
WELCOME!
Summer School will begin at
8:30 a.m. PDT

8:30-8:35 a.m.	**Introduction**	**Javier Flores**
8:35-9:35	Experimental Design	Damon Leach
9:35-10:35	Missing Data and Imputation	Moses Obiri
10:35-10:45	Networking Break	
10:45-11:45	Unsupervised Learning	Javier Flores

EMSL

Summer School Day 2: Data Science 101

Damon Leach & Javier Flores
Biostatistics & Data Science

07.25.2023

8:30-8:35 a.m.	Introduction	Javier Flores
8:35-9:35	**Experimental Design**	**Damon Leach**
9:35-10:35	Missing Data and Imputation	Moses Obiri
10:35-10:45	Networking Break	
10:45-11:45	Unsupervised Learning	Javier Flores

EMSL

Experimental Design and Simple Tests

Damon Leach

Instructor Intro

Damon Leach

Biostatistician

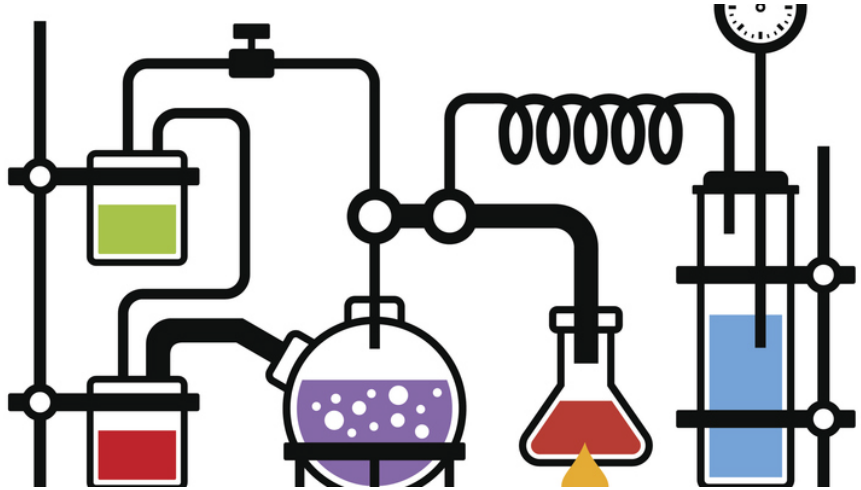
- Biostatistics, Statistics, R, Data Visualization, R Package Development
- Damon.Leach@pnnl.gov
- linkedin.com/in/damon-leach

Outline

- Observational vs Experimental Studies
- Terminology
- Common Experimental Designs
- Simple Statistical Tests
- Non-Standard Experimental Designs

Observational vs Experimental

- Observational Study
  - Conditions influencing response not under the control of investigator
  - Cannot randomly assign treatments
  - Potential confounding
  - Cannot infer cause-and-effect relationship
- Randomized Experiment
  - Can infer cause-and-effect relationship*



General Steps to Experimental Design

1. Choosing a biological system
   - Balancing physiological relevance, reproducibility, and complexity
2. Determining hypotheses to be tested
   - Treatments
   - Time points
3. Selecting data to be generated or collected
4. Determining number of replicates

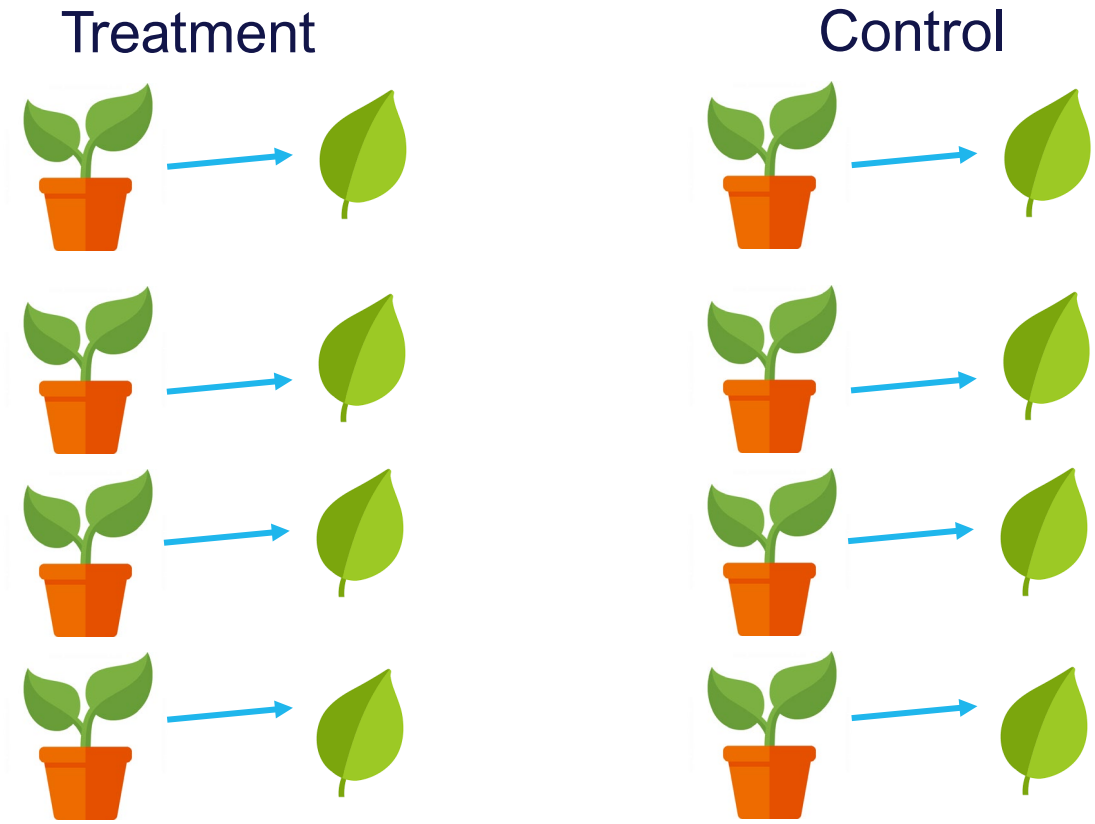
Terminology

- Treatment: the procedures being studied
- Experimental Unit: unit to which a treatment is applied on which we wish to make inference
- Observational Unit: unit on which a measurement is taken
- Response: the outcome that is measured

Terminology cont.

Experimental Unit vs Observational Unit

- If 1:1 e.u. to o.u. mapping → standard statistics



Terminology cont.

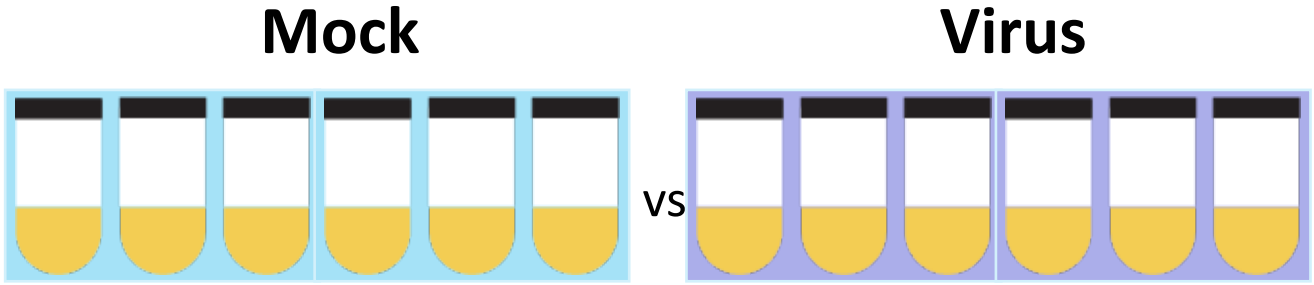
- Randomization: assigning treatments to experimental units in a probabilistic manner
- Control: Baseline treatment (e.g. placebo)
- Confounding: When one factor's effect cannot be distinguished from another factor's effect on response
  - Goal is to reduce this confounding

Major Components of Experimental Design

- Treatment Structure
  - Treatments to be studied
- Design Structure
  - Grouping experimental units into uniform blocks
- Randomization
  - Avoid systematic bias
  - Does not have to be complicated
  - May not know what factors could be confounding

Treatment Structures

One-Way Design (One-Way ANOVA)



Treatment Structures cont.

Two-Way Design (Two-Way ANOVA)

- e.g. Crop (2 levels) and Location (2 levels) – all 4 combinations observed

Crop 1, Location A	**Crop 1, Location B**
Crop 2, Location A	**Crop 2, Location B**

- We can have as many factors as desired for n-way ANOVA, but not always a good idea

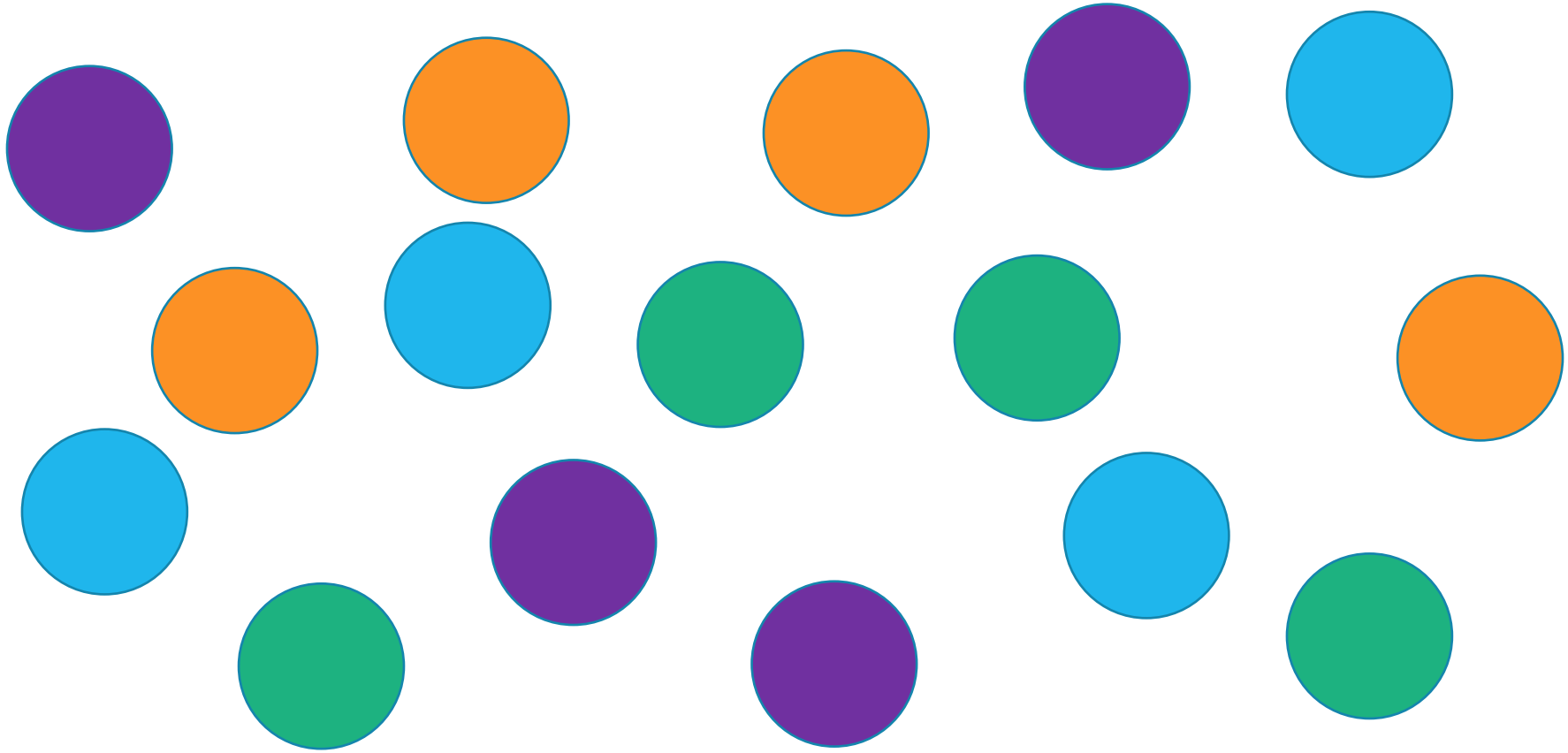
Design Structures

- Completely Randomized Design (CRD)
- Randomized Complete Block Design (RCB)
- Incomplete Block Design
- Latin Square Design

Design Structures

Completely Randomized Design (CRD)

- Treatments applied randomly to experimental units



Design Structures cont.

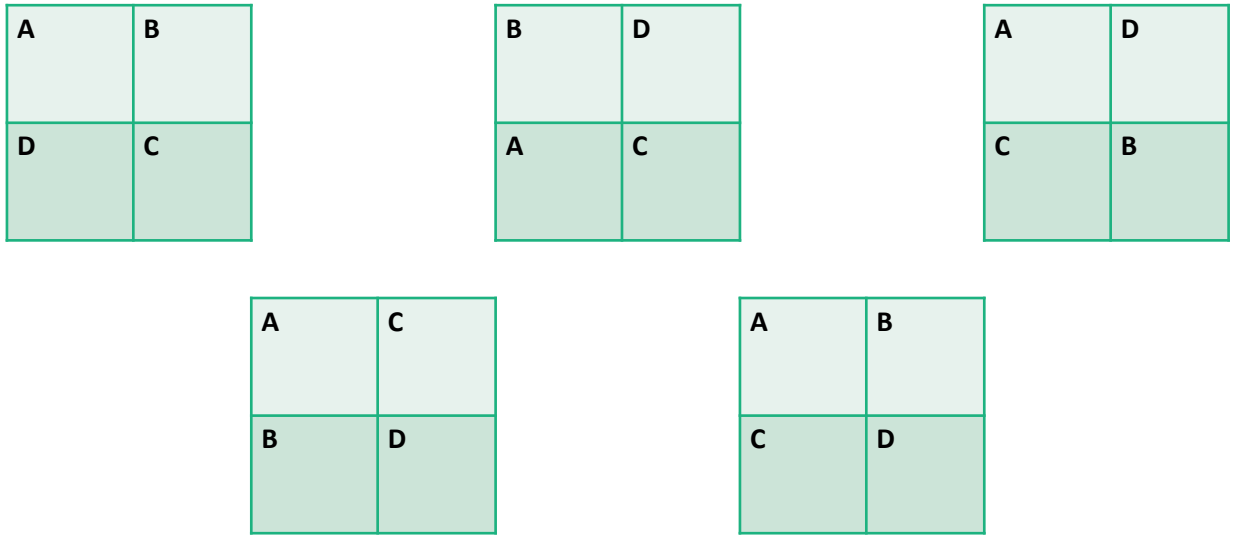
Blocking

- Create sets of groups ***BEFORE*** running experiment
- Reduces variance
- Units within same block expected to have similar responses (in contrast to units not in the same block)
- Used in many design structures

Design Structures cont.

Randomized Complete Block Design (RCB)

- Experimental units grouped into blocks
  - Could be similar traits (age, sex, etc.)
- Different treatments tested randomly assigned to units in each block
- Blocks on one variable



- Can also be an incomplete block design

Design Structures cont.

Latin Square Design

- More restrictive than RBD
- Blocks on two variables
- Total number of blocks is number of treatments
- Each treatment appears only once in each row and column

A	B	C	D
D	A	B	C
C	D	A	B
B	C	D	A

How Many Samples?

- Replication is needed to establish statistical significance in any analysis
  - Increased number of replicates is necessary to guard against loss of statistical significance due to sample-to-sample variation or other technical problems
  - Balance between cost and quality data
- Different types of replicates capture different sources of variability
  - Technical/Injection
  - Biological
- Consider adding more replicates per group rather than a new group to increase statistical power in discovery phase
- Statistical power is quantity typically calculated to help determine the number of replicates that should be run

Why Bother?

- Scientifically Meaningful Effect
  - Want enough samples to detect a scientifically meaningful effect
- Money
  - Undersized study wastes resources by not having capacity to produce useful results
  - Oversized study uses more resources than necessary
- Ethical issues when using live subjects (humans, etc.)
- Grant reviewers are looking for sample size and power calculations

Hypotheses

- Null Hypothesis (what we assume to be true)
  - H_0
  - “There is no difference between treatments”
- Alternative Hypothesis (what we want to show)
  - H_a
  - “There is a difference between treatments”
- “Innocent until proven guilty”
  - Assume H_0 is true until/unless we have enough evidence in the data in favor of H_a

Hypotheses cont.

- H_0 & H_A stated in terms of some population parameter, p
- Different types of hypotheses, e.g.

2-sided	1-sided	1-sided
$H_0: p = p_0$	$H_0: p \geq p_0$	$H_0: p \leq p_0$
$H_A: p \neq p_0$	$H_A: p < p_0$	$H_A: p > p_0$

- Note:
  - "=" sign always included in H_0
  - H_0 & H_A must contradict each other

P-value

- The probability that an observed outcome is due to chance
  - High p-value: more likely to be due to chance
  - Low p-value: less likely to be due to chance
- Threshold often to be 0.05
  - P-value > 0.05 -> fail to reject the null hypothesis
  - P-value < 0.05 -> Reject the null hypothesis
  - ***NEVER*** accept the null hypothesis

Errors in Hypothesis Testing

- Rejecting H_0 in favor of H_A does not guarantee that H_A is true despite very strong evidence
- For any hypothesis test, there are 2 kinds of errors we can make

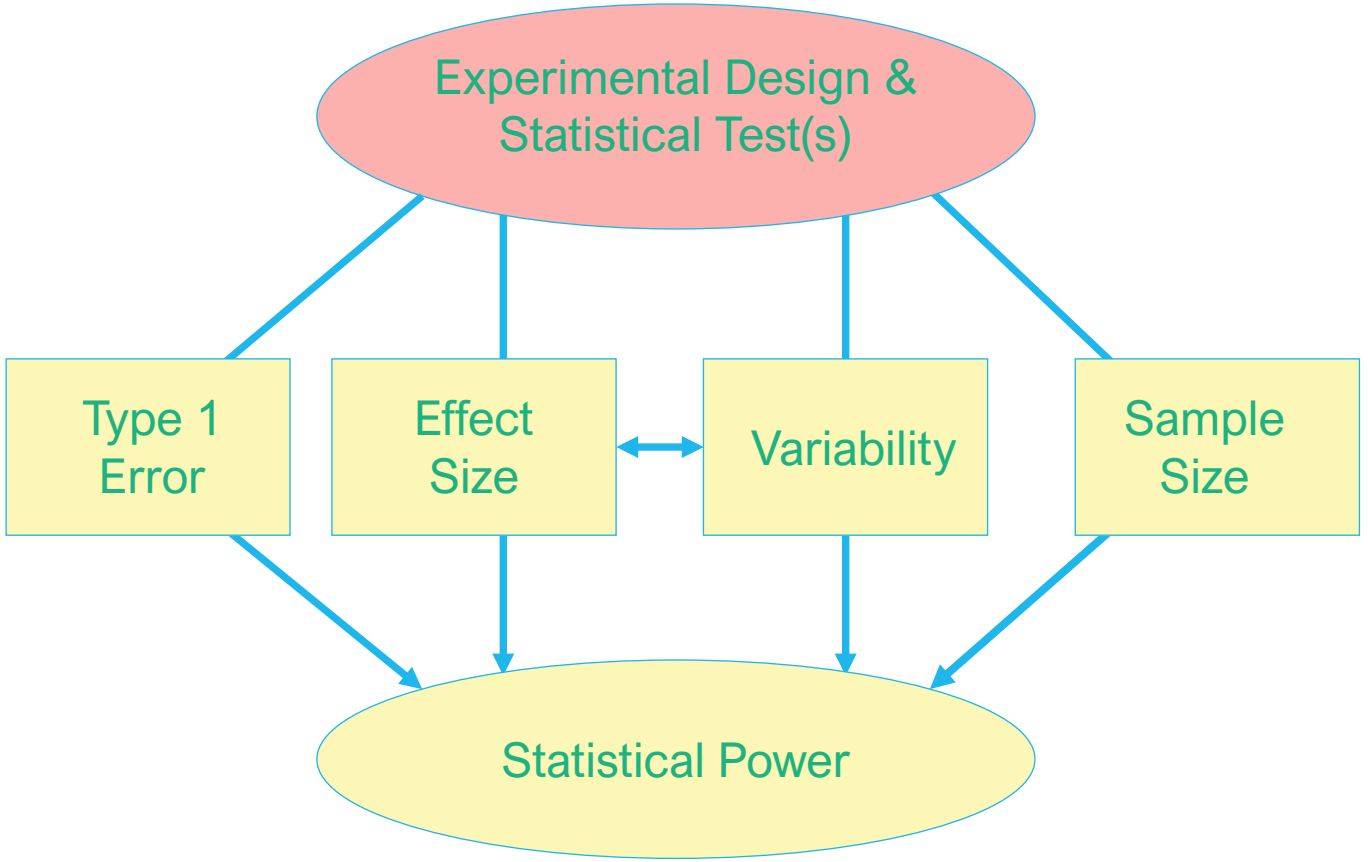
		Decision	
		Fail to reject H_0	Reject H_0
Truth	H_0	Correct decision	Type I error = α (false positive)
	H_A	Type II error (false negative)	Correct decision (power)

Errors in Hypothesis Testing cont.

- By construction, hypothesis testing limits the rate of Type I errors (false positives) to a significance level, α
  - We choose ahead of time what significance level we will test at
  - Multiple tests on different attributes of the same data require an adjustment in order to preserve the significance level
    - E.g. testing the salmon for levels of multiple chemicals requires an adjustment
    - The more tests we do, the more likely we are to find a false positive
- Type II error rate (β) is a function of sample size, significance level, & effect size
- Trade-off between Type I and II error rates

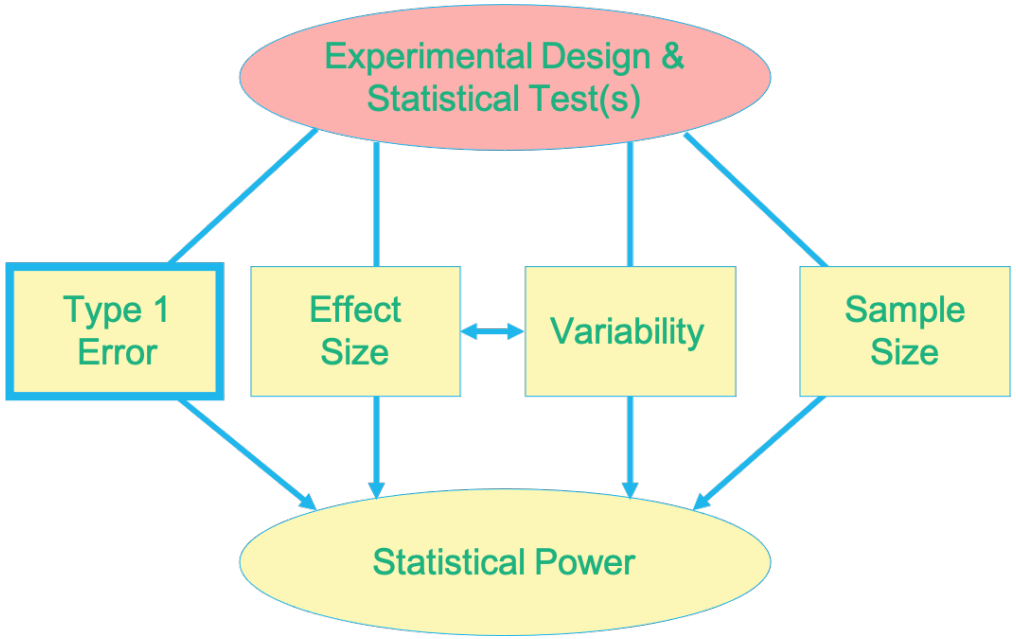
What is Statistical Power

- Statistical Power (1-β) – probability of rejecting the null hypothesis, when the alternative hypothesis is true



Type I Error

- Tradeoff between Type 1 and 2 errors
  - Inversely related
- Lower Type 1 error → Lower power
  - Holding other factors constant
- Typical values: 0.05, 0.1
- Domain dependent

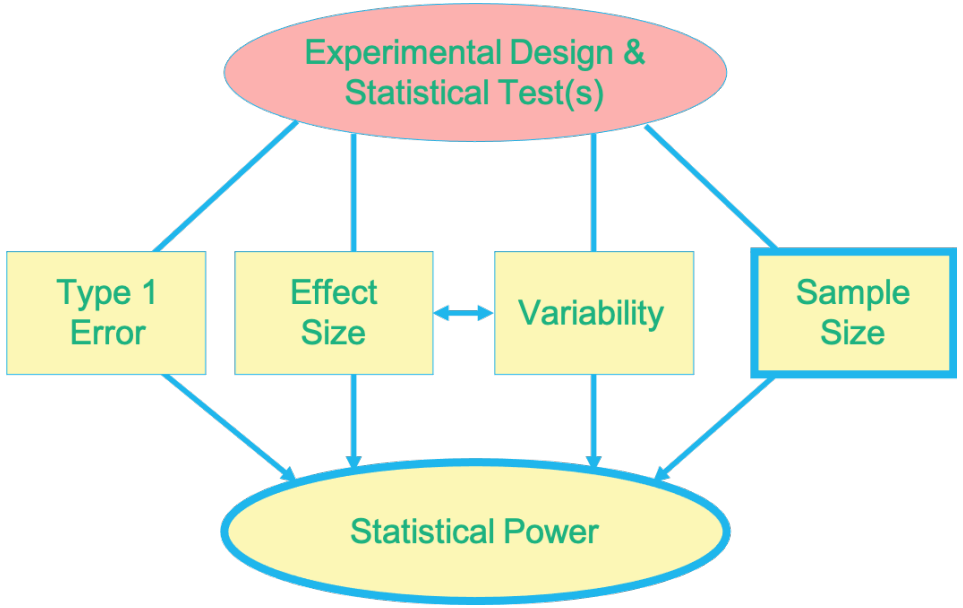


Effect Size and Variability

- Effect Size – desired detectable difference (if a difference exists)
- Variability – variability of the parameter being estimated
  - If evaluating difference in means → variance of values from same group
- Ratio of Effect Size/Variability determines power
- More variability → less power
- Smaller effect size → less power
- Typical values?
  - Determined by data and/or domain knowledge

Sample Size and Power

- Usually the quantities we want to estimate
  - If I have X number of samples, what is my power?
  - How many samples do I need for 80% power?
- Most domains, we have one estimate of variability → one power calculation
- Biology ‘omics data often have multiple response variables → multiple power calculations



Simple Tests

Quantitative vs Qualitative

- Quantitative Test
  - Are there differences in the mean abundance of each biomolecule between the treatments/groups?
- Qualitative Test
  - Are patterns of presence/absence for each biomolecule associated with treatment/group?

1	Group1_1	Group1_2	Group1_3	Group1_4	Group2_1	Group2_2	Group2_3	Group2_4
A	16.4	16.9	16.2	16.7	16.9	17.2	17.5	17.9
B	NA	NA	NA	NA	17.5	16.9	17.3	17.1

Quantitative Tests

ANOVA (F-Test)

F-test

$H_0: \mu_a = \mu_b = \mu_c$
$H_A: At\ least\ one\ \mu_i \neq \mu_j$

Post-Hoc Pairwise Comparisons

$H_0: \mu_a = \mu_b\ (\mu_a - \mu_b = 0)$
$H_A: \mu_a \neq \mu_b$

p-value

log FC = $\log \frac{\mu_a}{\mu_b}$

$H_0: \mu_a = \mu_c$
$H_A: \mu_a \neq \mu_c$

p-value

log FC = $\log \frac{\mu_a}{\mu_c}$

$H_0: \mu_b = \mu_c$
$H_A: \mu_b \neq \mu_c$

p-value

log FC = $\log \frac{\mu_b}{\mu_c}$

Assumptions

- Assumptions
  - Independence
  - Constant variance
  - Normality
- If assumptions not met?
  - Potentially transform the data
  - Boxcox transformation

Outliers

- Response is extreme in comparison to other responses with similar predictors
- Use caution when removing outliers

Multiple Comparisons

- Type I vs Type II error rate
- Reduce the Type I error rate with multiple comparisons analyses
- If you run many many tests it is likely that something will be found significant at 5% threshold (even if it really isn't)
- Bonferroni, tukey, dunnet, etc.
- No best procedure, just best for your situation (exploratory or confirmatory, etc, working with just comparisons to control, etc., some require to be pre-planned, others are post-hoc)

Quantitative Tests cont.

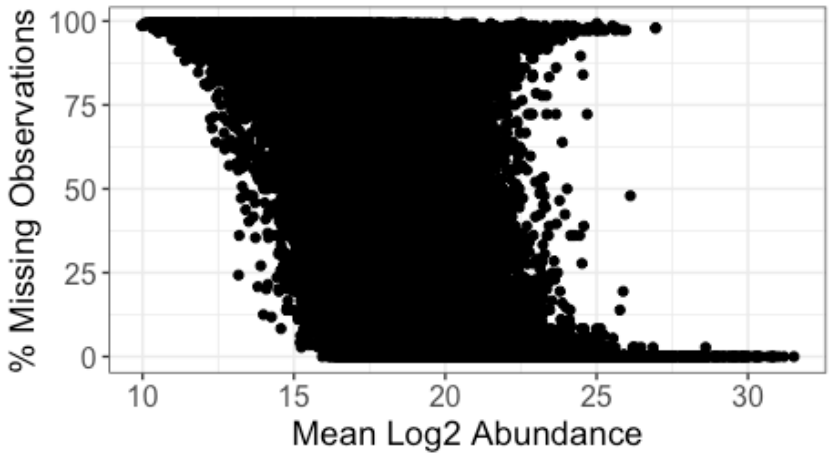
- (adjusted) p-values
  - Using p-value $\leq \alpha$ →$\leq \alpha$% Type 1 error rate
  - Multiple tests → error rate is inflated
  - Multiple tests for a biomolecule

Method Name	Appropriate Comparison	ANOVA	IMD
Bonferroni	Both	√	√
Dunnett	Case-vs-control	√	
Holm	Both	√	√
Tukey	All pairwise	√	

Qualitative Tests

G-Test

- What if we don't have enough observations to conduct a quantitative test?
- Determine if proportion of missing values are associated with treatment group, compared to random chance



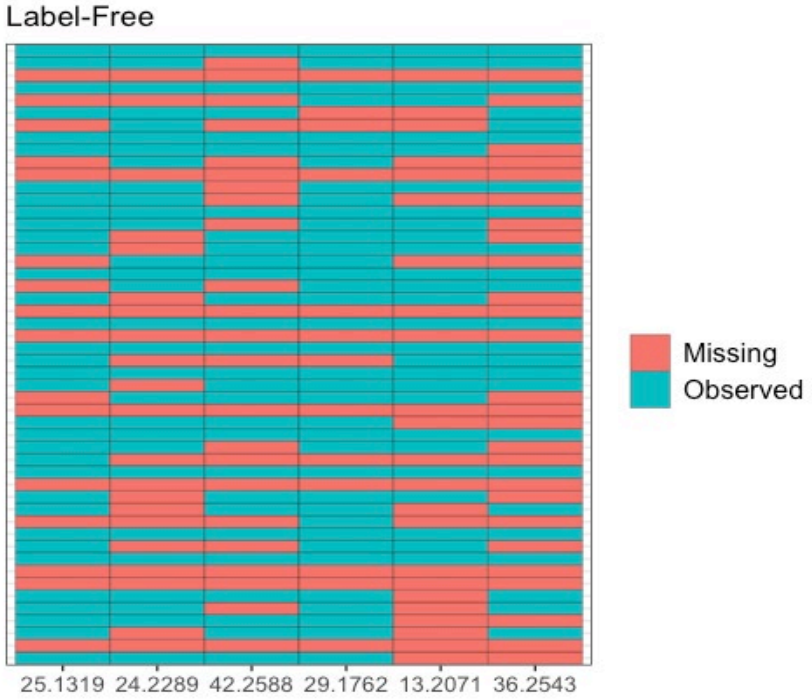
	Present	**Absent**	**Total**
Treatment A	1	4	**5**
Treatment B	3	1	**4**
Total	**4**	**5**	

Qualitative Tests cont.

G-Test

- Determine if proportion of missing values are associated with treatment group, compared to random chance
- Fisher's test of independence with correction for small sample size

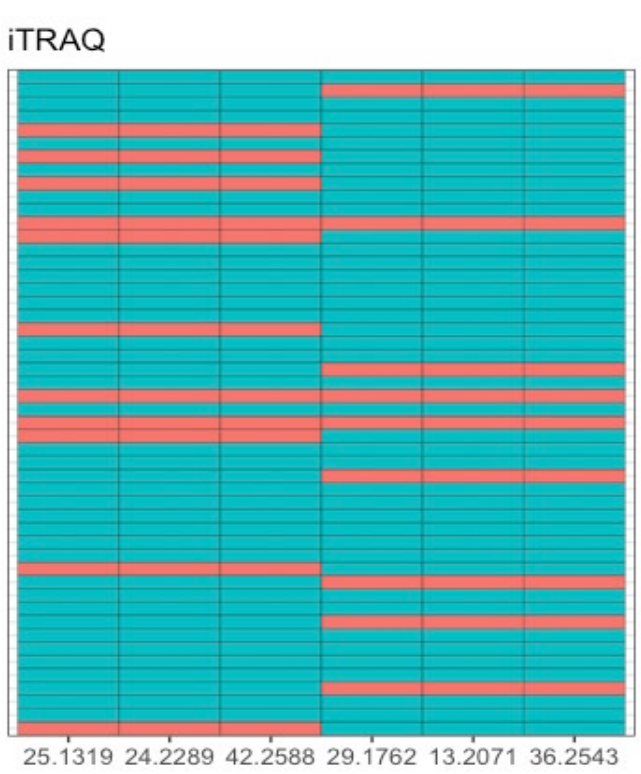
	Present	**Absent**	**Total**
Treatment A	1	4	**5**
Treatment B	3	1	**4**
Total	**4**	**5**	

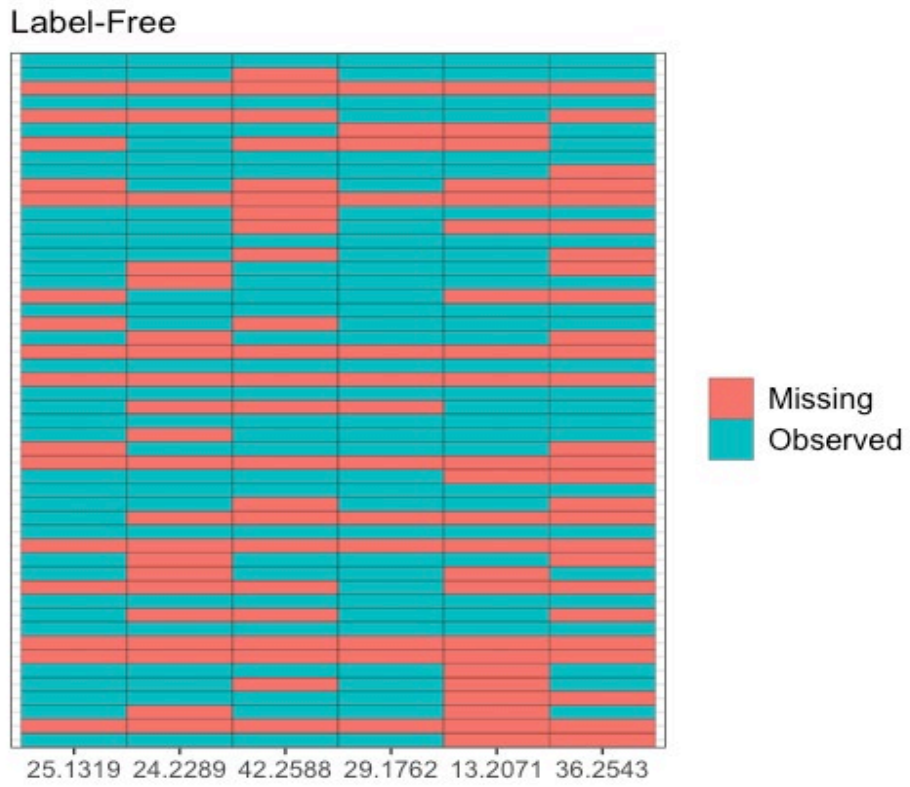


Qualitative Tests cont.

G-Test

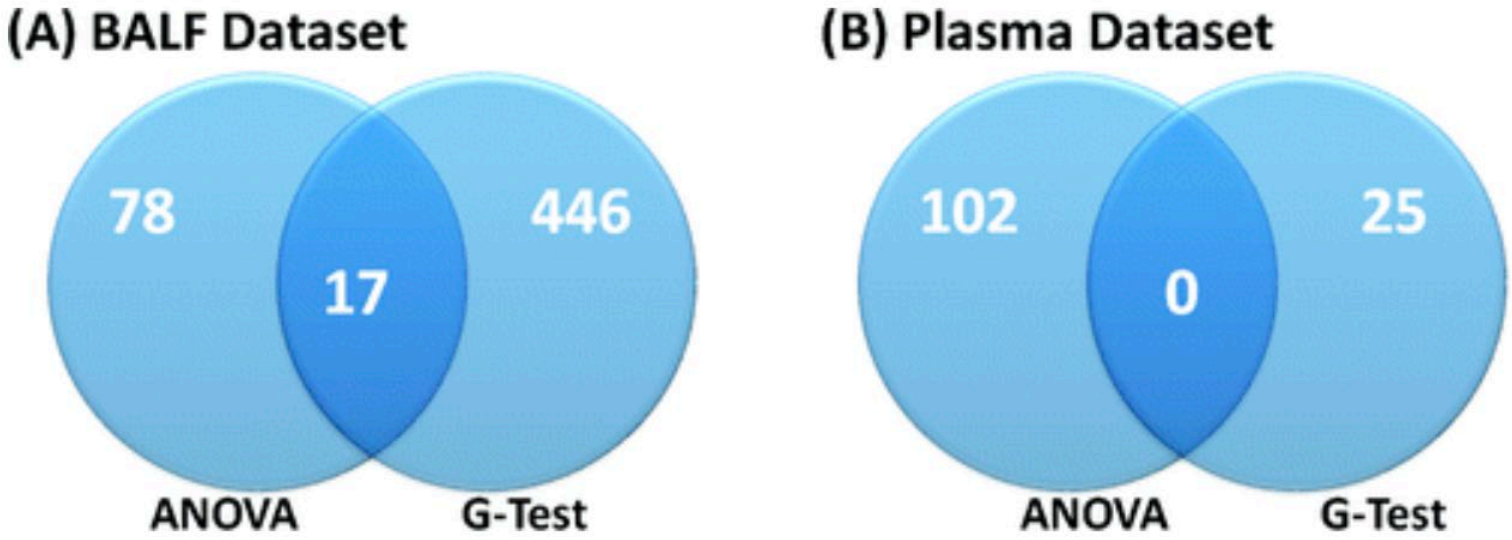
- n ≥ 3 replicates per treatment group
- Not typically applicable for metabolomics data
  - Not applicable for isobaric labeled proteomics data





Using Both Quantitative and Qualitative Metrics

- ANOVA and then G-test



Webb-Robertson et al (2010). *Journal of proteome research*.

Non-Standard Experimental Designs

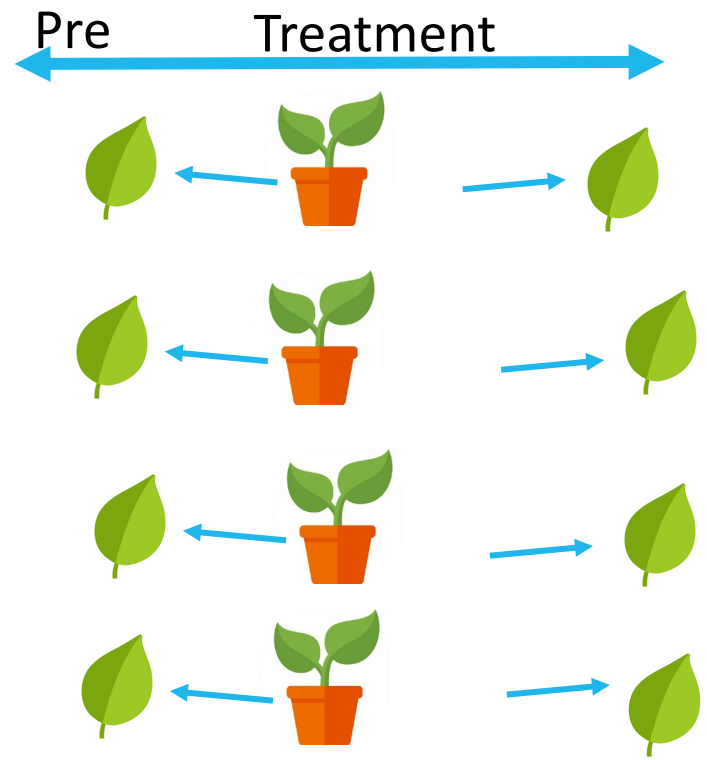
Seek a statistician when…

- **Before you plan your experiment!**
- Non-independence
  - Time course
  - Repeated measures (Mouse litters)
  - Numerous samples from one soil core
- Data not normally distributed
  - Censored data (survival analyses)

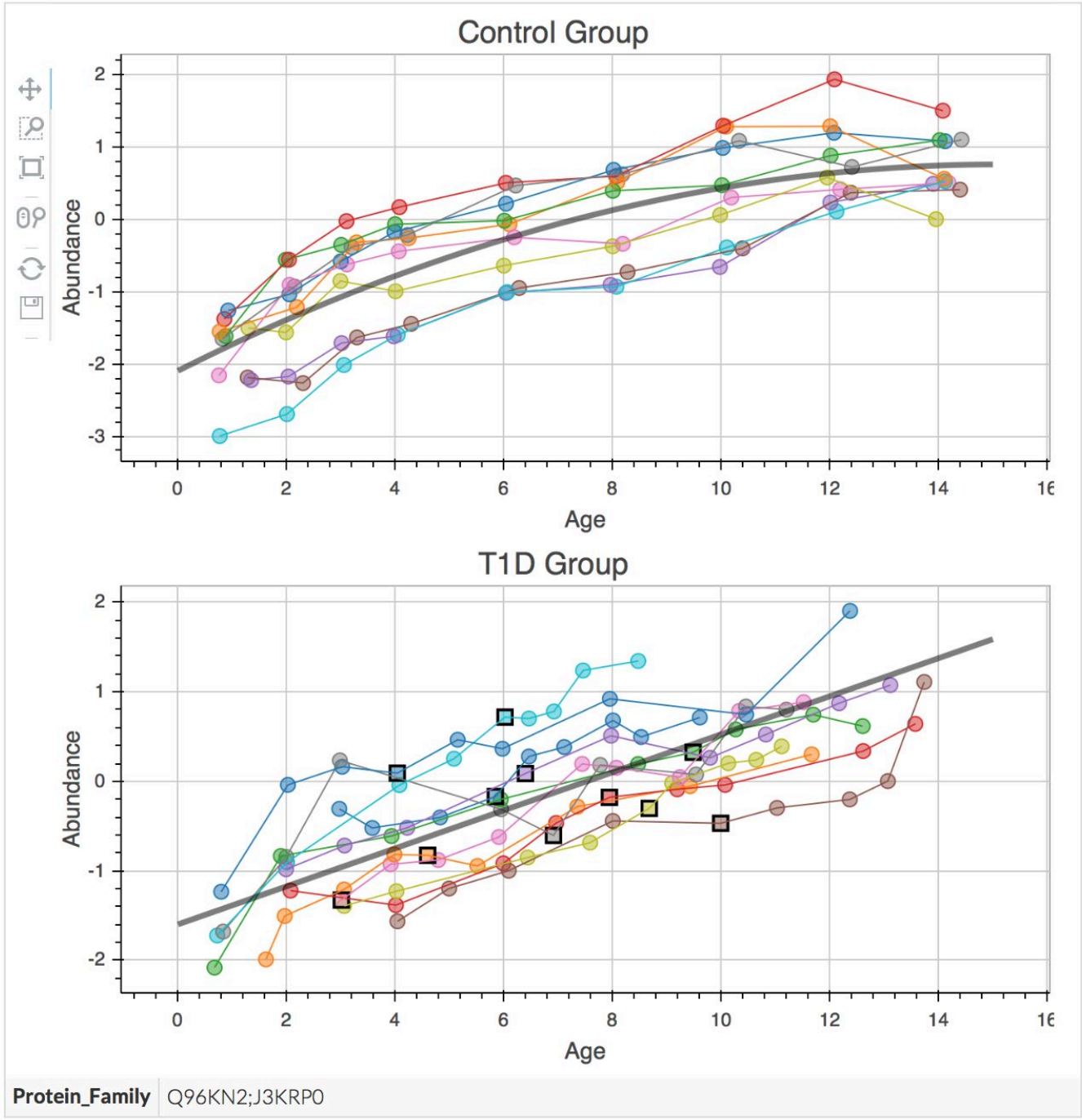
Non-Standard Experimental Designs cont.

Paired Study

- If before, after study → standard paired statistics



Non-Standard Experimental Designs cont.



Main Takeaways

- Experiments are only as valid as their design
  - An experiment with proper design can help detect cause-and-effect relationships
- There are quantitative and qualitative tests for analyses
- Replication is essential for statistical significance
- Seek a statistician before starting an experiment

Other Considerations

- What if my samples sizes are already determined?
  - Calculate power based on expected effect sizes
    - What power will I have to detect a two-fold change in mean expression?
  - Calculate detectable effect size based on required power
    - Given sample sizes and 80% power, we can detect 2.5 fold-change
    - Is this effect size in the realm of what is reasonably expected from the data?
- What if I don’t have preliminary experiment data?
  - Power can be calculated for some hypothesis tests under a “worst case scenario”
    - E.g. tests in proportions
  - Identify variability values from literature
  - Use data from other study with closest sample properties

Other Considerations cont.

- Will you have missing data (e.g. proteomics)?
  - Power calculations assume no missing data
- Are you testing more than two groups?
  - Multiple test correction
- Hypotheses not dealing with means (e.g. trend analysis over time)
  - Variability is not straightforward calculation
  - Example data is key
- Tips
  - Contact your favorite statistician before you plan replicates
  - Groups vs Replicates
    - If determining how to allocate resources - More replicates per group, rather than adding more groups

Thank You!

8:30-8:35 a.m.	Introduction	Javier Flores
8:35-9:35	Experimental Design	Damon Leach
9:35-10:35	**Missing Data and Imputation**	**Moses Obiri**
10:35-10:45	Networking Break	
10:45-11:45	Unsupervised Learning	Javier Flores

EMSL

Missing Data and Imputation

Jen Huckett, PhD
Data scientist/Statistician

Agenda

Agenda

- Missing data
  - Nature of missingness
  - Implications for inference based on missing data
- Imputation
  - Common imputing methods & considerations
  - Implications of using imputed data in subsequent inference
  - Multiple imputation

Up Front

Definitions

- What is “missing data”?
  - No data value is stored for the variable in an observation
  - Various reasons for & types of missing data
- What is “imputation”?
  - Filling in the missing data
  - Numerous methods, selection depends on type(s) of missing data
  - Multiple imputation is best practice
    - Impute (result is *m* complete data sets)
    - Analyze (leading to *m* analysis results)
    - Pool *m* results

https://en.wikipedia.org/wiki/Missing_data

https://en.wikipedia.org/wiki/Imputation_(statistics)

Missing Data

What is missing data?

- Unobserved, not captured, not applicable, NA, etc.
- Often appear as blank cell in a table or as NA, N/A, n/a,…

Education	Number Earners	No Earners	Work Hours	Work Months	Earnings	Retirement	Interest	Assistance
4	1	0	55	12	84	0.7	0.2	0
4	2	0	40	12		0	0	0
4	1	0	8	11	85	12	5	0
4	1	0	75	12	135	0	0.1	0
4	1	0	43	0	0	NA	NA	NA
4	2	0	40	12	92	0	1.5	0
2	0	1	0	0	0	NA	0	0
4	1	0	40	7	35	0	0.65	0
1	0	1	0	0	0	NA	NA	2
4	1	0	30	8	14	0	1	0
4	2	0	35	12		0	0	10
2	1	0	5	4	5	NA	0.5	0
2	2	0	40	6	12	0	0	NA
3	1	0	40	12	25	0	0.1	0

Each row is an observation (a.k.a. record, case, unit,…)

Highlighted cells represent the missing values

Missing Data

- The Nature of Missingness

- Missing completely at random (MCAR)
  - Missingness is independent of observed and unobserved variables
  - Analysis of complete cases (throwing out cases with NAs) produces unbiased results
  - Strong and often unrealistic assumption
- Missing at random (MAR)
  - Missingness is systematically related to observed data
  - Analysis of complete cases may or may not produce unbiased results
  - Proper accounting for known factors can produce unbiased results
- Missing not at random (MNAR)
  - Missingness is systematically related to unobserved predictors or the missing value itself (e.g., censoring)
  - Analysis of complete cases may produce unbiased results but more likely results will be biased

http://www.stat.columbia.edu/~gelman/arm/missing.pdf

Missing Data

Path Forward

- Analysis using complete cases only (no imputation)
  - Complete-case analysis
  - Available-case analysis
  - Nonresponse weighting

Imputation

Path Forward (cont.)

- Imputation to enable analysis methods that retain all observations
  - Mean imputation
    - Overall mean
    - Means of subpopulations
    - Regression predictions
    - Matching & hot deck
  - Indicator that indicates missingness of predictors
    - Additional category within existing categorical variable
    - Additional variable associated with continuous predictor)
  - Random imputation of single or multiple variables
    - Draw from a probability distribution
    - Regression + error drawn from a probability distribution

Imputation

Overall mean imputation

- Impute earnings = average(84, 85, 135, 0, 92, 0, 35, 0, …) = 52.2
- Impute retirement = average(0.7, 0, 12, 0, 0, 0, 0, …) = 2.2
- Impute interest = average(0.2, 0, 5, 0.1, 1.5, 0, 0.65,…) = 1.7
- Impute assistance = average(0, 0, 0, 0, …) = 0.7

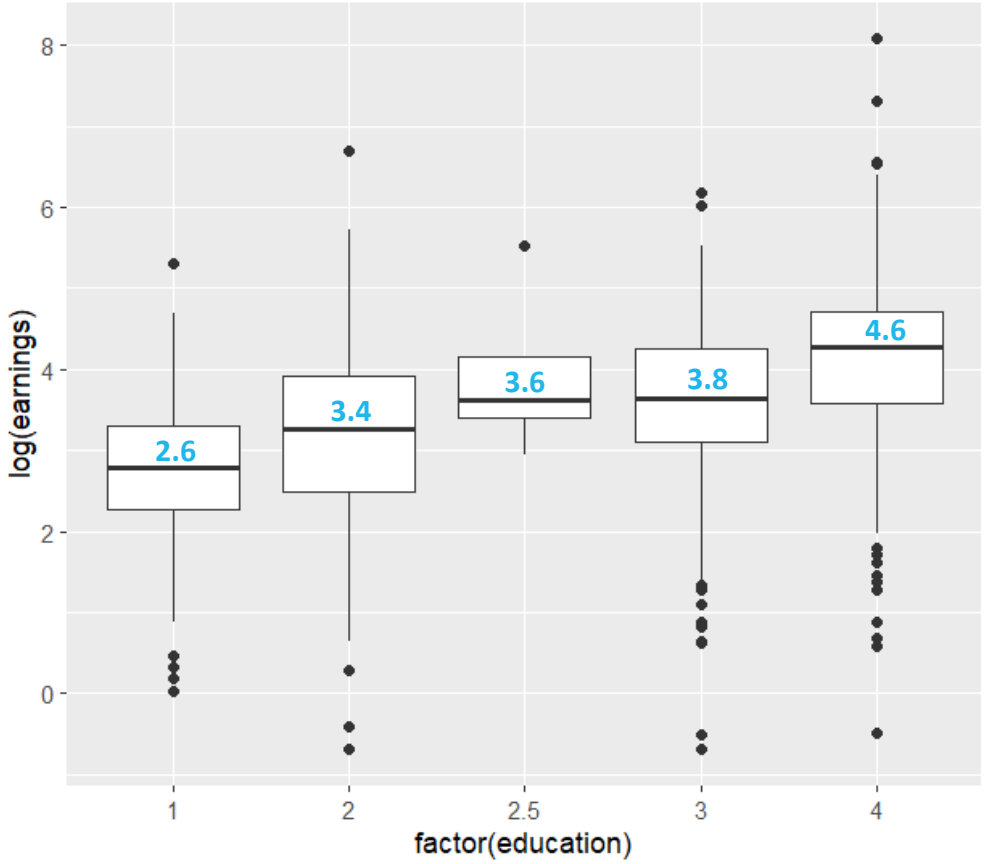
Education	Number Earners	No Earners	Work Hours	Work Months	Earnings	Retirement	Interest	Assistance
4	1	0	55	12	84	0.7	0.2	0
4	2	0	40	12	41	0	0	0
4	1	0	8	11	85	12	5	0
4	1	0	75	12	135	0	0.1	0
4	1	0	43	0	0	2.2	1.7	0.7
4	2	0	40	12	92	0	1.5	0
2	0	1	0	0	0	2.2	0	0
4	1	0	40	7	35	0	0.65	0
1	0	1	0	0	0	2.2	1.7	2
4	1	0	30	8	14	0	1	0
4	2	0	35	12	41	0	0	10
2	1	0	5	4	5	2.2	0.5	0
2	2	0	40	6	12	0	0	0.7
3	1	0	40	12	25	0	0.1	0

Assign overall average of observed values for each variable to fill in missing values

Imputation

Subpopulation mean imputation

- Overall mean earnings did not account for increases with education



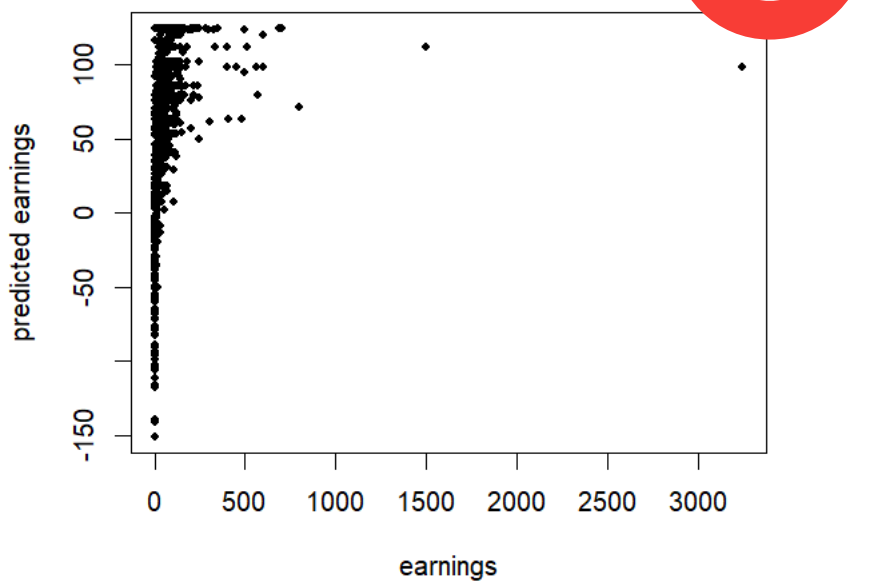
Assign average of observed values within education level for each variable to fill in missing earnings values

Imputation

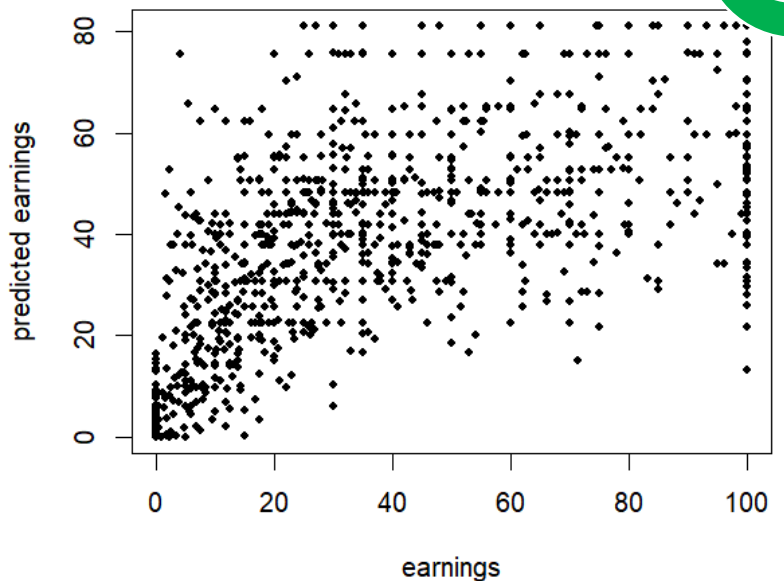
Regression prediction imputation

- Means did not account for info available from other variables
- Regression model could!
- *earnings = f(male, over65, white , immig, educ, workmos, workhrs, assistance)*

- Linear scale
- R^2=0.08



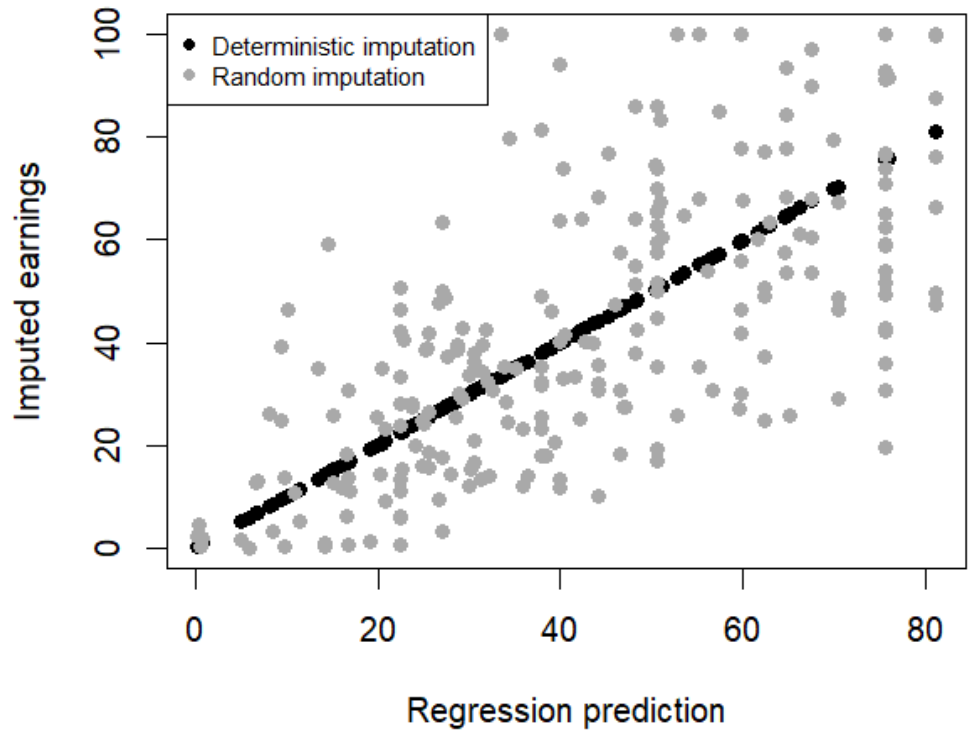
- Square root scale and top coding
- R^2=0.44



Imputation

Regression predictions as imputed values

- Impute predicted earnings to fill in missing earnings (deterministic imputation)
- What about uncertainty in model & predictions?
- Draw imputed values from a normal distribution (random imputation)
  - mean = predicted value
  - standard deviation = prediction standard error



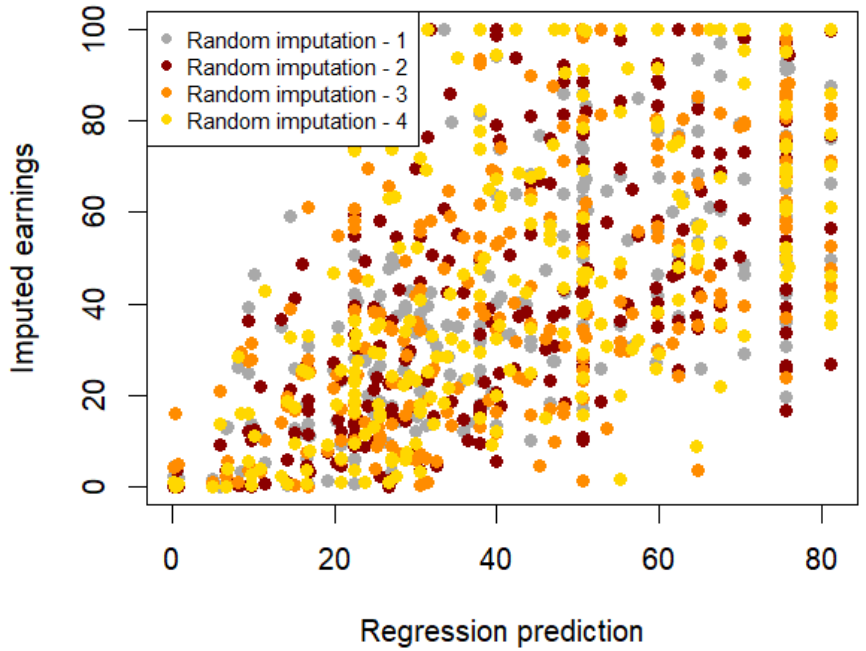
Considerations

- Imputation model matters
  - Simple models shown today
  - More complex models and algorithms can be used
    - Multistage imputation: combine logical rules with one or more imputation methods
    - Multivariate or iterative imputation to impute multiple variables
- Incorporate uncertainty in imputed values
  - Which realization of random imputations should I use?
  - All of them (multiple imputation)

Multiple Imputation

What is multiple imputation?

- Rather than replacing each missing value with one randomly imputed value, replace each with several imputed values reflecting uncertainty about the imputation model
- Analyze each complete dataset
  - Analysis results will vary, reflecting impacts of imputation uncertainty
  - Which results should I use?
  - Use them all to account for variation within & between imputed data & results



- Example:
  - Impute M=5 complete datasets (m=1, 2, 3, 4, 5)
  - Apply regression analysis to each: estimate coefficients $\hat{\beta}_m$ and standard errors s_m for each m=1,…,5 dataset & analysis
  - Overall estimate: $\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m$
  - Variance estimate: $V = W + \left(1 + \frac{1}{M}\right)B$ where $B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}_m - \hat{\beta})$ and $W = \frac{1}{M}\sum_{m=1}^{M} s_m^2$

Proteomics Example

Data Processing

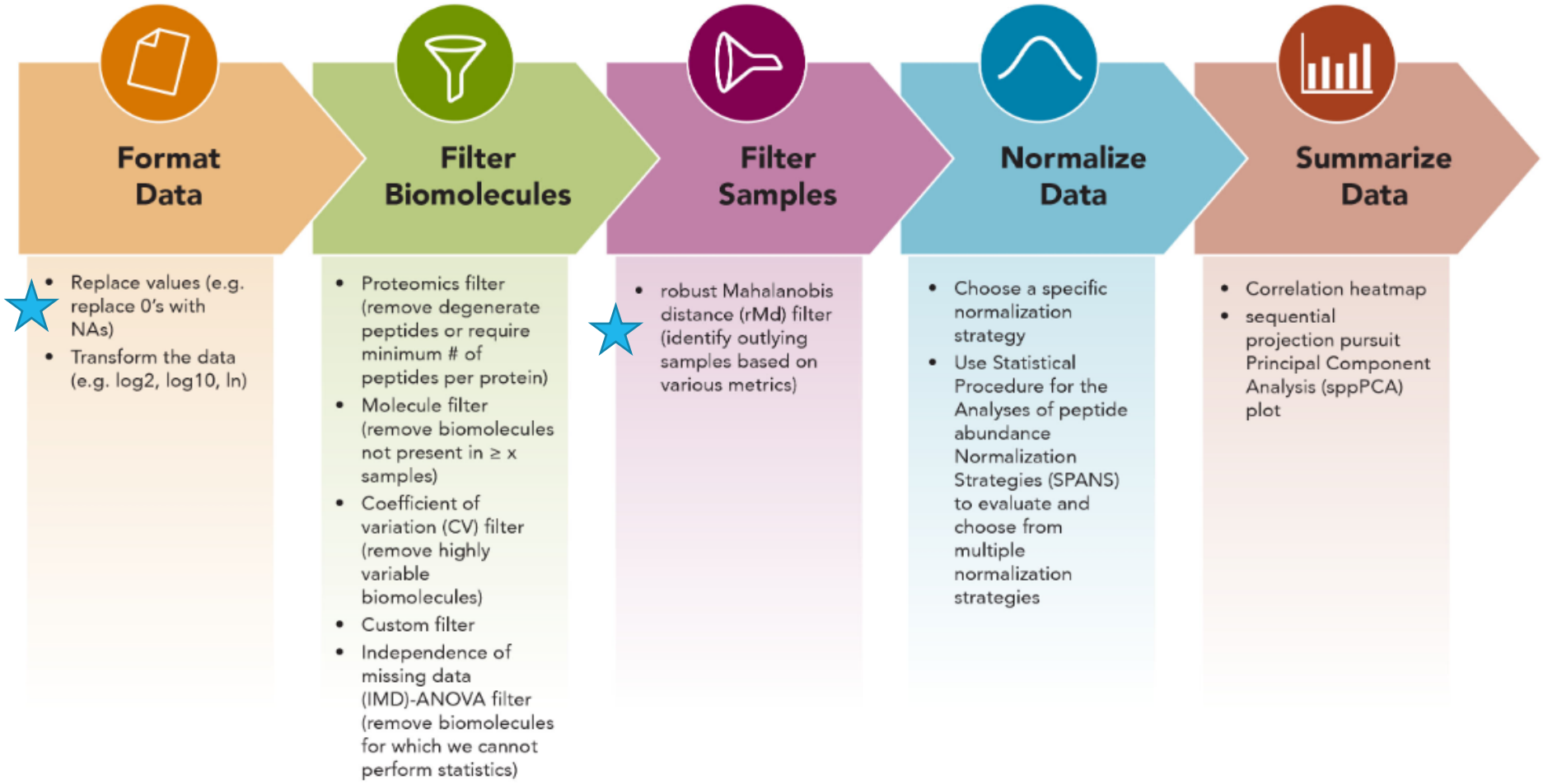


Figure 1. Quality control and processing workflow in pmartR package.

Stratton, K., Bramer, L. 2023. Typical Processing Workflow.
https://pmartr.github.io/pmartR/articles/Typical_Processing_Workflow.html

Data Processing

Summarize peptide data

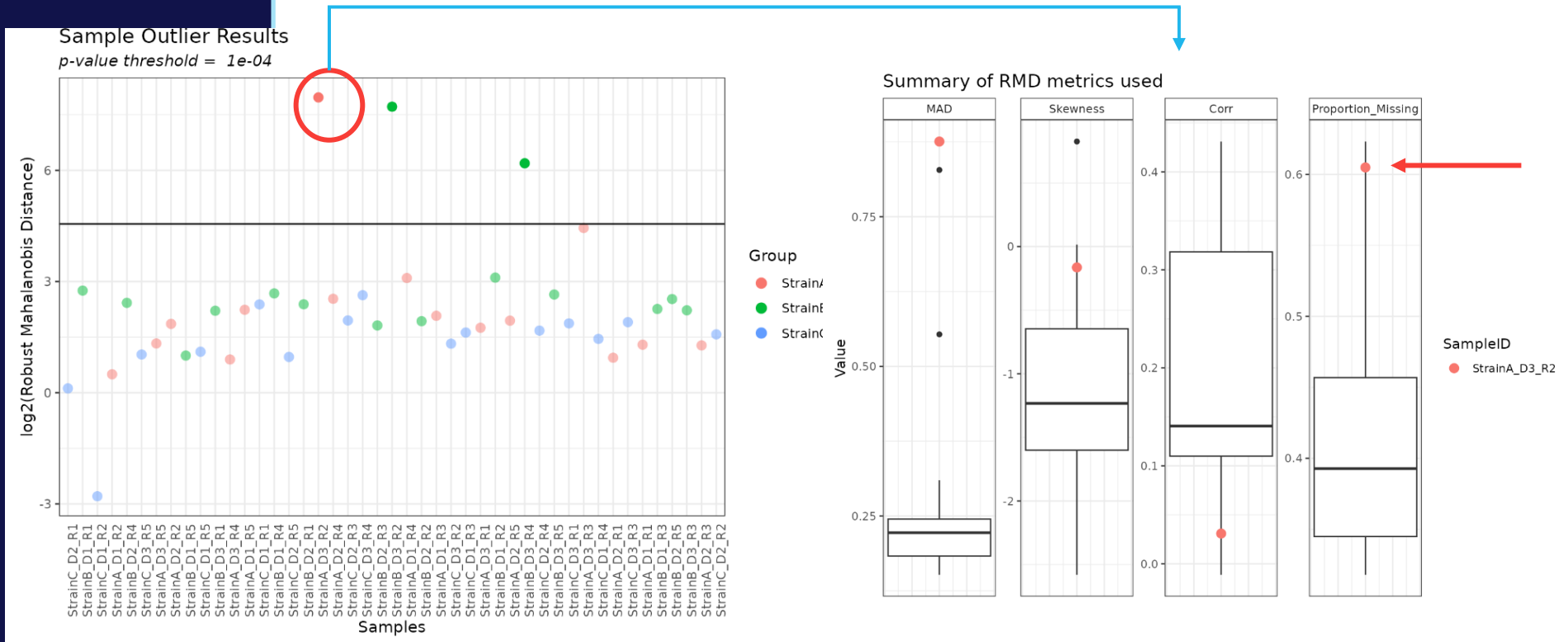
- # unique sample IDs
- # unique proteins
- # unique peptides
- # missing
- Proportion missing
- Samples per group

```
## Class                          isobaricpepData
## Unique SampleIDs (f_data)                   45
## Unique Peptides (e_data)                215220
## Unique Proteins (e_meta)                 16259
## Missing Observations                   5541265
## Proportion Missing                       0.572
## Samples per group: StrainC                  15
## Samples per group: StrainB                  15
## Samples per group: StrainA                  15
```

Stratton, K., Bramer, L. 2023. Typical Processing Workflow.
https://pmartr.github.io/pmartR/articles/Typical_Processing_Workflow.html

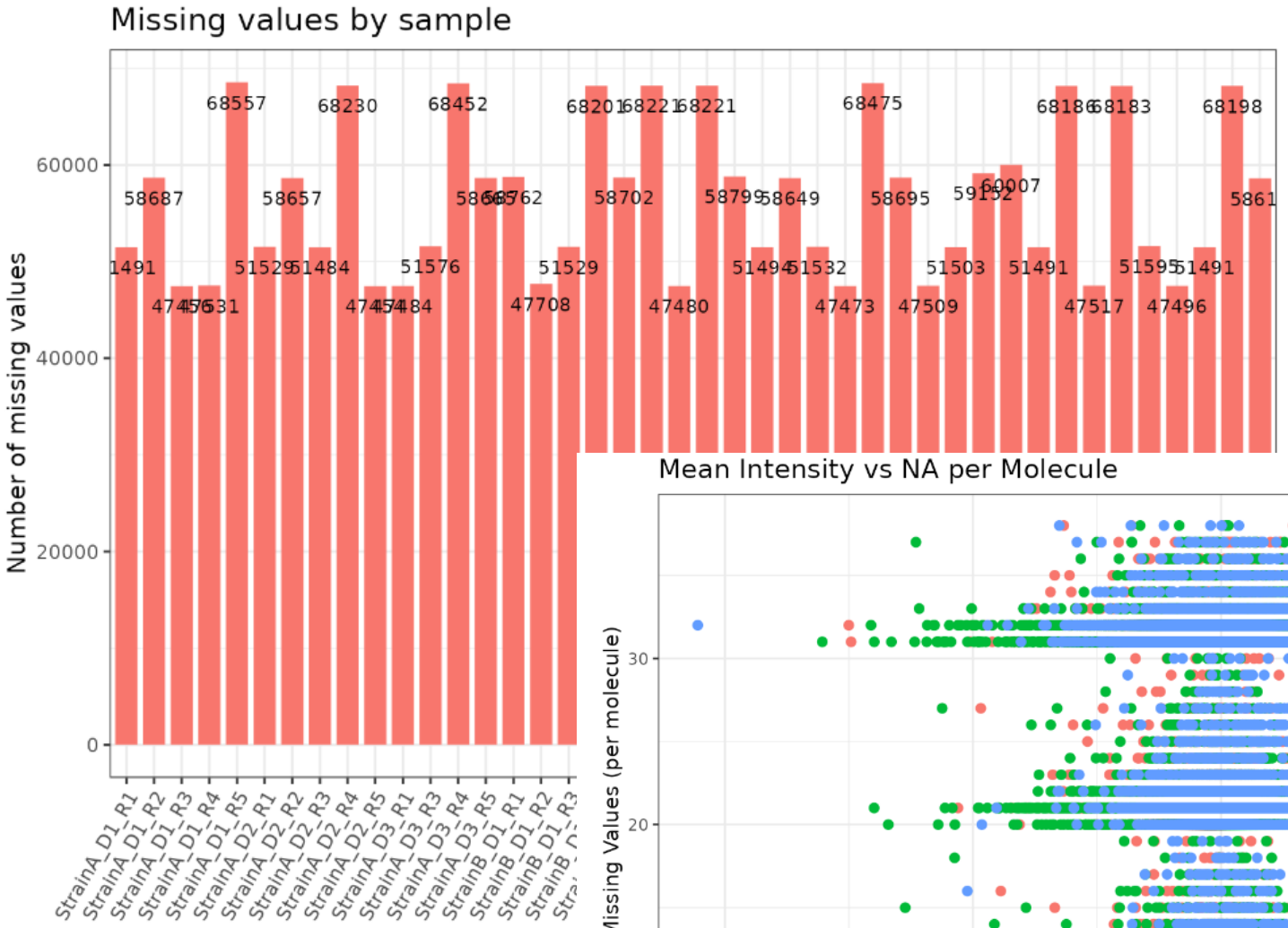
Data Processing

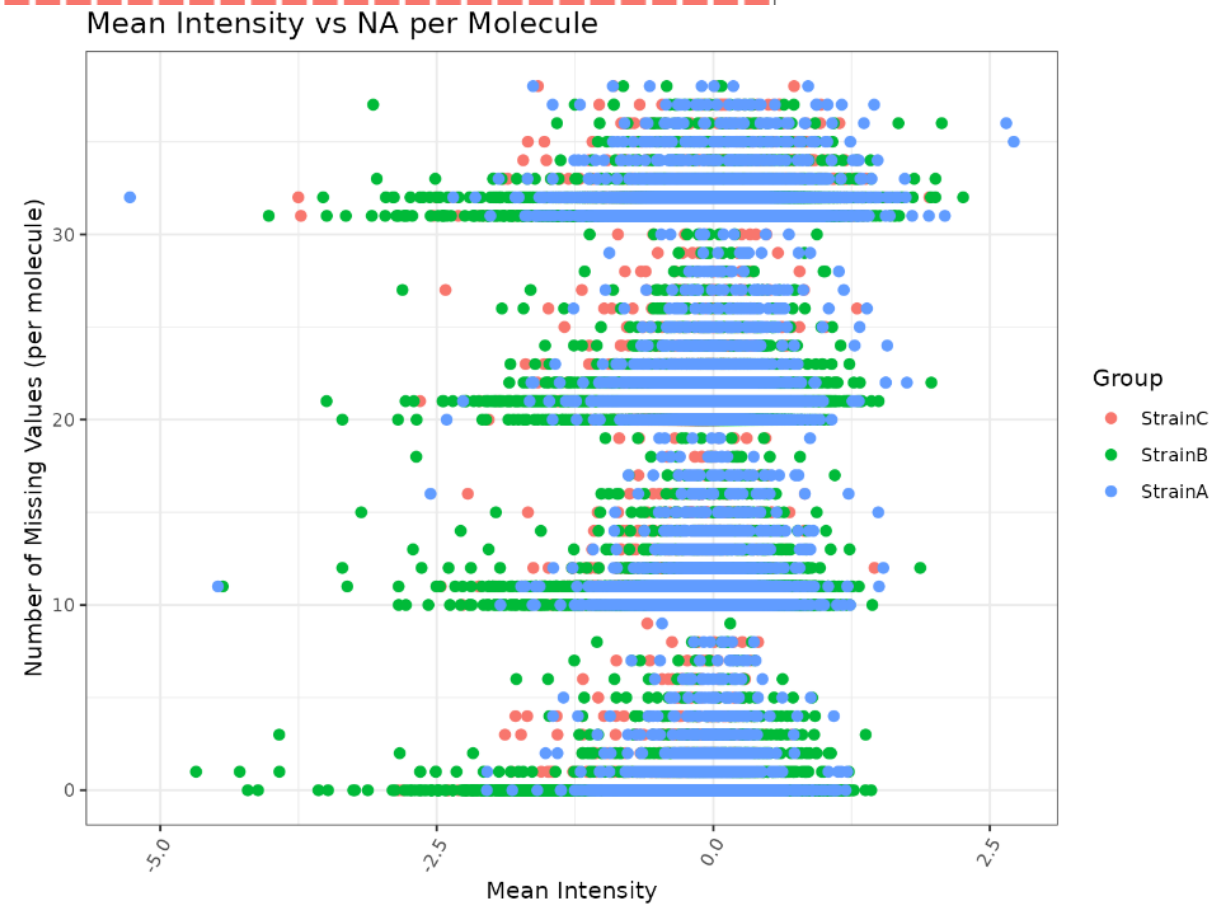
- Identify **samples** that are potential outliers or anomalies (due to sample quality, preparation, or processing circumstances) using robust Mahalanobis distance (rMd) score based on 2-5 o fhte following metrics
  - Correlation
  - Proportion of data that is missing (“Proportion_Missing”)
  - Median absolute deviation (“MAD”)
  - Skewness
  - Kurtosis



May remove **samples** with higher proportion missing than other samples

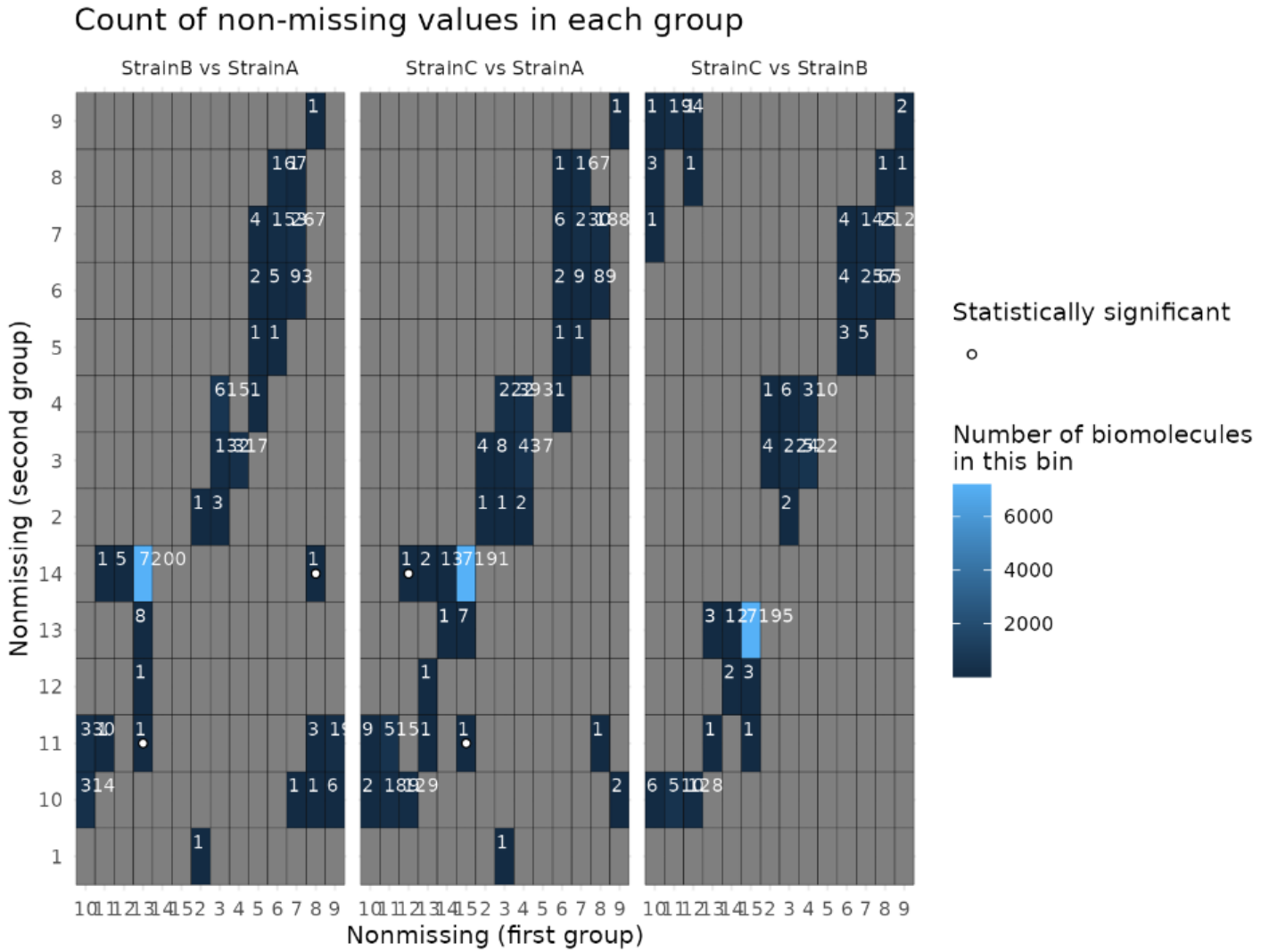
Explore Patterns of Missing Data





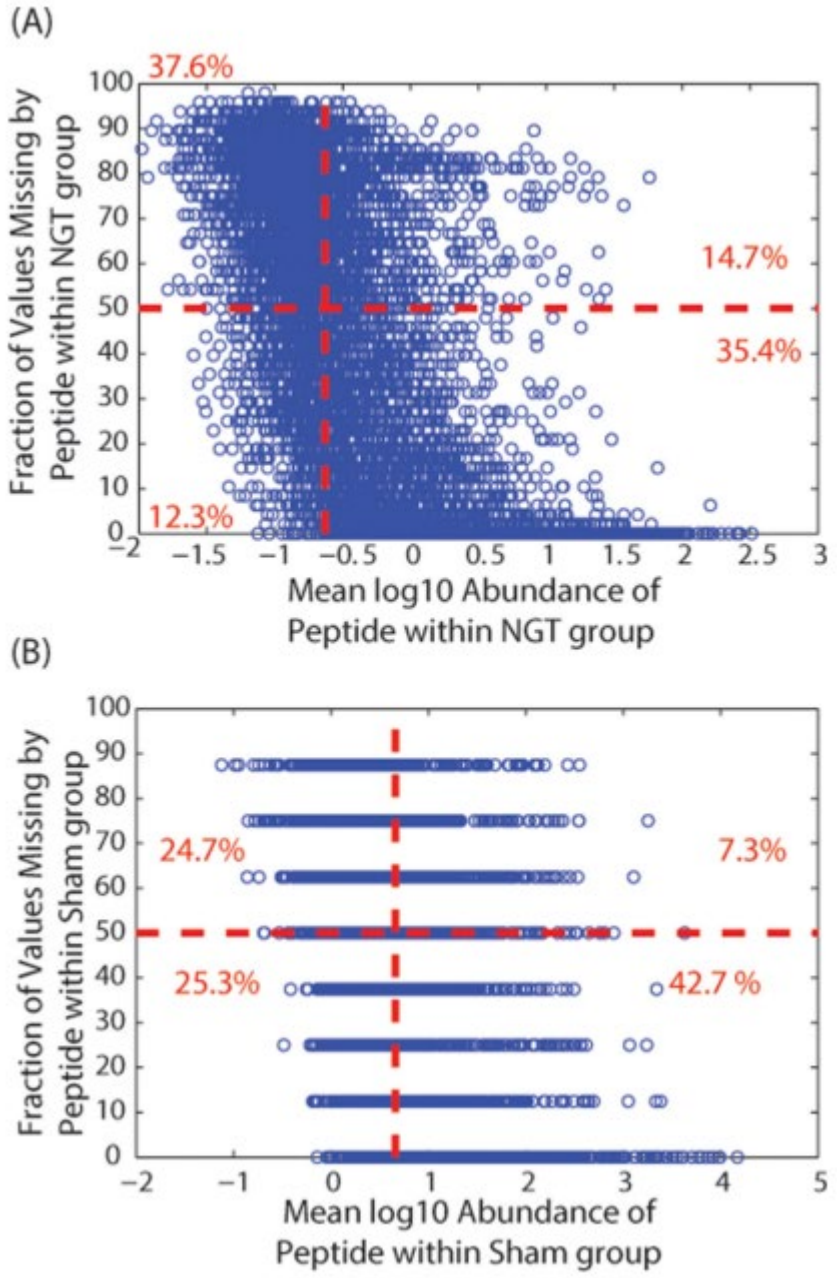
Statistically Compare Missingness Between Groups

- Like ANOVA but instead of comparing means & variance, g-test compares rates of missing data between groups



Patterns of Missingness

- Human plasma (top)
  - n = 23
  - Negative relationship between missing values and log10 mean intensity
  - Correlation = −0.51
- Mouse lung (bottom)
  - n = 8
  - Negative relationship between peptide missing data and intensity
  - Correlation = −0.40



Webb-Robertson, B., et al. 2015. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J Proteome Res.* 2015 May 1; 14(5): 1993–2001. doi:10.1021/pr501138h

Nature of Missing Data

- Peptide peak intensity and amount of missing data have been previously shown to be negatively correlated [26,27]
  - Potentially due to left-censoring of data. Under this assumption, the following should be true:
    - Only low-abundant peptides should have missing values.
    - Fraction of missing values should increase as the peptide intensity decreases.
- Don’t see this uniformly though
  - **Not all** peptides of low intensity have large amounts of missing values and likewise **not all** highly abundant peptides have high coverage.
  - Although there is a relationship between peptide intensity and missing values many peptides exhibit other behavior
- Conclusion: missing values are a combination of NMAR and MAR data.

Webb-Robertson, B., et al. 2015. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J Proteome Res.* 2015 May 1; 14(5): 1993–2001. doi:10.1021/pr501138h

Imputation

- A variety of imputation algorithms have been developed and discussed in the literature. [12,14–16,20,27,35–38]
- Algorithms grouped into three categories:
  - (1) imputation by a single-digit replacement
  - (2) imputation based on local structures in datasets
  - (3) imputation based on global structures in datasets
- Various methods can be used to execute each:
  - (1) replace missing with limit of detection (LOD), half the minimum observed among all peptides (LOD1), half the minimum for each peptide (LOD2), random tail imputation (RTI)
  - (2) K nearest neighbors (KNN), local least squares (LLS), least-squares adaptive (LSA), regularized expectation maximization (REM), model-based imputation (MBI)
  - (3) Probabilistic principal component analysis (PPCA), Bayesian principal components analysis (BPCA)

Webb-Robertson, B., et al. 2015. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J Proteome Res.* 2015 May 1; 14(5): 1993–2001. doi:10.1021/pr501138h

Evaluating the Imputation

- Dilution series
  - Dataset created with known dilution ratios produces peptides with known expected ratio.
  - Compare expected to actual based on analysis data set (observed combined with imputed data) for each imputation method.
  - Determine ‘best’ imputation method
    - How to compare?
    - Coefficient of variation (CV) of root-mean-square error (RMSE) to measure deviation of observed values from expected values each peptide and protein
- Experimental data
  - Collected on real samples, with no estimates of actual values (no ground truth).
  - If you have two sets of sample from known and distinct experimental groups → evaluate based on classification accuracy (classification method applied to analysis data set)
  - If not – what are your options?

Webb-Robertson, B., et al. 2015. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J Proteome Res.* 2015 May 1; 14(5): 1993–2001. doi:10.1021/pr501138h

Questions?

Dr. Moses Obiri
Statistician—Data Scientist
Moses.Obiri@pnnl.gov


Networking Break
10:35 – 10:45 a.m.

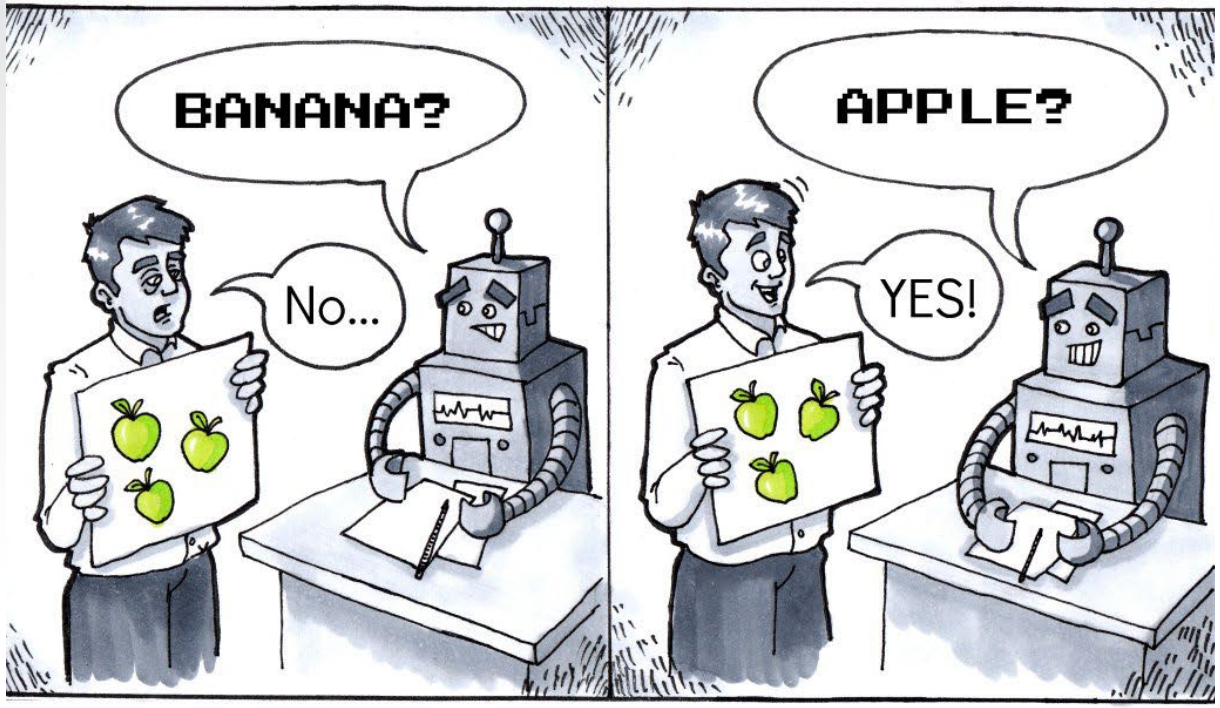
8:30-8:35 a.m.	Introduction	Javier Flores
8:35-9:35	Experimental Design	Damon Leach
9:35-10:35	Missing Data and Imputation	Moses Obiri
10:35-10:45	Networking Break	
10:45-11:45	**Unsupervised Learning**	**Javier Flores**

EMSL

Unsupervised Learning

Javier E. Flores, PhD
Biostatistician

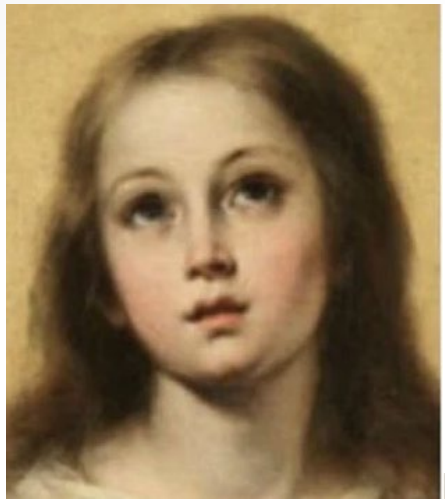
What is unsupervised learning?



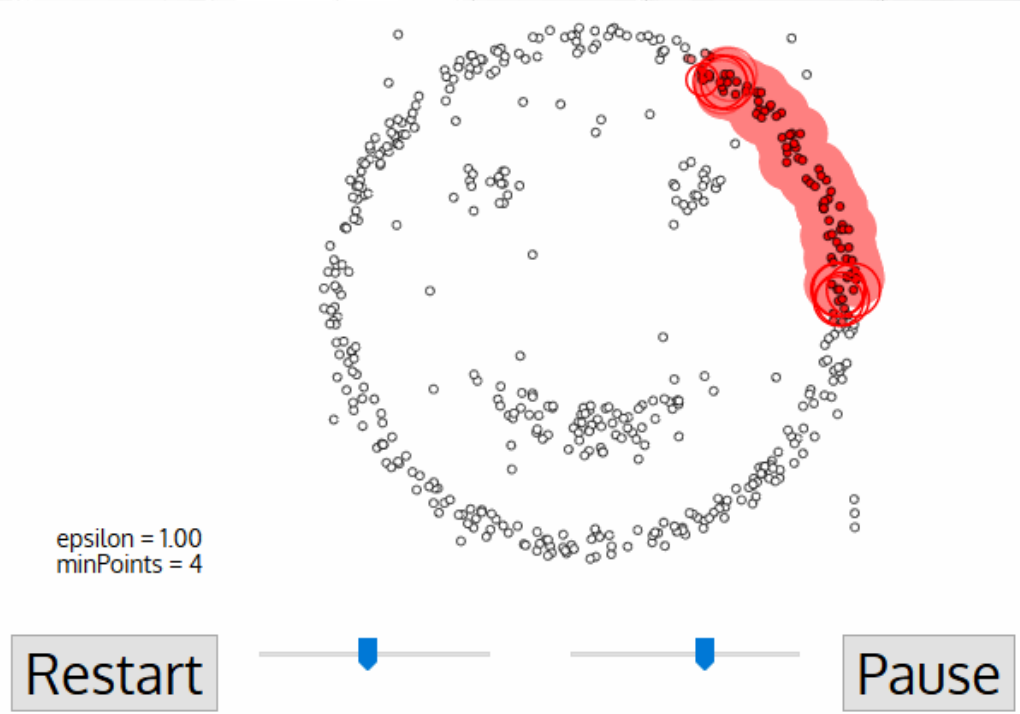
Supervised Learning

Categorizations of unsupervised learning techniques

Dimension Reduction / Feature Extraction



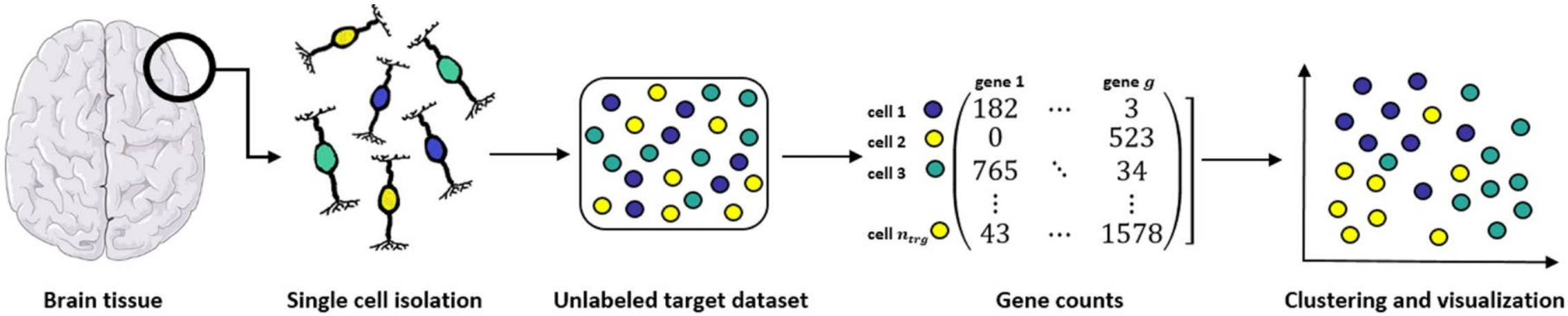
Clustering



Source: http://arogozhnikov.github.io/images/opera/post/clustering-dbscan-smiley.gif

Terminology

- **Sample**: the item(s) or unit(s) of analysis.
- **Features**: the distinct traits or characteristics that describe each sampled unit.
  - The collection of p features are represented in the data as a p-dimensional **feature vector**
- **Data**: the collection of samples and their features.
- **Feature extraction**: a transformation of the p-dimensional feature vector to a p'-dimensional feature vector, where $p' \ll p$.
- **Unsupervised Learning**: techniques that enable pattern identification within the data without prior knowledge (labels) on the samples.



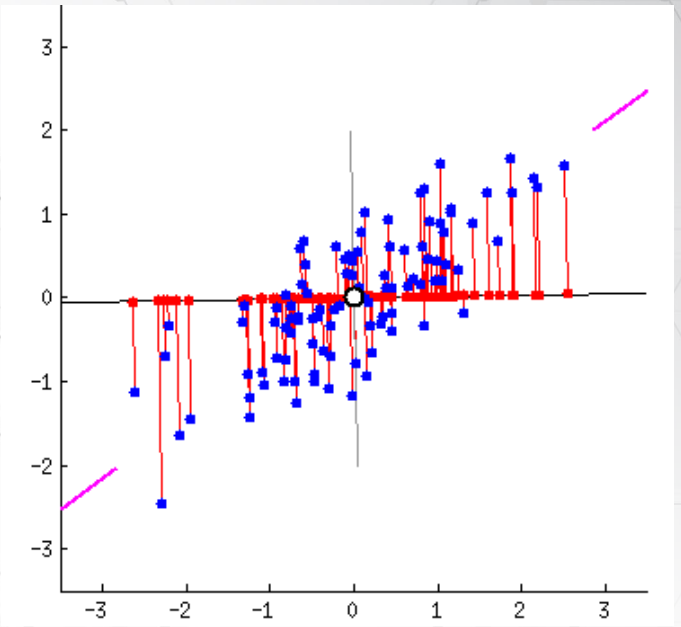
Source: Mieth, B., Hockley, J.R.F., Görnitz, N. *et al.* Using transfer learning from prior reference knowledge to improve the clustering of single-cell RNA-Seq data. *Sci Rep* **9,** 20353 (2019). https://doi.org/10.1038/s41598-019-56911-z

Dimension Reduction Techniques

Principal Component Analysis (PCA)

- Given a set of p features, PCA will construct p **principal components** that are linear combinations of the original set of features.
  - $\mathrm{PC}_1 = w_{11}Feature_1 + w_{12}Feature_2 + \cdots + w_{1p}Feature_p$
  - $\mathrm{PC}_2 = w_{21}Feature_1 + w_{22}Feature_2 + \cdots + w_{2p}Feature_p$

- Principal components are found such that each are **uncorrelated** and ordered according to the proportion of variability that they explain.
  - The linear combinations (w_{ij}) that define each component correspond to the orthogonal **eigenvectors** of the data covariance matrix.
  - The **eigenvalues** of the data covariance matrix reflect the proportion of variability captured by the corresponding principal component.



Source: https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

- In practice, one retains only the first p' principal components, where $p' < p$. The number of components p' is chosen according to the cumulative variability they explain (e.g. 80%).

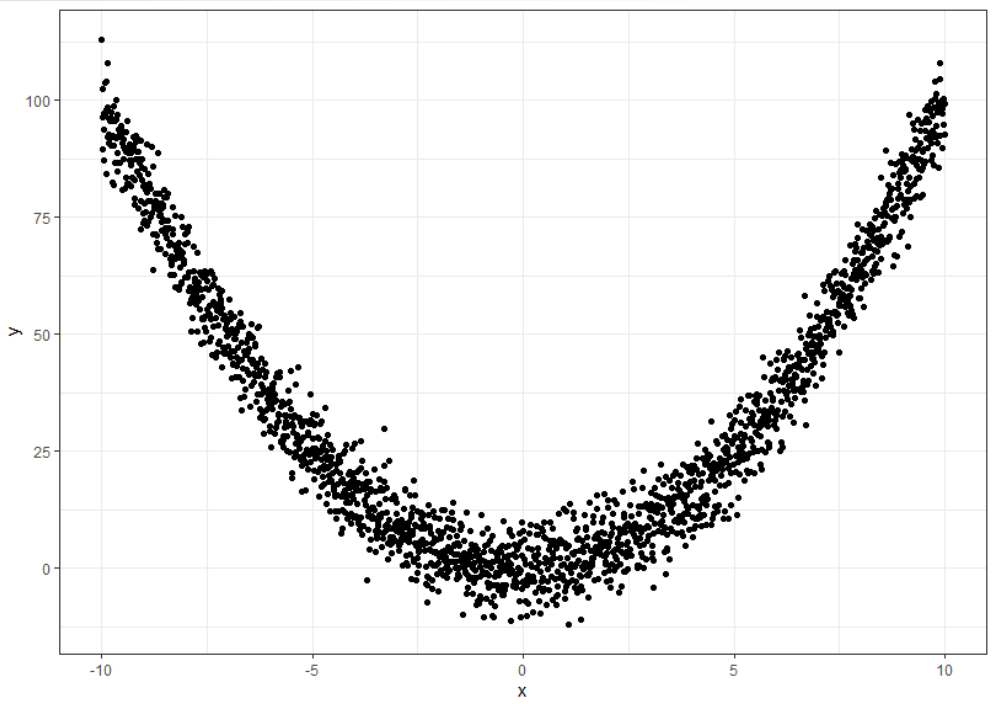
	Eigenvalue	Percentage of Variance	Cumulative
1	4.00644	44.52%	44.52%
2	1.635	18.17%	62.68%
3	1.12792	12.53%	75.22%
4	0.95466	10.61%	85.82%
5	0.46384	5.15%	90.98%
6	0.32513	3.61%	94.59%
7	0.27161	3.02%	97.61%
8	0.11629	1.29%	98.90%
9	0.09911	1.10%	100.00%

Source: https://www.originlab.com/doc/Tutorials/Principal-Component-Analysis

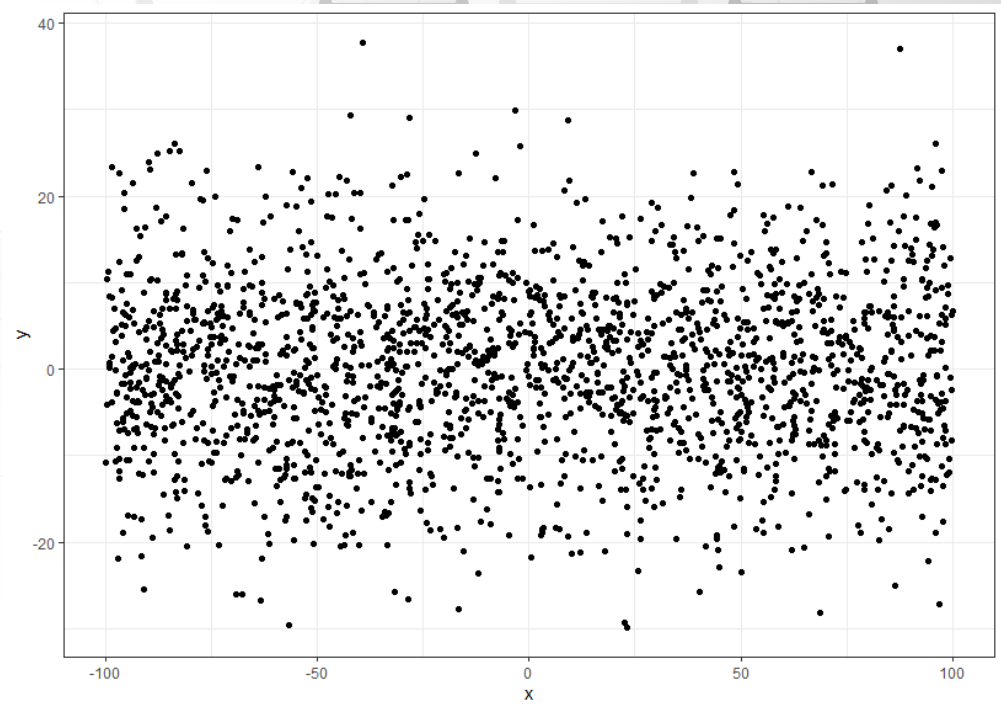
- PCA requires that each of the original p features are numeric variables.
  - Prior to performing PCA, the original data must be centered and scaled such that each feature has a mean of 0 and variance of 1.
  - If each of the p original features are normally distributed, the resulting principal components are independent.

Uncorrelated ≠ Independent

Correlation: -0.002

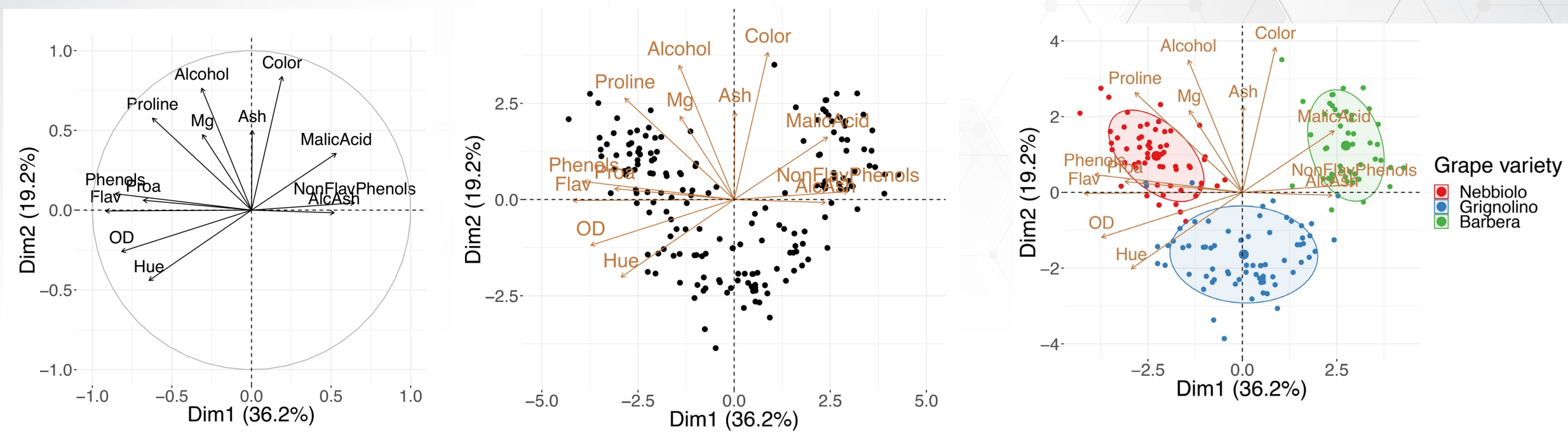


Correlation: 0.008



- If two variables are independent, they are uncorrelated. In general, the reverse is not true: uncorrelated variables are not necessarily independent.
- However, if the data are normally distributed, lack of correlation **does** imply independence.

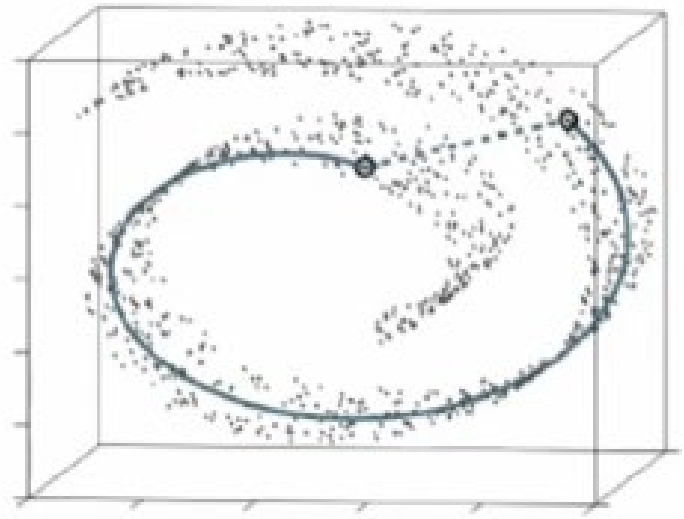
PCA: Example



Source: Nguyen LH, Holmes S (2019) Ten quick tips for effective dimensionality reduction. PLoS Comput Biol 15(6): e1006907. https://doi.org/10.1371/journal.pcbi.1006907

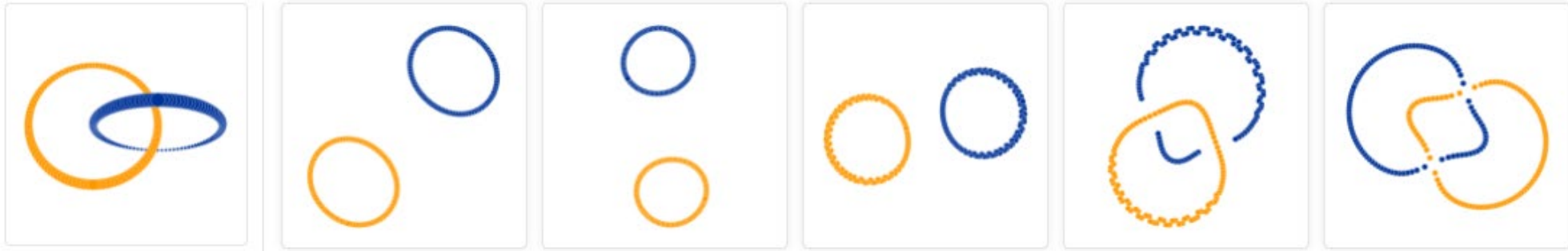
t-Distributed Stochastic Neighbor Embedding (t-SNE)

- PCA is a **linear** dimension reduction technique that may fail to adequately represent data that are non-linear in structure.
- t-SNE is a **non-linear** dimension reduction technique suitable for visualizing such non-linear, high-dimensional data in lower dimensions.
- Implementation of t-SNE requires that one specify a parameter called the **perplexity** that balances between local and global representation of data structure.

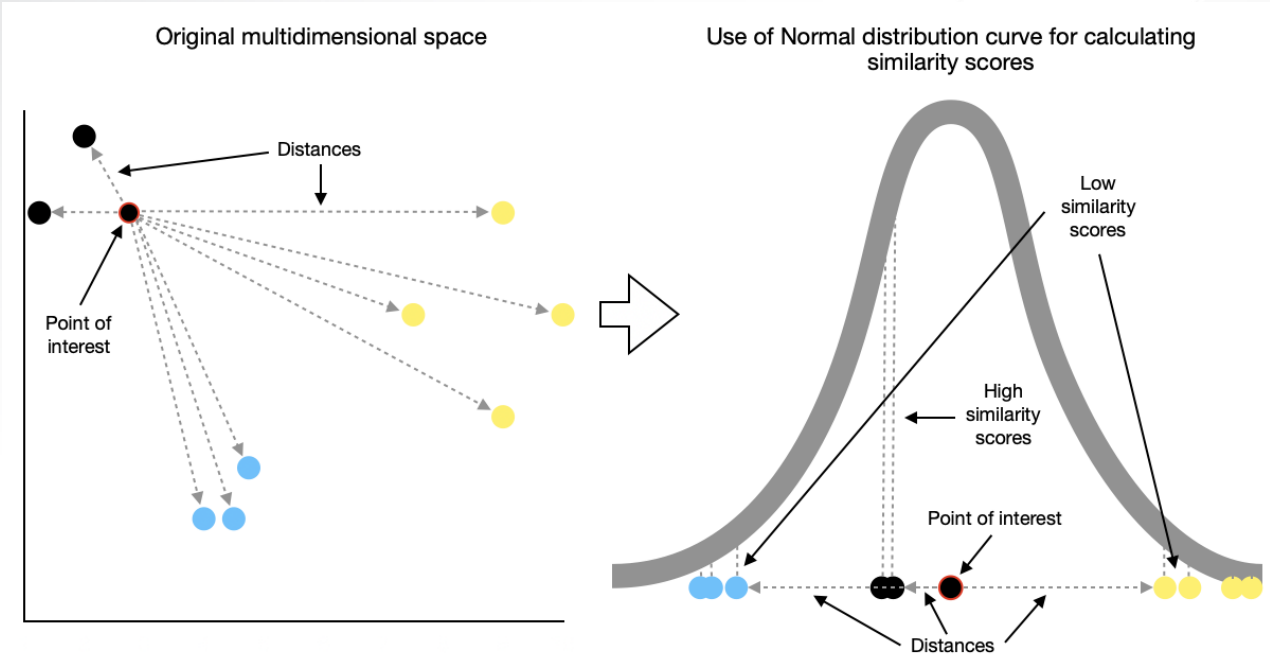


Source: https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1

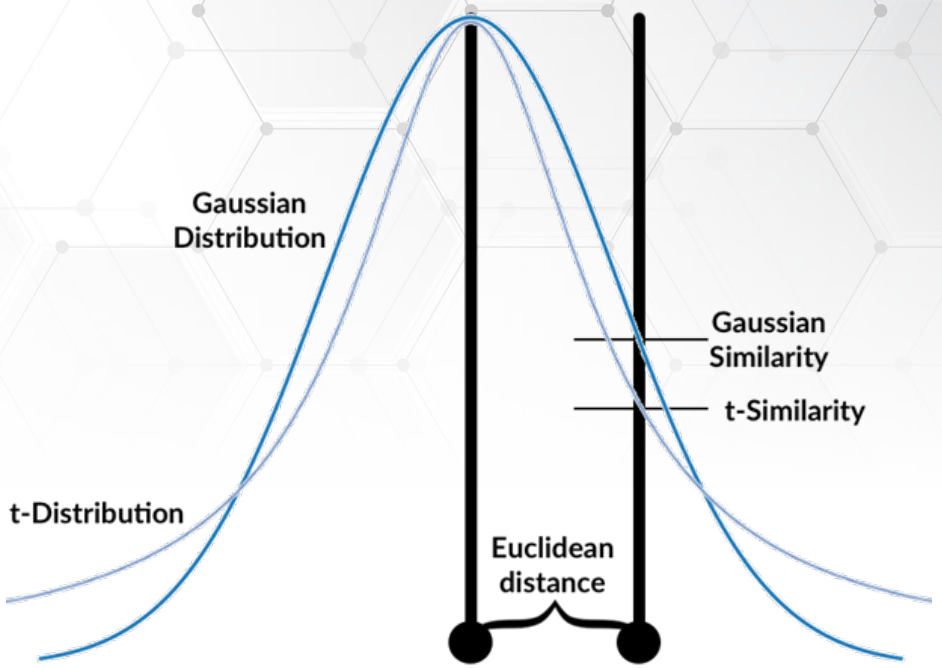
Local structure ⟷ Global structure



t-Distributed Stochastic Neighbor Embedding (t-SNE)

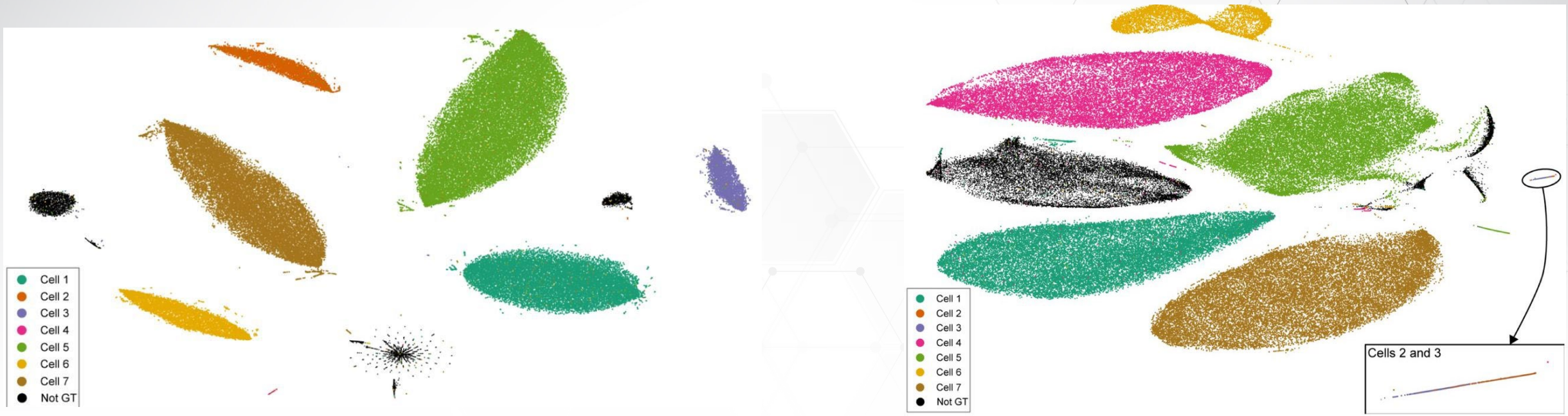


Source: https://towardsdatascience.com/t-sne-machine-learning-algorithm-a-great-tool-for-dimensionality-reduction-in-python-ec01552f1a1e



Source: https://hub.packtpub.com/using-autoencoders-for-detecting-credit-card-fraud-tutorial/

t-SNE: Example



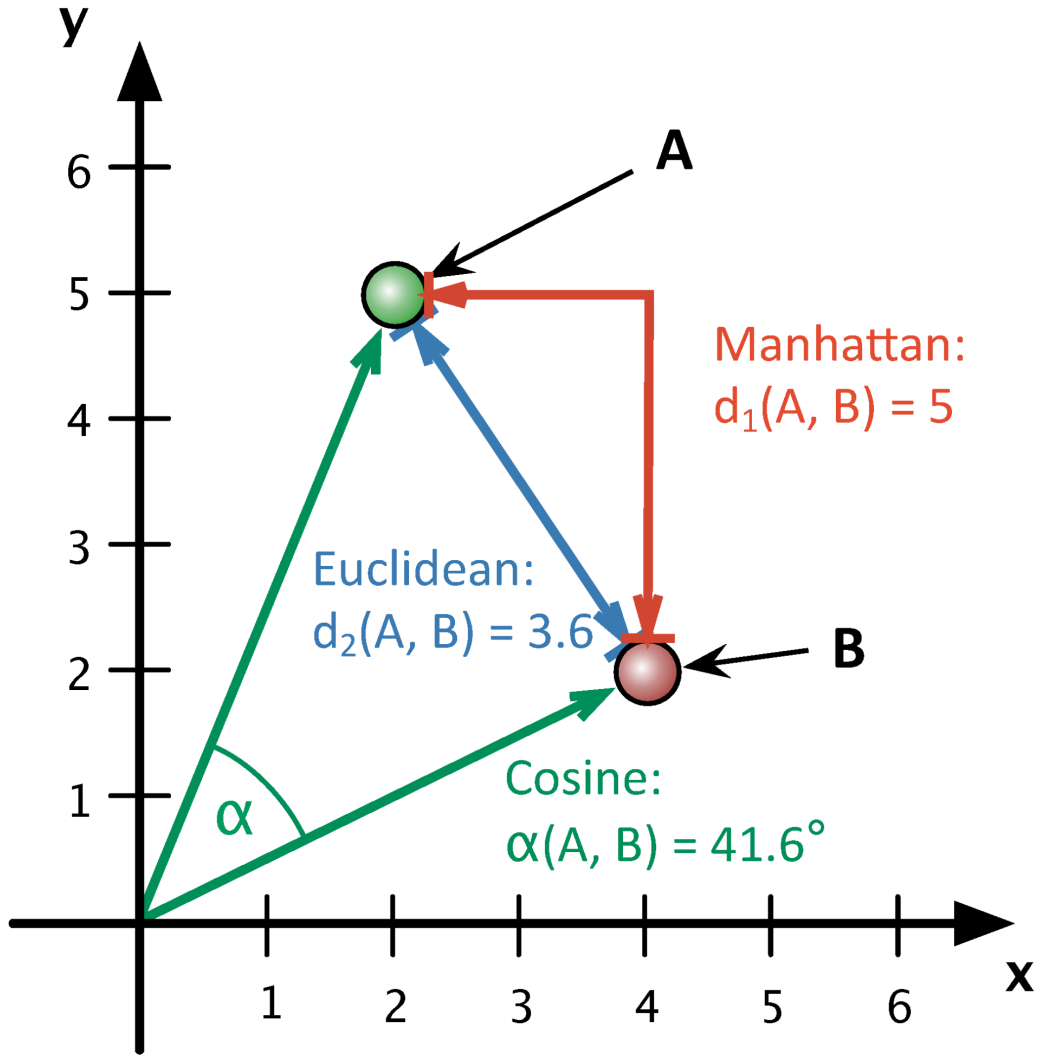
Source: George Dimitriadis, Joana P. Neto, Adam R. Kampff; t-SNE Visualization of Large-Scale Neural Recordings. *Neural Comput* 2018; 30 (7): 1750–1774. doi: https://doi.org/10.1162/neco_a_01097

- "Cluster" sizes have no meaning in t-SNE plots, nor do distances between them.
- Since t-SNE is a non-deterministic algorithm, different runs with the same hyperparameters may yield different results.
- In contrast to PCA, the lower-dimensional mappings of t-SNE have no interpretation.
  - Related to this point, t-SNE mappings may not be applied to new data whereas PCA loadings can.

Clustering Techniques

Measuring Dissimilarity

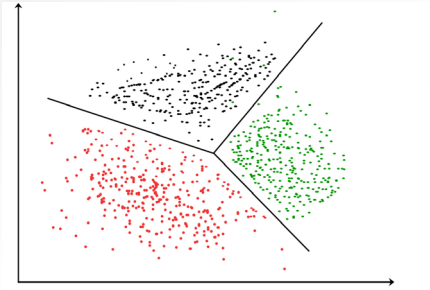
- Fundamentally, clustering algorithms rely upon some quantitative measure of dissimilarity to inform how sampled units are clustered.
  - In general, clustering algorithms aim to minimize the measured dissimilarity within clustered groups
- Several different metrics exist, and the choice of metric is largely data-dependent.
- Assuming a p-dimensional feature vector, the following are a few commonly used dissimilarity metrics:
  - **Euclidean**: $d(x_i, x_{i'}) = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2}$
  - **Manhattan**: $d(x_i, x_{i'}) = \sum_{j=1}^{p}|x_{ij} - x_{i'j}|$
  - **Correlation**: $d(x_i, x_{i'}) = 1 - corr(x_i, x_{i'})$



Source: https://community.insaid.co/hc/en-us/articles/360052305633-What-is-Euclidean-And-Manhattan-distances-in-KNN-

Clustering Approaches

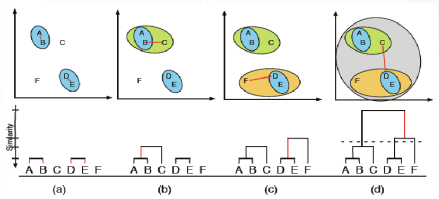
Partitional



Source: https://www.geeksforgeeks.org/clustering-in-machine-learning/

K-Means

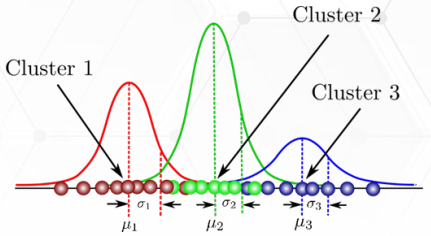
Linkage/Hierarchical



Source: Janssen, Peter & Walther, Carsten & Lüdeke, M.. (2012). Cluster Analysis to Understand Socio-Ecological Systems: A Guideline.

Agglomerative/Divisive

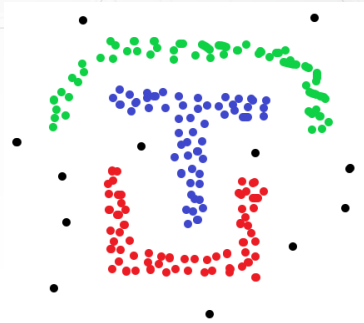
Model-Based



Source: https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

Gaussian Mixture Modeling (GMM)

Density-Based



Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

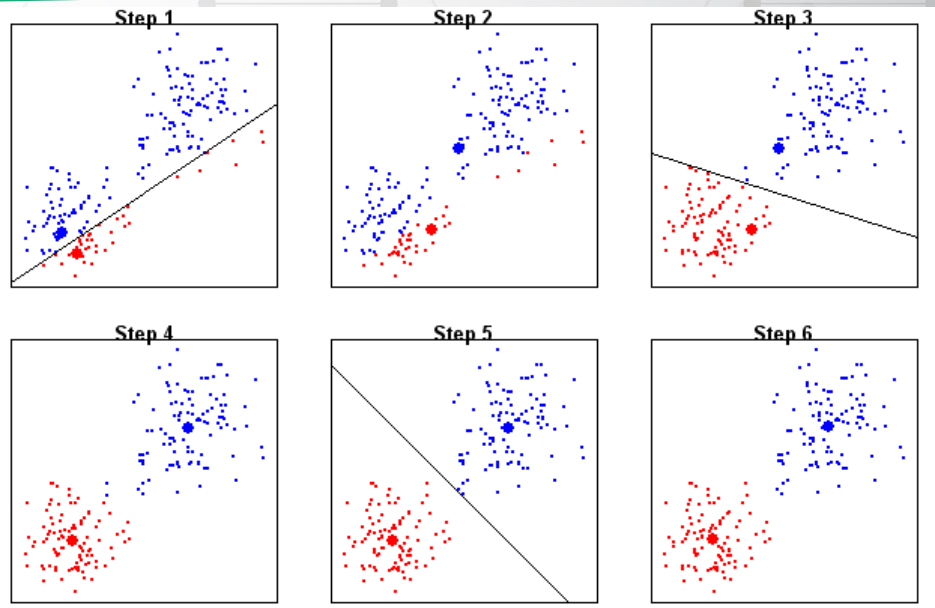
K-Means

- Provided specification of the number of clusters k, the k-means algorithm:
  1. Randomly chooses k **centroids** as initial cluster centers.
  2. Assigns each datapoint to cluster of nearest centroid.
  3. Updates centroid assignment to the mean of datapoints in each cluster.
  4. Repeats 2-3 until assignments no longer change (convergence).

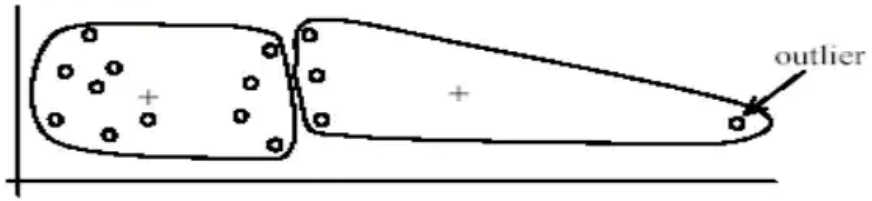
- Given the random initialization, setting a random number generator seed prior to running is essential for reproducibility.

- The k-means algorithm seeks to minimize the variability of data within clusters.
  - Convergence is only guaranteed when Euclidean distance is used.

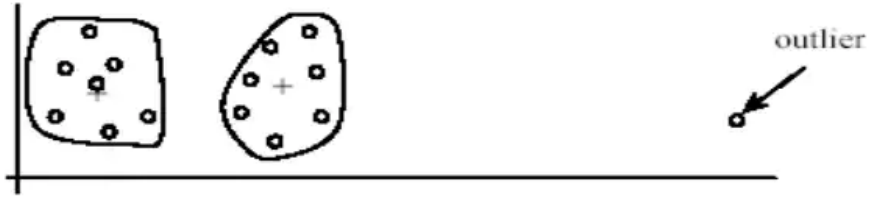
- The k-means algorithm is sensitive to outliers and differences in scale.
  - Data should be centered and scaled prior to implementation
  - If outliers are present, consider implementing the partition around medoids (PAM) algorithm.
  - For categorical data, the k-modes algorithm should be implemented.



Source: https://medium.com/@dilekamadushan/introduction-to-k-means-clustering-7c0ebc997e00



(A): Undesirable clusters

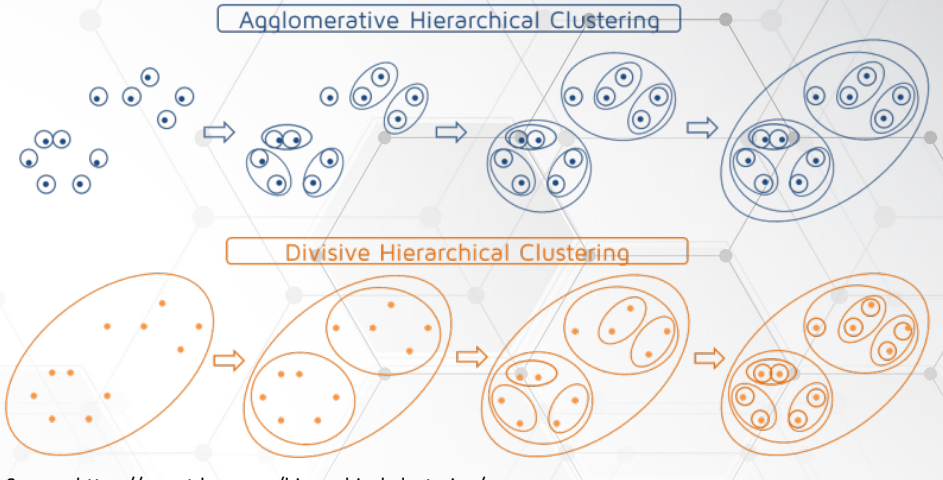


(B): Ideal clusters

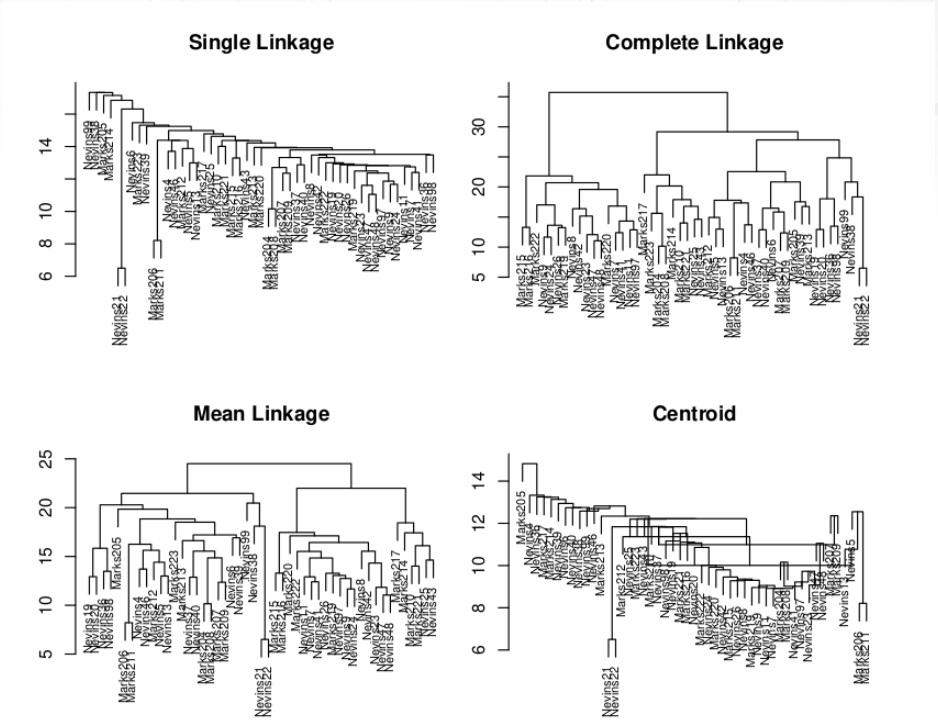
Source: https://www.slideshare.net/anilyadav5055/15857-cse422-unsupervisedlearning

Agglomerative/Divisive Hierarchical Clustering

- When smaller-scale clustering is expected within larger clusters, hierarchical algorithms should be considered:
  - Observations are iteratively grouped (agglomerative) or divided (divisive) according to measured dissimilarity.
- Implementation of these algorithms requires specification of
  - The number of clusters k
  - The dissimilarity metric
  - The linkage method (i.e. criteria for merging/dividing clusters)
- Several linkage methods exist, each of which may lead to different clusters. Common methods include
  - **Single Linkage**: minimum distance between points in different clusters. Prone to chaining.
  - **Complete Linkage**: maximum distance between points in different clusters. Tends to yield compact, spherical clusters.
  - **Mean Linkage**: average of all distances between points in different clusters. Compromise between single and complete linkage.
  - **Centroid Linkage**: distance between cluster centroids. Often similar to mean linkage.



Source: https://quantdare.com/hierarchical-clustering/



Source: Smith, B. (2018). *Section 2: Machine Learning for Biomedical Data* [PowerPoint slides].

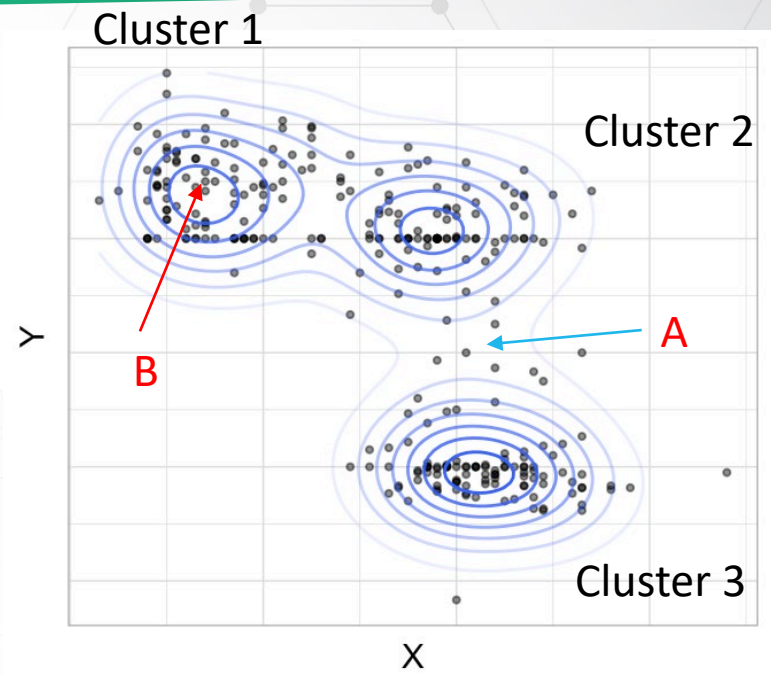
Gaussian Mixture Modeling (GMM)

- In contrast to previous approaches, GMM takes a statistical/probabilistic approach to clustering:
  - GMM assumes the data arise from a mixture of k multivariate normal distributions.
  - The expectation-maximization (EM) algorithm is used to estimate model parameters.

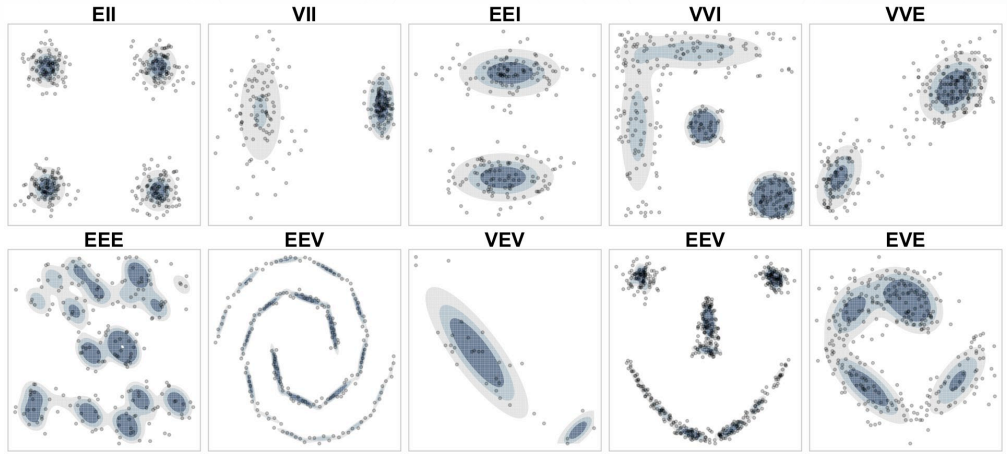
- Cluster assignments obtained from GMM are **soft**.
  - Unlike for previous approaches, cluster membership of each datapoint is characterized by a quantified level of uncertainty.

x_i	Cluster 1	Cluster 2	Cluster 3
A	0.002	0.499	0.499
B	0.999	0.001	0.000

- GMM clustering may better capture irregular, non-spherical clustering structures depending on user specification of k and the covariance structure.

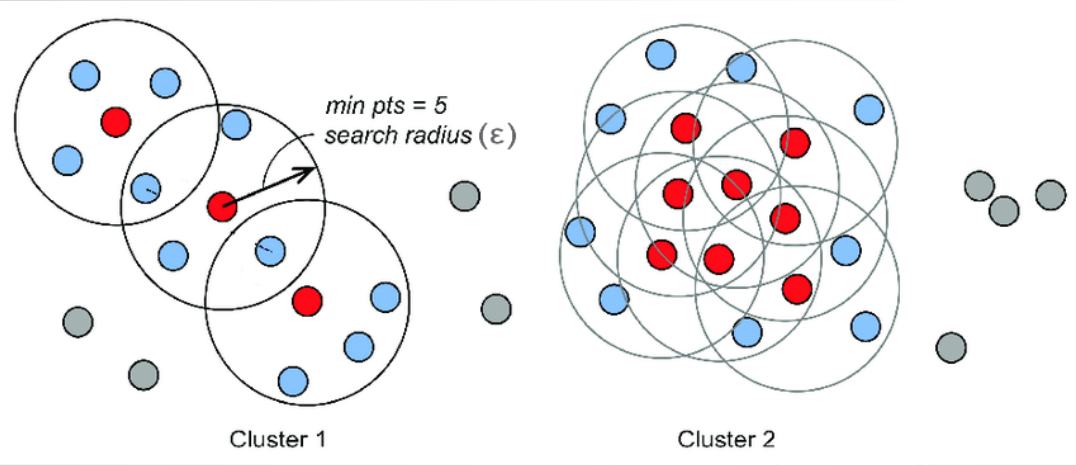


Source: https://bradleyboehmke.github.io/HOML/model-clustering.html

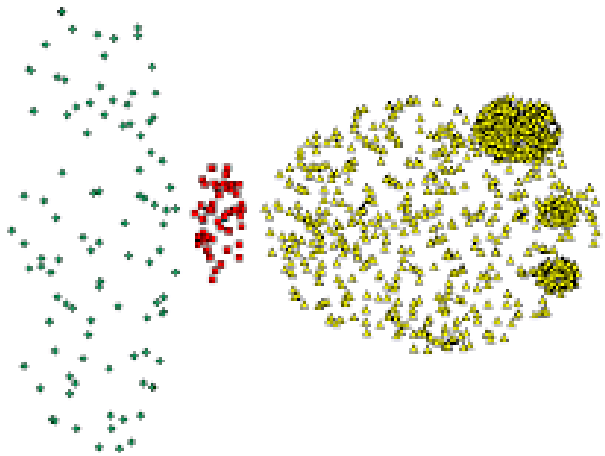


Source: https://bradleyboehmke.github.io/HOML/model-clustering.html

DBSCAN



Source: DiFrancesco, P.-M.; Bonneau, D.; Hutchinson, D.J. The Implications of M3C2 Projection Diameter on 3D Semi-Automated Rockfall Extraction from Sequential Terrestrial Laser Scanning Point Clouds. *Remote Sens.* **2020**, *12*, 1885. https://doi.org/10.3390/rs12111885



Source: https://web.iitd.ac.in/~bspanda/dbscan%20clustering.pdf

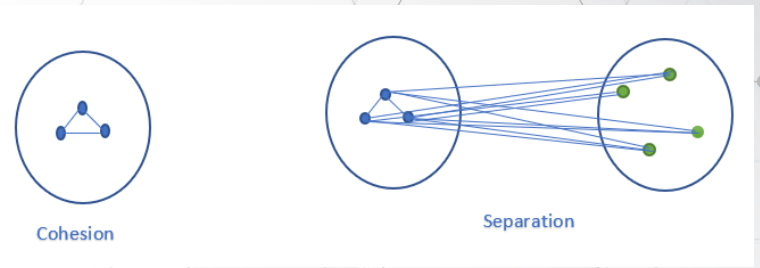
- Unlike previously discussed approaches, DBSCAN does not require specification of the number of clusters. Instead, users specify two parameters, “Min pts” and ε.
- Based on these parameters, DBSCAN defines clusters in the following way:
  - Identify all points that have “Min pts” datapoints within a distance of ε. These points are core points.
  - Core points that are **density-connected** are assigned to the same cluster.
  - Points that do not have “Min pts” datapoints within ε are classified as either border points or noise points.
    - Border points are within ε of a core point. Noise points are not.
  - Border points are assigned to the cluster of the nearest core point, noise points are unclustered.
- DBSCAN can find clusters of any shape and is robust to outliers.
  - However, DBSCAN cluster quality may be hindered by improper choice of distance metric (e.g. using Euclidean distance for high-dimensional data)
  - DBSCAN assumes uniform densities across clusters.

How to choose k and other hyperparameters?

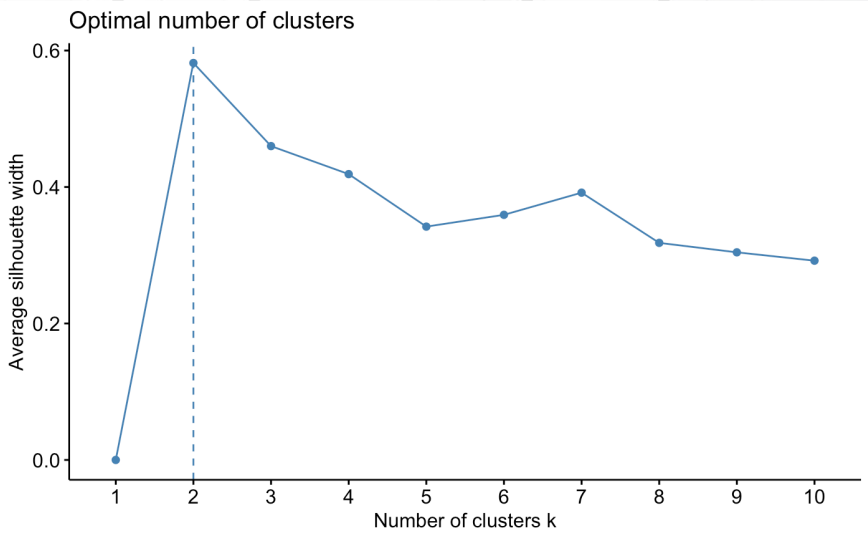
- Work with a statistician/data scientist! ☺

- Compute cluster **validation metrics** over a grid of parameter choices.
  - **Internal**: These metrics quantify cluster cohesion and separation. They may also measure connectivity.
  - **External**: These metrics quantify the agreement between clusters and some externally provided set of labels.

- Dozens of validation metrics exist, none of which are uniformly "best" in all situations.
  - In practice, several metrics are often computed.
  - For the purposes of parameter selection, the Dunn Index, Davies-Bouldin Index, silhouette coefficient, and gap statistic are often used.

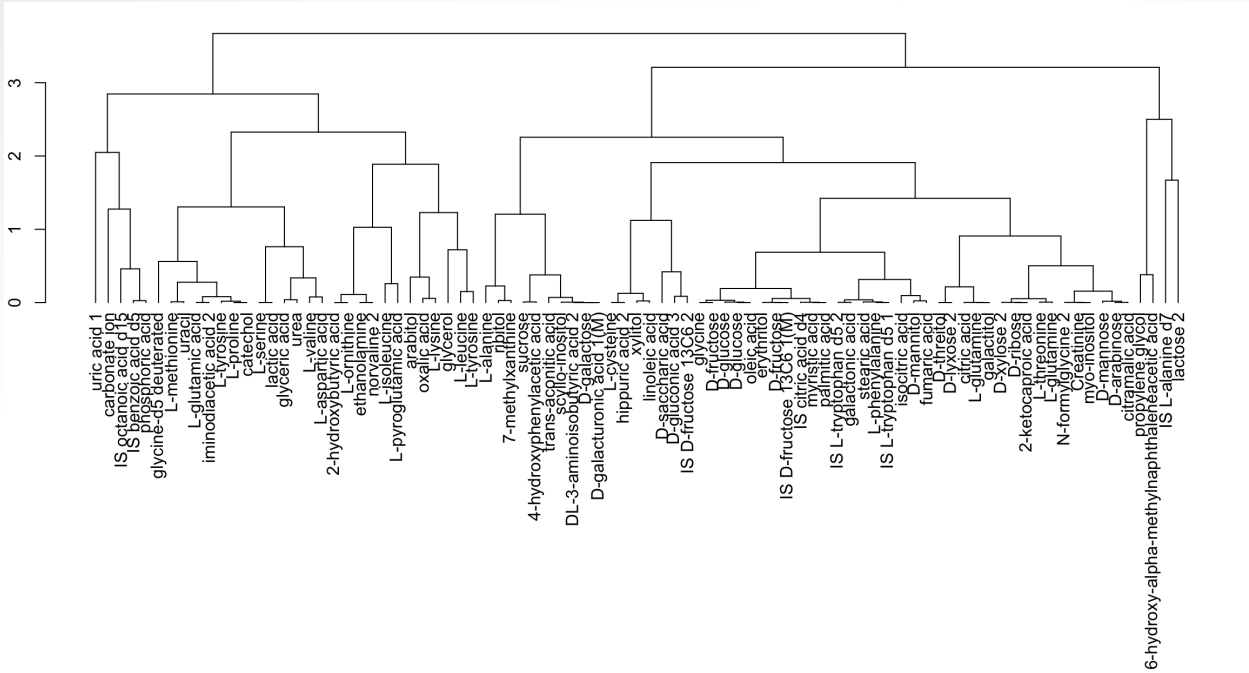


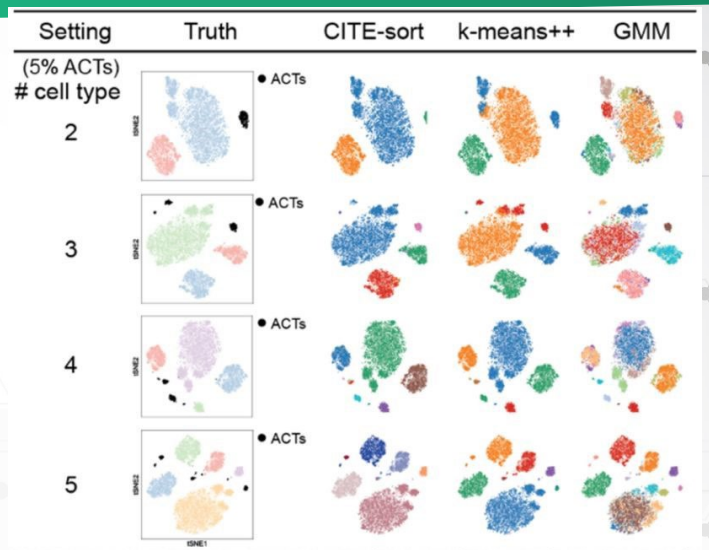
Source: https://towardsdatascience.com/are-the-clusters-good-1567de6a9524



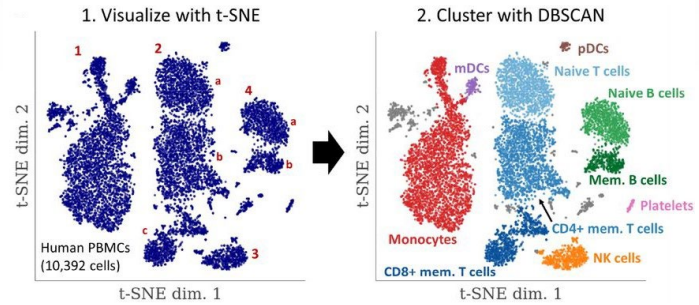
Source: https://rpkgs.datanovia.com/factoextra/reference/fviz_nbclust.html

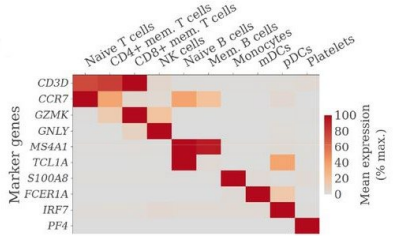
Clustering in Action!





Source: Qiuyu Lian, Hongyi Xin, Jianzhu Ma, Liza Konnikova, Wei Chen, Jin Gu, Kong Chen, Artificial-cell-type aware cell-type classification in CITE-seq, *Bioinformatics*, Volume 36, Issue Supplement_1, July 2020, Pages i542–i550, https://doi.org/10.1093/bioinformatics/btaa467





Source: Wagner, F. (2020). "Straightforward clustering of single-cell RNA-Seq data with t-SNE and DBSCAN." bioRxiv: 770388.

Thank you!

Afternoon Session

1:15-2:30 p.m.	Experimental Design in Practice	Damon Leach
2:30-2:40	Networking Break	
2:40-4:00	The Power of Simulation: Using Simulation to Answer Statistical Questions	Javier Flores