



# STA 426 First Lecture

- survey
- course structure
- Molecular Biology lecture (Hubert)
- R survey + computing + Exercise 1



# COVID19 Policy

Statistical Bioinformatics // Institute of Molecular Life Sciences

- COVID certificates (or equivalent) required to attend lectures in person; please send yours to Mark via Slack DM
- Masks are not required, but **highly** recommended
- Lecturers (speakers) are allowed to speak without a mask; space-permitting, we leave the first 1-2 rows empty
- The lecture hall to be aired regularly

<https://github.com/sta426hs2021/material/>



## Today's structure

**9.00-9.45:** Survey + Course Structure (Mark)

**10.00-10.45:** Introduction to Molecular Biology (Hubert)

**11.00-11.45:** Computing + R quiz + Rmarkdown exercise



## Survey: Statistical Insight

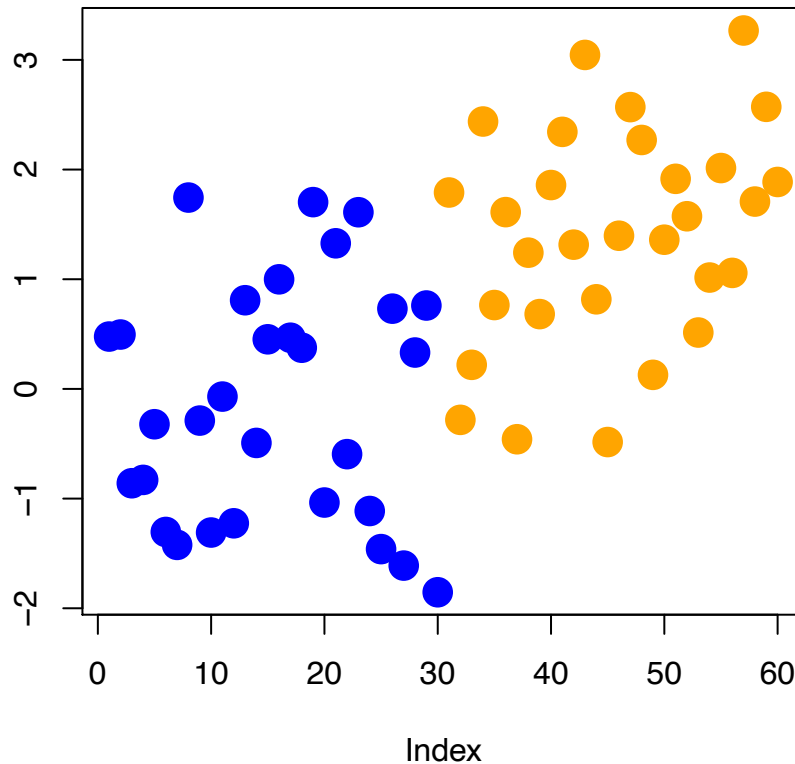
movo.ch

Token: ZE HI ZO SO

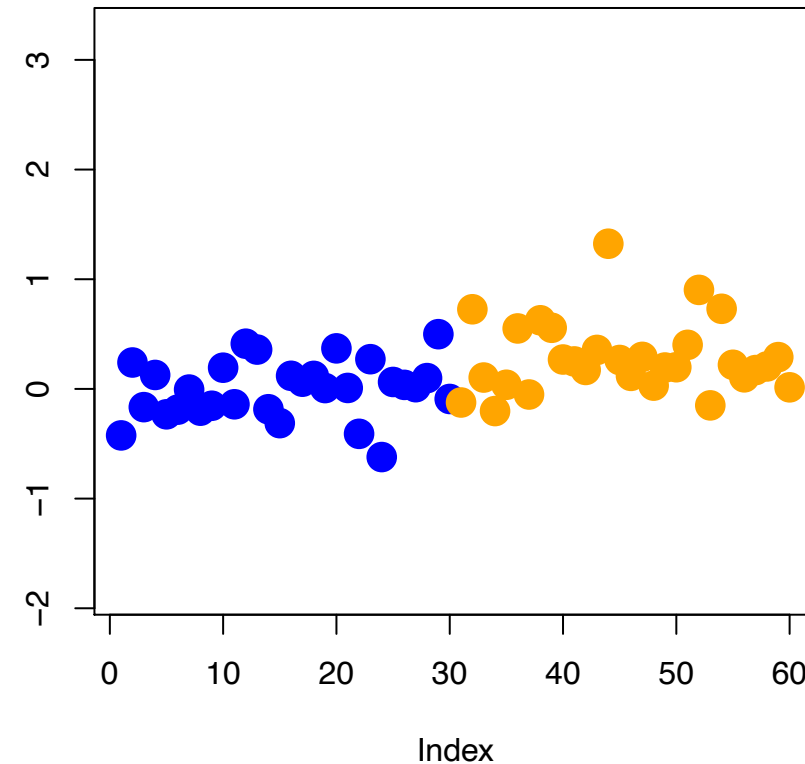


Question 1: Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?

**A**

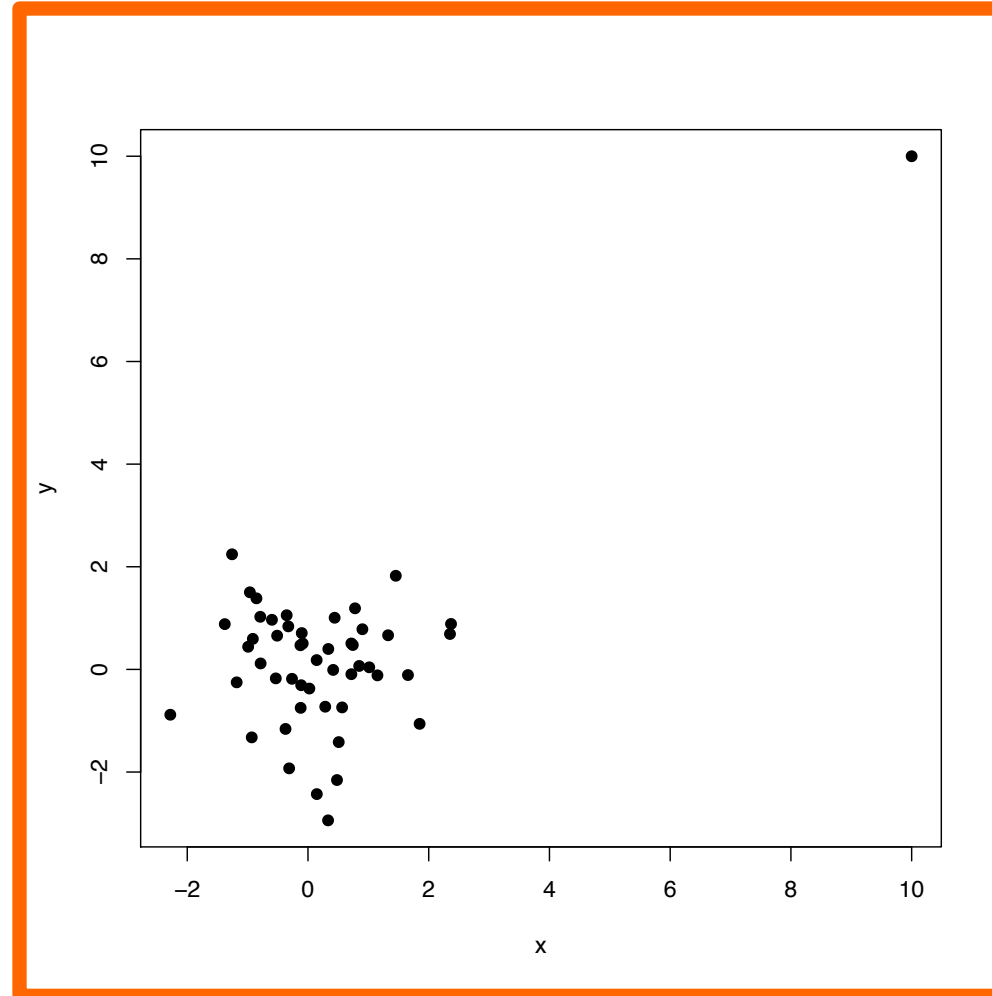


**B**





Question 2: In your view, what best describes the associations shown in the plot of 'x' and 'y' ?





Question 4: Given this design matrix, describe the experimental design.

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Question 6: Of these equations, which one resembles the standard two sample t-test ?

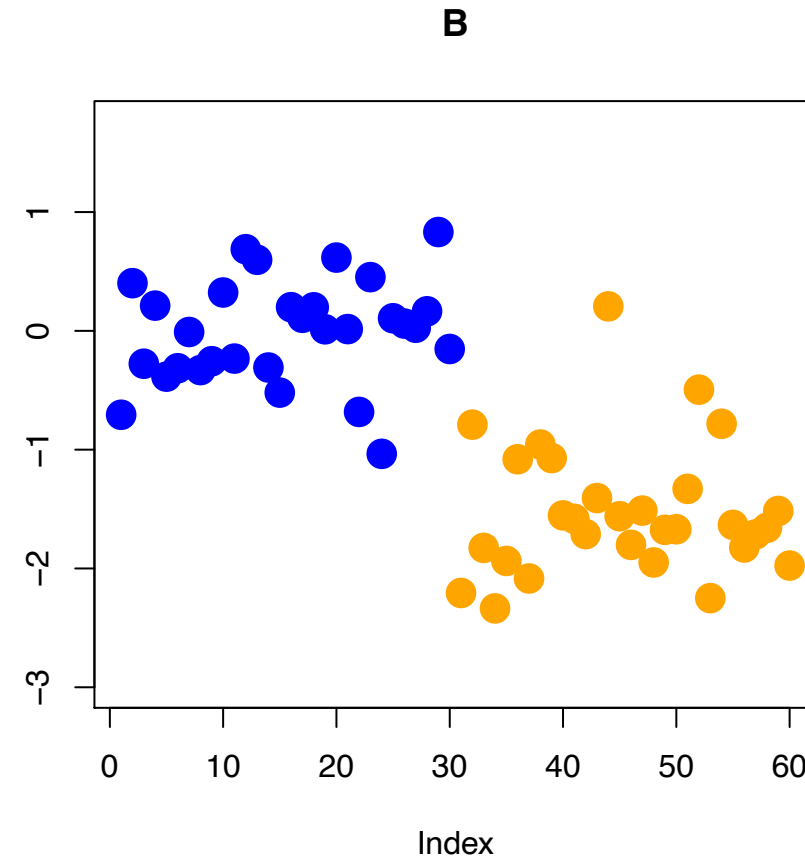
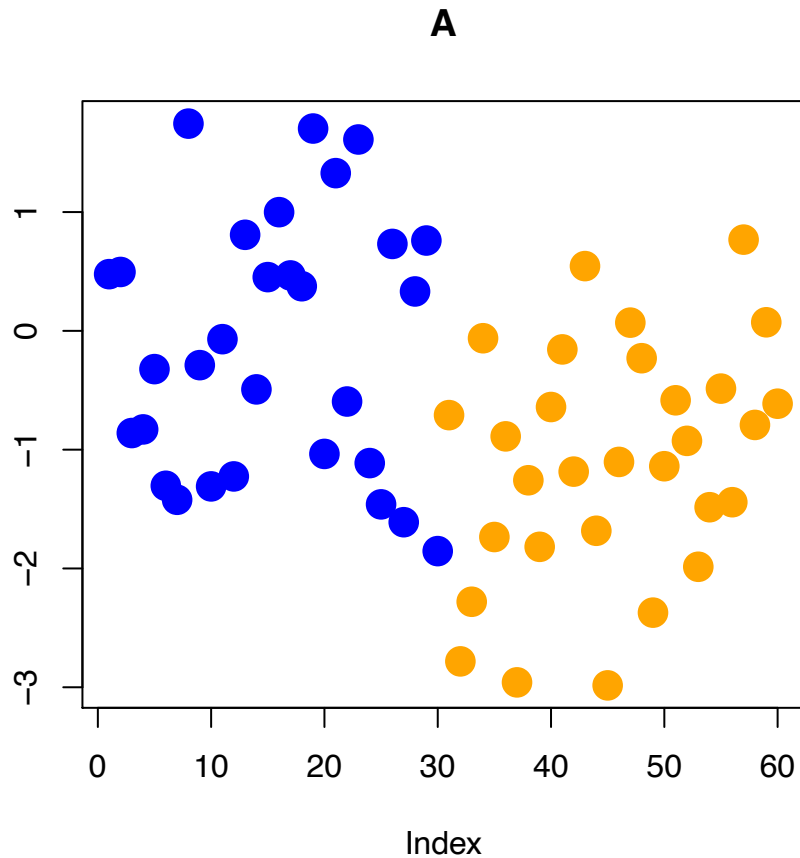
1 
$$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

2 
$$\sum^k \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

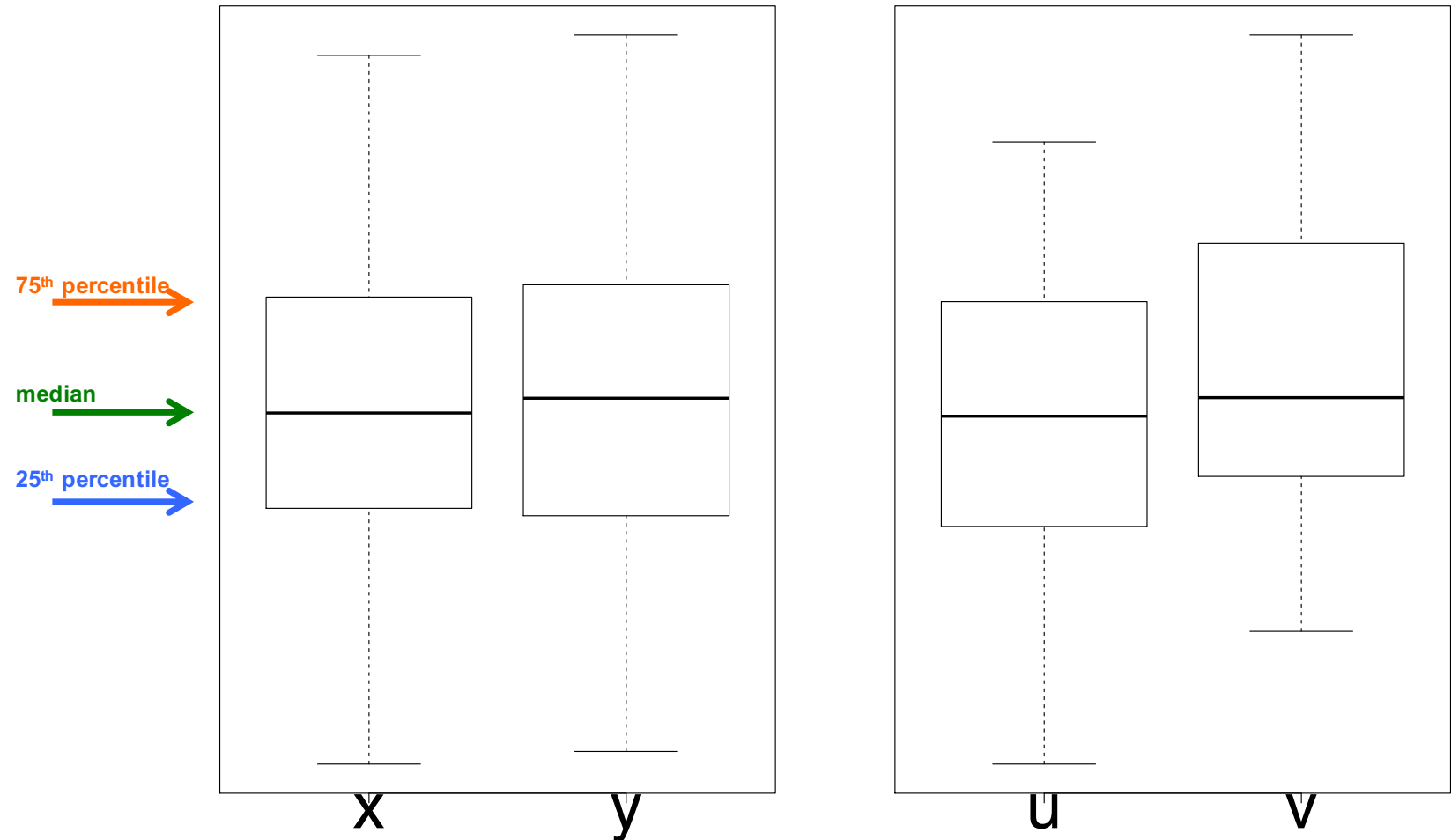
3 
$$\frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$



Question 7: Which plot highlights more (statistical) evidence for a change in the population means (between orange and blue)?



Question 8: Given these boxplots, which of two underlying distributions are more similar?





## Course communication

- Video situation: still unfolding, will keep you updated. Likely, we will have synchronous lectures, maybe also full recordings
- Slack: vast majority of communication happens here (note: all invitations were sent to UZH email addresses)
- Except for exceptional circumstances, **no emails please**; communicate on Slack only (maybe later on GitHub)
- Slack policy: unless private, ask questions in a public channel (please note: *good questions get good answers*); use threads when relevant; good manners/behaviours are expected



## Course evaluation

1. Journal club presentation	20%
2. Project	50%
3. Exercises	30%
4. Technology day (participation)	0% or -10%



## Rough structure of lecture/exercise time

Monday mornings: we will run X.00-X.45; X in {9,10,11}

- Lectures and Exercises
- Lecture/journal club presentation (9.00-whenever)
- Remaining time: free (can be used to work on exercises; we are available for questions)



## M.Sc. thesis projects

If you are:

- in a M.Sc. programme (ETHZ or UZH)
- have a solid background/experience in mathematics / statistics / computation
- have an interest in research in this field (“statistical bioinformatics”)
- looking for a thesis project

→ Discuss a project in my lab

## Critical skills needed by statisticians (Jeffrey Leek's words):

With all the excitement going on around statistics, there is also increasing diversity. It is increasingly hard to define “statistician” since the definition ranges from [very mathematical](#) to [very applied](#). An obvious question is: what are the most critical skills needed by statisticians?

So just for fun, I made up my list of the top 5 most critical skills for a statistician by my own definition. They are by necessity very general (I only gave myself 5).

1. **The ability to manipulate/organize/work with data on computers** - whether it is with excel, R, SAS, or Stata, to be a statistician you have to be able to work with data.
2. **A knowledge of exploratory data analysis** - how to make plots, how to discover patterns with visualizations, how to explore assumptions
3. **Scientific/contextual knowledge** - at least enough to be able to abstract and formulate problems. This is what separates statisticians from mathematicians.
4. **Skills to distinguish true from false patterns** - whether with p-values, posterior probabilities, meaningful summary statistics, cross-validation or any other means.
5. **The ability to communicate results to people without math skills** - a key component of being a statistician is knowing how to explain math/plots/analyses.



## Learning outcomes (in my words)

- Understand the fundamental “scientific process” in the field of Statistical Bioinformatics
- Be equipped with the skills / tools to preprocess genomic data (Unix, Bioconductor, mapping, etc.) and ensure reproducible research (R / markdown)
- Have a general knowledge of (some) **types** of data and **biological applications** encountered with high throughput genomic data
- Have the general knowledge of the range of statistical methods that get used with microarray and sequencing data
- Gain the ability to apply statistical methods / knowledge / software to a collaborative biological project
- Gain the ability to critical assess the statistical bioinformatics literature
- Write a coherent summary of a bioinformatics problem and it’s solution in statistical terms





## The semester-long course structure (subject to change)

Date	Lecturer	Topic	Exercise	JC1	JC2
20.09.2021	Mark + Hubert	admin; mol. bio. basics	R markdown; git(hub)		
27.09.2021	Mark	interactive technology/statistics session	group exercise: technology pull request		
04.10.2021	Hubert	NGS intro; exploratory data analysis	EDA in R		
11.10.2021	Hubert	mapping	Rsubread		
18.10.2021	Mark	limma + friends	linear model simulation + design matrices		
25.10.2021	Hubert	RNA-seq quantification	RSEM	X	X
01.11.2021	Mark	edgeR+friends 1	basic edgeR/voom	X	X
08.11.2021	Mark	edgeR+friends 2	advanced edgeR/voom	X	X
15.11.2021	TBA	hands-on session #1: RNA-seq	FASTQC/Salmon/etc.	X	X
22.11.2021	Hubert	single-cell 1: preprocessing, dim. reduction, clustering	clustering	X	X
29.11.2021	TBA	hands-on session #2: cytometry	cytof null comparison	X	X
06.12.2021	Mark	single-cell 2: clustering, marker gene DE	marker gene DE	X	X
13.12.2021	Pierre-Luc	hands-on session #3: single-cell RNA-seq (cell type definition, differential state)	full scRNA-seq pipeline	X	X
20.12.2021	Mark	loose ends: HMM, EM, robustness	segmentation, peak finding	X	X



## Expectations: **journal club** presentation

- 20-25 minutes (+5 minutes discussion)
- MUST:
  - ➔ be a paper about a **statistical** method in bioinformatics
  - ➔ be approved by Mark/Hubert
- Should:
  - ➔ describe the biological context and/or data collected
  - ➔ describe the (new) model used
  - ➔ describe comparisons to existing methods
- Should not:
  - ➔ be one of the papers discussed in detail in lectures: limma, edgeR, DEXSeq, etc.
- (since 2017) feedback forms from fellow students



## Expectations: **project**

- ~10-15 page report, with R code in line (e.g. **knitR** / **Rmarkdown**)
- Describe the biological setting, statistical analysis, exploratory analysis with publication-quality graphics embedded
- Three possibilities:
  - Comparison of statistical methods (simulation / reference data + metrics)
  - Reproduce an analysis from a paper from the raw data
  - Real collaborative project with FGCZ or a local laboratory
- Be strategic: work on something related to your interests!
- Typically due at end of first working week of January



## Expectations: **exercises**

- There will be an exercise **every** week
- Across 14 weeks, the *best 9* exercises are counted towards the 30%



## Soft technical skills needed (developed) in this course ...

- **Data Science!**
- Use unix-like operating system to run command-line programs
- Options:
  - use your own computer (if Windows, use cygwin)
  - use [renkulab.io](https://renkulab.io)
- R: from the command line or RStudio (<https://rstudio.com/>); getting help; creating workflows; how to make publication-quality graphics (ggplot2); knitr/Rmarkdown
- Bioconductor – [www.bioconductor.org](http://www.bioconductor.org)
- git/github
- bioconda/Docker (cloud computing)



# Hubert's lecture



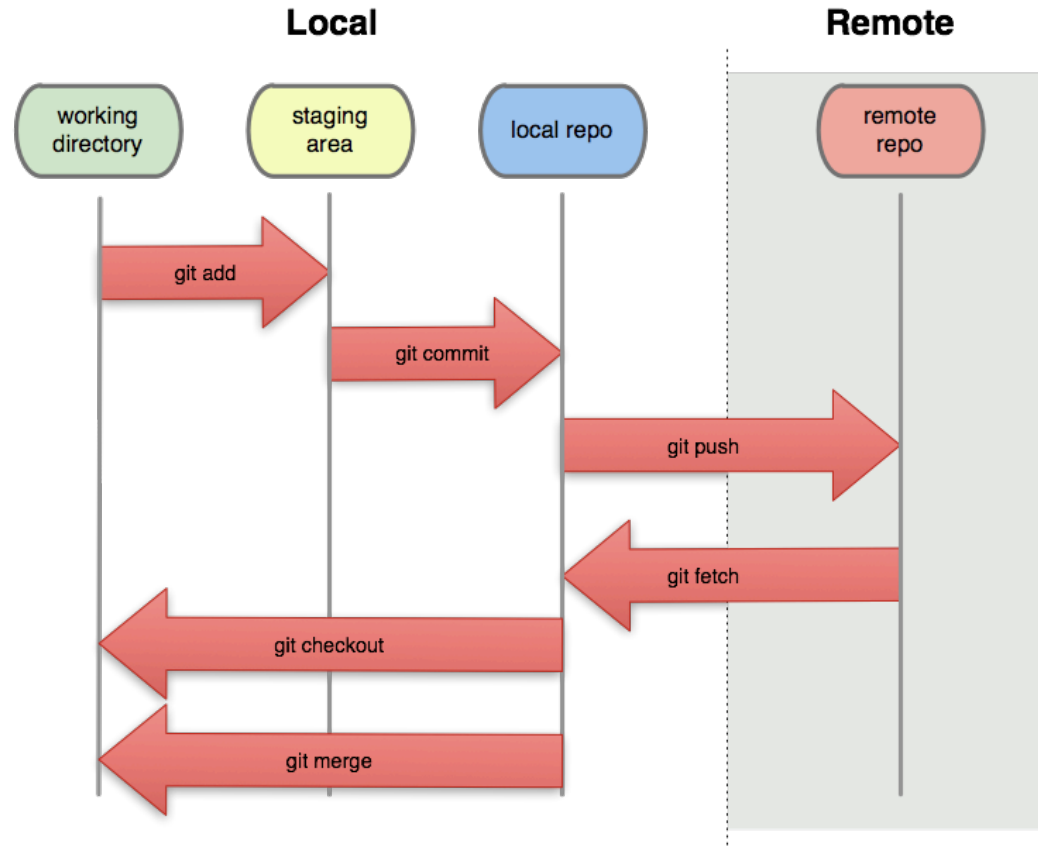
## Demos:

- git/github
- [renkulab.io](https://renkulab.io) (fork a project, start a session)
- RStudio/Rmarkdown

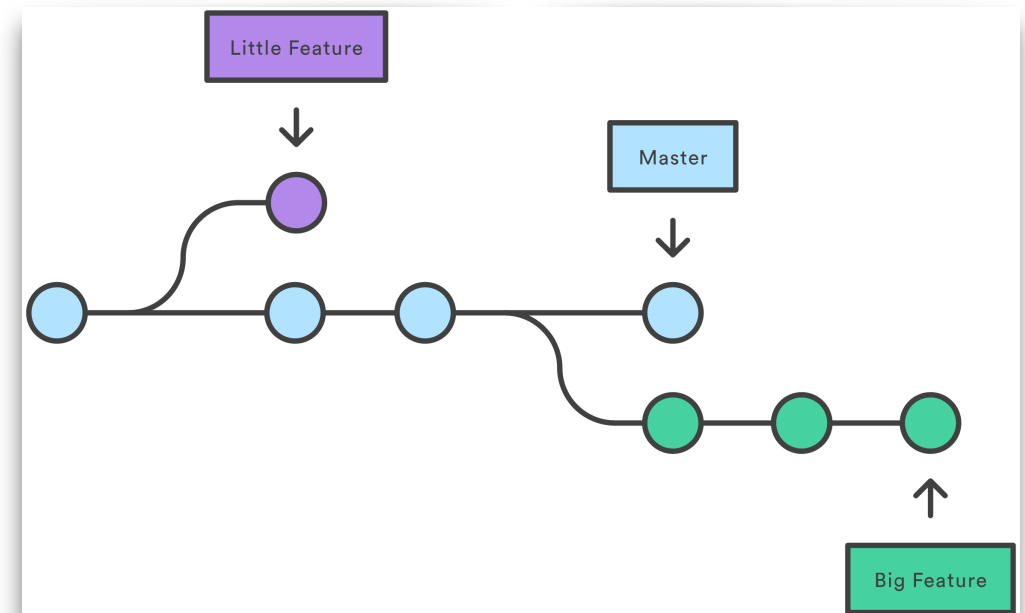


```
git clone
git pull
git status
git branch
git commit
git add
git checkout
git push
```

## Quick intro to Git/Github (version control)



### Branching







# Exercise 1

Part a: GitHub

Part b: R/knitR/Rmarkdown



Note: all homework submissions occur via github

## Week 1 Exercise (part a):

1. If you haven't already, create an account at [github.com/join](https://github.com/join); give userid to Mark via <https://forms.gle/4dhWdcPa9TfsMhHMA>
2. Acquaint yourself with git / github (gitlab) [1]; make sure you can check in (push) / out (pull / clone) files from command line or app [2].
3. Create a new public git repository, add a README.md (using markdown [3]) and add some content; include an image; include a web link, etc.
4. Add an Issue to the 'material' repo [4] with a link to your repo (you can delete the repo after I've closed the issue, if you want)

[1] <https://gist.github.com/andrewpmiller/9668225>

[2] <https://confluence.atlassian.com/stash/basic-git-commands-278071958.html>

[3] <http://markdowntutorial.com/>

[4] <https://github.com/sta426hs2020/material>

## Rmarkdown / knitr for executable documents / reproducibility

### Week 1 Exercise (part b):

1. Test your R knowledge here: <https://forms.gle/NjpFPW5G6Y8viU3C6>
2. Acquaint yourself with knitr PDF/HTML Rmarkdown documents [1].
3. Create an HTML document that samples 100 values from a log-normal distribution (say,  $\mu=1$ ,  $\sigma=.25$ ); create a histogram of the distribution and the distribution on the log scale; report the mean and variance of the sample in line in the text. In general, do not just dump the R code and plots in the HTML document; break your exercises into sections with headings, add some text and make it readable.
4. Add the RMD and HTML files to your repo from Exercise part a.

[1] <https://www.r-bloggers.com/r-markdown-and-knitr-tutorial-part-1/>