	Monday	Tuesday	Wednesday	Thursday	Friday
9:00	Hands-on: Unix and R hands-on workshop (Ruoqiao Chen/Bin Chen)	Introduction to Big Data and AI (Bin Chen)	Intro to single-cell "omics"	Hands-on: System biology and modeling using Julia (Bhattacharya)	Single-cell trajectories, networks, and RNA velocity (Bhattacharya)
10:00			Hands-on: Single-cell data portals (Nault)		
11:00			Pharmacological application of single-cell "omics" (Nault)		
12:00	Lunch on your own	Lunch on your own	Lunch on your own	Lunch on your own (return at 1:30PM)	Provided lunch & Group photo
13:00	RNA-Seq workshop (Rama Shankar/Bin Chen)	Transcriptomics-based drug discovery workshop (Dmitry Leshchiner/Bin Chen)	Introduction to System biology and modeling (Bhattacharya)	Hands-on: Single-cell QC, data structure, and visualization hands-on workshop (Nault)	Q&A and Round Table discussion (location TBD)
14:00					
15:00					Intro to spatial transcriptomics Hands-on: Exploring spatial transcriptomic datasets (Nault)
16:00					
17:00					

Location: All bootcamp activities will be in ISTB 1404 unless noted otherwise.

Instructors:

Bin Chen

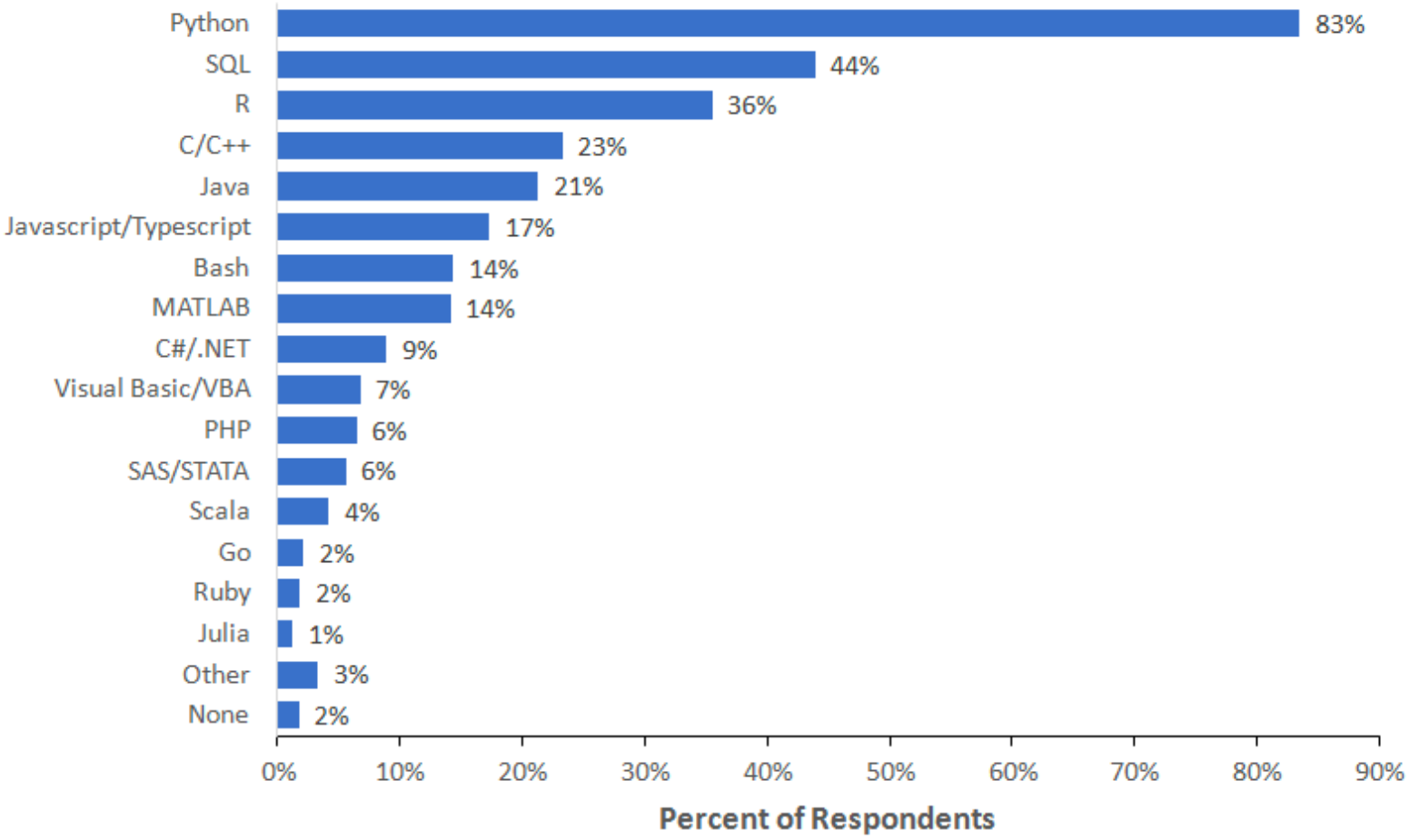
Rance Nault

Sudin Bhattacharya

Day 1 Morning

- Introduction to R and basic statistics (30mins, Bin Chen)
- Unix/R tutorial (90mins, Ruoqiao Chen)
- Practice (60mins)

What programming language do you use on a regular basis?



Note: Data are from the 2018 Kaggle Machine Learning and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018. A total of 18827 respondents answered the question.

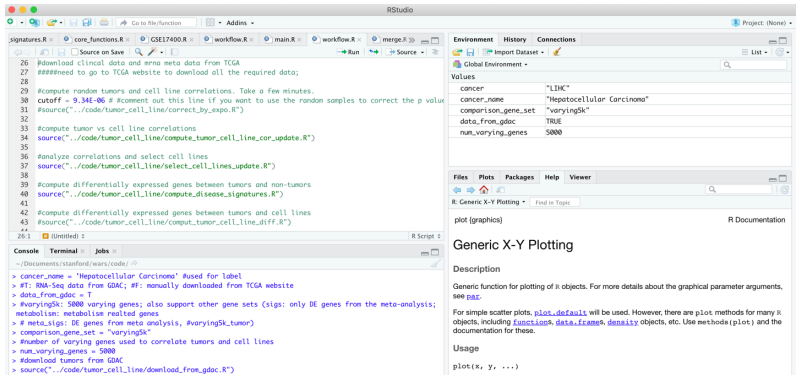
BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING

R resources

- Tutorial
  - http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual/
  - https://rstudio.com/resources/cheatsheets/
  - https://www.r-bloggers.com/
  - http://rafalab.github.io/pages/harvardx.html
- Troubleshooting
  - https://stackoverflow.com/
  - https://www.biostars.org/
- GitHub code repository

R

- Install R
- Install Rstudio



RStudio

Basic grammar:

- Variable
- Data type
- Data input/output
- Basic operator
- Conditionals and loops
- Function

R Packages
>10K

Base

customized

- Data manipulation: dplyr
- Data visualization: ggplot
- Web developer: shiny
- Documentation: Rmarkdown
- Text processing: stringr
- Machine learning: e1071

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

~2000

RStudio

signatures.R | core_functions.R | GSE17400.R | workflow.R | main.R | workflow.R | merge.R

```r
#download clincal data and mrna meta data from TCGA
#####need to go to TCGA website to download all the required data;

#compute random tumors and cell line correlations. Take a few minutes.
cutoff = 9.34E-06 # #comment out this line if you want to use the random samples to correct the p value
#source("../code/tumor_cell_line/correct_by_expo.R")

#compute tumor vs cell line correlations
source("../code/tumor_cell_line/compute_tumor_cell_line_cor_update.R")

#analyze correlations and select cell lines
source("../code/tumor_cell_line/select_cell_lines_update.R")

#compute differentially expressed genes between tumors and non-tumors
source("../code/tumor_cell_line/compute_disease_signatures.R")

#compute differentially expressed genes between tumors and cell lines
#source("../code/tumor_cell_line/comput_tumor_cell_line_diff.R")
```

26:1 (Untitled) R Script

Console | Terminal | Jobs

~/Documents/stanford/wars/code/

```
> cancer_name = 'Hepatocellular Carcinoma' #used for label
> #T: RNA-Seq data from GDAC; #F: manually downloaded from TCGA website
> data_from_gdac = T
> #varying5k: 5000 varying genes; also support other gene sets (sigs: only DE genes from the meta-analysis;
 metabolism: metabolism realted genes
> # meta_sigs: DE genes from meta analysis, #varying5k_tumor)
> comparison_gene_set = "varying5k"
> #number of varying genes used to correlate tumors and cell lines
> num_varying_genes = 5000
> #download tumors from GDAC
> source("../code/tumor_cell_line/download_from_gdac.R")
```

Environment | History | Connections

Global Environment

Values

cancer	"LIHC"
cancer_name	"Hepatocellular Carcinoma"
comparison_gene_set	"varying5k"
data_from_gdac	TRUE
num_varying_genes	5000

Files | Plots | Packages | Help | Viewer

R: Generic X-Y Plotting

plot {graphics} R Documentation

Generic X-Y Plotting

Description

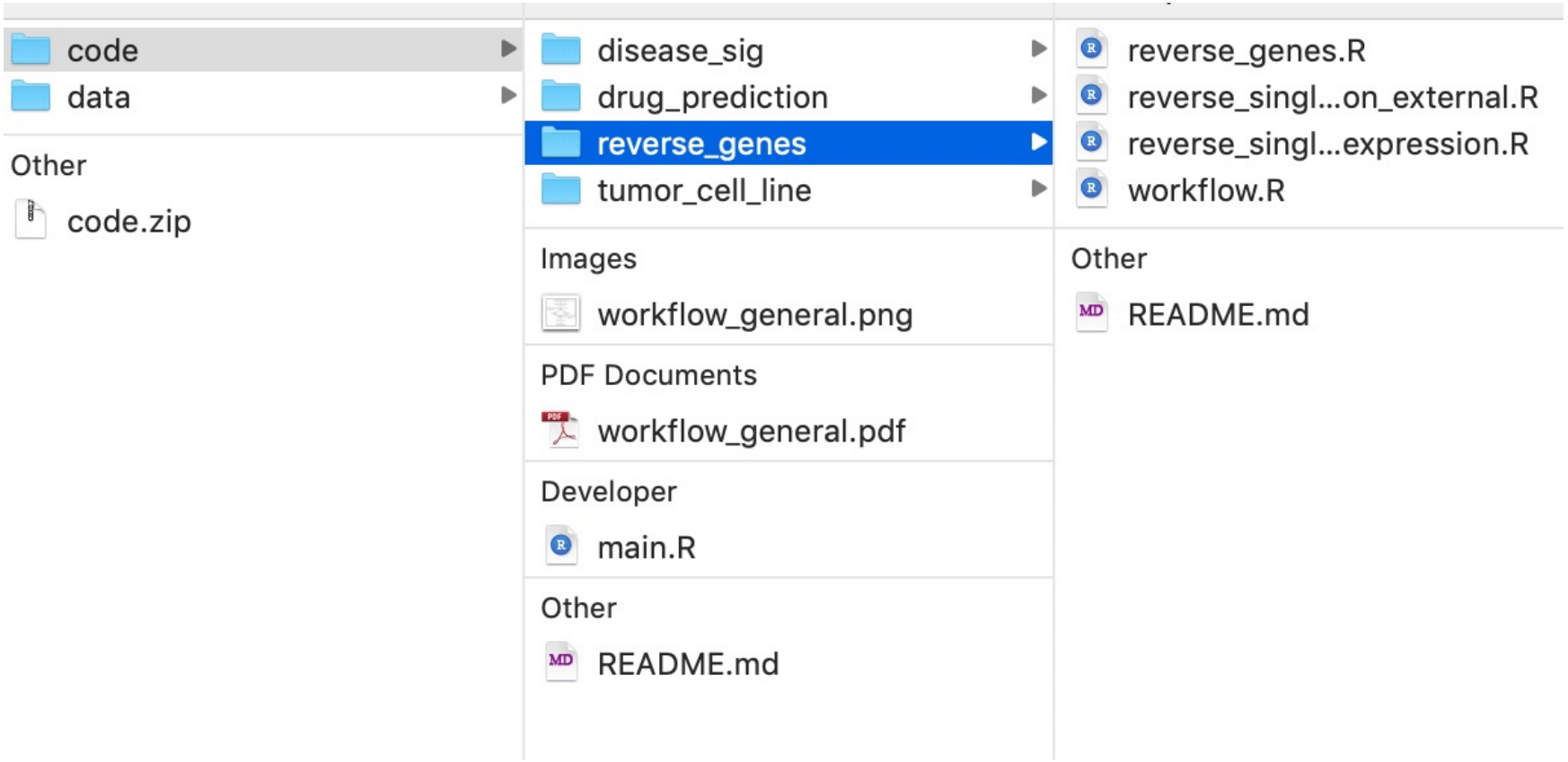
Generic function for plotting of R objects. For more details about the graphical parameter arguments, see par.

For simple scatter plots, plot.default will be used. However, there are plot methods for many R objects, including functions, data.frames, density objects, etc. Use methods(plot) and the documentation for these.

Usage

```
plot(x, y, ...)
```

Demo R project



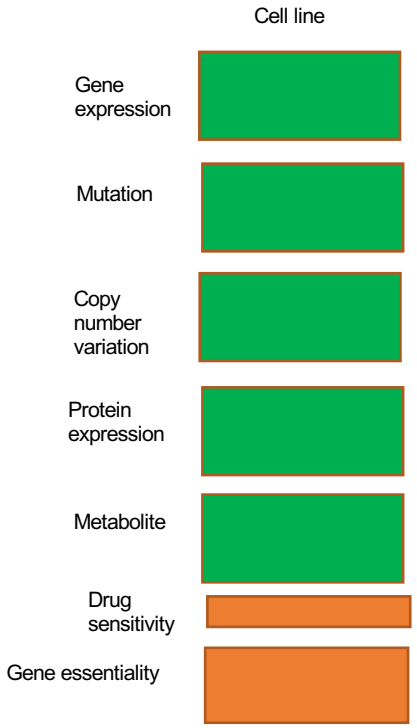
https://github.com/Bin-Chen-Lab/HCC_NEN

OCTAD_cell_line.RData

octad_cell_line_features

	id	name	type
VPS13D	mutation_VPS13D	VPS13D	mutation
AADACL4	mutation_AADACL4	AADACL4	mutation
TMEM57	mutation_TMEM57	TMEM57	mutation
ZSCAN20	mutation_ZSCAN20	ZSCAN20	mutation
POU3F1	mutation_POU3F1	POU3F1	mutation
VAV3	mutation_VAV3	VAV3	mutation

octad_cell_line_matrix



octad_cell_line_meta

DepMap_ID	stripped_cell_line_name	CCLE.Name	alias	COSMIC_ID	lineage	lineage_subtype
ACH-000001	NIHOVCAR3	NIHOVCAR3_OVARY	OVCAR3	905933	ovary	ovary_adenocarcinoma
ACH-000002	HL60	HL60_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE		905938	leukemia	AML
ACH-000003	CACO2	CACO2_LARGE_INTESTINE	CACO2, CaCo-2	NA	colorectal	
ACH-000004	HEL	HEL_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE		907053	leukemia	AML
ACH-000005	HEL9217	HEL9217_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE		NA	leukemia	AML
ACH-000006	MONOMAC6	MONOMAC6_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE		908148	leukemia	AML

Data input and output

Data Type

- Numeric
- Character
- Logical
- Factor

Data Type

- Vector
- Data.frame
- Matrics
- Arrays
- List
- RData

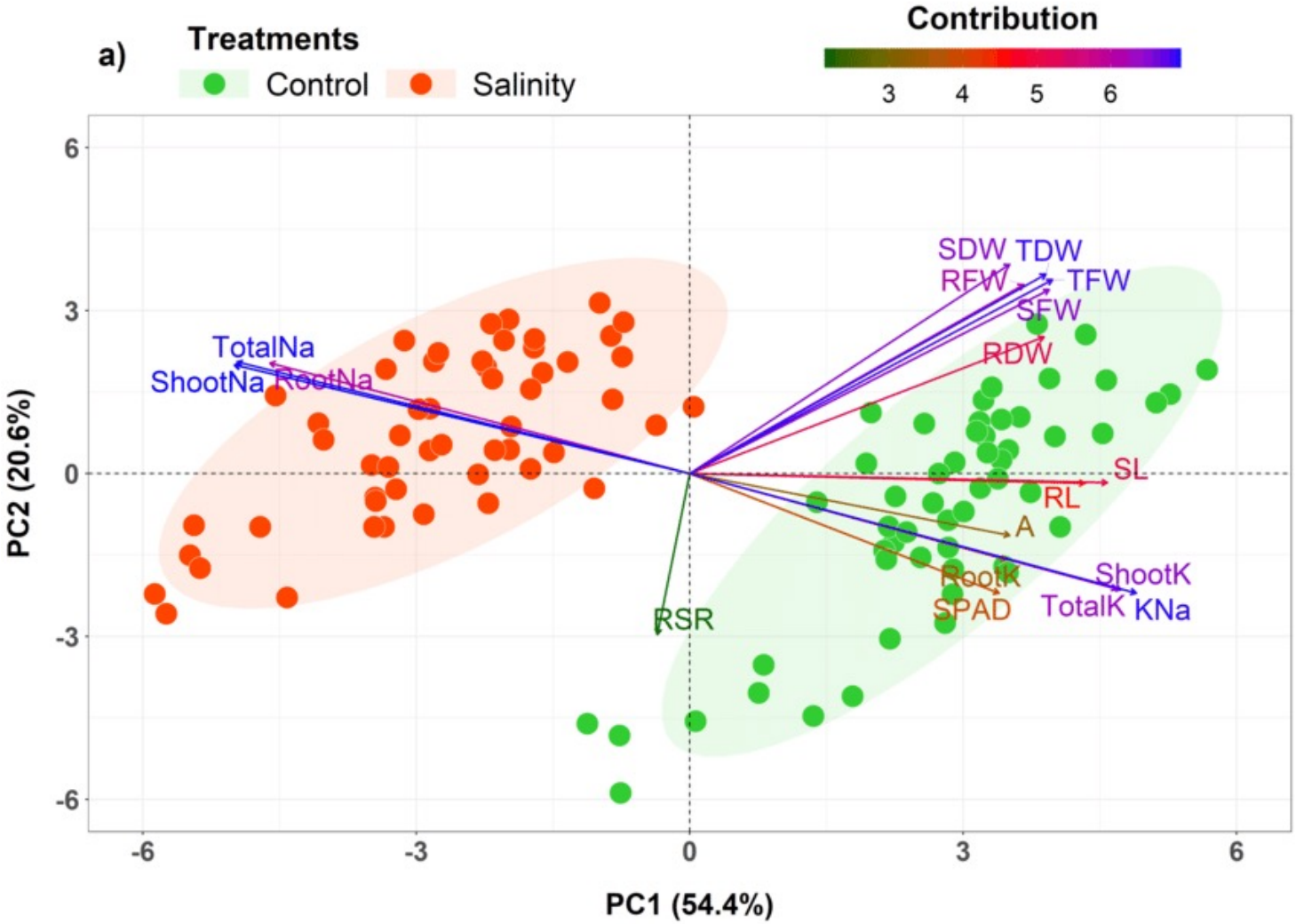
Subsetting

Basic Operators and Calculations

Data summary

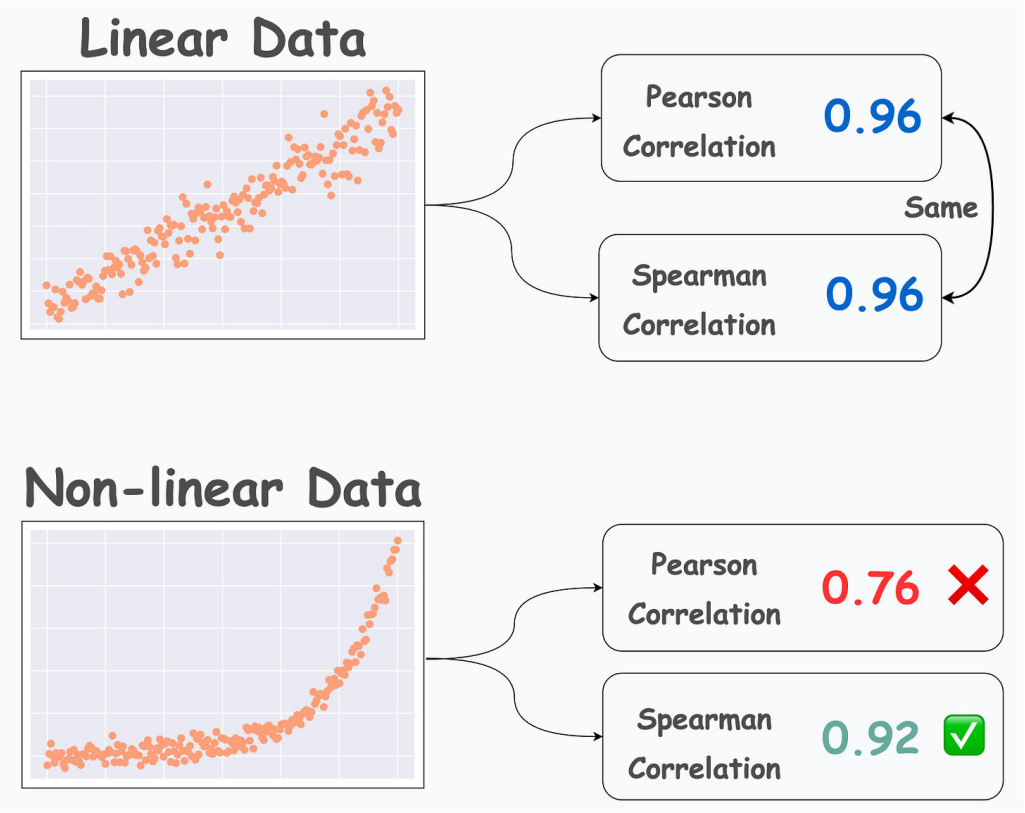
Data Visualization

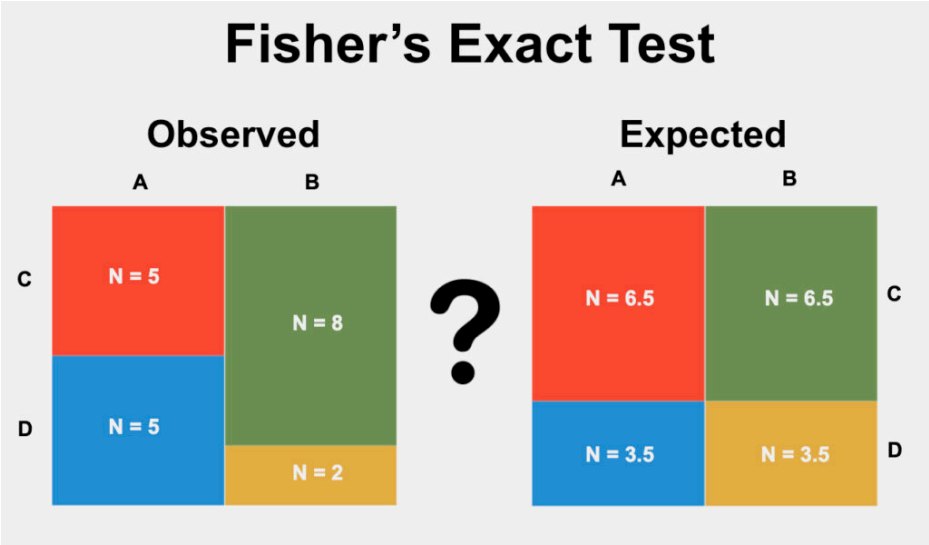
- PCA



Correlation analysis

- Continuous data
  - Pearson Correlation
  - Spearman Correlation
- Categorical Data
  - Fisher test
  - Chi-square test





Statistical test

- T-test
- Wilcoxon signed-rank test