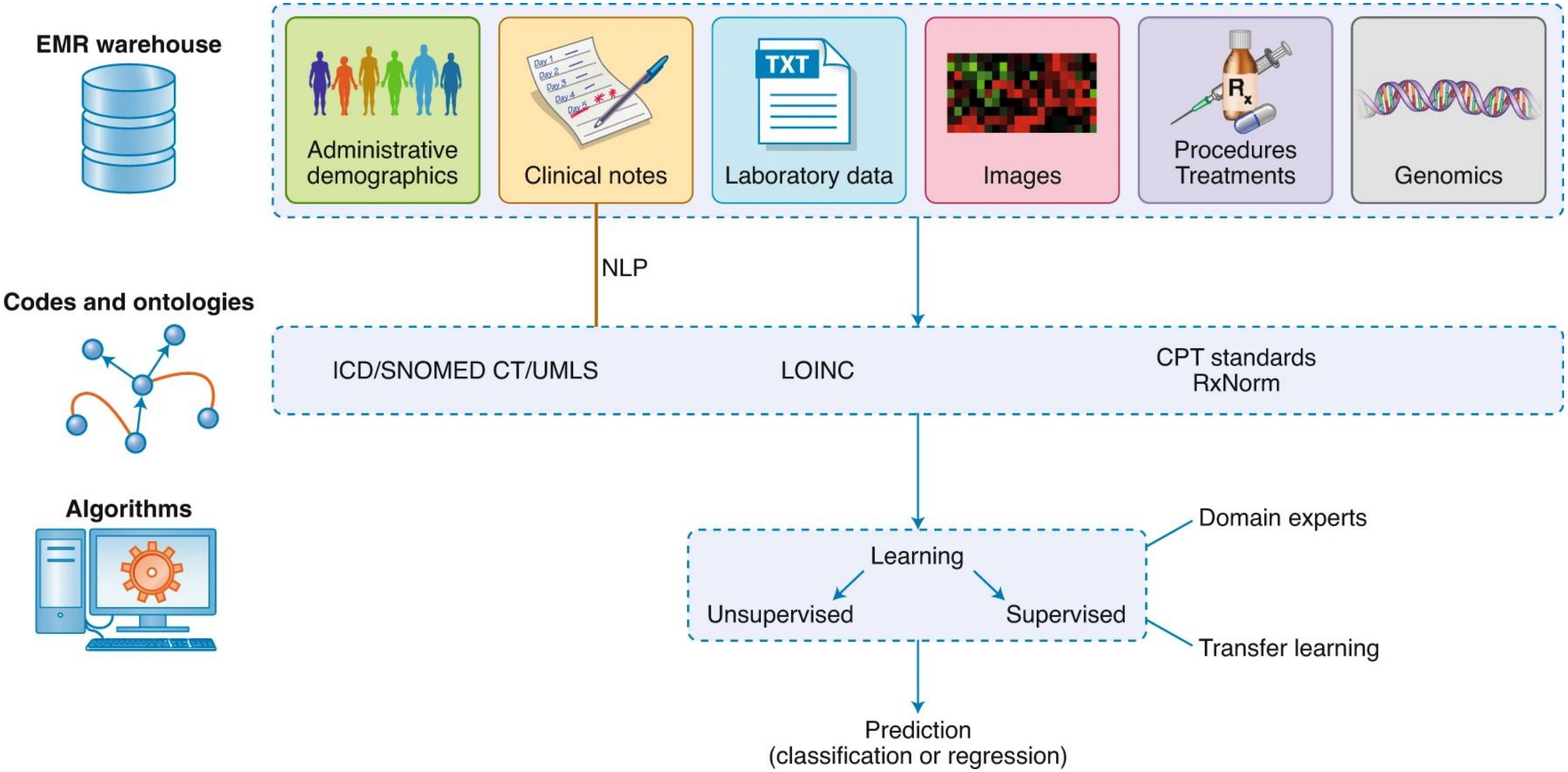
OK, what is **Health** Data Science?

Data Science applied to Health Data



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

Data Science applied to Health Data



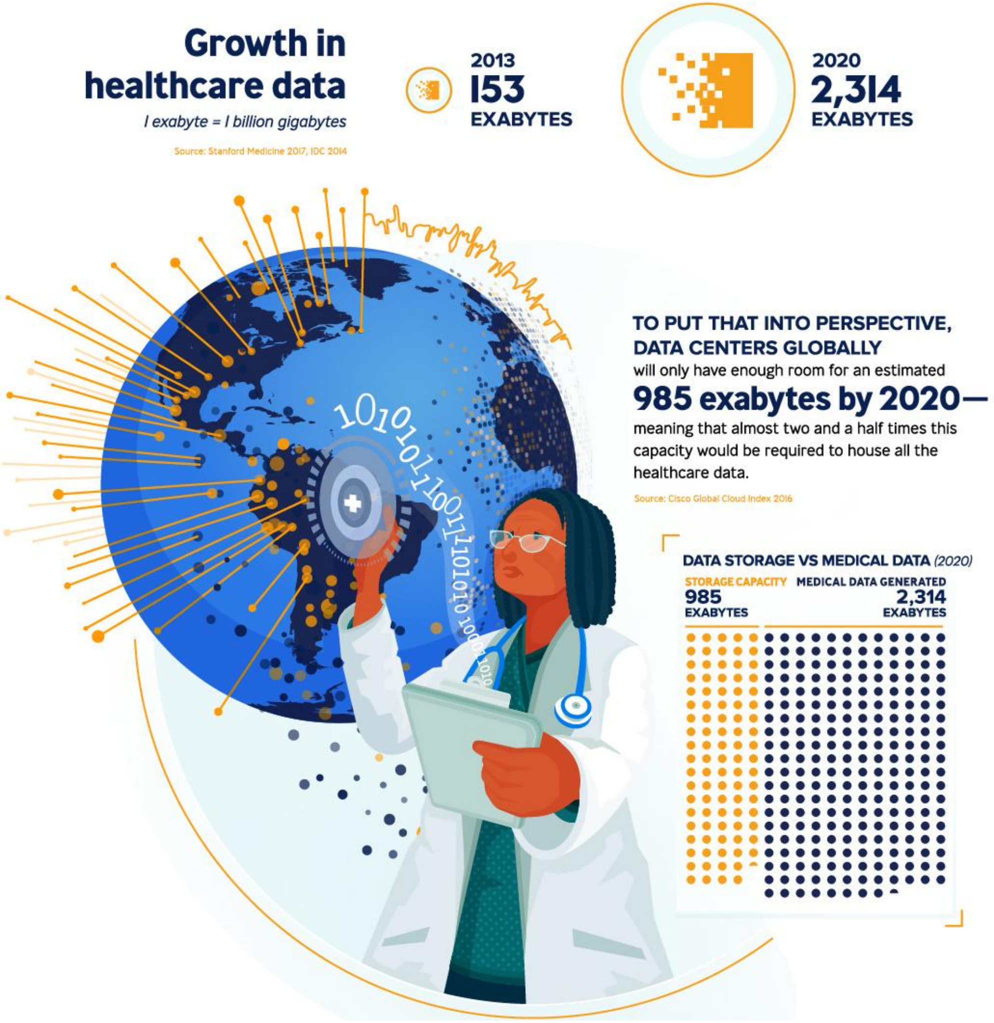
https://www.nature.com/articles/s41588-020-0698-y/figures/2

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


Growth in healthcare data
I exabyte = I billion gigabytes
Source: Stanford Medicine 2017, IDC 2014
2013
153
EXABYTES
2020
2,314
EXABYTES
1010101101100111011101010100101010

TO PUT THAT INTO PERSPECTIVE, DATA CENTERS GLOBALLY will only have enough room for an estimated **985 exabytes by 2020—** meaning that almost two and a half times this capacity would be required to house all the healthcare data.

Source: Cisco Global Cloud Index 2016

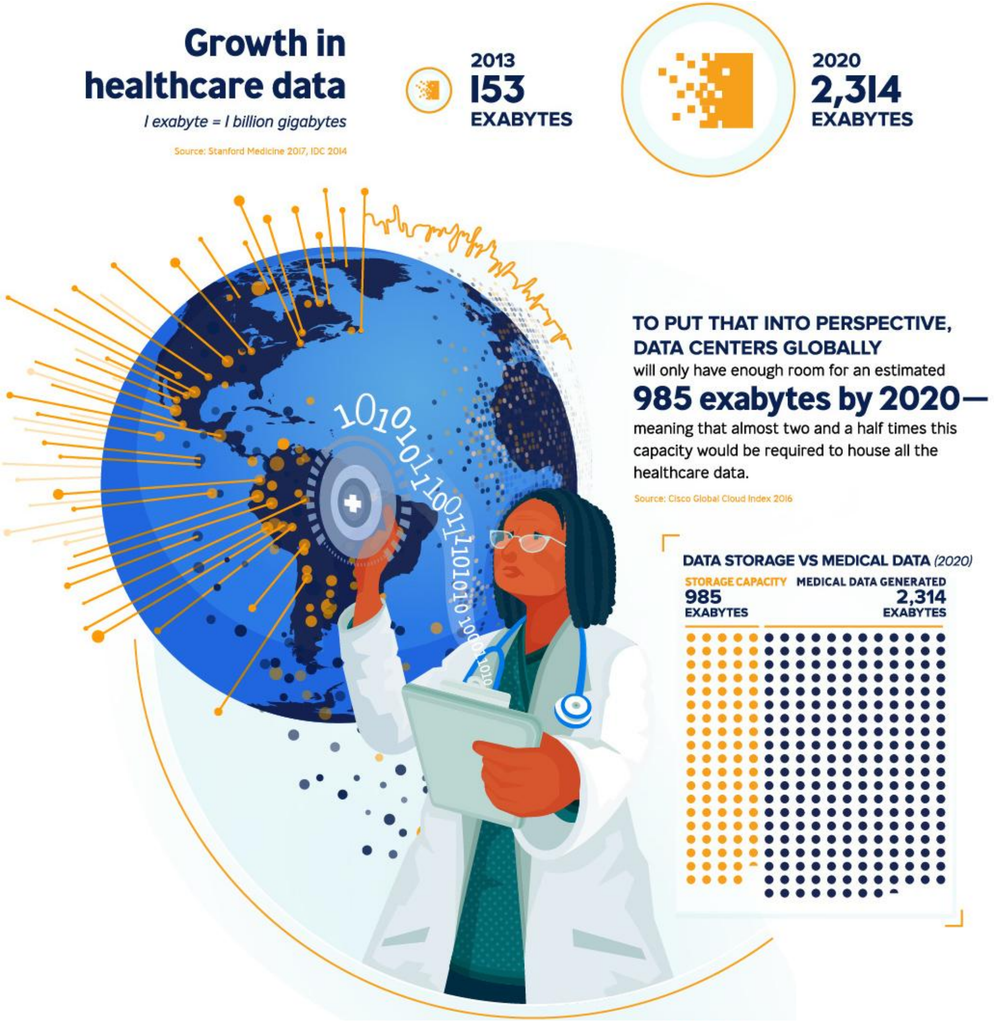
DATA STORAGE VS MEDICAL DATA (2020)
STORAGE CAPACITY
985
EXABYTES
MEDICAL DATA GENERATED
2,314
EXABYTES

https://www.visualcapitalist.com/big-data-healthcare/

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**



Growth in healthcare data

I exabyte = I billion gigabytes

Source: Stanford Medicine 2017, IDC 2014

TO PUT THAT INTO PERSPECTIVE, DATA CENTERS GLOBALLY will only have enough room for an estimated **985 exabytes by 2020**— meaning that almost two and a half times this capacity would be required to house all the healthcare data.

Source: Cisco Global Cloud Index 2016

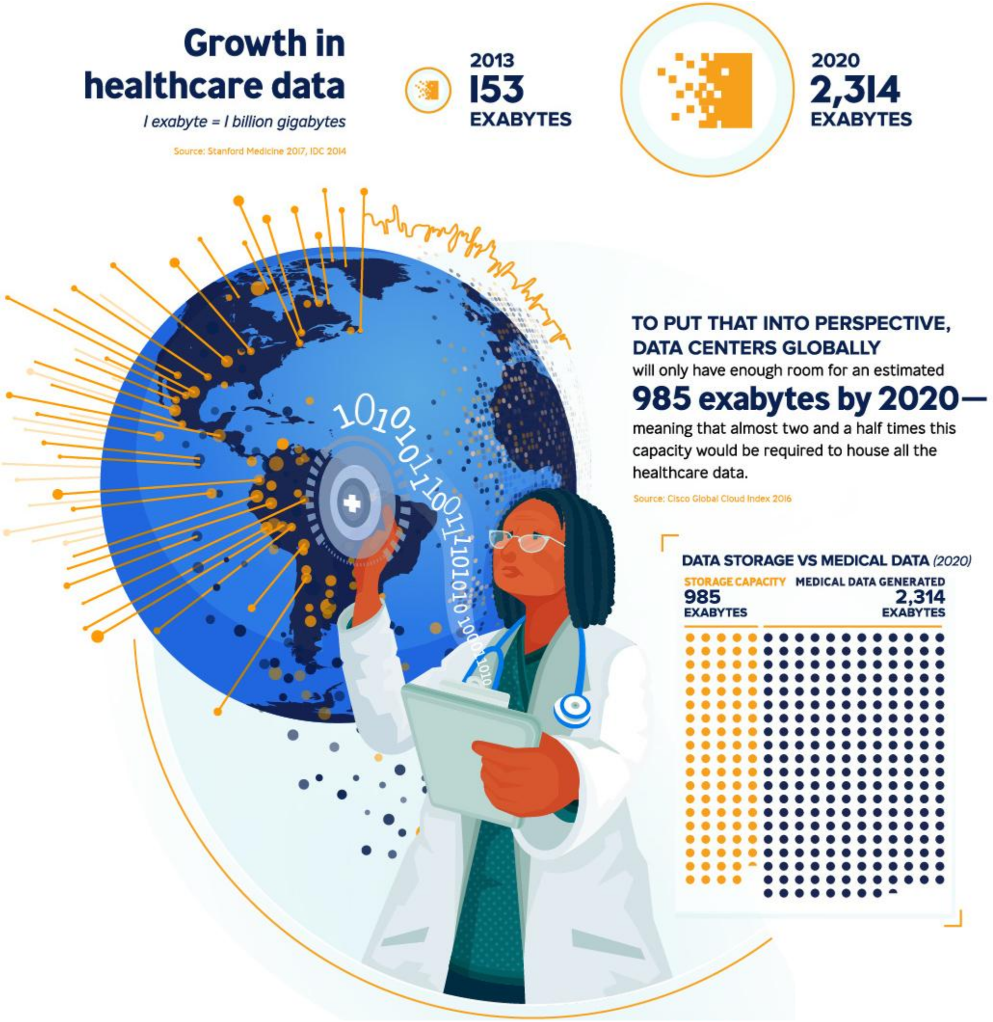
DATA STORAGE VS MEDICAL DATA *(2020)*	
STORAGE CAPACITY	MEDICAL DATA GENERATED
985 EXABYTES	2,314 EXABYTES

https://www.visualcapitalist.com/big-data-healthcare/

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


Growth in healthcare data
I exabyte = I billion gigabytes
Source: Stanford Medicine 2017, IDC 2014
2013
153
EXABYTES
2020
2,314
EXABYTES
1010101101100111010100

TO PUT THAT INTO PERSPECTIVE, DATA CENTERS GLOBALLY will only have enough room for an estimated **985 exabytes by 2020—** meaning that almost two and a half times this capacity would be required to house all the healthcare data.

Source: Cisco Global Cloud Index 2016

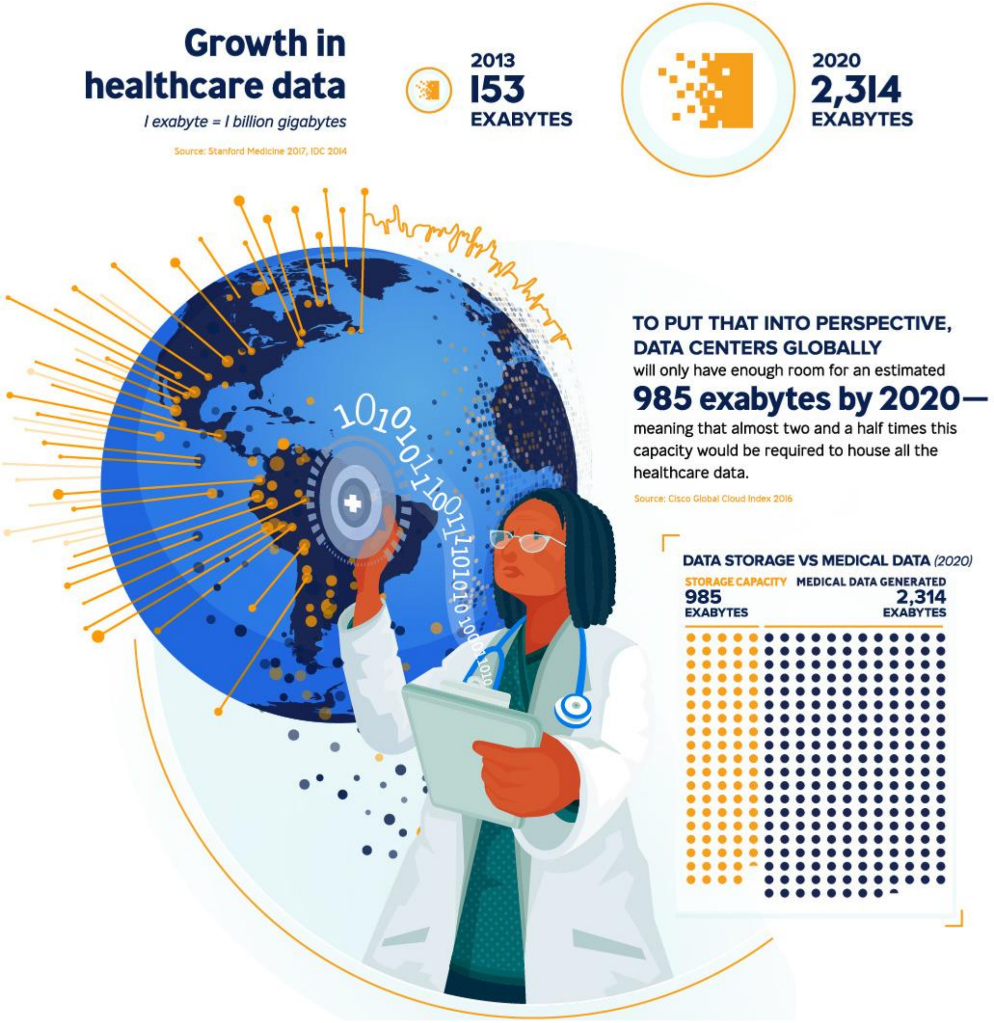
DATA STORAGE VS MEDICAL DATA (2020)
STORAGE CAPACITY
985
EXABYTES
MEDICAL DATA GENERATED
2,314
EXABYTES

https://www.visualcapitalist.com/big-data-healthcare/

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



Growth in healthcare data

1 exabyte = 1 billion gigabytes

Source: Stanford Medicine 2017, IDC 2014

TO PUT THAT INTO PERSPECTIVE, DATA CENTERS GLOBALLY will only have enough room for an estimated **985 exabytes by 2020—** meaning that almost two and a half times this capacity would be required to house all the healthcare data.

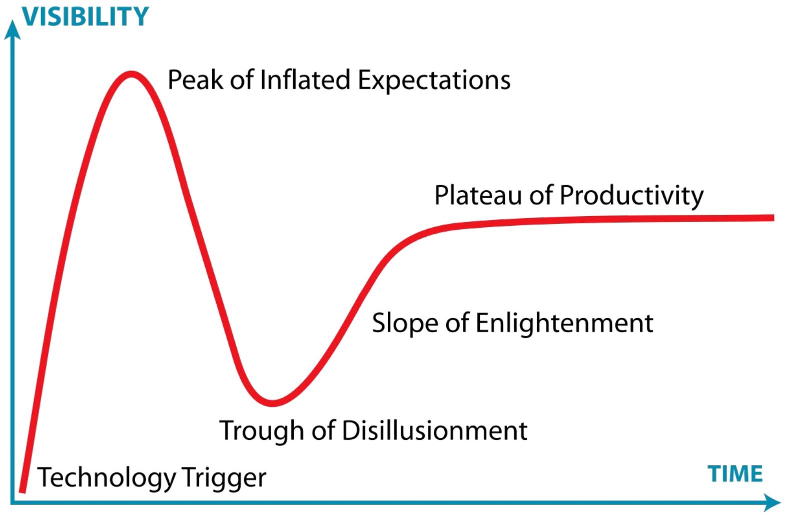
Source: Cisco Global Cloud Index 2016

DATA STORAGE VS MEDICAL DATA *(2020)*

https://www.visualcapitalist.com/big-data-healthcare/

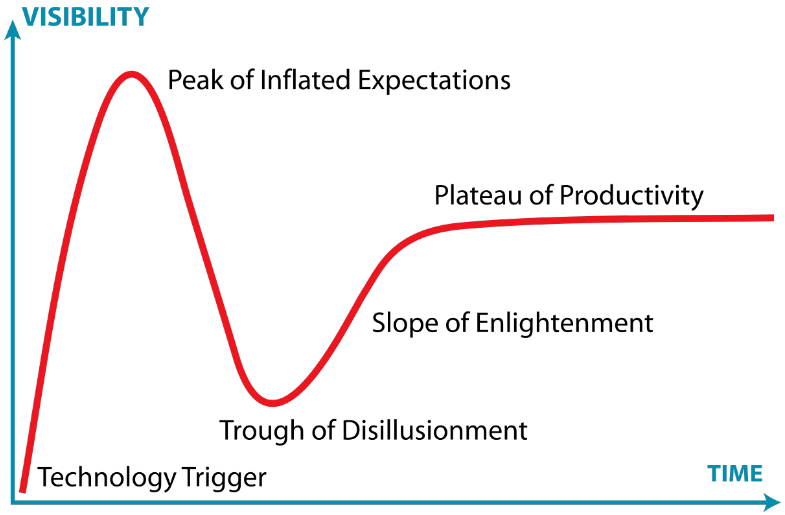
(Some) Challenges of Health Data Science

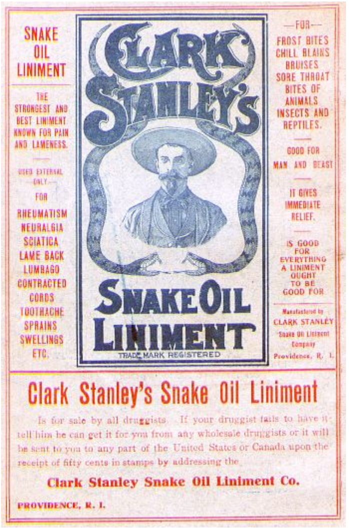
- Lots of hype



(Some) Challenges of Health Data Science

- Lots of hype
- Lots of grifters




SNAKE OIL LINIMENT
THE STRONGEST AND BEST LINIMENT KNOWN FOR PAIN AND LAMENESS.
USED EXTERNAL ONLY
FOR
RHEUMATISM
NEURALGIA
SCIATICA
LAME BACK
LUMBAGO
CONTRACTED CORDS
TOOTHACHE
SPRAINS
SWELLINGS
ETC.
CLARK STANLEY'S
SNAKE OIL LINIMENT
TRADE MARK REGISTERED
—FOR—
FROST BITES
CHILL BLAINS
BRUISES
SORE THROAT
BITES OF ANIMALS INSECTS AND REPTILES.
GOOD FOR MAN AND BEAST
IT GIVES IMMEDIATE RELIEF.
IS GOOD FOR EVERYTHING A LINIMENT OUGHT TO BE GOOD FOR
Manufactured by CLARK STANLEY Snake Oil Liniment Company Providence, R. I.
Clark Stanley's Snake Oil Liniment

Is for sale by all druggists. If your druggist fails to have it tell him he can get it for you from any wholesale druggists or it will be sent to you to any part of the United States or Canada upon the receipt of fifty cents in stamps by addressing the

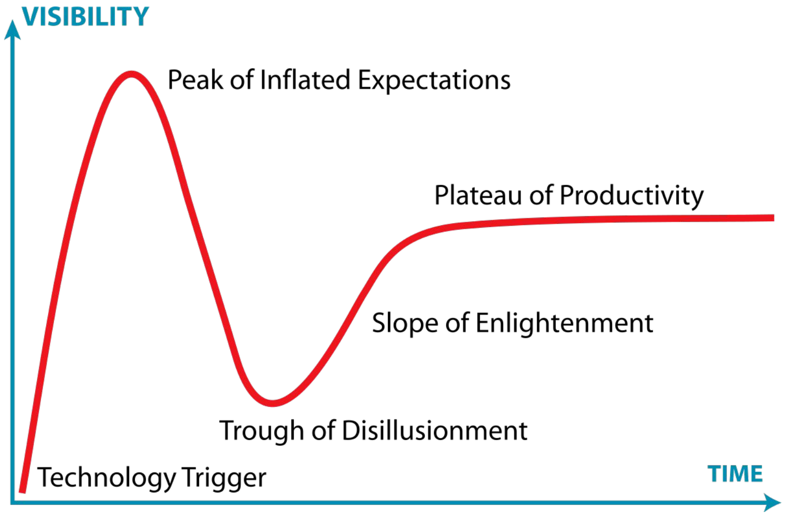
Clark Stanley Snake Oil Liniment Co.

PROVIDENCE, R. I.

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

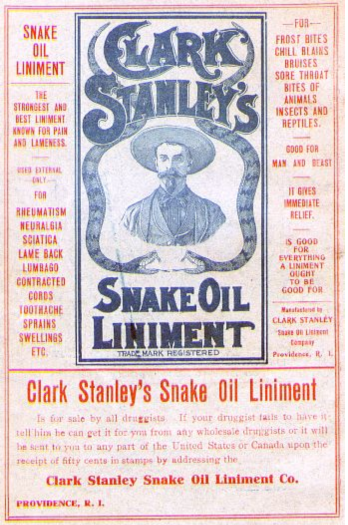
(Some) Challenges of Health Data Science

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



GARBAGE
DATA
GREAT
MODEL
GARBAGE
RESULTS

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



SNAKE OIL LINIMENT

THE STRONGEST AND BEST LINIMENT KNOWN FOR PAIN AND LAMENESS.

USED EXTERNAL ONLY

FOR RHEUMATISM NEURALGIA SCIATICA LAME BACK LUMBAGO CONTRACTED CORDS TOOTHACHE SPRAINS SWELLINGS ETC.

CLARK STANLEY'S

SNAKE OIL LINIMENT

TRADE MARK REGISTERED

—FOR— FROST BITES CHILL BLAINS BRUISES SORE THROAT BITES OF ANIMALS INSECTS AND REPTILES.

GOOD FOR MAN AND BEAST

IT GIVES IMMEDIATE RELIEF.

IS GOOD FOR EVERYTHING A LINIMENT OUGHT TO BE GOOD FOR

Manufactured by CLARK STANLEY Snake Oil Liniment Company Providence, R. I.

Clark Stanley's Snake Oil Liniment

Is for sale by all druggists. If your druggist fails to have it tell him he can get it for you from any wholesale druggists or it will be sent to you to any part of the United States or Canada upon the receipt of fifty cents in stamps by addressing the

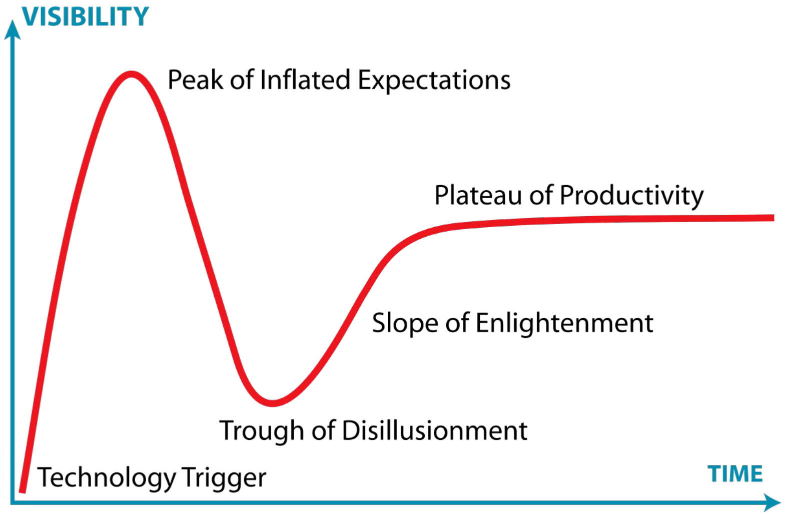
Clark Stanley Snake Oil Liniment Co.

PROVIDENCE, R. I.

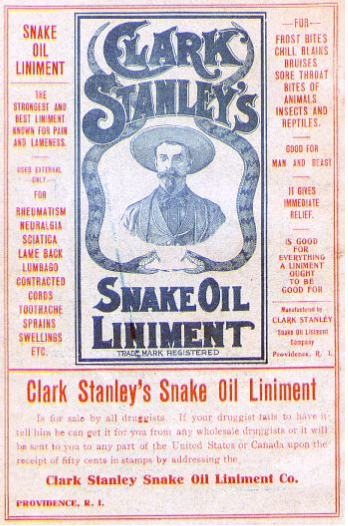
https://upload.wikimedia.org/wikipedia/commons/9/94/Gartner_Hype_Cycle.svg
https://commons.wikimedia.org/wiki/File:Clark_Stanley%27s_Snake_Oil_Liniment.png

(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/



SNAKE OIL LINIMENT

THE STRONGEST AND BEST LINIMENT KNOWN FOR PAIN AND LAMENESS.

USED EXTERNAL ONLY

FOR RHEUMATISM NEURALGIA SCIATICA LAME BACK LUMBAGO CONTRACTED CORDS TOOTHACHE SPRAINS SWELLINGS ETC.

CLARK STANLEY'S SNAKE OIL LINIMENT

TRADE MARK REGISTERED

—FOR— FROST BITES CHILL BLAINS BRUISES SORE THROAT BITES OF ANIMALS INSECTS AND REPTILES.

GOOD FOR MAN AND BEAST

IT GIVES IMMEDIATE RELIEF.

IS GOOD FOR EVERYTHING A LINIMENT OUGHT TO BE GOOD FOR

Manufactured by CLARK STANLEY Snake Oil Liniment Company Providence, R. I.

Clark Stanley's Snake Oil Liniment

Is for sale by all druggists. If your druggist fails to have it tell him he can get it for you from any wholesale druggists or it will be sent to you to any part of the United States or Canada upon the receipt of fifty cents in stamps by addressing the

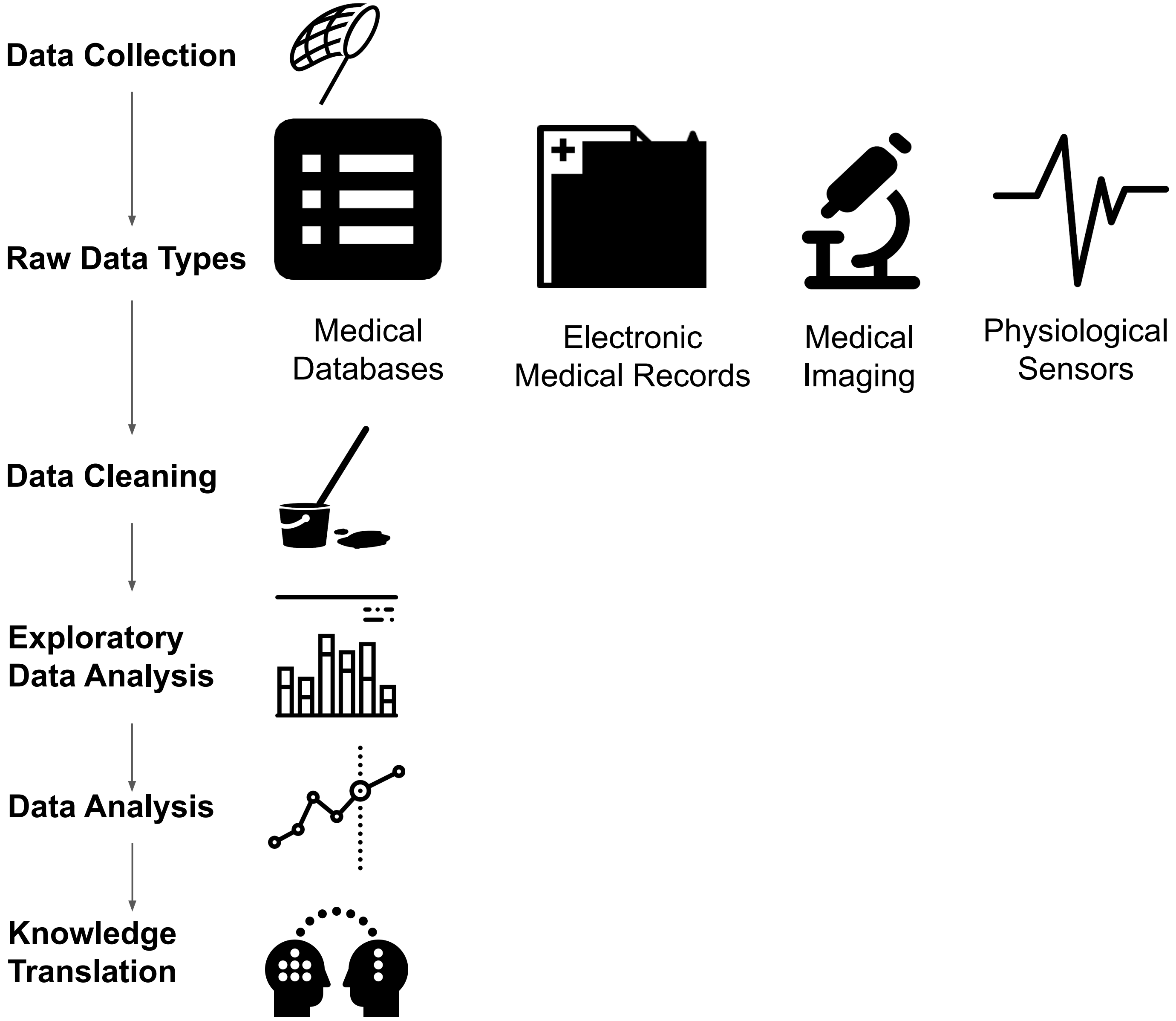
Clark Stanley Snake Oil Liniment Co.

PROVIDENCE, R. I.

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?



Data Collection

Raw Data Types

Medical Databases

Electronic Medical Records

Medical Imaging

Physiological Sensors

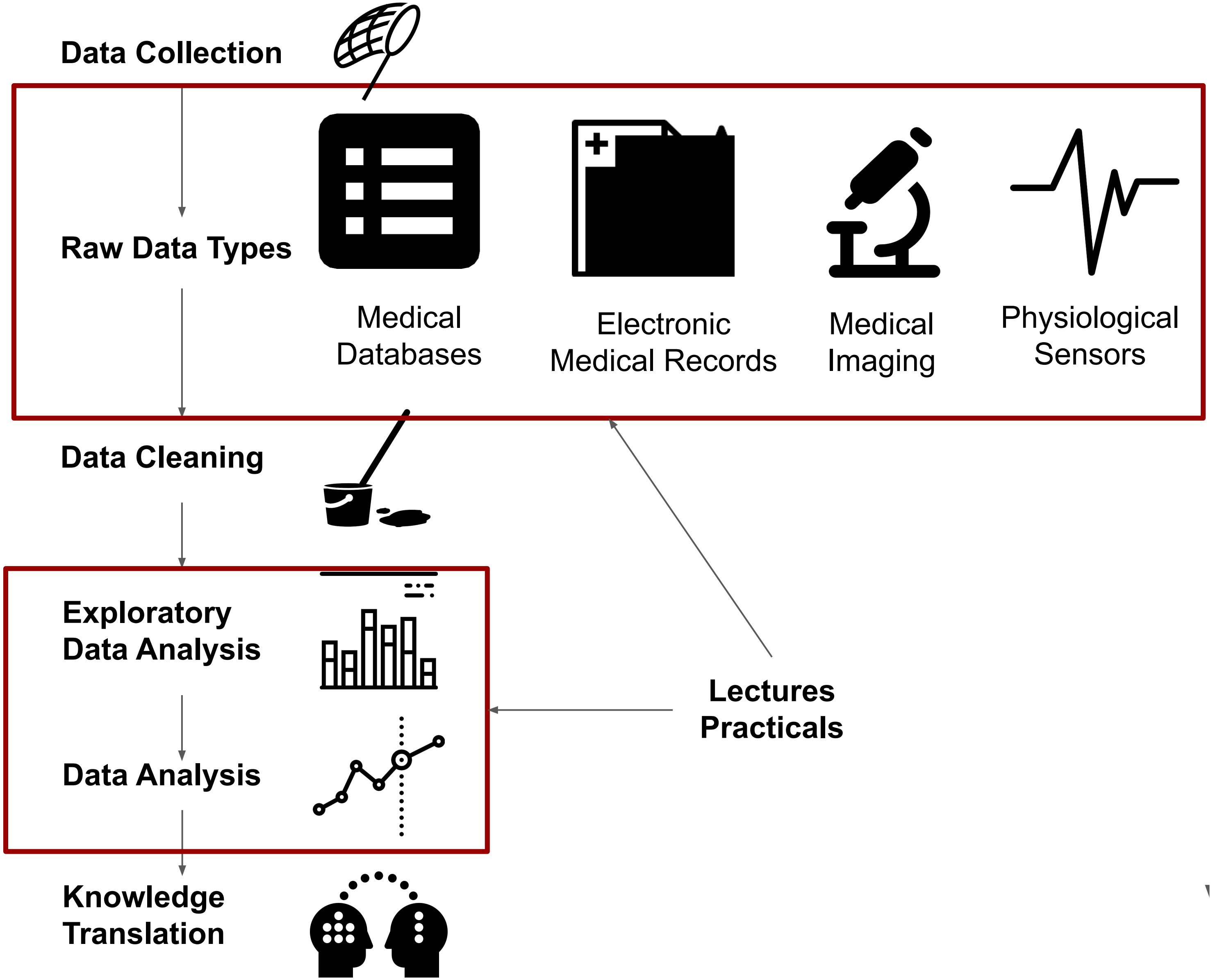
Data Cleaning

Exploratory Data Analysis

Data Analysis

Knowledge Translation

What parts of health data science will this course cover?



Data Collection

Raw Data Types

Medical Databases

Electronic Medical Records

Medical Imaging

Physiological Sensors

Data Cleaning

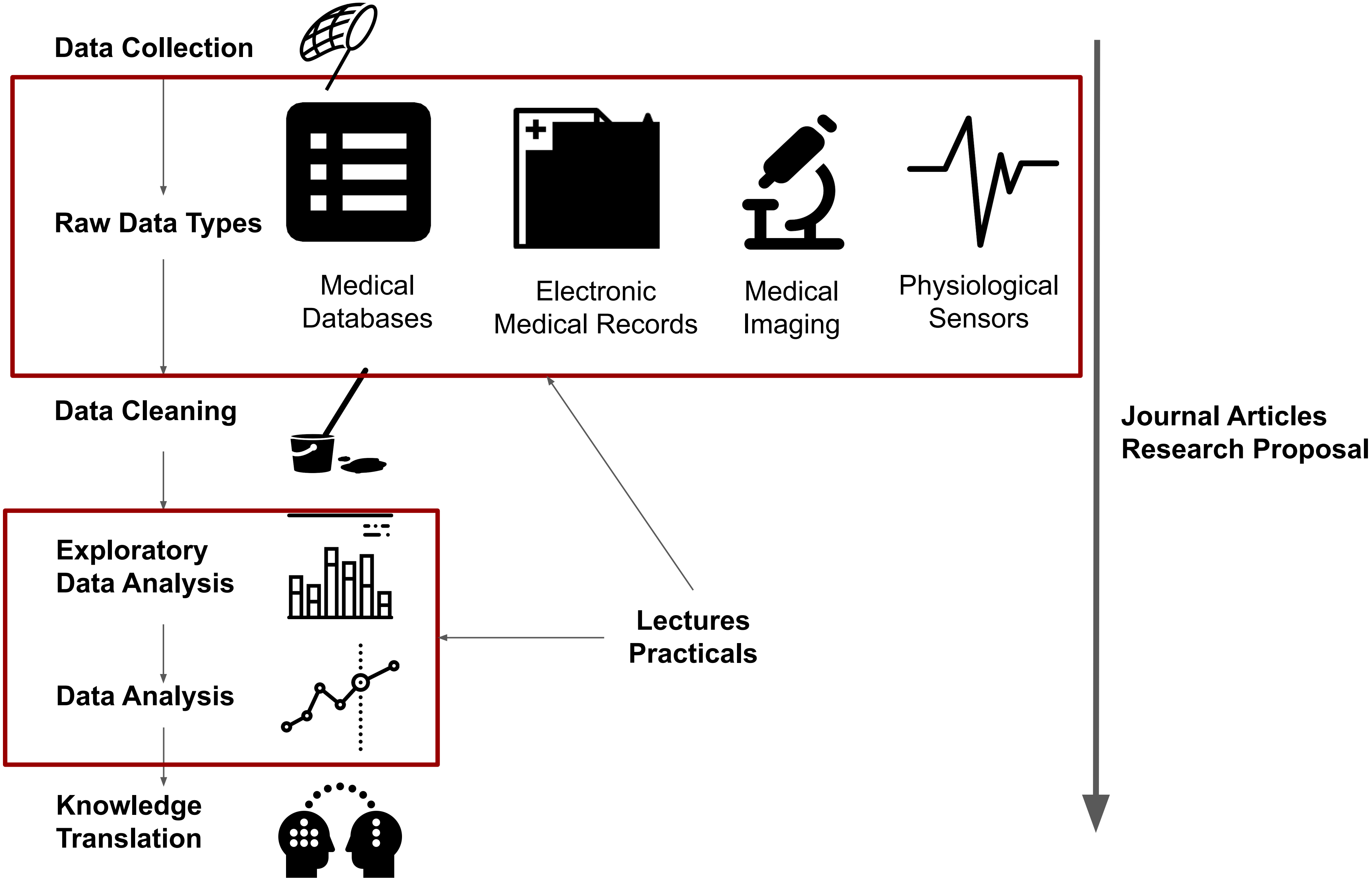
Exploratory Data Analysis

Data Analysis

Knowledge Translation

Lectures Practicals

What parts of health data science will this course cover?



Data Collection

Raw Data Types

Medical Databases

Electronic Medical Records

Medical Imaging

Physiological Sensors

Data Cleaning

Exploratory Data Analysis

Data Analysis

Knowledge Translation

Lectures
Practicals

Journal Articles
Research Proposal

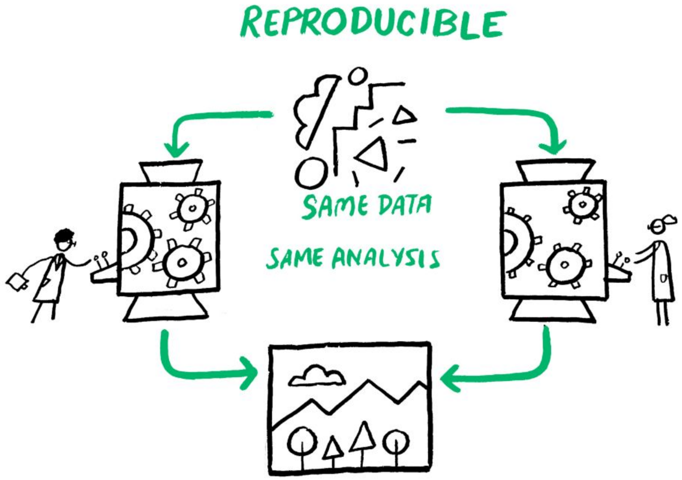
Let’s take a 5 minute break!

Tools for Reproducible Health Data Science

Rstudio, Rmarkdown, Git

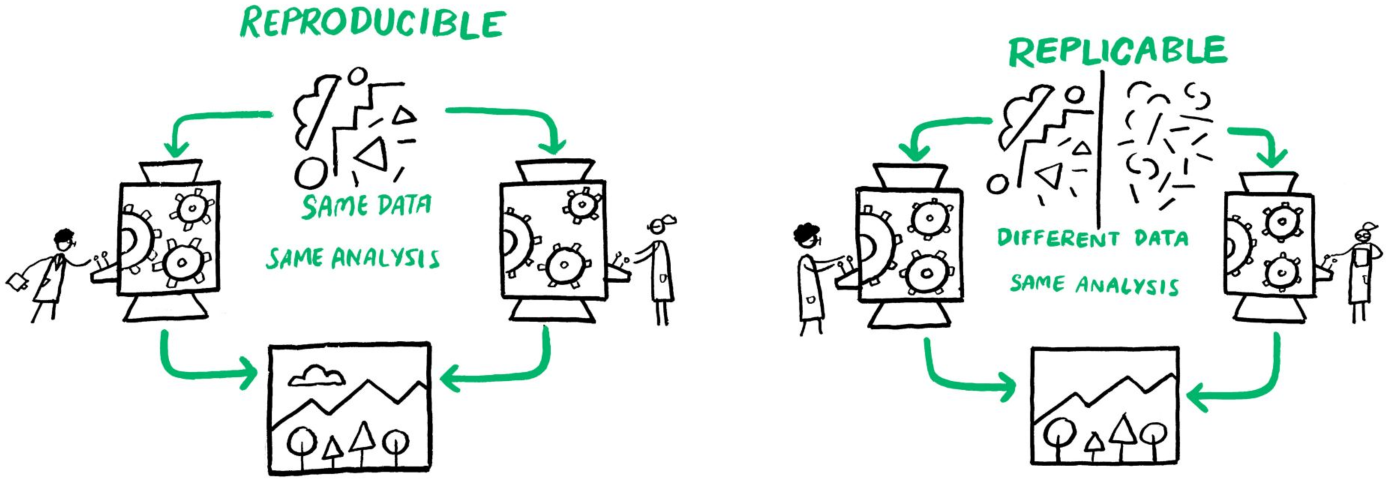
Why do we care about reproducibility?

Reproducibility should be the bare minimum



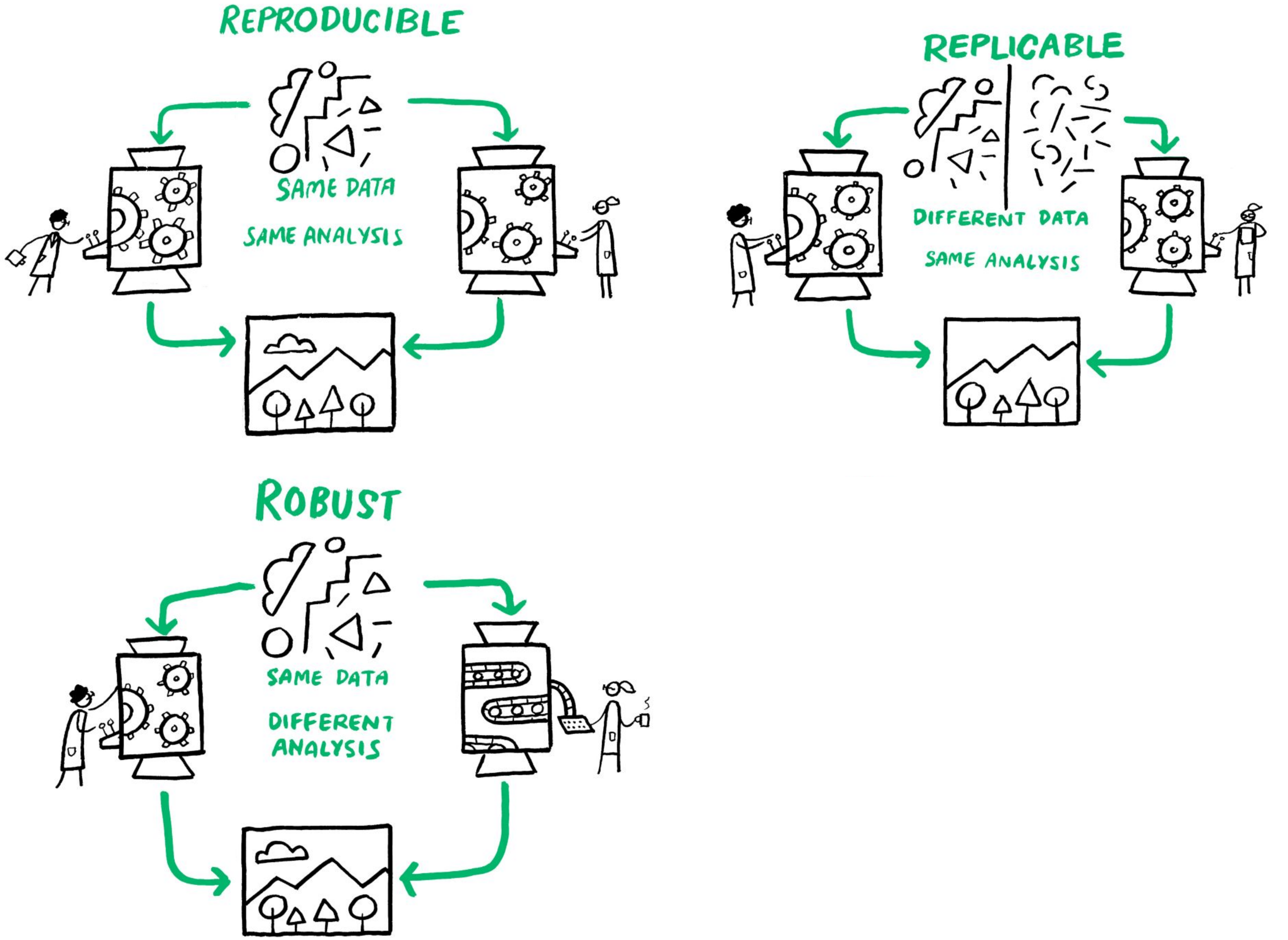
oliviergimenez.github.io/reproducible-science-workshop

Reproducibility should be the bare minimum



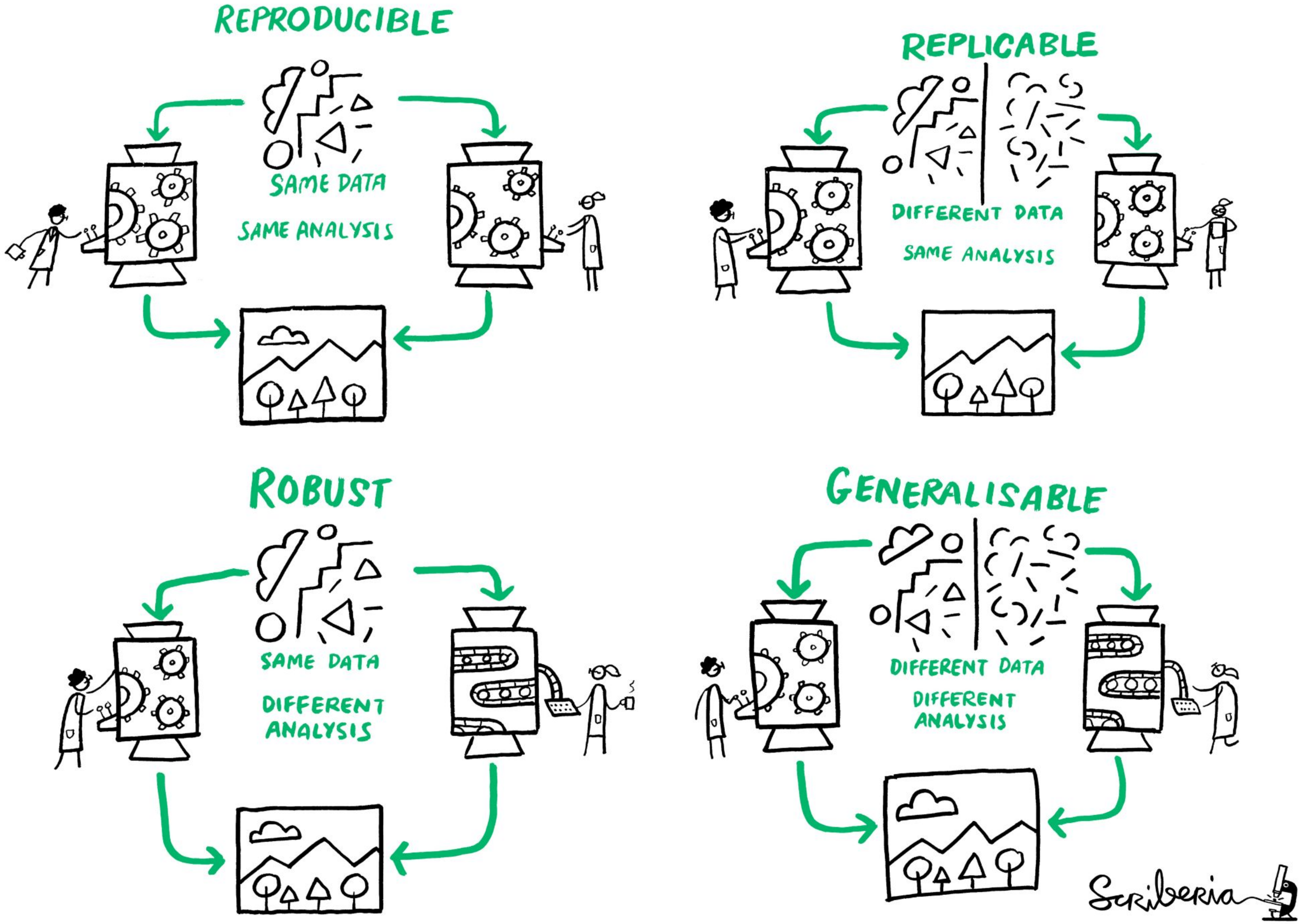
oliviergimenez.github.io/reproducible-science-workshop

Reproducibility should be the bare minimum



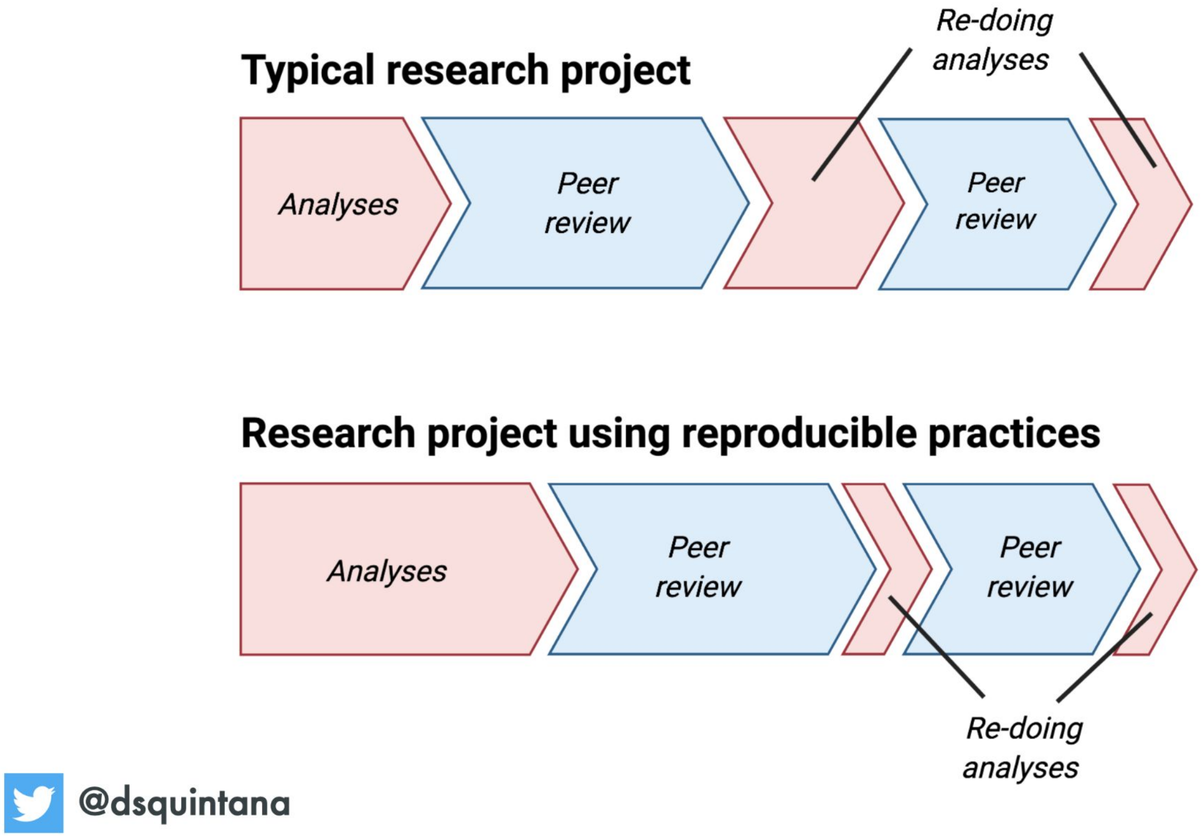
oliviergimenez.github.io/reproducible-science-workshop

Reproducibility should be the bare minimum



oliviergimenez.github.io/reproducible-science-workshop

Makes your own life easier



oliviergimenez.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

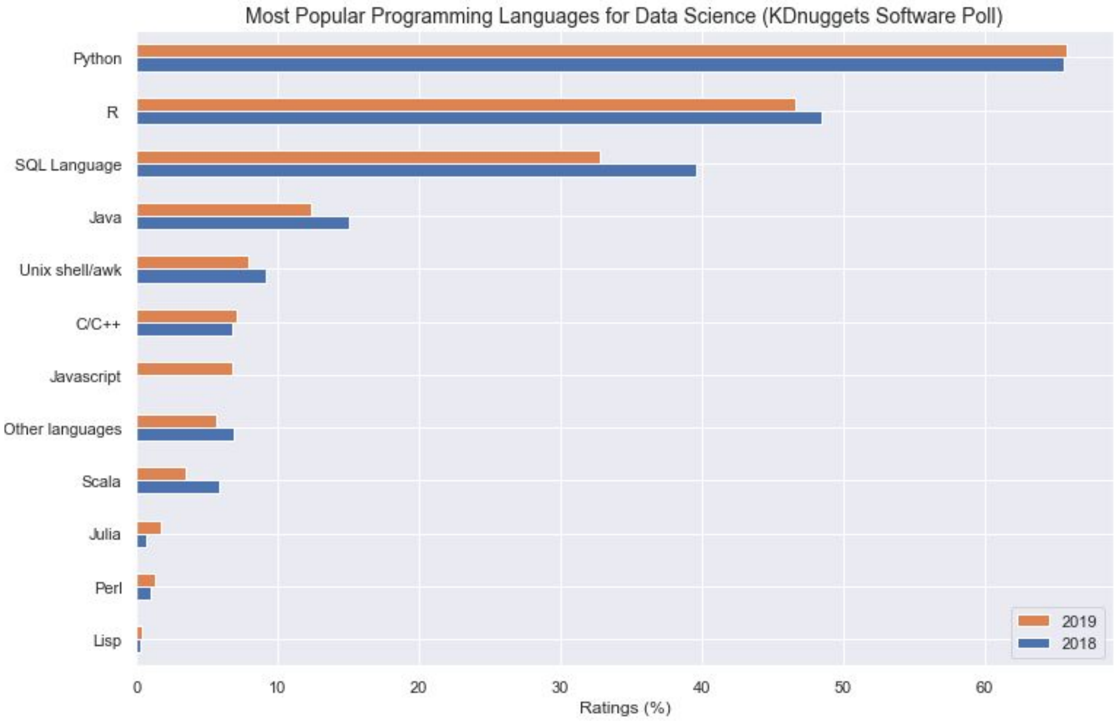
Reproducibility checklist

- 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?

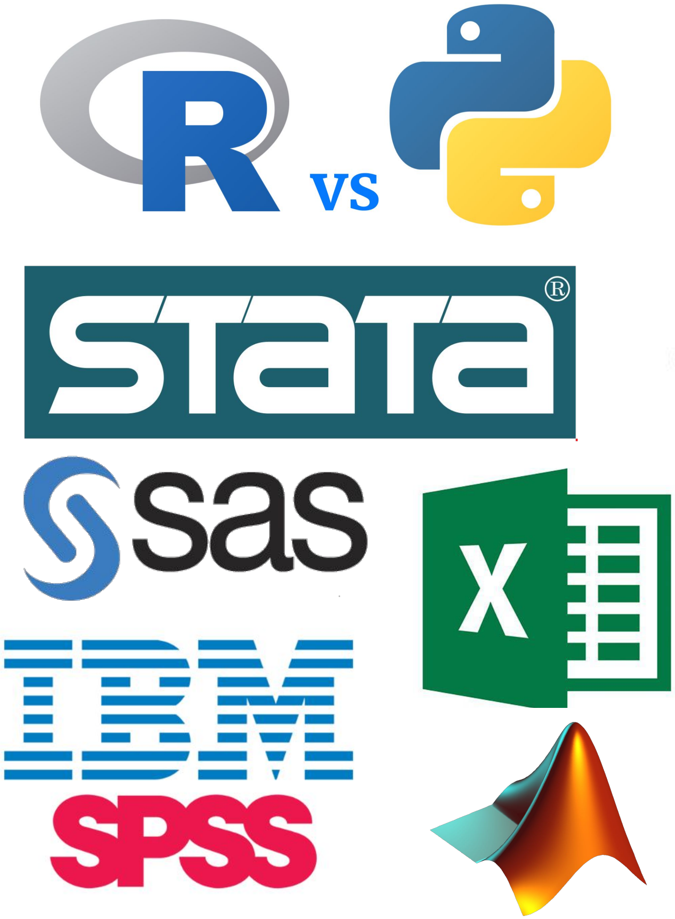
Choose a language that makes it easy to do most/all of your analysis

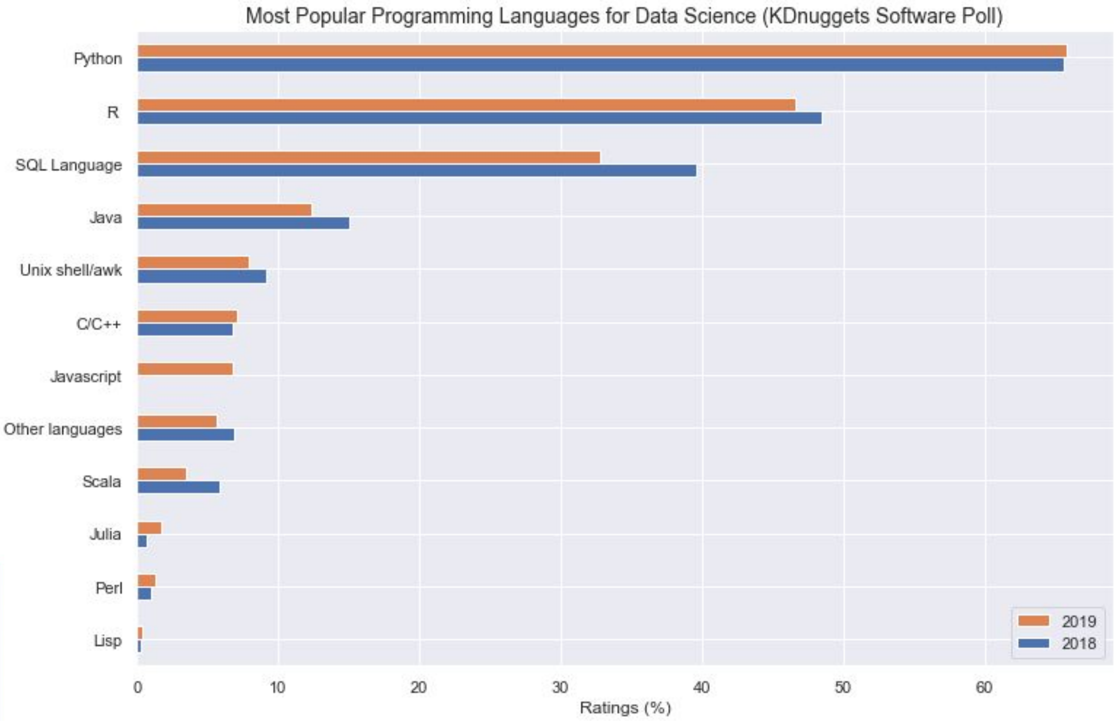
Choose a language that makes it easy to do most/all of your analysis



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

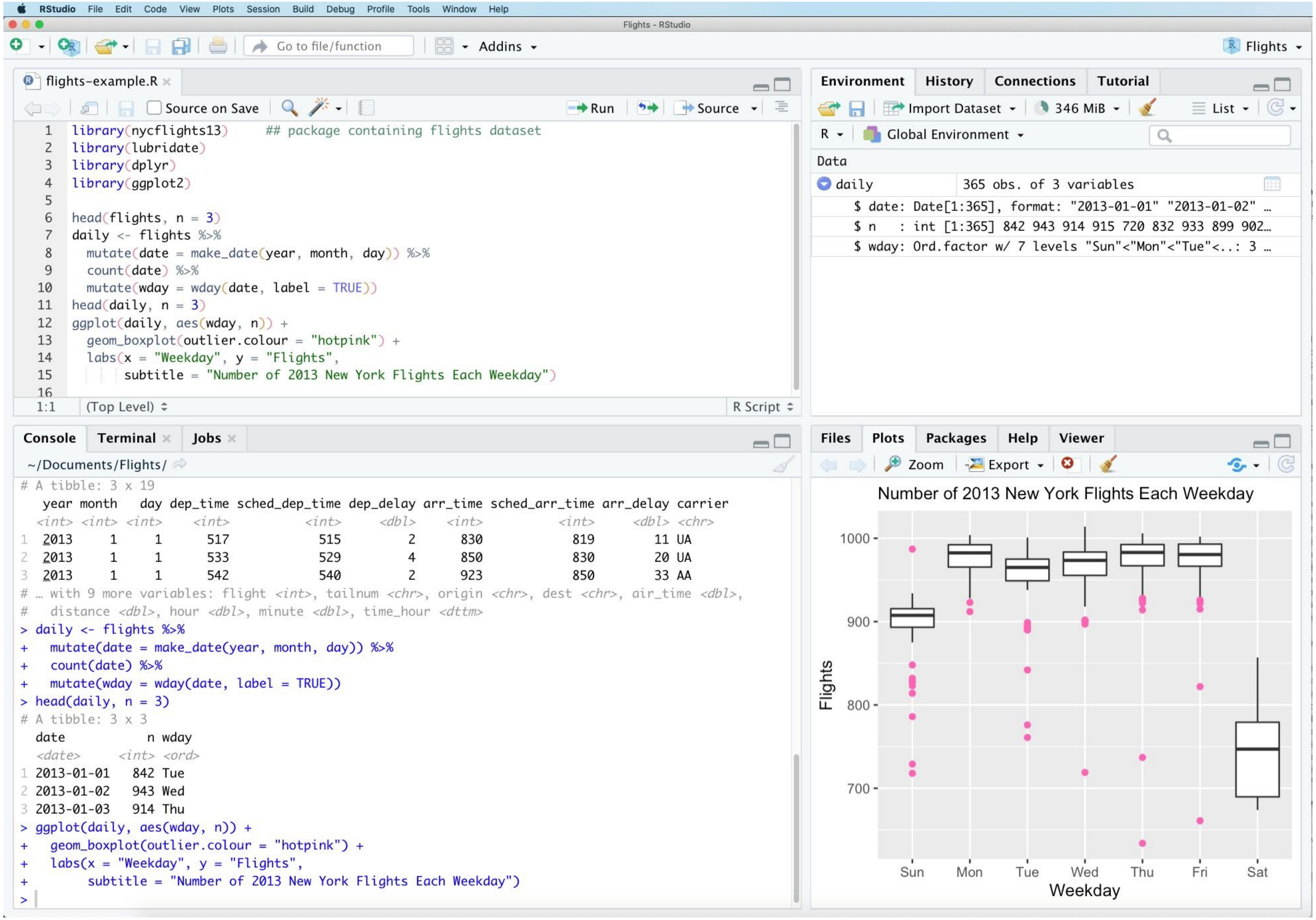




https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

Use a data science focused IDE: Rstudio

set.seed()
sessionInfo()



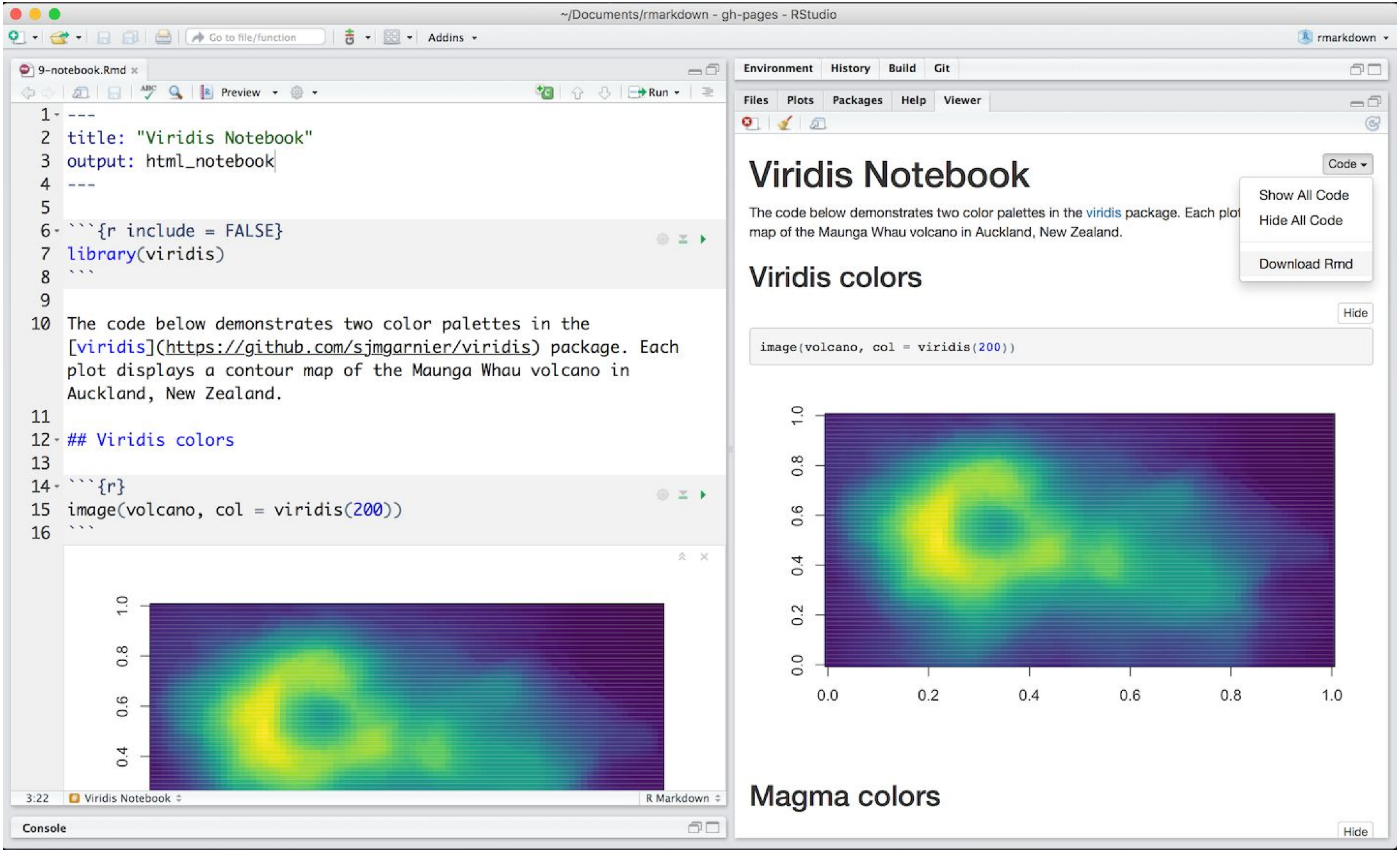
```r
library(nycflights13)    ## package containing flights dataset
library(lubridate)
library(dplyr)
library(ggplot2)

head(flights, n = 3)
daily <- flights %>%
  mutate(date = make_date(year, month, day)) %>%
  count(date) %>%
  mutate(wday = wday(date, label = TRUE))
head(daily, n = 3)
ggplot(daily, aes(wday, n)) +
  geom_boxplot(outlier.colour = "hotpink") +
  labs(x = "Weekday", y = "Flights",
       subtitle = "Number of 2013 New York Flights Each Weekday")
```

Data

```
daily          365 obs. of 3 variables
  $ date: Date[1:365], format: "2013-01-01" "2013-01-02" …
  $ n   : int [1:365] 842 943 914 915 720 832 933 899 902…
  $ wday: Ord.factor w/ 7 levels "Sun"<"Mon"<"Tue"<..: 3 …
```

```
# 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 <dttm>
> 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")
>
```

Use notebooks to document analyses: Rmarkdown



~~~~
---
title: "Viridis Notebook"
output: html_notebook
---

```{r include = FALSE}
library(viridis)
```


The code below demonstrates two color palettes in the
[viridis](https://github.com/sjmgarnier/viridis) package. Each
plot displays a contour map of the Maunga Whau volcano in
Auckland, New Zealand.

## Viridis colors

```{r}
image(volcano, col = viridis(200))
```
~~~~

Viridis Notebook

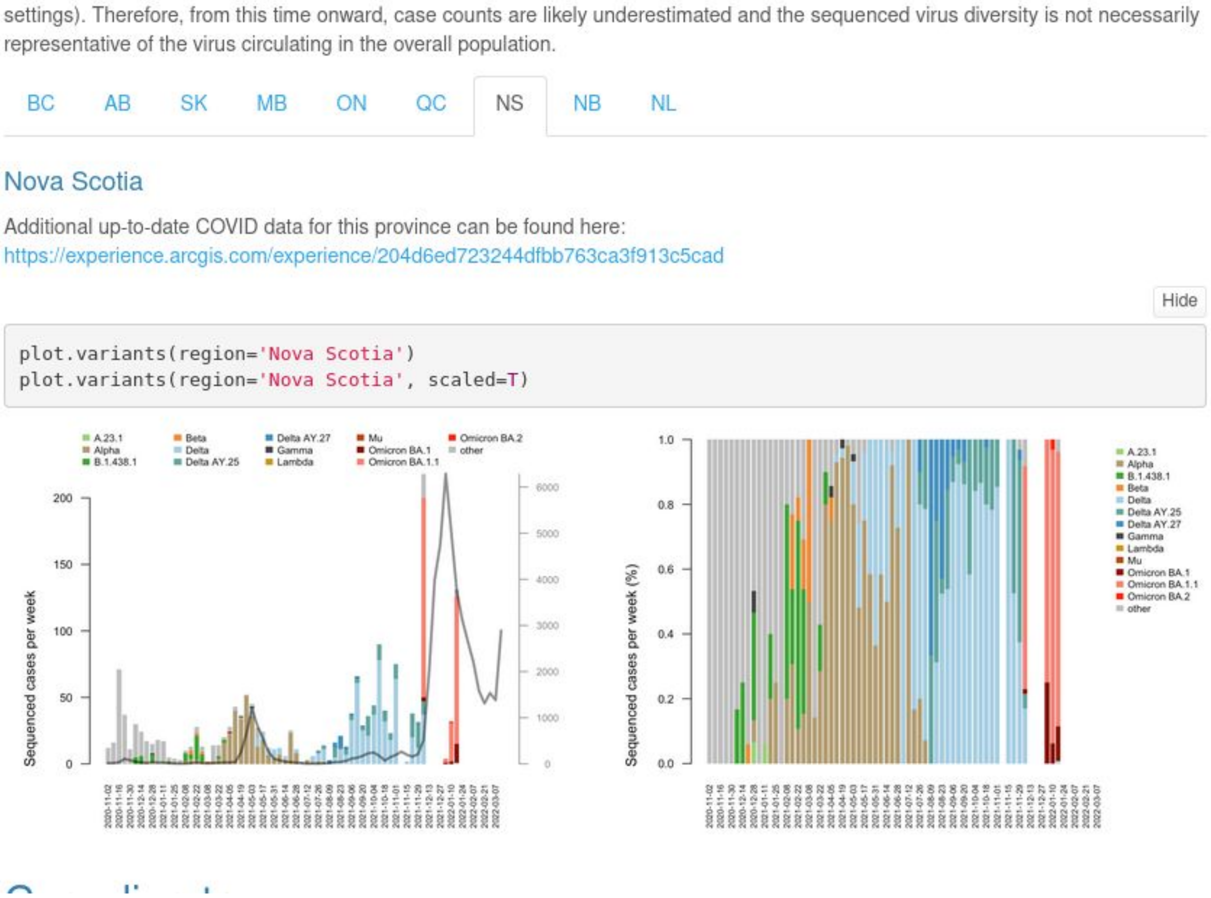
The code below demonstrates two color palettes in the viridis package. Each plot map of the Maunga Whau volcano in Auckland, New Zealand.

Viridis colors

```
image(volcano, col = viridis(200))
```

Magma colors

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.

Nova Scotia

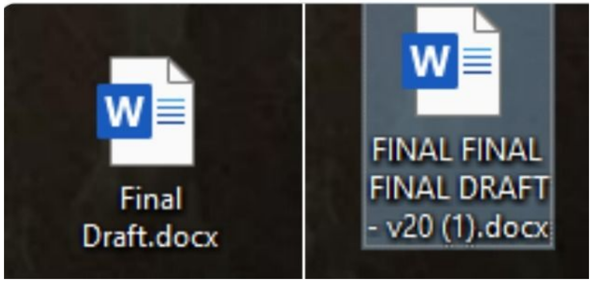
Additional up-to-date COVID data for this province can be found here:
https://experience.arcgis.com/experience/204d6ed723244dfbb763ca3f913c5cad

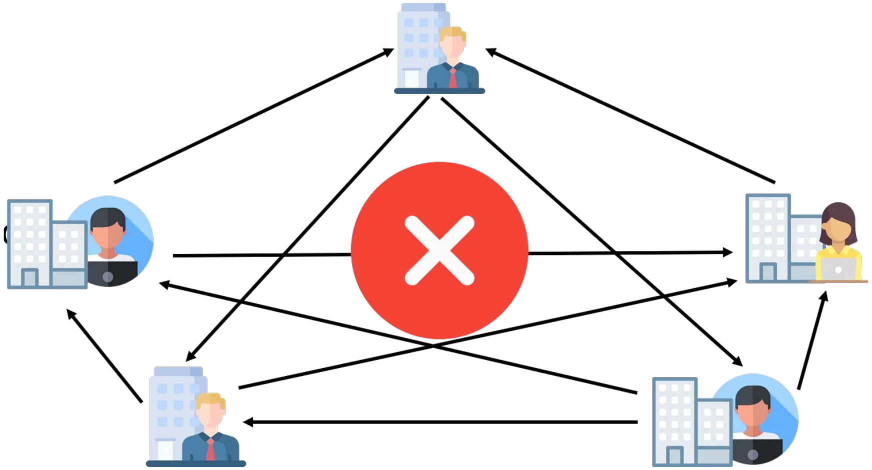
```
plot.variants(region='Nova Scotia')
plot.variants(region='Nova Scotia', scaled=T)
```

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

Use standard version control systems

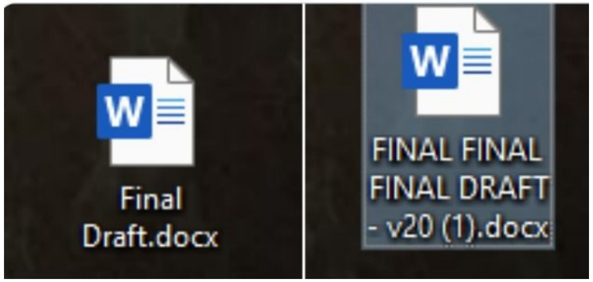
- Ever had a nightmare of versioning even when just you?
- Add more people and the chaos grows exponentially!





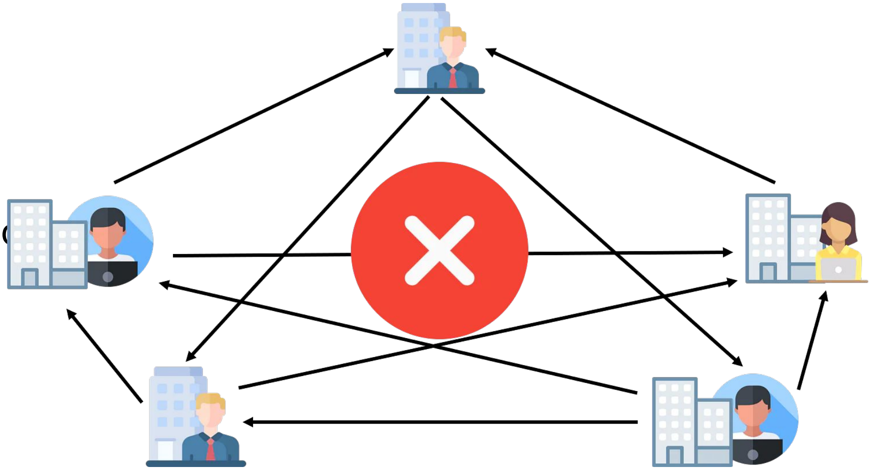
Use standard version control systems

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

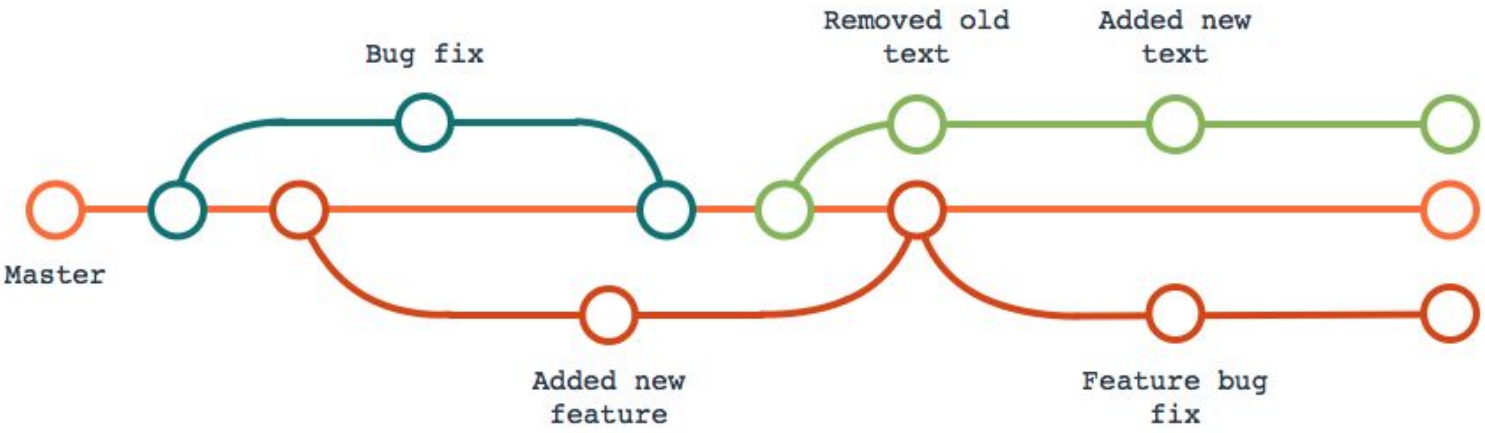


Version control let’s you:

- Revert mistakes
- Acts as a comprehensive backup
- Let’s you maintain multiple versions of your analysis
- Let’s you compare different versions of your co
- 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

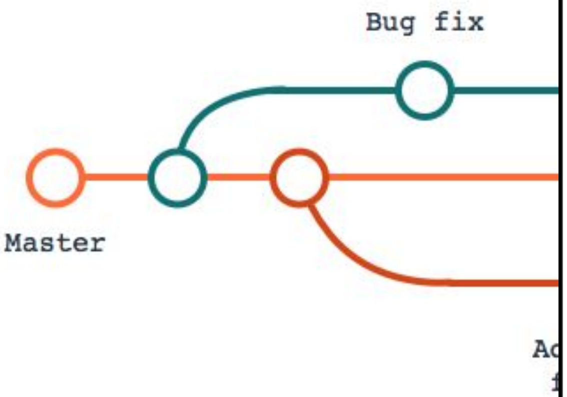


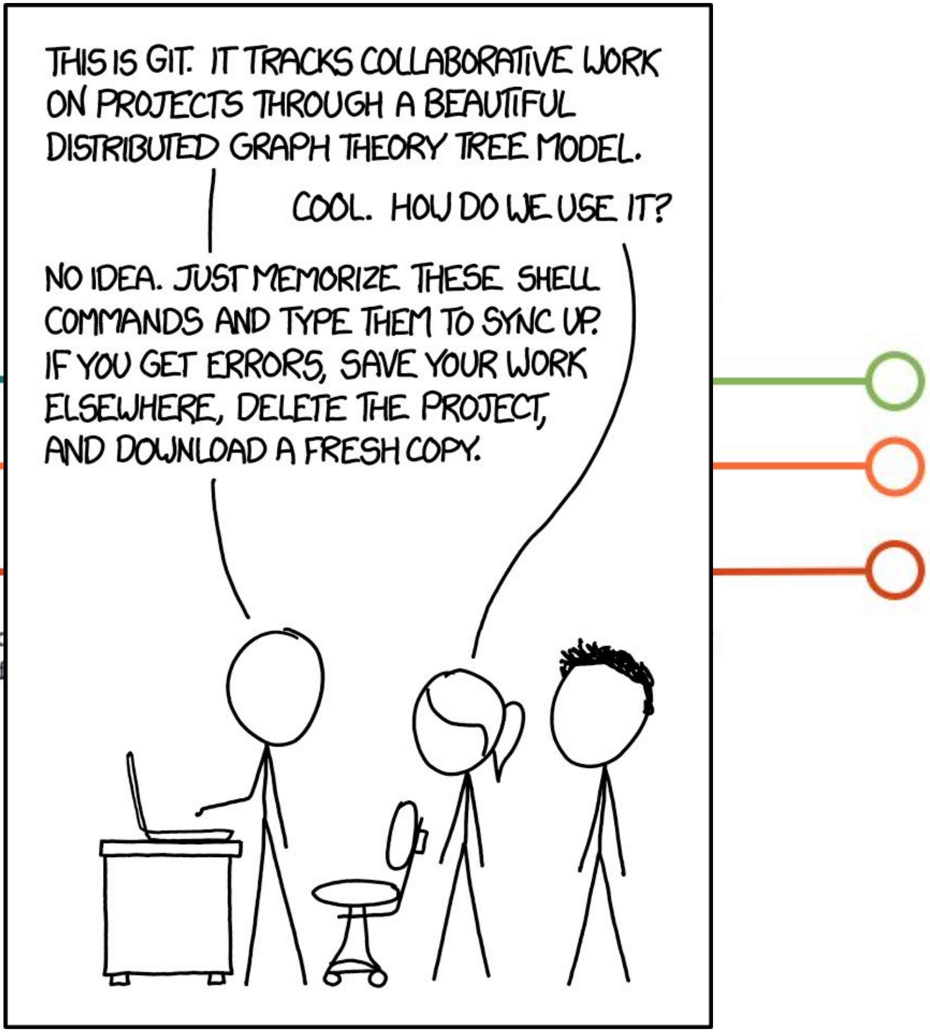
Git Version Control



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

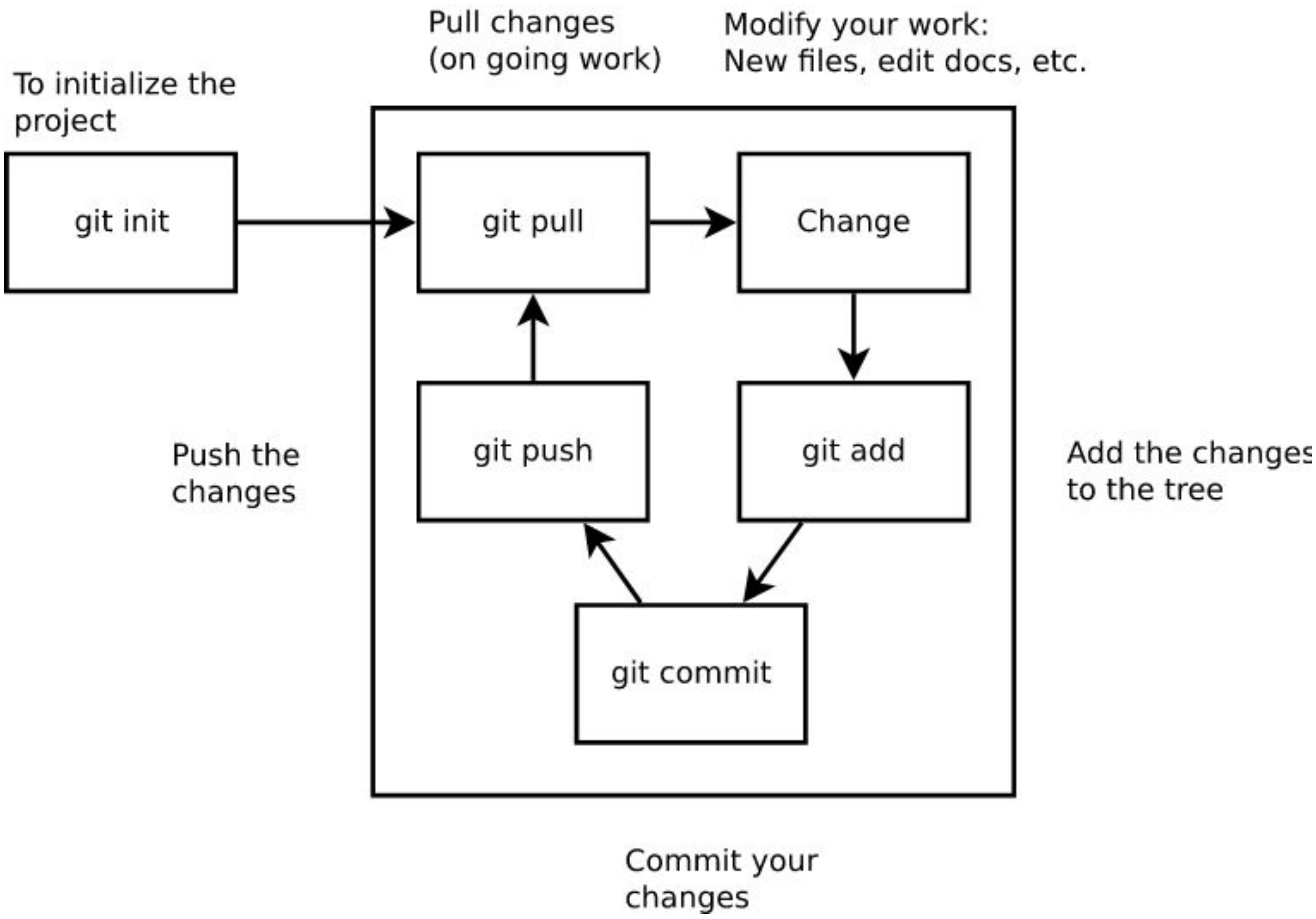
Git Version Control




THIS IS GIT. IT TRACKS COLLABORATIVE WORK ON PROJECTS THROUGH A BEAUTIFUL DISTRIBUTED GRAPH THEORY TREE MODEL.
COOL. HOW DO WE USE IT?
NO IDEA. JUST MEMORIZE THESE SHELL COMMANDS AND TYPE THEM TO SYNC UP. IF YOU GET ERRORS, SAVE YOUR WORK ELSEWHERE, DELETE THE PROJECT, AND DOWNLOAD A FRESH COPY.

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

Git Workflow



To initialize the project

git init

Pull changes (on going work)

git pull

Modify your work: New files, edit docs, etc.

Change

git add

Add the changes to the tree

git commit

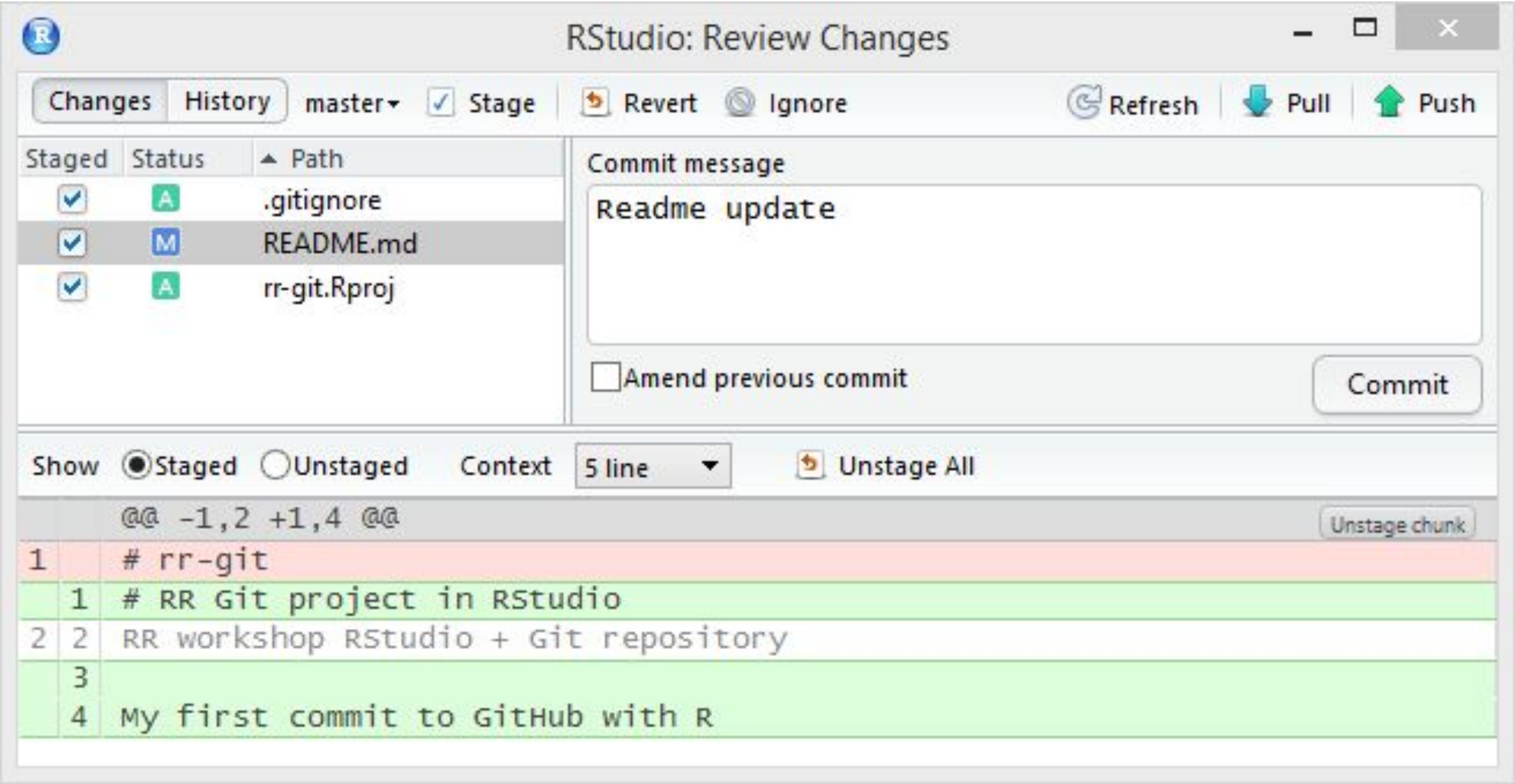
Commit your changes

git push

Push the changes

https://uscbiostats.github.io/PM566/slides/02-version-control/slides.html#8

Git is integrated into Rstudio!



Staged	Status	Path
☑	A	.gitignore
☑	M	README.md
☑	A	rr-git.Rproj

Commit message

Readme update

□ Amend previous commit

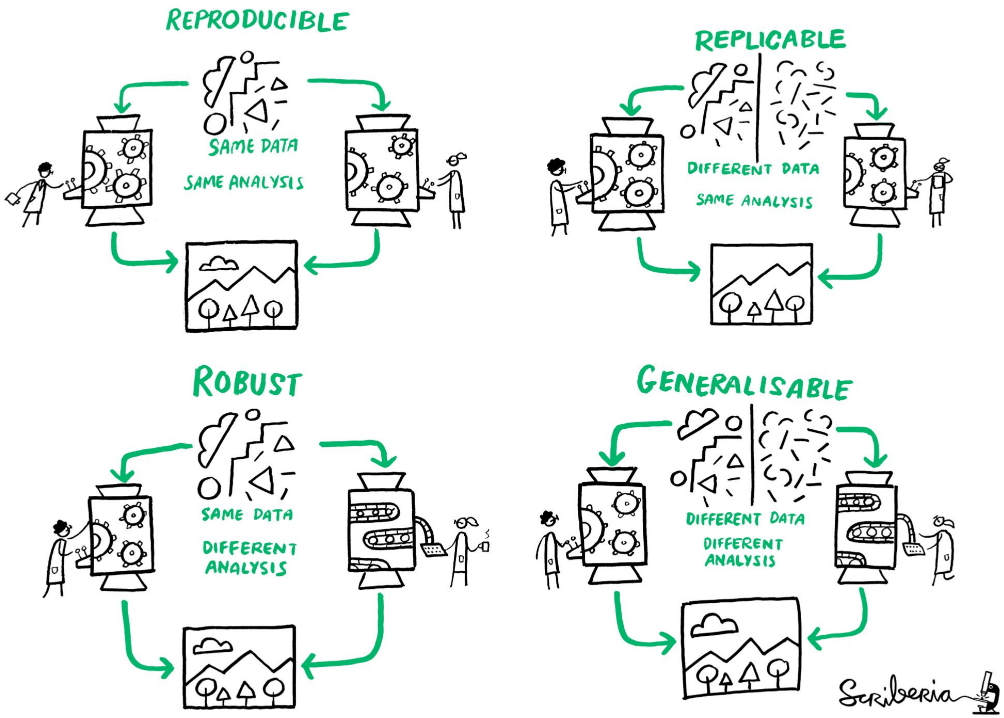
Show ● Staged ○ Unstaged Context 5 line

```
     @@ -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
```

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

- **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