Intro to STA 426 course structure + Some GitHub/R/knitR + Some Molecular Biology

Today's structure

9.00-9.45 (Y34-F-01): Ice Breakers + Survey + Course Structure (Mark)

10.00-10.15 (Y01-F-50): Computer room Orientation (Tina)

10.15-10.45 (Y01-F-50): Troubleshooting logins + R quiz + Rmarkdown exercise

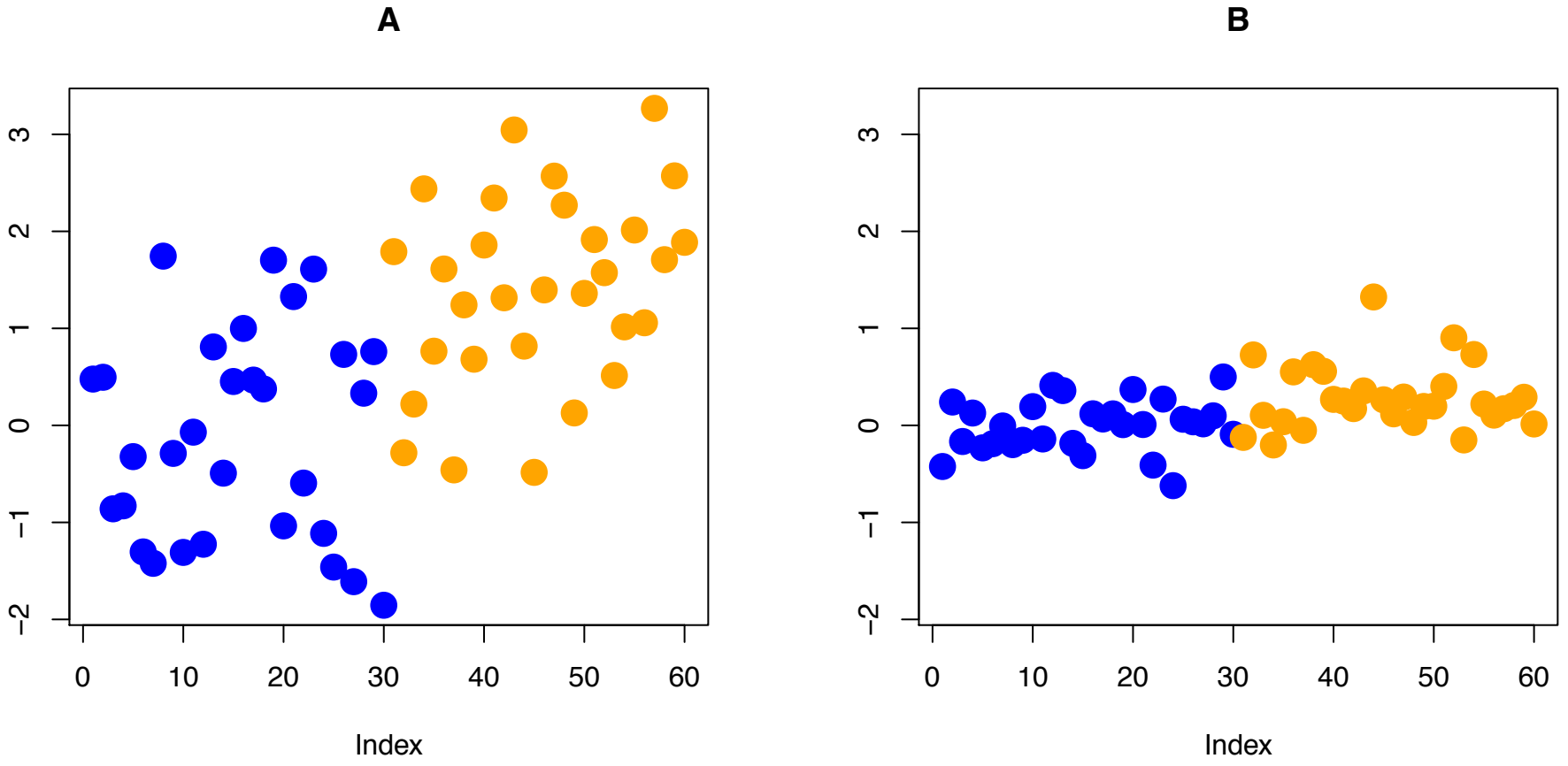
11.00-11.30 (Y01-F-50): Introduction to Molecular Biology (Hubert)

Survey: Statistical Insight

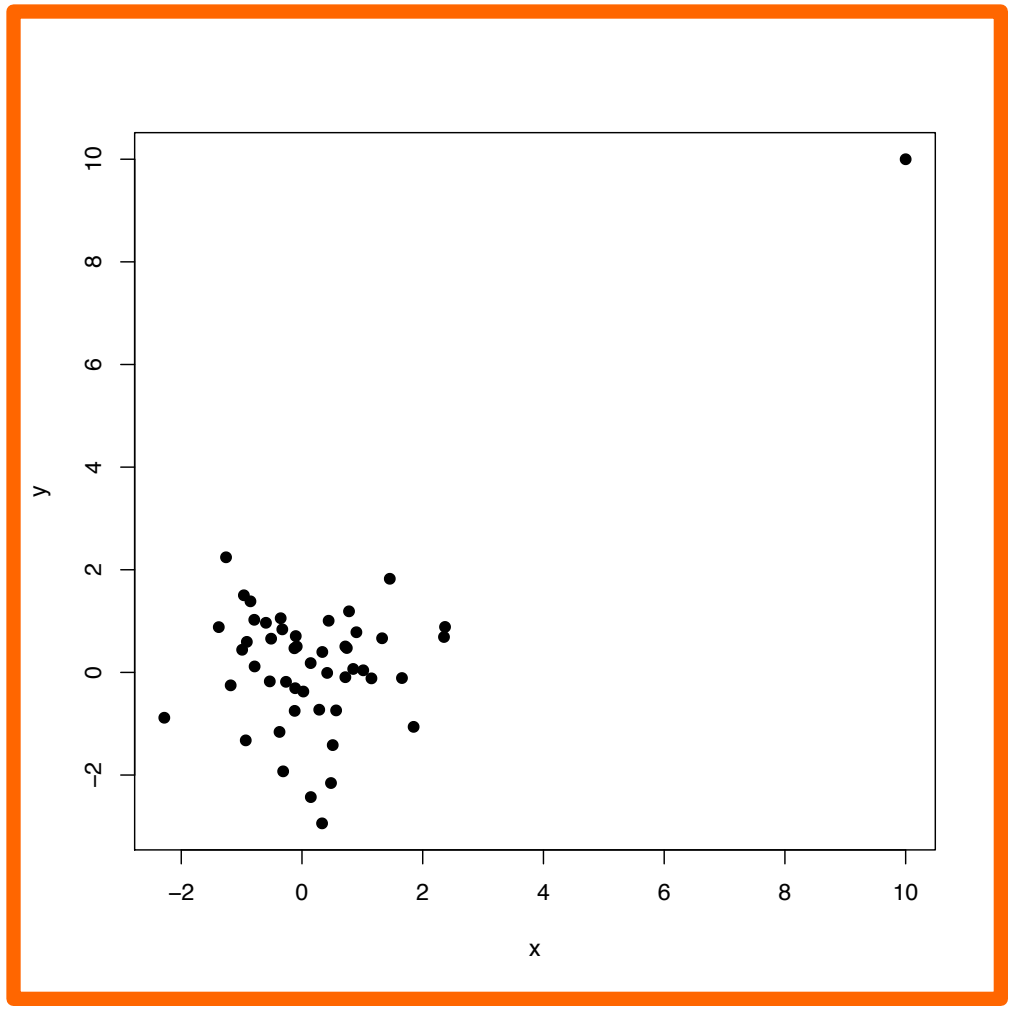
movo.ch

Token:

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



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

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

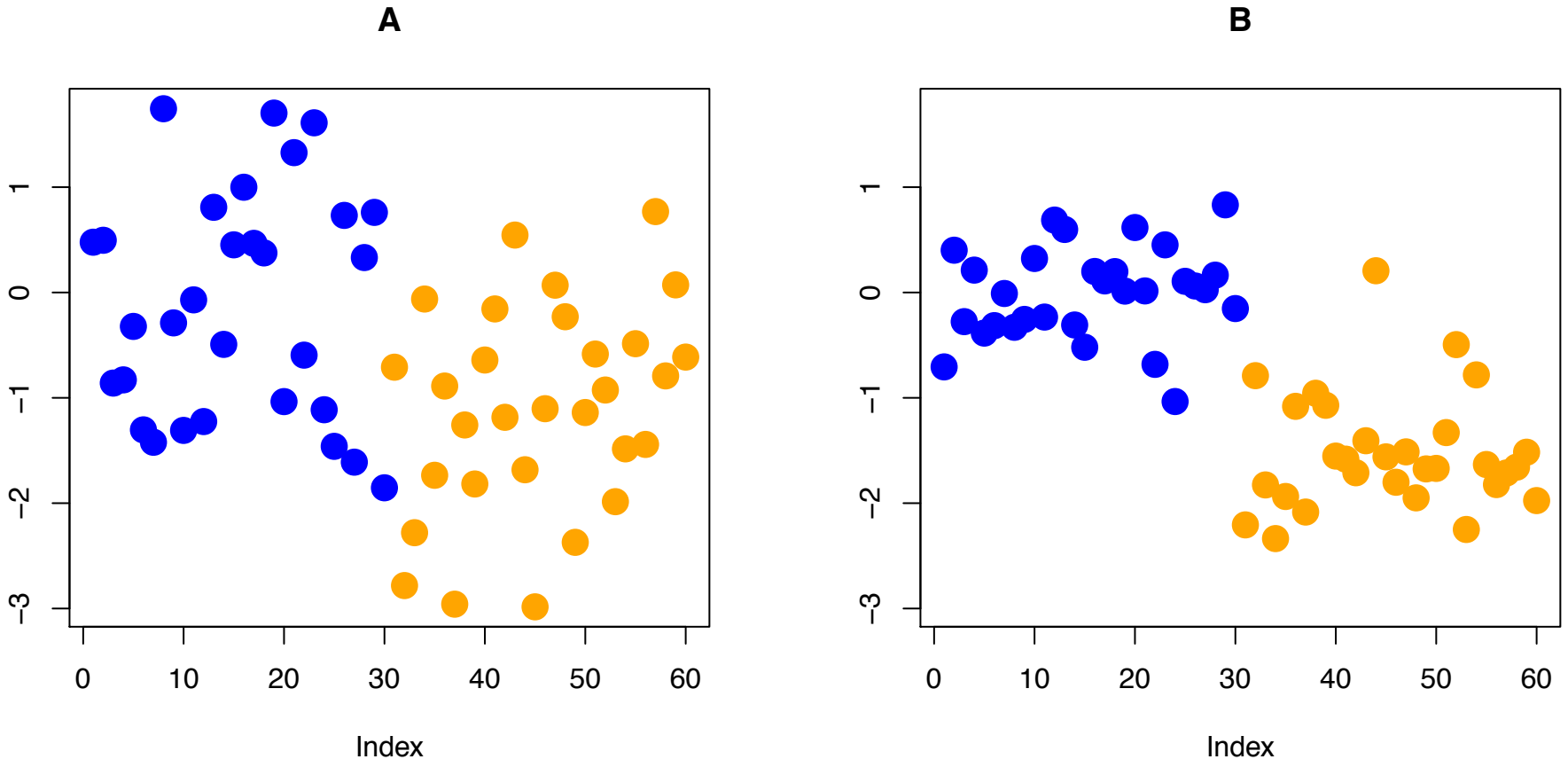
2

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

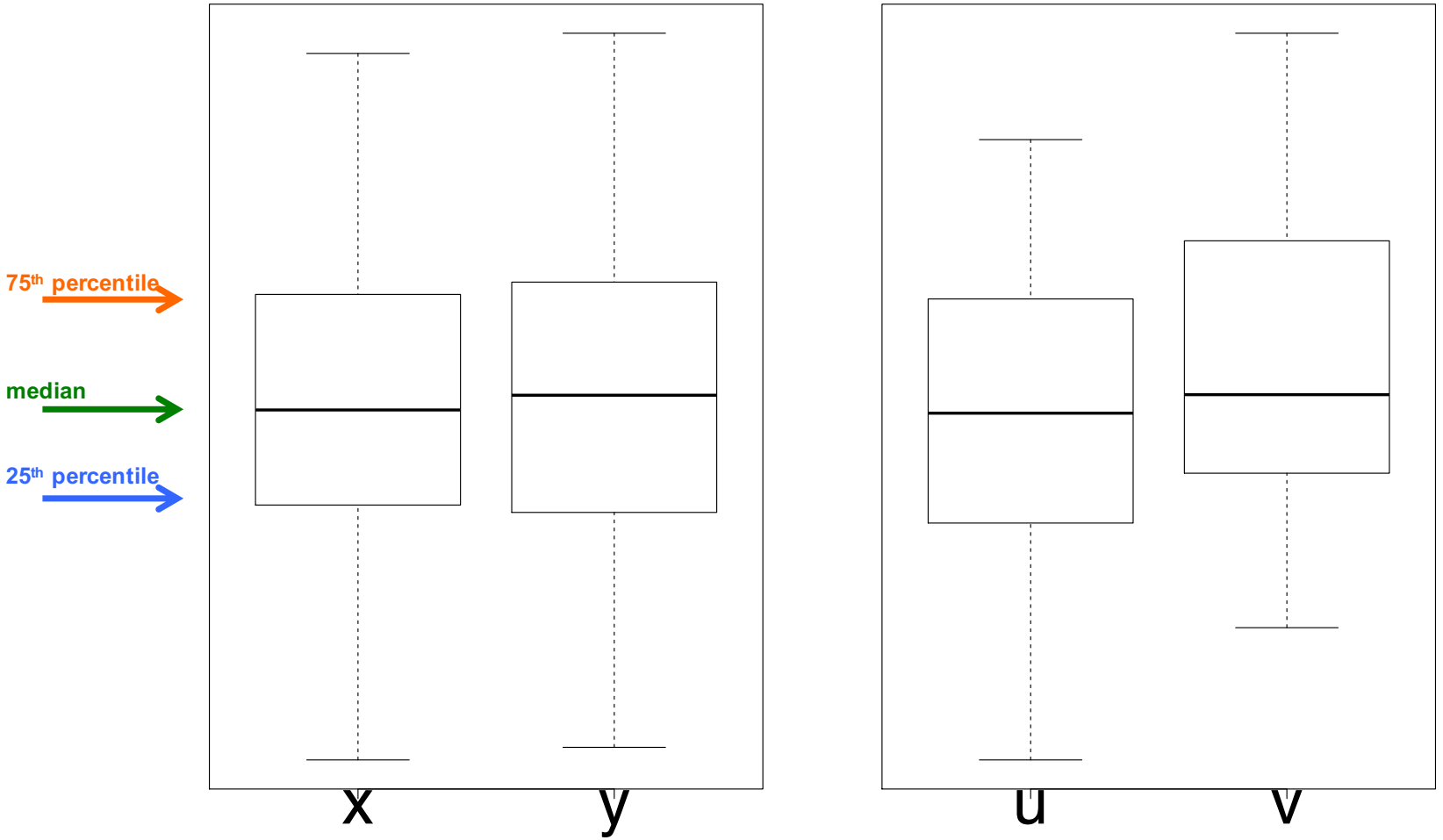
3

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

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?



Rough structure of lecture/exercise time

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

- Lectures: Y34-F-01
- Exercises: Y01-F-50
- Lecture/journal club presentation (9.00-whenever)
- Remaining time: in the computer lab (Y01-F-50) doing exercises/project

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

http://simplystatistics.org/2013/07/03/repost-the-5-most-critical-statistical-concepts/

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

Course evaluation

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

The semester-long course structure (subject to change)

Date	Lecturer	Topic	Exercise	JC1	JC2
17.09.2018	Mark + Hubert	admin; mol. bio. basics	R markdown; git(hub)		
24.09.2018	Hubert	NGS intro; exploratory data analysis	EDA in R		
01.10.2018	Mark + Hubert	interactive technology/statistics session	group exercise: technology pull request		
08.10.2018	Hubert	mapping	Rsubread		
15.10.2018	Mark	limma + friends	linear model simulation + design matrices		
22.10.2018	Hubert	RNA-seq quantification	RSEM		
29.10.2018	Charlotte	hands-on session #1: RNA-seq	FASTQC/Salmon/etc.	X	X
05.11.2018	Mark	edgeR+friends 1	basic edgeR/voom		
12.11.2018	Mark	edgeR+friends 2	GLM/DEXSeq		
19.11.2018	Mark	single-cell dim. reduction + clustering; FDR	conquer		
26.11.2018	Lukas	hands-on session #2: cytometry	cytof null comparison	X	X
03.12.2018	Hubert	classification	MLInterfaces		
10.12.2018	Mark	loose ends: HMM, EM, robustness	segmentation, peak finding		
17.12.2018	Mark	hands-on session #3: single-cell RNA-seq	full scRNA-seq pipeline	X	X

Expectations: **journal club** presentation

- 20-25 minutes (+5 minutes discussion)
- MUST:
  - ➡ be a paper about a **statistical** method in genomics
  - ➡ be approved by Mark/Hubert
- Should:
  - ➡ describe the biological context
  - ➡ 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 / independent 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

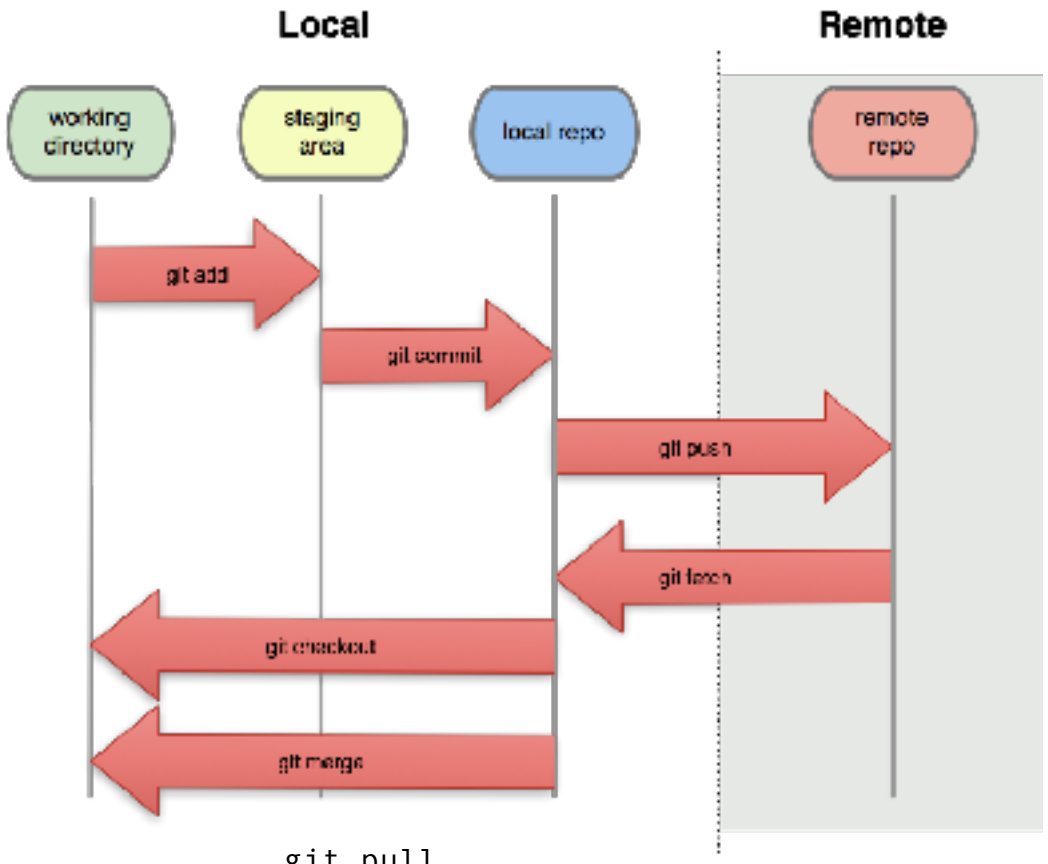
Expectations: **exercises**

- There will be an exercise **every** week
- Across 14 weeks, the *best 9* exercises are counted towards the 30% (today's exercise is a gift)

Soft technical skills needed (developed) in this course …

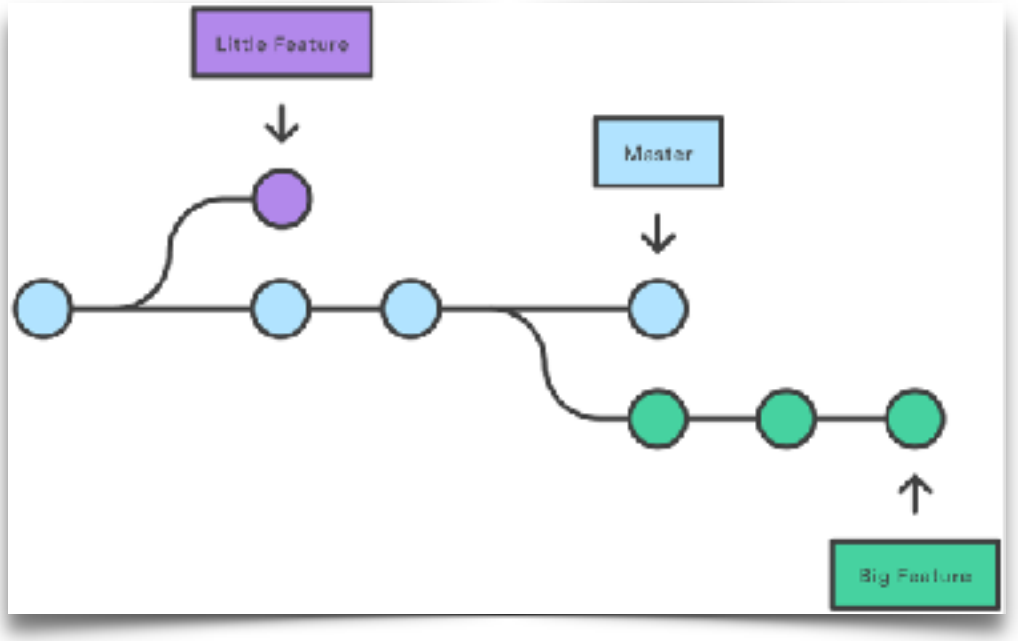
- **Data Science!**
- Use unix-like operating system to run command-line programs
- Options:
  - use your own Linux/MacOSX computer; N.B.: you may be able to do everything from Windows (e.g., cygwin)
  - use the Macs in Y01-F-50
- R: from the command line or R studio; getting help; creating workflows; how to make publication-quality graphics; knitr/Rmarkdown
- Bioconductor – www.bioconductor.org

Quick intro to Git/Github (version control)




```
git pull
git status
git branch
```

Branching



https://blog.seibert-media.net/blog/2015/07/31/git-mit-branches-arbeiten-git-branch/

https://skylabcoders.github.io/bootcamp-winter2017/?full#17

GitHub + knitR exercise

All homework submissions occur via github

Homework (part 1):

1. Create an account at github.com (Slack —> Mark with your userid)
2. Acquaint yourself with git/github [1]
3. Make sure you know how to check in / out files from command line or app [2]
4. Create a new public repository, add a README.md (using markdown [3]) and add some content
   - Include an image; Include a web link
   - add an Issue to the materials repo to let me know that you've done it
   - (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/

Rmarkdown / **knitR** for executable documents / reproducibility

Homework (part 2):

Acquaint yourself with knitR PDF/HTML Rmarkdown documents [1]:

1. Create an HTML/PDF 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.
   - Do not just dump the R code and plots in the HTML/PDF document; add some text and headings and make it readable (i.e., the document should be self-explanatory)