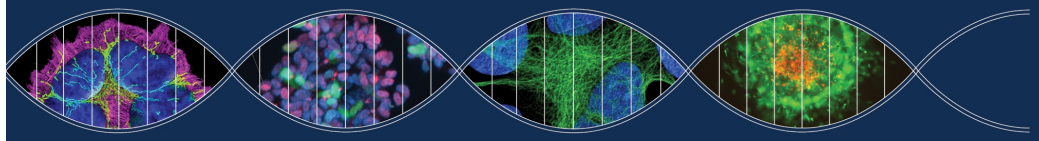
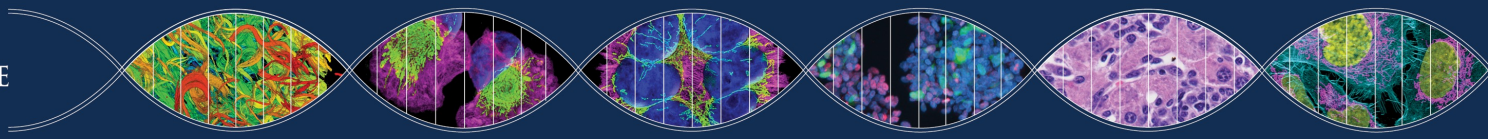
Bioinformatics @ Wistar



THE WISTAR INSTITUTE



NCI Cancer Center

A Cancer Center Designated by the National Cancer Institute

Wistar Institute Bioinformatics Team

Jozef Madzo
Co-director

Andrew Kossenkov
Co-director

Jayamanna (Priya) Wickramasinghe
(Associate) Managing Director

Bhanu Karisetty
Bioinformatic Analyst

Ying (Julia) Ye
Bioinformatic Analyst

Bioinformatics pipelines:

Gene Regulation: *Transcriptomics / Epigenetics*

- RNA-seq (expression)
- ChIP-seq / CUT&RUN (histone marks / TF)
- ATAC-seq (chromatin accessibility)
- Hi-C
- RRBS, WGB

Genomics: *Variant detection*

- Exome sequencing
- Low-pass WGS
- RNA-editing

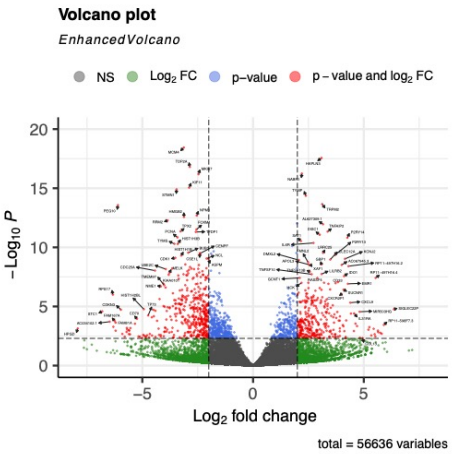
Single-Cell experiment

- scRNA-seq
- scATAC-seq
- Single nuclei RNA-seq (snucRNA-seq)

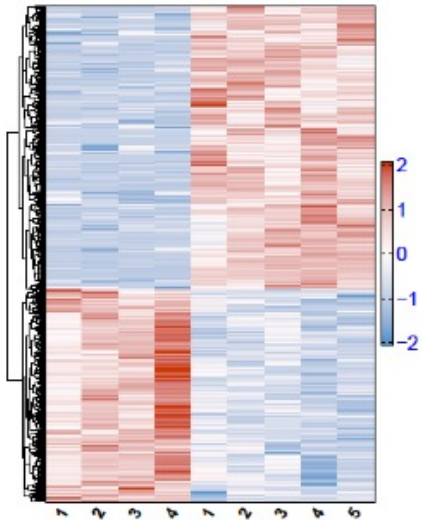
Long read sequencing

- PacBio HiFi sequencing
- ONT

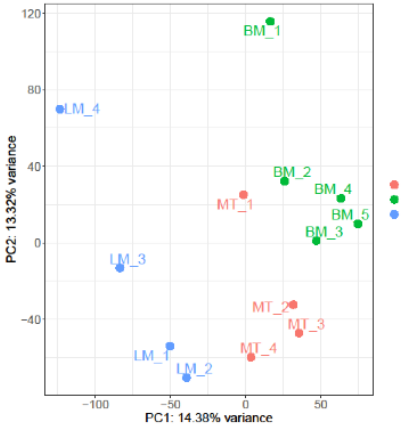
Example of RNA-seq analysis



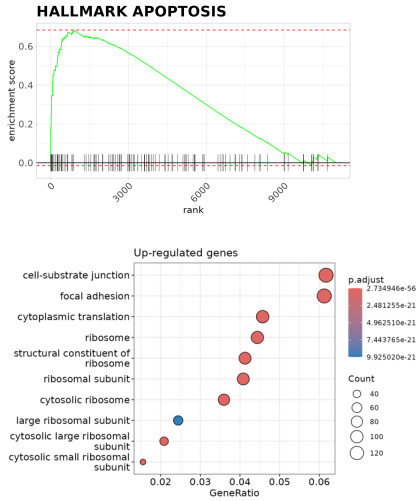
Volcano plot
Detection of differentially expressed genes between conditions



Heatmap plot
Hierarchical clustering of samples based on gene expression



PCA plot
Visualization of relative differences in samples in reduced dimensions (2D)



Pathway analysis
Looking for a significantly enriched pathway with GSEA and/or Gene Ontology

Calendar

Tue – theoretical lecture

Thr – practical / coding exercise

Week	Day	Lecture Title
1	6-Jan	Orientation and Basics of R
	8-Jan	Orientation and Basics of R
2	13-Jan	Basic R programming
	15-Jan	Basic R programming practice section
3	20-Jan	Statistics I: Scientific Foundations
	22-Jan	Data wrangling and data visualization
4	27-Jan	Statistics II: Practical application
	29-Jan	Statistic test practice section
5	3-Feb	Theoretical basis for sequencing (Sonali Majumdar)
	5-Feb	Midterm exam review
6	10-Feb	Gene Regulation (Alessandro Gardini)
	12-Feb	Introduction to RNA-seq
7	17-Feb	Epigenetics & Gene Regulation (TBD)
	19-Feb	RNA-seq part II
8	24-Feb	Noncoding RNA and Repetitive Elements (Simon Chu)
	26-Feb	Introduction to ChIP-seq
9	3-Mar	Single Cell sequencing (Avi Srivastava)
	5-Mar	ChIP-seq part II
10	10-Mar	Spatial Transcriptomics (Andrew Kossenkov)
	12-Mar	No Class- Spring Break
11	17-Mar	TBD
	19-Mar	Intro into DNA methylation Analysis
12	24-Mar	RNA modification (Bin Tian)
	26-Mar	DNA methylation Analysis part II
13	31-Mar	New Emerging Technologies
	2-Apr	Single cell sequencing and RNA-seq Analysis
14	7-Apr	Leveraging public online tools practice section
	9-Apr	Public tools practice section
15	14-Apr	Review of practical skills
	16-Apr	Final exam review
	21-Apr	Final class and evaluations

How this is going to work

- Objective: learn **R** / **Linux** and independent work with RNAseq data
- Have to attend the lecture (Thursday practical lecture)
- Office hours are Thursday 3.30pm - 4.45pm (Stay on zoom)
- Optional if you need help with something; can also get help through email or Slack
- If there will be work assigned on Thursday, that will be due next week

Software

R

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Get Involved: Contributing
Developer Pages
R Blog

R Foundation

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 4.4.3 (Trophy Case)** has been released on 2025-02-28.
- The **useR! 2025** conference will take place at Duke University, in Durham, NC, USA, August 8-10.
- **R version 4.4.2 (Pile of Leaves)** has been released on 2024-10-31.
- We are deeply sorry to announce that our friend and colleague Friedrich (Fritz) Leisch has died. Read our tribute to Fritz here.
- **R version 4.3.3 (Angel Food Cake)** (wrap-up of 4.3.x) was released on 2024-02-29.
- You can support the R Foundation with a renewable subscription as a supporting member.

RStudio

posit PRODUCTS OPEN SOURCE USE CASES PARTNERS LEARN & SUPPORT ABOUT

DOWNLOAD

RStudio Desktop

Used by millions of people weekly, the RStudio integrated development environment (IDE) is a set of tools built to help you be more productive with R and Python.

Don't want to download or install anything? Get started with RStudio on Posit Cloud for free. If you're a professional data scientist looking to download RStudio and also need common enterprise features, don't hesitate to book a call with us.

Want to learn about core or advanced workflows in RStudio? Explore the RStudio User Guide or the Getting Started section.

1: Install R

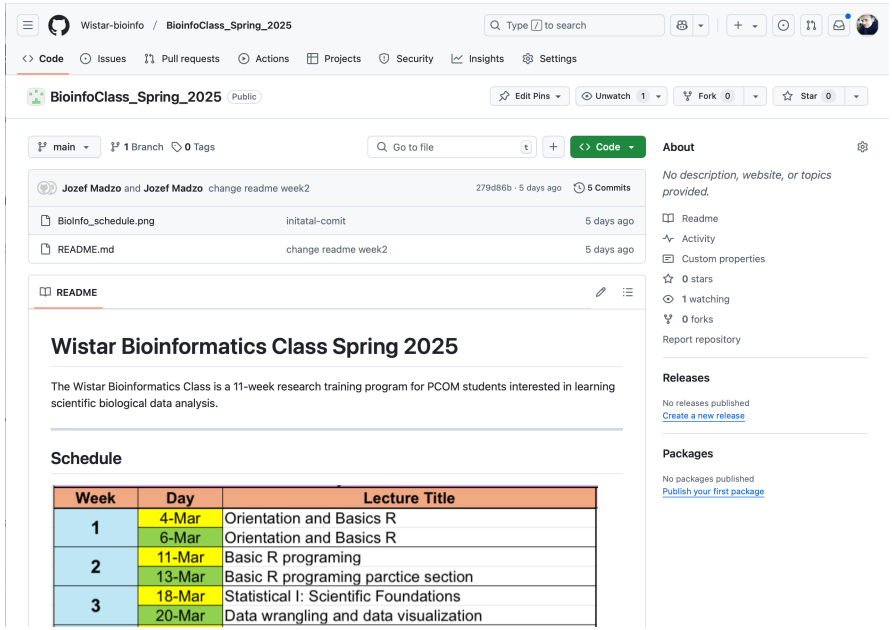
RStudio requires R 3.6.0+. Choose a version of R that matches your computer's operating system.

2: Install RStudio

DOWNLOAD RSTUDIO DESKTOP FOR MACOS 13+

Resources

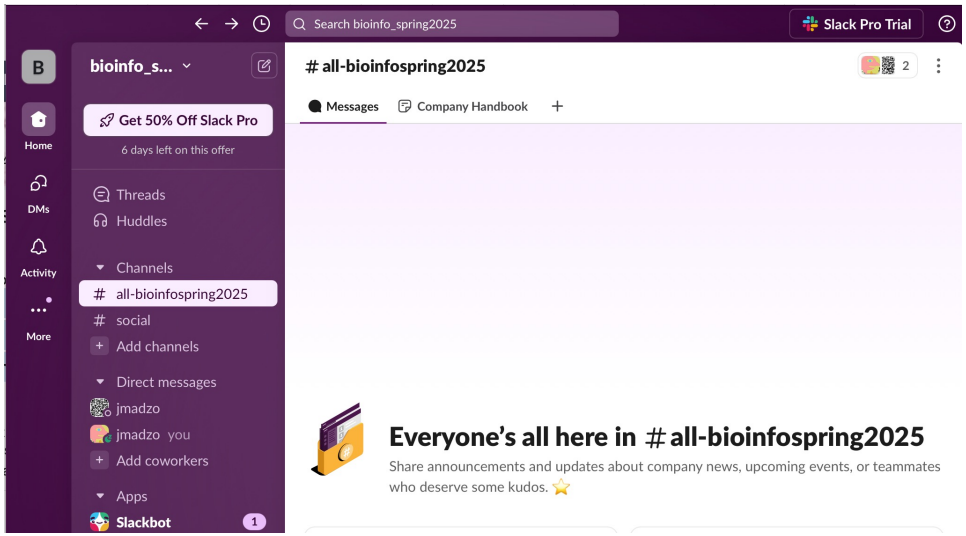
Github



Week	Day	Lecture Title
1	4-Mar	Orientation and Basics R
	6-Mar	Orientation and Basics R
2	11-Mar	Basic R programing
	13-Mar	Basic R programing parctice section
3	18-Mar	Statistical I: Scientific Foundations
	20-Mar	Data wrangling and data visualization

https://github.com/Wistar-bioinfo/BioinfoClass_Spring_2025

Slack



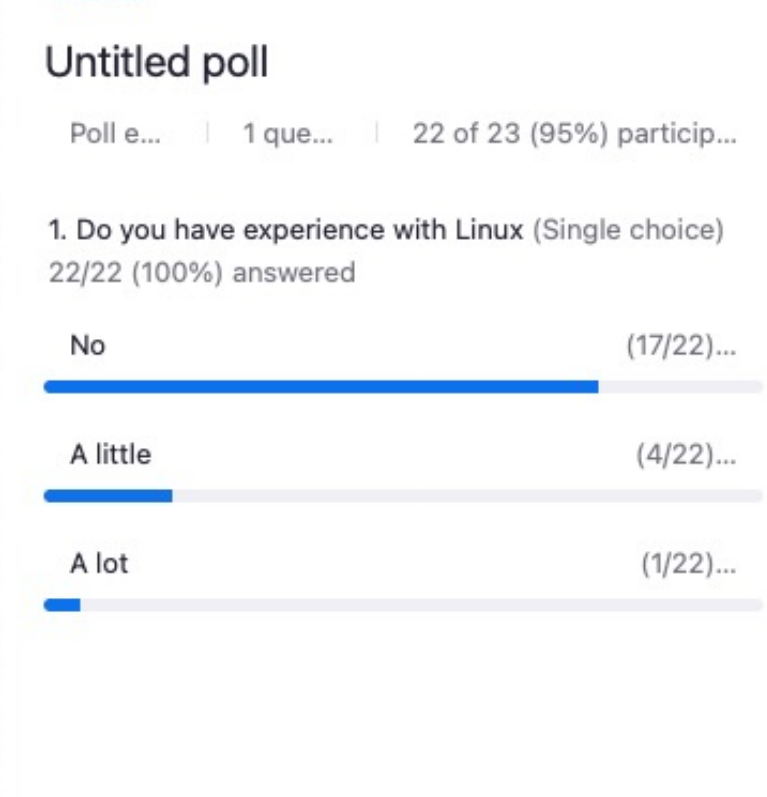
https://join.slack.com/t/bioinfospring2025/shared_invite/zt-31aqz3j8g-Ha7eNObbwzGDP1U9Wr7wRg

You can also email me at jmadzo@wistar.org

Poll

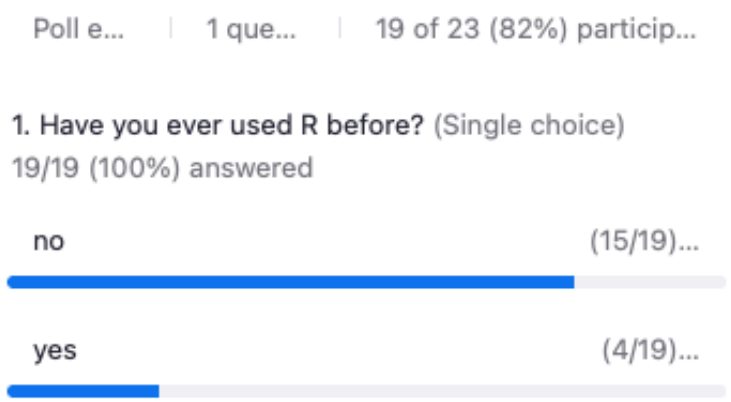
Poll

Do you have experience with Linux?



Poll

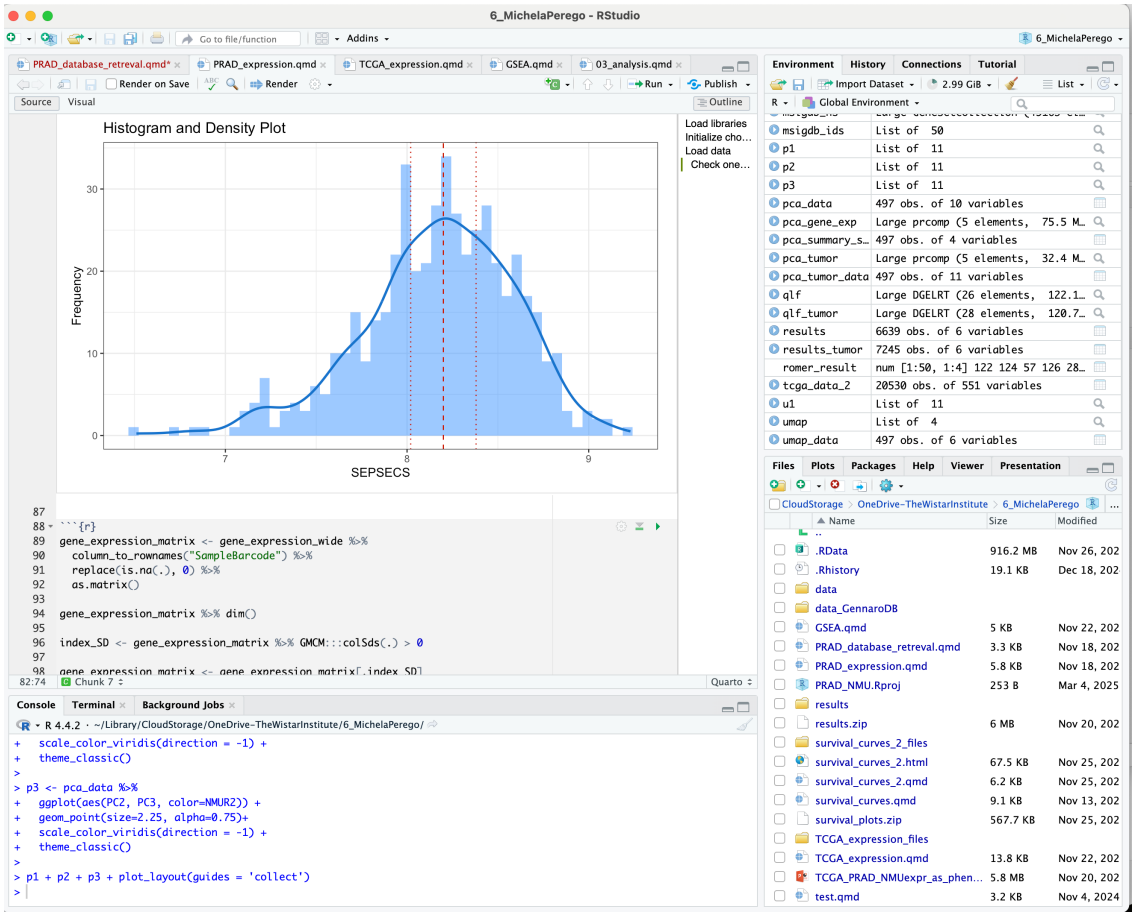
Do you have experience with R?



R Programming Language

- Open-source scripting language developed for statistical analysis
- Was originally known as S
- S was developed in 1975
- S was reimplemented as R in 1993
- Heavily used in bioinformatics, with our own archive Bioconductor, for biology-related packages

RStudio environment



R Studio®

The Tidyverse and RStudio environment

- “The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.”
- Keep everything simple
- Use existing data structures instead of custom, aka, use tidy data
- Functions should do one thing well
- Glue the simple things together; simple things put together are more powerful than one complex thing
- Design for humans

What is markdown (and Rmarkdown)?

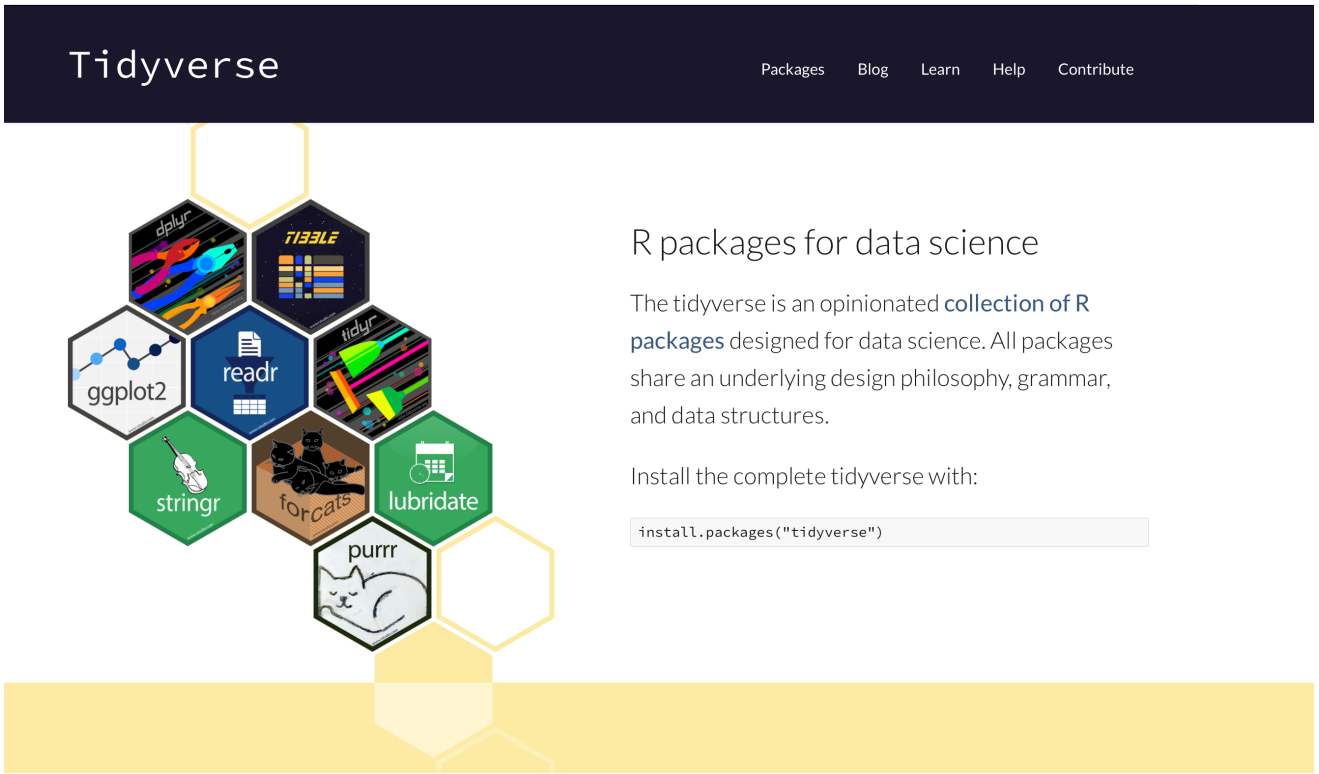
- Markdown is just text, with a few optional symbols that allows a markdown interpreter to make it look good. Goal is to have something that still is human readable even without the interpreter.
- Rmarkdown is markdown for R
- All the features of markdown, with extras
- Intention is to make documenting data analysis easy
- Execute code in Rmarkdown files (cannot do this in markdown)
- Can also knit Rmarkdown files into other files
- html, pdf, or Microsoft word reports
- Can make websites and slides with Rmarkdown

Why use markdown and Rmarkdown?

- Bioinformatics is mostly on a Linux machine using the command line terminal. The rest of the universe uses Macs or PCs.
- This causes a bunch of problems
- operating systems don't talk to each other easily
- files are in proprietary formats (no Microsoft anything on Linux)
- files are not readable in plain text (and terminal needs them to be!)
- Markdown solves all the problems
- Simple
- readable by every machine in both GUI and terminal AND humans
- allows some simple formatting to increase human readability

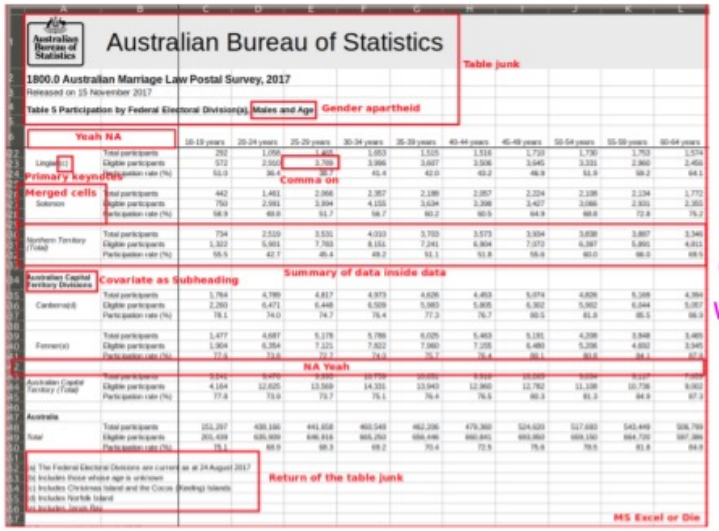
Demo

Demo Introduction to Data Wrangling with `dplyr`



Tidy Data

untidy data



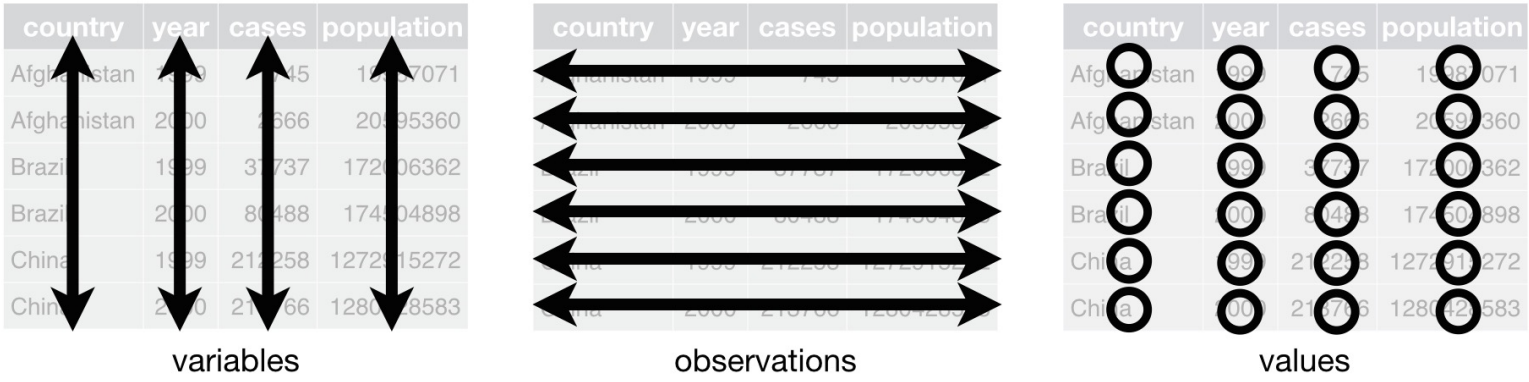
data wrangling

tidy data

https://medium.com/@miles.mcbain/tidying-the-australian-same-sex-marriage-postal-survey-data-with-r-5d35cea07962

https://jhudatascience.org/tidyversecourse/wrangle-data.html

Tidy Data



1. Each variable is in a column.
2. Each observation is a row.
3. Each value is a cell.

Data Wrangling

data wrangling – organizing your data into the form you want

• Everyone spends most of their time on wrangling ! It's hard!

• The tidyverse makes it much easier though; that's it's primary purpose

Tidy Data

wide data

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Height_inches	Weight_lbs	Insulin	Glucose
2	1004	Smith	Jane	65	180	0.60	163
3	4587	Nayef	Mohammed	75	215	1.46	150
4	1727	Doe	Janice	62	124	0.72	177
5	6879	Jordan	Alex	77	160	1.23	205

Tidy Data

long data

	A	B	C
1	**ID**	**Variable**	**Value**
2	1004	LastName	Smith
3	4587	LastName	Nayef
4	1727	LastName	Doe
5	6879	LastName	Jordan
6	1004	FirstName	Jane
7	4587	FirstName	Mohammed
8	1727	FirstName	Janice
9	6879	FirstName	Alex
10	1004	Height_inches	65
11	4587	Height_inches	75
12	1727	Height_inches	62
13	6879	Height_inches	77
14	1004	Weight_lbs	180
15	4587	Weight_lbs	215
16	1727	Weight_lbs	124
17	6879	Weight_lbs	160
18	1004	Insulin	0.60
19	4587	Insulin	1.46
20	1727	Insulin	0.72
21	6879	Insulin	1.23
22	1004	Glucose	163
23	4587	Glucose	150
24	1727	Glucose	177
25	6879	Glucose	205

Tidy Data

long data

	A	B	C
1	**ID**	**Variable**	**Value**
2	1004	LastName	Smith
3	4587	LastName	Nayef
4	1727	LastName	Doe
5	6879	LastName	Jordan
6	1004	FirstName	Jane
7	4587	FirstName	Mohammed
8	1727	FirstName	Janice
9	6879	FirstName	Alex
10	1004	Height_inches	65
11	4587	Height_inches	75
12	1727	Height_inches	62
13	6879	Height_inches	77
14	1004	Weight_lbs	180
15	4587	Weight_lbs	215
16	1727	Weight_lbs	124
17	6879	Weight_lbs	160
18	1004	Insulin	0.60
19	4587	Insulin	1.46
20	1727	Insulin	0.72
21	6879	Insulin	1.23
22	1004	Glucose	163
23	4587	Glucose	150
24	1727	Glucose	177
25	6879	Glucose	205

Tidy Data

reshaping data

wide data

	A	B	C	D	E	F	G
1	ID	LastName	FirstName	Height_inches	Weight_lbs	Insulin	Glucose
2	1004	Smith	Jane	65	180	0.60	163
3	4587	Nayef	Mohammed	75	215	1.46	150
4	1727	Doe	Janice	62	124	0.72	177
5	6879	Jordan	Alex	77	160	1.23	205

long data

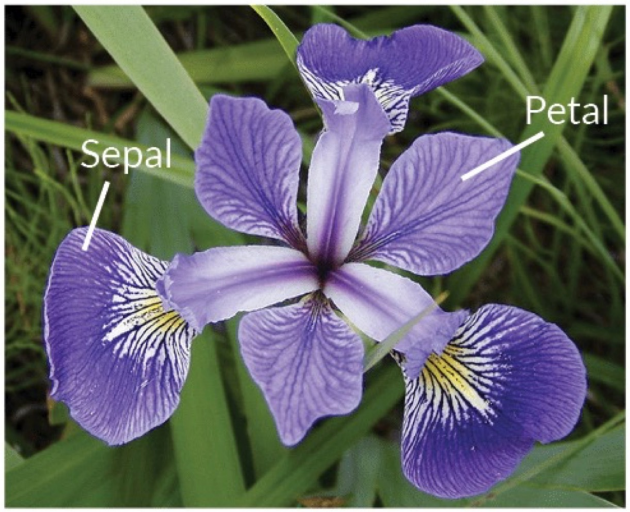
	A	B	C
1	**ID**	**Variable**	**Value**
2	1004	LastName	Smith
3	4587	LastName	Nayef
4	1727	LastName	Doe
5	6879	LastName	Jordan
6	1004	FirstName	Jane
7	4587	FirstName	Mohammed
8	1727	FirstName	Janice
9	6879	FirstName	Alex
10	1004	Height_inches	65
11	4587	Height_inches	75
12	1727	Height_inches	62
13	6879	Height_inches	77
14	1004	Weight_lbs	180
15	4587	Weight_lbs	215
16	1727	Weight_lbs	124
17	6879	Weight_lbs	160
18	1004	Insulin	0.60
19	4587	Insulin	1.46
20	1727	Insulin	0.72
21	6879	Insulin	1.23
22	1004	Glucose	163
23	4587	Glucose	150
24	1727	Glucose	177
25	6879	Glucose	205

Data Wrangling

`dplyr`

- mutate() adds new variables that are functions of existing variables
- select() picks variables based on their names.
- filter() picks cases based on their values.
- summarise() reduces multiple values down to a single summary.
- arrange() changes the ordering of the rows.
- You also need to know group_by() which allows you do any function by group.

Demo Data



Iris Versicolor

Iris Setosa

Iris Virginica