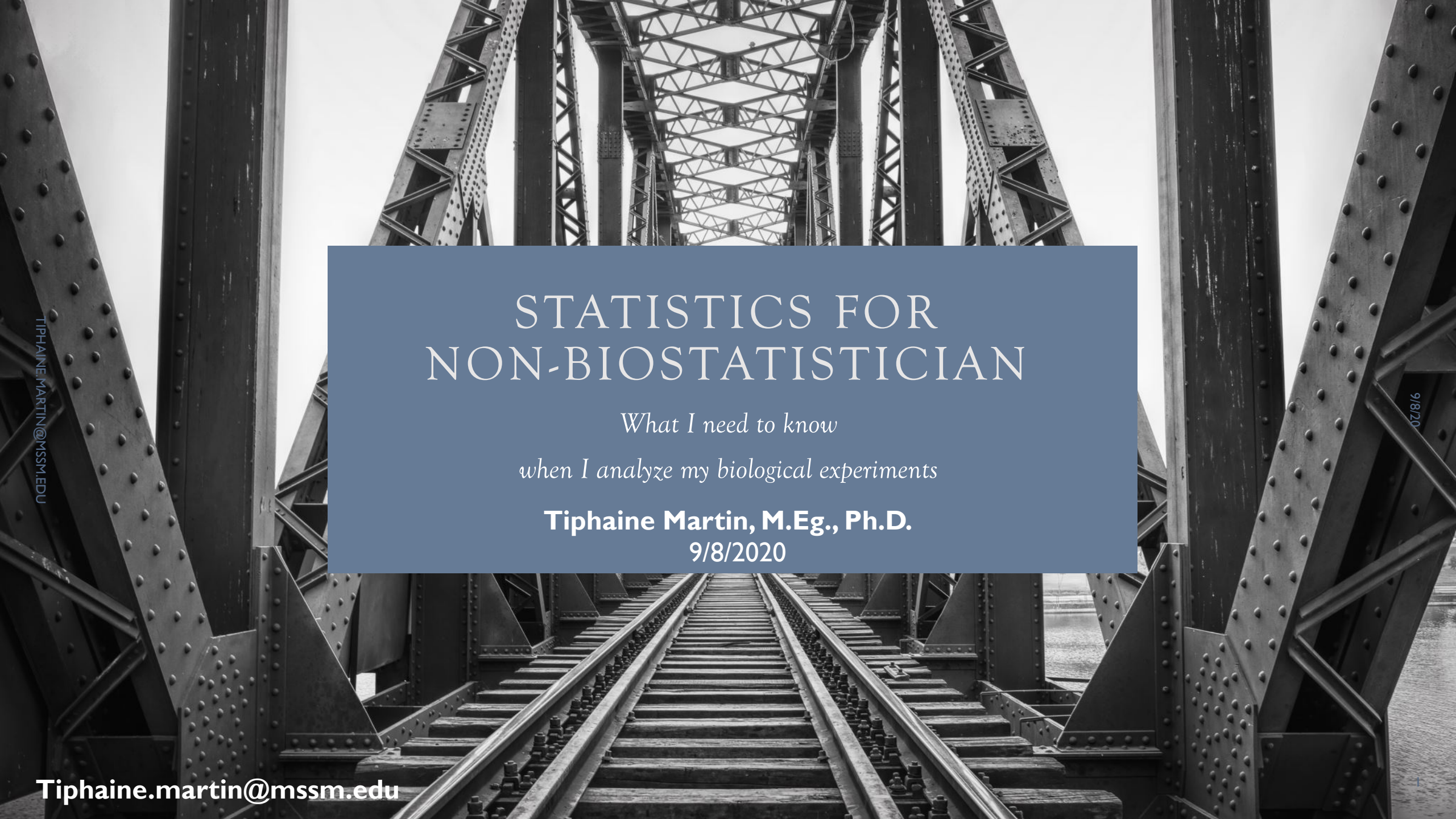


STATISTICS FOR NON-BIOSTATISTICIAN

What I need to know

when I analyze my biological experiments

Tiphaine Martin, M.Eg., Ph.D.

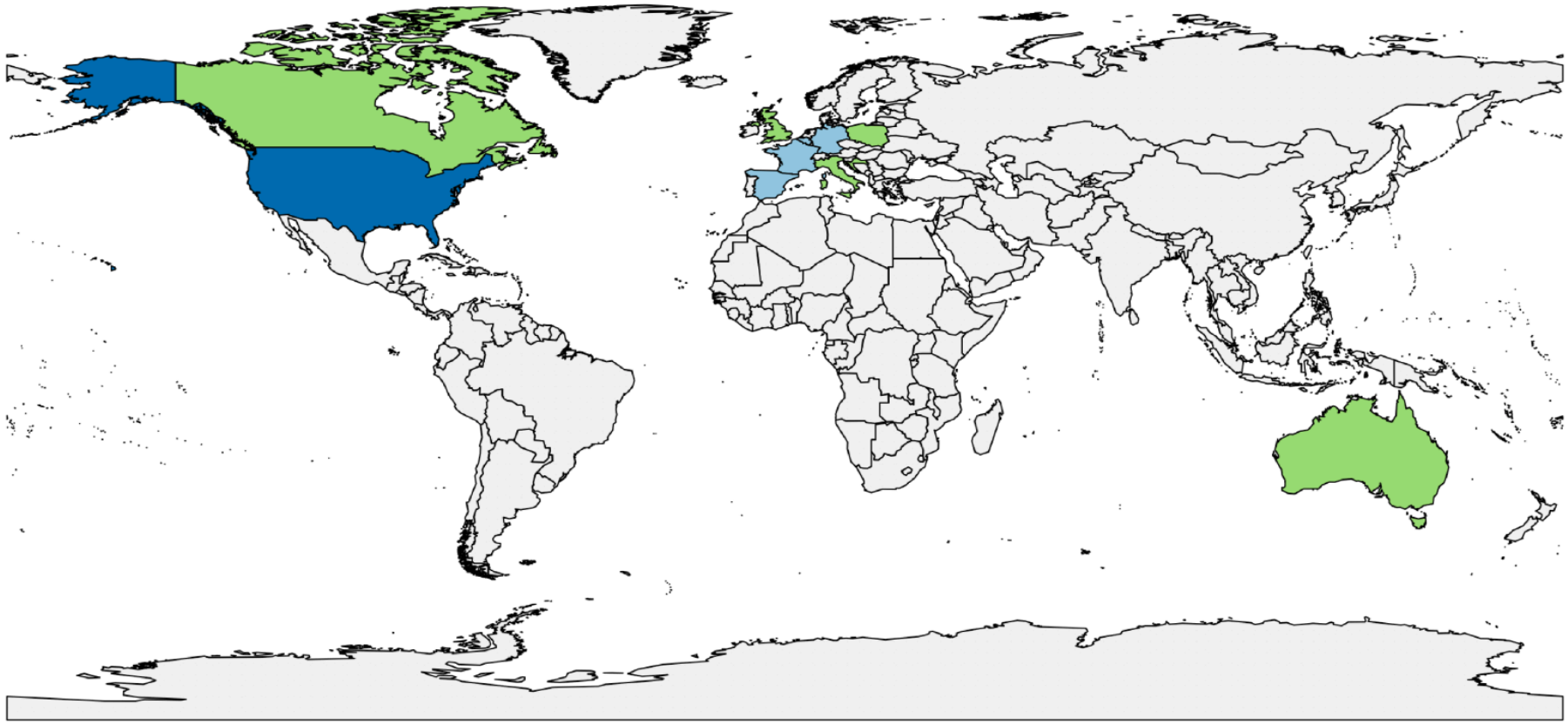
9/8/2020

TIPHAINE.MARTIN@MSSM.EDU

9/8/20

Tiphaine.martin@mssm.edu

Bioinformatician/Biostatistician - Multi-omic/EHR



Postdoc - Multi-omic/EHR/human – liquid biopsy/BioMe biobank, microbiome

PhD - Multi-omic/EHR/human – autoimmune thyroid diseases/rare diseases (hematology) – CHARGE consortium (epigenome)/EpiTwin/Blueprint Project/Interval Project/UK10K/TwinsUK/Genomics England

Research Engineer – Comparative genomics/yeast – Marine genomics – AI/robotics – human/virus EGI/ELIXIR/Genolevures/ BioMerieux/It-omics

TIPHAINE.MARTIN@MSSM.EDU

OUTLINE

- Visualization
- Basic statistics
- Artificial Intelligence vs Machine learning
- Examples

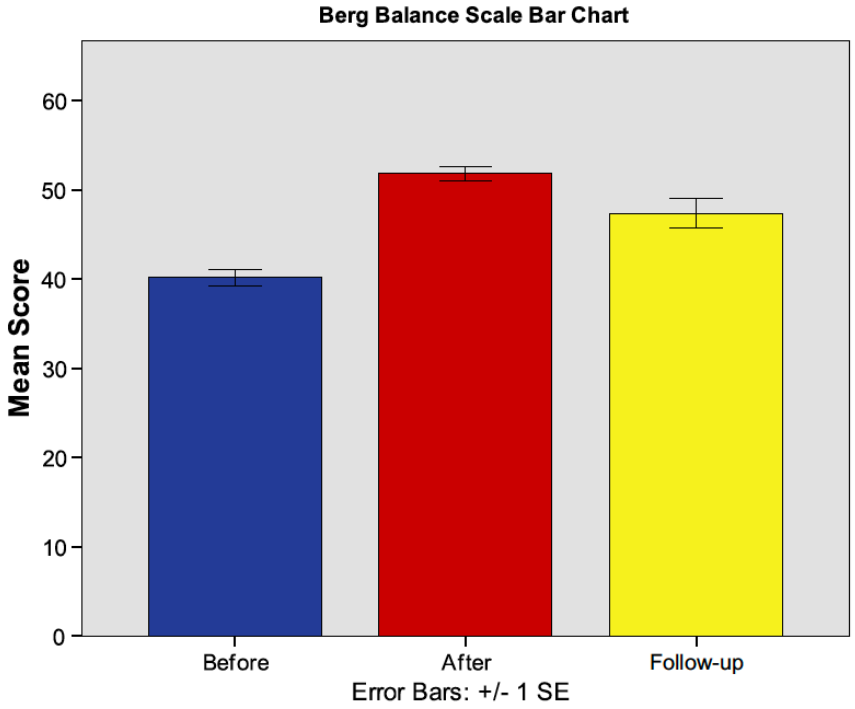
TIPHAINE.MARTIN@MSSM.EDU

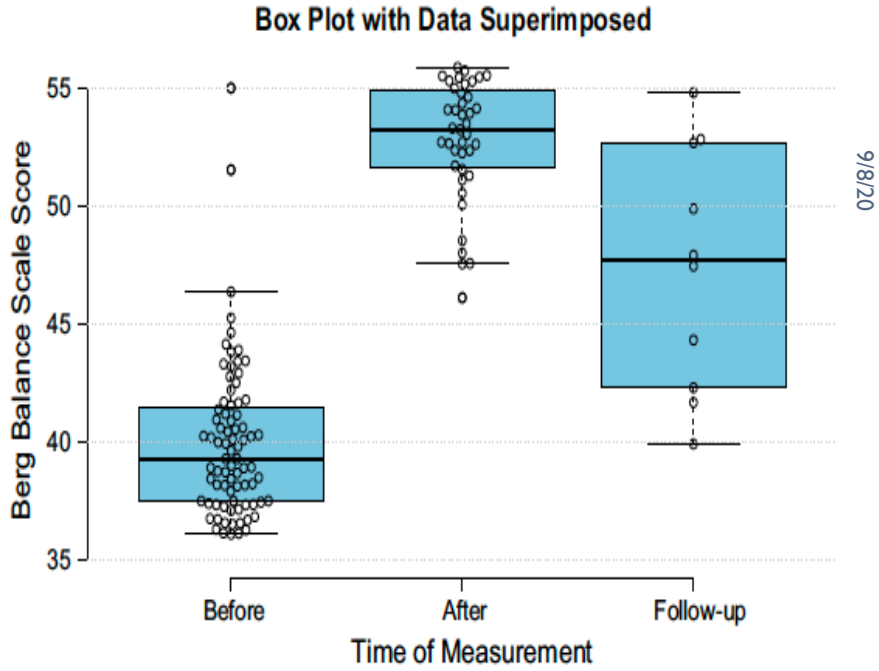
VISUALIZATION

TIPHAINE.MARTIN@MSSM.EDU

VISUALIZATION

Which one do you prefer? And why ?



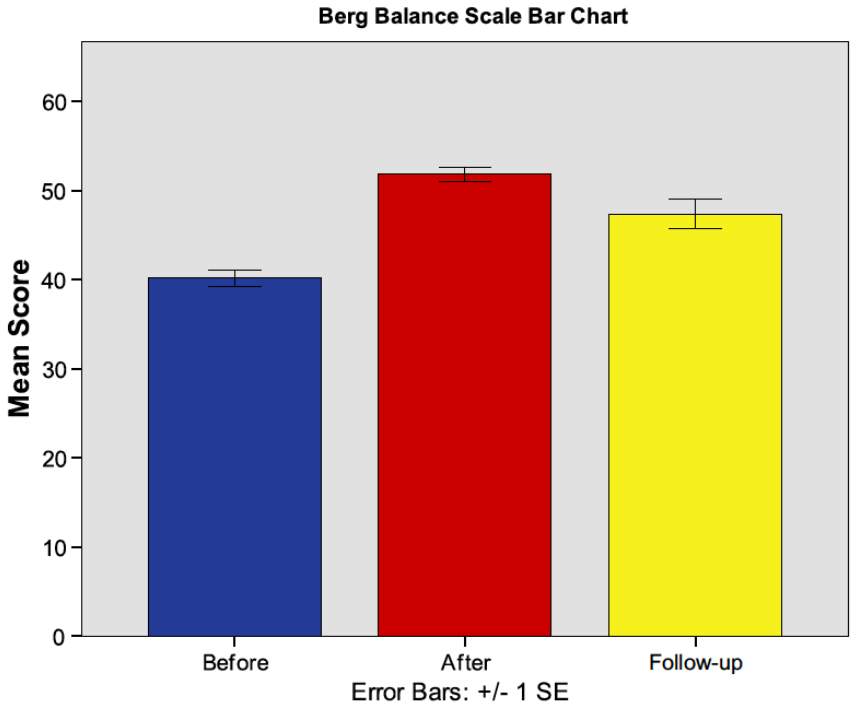


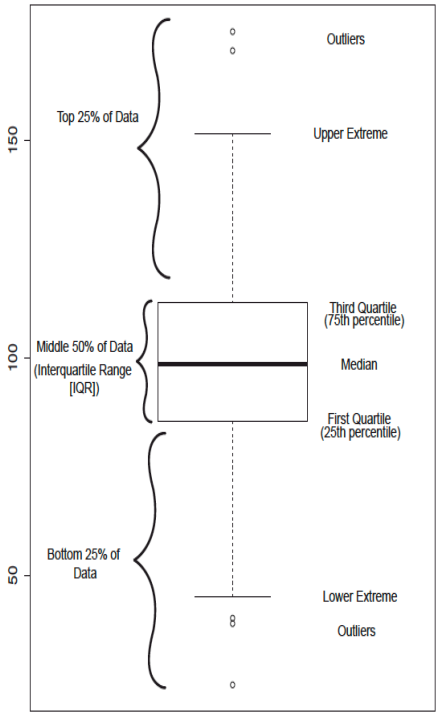
TIPHAINE.MARTIN@MSSM.EDU

R.L. Nuzzo / PM R 8 (2016) 268-272

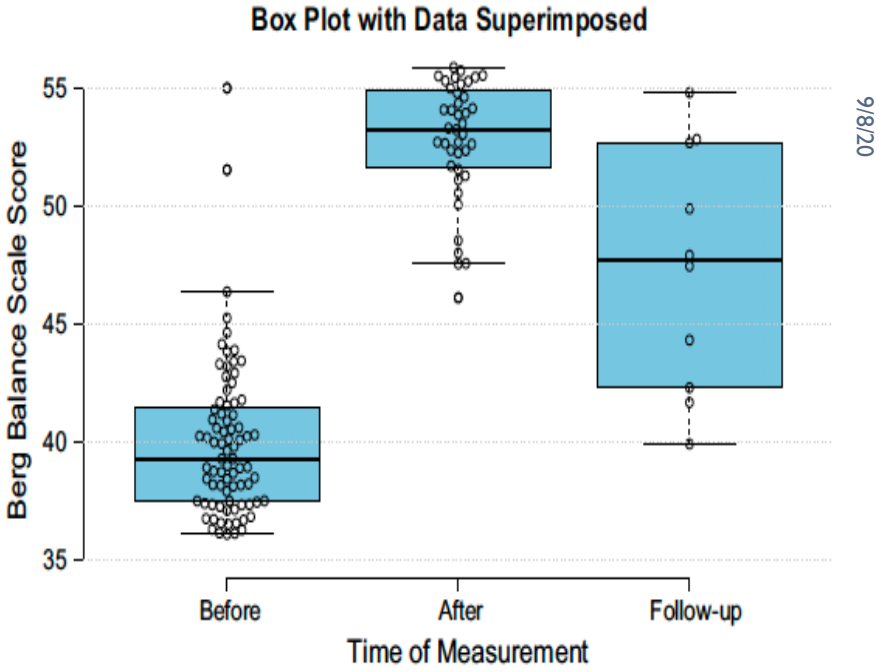
VISUALIZATION

To avoid





Better

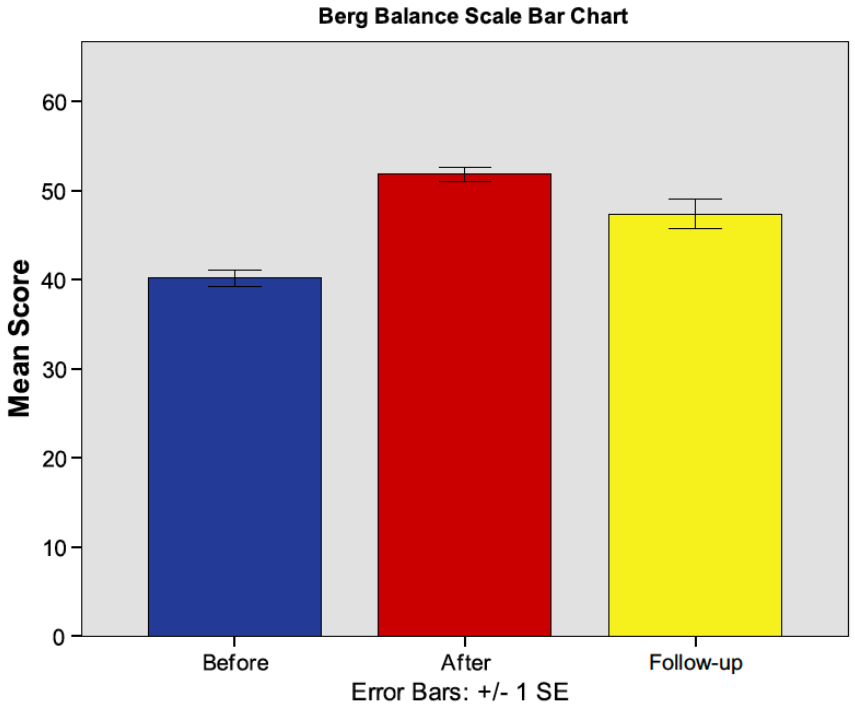


R.L. Nuzzo / PM R 8 (2016) 268-272

TIPHAINE.MARTIN@MSSM.EDU

VISUALIZATION

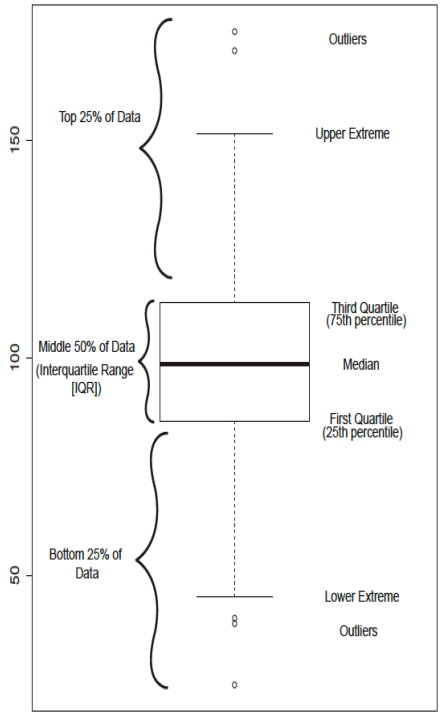
To avoid

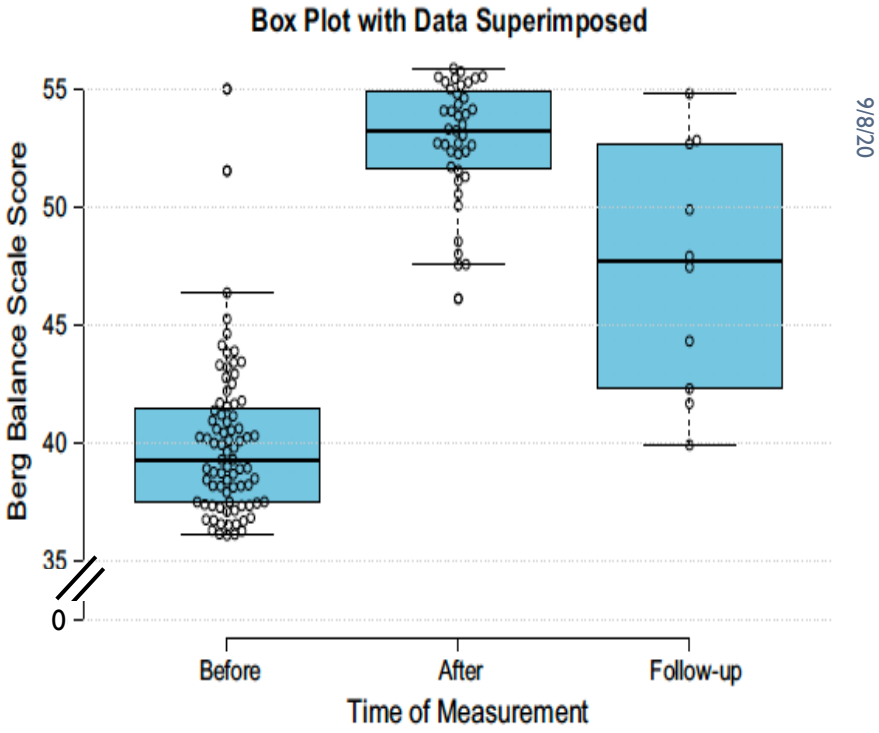


Better, but …

Need to start from zero

Add P-value in the figure



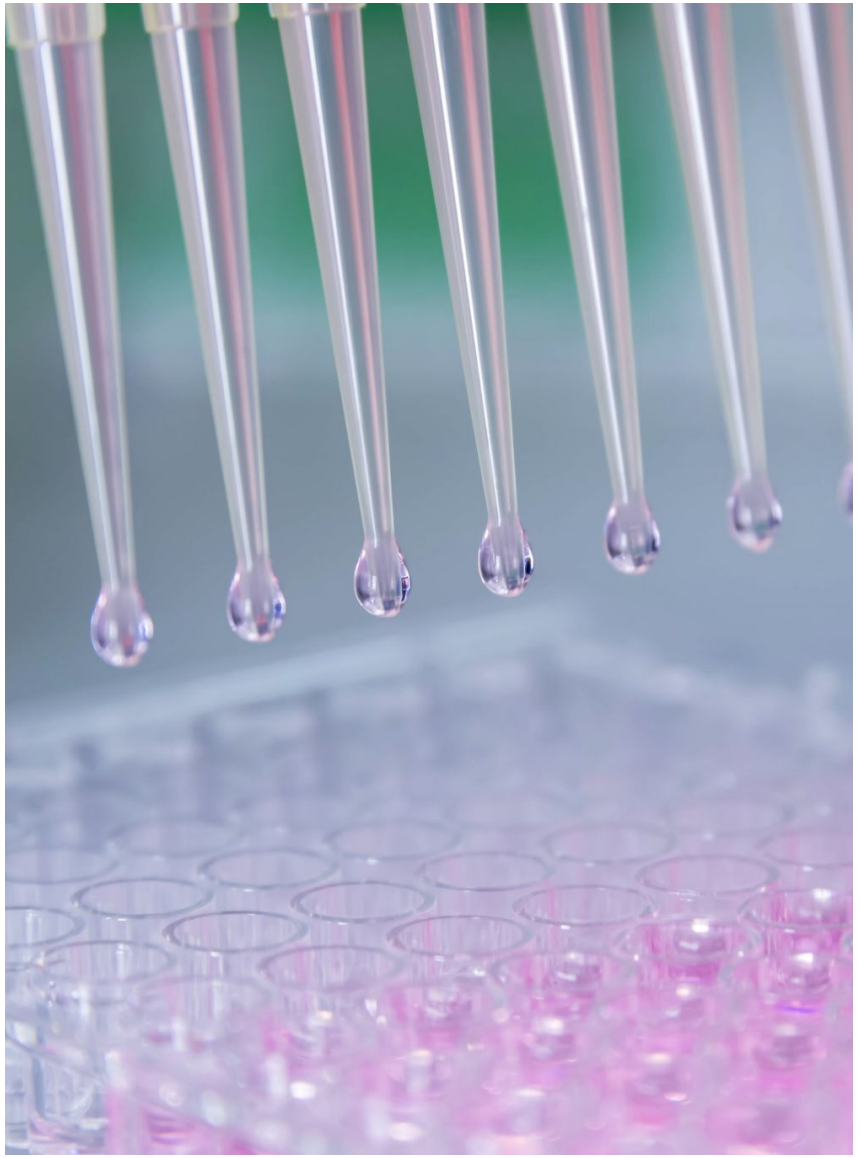


TIPHAINE.MARTIN@MSSM.EDU

R.L. Nuzzo / PM R 8 (2016) 268-272

VISUALIZATION: NOT TO FORGET

- Give information about number of replicates
  - **Biological** replicates and **technical** replicates ?



TIPHAINE.MARTIN@MSSM.EDU

BIOLOGICAL AND TECHNICAL REPLICATE?

TIPHAINE.MARTIN@MSSM.EDU

A **technical** replicate is when you test the same sample multiple times - it's used to test the **variability in the testing protocol itself**.

A **biological** replicate is where you perform the same test on multiple samples of the same material / type of cells / tissue. The samples are different (e.g. not the same day) but are expected to be very similar (if not identical) with regard to the test. Biological replicates are used to test the **variability between samples** that were selected on the basis of being otherwise identical.

For example, if you did an RNA prep for a cell sample and placed part of that prep in multiple wells of a microtiter plate, that's a **technical replicate**.

If you then took multiple samples (eg. Different days) of the cells and made an RNA prep for each and plated them, then those are **biological replicates**. Examining the technical replicates allows you to quantify the variation internal to a test (say qPCR) and while the later would allow you to see variation in the prep procedure or sample-taking.

VISUALIZATION: NOT TO FORGET IN THE LEGEND

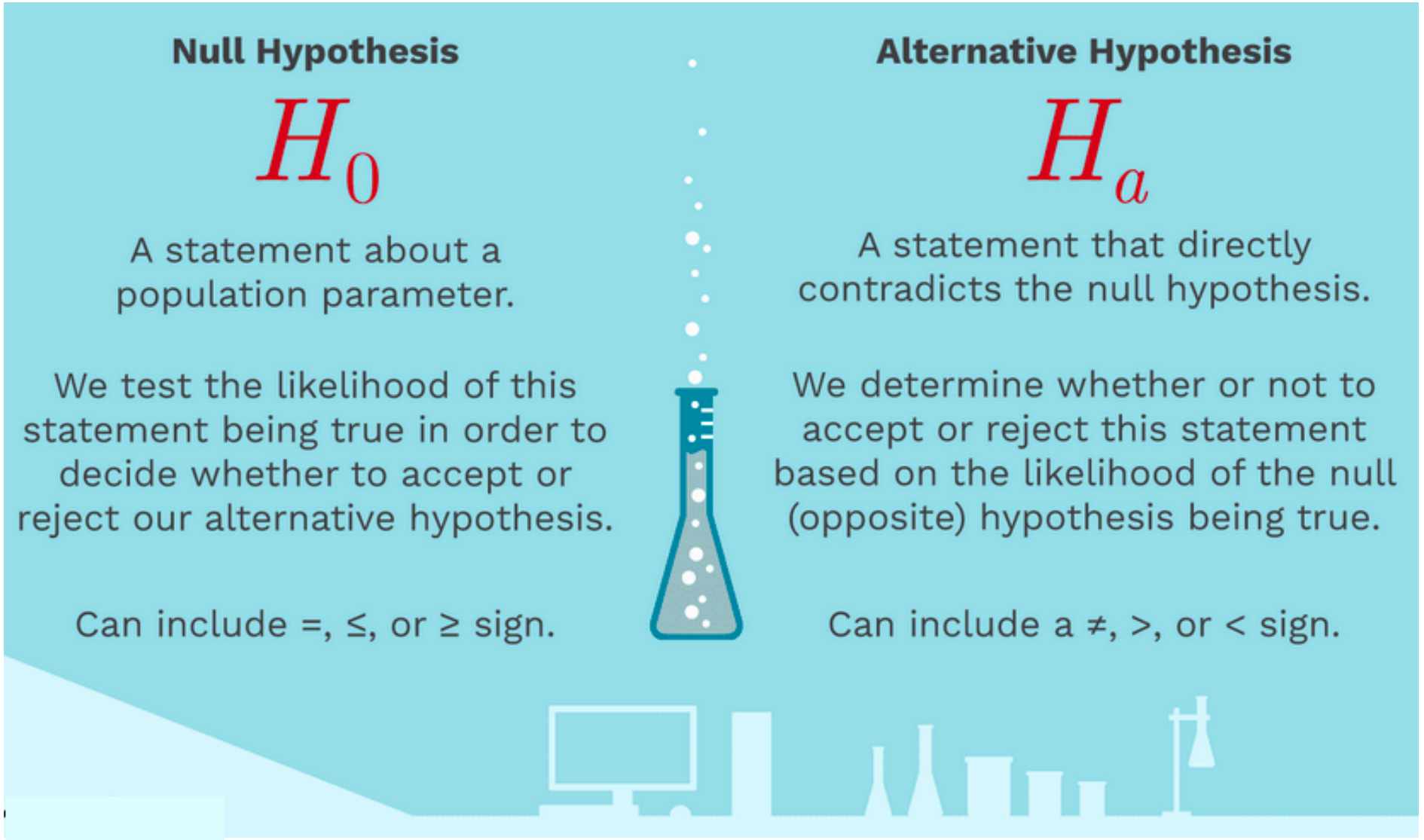
- Inform about the choice of the statistical methods
  - **Paired- or non-paired methods**
  - **One-tailed tests or two-tailed tests**
  - **Value of P-value** and not *. **, *** or NS
  - If multiple testing, provide **adjusted Pvalue** using which method or nominal Pvalue (put information in the legend)

TIPHAINE.MARTIN@MSSM.EDU

STATISTICAL ANALYSIS

STEPHAINE.MARTIN@MSSM.EDU

WHAT IS THE HYPOTHESIS NULL VS ALTERNATIVE?



TIPHAINE.MARTIN@MSSM.EDU

TESTS BETWEEN TWO GROUPS

Student's T test

- Assumption :
  - normal distribution
  - Unrelated data
- Need to have large data
- a normal distribution

- If paired data,
  - use **paired sample *t*-test**

Wilcoxon-Mann-Whitney's test

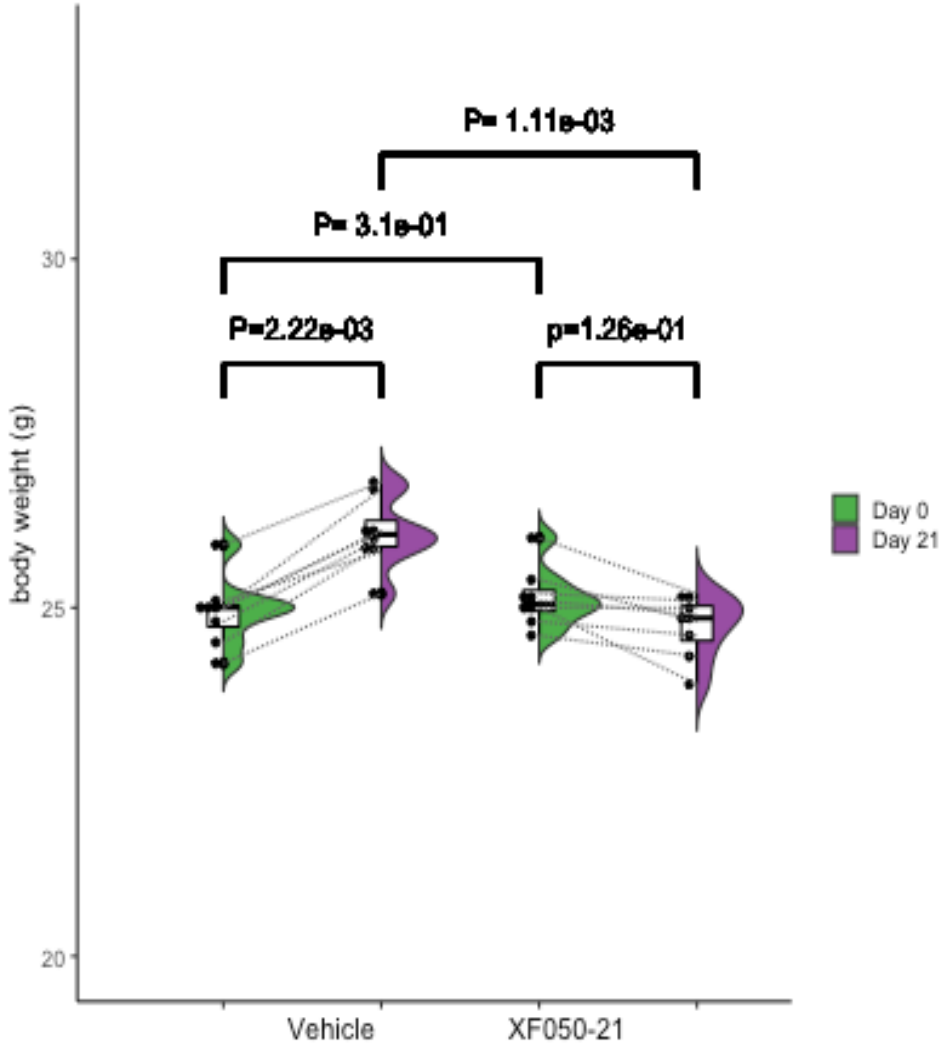
- Assumption :
  - Unrelated data
  - No assumption on the distribution of data
- Non-parametric test

- If paired data,
  - use **Wilcoxon signed-rank test**

TIPHAINE.MARTIN@MSSM.EDU

EXAMPLE OF VISUALIZATION FOR PAIRED-DATA

Which test(s) did I use here ?



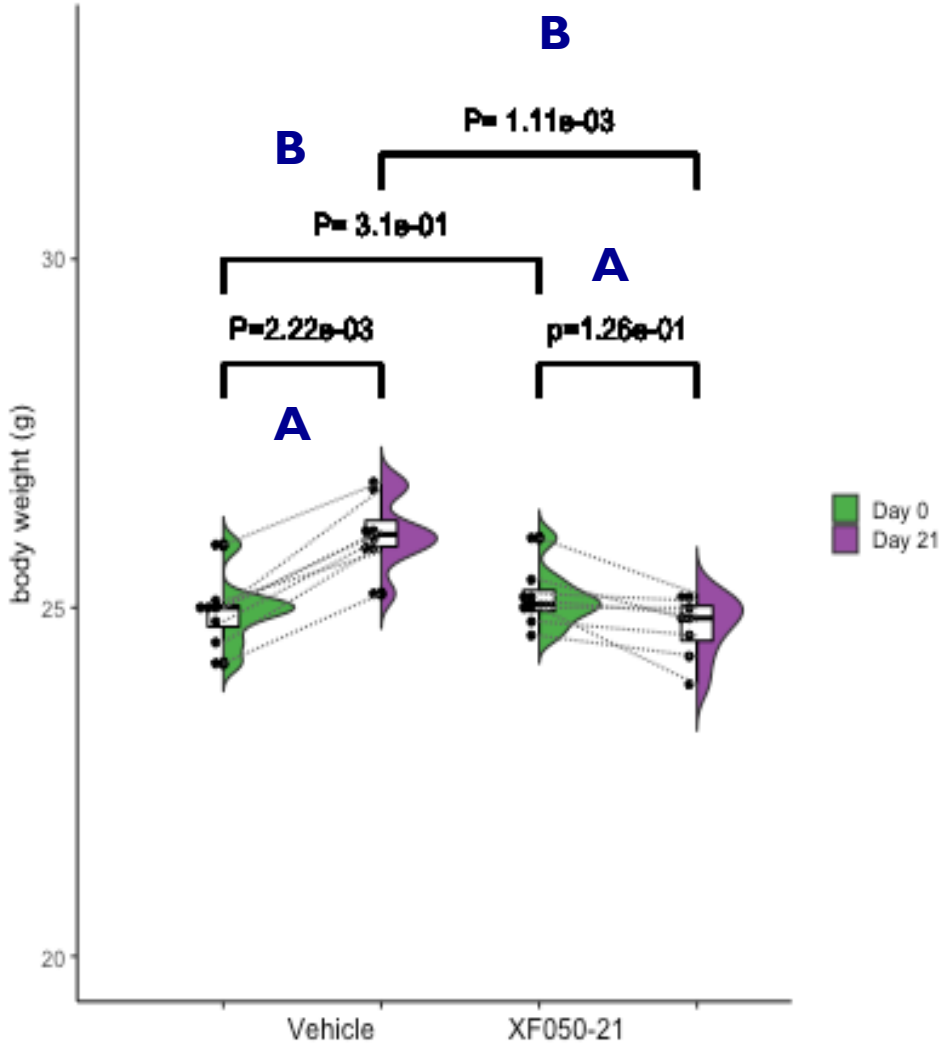
TIPHAINE.MARTIN@MSSM.EDU

EXAMPLE OF VISUALIZATION FOR PAIRED-DATA

Which test(s) did I use here ?

A) Two-tailed Wilcoxon signed-rank test

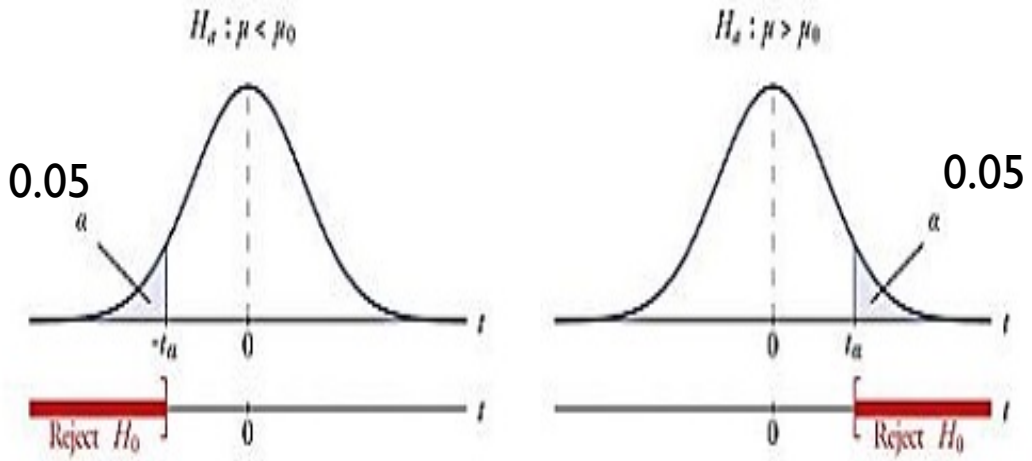
B) Two-tailed Wilcoxon-Mann-Whitney's test



TIPHAINE.MARTIN@MSSM.EDU

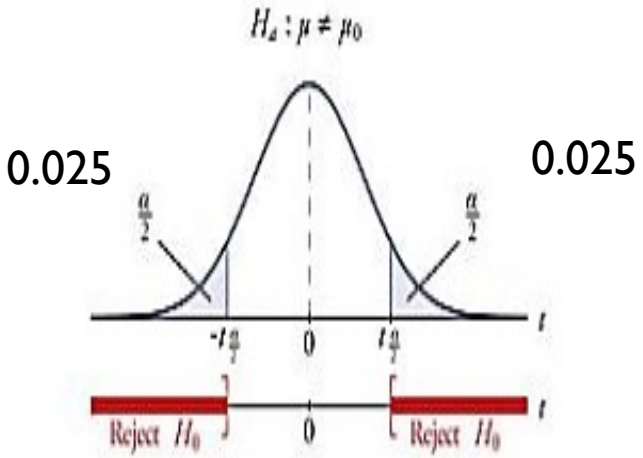
ONE-TAILED TEST VERSUS TWO-TAILED TEST

- One-tailed tests:
  - "is the sample mean greater than u or P_u?"
  - "is the sample mean less than u or P_u?"



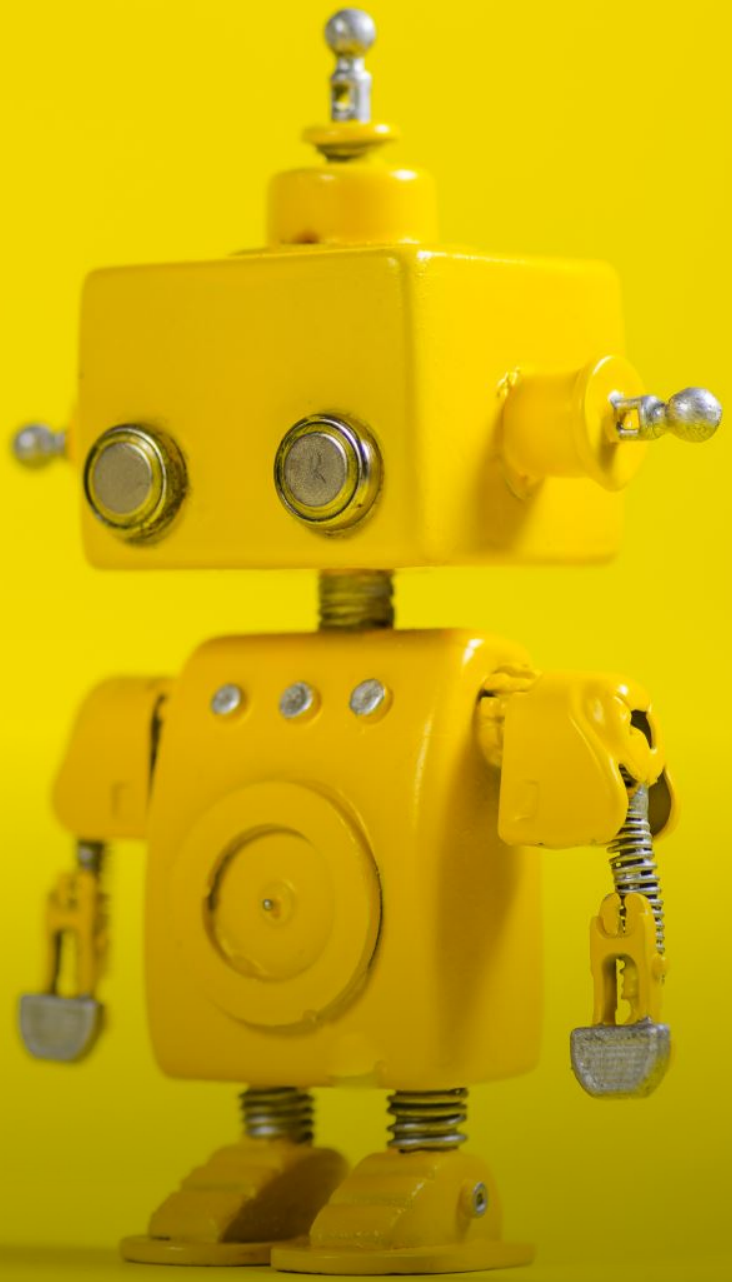
Is the hypothesis null or alternative written here?

- Two-tailed test:
  - "is there a significant difference?"

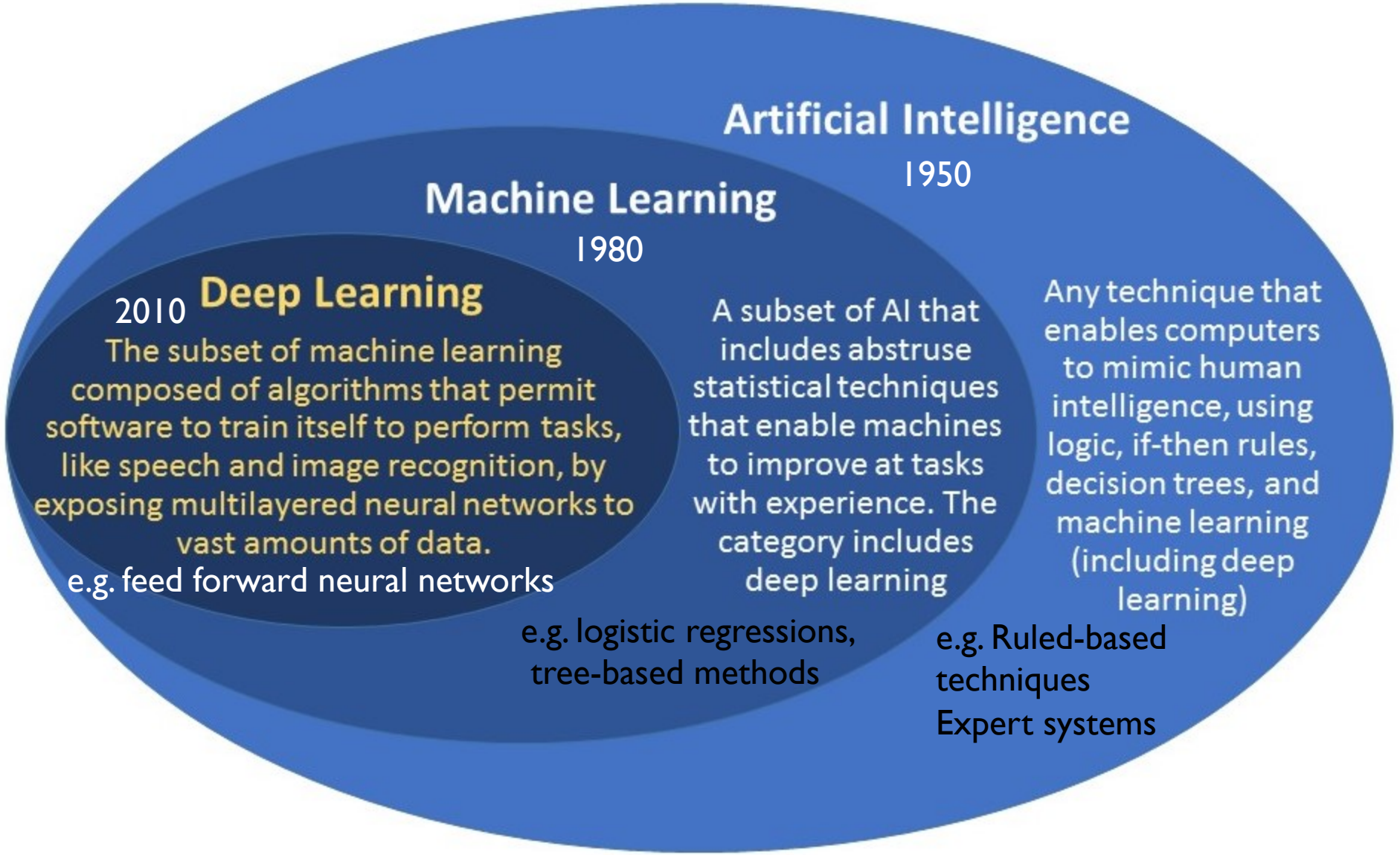


TIPHAINE.MARTIN@MSSM.EDU

ARTIFICIAL INTELLIGENCE (AI)



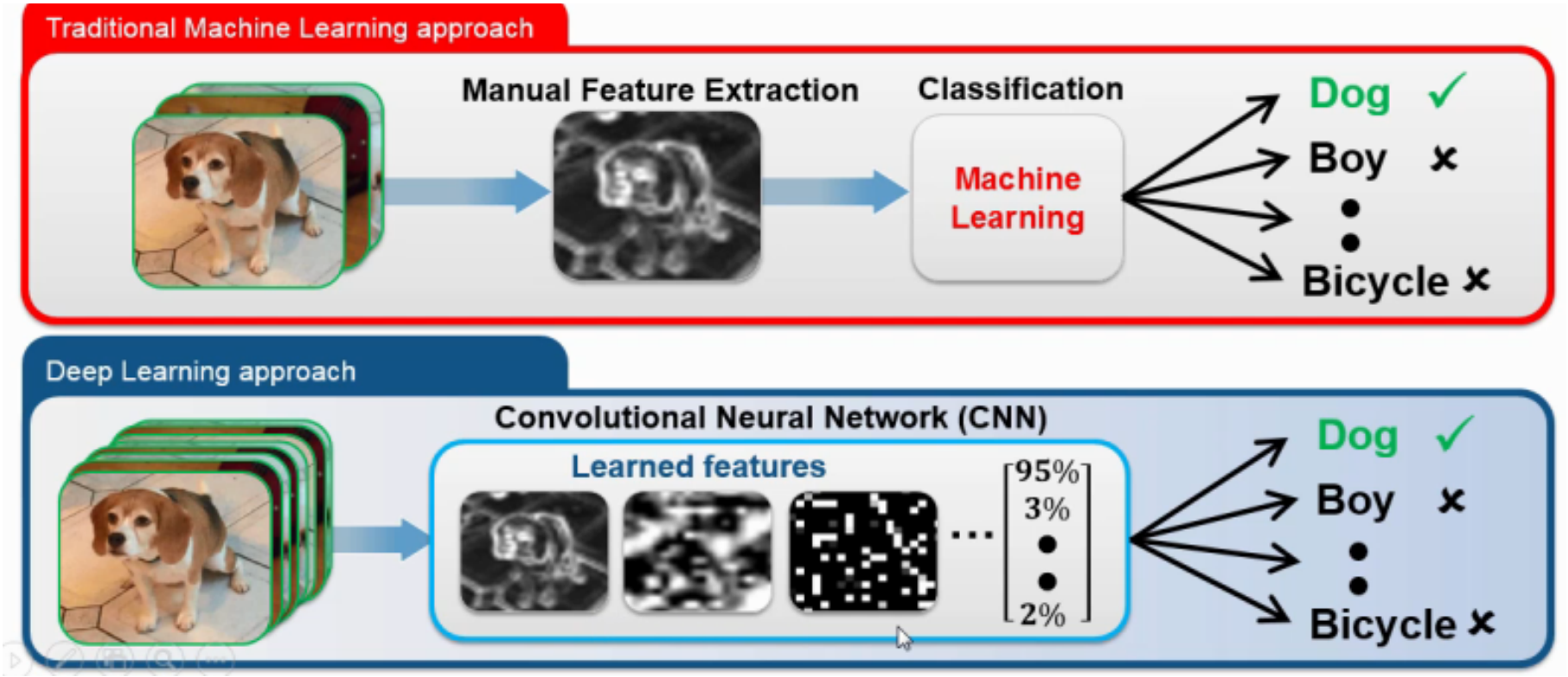
TIPHAINE.MARTIN@MSSM.EDU



TIPHAINE.MARTIN@MSSM.EDU

MACHINE LEARNING (ML) VERSUS DEEP LEARNING (DL)

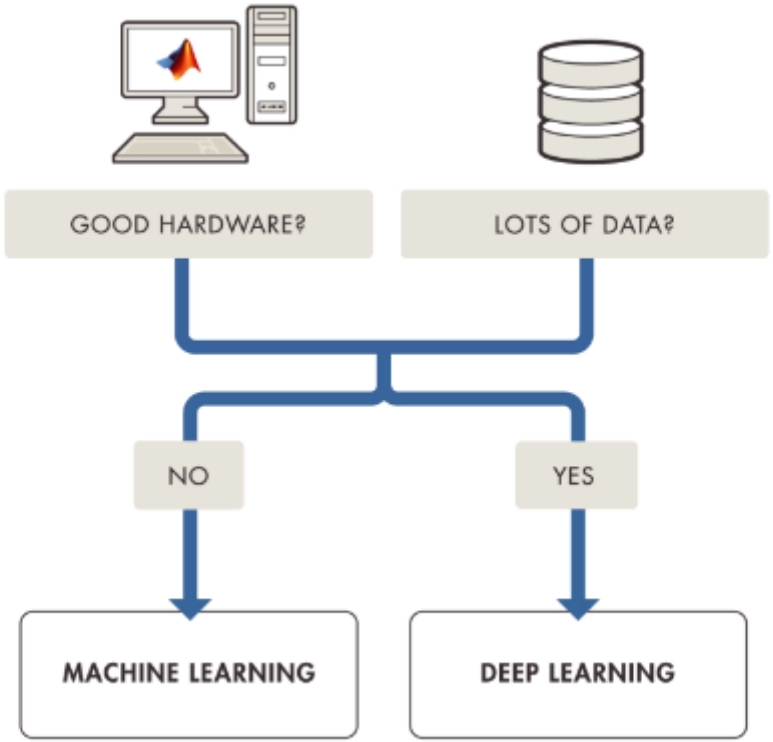
- Both could be used to classify (scores), predict (probabilities) or detect of features (binary)



TIPHAINE.MARTIN@MSSM.EDU

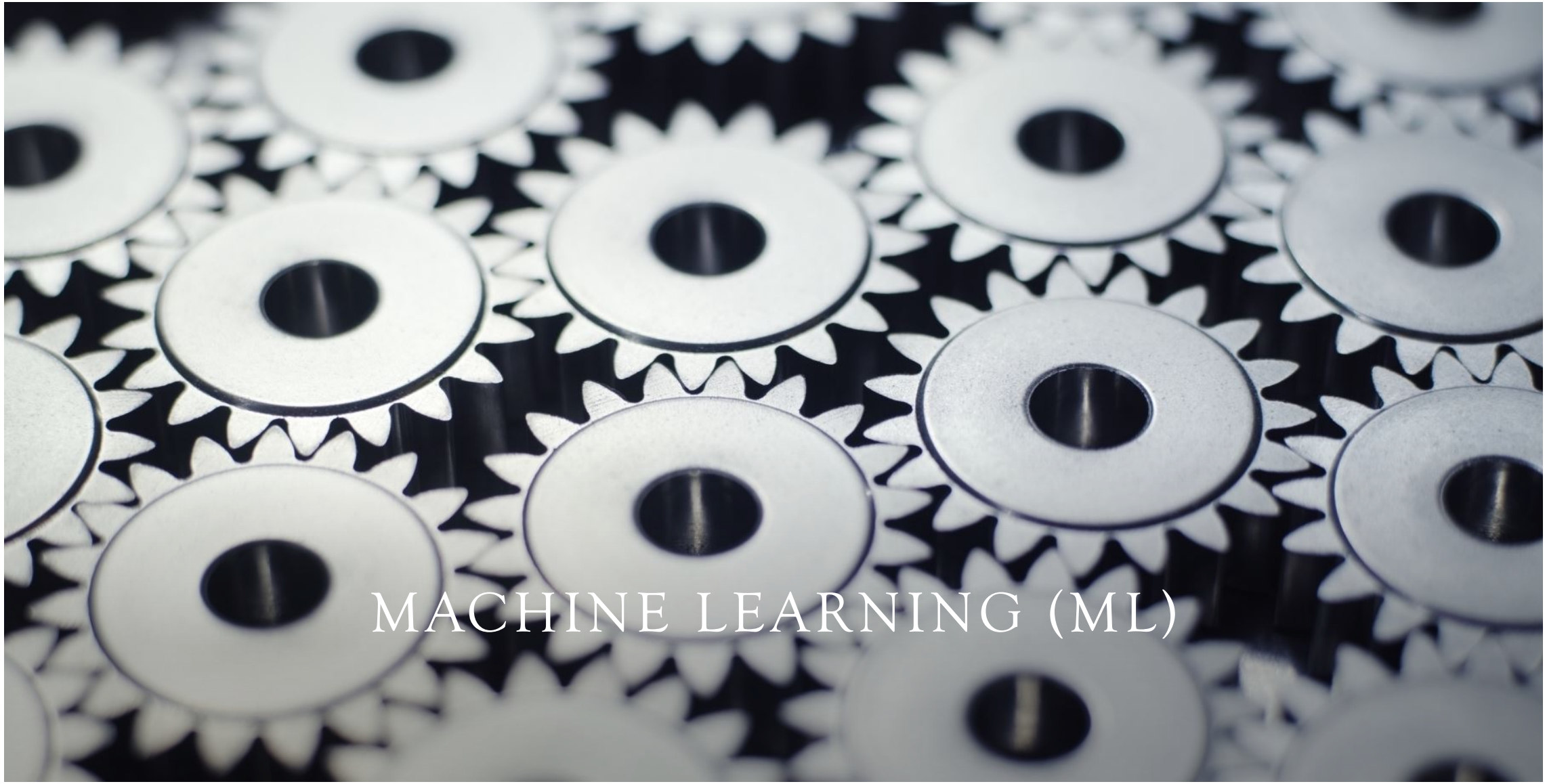
ML VERSUS DL

	Machine Learning	Deep Learning
Training dataset	Small	Large >1000
Choose your own features	Yes	No
# of classifiers available	Many	Few
Training time	Short	Long GPU



- ML : Find the good features is time consuming and can be challenging
- DL : have enough pictures to train neural networks and training is time consuming as well as the definition of layers can be challenging

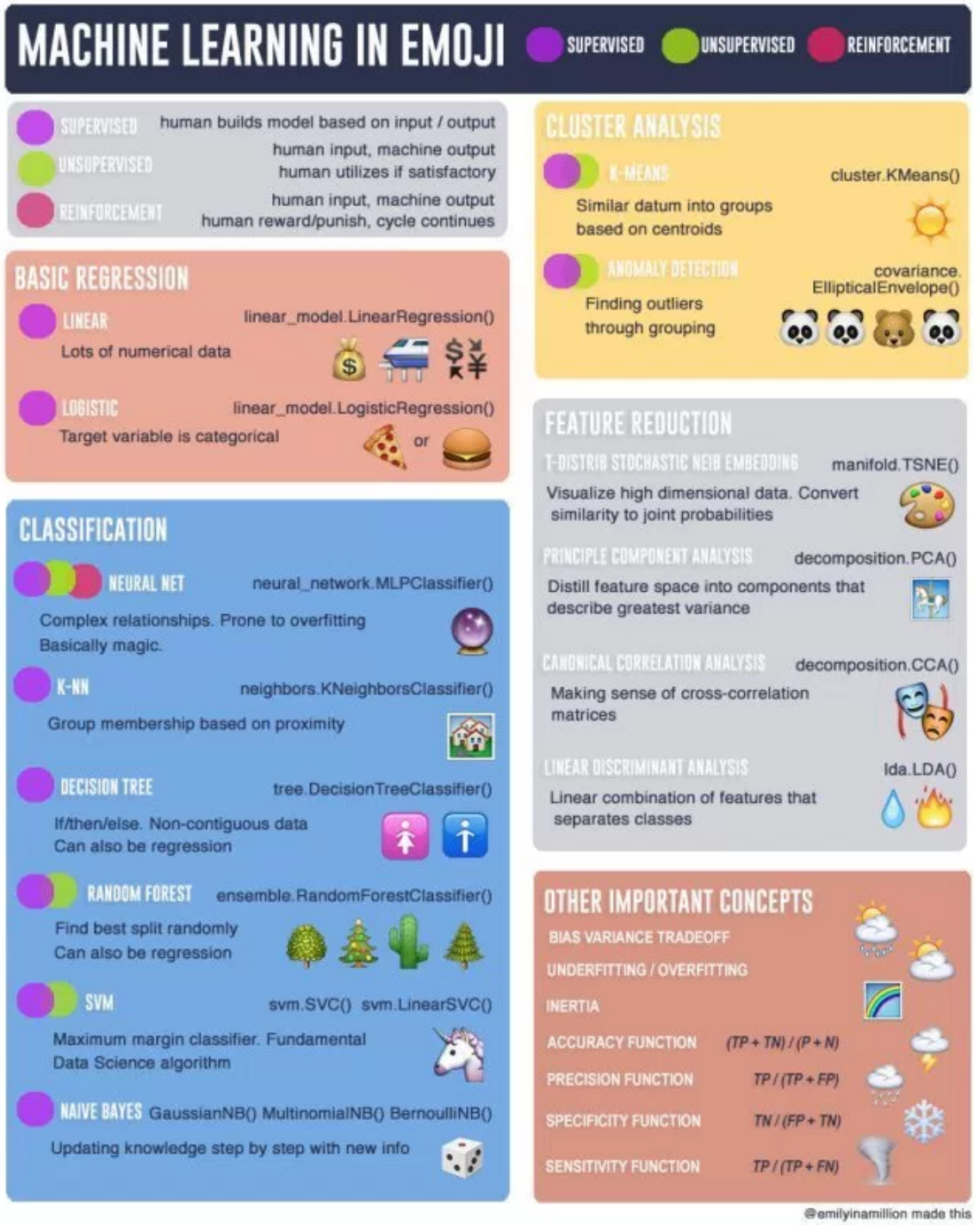
TIPHAINE.MARTIN@MSSM.EDU



MACHINE LEARNING (ML)

TIPHAINE.MARTIN@MSSM.EDU

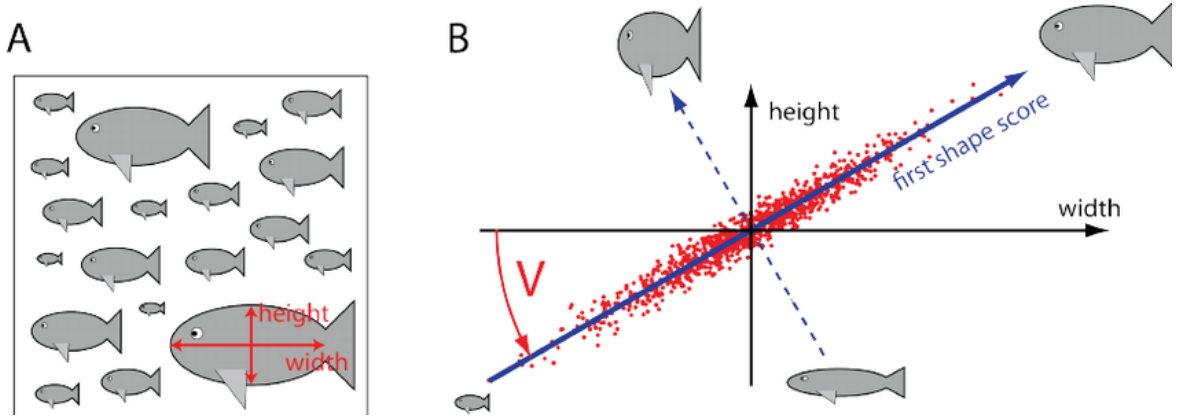
You are like Mr. Jourdain in "*The Bourgeois Gentleman*" *(Moliere, 1670) with the prose, but instead for you, this is ML*



TIPHAINE.MARTIN@MSSM.EDU

PRINCIPAL COMPONENT ANALYSIS (PCA)

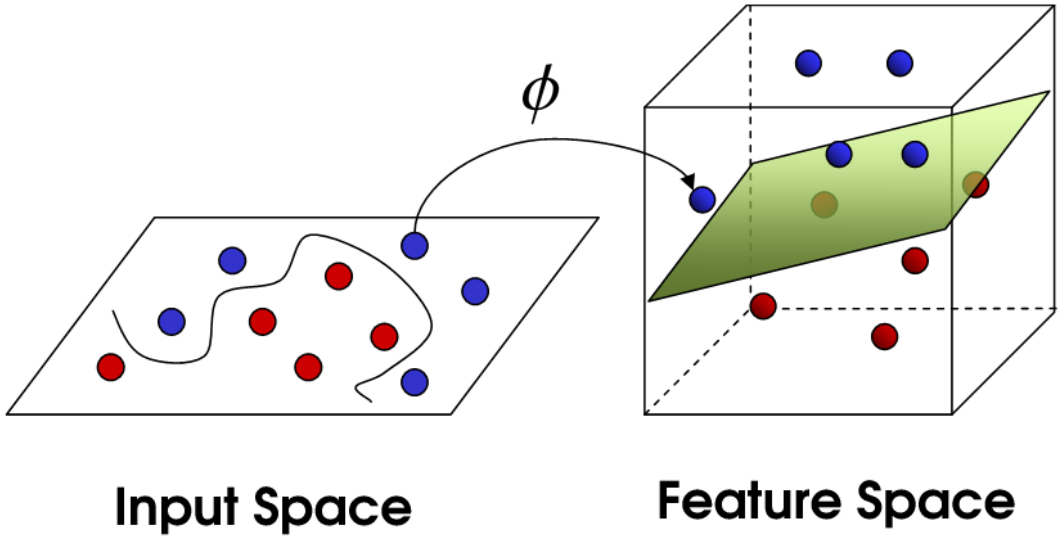
- **Unsupervised**
- **Dimension reduction**
- **Orthogonal transformation** to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables
- such that the **greatest variance** by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on

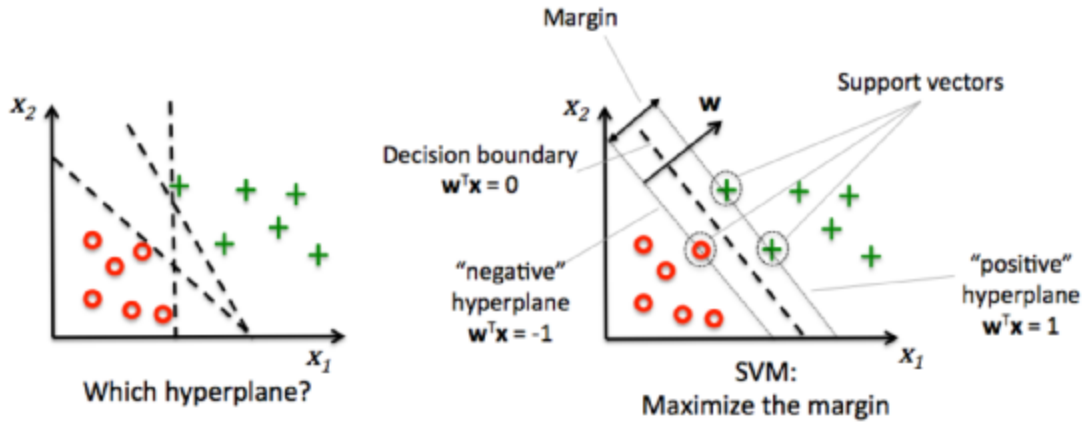


TIPHAINE.MARTIN@MSSM.EDU

SUPPORT VECTOR MACHINE (SVM)

- **Supervised**
- used for classification and regression analysis
- Model linear or not linear





TIPHAINE.MARTIN@MSSM.EDU

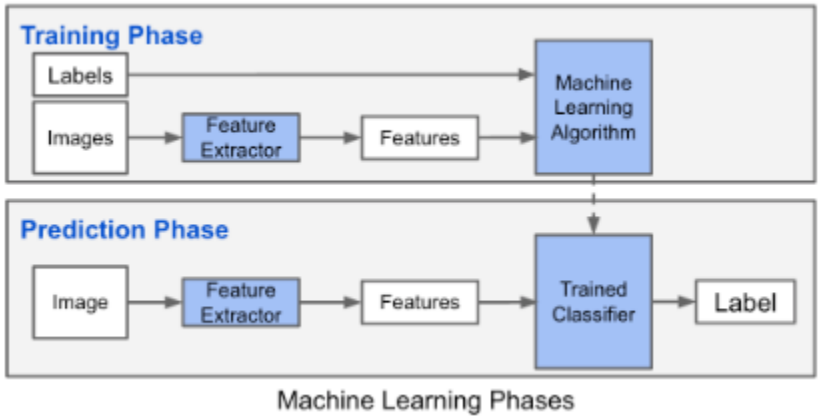
MACHINE LEARNING TO DEEP LEARNING

TYPE	NAME	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Linear	Linear regression	**The "best fit" line** through all data points. Predictions are numerical.	Easy to understand -- you clearly see what the biggest drivers of the model are.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
	Logistic regression	The adaptation of **linear regression** to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	X Sometimes too simple to capture complex relationships between variables. X Tendency for the model to "overfit".
Tree-based	Decision tree	A graph that uses a **branching method** to match all possible outcomes of a decision.	Easy to understand and implement.	X Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
	Random Forest	Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but **by combining them we get better overall performance.**	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	X Can be slow to output predictions relative to other algorithms. X Not easy to understand predictions.
	Gradient Boosting	Uses even weaker decision trees, that are increasingly **focused on "hard" examples.**	High-performing.	X A small change in the feature set or training set can create radical changes in the model. X Not easy to understand predictions.
Neural networks	Neural networks	Mimics the behavior of the brain. Neural networks are interconnected neurons that pass messages to each other. Deep learning uses several **layers of neural networks put one after the other.**	Can handle extremely complex tasks - no other algorithm comes close in image recognition.	X Very, very slow to train, because they have so many layers. Require a lot of power. X Almost impossible to understand predictions.

TIPHAINE.MARTIN@MSSM.EDU

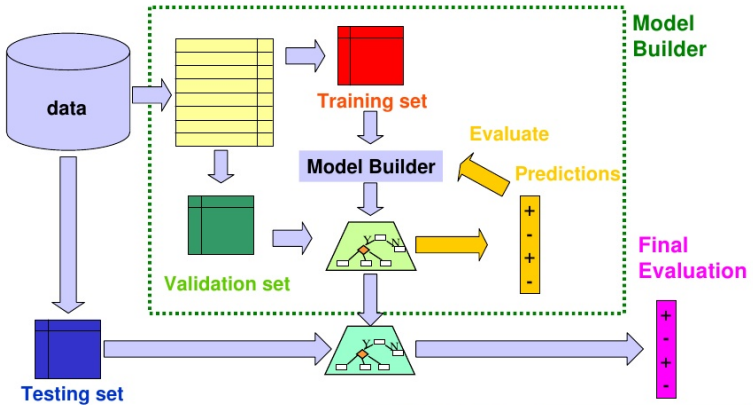
MACHINE LEARNING

- Classification/Prediction/Detection :
  - Unsupervised (unlabeled data)
  - Reinforced (reward)
  - Supervised (labeled data)
- Need to define and extract features
- Training/validation/test



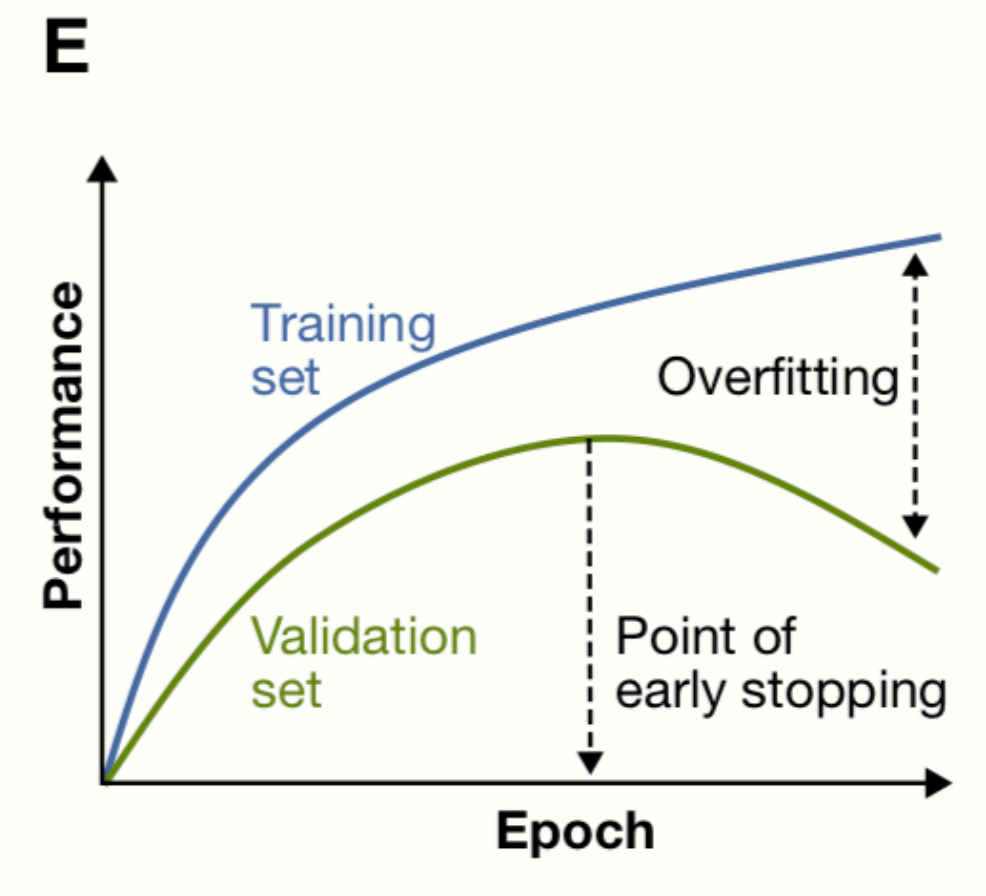
Machine Learning Phases

Traditional Machine Learning Flow



TIPHAINE.MARTIN@MSSM.EDU

OVERFITTING

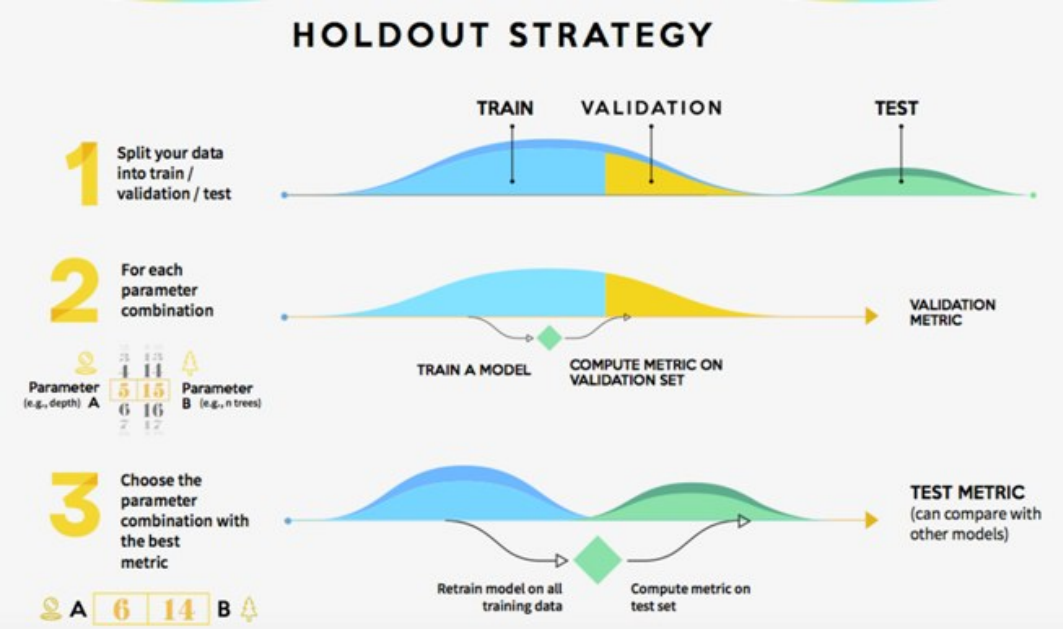


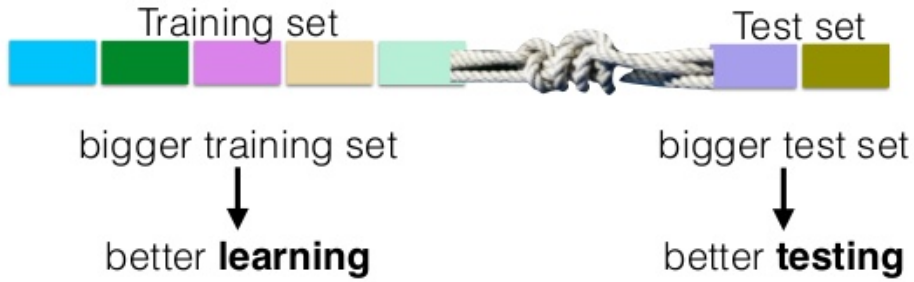
- Results from a too complex model relative to the size of the training set

- Reducing this issue by
  - decreasing the model complexity, or
  - increasing the size of the training set

TIPHAINE.MARTIN@MSSM.EDU

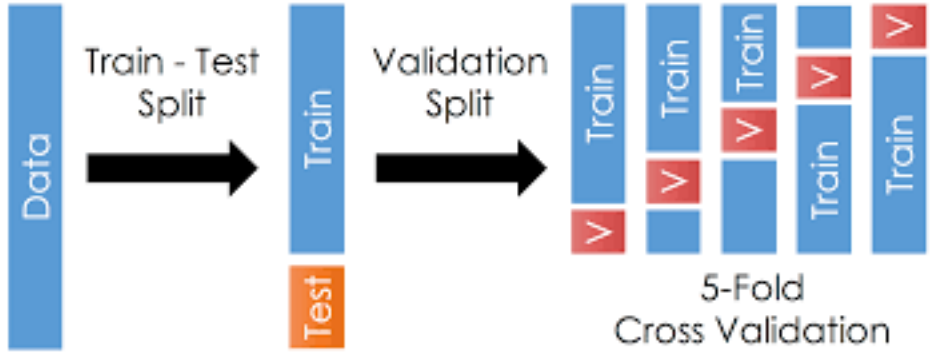
Cross-validation (computational)

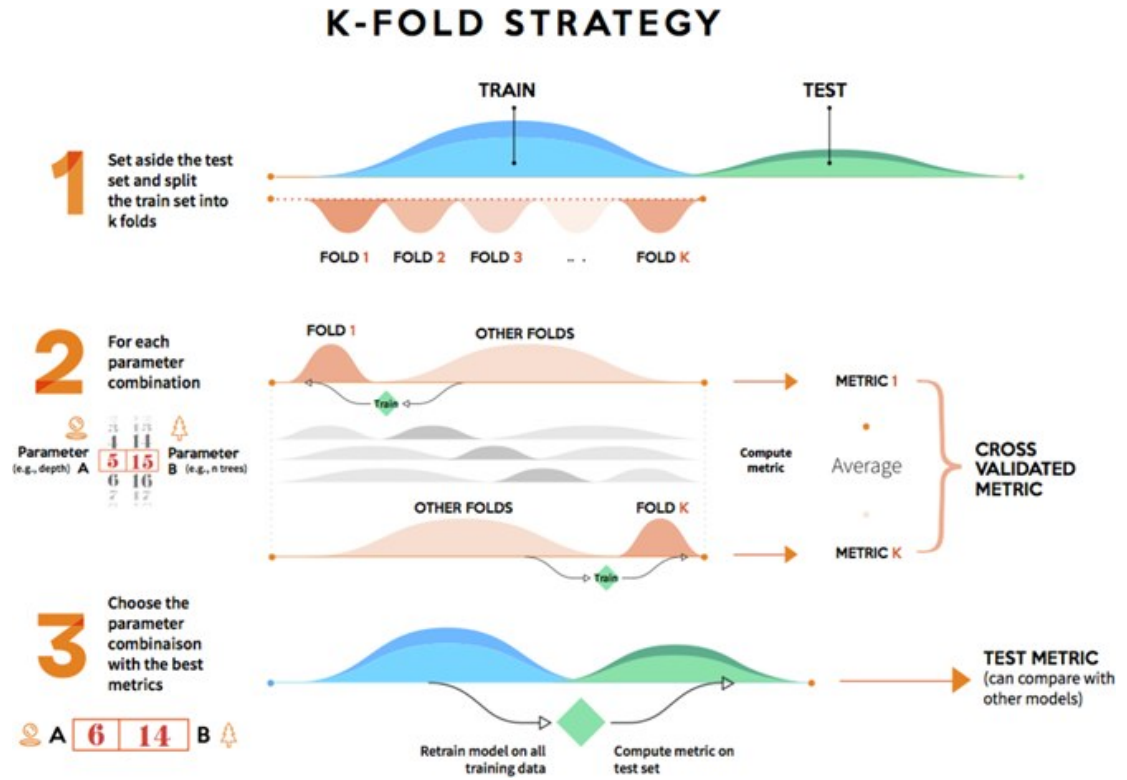




Key: Train & test sets must be **disjoint.**
And the dataset or sample size is fixed.
They grow at the expense of each other!

cross-validate to maximize both





TIPHAINE.MARTIN@MSSM.EDU