BINF 5003: Data Mining, Modeling, and Biostatistics

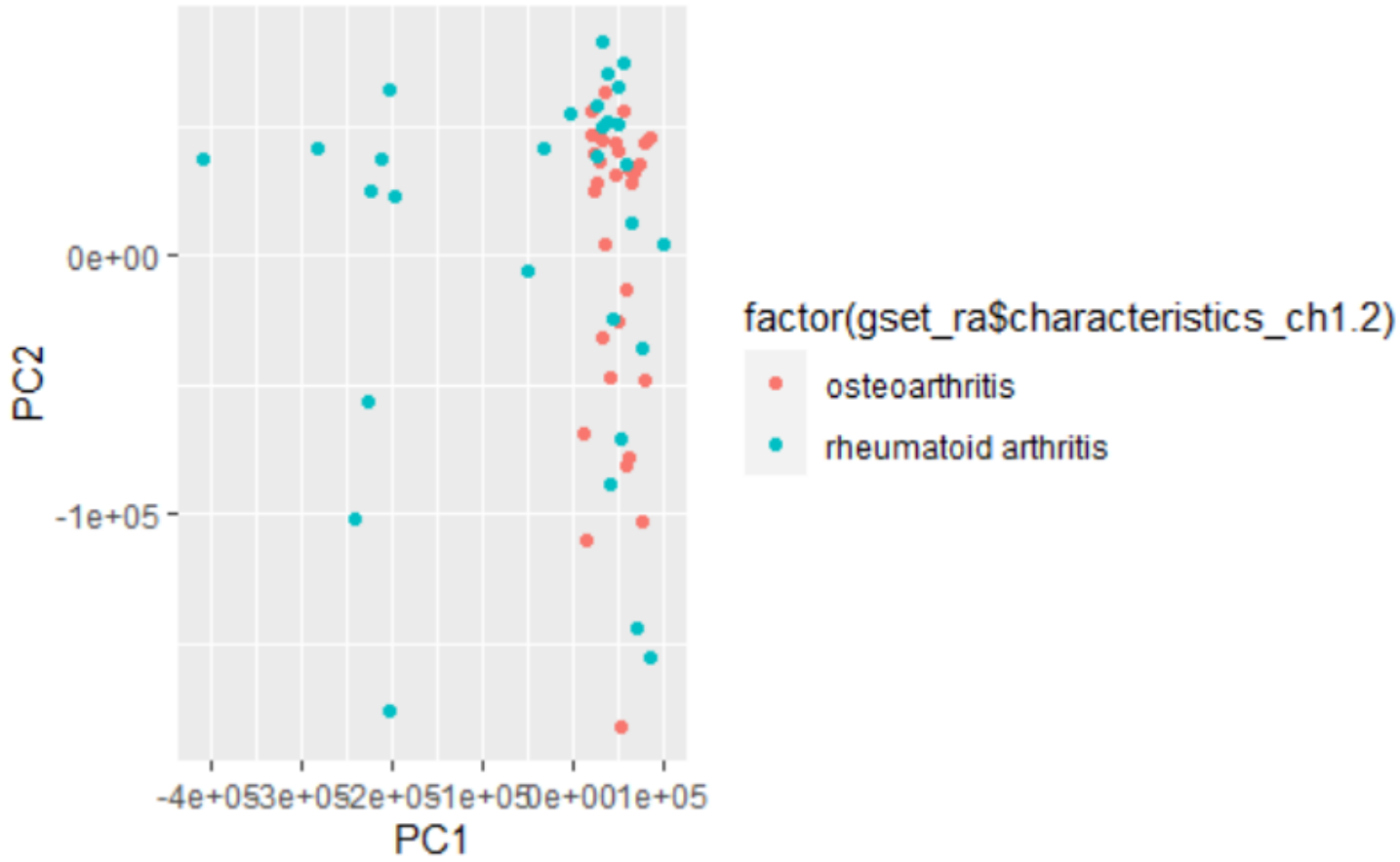
Week 10

Module 6 – Dimensionality Reduction

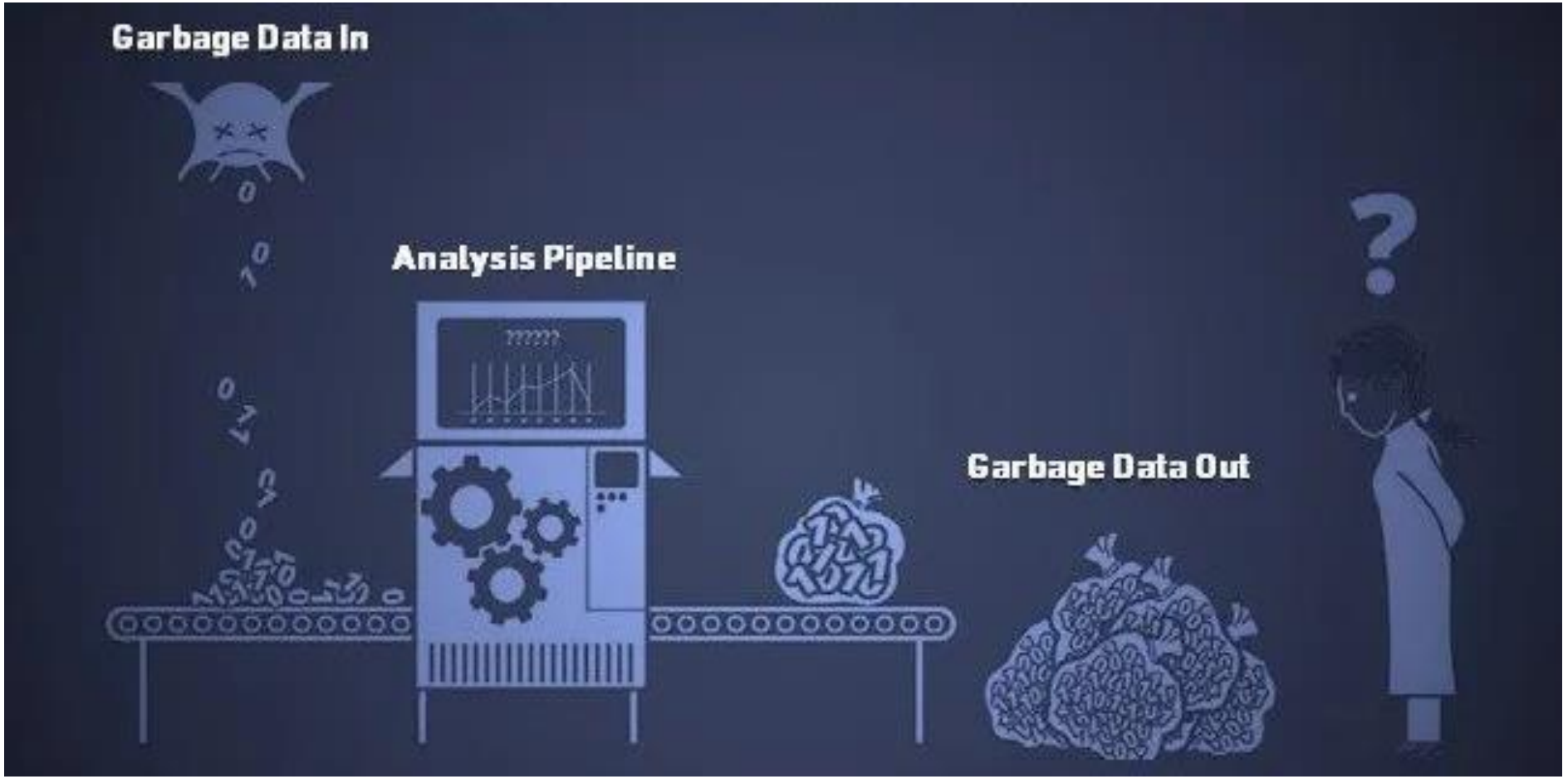
Overview

- Review dimensionality reduction
- Impact of collecting or processing data in batches
  - What can (and cannot) be done to address this computationally
- False positive and false negatives
- Introduce expression workflow

osteoarthritis vs rheumatoid arthritis



Conclusions? Concerns?



All patients affected by RA are female

	osteoarthritis	rheumatoid arthritis
sex: f	10	30
sex:m	20	0

Data collection was done on three days

	osteoarthritis	rheumatoid arthritis
Dec 03 2008	2	10
Dec 04 2008	28	5
Dec 05 2008	0	15

Main concerns and what can we do about it

- All affected patients are female
  - Conclusions about difference between OA and RA will only be applicable to females
  - Since we don’t have any affected male patients, we cannot make any conclusions about this group

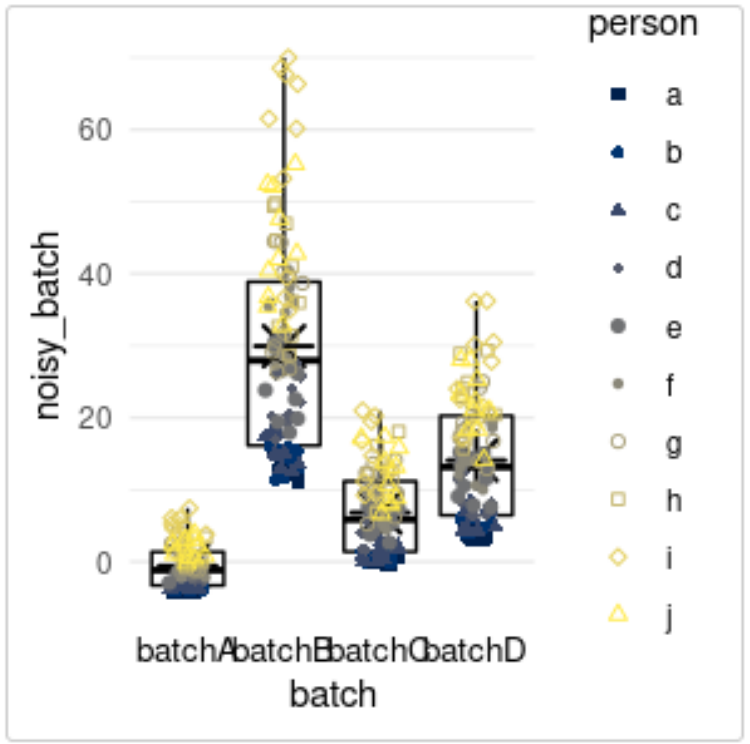
- Data collection was on different days
  - Most collections have samples from both phenotypes – woo hoo!
  - Last collection only has RA patients – best to discard but we are working with limited samples.

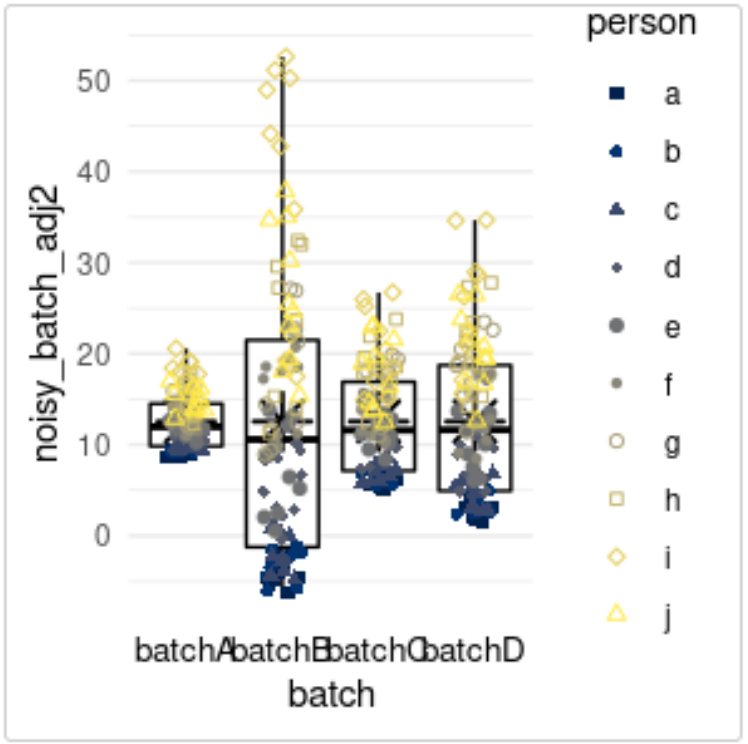
Batch Correction

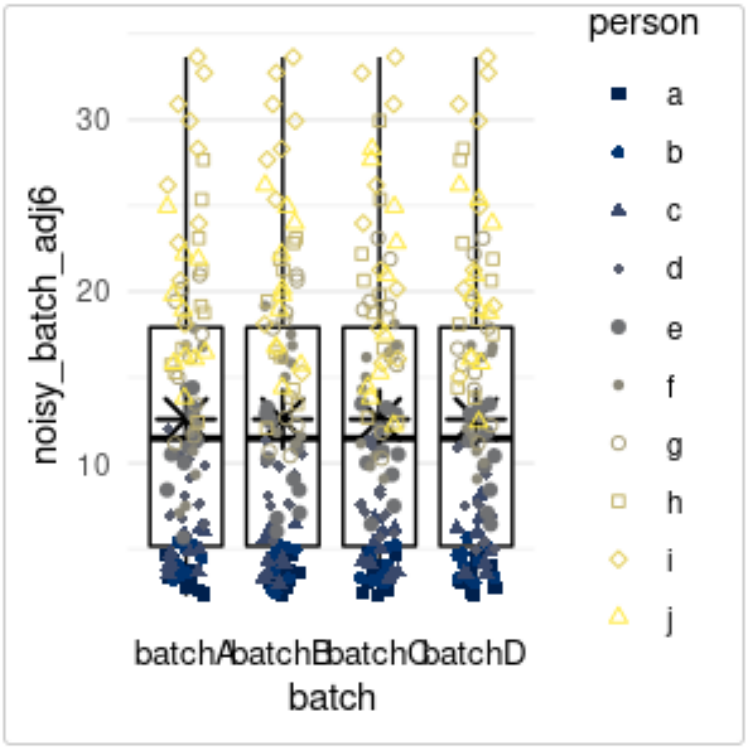
- Compares the means between the two groups and applies adjusts each batch so that each group's mean is equivalent to the overall mean

- Often does not take experimental group into account – differences between variables will persist

Batch Correction







Phenotype representation should be similar in all groups

- Means may (and likely will!) change depending on the phenotype composition

- There are strategies for taking confounders into account for the adjustment

Negative expression data?

- During the correction, some expression values may be negative
- Biologically – this makes no sense!
- Be careful of what statistical tests you are using, some can not tolerate negative or zero values

Batch correction always adds an artificial element of noise to the data

Always best to collect high quality data to minimize informatic transformations required

Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies

Tristan Zindler ✉, Helge Frieling, Alexandra Neyazi, Stefan Bleich & Eva Friedel

BMC Bioinformatics **21**, Article number: 271 (2020) | Cite this article

Tips for applying batch control

- Investigate the data for covariates and batches
  - Get information about the data collection process too!
- Always compare back to the uncorrected dataset
- Visualization tool are important for understanding the changes to your data

False results

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

Differential gene expression

- Identify differences in the transcriptome (gene expression) across samples
  - E.g., differences between multiple biological conditions (e.g. drug treated vs. untreated samples)

- Many possible tools, depending on the type of data & your questions
  - DESeq2 (RNA-seq)
  - limma (microarray)
  - edgeR (RNA-seq)

Type of data: microarray versus RNA-seq

- Microarray: detects expression of thousands of genes
  - Hybridisation method

- RNA-Seq: detects presence and quantity of RNA
  - Higher resolution, better detection of transcriptome
  - More expensive

Wrap-up

- Bioinformatics should not be a correction method or afterthought, it should be considered and built into the plan at the start
- Any computational correction or transformation will add noise to the dataset
- There needs to be balanced representation in each batch – computational batch correction cannot save all datasets
- Be conscious of false positive and false negatives
- Microarrays (expression data) and RNA-seq (sequence counts) both use a bioinformatics workflow