# Data visualization

Lecture 1

## The Seven Stages of Visualizing Data

**17.01.2022**

Phuc Loi Luu, PhD
p.luu@garvan.org.au
luu.p.loi@googlemail.com

# Roadmap of today lecture

- Why Data Display Requires Planning?
- Process of 7 stages in visualizing data
- Example of Methylation Risk Score (Example 1): from stage 1 to 6
- Example of MethPanel (Example 2): focus on stage 7 (interact)

# Why Data Display Requires Planning?

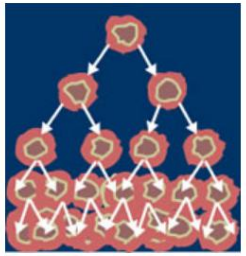
- Too Much Information
- Data Collection
- Thinking About Data
- Data Never Stays the Same
- What is the Question?
- A Combination of Many Disciplines

# Too Much Information: breast cancer (1)

## Breast Cancer Data Set

*Download*: Data Folder, Data Set Description

**Abstract**: Breast Cancer Data (Restricted Access)



| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 286 | **Area:** | Life |
| **Attribute Characteristics:** | Categorical | **Number of Attributes:** | 9 | **Date Donated** | 1988-07-11 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes | **Number of Web Hits:** | 604005 |

### Attribute Information:

1. Class: no-recurrence-events, recurrence-events
2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
3. menopause: lt40, ge40, premeno.
4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
6. node-caps: yes, no.
7. deg-malig: 1, 2, 3.
8. breast: left, right.
9. breast-quad: left-up, left-low, right-up, right-low, central.
10. irradiat: yes, no.

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

## Breast Cancer Coimbra Data Set

*Download*: Data Folder, Data Set Description

**Abstract**: Clinical features were observed or measured for 64 patients with breast cancer and 52 healthy controls.

| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 116 | **Area:** | Life |
| **Attribute Characteristics:** | Integer | **Number of Attributes:** | 10 | **Date Donated** | 2018-03-06 |
| **Associated Tasks:** | Classification | **Missing Values?** | N/A | **Number of Web Hits:** | 129855 |

### Attribute Information:

1) ID number
2) Outcome (R = recur, N = nonrecur)
3) Time (recurrence time if field 2 = R, disease-free time if field 2 = N)
4-33) Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

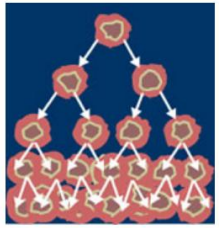
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

# Too Much Information: breast cancer (2)

## Breast Cancer Wisconsin (Diagnostic) Data Set

*Download*: Data Folder, Data Set Description

**Abstract**: Diagnostic Wisconsin Breast Cancer Database



| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 569 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 32 | **Date Donated** | 1995-11-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 1650333 |

### Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:

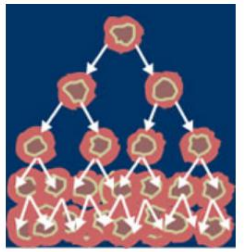
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

## Breast Cancer Wisconsin (Prognostic) Data Set

*Download*: Data Folder, Data Set Description

**Abstract**: Prognostic Wisconsin Breast Cancer Database



| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 198 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 34 | **Date Donated** | 1995-12-01 |
| **Associated Tasks:** | Classification, Regression | **Missing Values?** | Yes | **Number of Web Hits:** | 249085 |

### Attribute Information:

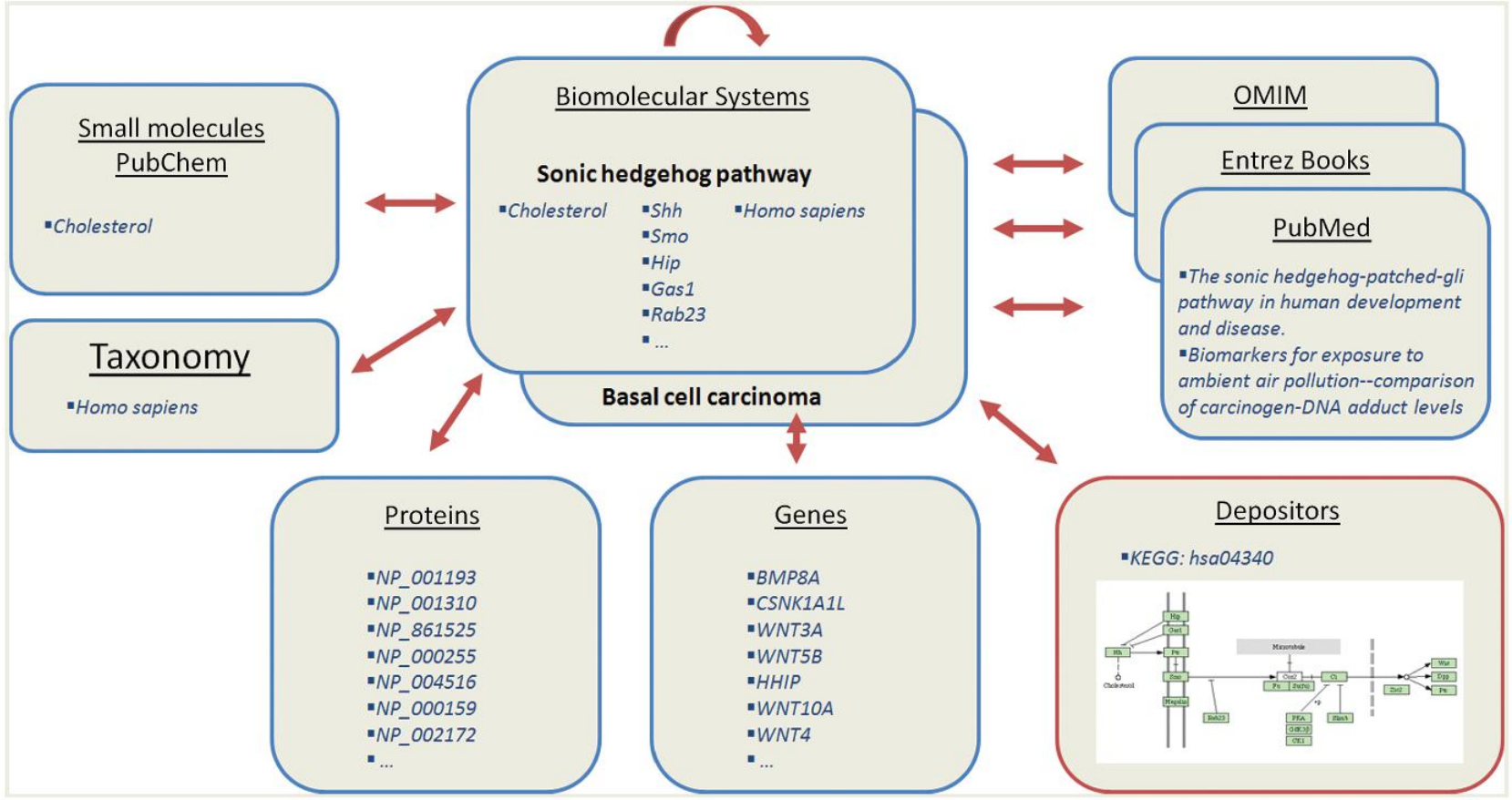
1) ID number
2) Outcome (R = recur, N = nonrecur)
3) Time (recurrence time if field 2 = R, disease-free time if field 2 = N)
4-33) Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Prognostic%29

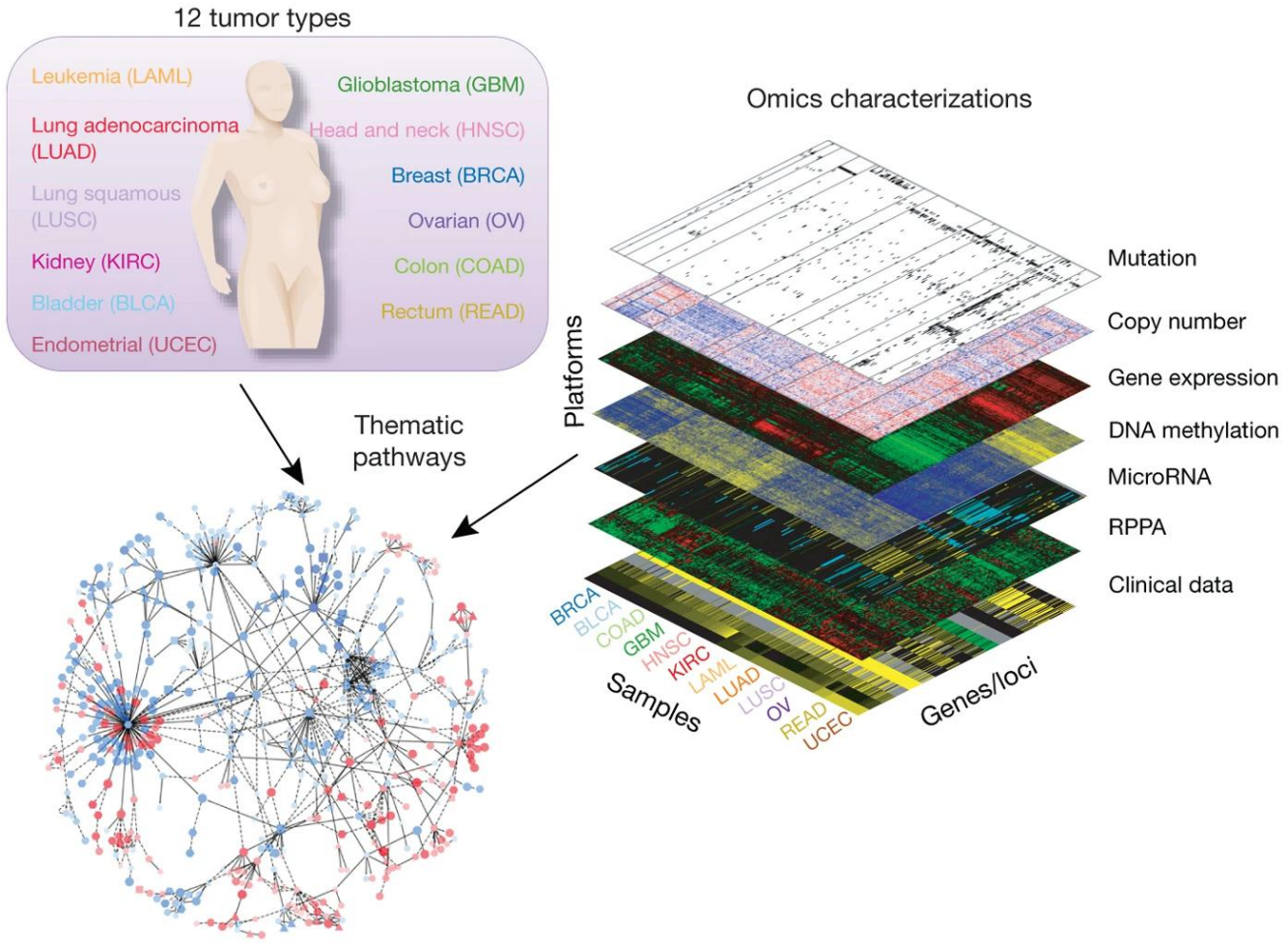
# Data Collection: NCBI (1)

- A lot of public data out there on internet
- Where to get the data?
- How to get them?
- Ethics!



# Data Collection: TCGA (2)

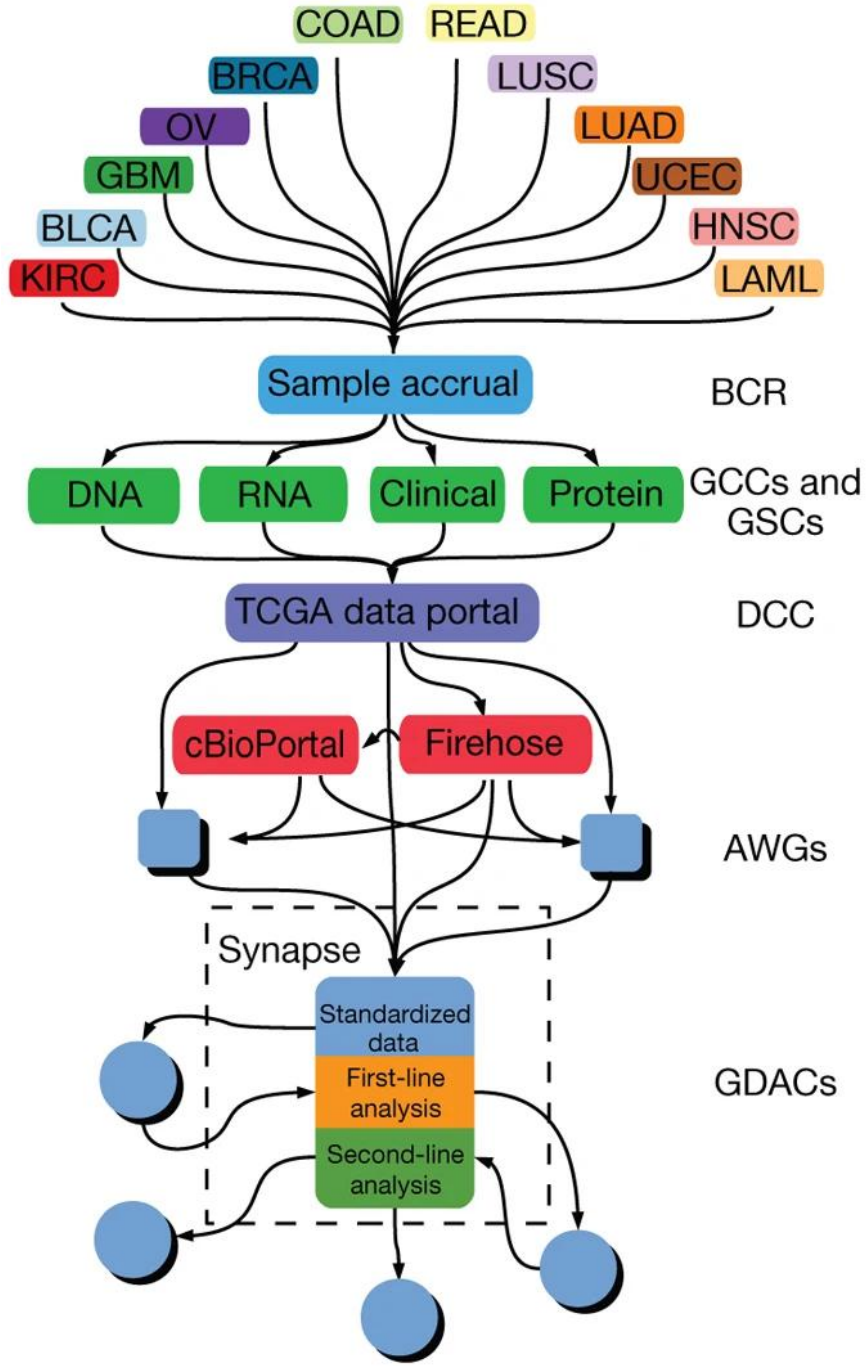
- A lot of public data out there on internet
- Where to get the data?
- How to get them?
- Connect or set up a database
- Ethics!




https://www.nature.com/articles/ng.2764

# Data Collection: TCGA (3)

- A lot of public data out there on internet
- Where to get the data?
- How to get them?
- Connect or set up a database
- Ethics!



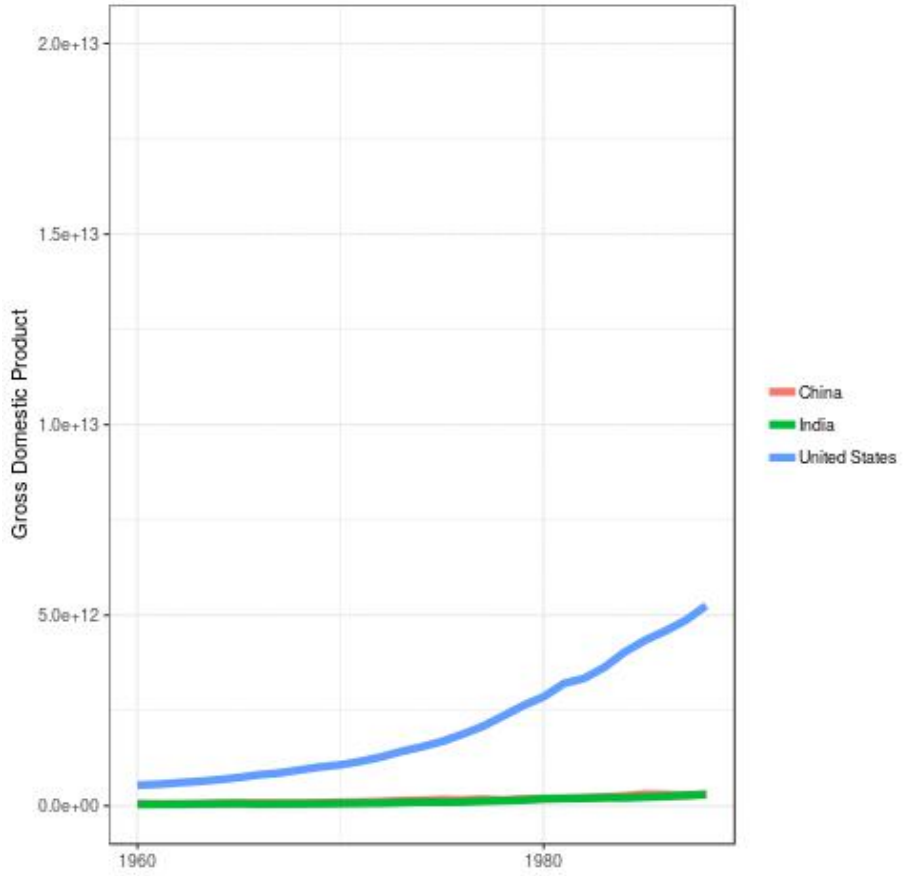

https://www.nature.com/articles/ng.2764

# Thinking About Data

- Personal/private data
- Accessed illegally
- For example credit card
- Law!

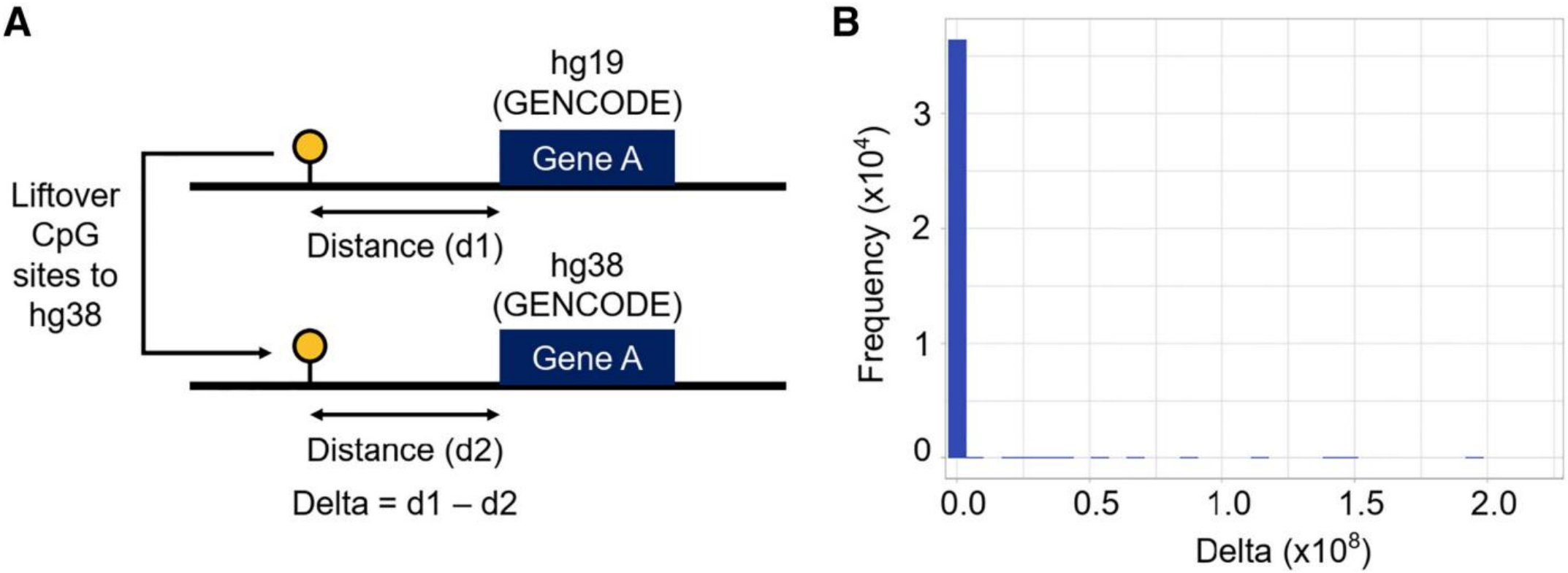
# Data Never Stays the Same

- Dataset is moving target.
- What happens when things start moving?
- How do we interact with “live” data?
- How do we unravel data as it changes over time?
- We might use animation to play back the evolution of a data set, or interaction to control what time span we’re looking at.
- How can we write code for these situations?




https://www.nature.com/articles/ng.2764

# What is the Question?



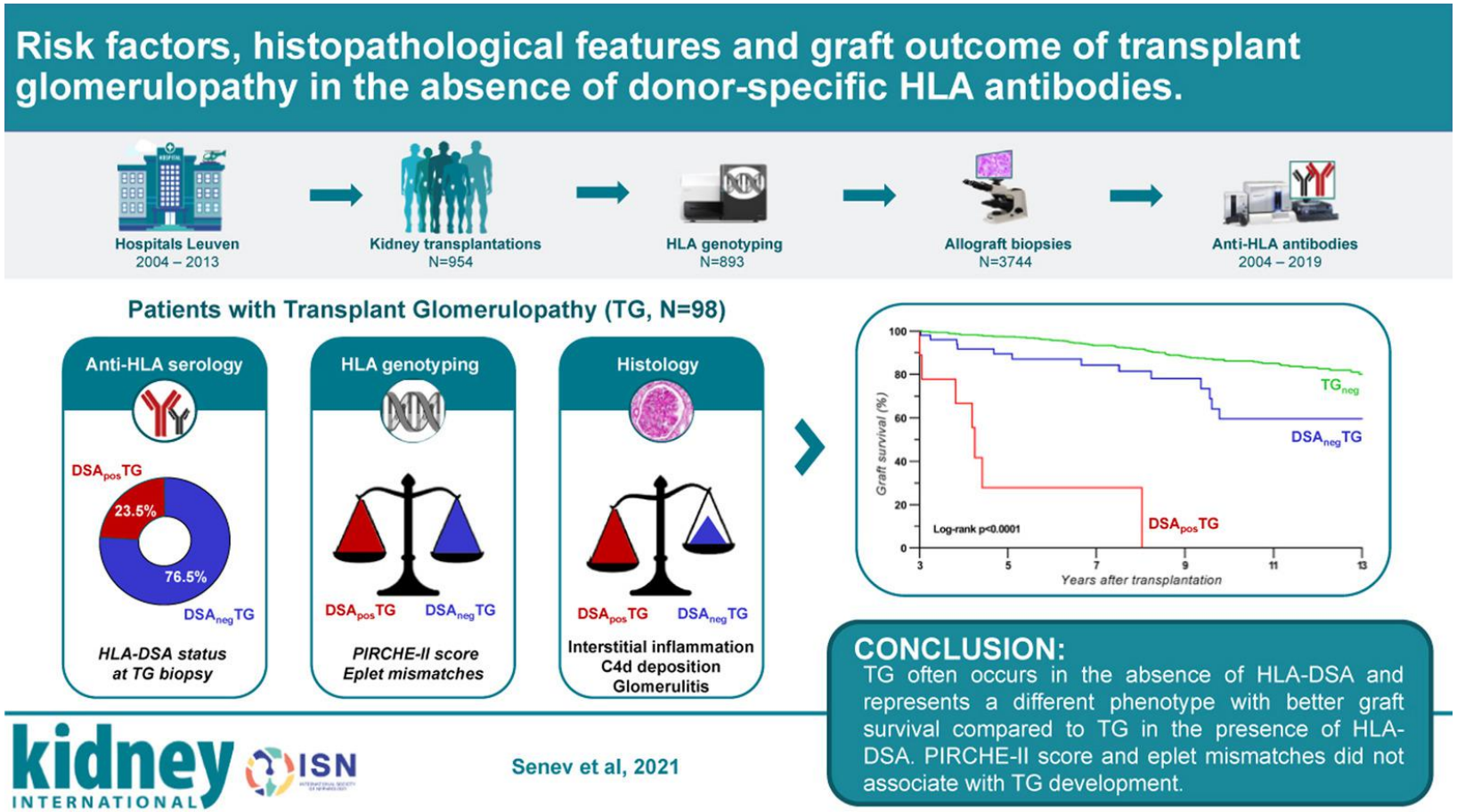

UCSC liftOver conversion of WGBS data.
(A) Steps of analysis to determine if CpG sites are mapped to the correct location.
(B) Distribution of delta values by subtracting distance between initial CpGs and gene coordinates on hg19 (d1) by distance between lifted CpGs and gene coordinates on hg38 (d2).

Luu, P. L., Ong, P. T., Dinh, T. P., & Clark, S. J. (2020). Benchmark study comparing liftover tools for genome conversion of epigenome sequencing data. NAR genomics and bioinformatics, 2(3), lqaa054.

# A Combination of Many Disciplines

- Given the complexity of data, using it to provide a meaningful solution requires insights from diverse fields:
  - statistics
  - data mining
  - graphic design
  - information visualization

# A Combination of Many Disciplines, example



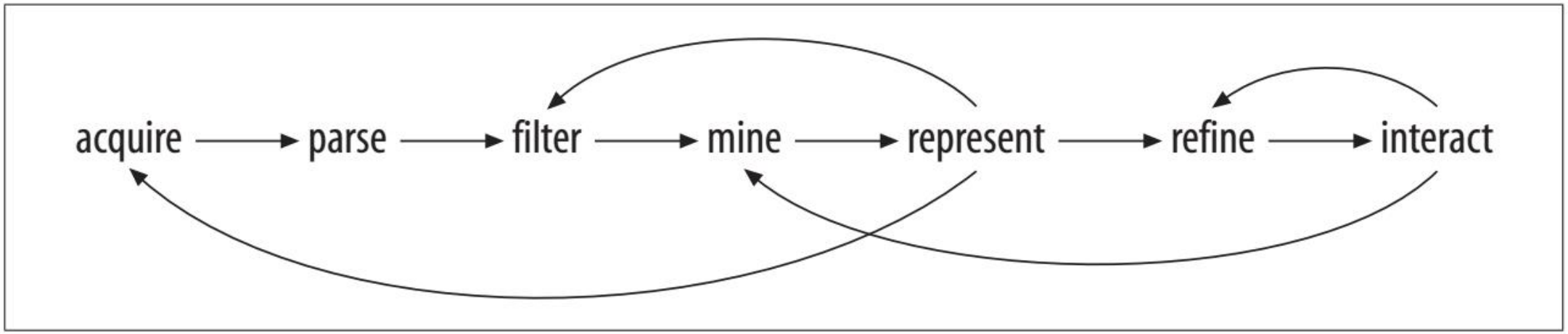

10.1016/j.kint.2021.01.029

# Process of 7 stages in visualizing data

The process of understanding data begins with a set of numbers and a question. The following steps form a path to the answer:

1. *Acquire*
   Obtain the data, whether from a file on a disk or a source over a network.
2. *Parse*
   Provide some structure for the data's meaning, and order it into categories.
3. *Filter*
   Remove all but the data of interest.
4. *Mine*
   Apply methods from statistics or data mining as a way to discern patterns or place the data in mathematical context.
5. *Represent*
   Choose a basic visual model, such as a bar graph, list, or tree.
6. *Refine*
   Improve the basic representation to make it clearer and more visually engaging.
7. *Interact*
   Add methods for manipulating the data or controlling what features are visible.

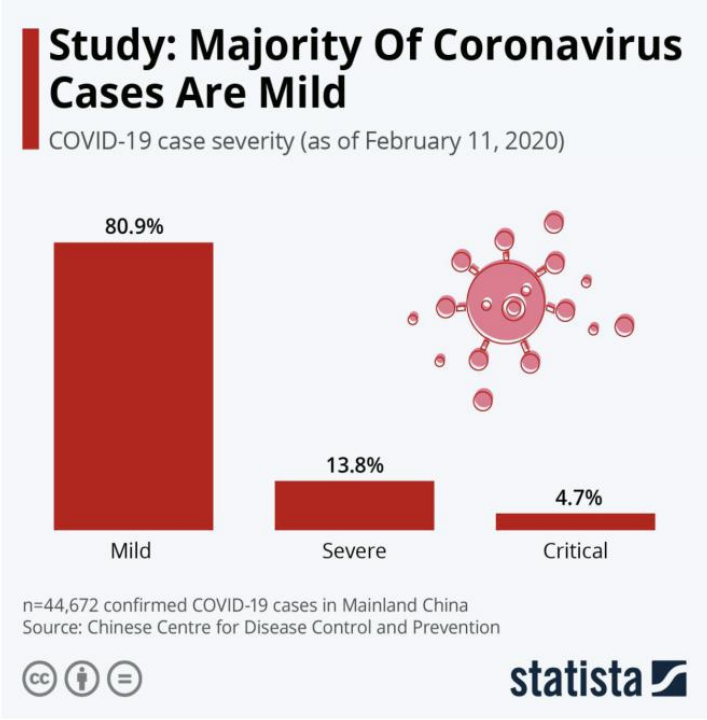
# Interactions between 7 stages



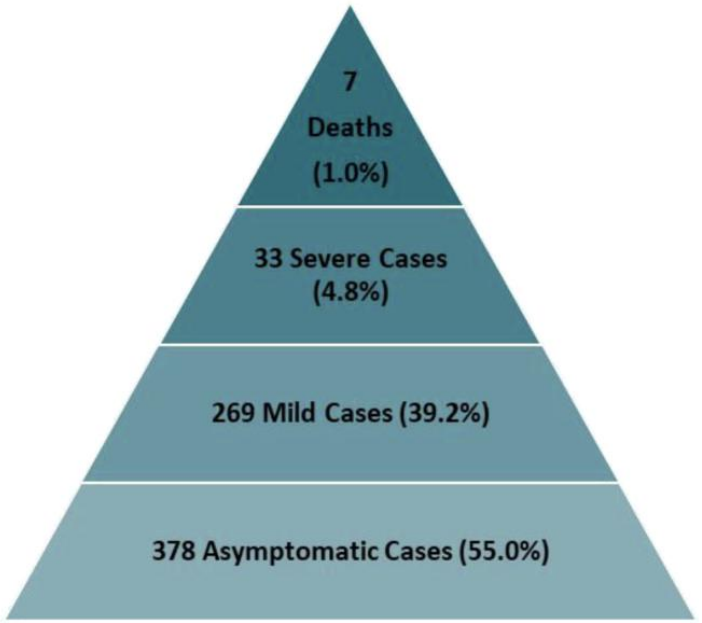
# Example 1: Methylation Risk Score (MRS)

## What is the question?

Developing Methylation Risk Score for stratification of COVID-19 serverity




https://www.statista.com/chart/20856/coronavirus-case-severity-in-china/




Adapted from Expert Taskforce for the COVID-19 Cruise Ship Outbreak. Proportion of fatal, severe, mild, and asymptomatic COVID-19 cases among 544 passengers and 143 crew.

https://www.cdc.gov/library/covid19/090120_covidupdate.html

# Example 1: Methylation Risk Score (MRS)

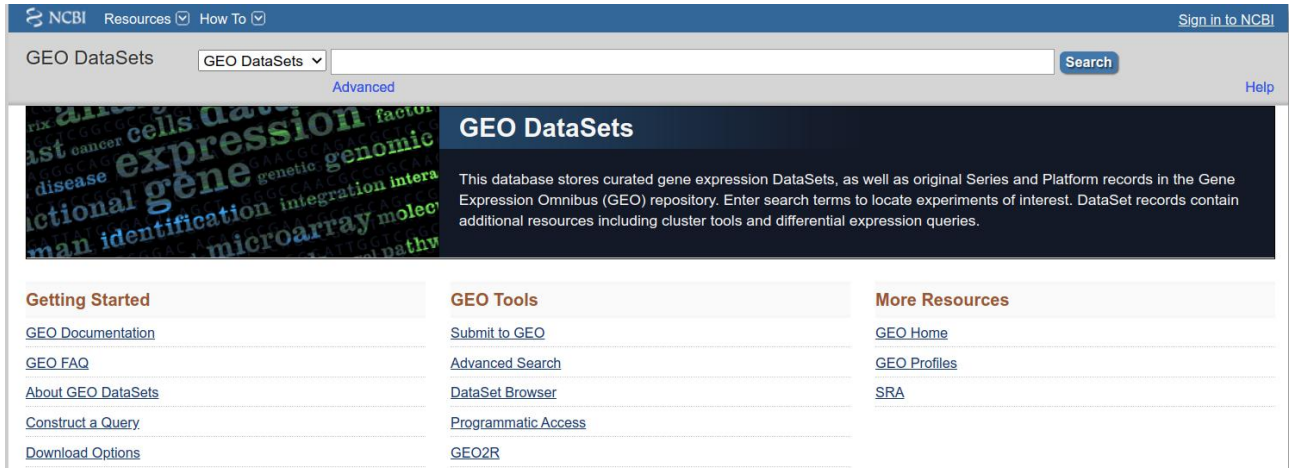
What is the question?

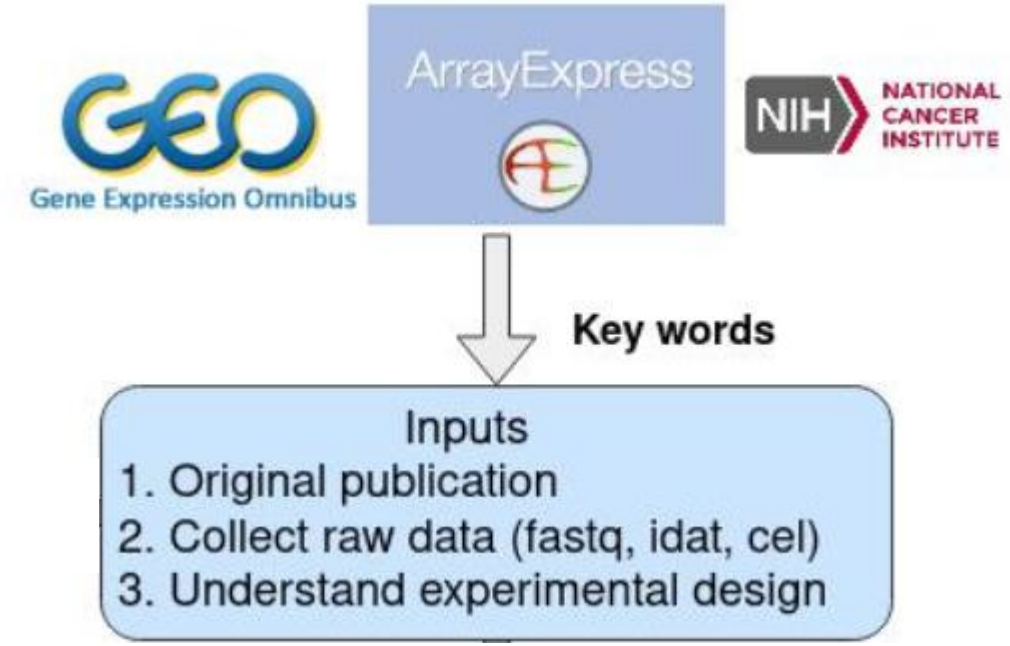
COVID-19 serverity = ICU + Dead

COVID-19 Mild = non-ICU + non-Dead

# Example 1: Methylation Risk Score (MRS)

## 1. **Acquire**: 850K/EPIC array + clinical record



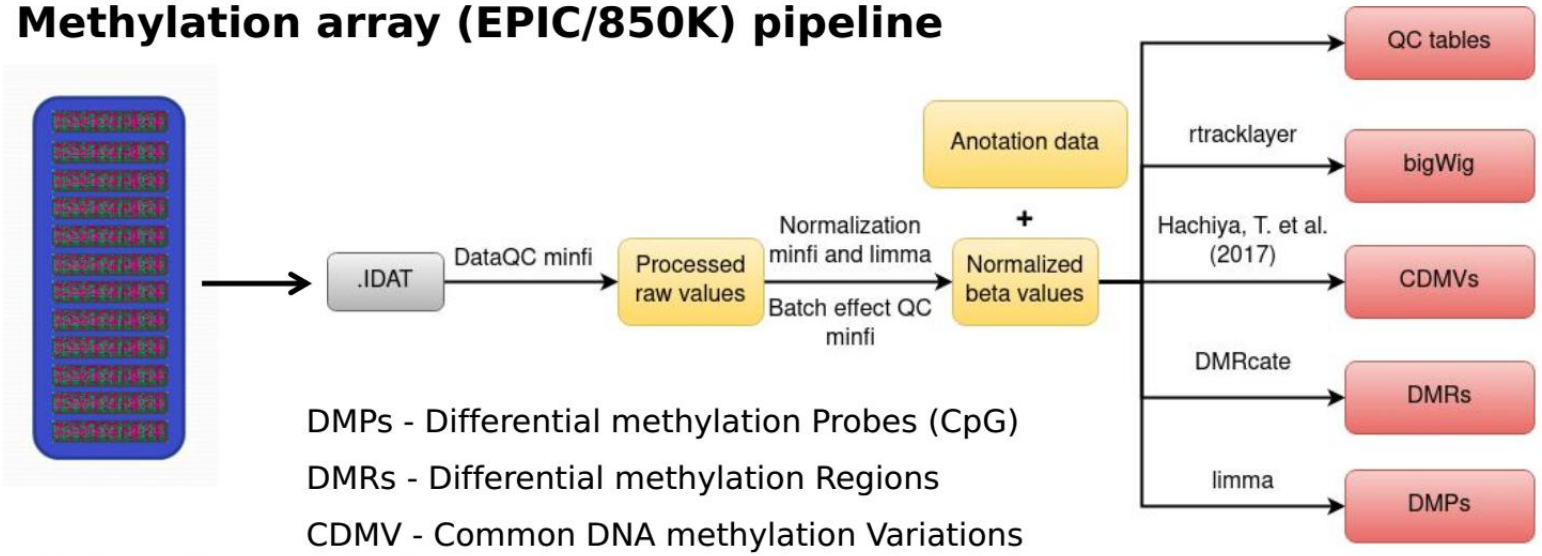




wget/curl + bash script

https://www.ncbi.nlm.nih.gov/gds

# Example 1: Methylation Risk Score (MRS)

## **2. Parse**: pre-processing for the right format



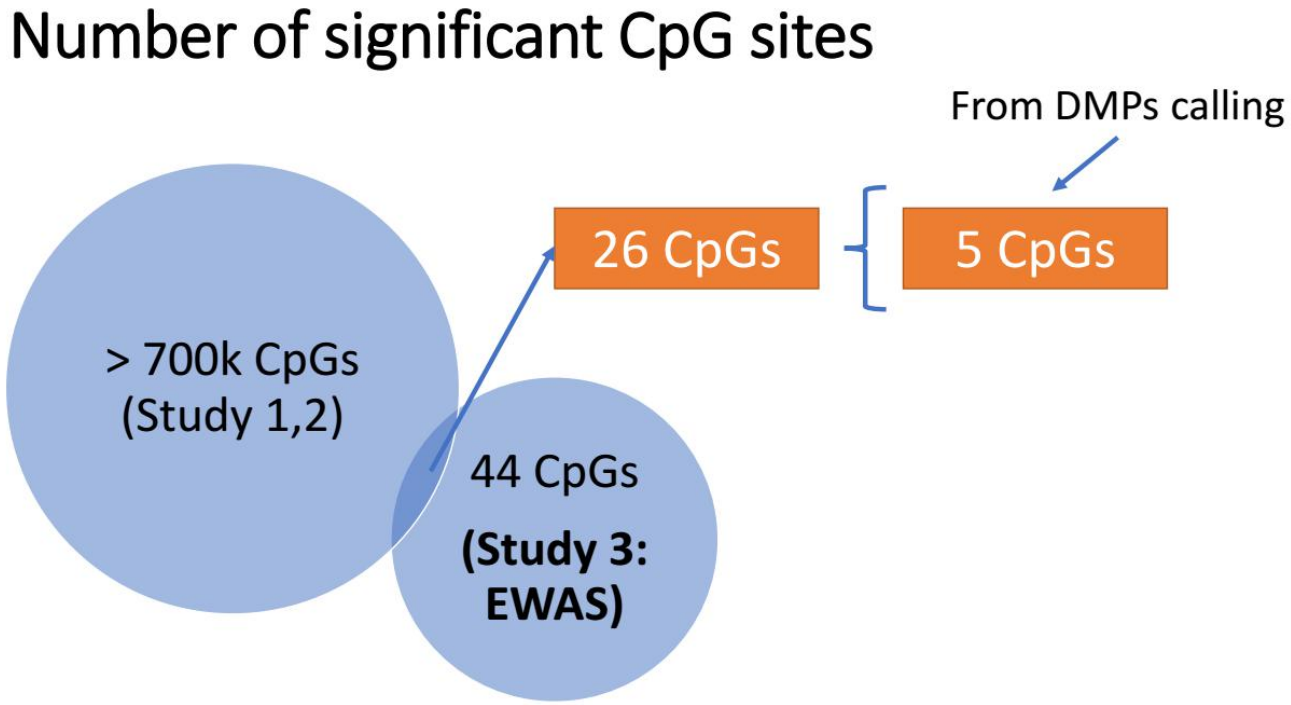

| Package | Version | Link |
|---|---|---|
| minfi | 1.38.0 | http://bioconductor.org/packages/release/bioc/html/minfi.html |
| limma | 3.48.1 | https://bioconductor.org/packages/release/bioc/html/limma.html |
| rtracklayer | 1.52.0 | https://bioconductor.org/packages/release/bioc/html/rtracklayer.html |
| DMRcate | 2.6.0 | https://bioconductor.org/packages/release/bioc/html/DMRcate.html |
| CMDV | | Hachiya, T. et al. (2017) - 10.1038/s41525-017-0016-5 |
| Anotation | | 1. Probes anotation: https://zwdzwd.github.io/InfiniumAnnotation<br>2. Cross reactive probes: https://github.com/sirselim/illumina450k_filtering |



R + bash script

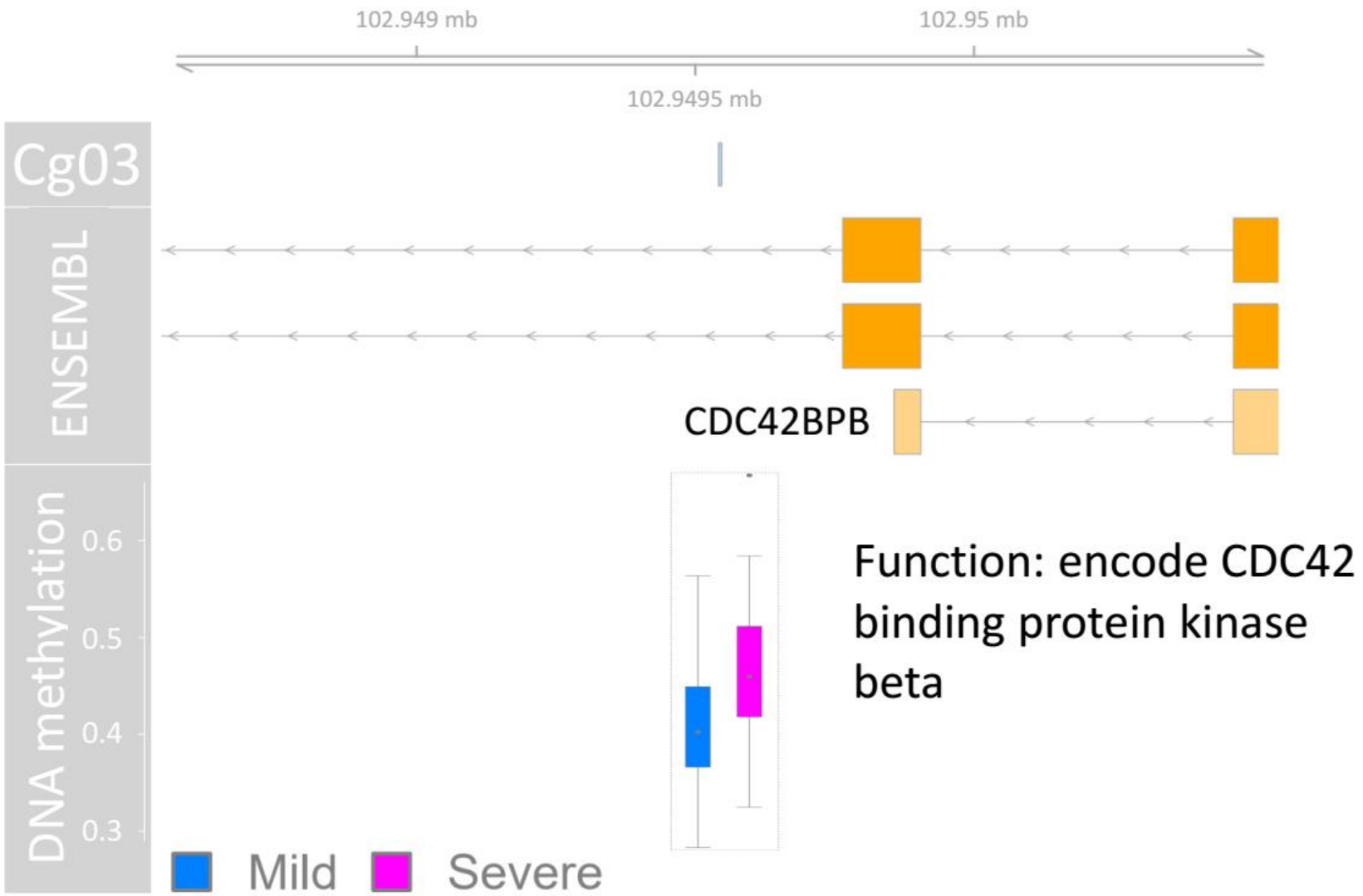
# Example 1: Methylation Risk Score (MRS)

**3. Filter**: remove portions (samples/probes) not relevant to the questions.




R script

# DNA methylation of probe at CDC42BPB




Function: encode CDC42 binding protein kinase beta

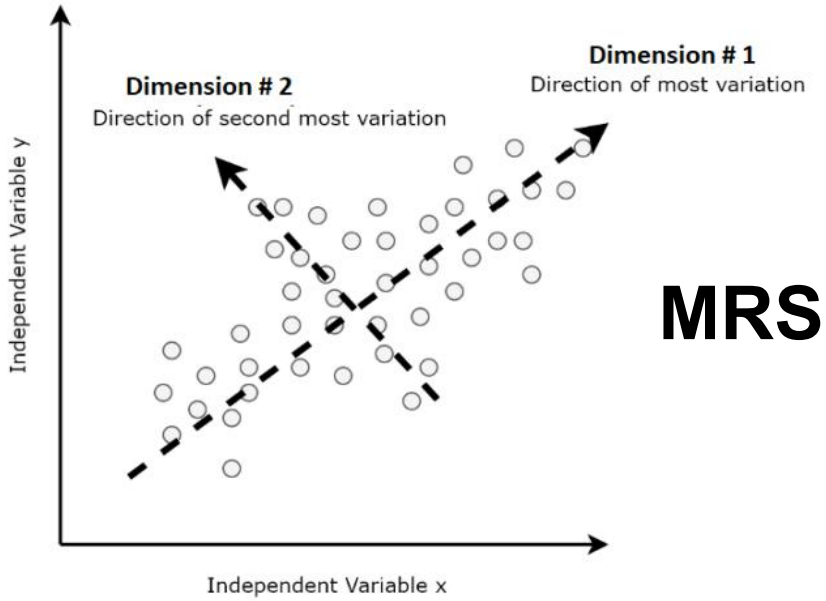
# CpG sites associated with gene functions

| SigGenesInSet | P.DE | Description |
|---|---|---|
| DDO | 0.001576367 | D-Amino acid metabolism |
| DDO | 0.011302778 | Alanine, aspartate and glutamate metabolism |
| CXCR2 | 0.0120096 | Viral protein interaction with cytokine and cytokine receptor |
| SGMS1 | 0.014132182 | Sphingolipid metabolism |
| DDO | 0.020980386 | Peroxisome |
| CXCR2 | 0.021004332 | Epithelial cell signaling in Helicobacter pylori infection |
| SGMS1 | 0.041698417 | Sphingolipid signaling pathway |
| CXCR2 | 0.045847039 | Cytokine-cytokine receptor interaction |
| CXCR2 | 0.052644258 | Chemokine signaling pathway |
| CXCR2 | 0.055835584 | Phospholipase D signaling pathway |
| SGMS1,DDO | 0.056711504 | Metabolic pathways |
| VIM | 0.059180432 | Epstein-Barr virus infection |
| VIM | 0.061226741 | MicroRNAs in cancer |
| CXCR2 | 0.069345635 | Human cytomegalovirus infection |
| CXCR2 | 0.084765607 | Endocytosis |

# Example 1: Methylation Risk Score (MRS)

## **4. Mine**: maths, statistics, and data mining

Dimension reduction




**MRS = PC1**

Methylation Index

$$\mathbf{MRS} = \frac{1}{n}\sum_{c}^{n} W_c \frac{\beta_{cs} - \mu_c}{\sigma_c}$$

R script

# Example 1: Methylation Risk Score (MRS)

**5. Represent**: Basic form that a set of data will take.

- Some data sets are shown as lists, others are structured like trees, and so forth.
- Summary statistics

| 1 | sample_ID | MRS | class | sex | age | race | study |
|---|---|---|---|---|---|---|---|
| 2 | COVID_72 | -0.792014279446939 | non-ICUd | Female | 50 | H | study1 |
| 3 | COVID_84 | -0.57576023419961 | non-ICUd | Female | 75 | W | study1 |
| 4 | COVID_96 | -0.687881386476828 | non-ICUd | Male | 51 | W | study1 |
| 5 | COVID_83 | -0.530784739208204 | ICUd | Female | 85 | W | study1 |
| 6 | COVID_95 | -0.705899758878854 | non-ICUd | Male | 49 | O | study1 |
| 7 | COVID_82 | -0.655856292308603 | non-ICUd | Male | 67 | W | study1 |
| 8 | COVID_94 | -0.880279408934101 | non-ICUd | Female | 24 | W | study1 |

R script

# Example 1: Methylation Risk Score (MRS)

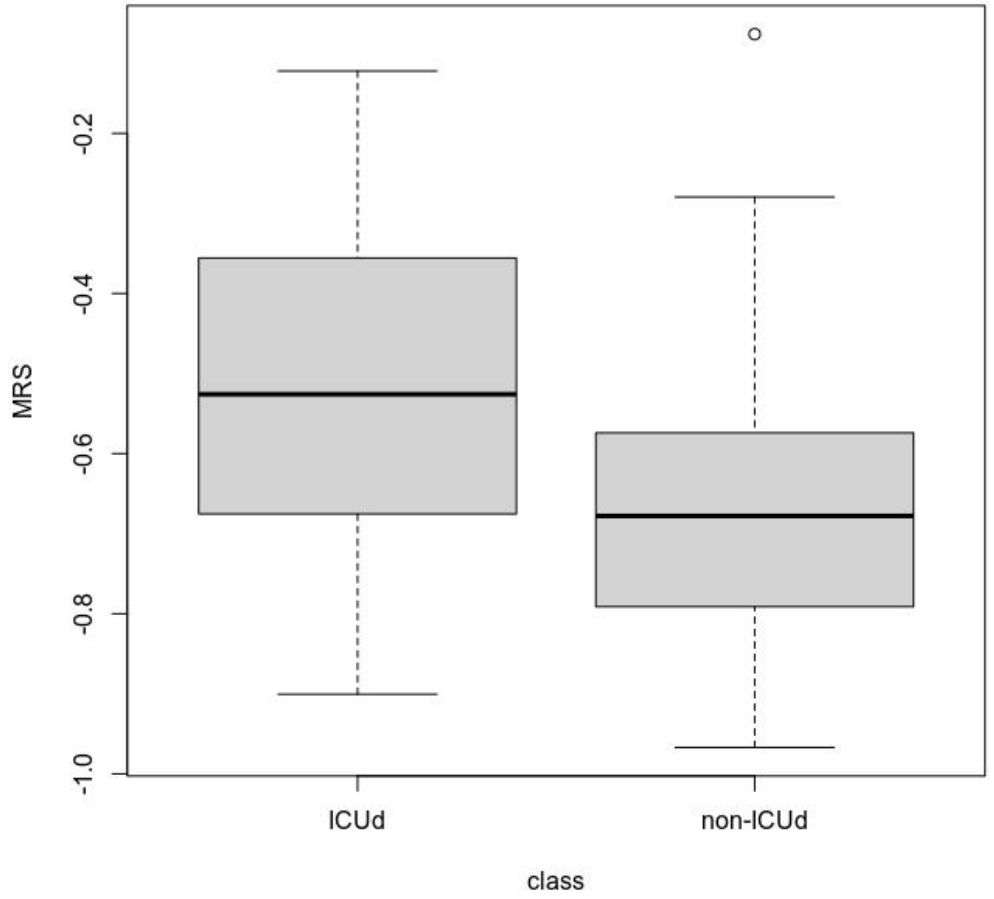
## 5. **Represent**: Summary statistics

```
> p <- '/home/phuluu/Documents/Lecture/Data_visualization/My_Lectures/lecture01/data/Visualization_Cont_Noncount_Data_/ICUd.tsv'
> dt <- read.table(p, header=T)
> head(dt)
  sample_ID        MRS    class    sex age race  study
1  COVID_72 -0.7920143 non-ICUd Female  50    H study1
2  COVID_84 -0.5757602 non-ICUd Female  75    W study1
3  COVID_96 -0.6878814 non-ICUd   Male  51    W study1
4  COVID_83 -0.5307847     ICUd Female  85    W study1
5  COVID_95 -0.7058998 non-ICUd   Male  49    O study1
6  COVID_82 -0.6558563 non-ICUd   Male  67    W study1
> summary(dt)
  sample_ID             MRS              class              sex
 Length:277         Min.   :-0.96706   Length:277         Length:277
 Class :character   1st Qu.:-0.74829   Class :character   Class :character
 Mode  :character   Median :-0.64055   Mode  :character   Mode  :character
                    Mean   :-0.61011
                    3rd Qu.:-0.47905
                    Max.   :-0.07605
      age            race              study
 Min.   :18.00   Length:277         Length:277
 1st Qu.:43.00   Class :character   Class :character
 Median :55.00   Mode  :character   Mode  :character
 Mean   :55.25
 3rd Qu.:70.00
 Max.   :94.00
```

R script

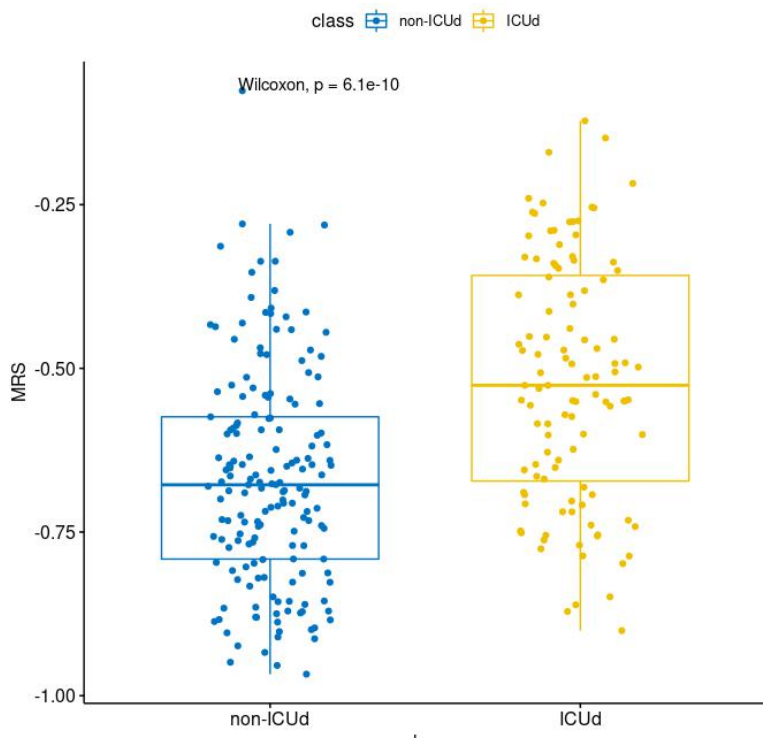
# Example 1: Methylation Risk Score (MRS)

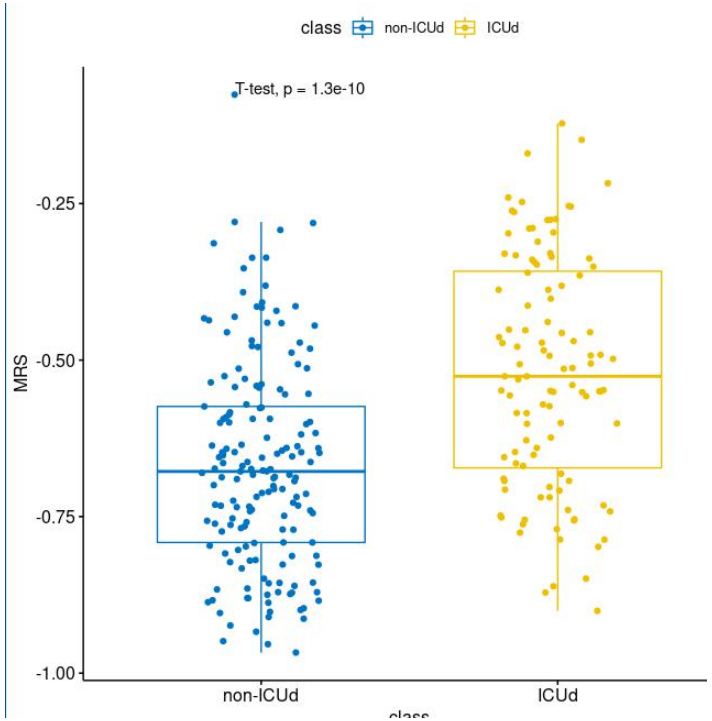
## **5. Represent**: Basic plot




boxplot(MRS ~ class, data=dt) and ttest (Giang)?

# Example 1: Methylation Risk Score (MRS)




boxplot(MRS ~ class, data=dt) and ttest (Giang)?

# Example 1: Methylation Risk Score (MRS)

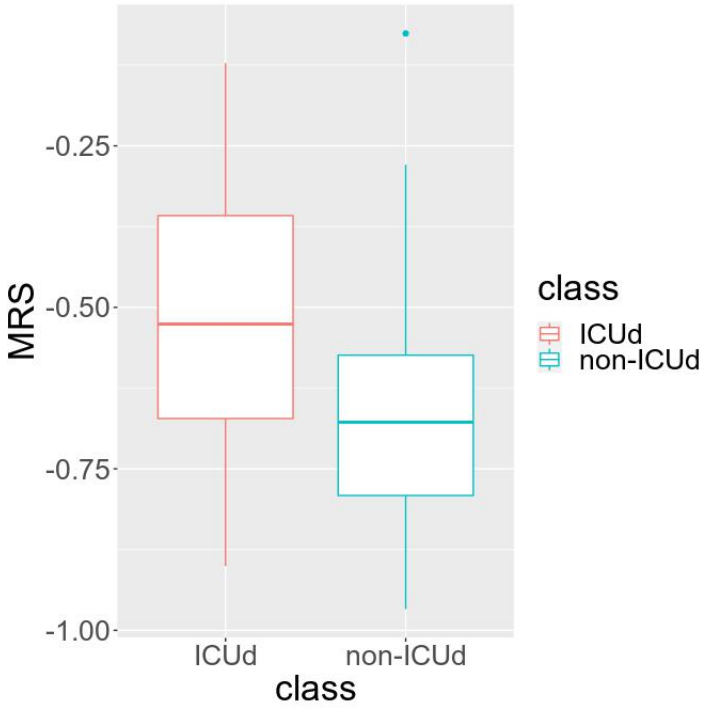
**6. Refine**: Graphic design methods are used to further clarify the representation by

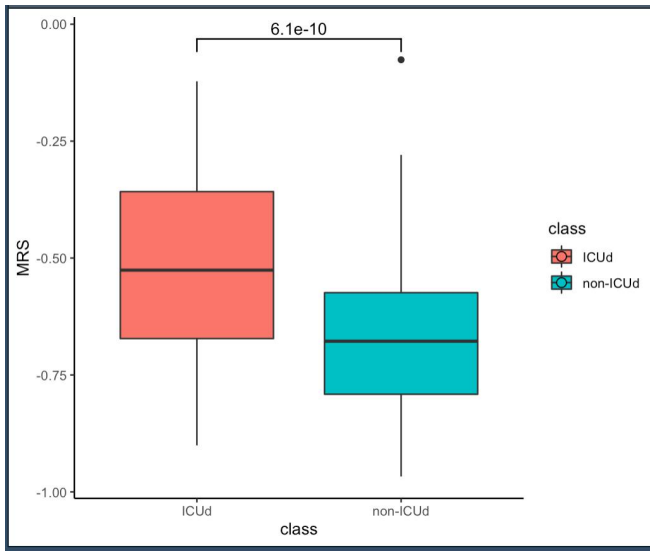
- calling more attention to particular data (establishing hierarchy)
- changing attributes (such as color) that contribute to readability

R script

# Example 1: Methylation Risk Score (MRS)

**6. Refine**: Graphic design methods are used to further clarify the representation



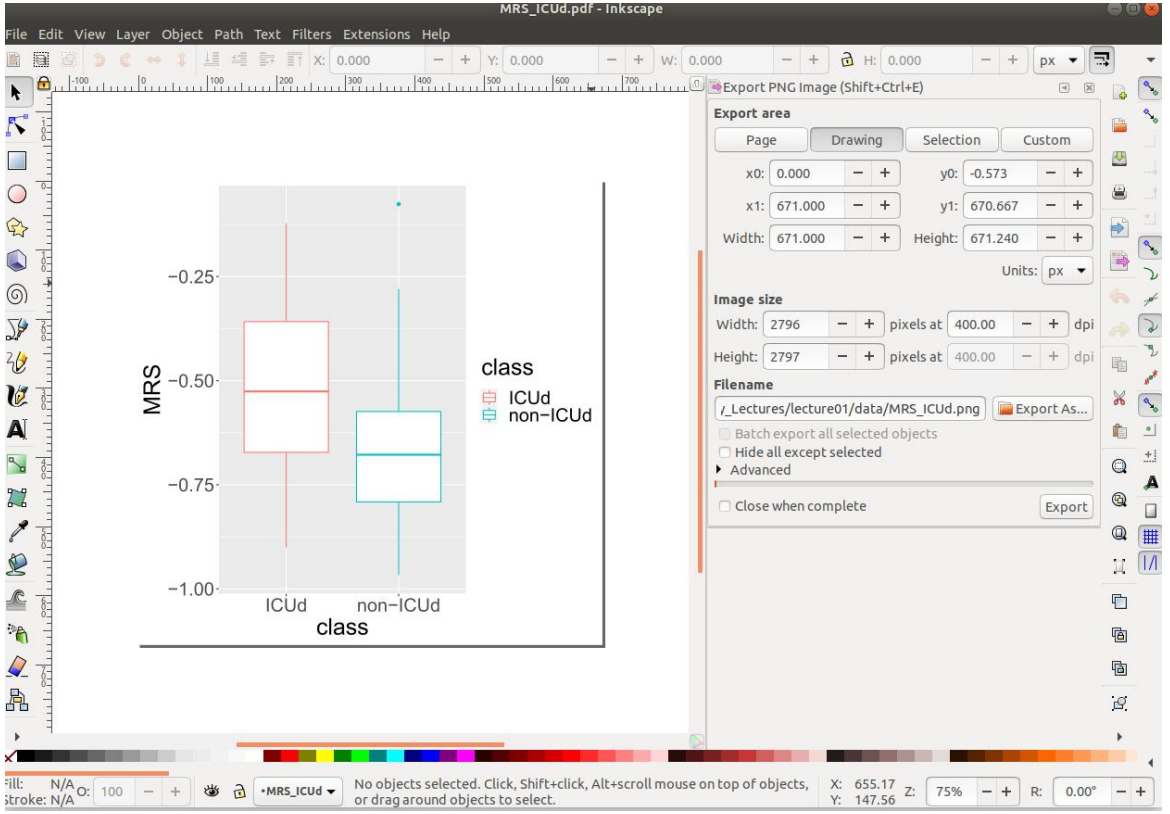




R script

```
library(ggplot2)
ggplot(data=dt, aes(x=class, y=MRS)) +
geom_boxplot(aes(color=class)) +
theme(text = element_text(size=25))
ggsave("MRS_ICUd.pdf")
```

# Example 1: Methylation Risk Score (MRS)

**6. Refine**: Graphic design methods are used to further clarify the representation

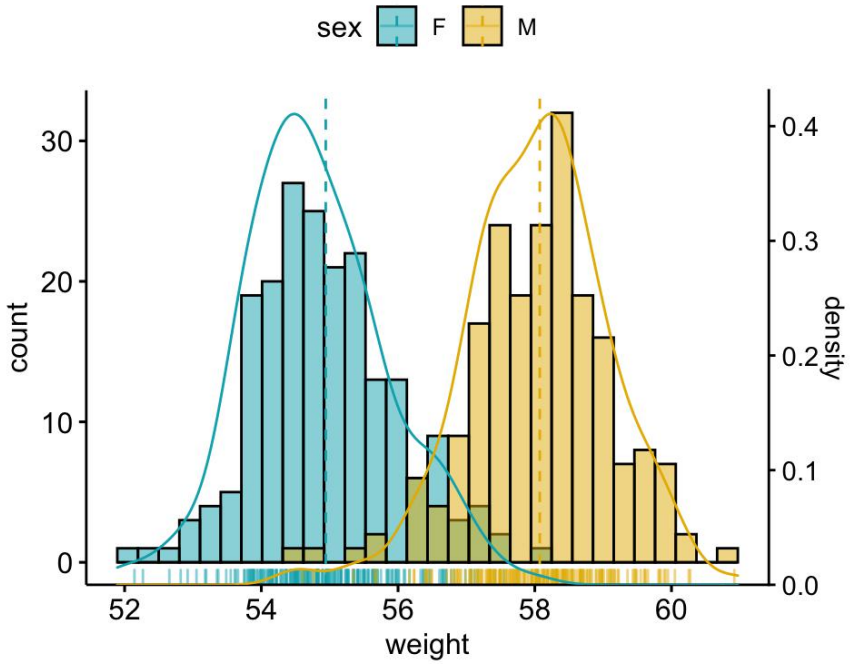


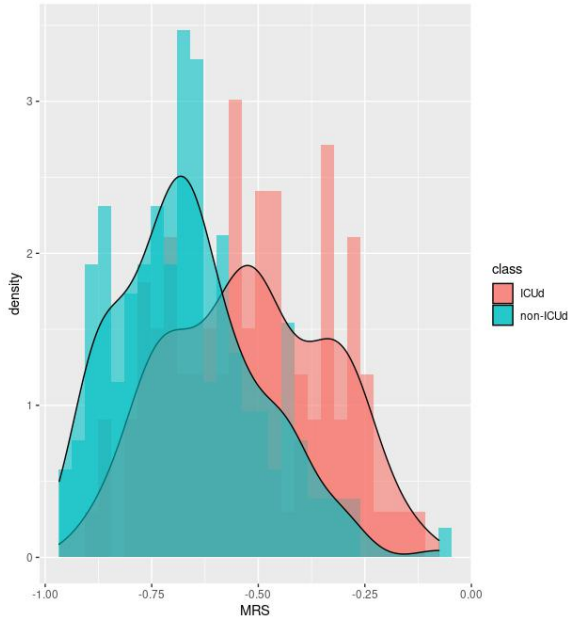

Inkscape

# Example 1: Methylation Risk Score (MRS)

**6. Refine**: Graphic design methods are used to further clarify the representation

How to plot like this graph, but content same as boxplot? (Thong Nho)



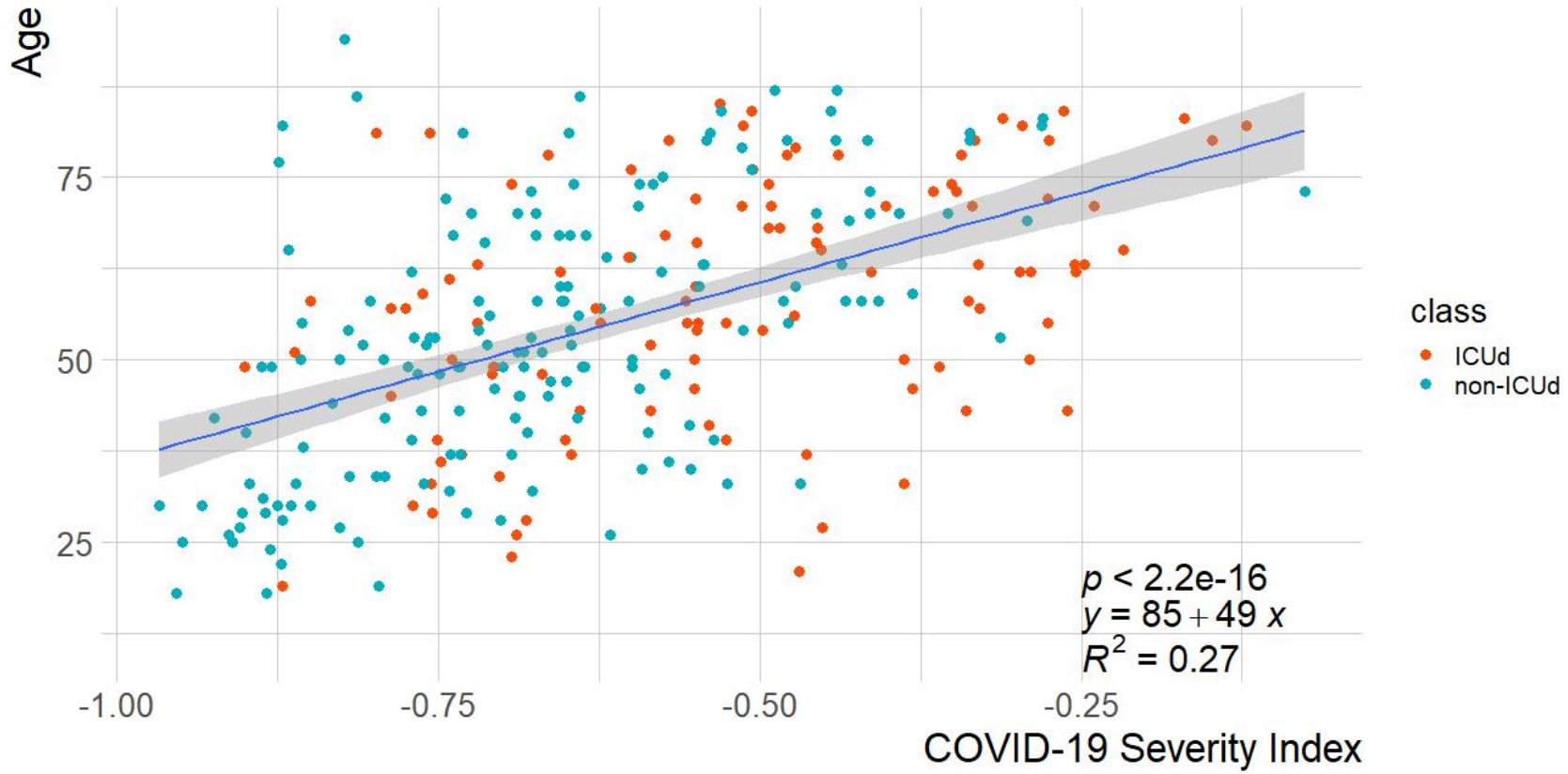

https://www.datanovia.com/en/blog/ggplot-histogram-with-density-curve-in-r-using-secondary-y-axis/

# Example 1: Methylation Risk Score (MRS)

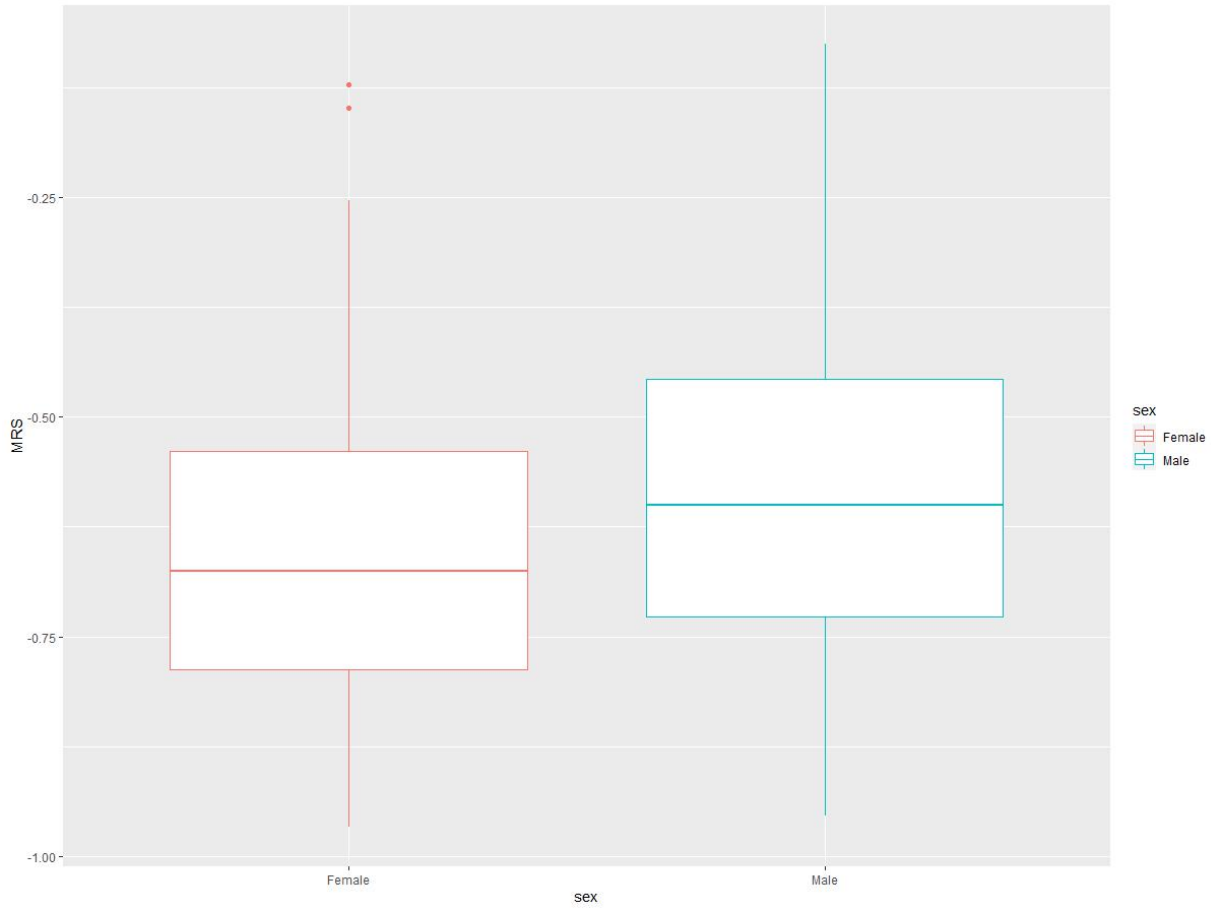
**Class tasks:**

1. MRS vs age: plot and regression (Thien)
2. MRS vs sex: plot and ttest (Hoang)
3. MRS vs race: plot and ANOVA (Nhuong)
4. MRS vs study: plot and ANOVA (Minh)
5. MRS vs age, race, sex and study: plot and regression (Giang)

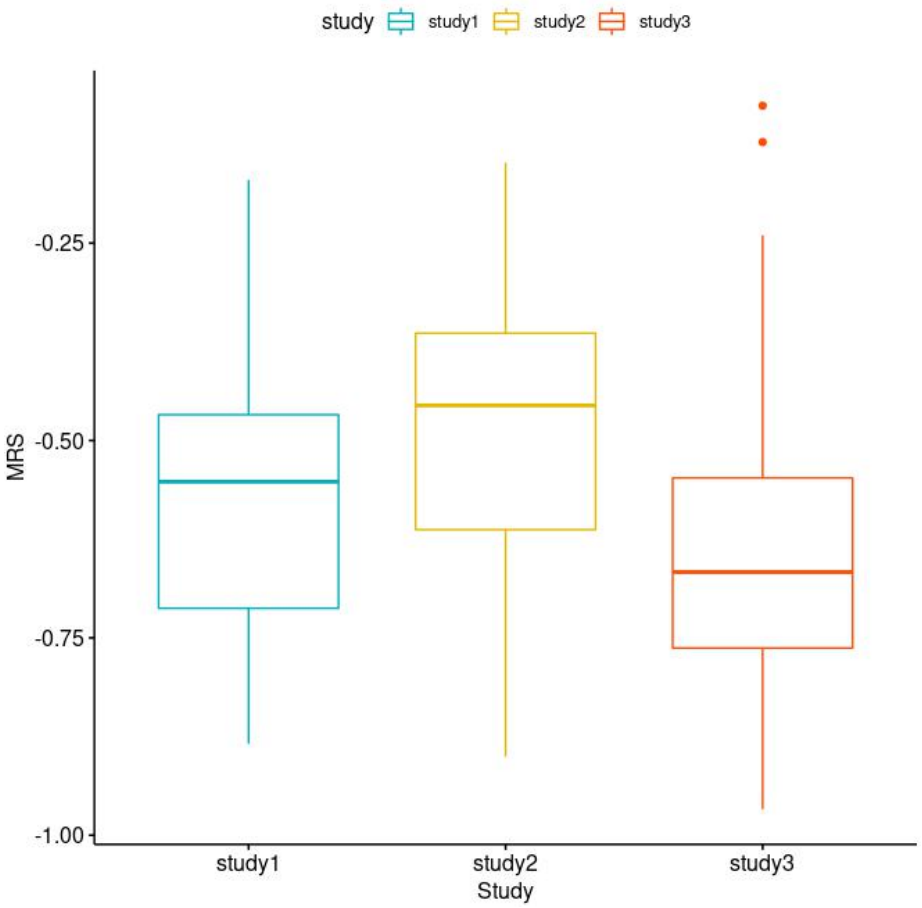
# MRS vs age: plot and regression (Thien)



# MRS vs sex: plot and ttest (Hoang)

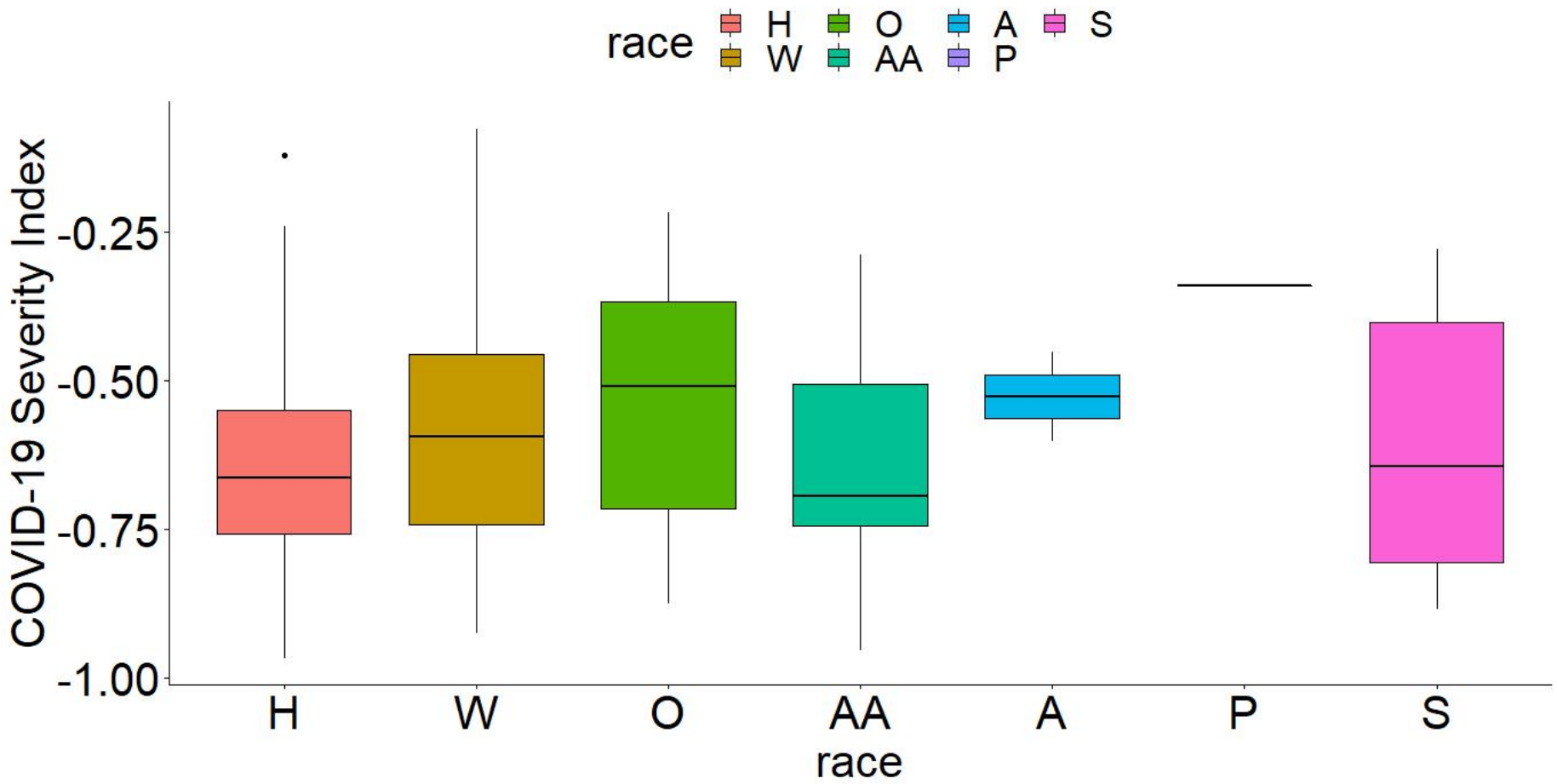


# MRS vs study: plot and ANOVA (Minh)




| study | count | mean | sd |
|---|---|---|---|
| <fct> | <int> | <dbl> | <dbl> |
| study1 | 100 | -0.571 | 0.181 |
| study2 | 15 | -0.518 | 0.226 |
| study3 | 162 | -0.643 | 0.180 |

# MRS vs race: plot and ANOVA (Nhuong)



# Example: Methylation Risk Score (MRS)

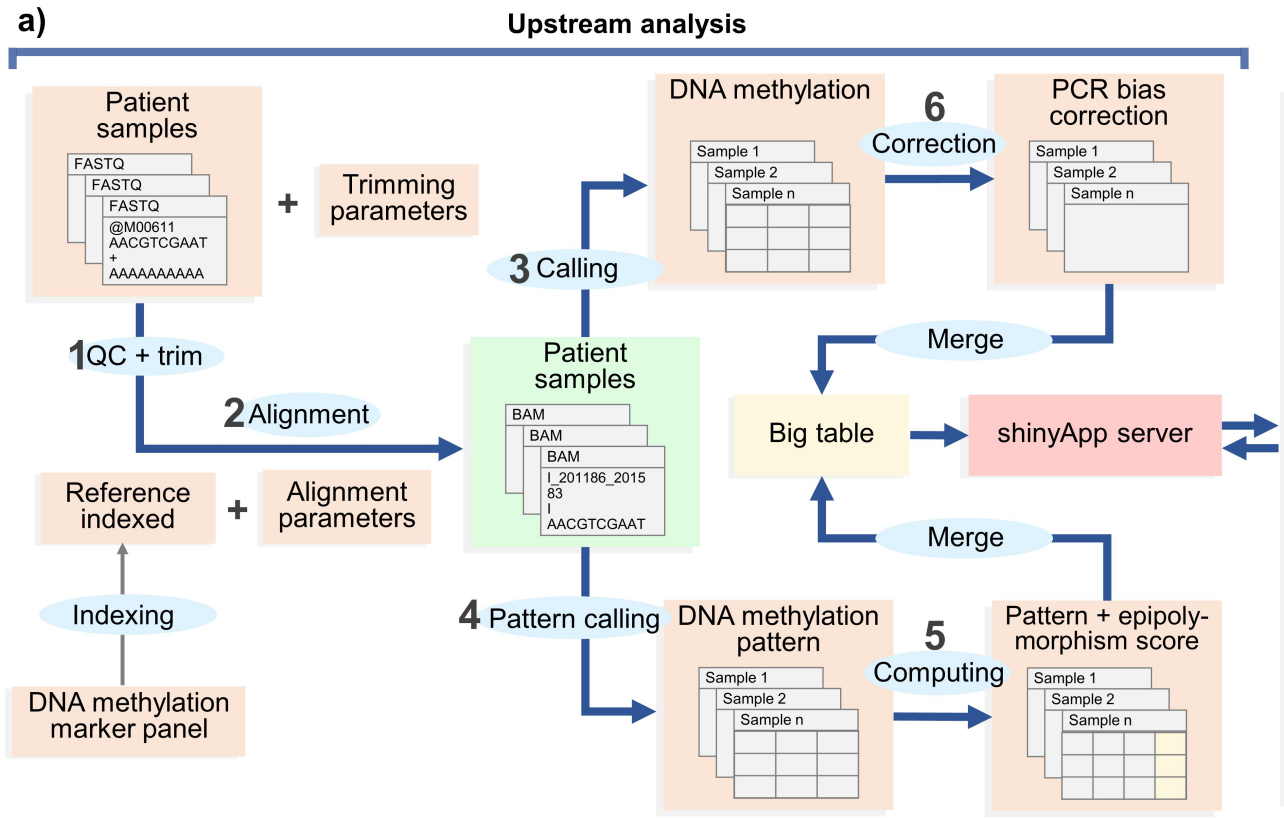
**7. Interact**: adds interaction, letting the user control or explore the data. Interaction might cover things like selecting a subset of the data or changing the viewpoint.

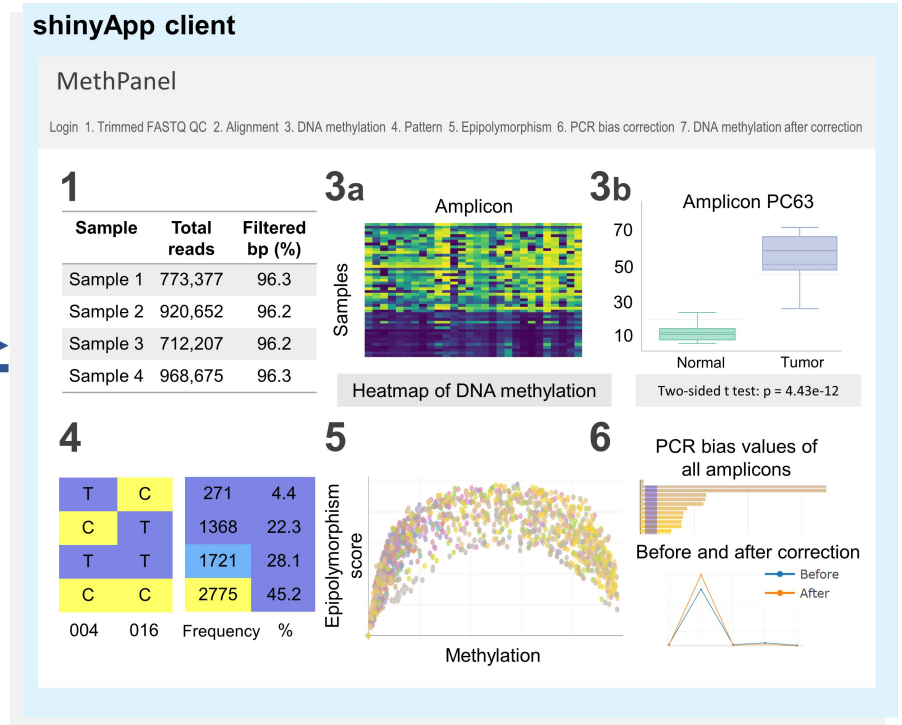
R script with loon package, plotly and ShinyApp

# Example 2: MethPanel

http://129.94.136.70/shiny/sample-apps/MethPanel/

Luu, P. L., Ong, P. T., Loc, T. T. H., Lam, D., Pidsley, R., Stirzaker, C., & Clark, S. J. (2021). MethPanel: a parallel pipeline and interactive analysis tool for multiplex bisulphite PCR sequencing to assess DNA methylation biomarker panels for disease detection. Bioinformatics, 37(15), 2198-2200.




https://github.com/thinhong/MethPanel

# Homework: Methylation Risk Score (MRS)

What is the question?

Developing Methylation Risk Score for stratification of COVID-19 serverity

COVID-19 serverity = COVID-GRAM percentage >= 50%

COVID-19 Mild = COVID-GRAM percentage < 50%

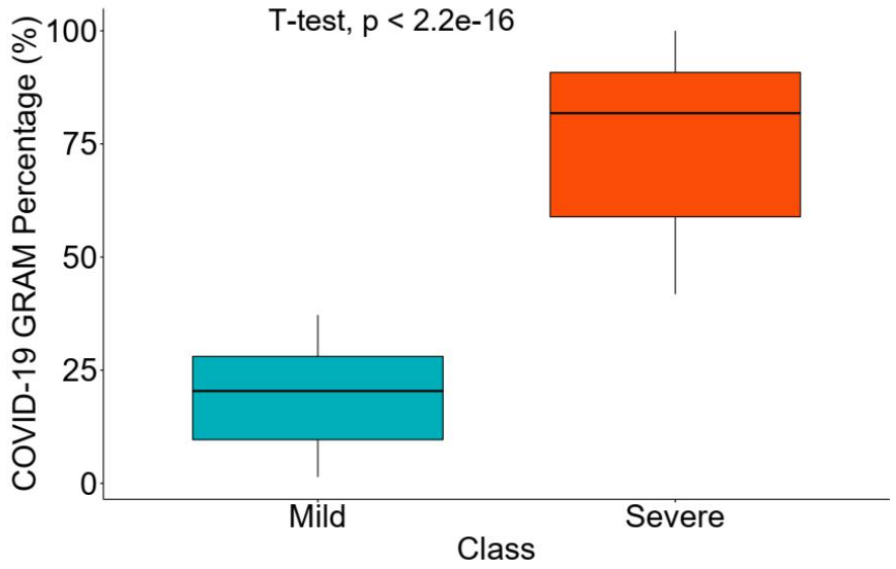
## What is COVID-19 GRAM Percentage?

**Including 10 clinical factors:**

1. Chest radiography abnormality
2. Age
3. Hemoptysis
4. Dyspnea
5. Unconsciousness
6. Number of comorbidities
7. Cancer history
8. Neutrophil-to-lymphocyte ratio
9. Lactate dehydrogenase
10. Direct bilirubin

**Mild/Severe**

- Severe: ≥ 50%
- Mild: < 50%



# Homework: Methylation Risk Score (MRS)

What is the question?

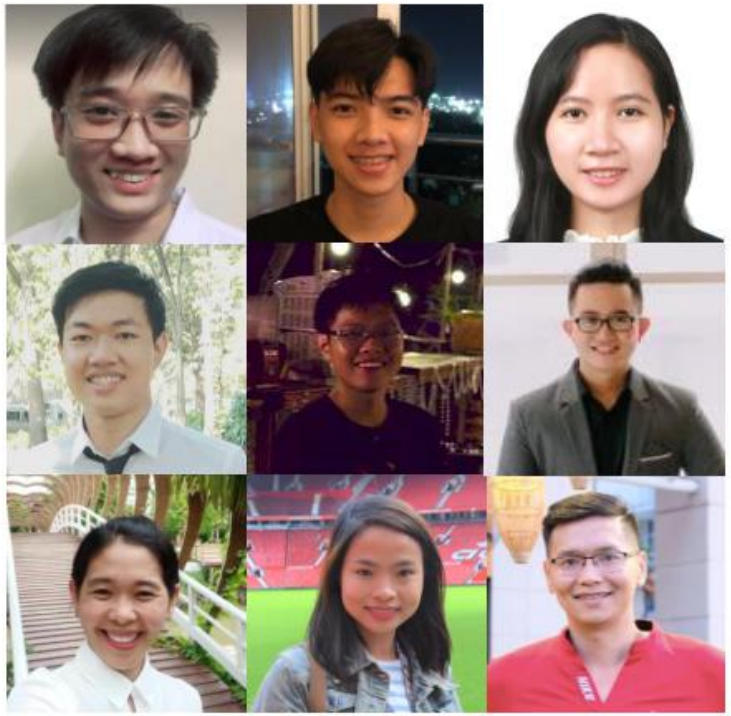
Developing Methylation Risk Score for stratification of COVID-19 serverity

COVID-19 serverity = COVID-GRAM percentage >= 50%

COVID-19 Mild = COVID-GRAM percentage < 50%

| 1 | sample_ID | gram | MRS | sex | age | race | study |
|---|---|---|---|---|---|---|---|
| 2 | COVID_96 | Mild | 0.342063866786641 | Male | 51 | W | study1 |
| 3 | COVID_95 | Mild | 0.535144608277981 | Male | 49 | O | study1 |
| 4 | COVID_82 | Mild | 0.386215415209122 | Male | 67 | W | study1 |
| 5 | COVID_94 | Mild | 0.080487319595148 | Female | 24 | W | study1 |
| 6 | COVID_73 | Severe | 0.193577864244823 | Male | 82 | W | study1 |
| 7 | COVID_85 | Severe | 0.71445633374692 | Male | 62 | H | study1 |
| 8 | COVID_97 | Severe | 0.684757326085255 | Male | 76 | W | study1 |

# Thank you The Team for the data generation!



Group members:

- Nhat-Thong Le
- Ba-Thien Tran
- My-Phung Duong, M.S
- Kim-Sang To, M.D
- Hung P. Le, M.D
- Phuc-Thinh Ong, M.D
- Thi-Tai-Nguyen Cao, Ph.D
- Thi-Thanh-Khuong Tran, Ph.D
- Phuc-Loi Luu, Ph.D

Clinical advice:

- Luong Thi My Hanh, M.D
- Members of VnPathoinformatics group

***All patients for their consent to donate blood samples and all researcher in the studies.***