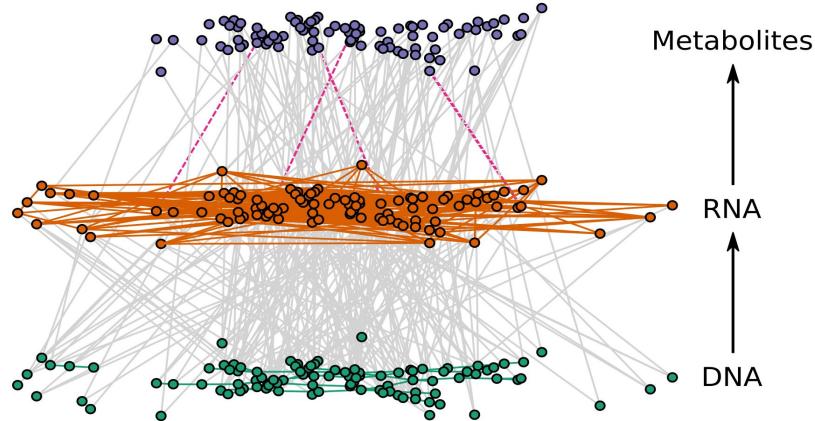


# Lecture 0: Introduction to Applied Research in Health Data Science

CSCI6XXX/CHE6XXX/CSCI4148  
(CSCI6093)

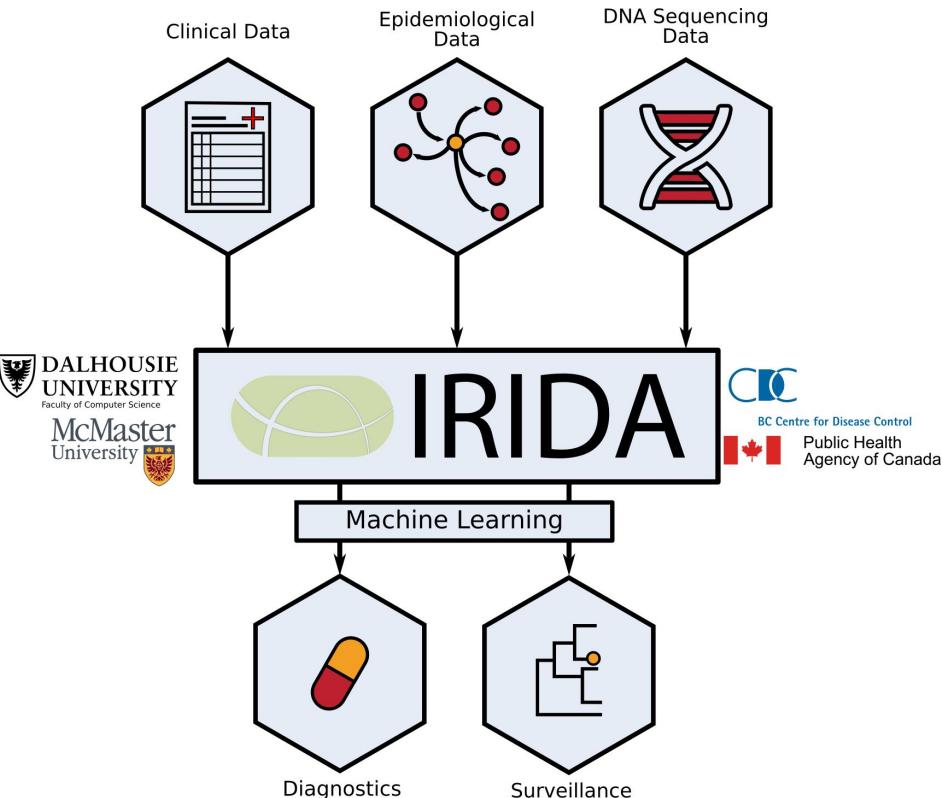
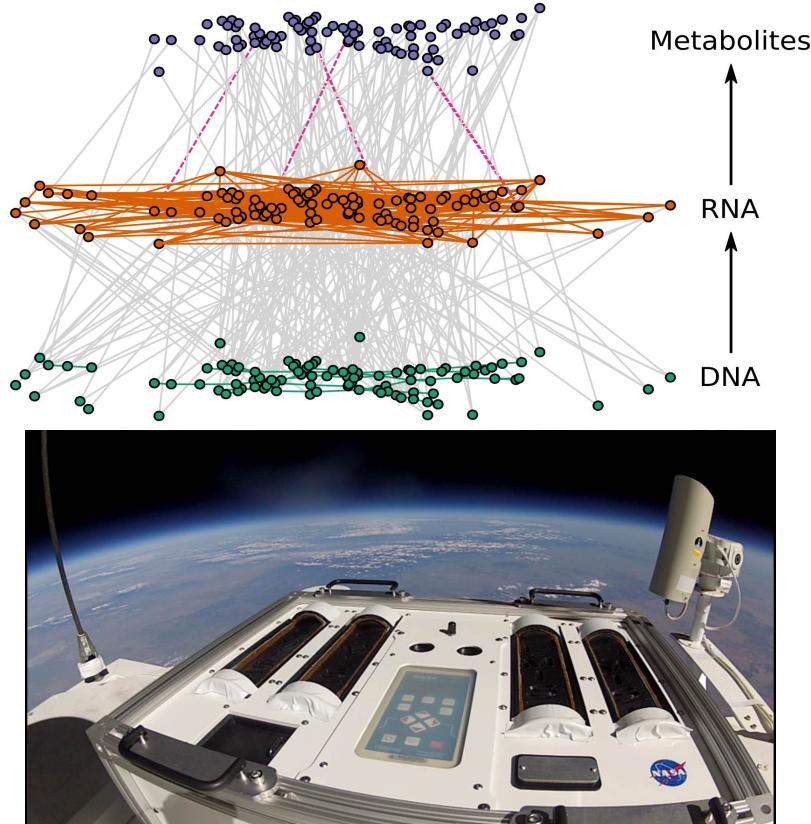
Finlay Maguire ([finlay.maguire@dal.ca](mailto:finlay.maguire@dal.ca))

# Why am I teaching this course?



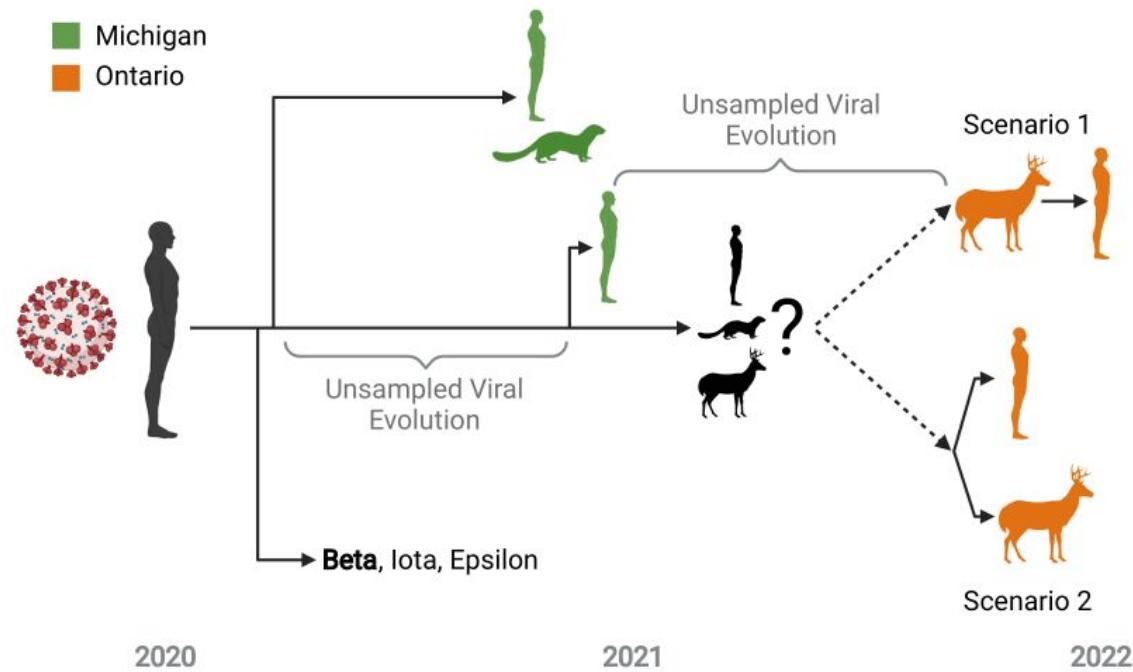
- **PhD (Bioinformatics):** using large noisy datasets to understand how microbial systems and mechanisms evolve.

# Why am I teaching this course?



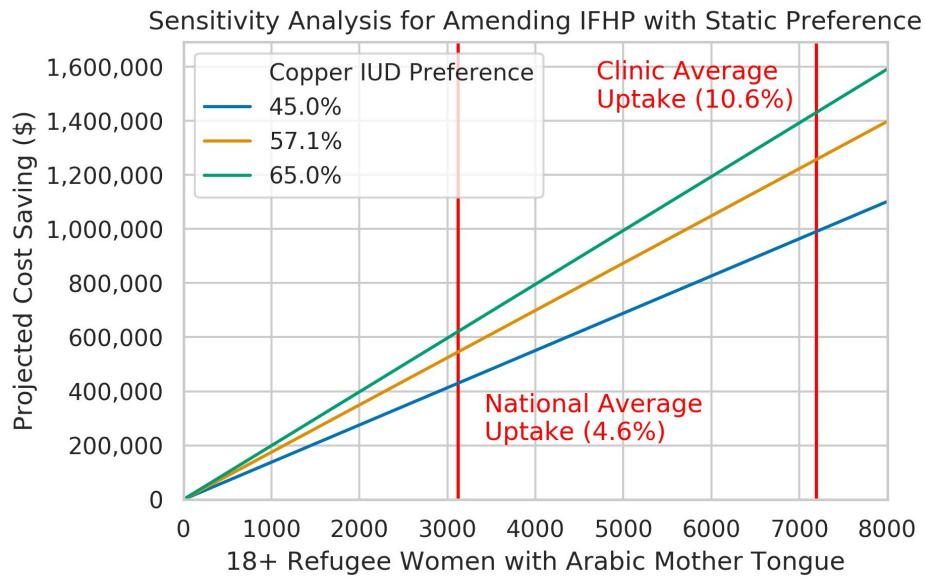
- **PhD (Bioinformatics)**: using large noisy datasets to understand how microbial systems and mechanisms evolve.
- **Postdoc (Genomic Epidemiology)**: using large noisy datasets to better diagnose, track and predict infectious diseases.

# Why am I teaching this course?

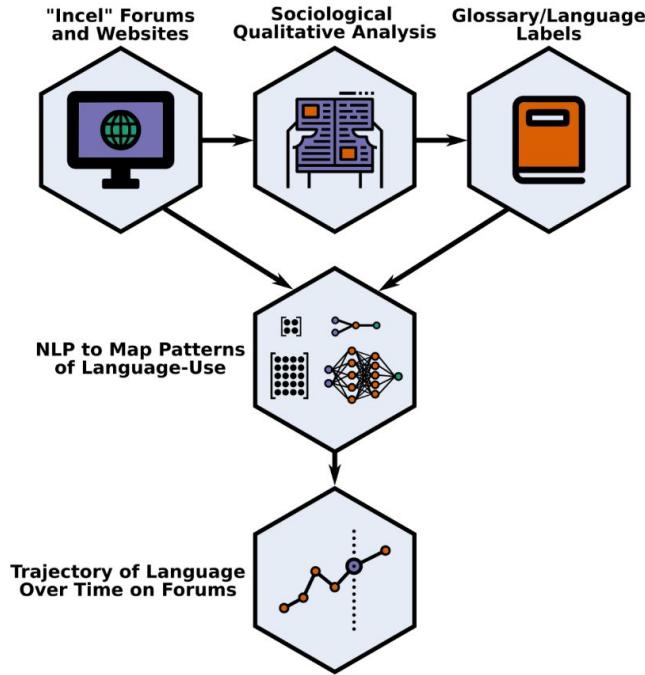


- **Research group:** using large noisy datasets:
  - Genomic epidemiology of infectious disease: **SARS-CoV-2, AMR**

# Why am I teaching this course?



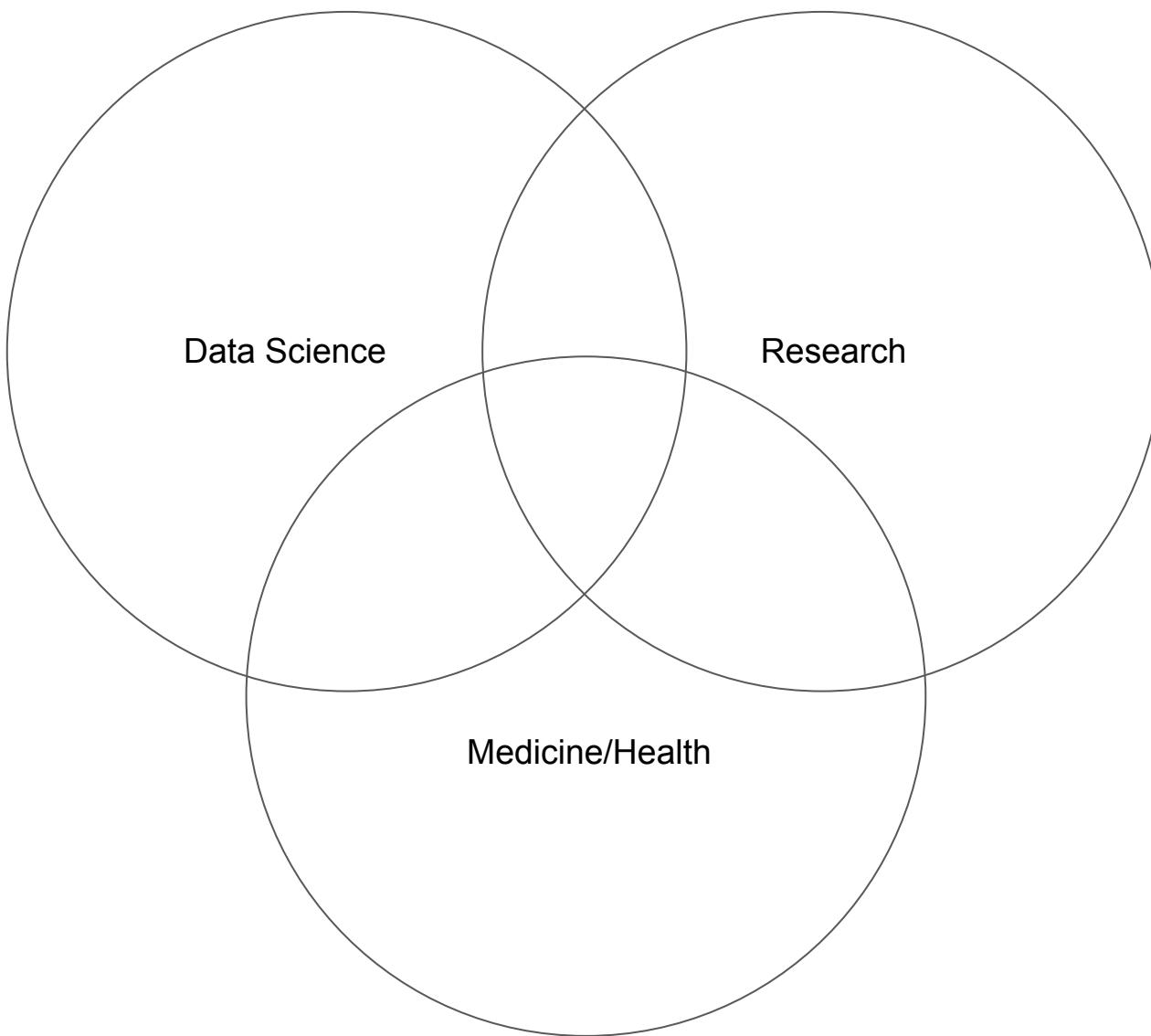
Modelling “Incel” Online Radicalisation via NLP



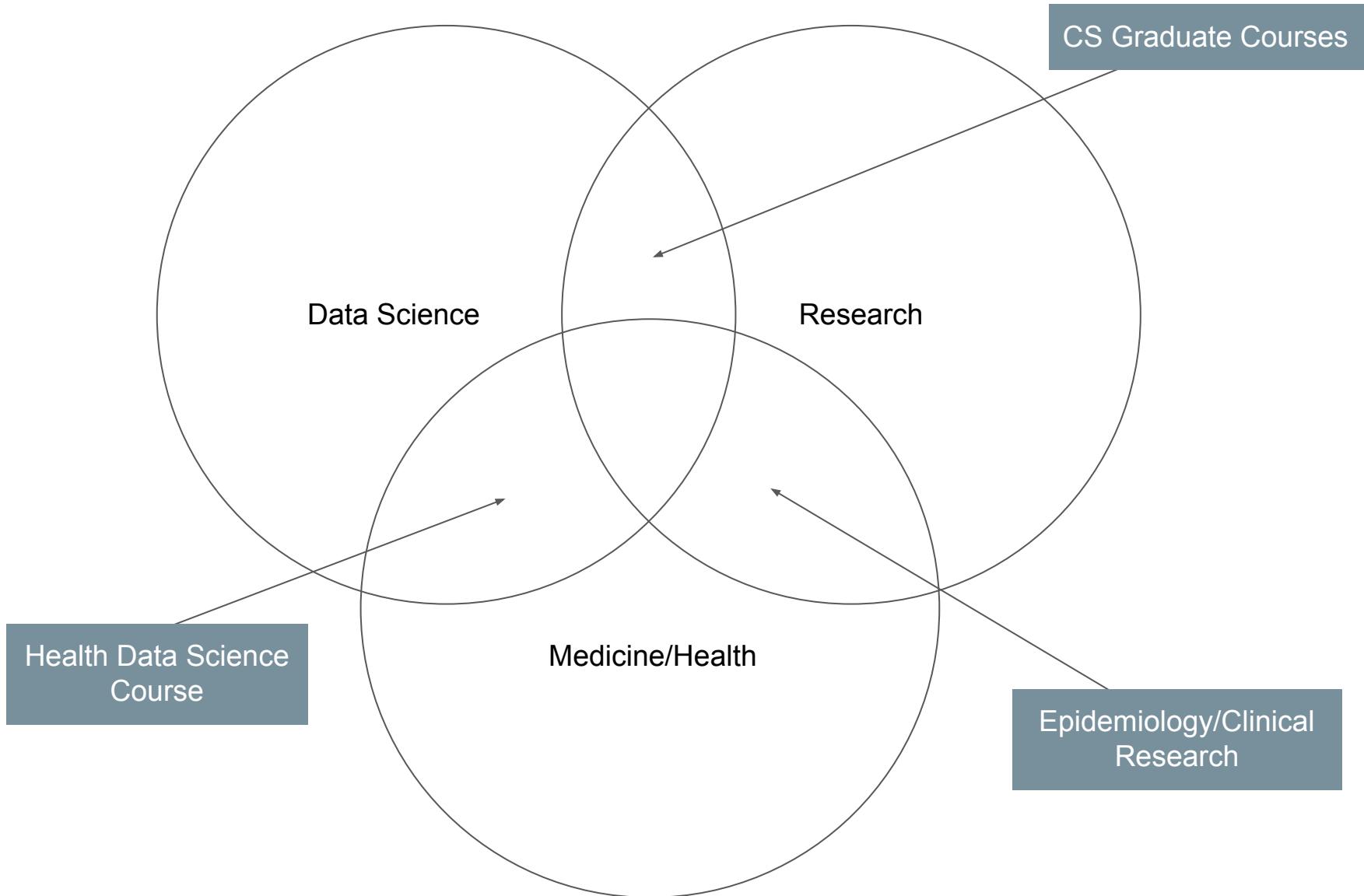
- **Research group:** using large noisy datasets:
  - Genomic epidemiology of infectious disease: **SARS-CoV-2, AMR**
  - Collaborations on socially/health focused problems: **refugee health, incel radicalisation, health inequality**

# Overview of course

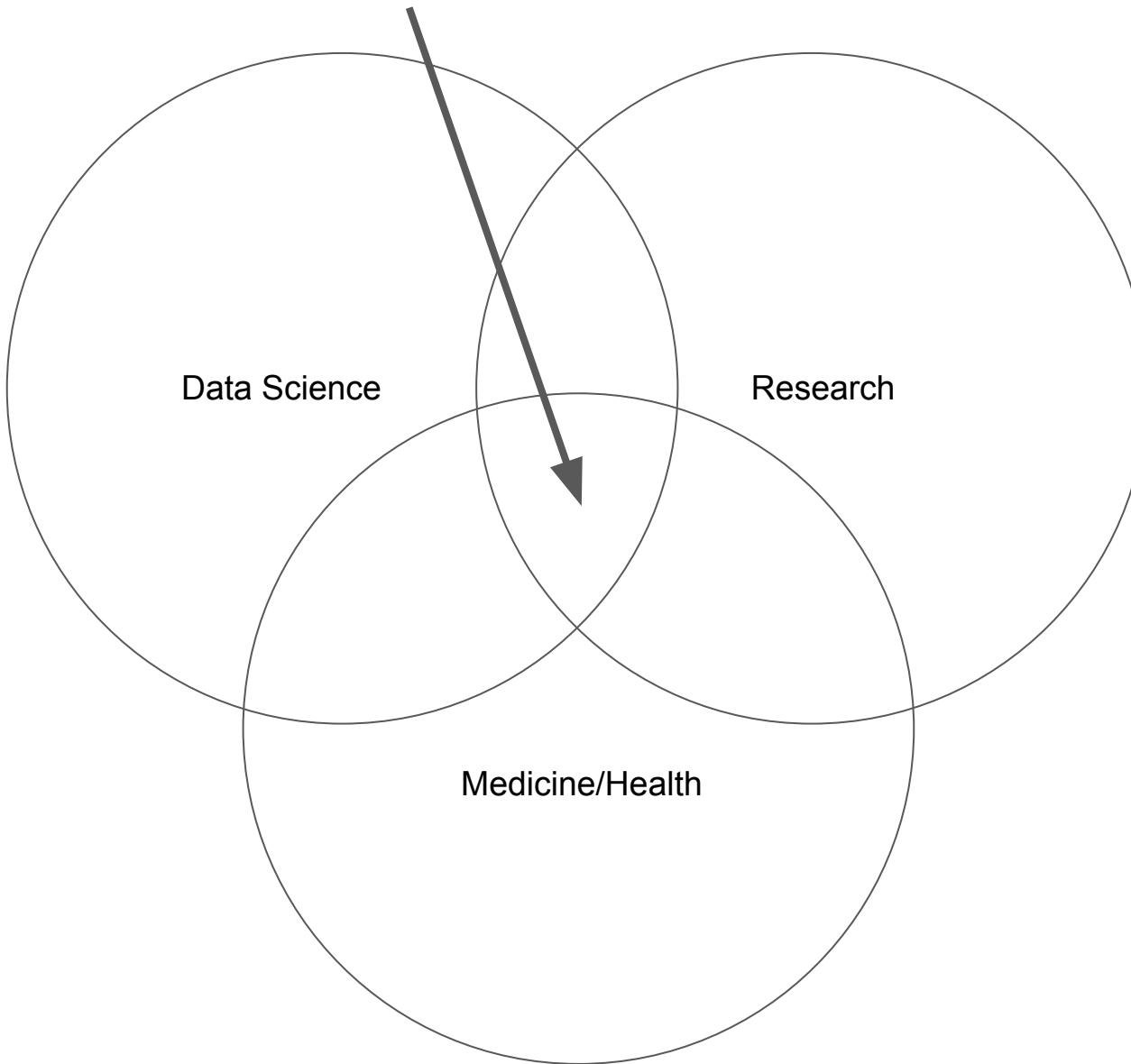
# Applied Research in Health Data Science



# 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 = "Density")
  if(orientation == "vertical")
    dx2 <- (dx - min(dx)) / max(dx)
    x[1.]
    dy2 <- (dx - min(dx)) / max(dy)
    y[1.]
    seqbelow <- rep(y[1.], length(dx))
    if(Fill == T)
      confshade(dx2, seqbelow, dy2
```



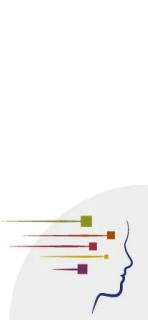
# Course Structure

## 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 = "Density")
if(orientation == "vertical")
  dx2 <- (dx - min(dx)) / max(dx)
  x[1.]
  dy2 <- (dx - min(dx)) / max(dy)
  y[1.]
  seqbelow <- rep(y[1.], length(dx))
  if(Fill == T)
    confshade(dx2, seqbelow, dy2)
```



## 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%)

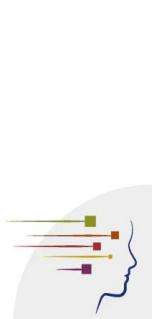
# Course Structure

## 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 = "Density")
if(orientation == "vertical")
  dx2 <- (dx - min(dx)) / max(dx)
  x[1.]
  dy2 <- (dx - min(dx)) / max(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:

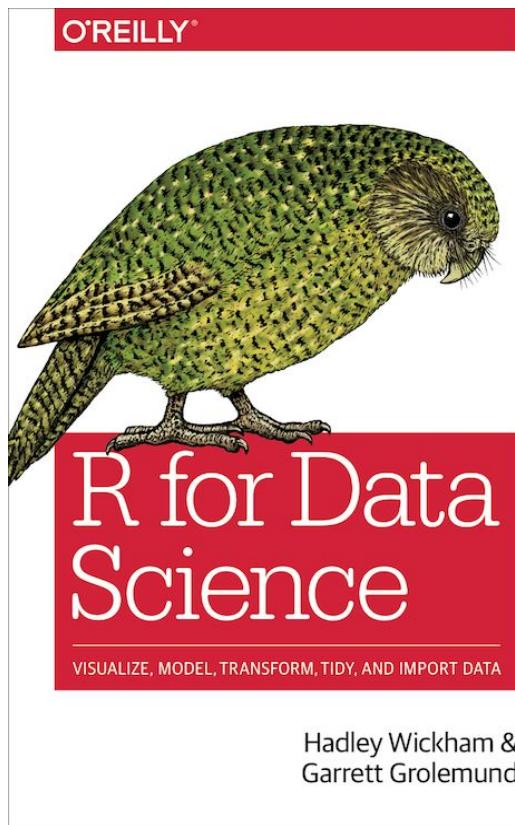
- **Class** (Wednesday/Friday)

### Assessment:

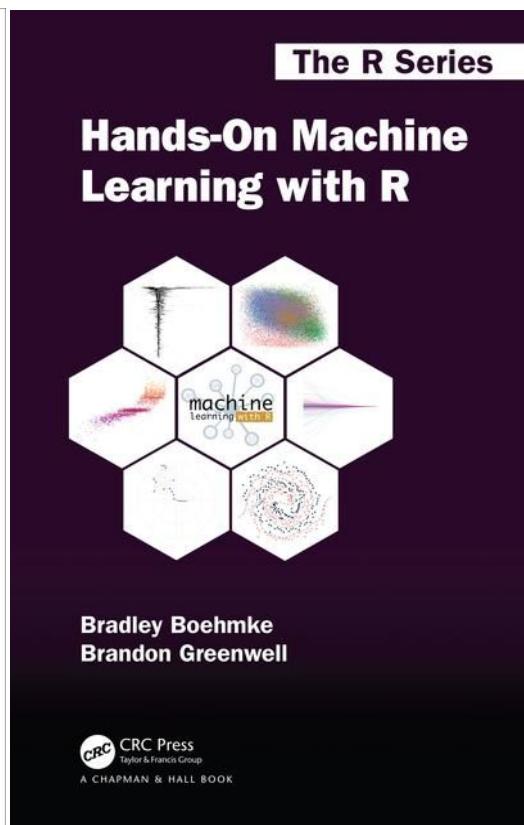
Presentation last full week of class (20%)

Submitted final day of class (20%)

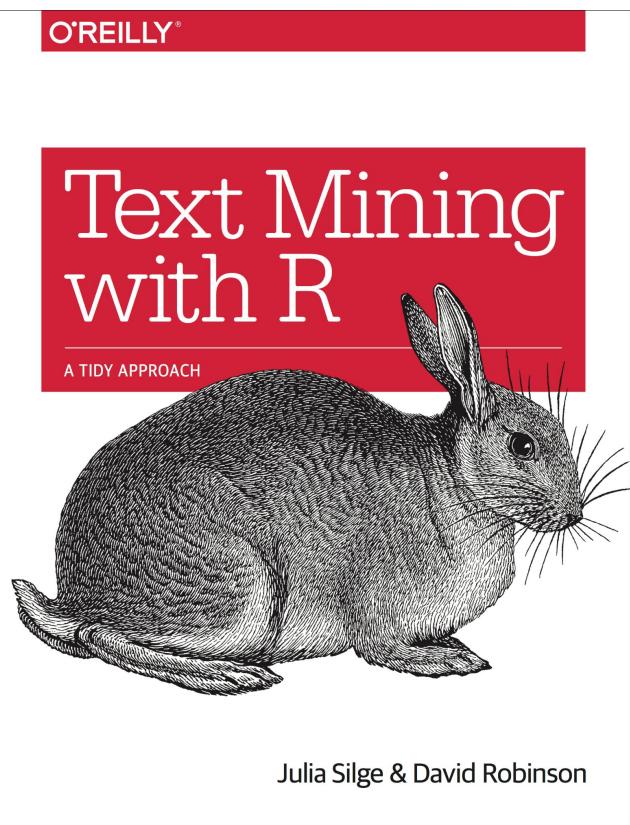
# Course Materials



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



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



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

# Course Website



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/](https://maguire-lab.github.io/health_data_science_research/)**

# Course Website



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/](https://maguire-lab.github.io/health_data_science_research/)**

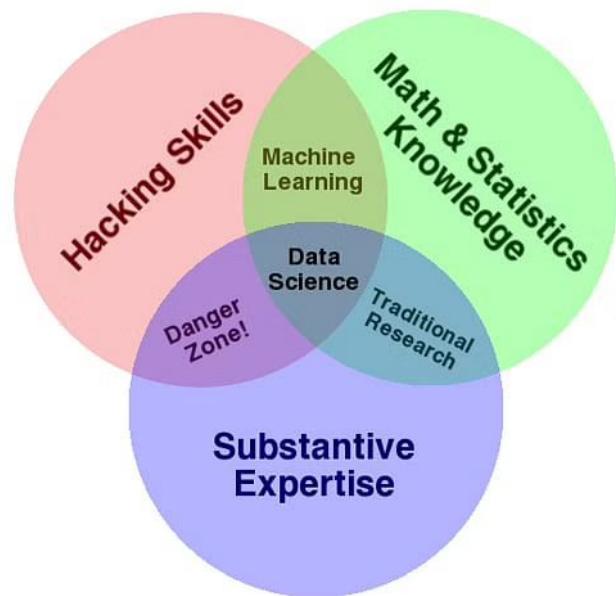
The screenshot shows the Brightspace course site for "CSCI6903 - Spec Grad Top Comp Sci (Sec 1) - 2022 Summer". The top navigation bar includes links for "Course Home", "Content", "Discussions", "Assessments", "My Tools", "Help", and "Course Admin". On the far left, there's a sidebar with "Announcements" and "Course Website" sections. The main content area features a large image of a vineyard. Below the image, the course title is displayed again. A "Updates" section is present, stating "There are no current updates for CSCI6903 - Spec Grad Top Comp Sci (Sec 1) - 2022 Summer". A footer at the bottom of the page says "I'm just putting together the main course website which will be linked here".

## Grades/Submissions:

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

What is ~~health~~ data science?

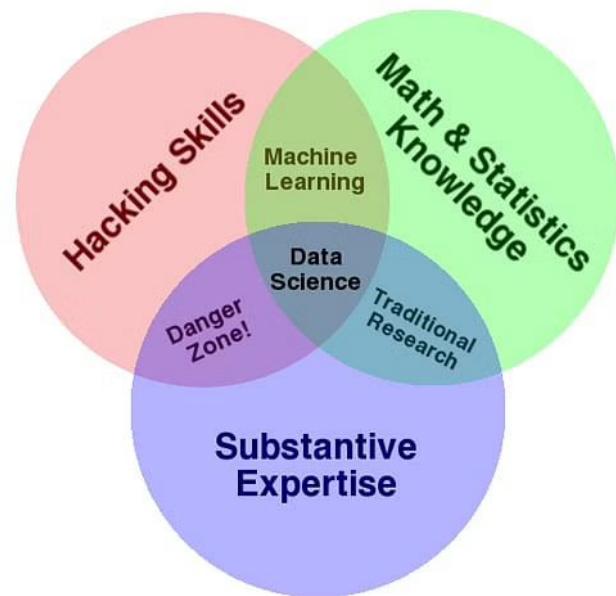
# Data Science: *Using Data to Better Understand Things in the Real World*



<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:

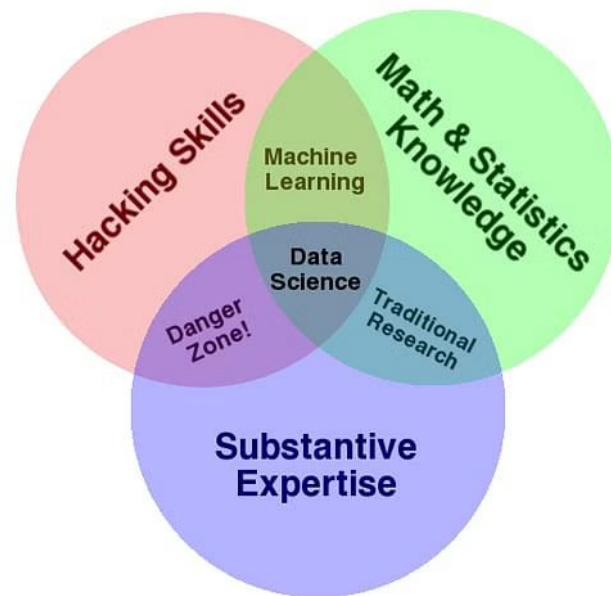


<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

# 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



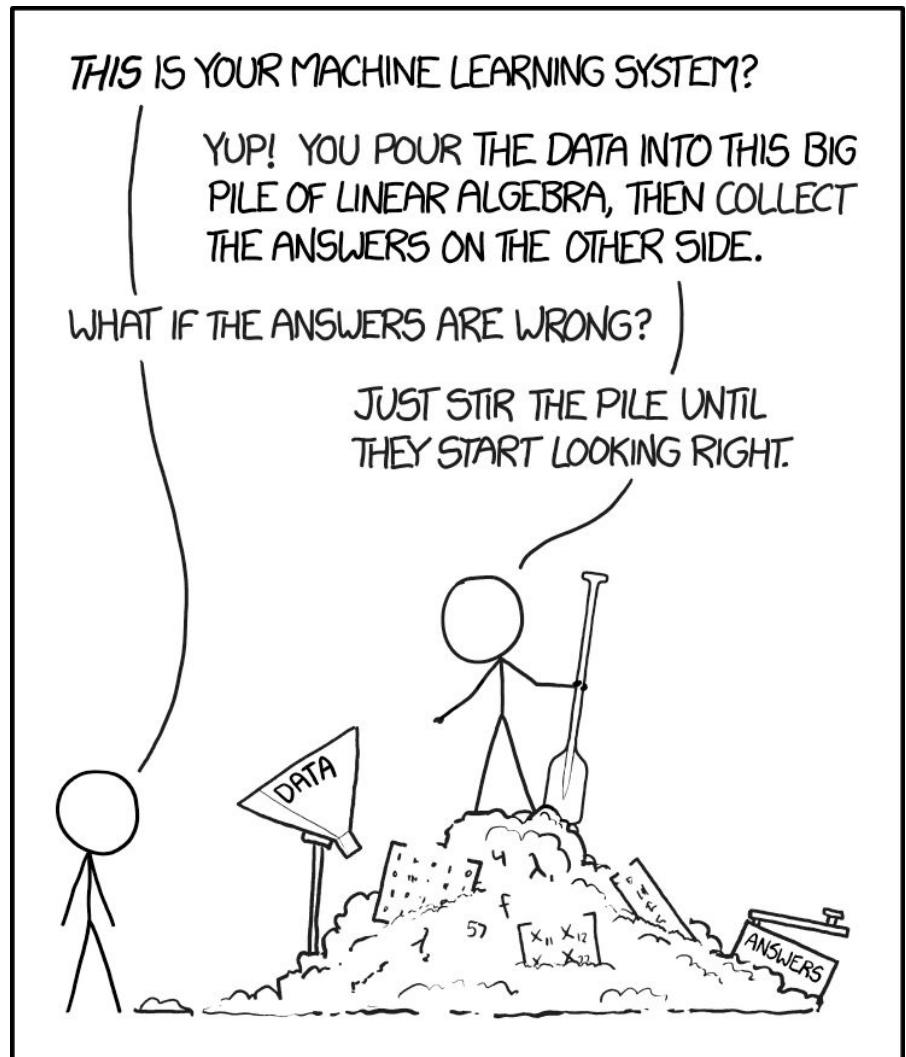
<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

So, it is just statistics?

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

## Pitfalls (can be):

- Less rigorous/principled
- Prone to reinventing the wheel



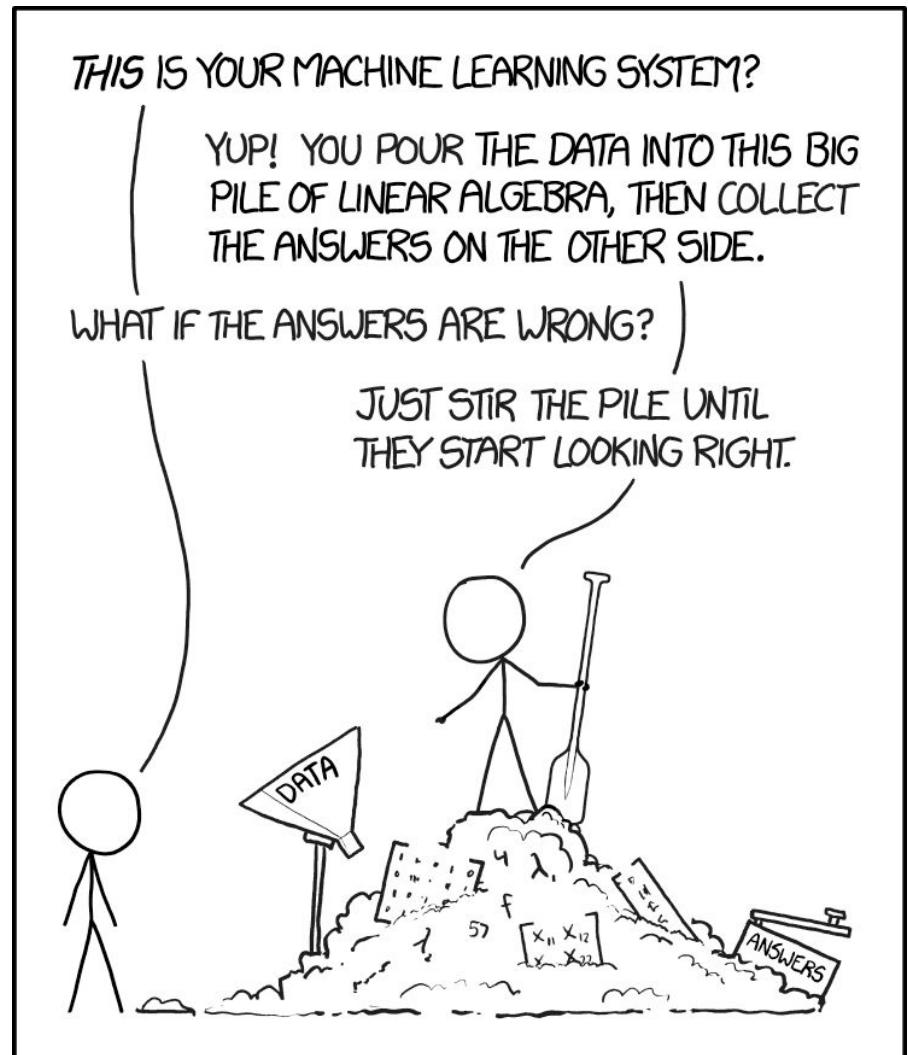
# Data Science (& Machine Learning): re-branded statistics

## Pitfalls (can be):

- Less rigorous/principled
- Prone to reinventing the wheel

## Benefits (can be):

- More flexible
- Less prescriptive/intimidating



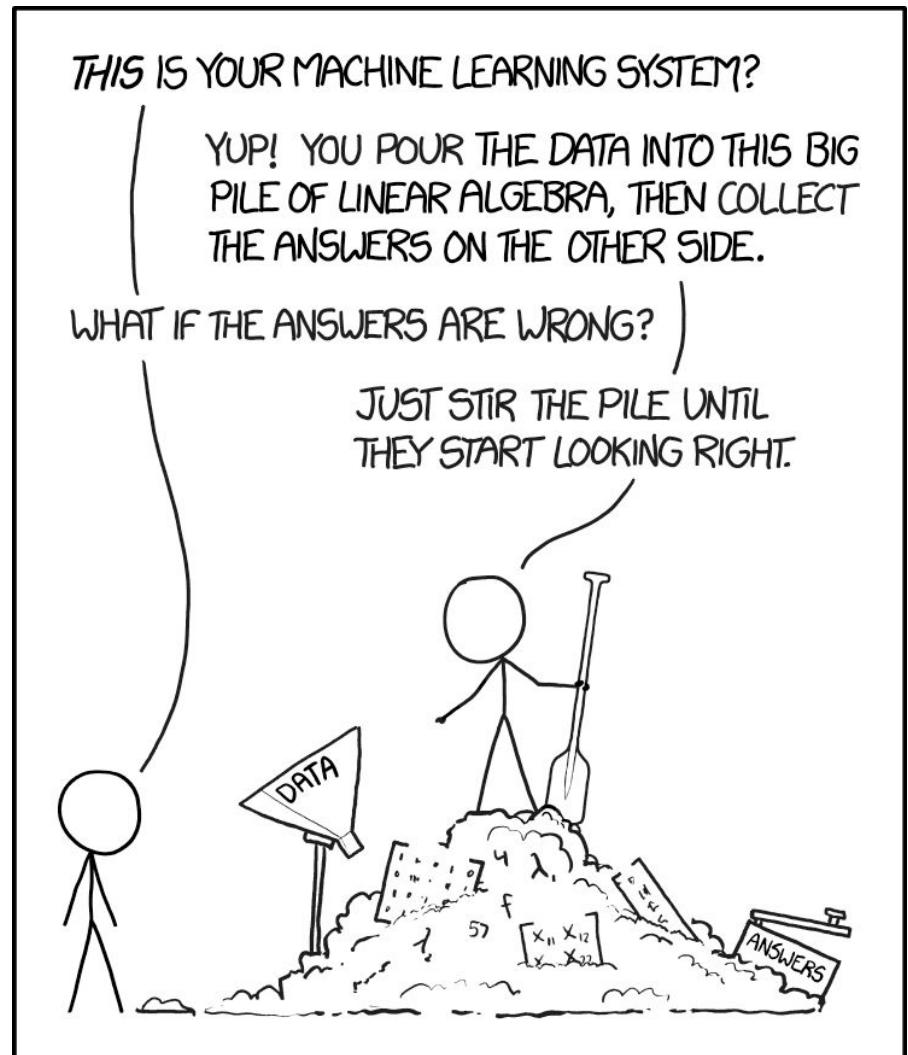
# Data Science (& Machine Learning): re-branded statistics

## Pitfalls (can be):

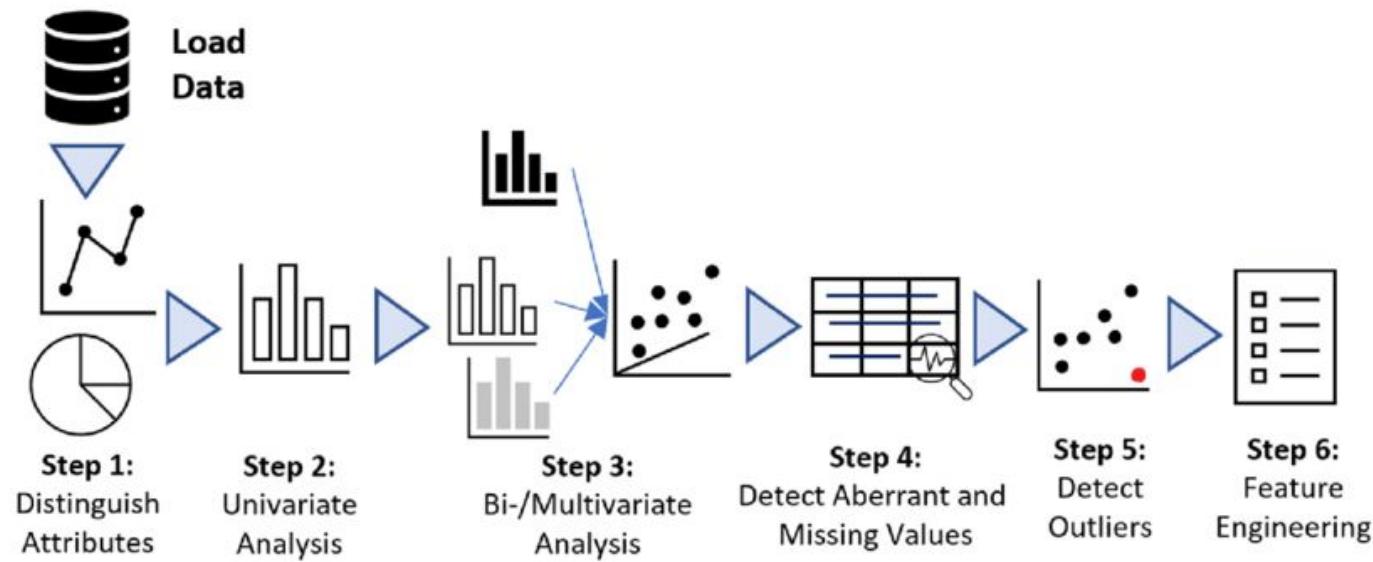
- Less rigorous/principled
- Prone to reinventing the wheel

## Benefits (can be):

- More flexible
- Less prescriptive/intimidating



# Data science centers exploratory data analysis



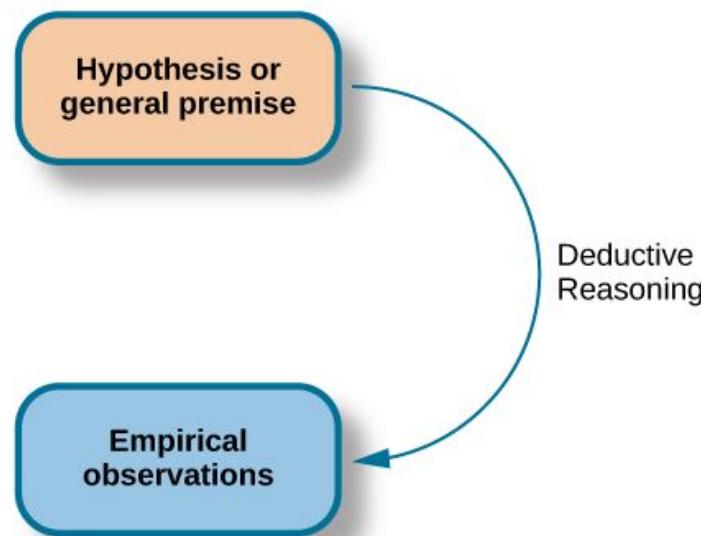
10.3390/su12124995

# 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



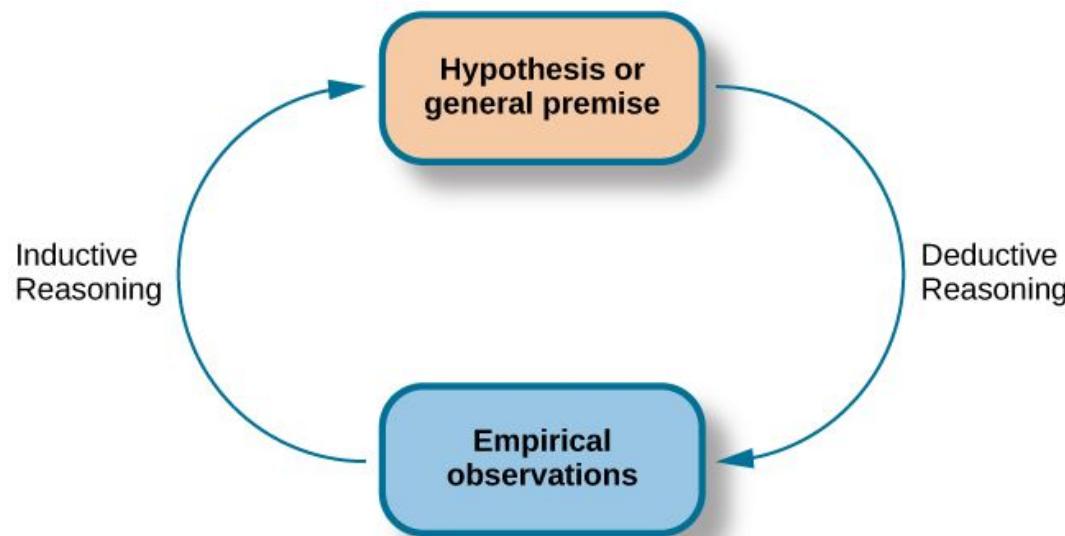
# Data science supports inductive approaches

## Deductive:

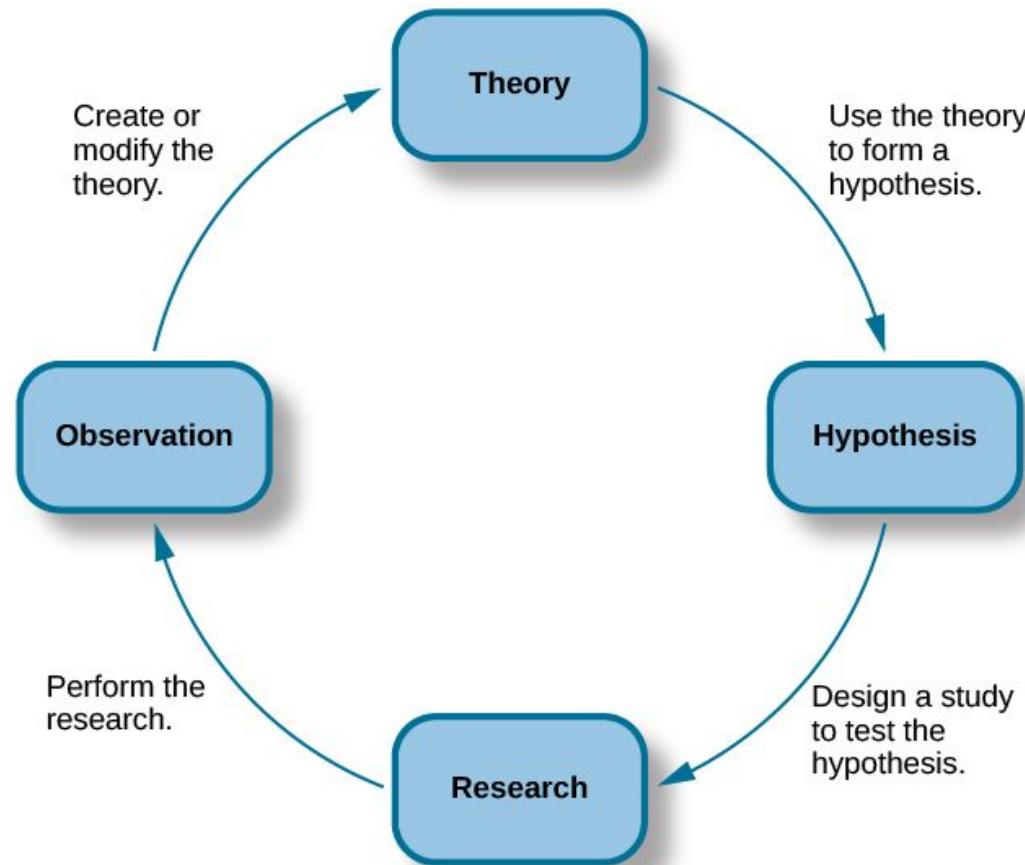
- “Condition X, causes Y”
- Collect data
- Perform frequentist statistical test
- Reject or confirm null hypothesis

## Inductive:

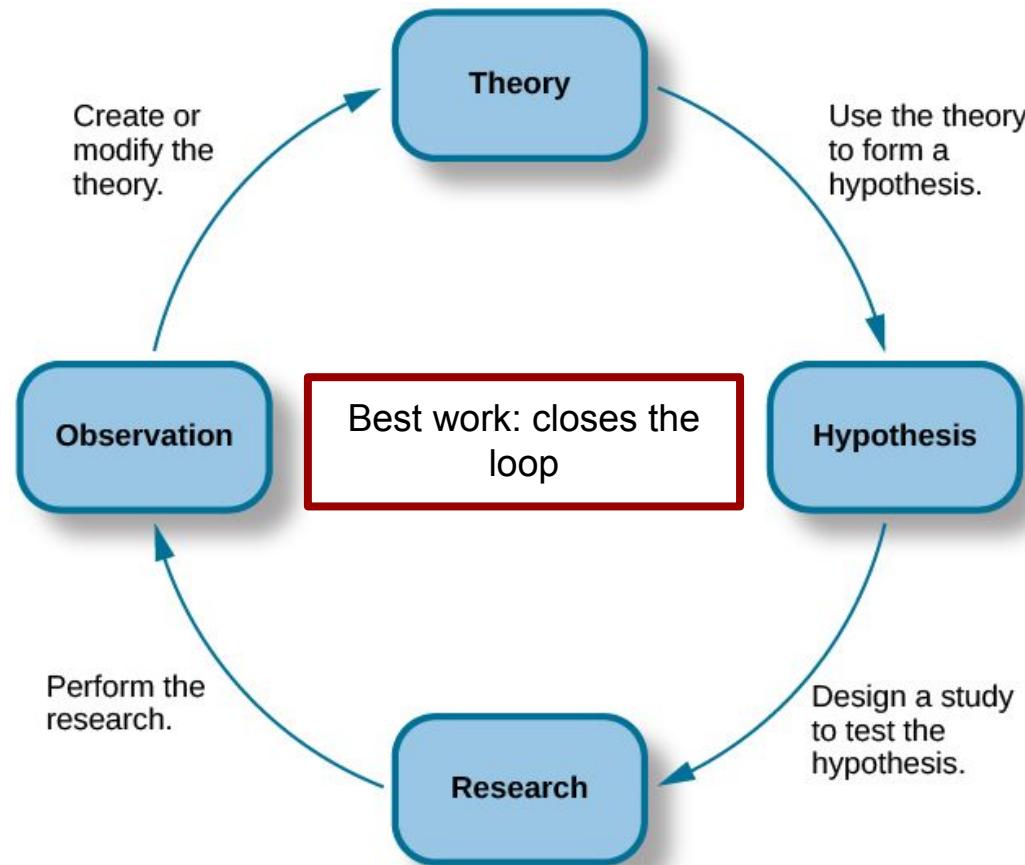
- Collect data
- Identify patterns in the data
- Observe X and Y seem connected somehow
- Quantify strength of association



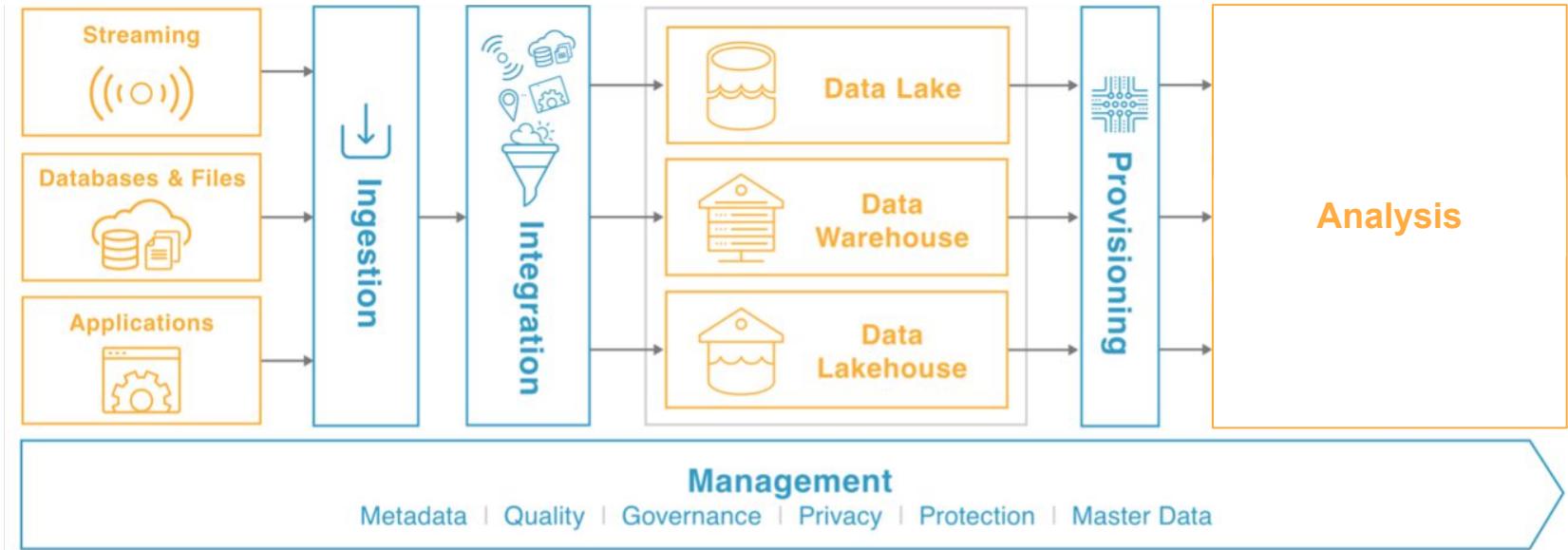
# Data science is more realistic



# Data science is more realistic

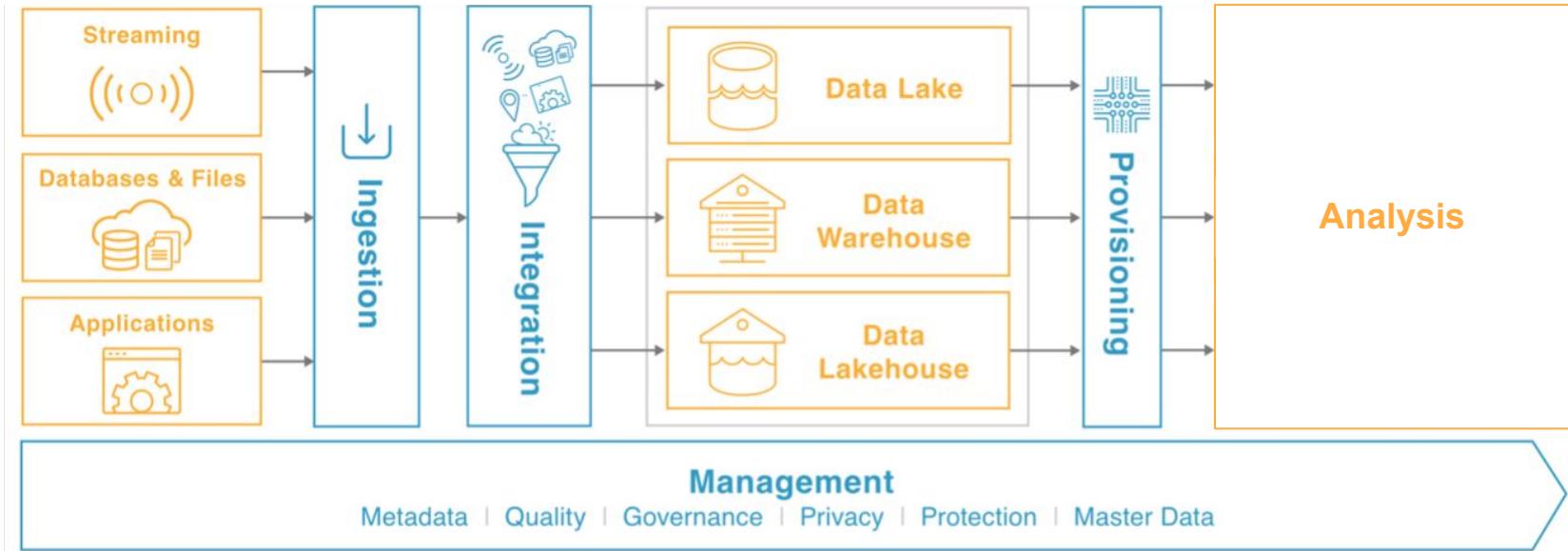


# Data science is integrated into a data ecosystem

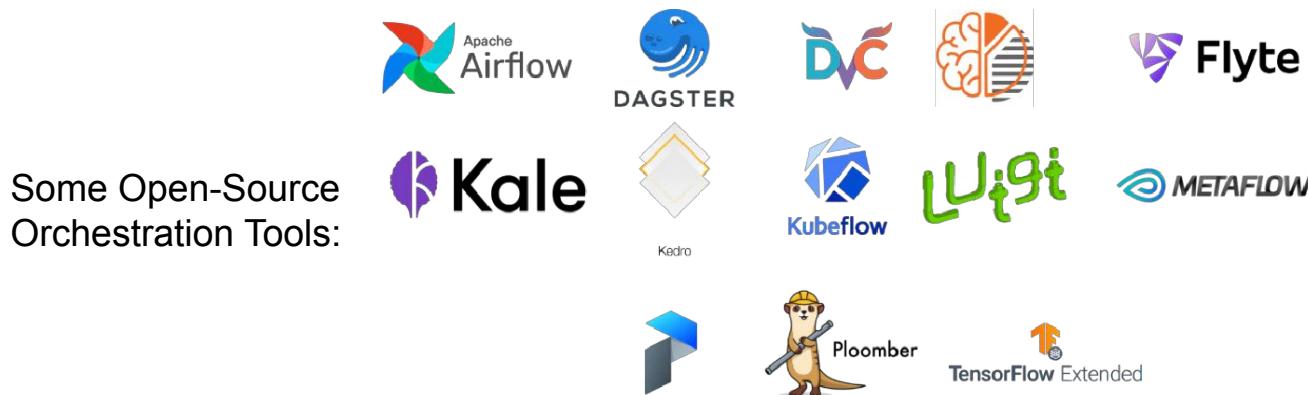


<https://www.2ndwatch.com/blog/what-is-a-data-pipeline-and-how-to-build-one/>

# Data science is integrated into a data ecosystem



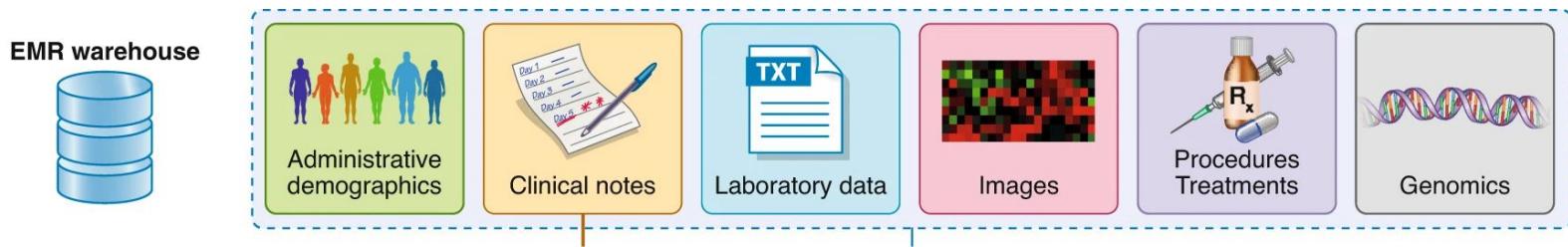
<https://www.2ndwatch.com/blog/what-is-a-data-pipeline-and-how-to-build-one/>



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

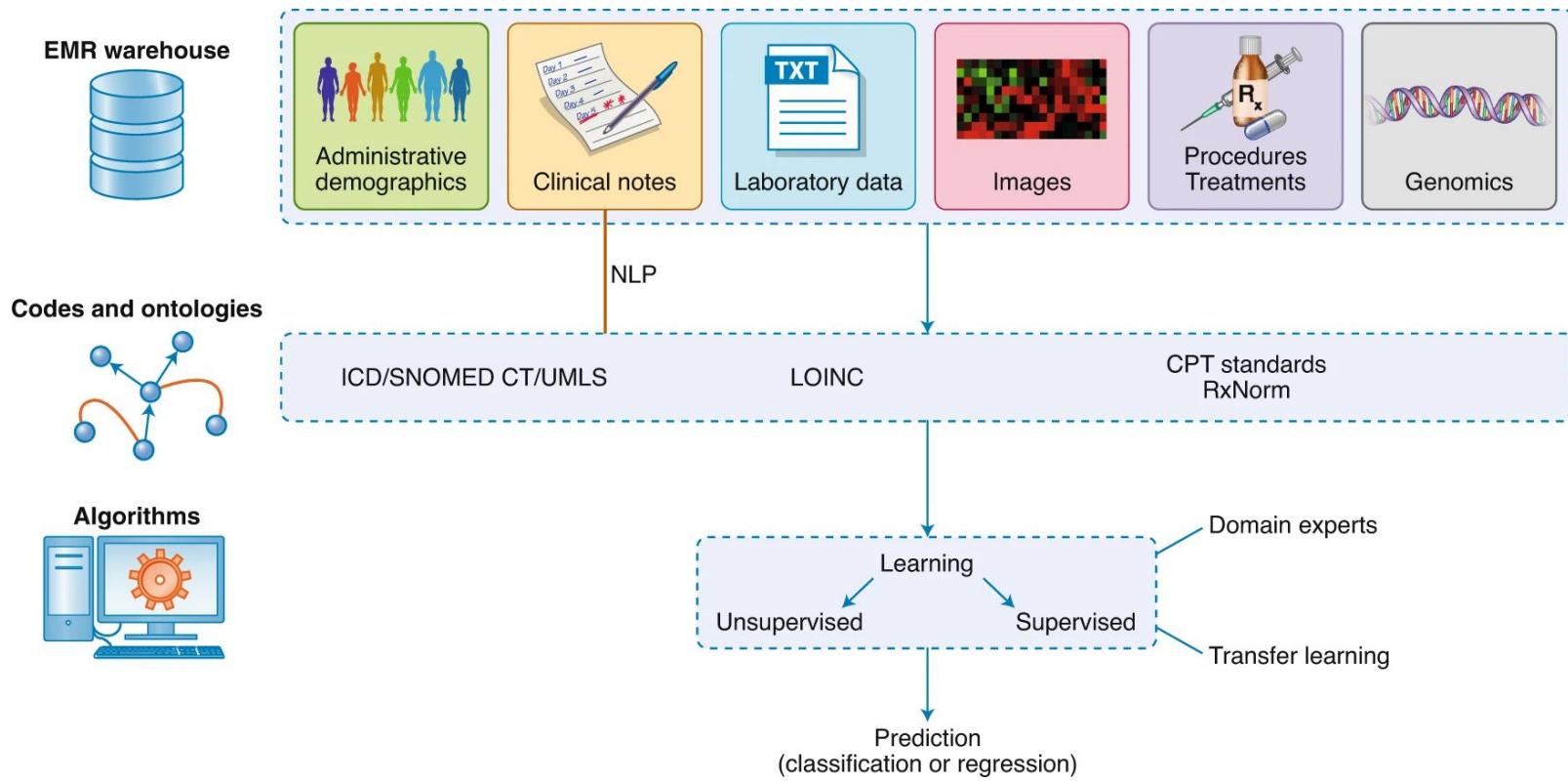
OK, what is **Health** Data Science?

# Data Science applied to Health Data



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

# Data Science applied to Health Data



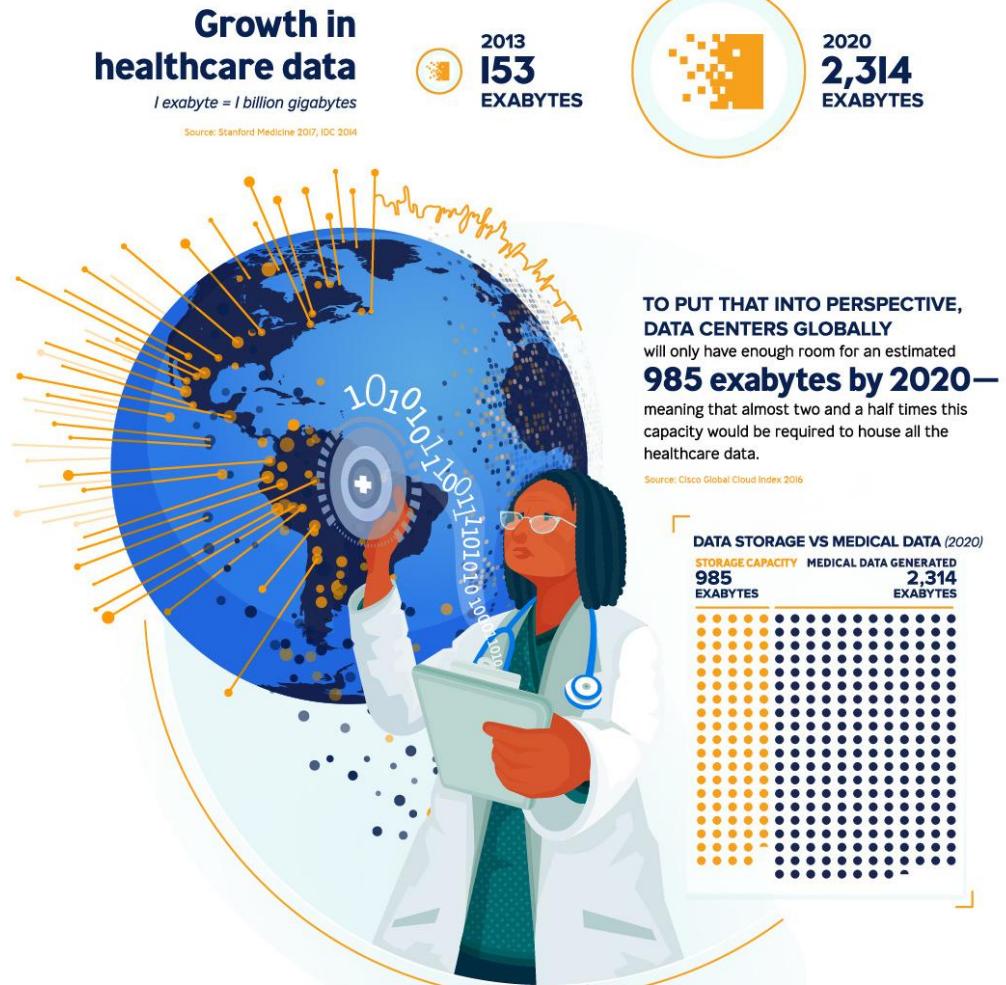
<https://www.nature.com/articles/s41588-020-0698-y/figures/2>

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

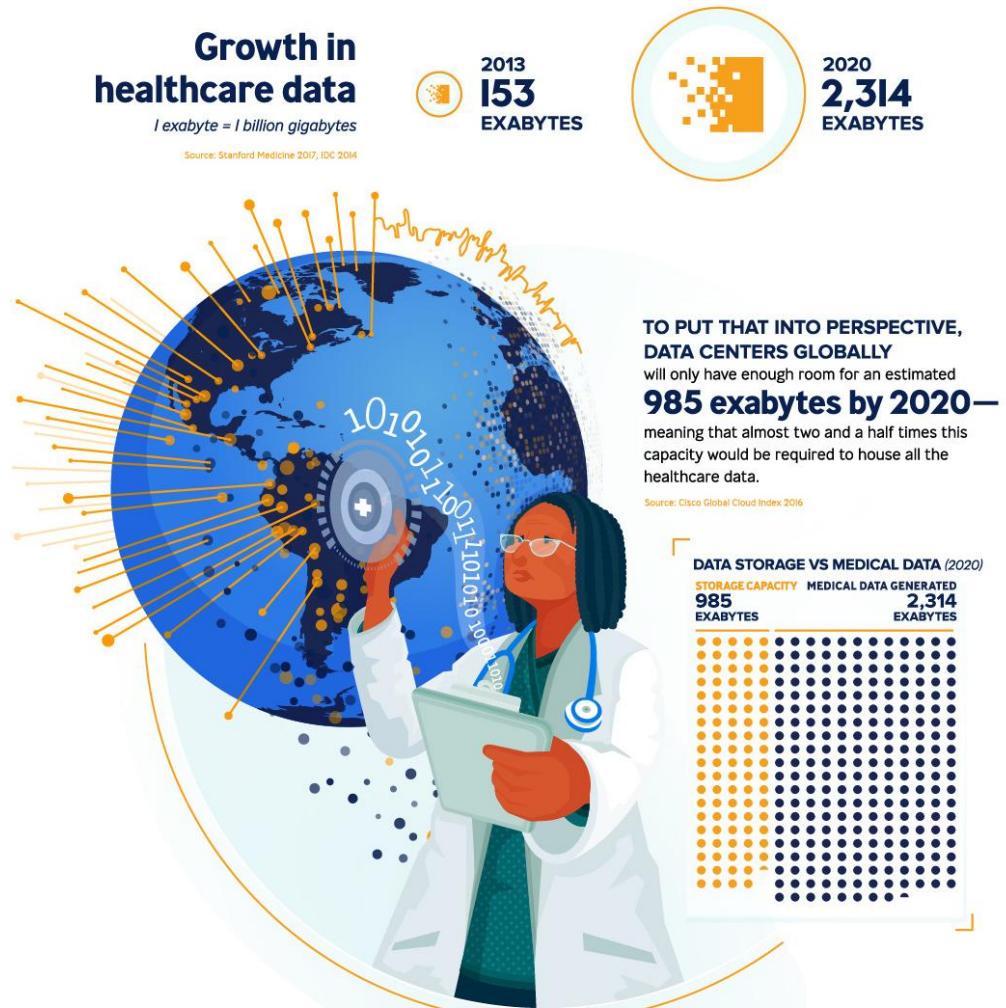


<https://www.visualcapitalist.com/big-data-healthcare/>

# 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**

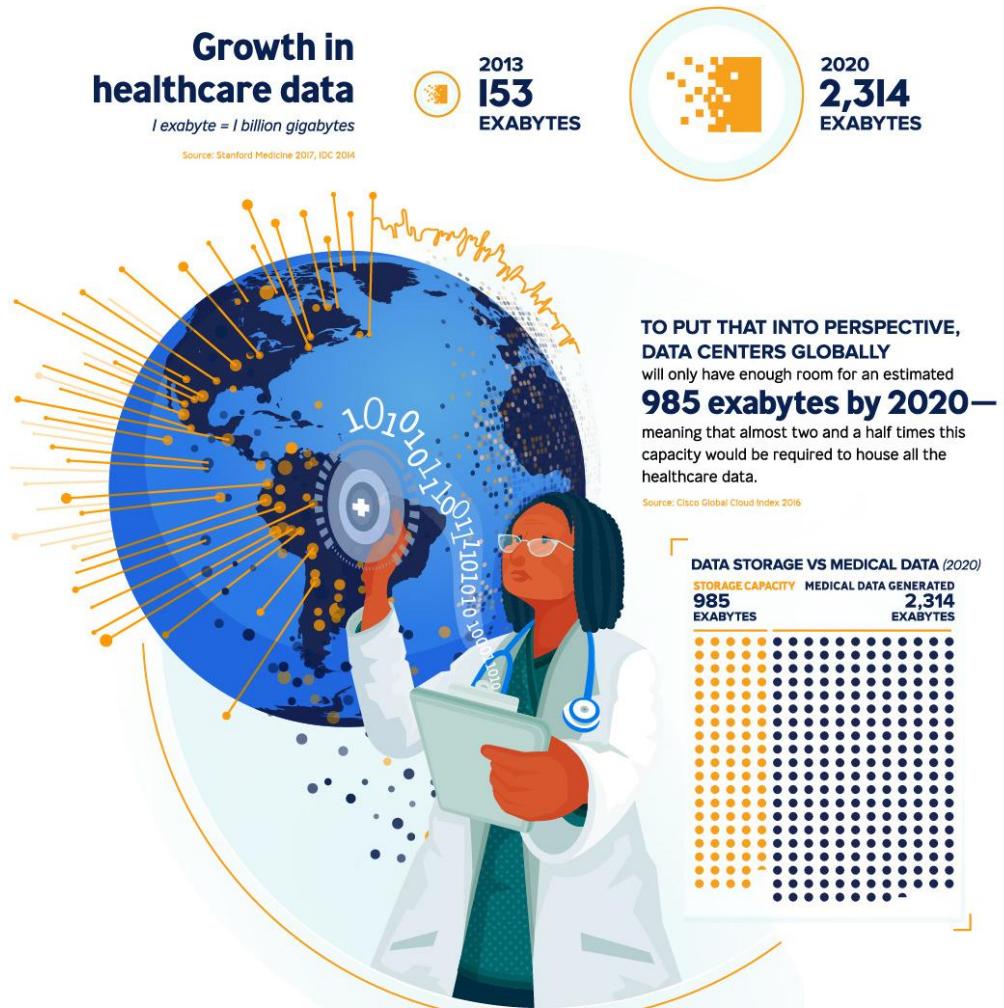


<https://www.visualcapitalist.com/big-data-healthcare/>

# 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

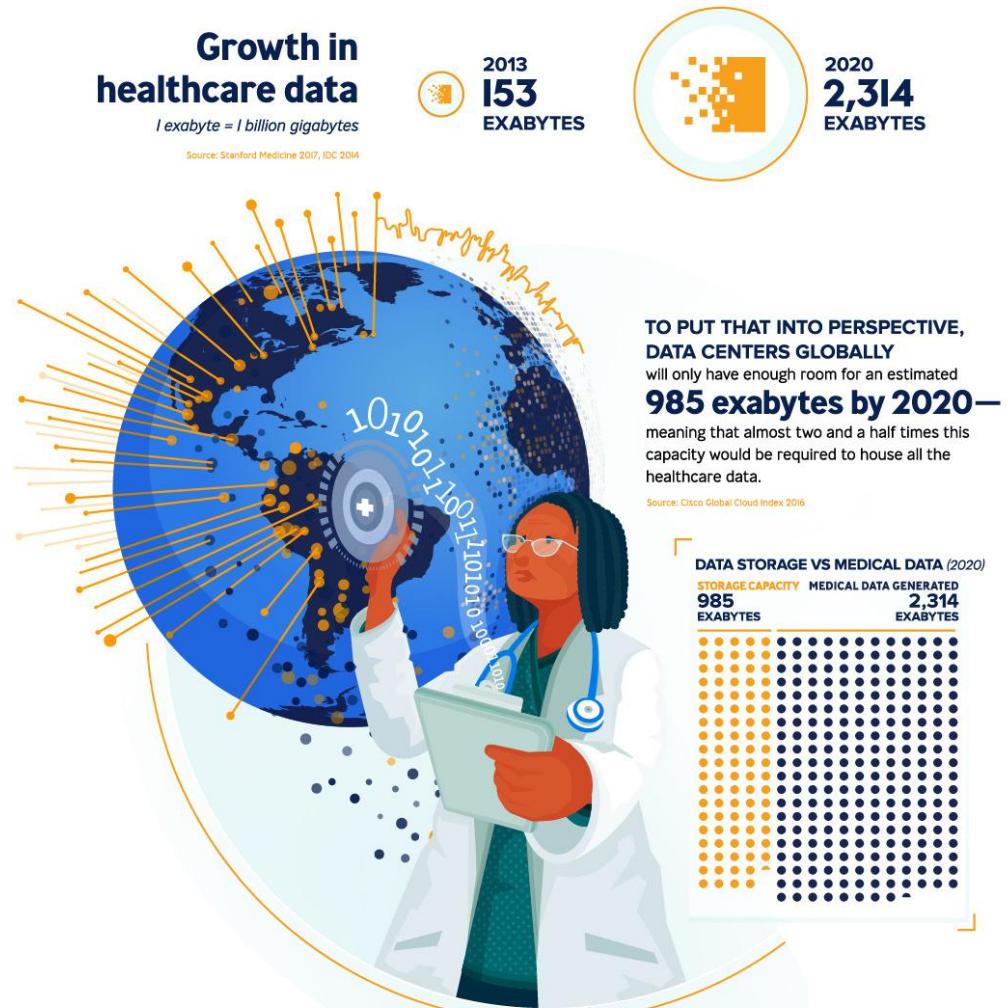


<https://www.visualcapitalist.com/big-data-healthcare/>

# Opportunity of Health Data Science

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

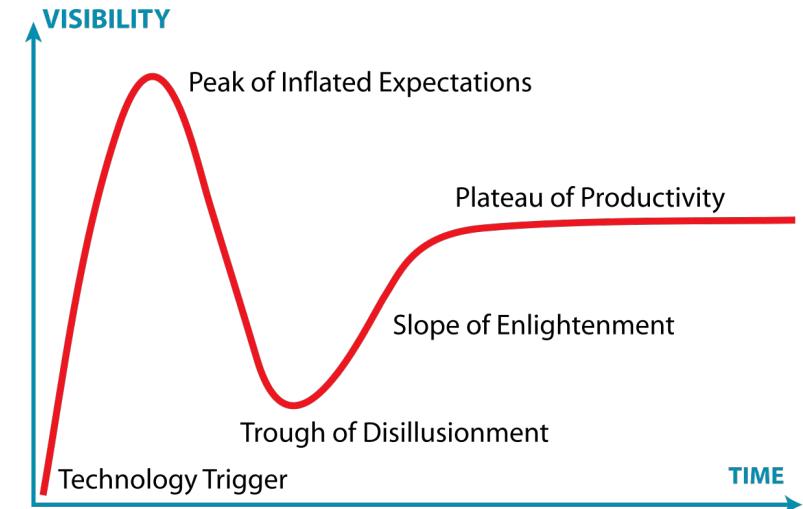
- Huge amounts of health data
- Many **interesting and important problems**
- Many domain experts desperate for data-related help with these problems
- Relative few skilled data science practitioners



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

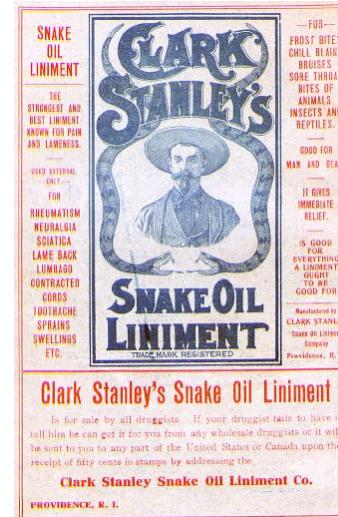
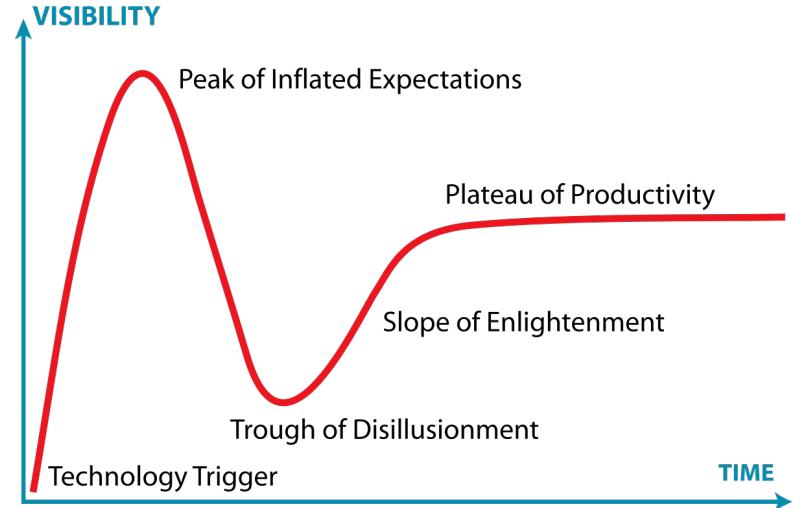
# (Some) Challenges of Health Data Science

- Lots of hype



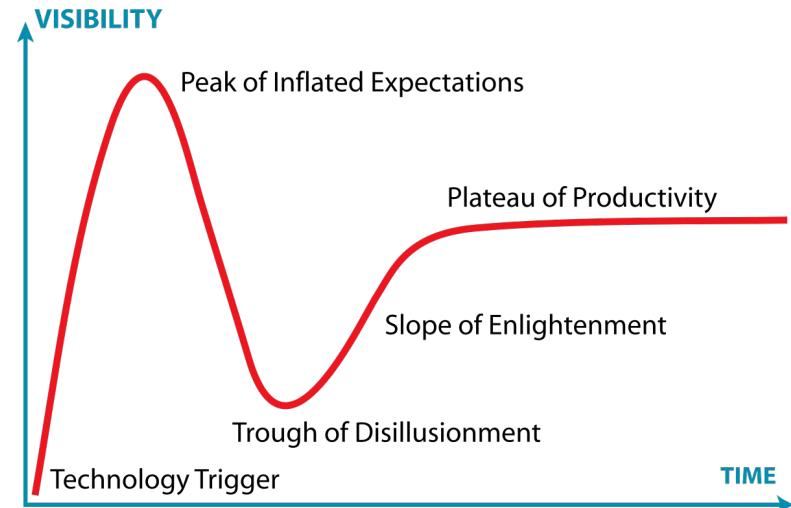
# (Some) Challenges of Health Data Science

- Lots of hype
- Lots of grifters

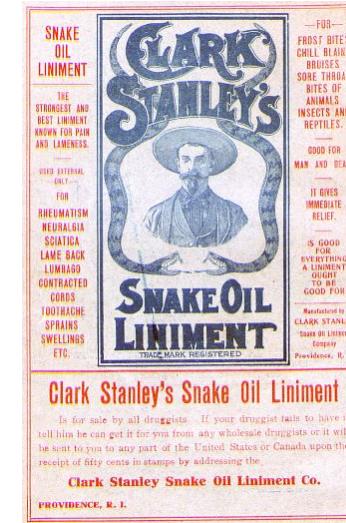


# (Some) Challenges of Health Data Science

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



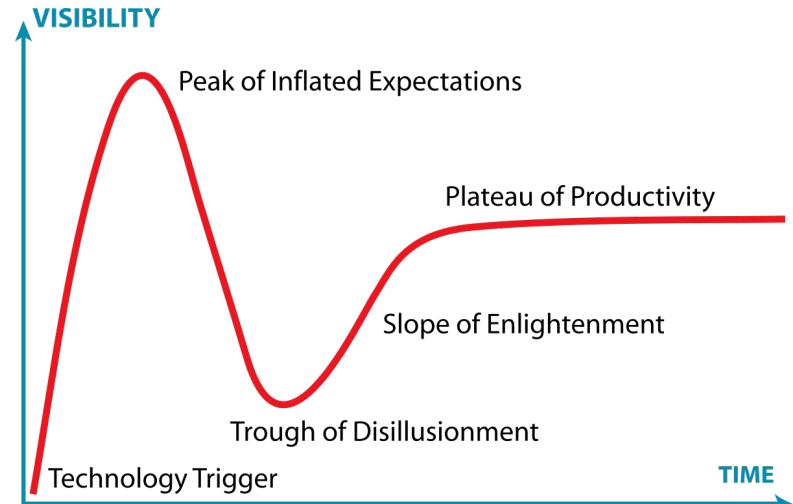
<https://www.r-bloggers.com/2019/08/new-course-learn-advanced-data-cleaning-in-r/>



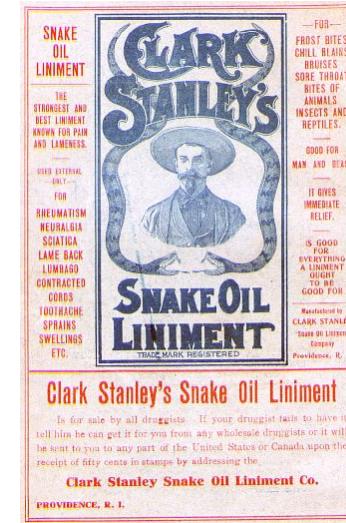
[https://upload.wikimedia.org/wikipedia/commons/9/94/Gartner\\_Hype\\_Cycle.svg](https://upload.wikimedia.org/wikipedia/commons/9/94/Gartner_Hype_Cycle.svg)  
[https://commons.wikimedia.org/wiki/File:Clark\\_Stanley%27s\\_Snake\\_Oil\\_Liniment.png](https://commons.wikimedia.org/wiki/File:Clark_Stanley%27s_Snake_Oil_Liniment.png)

# (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



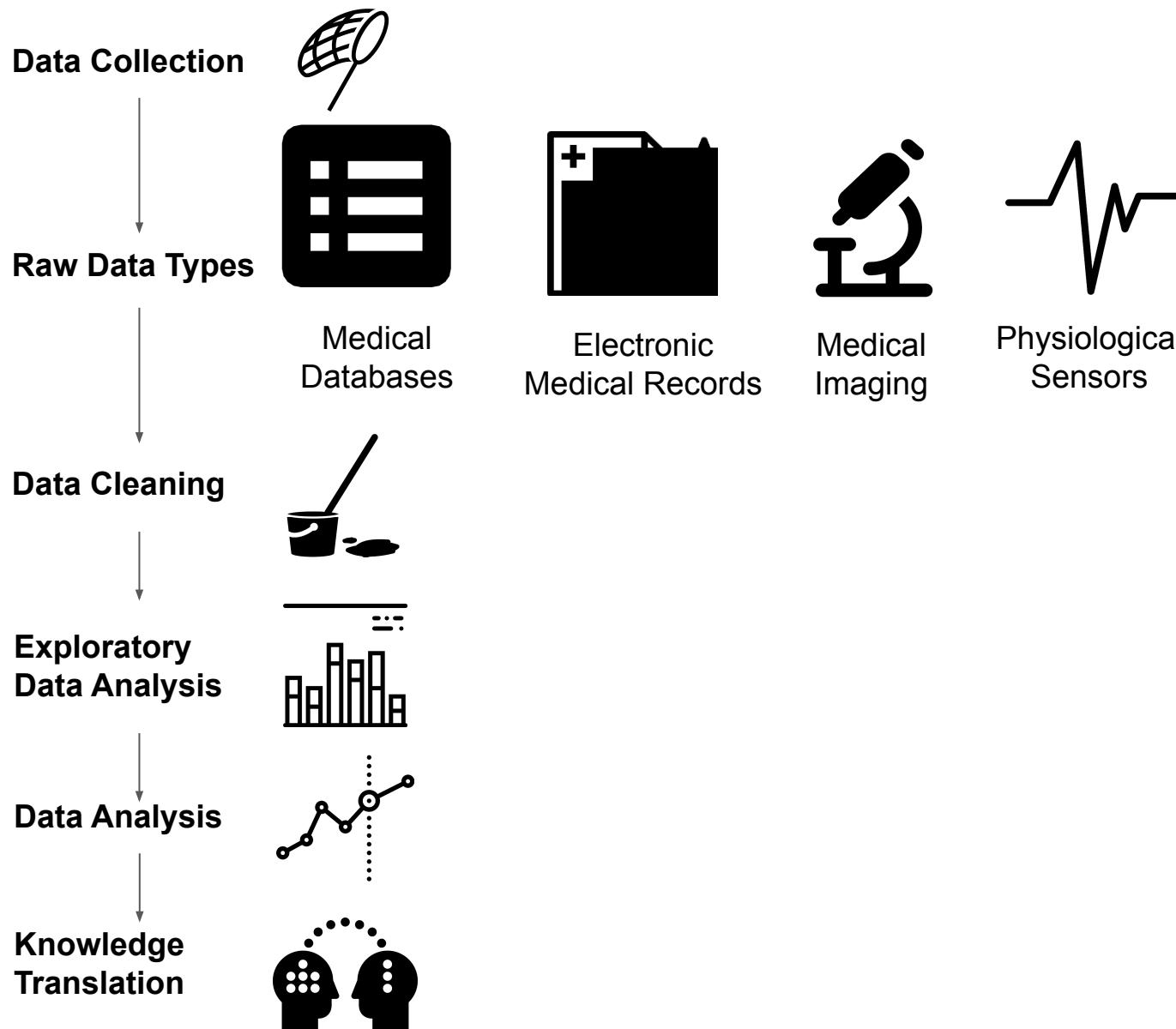
<https://www.r-bloggers.com/2019/08/new-course-learn-advanced-data-cleaning-in-r/>



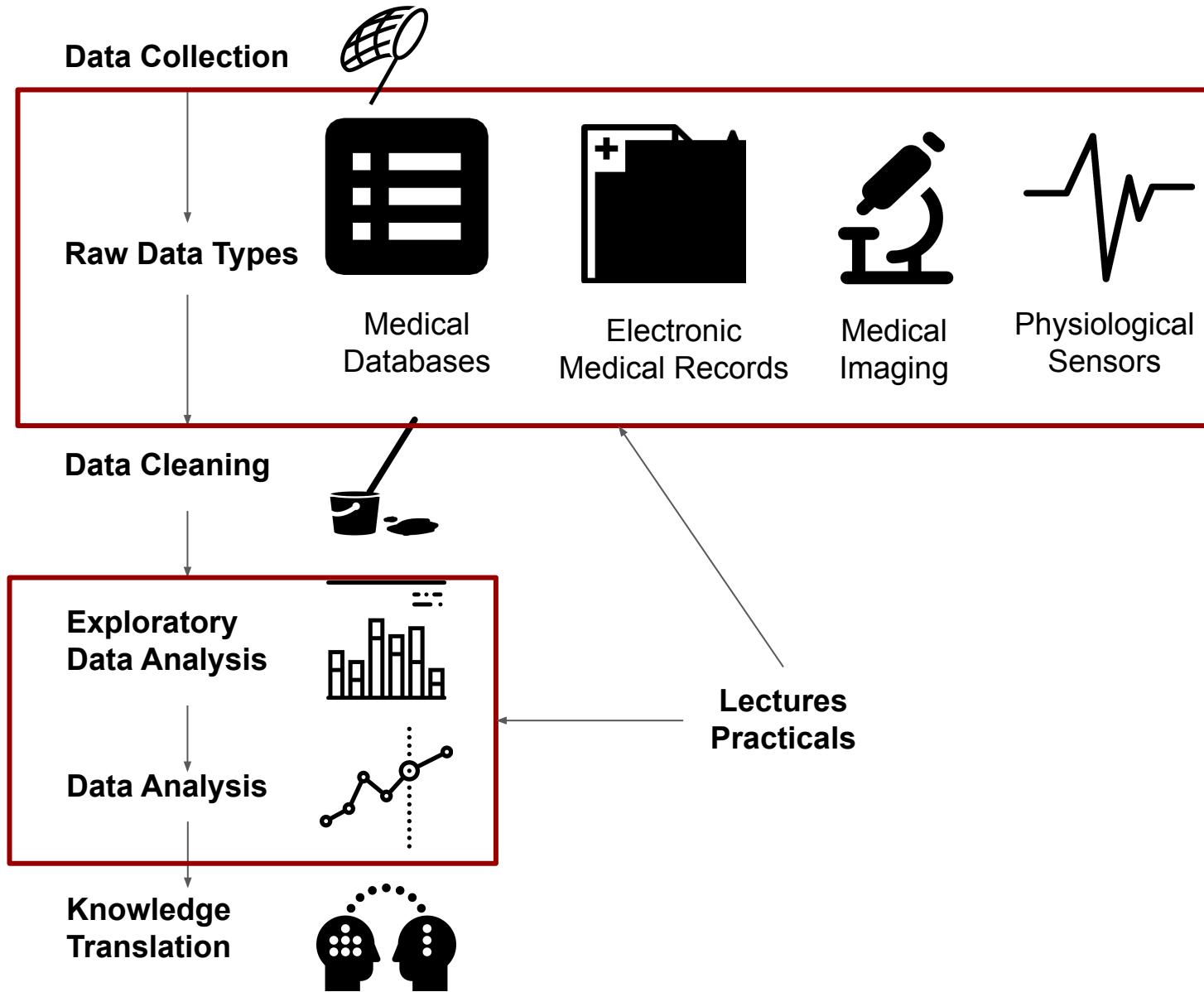
[https://upload.wikimedia.org/wikipedia/commons/9/94/Gartner\\_Hype\\_Cycle.svg](https://upload.wikimedia.org/wikipedia/commons/9/94/Gartner_Hype_Cycle.svg)  
[https://commons.wikimedia.org/wiki/File:Clark\\_Stanley%27s\\_Snake\\_Oil\\_Liniment.png](https://commons.wikimedia.org/wiki/File:Clark_Stanley%27s_Snake_Oil_Liniment.png)

What parts of health data science will this course cover?

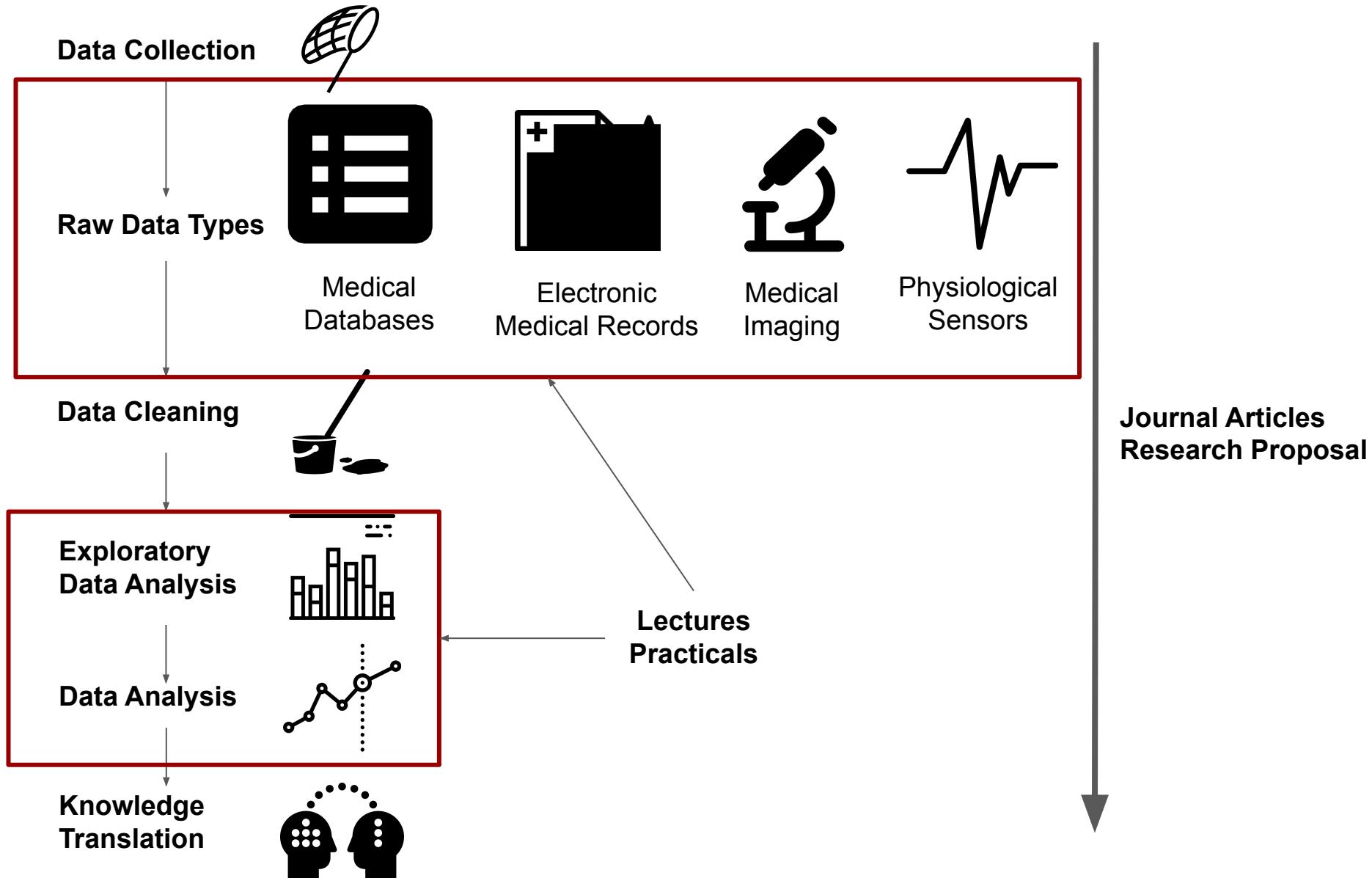
# What parts of health data science will this course cover?



# What parts of health data science will this course cover?



# What parts of health data science will this course cover?



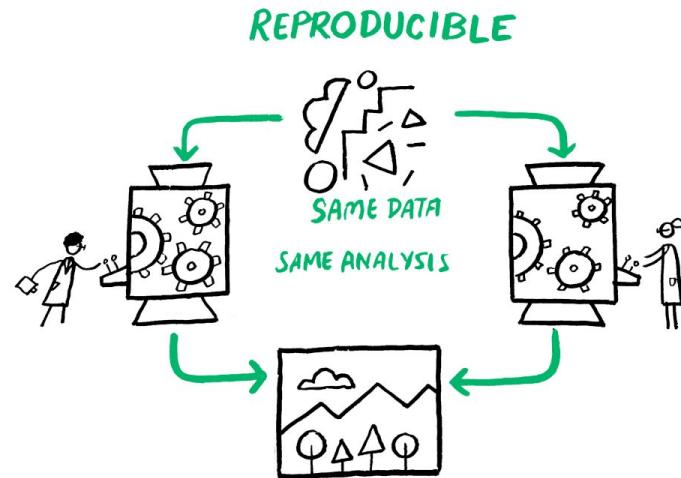
Let's take a 5 minute break!

# Tools for Reproducible Health Data Science

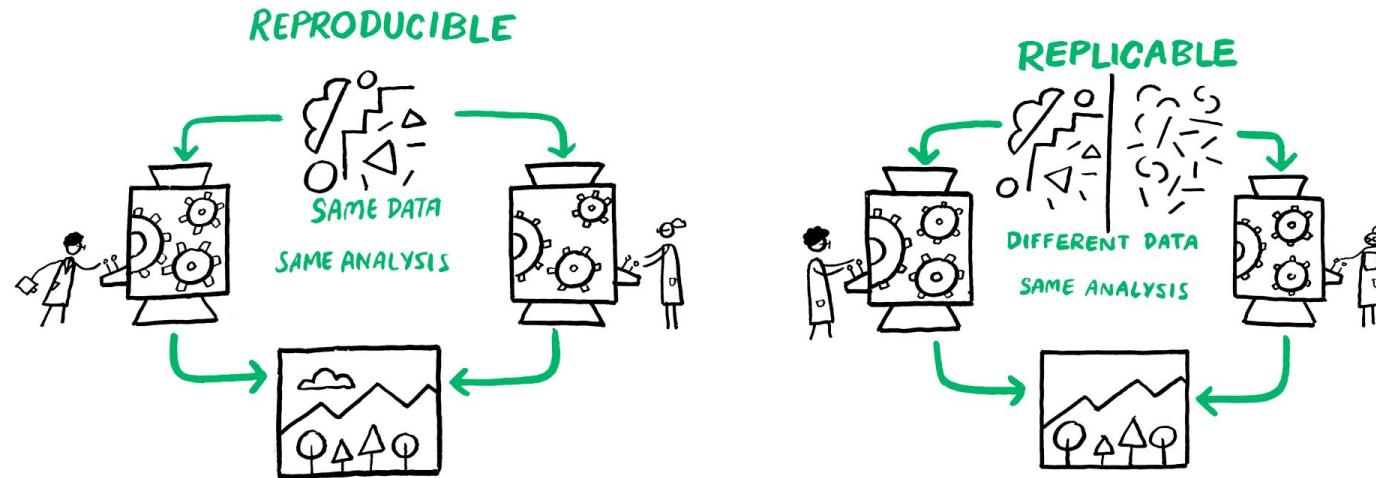
Rstudio, Rmarkdown, Git

Why do we care about reproducibility?

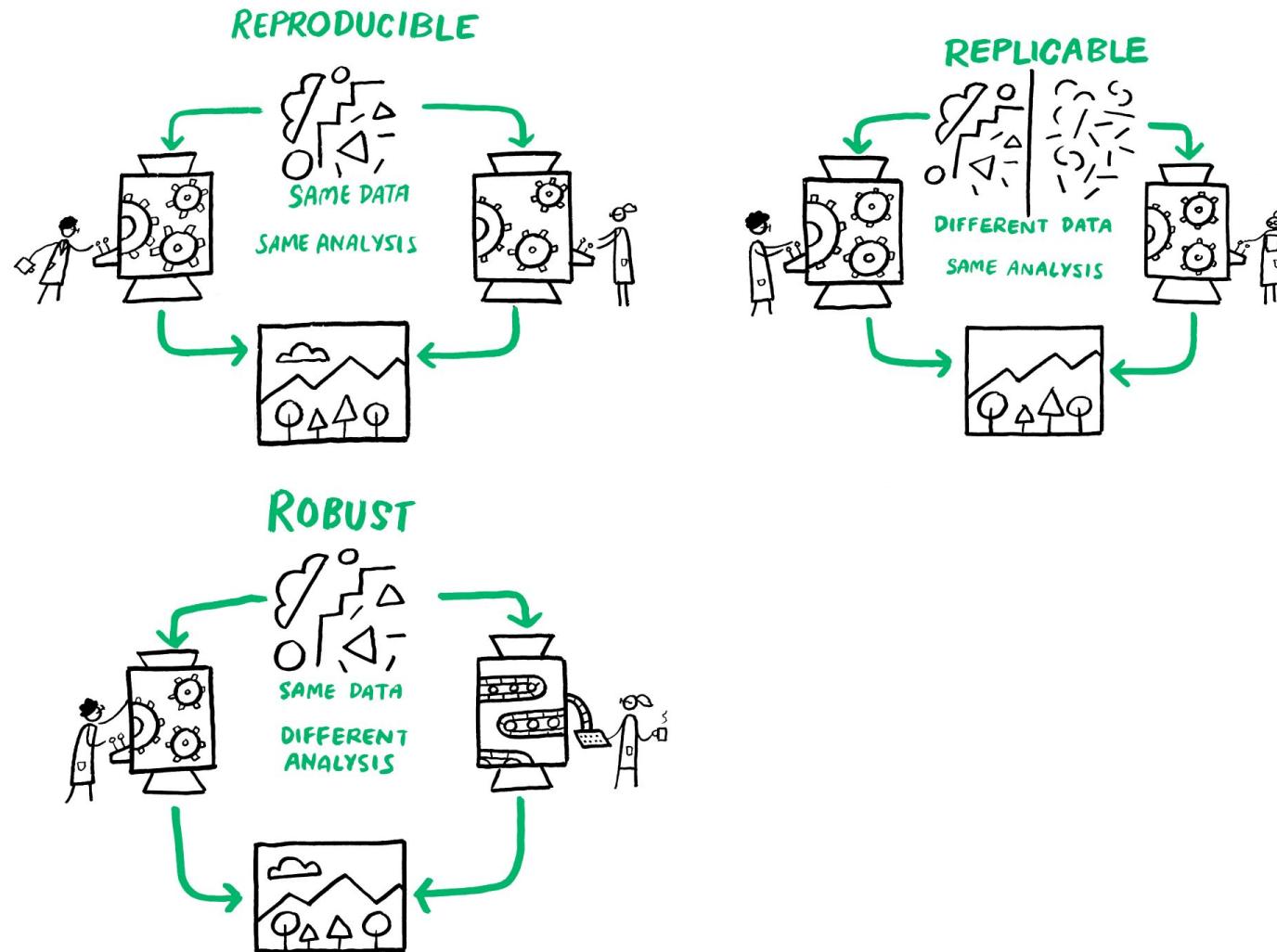
# Reproducibility should be the bare minimum



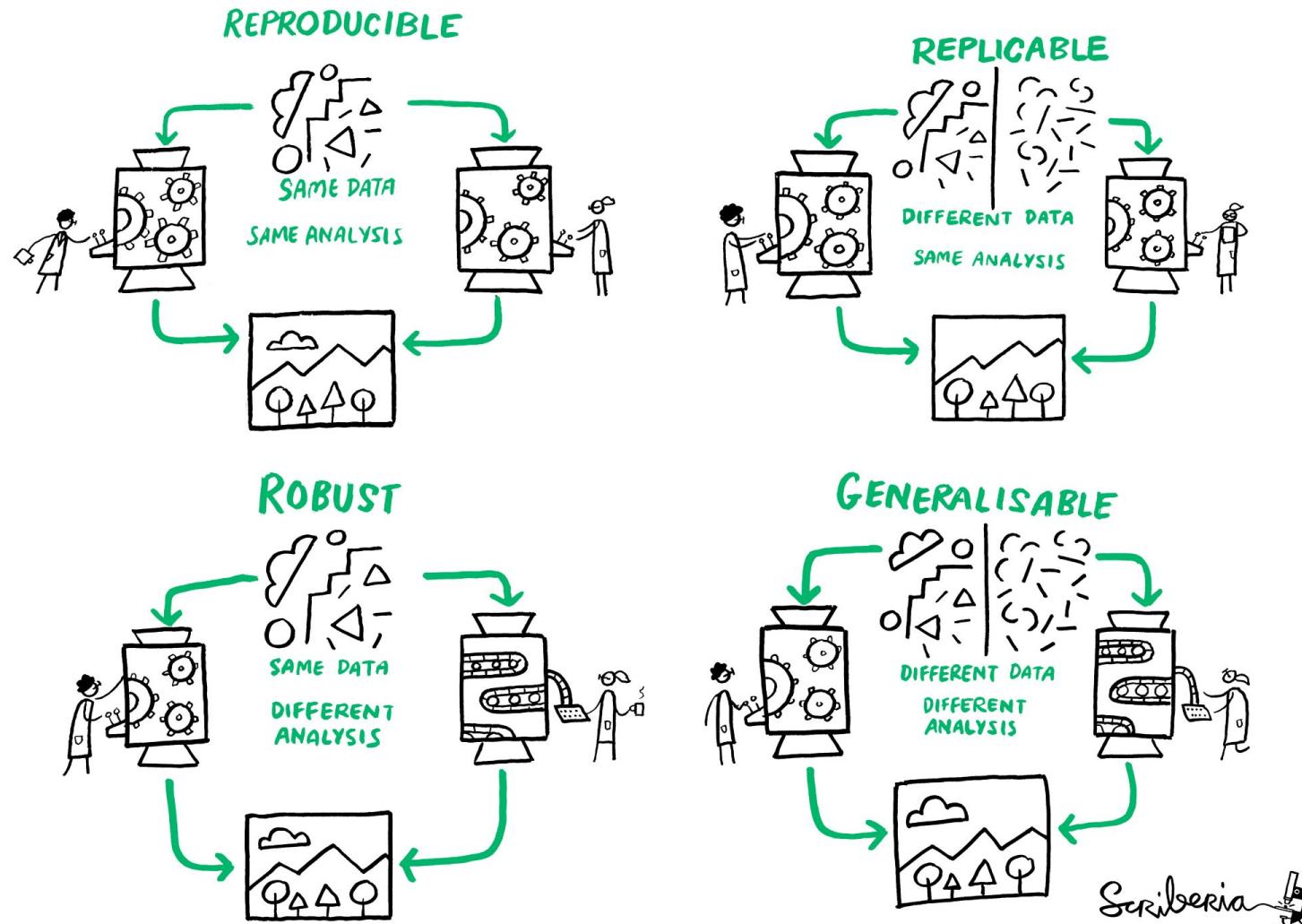
# Reproducibility should be the bare minimum



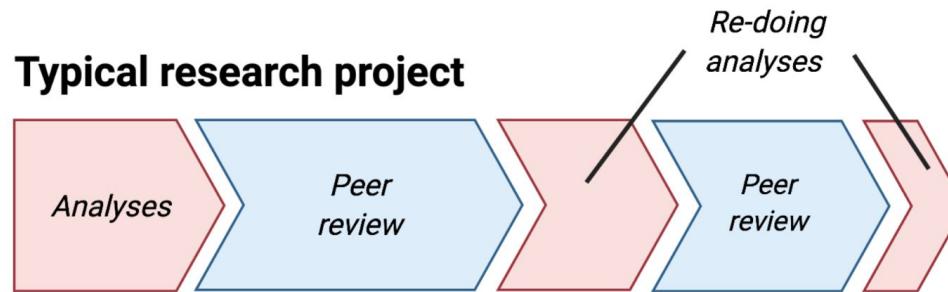
# Reproducibility should be the bare minimum



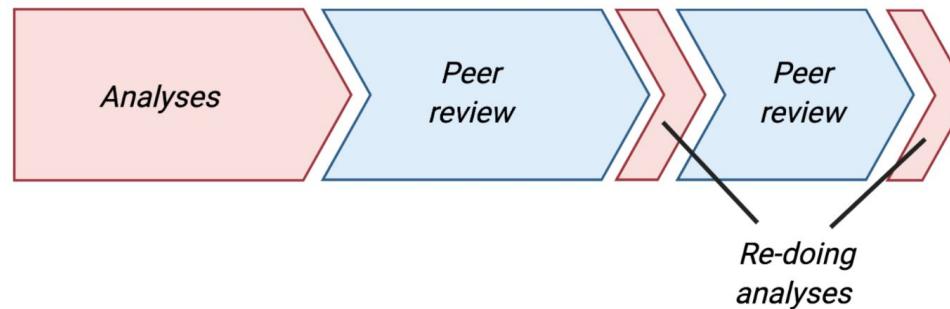
# Reproducibility should be the bare minimum



# Makes your own life easier



## Research project using reproducible practices



@dsquintana

[oliviergimenez.github.io/reproducible-science-workshop](http://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

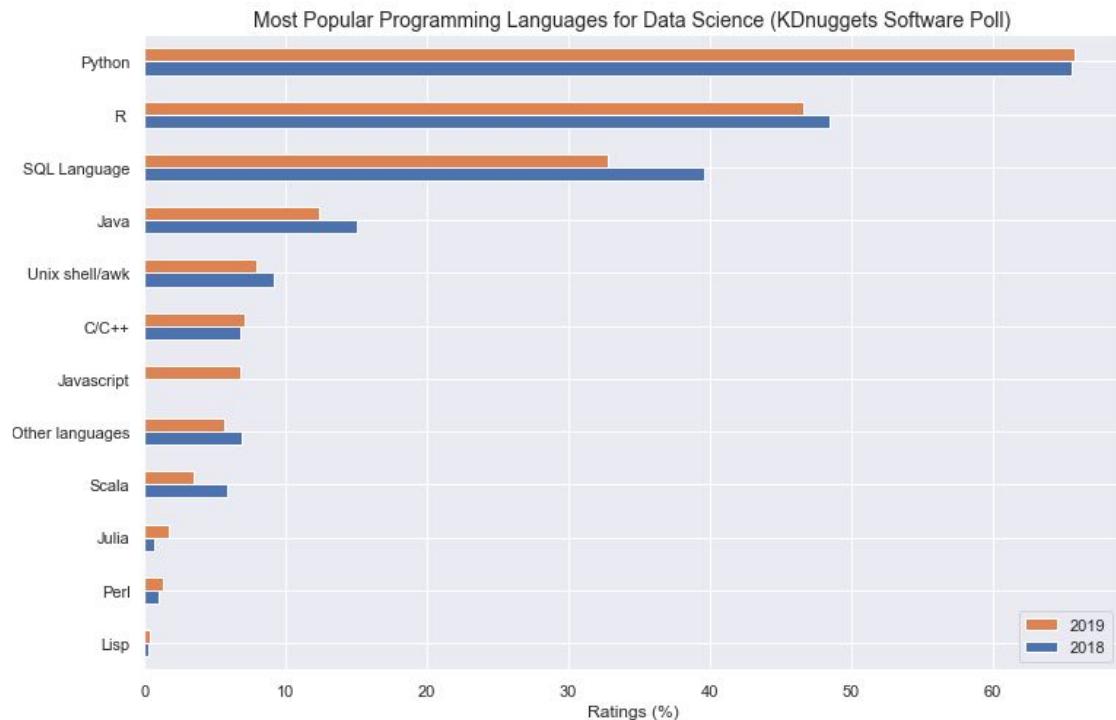
# Reproducibility checklist

- 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?

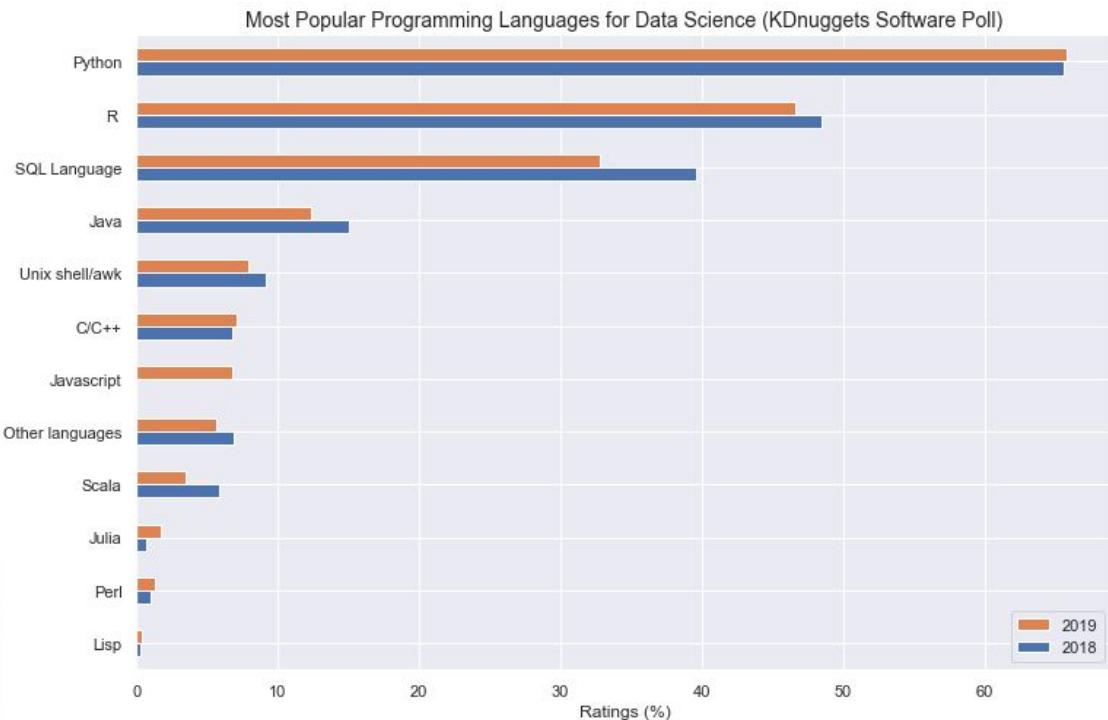
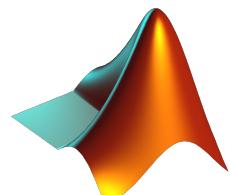
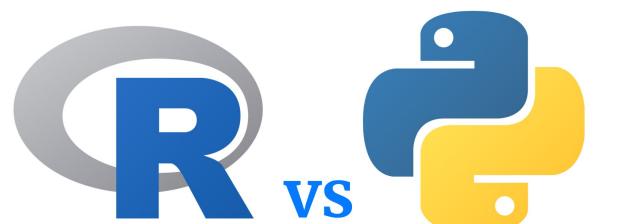
Choose a language that makes it easy to do most/all of your analysis

# 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()`

The screenshot shows the RStudio interface with the following components:

- Code Editor:** The "flights-example.R" script is open, displaying R code for loading the "nycflights13" dataset and creating a boxplot. The code includes imports for `nycflights13`, `lubridate`, `dplyr`, and `ggplot2`. It then filters the data to get daily counts, adds a weekday column, and creates a boxplot titled "Number of 2013 New York Flights Each Weekday".
- Console:** Shows the execution of the R code. It starts by loading the packages, then defining the "daily" tibble, and finally running the ggplot command.
- Environment:** Shows the "daily" tibble in the Global Environment, which has 365 observations and 3 variables: date (Date), n (int), and wday (ord.factor).
- Plots:** A boxplot titled "Number of 2013 New York Flights Each Weekday" is displayed. The y-axis is labeled "Flights" and ranges from 700 to 1000. The x-axis is labeled "Weekday" and shows categories for Sun, Mon, Tue, Wed, Thu, Fri, and Sat. The plot shows that flight volumes are highest on Monday and Friday, and lowest on Saturday.

# Use notebooks to document analyses: Rmarkdown

~/Documents/rmarkdown - gh-pages - RStudio

rmarkdown

9-notebook.Rmd

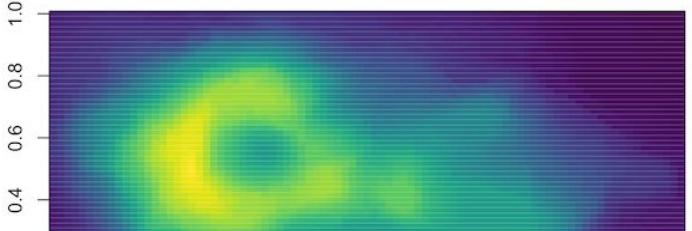
```
1 ---  
2 title: "Viridis Notebook"  
3 output: html_notebook  
4 ---  
5  
6 ```{r include = FALSE}  
7 library(viridis)  
8 ...  
9  
10 The code below demonstrates two color palettes in the  
11 [viridis](https://github.com/sjmgarnier/viridis) package. Each  
12 plot displays a contour map of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

Viridis Notebook

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

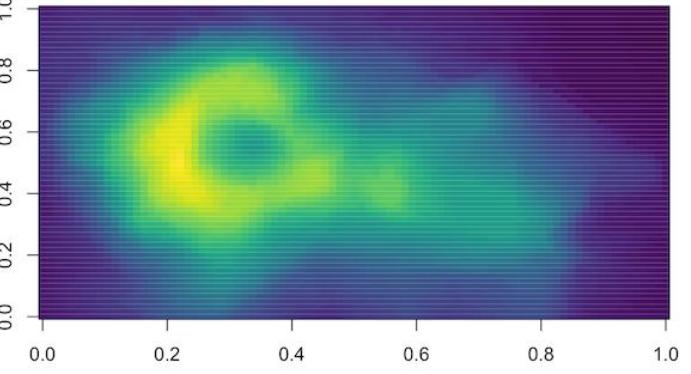
Viridis colors

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



Magma colors

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



Show All Code  
Hide All Code  
Download Rmd

Code

Environment History Build Git

Files Plots Packages Help Viewer

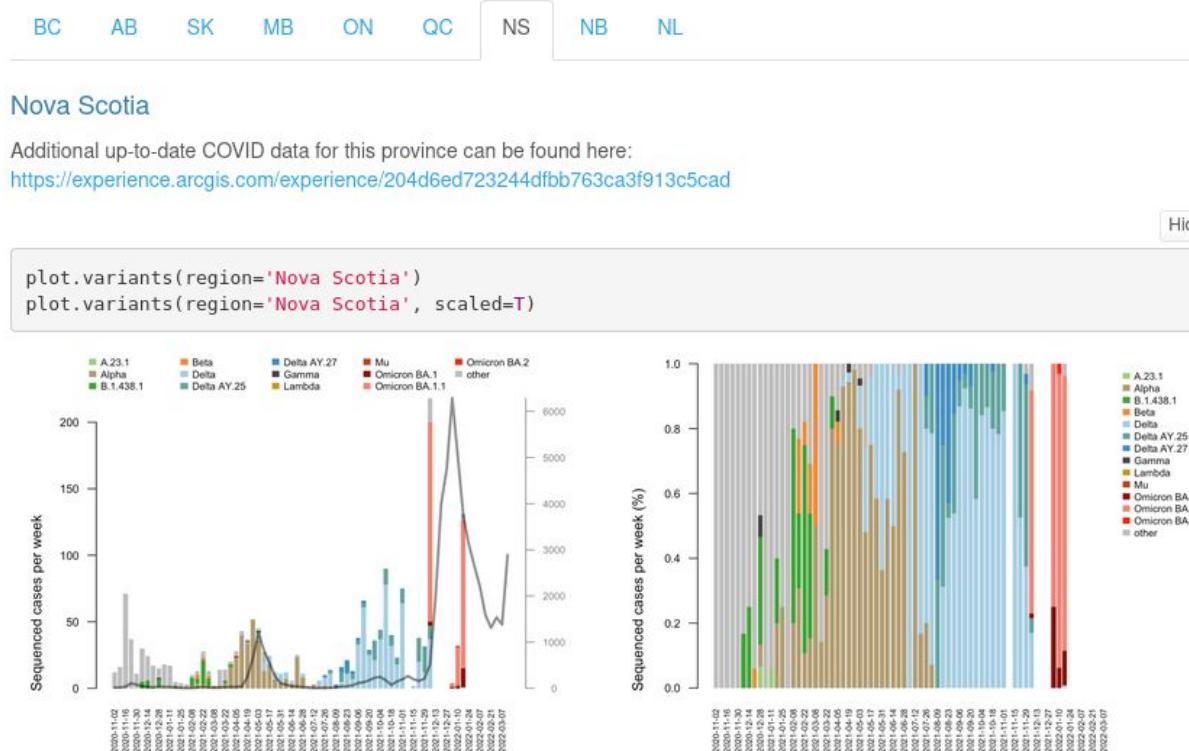
Console

R Markdown

3.22 Viridis Notebook Hide

# Use notebooks to document analyses: Rmarkdown

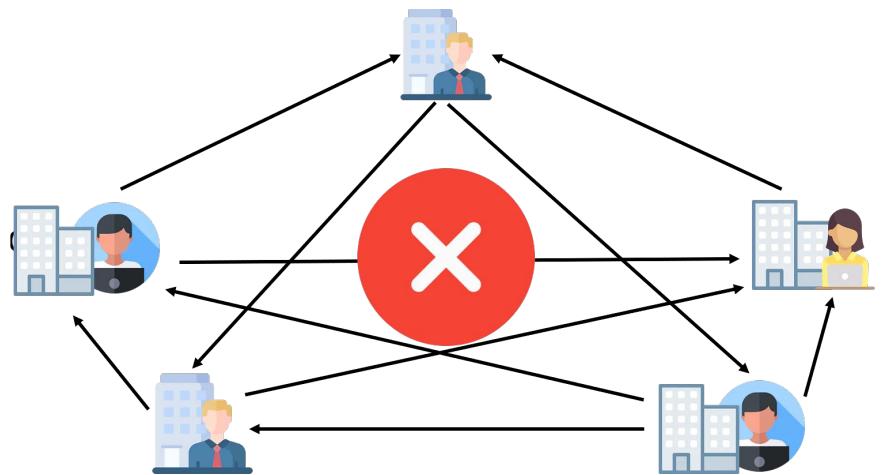
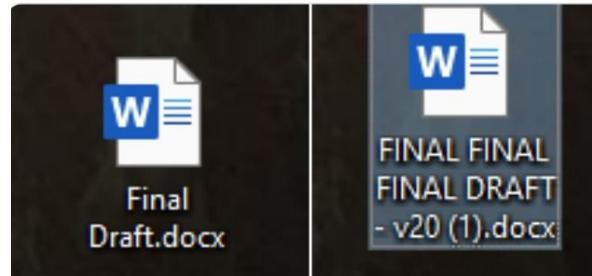
settings). Therefore, from this time onward, case counts are likely underestimated and the sequenced virus diversity is not necessarily representative of the virus circulating in the overall population.



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

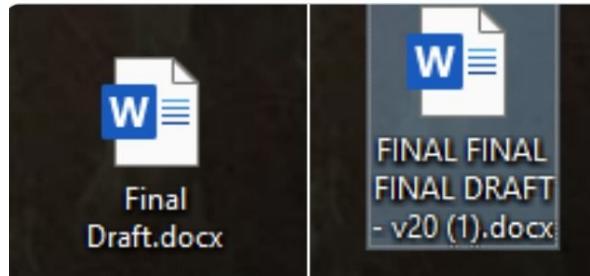
# Use standard version control systems

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



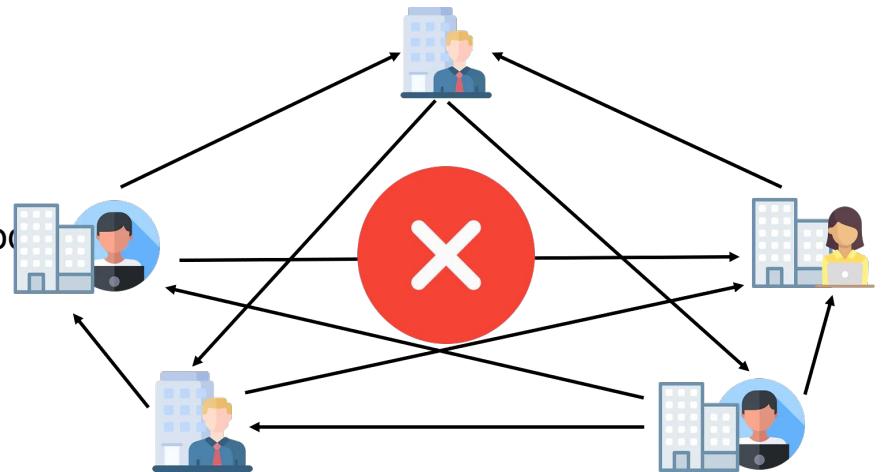
# Use standard version control systems

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

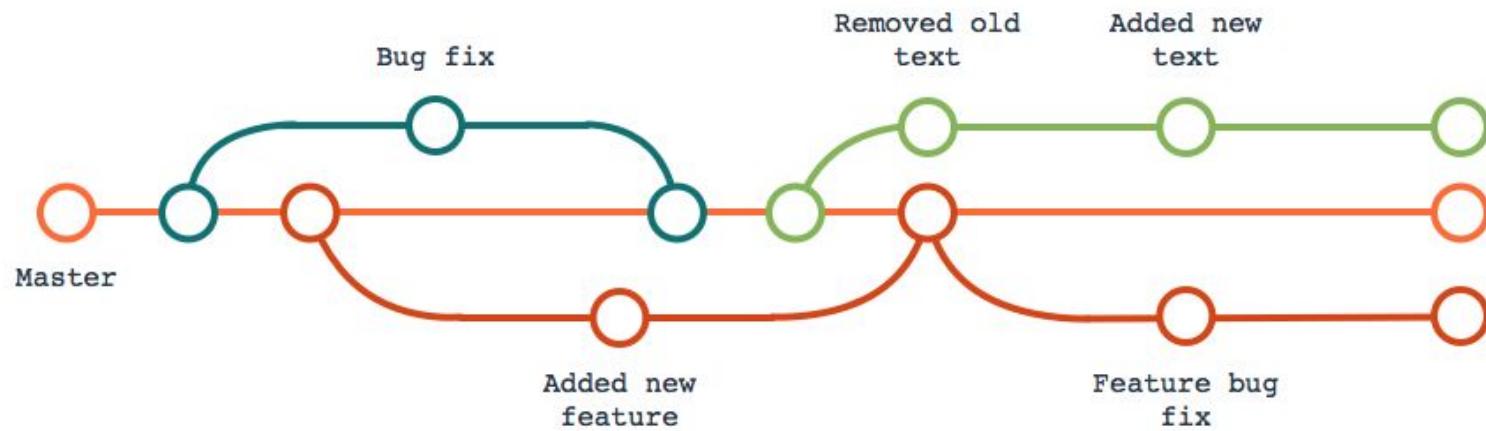


Version control let's you:

- 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

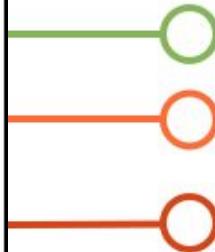


- Most popular
- Decentralised
- Designed for
- GitLab/GitHub Services

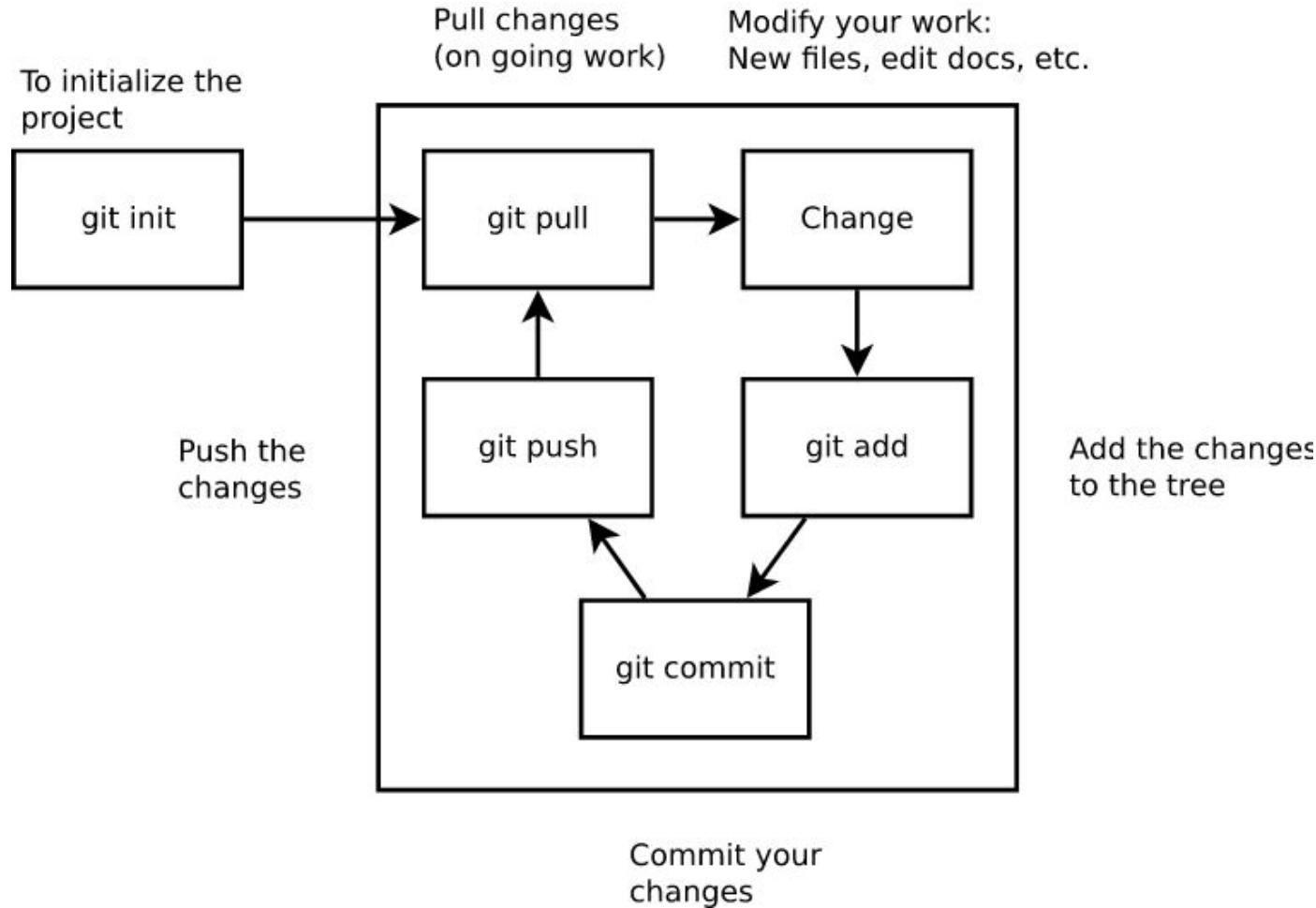
THIS IS GIT. IT TRACKS COLLABORATIVE WORK ON PROJECTS THROUGH A BEAUTIFUL DISTRIBUTED GRAPH THEORY TREE MODEL.

COOL. HOW DO WE USE IT?

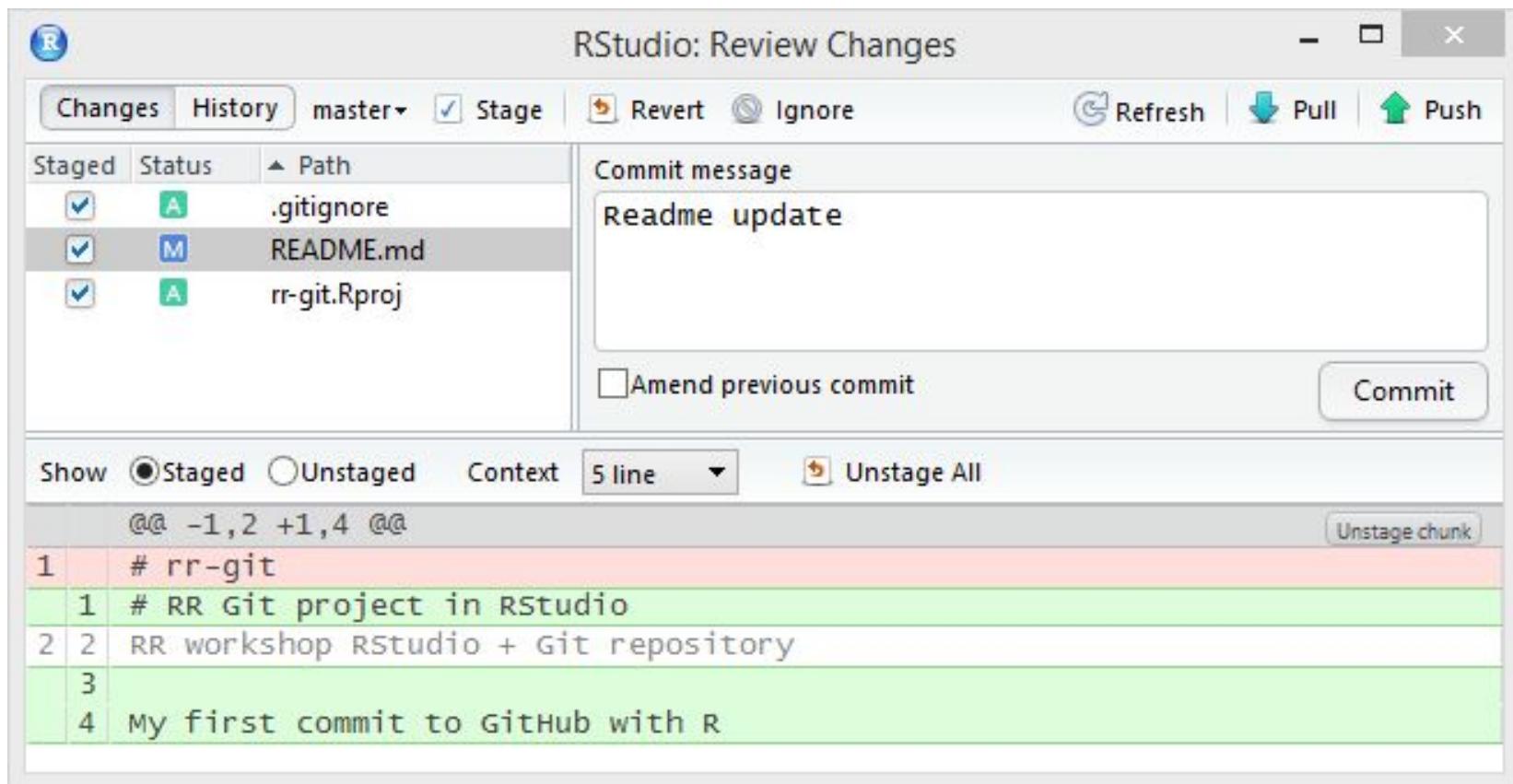
NO IDEA. JUST MEMORIZE THESE SHELL COMMANDS AND TYPE THEM TO SYNC UP.  
IF YOU GET ERRORS, SAVE YOUR WORK ELSEWHERE, DELETE THE PROJECT, AND DOWNLOAD A FRESH COPY.



# Git Workflow



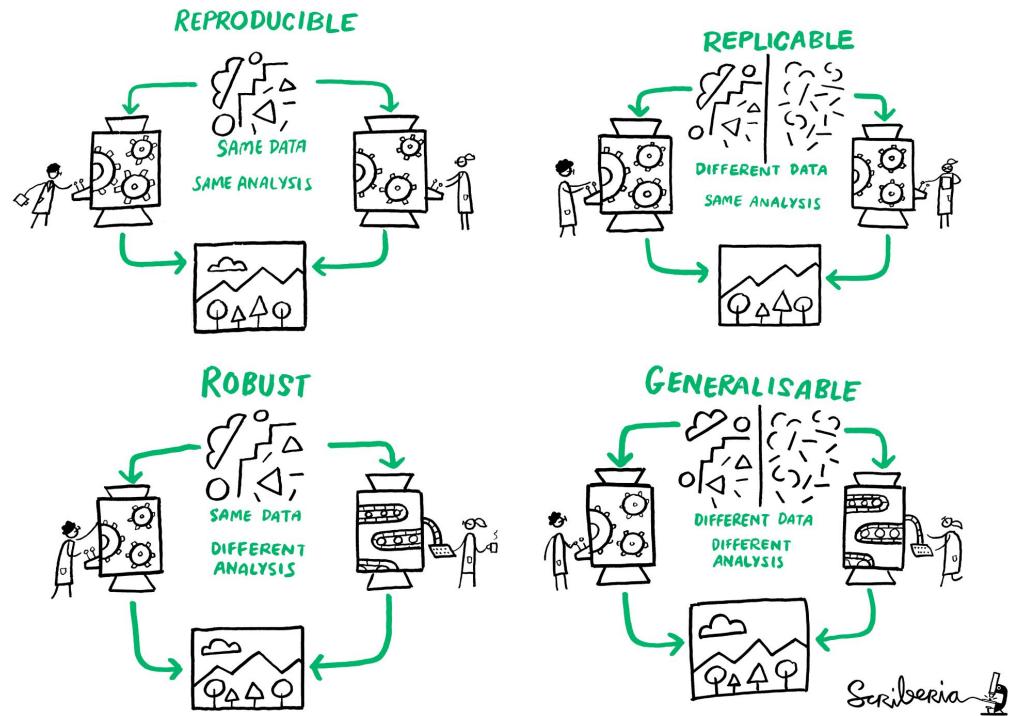
# Git is integrated into Rstudio!



# 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



# 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>