Data Science and Machine Learning - A Practical Approach

10 Recommendations for Machine Learning Projects

Hanna Bugler
Biomedical Engineering PhD Student
Harris Lab and AI^2 Lab

Winter 2025

UNIVERSITY OF CALGARY

Overview

Learning Goals

1. **Data Science**

The importance of exploring the data

2. **Python Programming**

Less is more, keep it organized

3. **Machine Learning**

Make the most of your resources

UNIVERSITY OF CALGARY

Learning Objectives

- Learn to how to critically assess and mitigate bias throughout the machine learning pipeline
- Learn about available Python and software resources for better data management in machine learning projects
- Learn the key features of scientific figures necessary for effective communication

#1 Data Science: Exploratory Data Analysis (EDA)

- What is EDA and why is it important?

#1 Data Science: Exploratory Data Analysis (EDA)

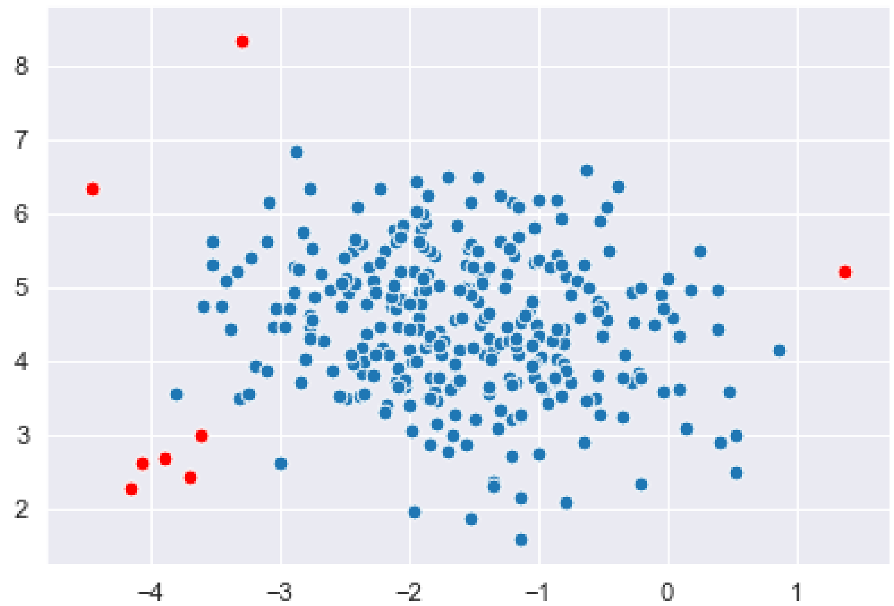
- What is EDA and why is it important?
  - **Data Cleaning** – knowing <u>when</u> and <u>why</u> a sample does not belong and how to account for it

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

	School ID	Name	Address	City	Subject	Marks	Rank	Grade
0	101.0	Alice	123 Main St	Los Angeles	Math	85.0	2	B
1	102.0	Bob	456 Oak Ave	New York	English	92.0	1	A
2	103.0	Charlie	789 Pine Ln	Houston	Science	78.0	4	C
3	NaN	David	101 Elm St	Los Angeles	Math	89.0	3	B
4	105.0	Eva	$\boxed{\text{NaN}}$	Miami	History	$\boxed{\text{NaN}}$	8	D
5	106.0	Frank	222 Maple Rd	$\boxed{\text{NaN}}$	Math	95.0	1	A

Elliptic Envelope Outlier Detection



https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/ https://www.geeksforgeeks.org/ml-handling-missing-values/ https://towardsdatascience.com/4-machine-learning-techniques-for-outlier-detection-in-python-21e9cfacb81d

UNIVERSITY OF CALGARY

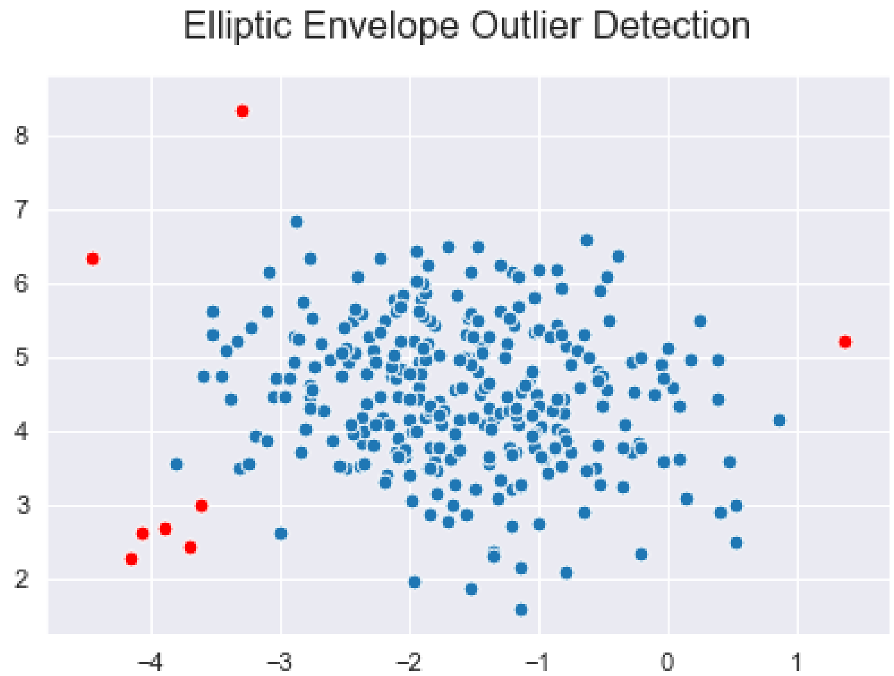
#1 Data Science: Exploratory Data Analysis (EDA)

- What is EDA and why is it important?
  - **Data Cleaning** – data imputation techniques including mean/mode, inter/extrapolation, model regression, etc.

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

	School ID	Name	Address	City	Subject	Marks	Rank	Grade
0	101.0	Alice	123 Main St	Los Angeles	Math	85.0	2	B
1	102.0	Bob	456 Oak Ave	New York	English	92.0	1	A
2	103.0	Charlie	789 Pine Ln	Houston	Science	78.0	4	C
3	NaN	David	101 Elm St	Los Angeles	Math	89.0	3	B
4	105.0	Eva	NaN	Miami	History	NaN	8	D
5	106.0	Frank	222 Maple Rd	NaN	Math	95.0	1	A

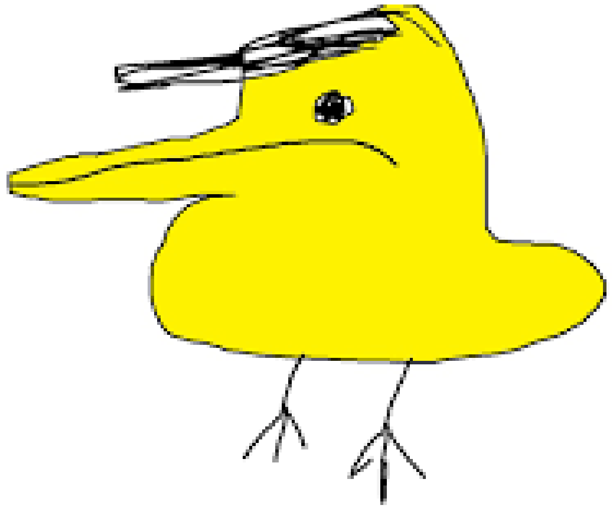


UNIVERSITY OF CALGARY

https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/ https://www.geeksforgeeks.org/ml-handling-missing-values/ https://towardsdatascience.com/4-machine-learning-techniques-for-outlier-detection-in-python-21e9cfacb81d

#1 Data Science: Exploratory Data Analysis (EDA)

- What is EDA and why is it important?
  - **Data Cleaning and Feature Engineering –** does the data follow prior assumptions, we may have prior knowledge of about the data, let's use it!



Example from the <u>Train</u> Distribution

Example from the <u>Test</u> Distribution

#1 Data Science: Exploratory Data Analysis (EDA)

- **My Challenge**: Good vs. Bad data can be easily defined by a single threshold

- **Recommendation:** Improve your domain specific knowledge (Literature & Lab mates)
- **Coding Recommendation**: Pandas_profiling

UNIVERSITY OF CALGARY

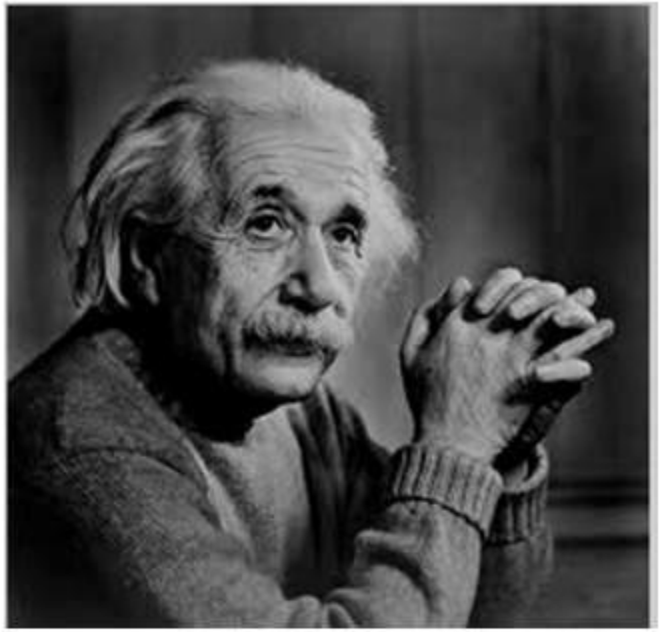
https://www.visitcalgary.com/things-to-do/stories-from-calgary/best-burgers-in-calgary

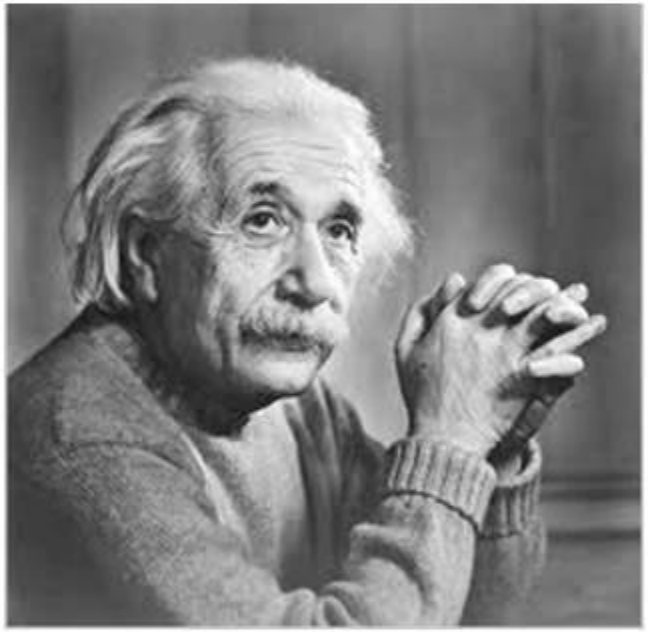
https://www.avenuecalgary.com/restaurants-food/best-burgers-in-calgary/

#2 Data Science: Processing

- What type of processing has the image undergone

(what was the transformation from its raw/original state)?



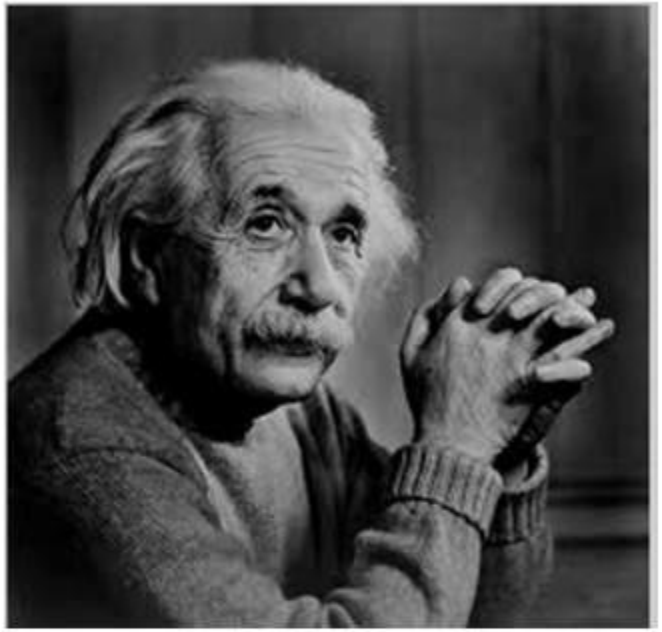


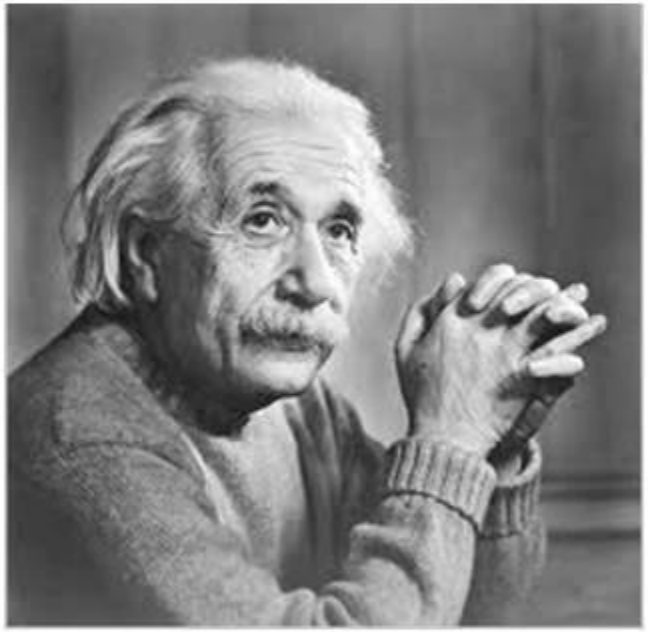
#2 Data Science: Processing

- What type of processing has the image undergone

(what was the transformation from its raw/original state)?

Contrast Enhanced





Make sure you know if your repository contains raw or preprocessed data

#2 Data Science: Processing

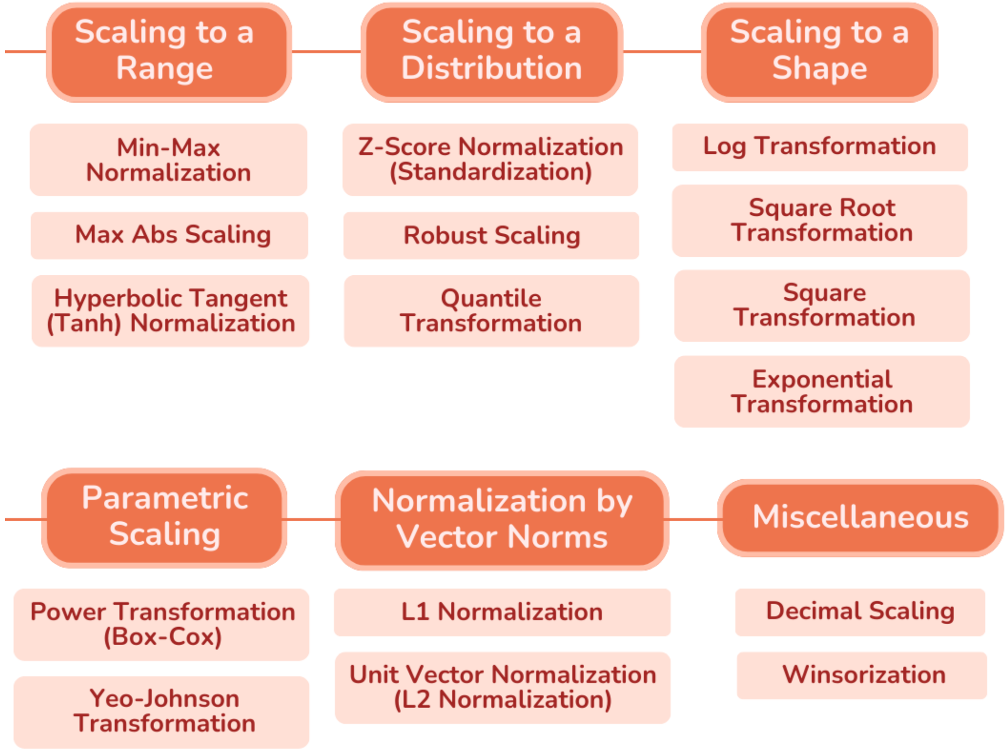
- Processing steps can vary by:
  - Data Representation
    - Pixel Scaling vs. Whitespace/Hidden character handling
  - Data Type
    - Complex vs. Magnitude Value Scaling
  - Group Dependent
    - Normalization vs. Standardization
  - Vendor Specific
    - Vendors often have their own proprietary pipelines

CT Orthopedic Segmentation

https://www.rsipvision.com/ct-segmentation-orthopedic-surgery/

#2 Data Science: Processing

- **My Challenge**: Unintended effects of 'standard' technique



- **Recommendation:** Understand the transformation and its repercussions on multiple input types
- **Coding Recommendation**: SciPy

UNIVERSITY OF CALGARY

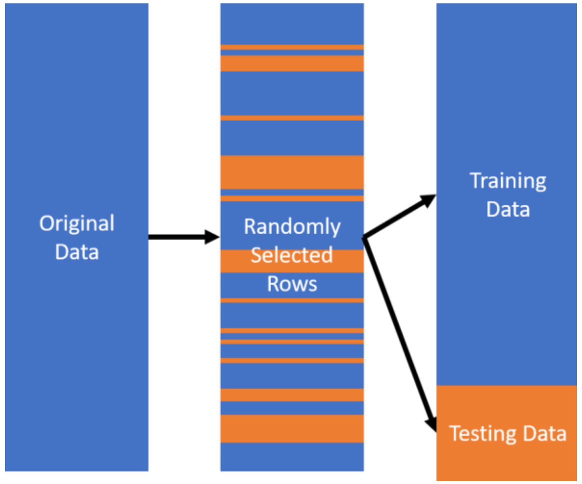
https://medium.com/@reinapeh/16-data-feature-normalization-methods-using-python-with-examples-part-1-of-3-26578b2b8ba6

#3 Data Science: Data Bias Mitigation

- What are some sources of bias in data splitting and feature engineering?

#3 Data Science: Data Bias Mitigation

- Data Splitting
  - **Leakage**: information is shared across training and testing → split data at lower levels
  - **Imbalanced datasets**: challenged by condition frequency and study recruitment protocols → k-fold cross-validation
  - **Hyperparameter overfitting**: when optimization is performed on test set → nested k-fold cross-validation

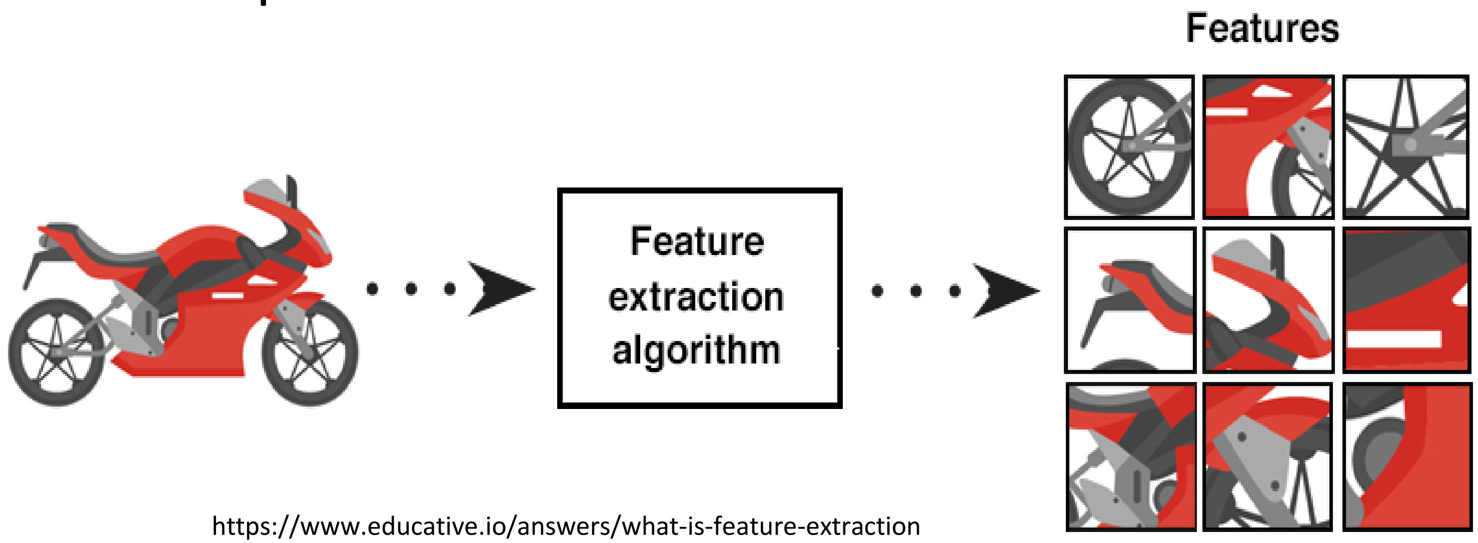


https://conlanscientific.com/posts/category/blog/post/avoiding-data-leakage-machine-learning/

UNIVERSITY OF CALGARY

#3 Data Science: Data Bias Mitigation

- Feature Engineering
  - **Feature removal**: while intuitive, it may remove useful patterns leveraged by the model → rigor, test different feature sets
  - **Feature rescaling**: normalization vs. standardization to improve learning → remove outliers, determine necessity
  - **Missing data**: may not be applied fairly across demographics → consider imputation vs. removal

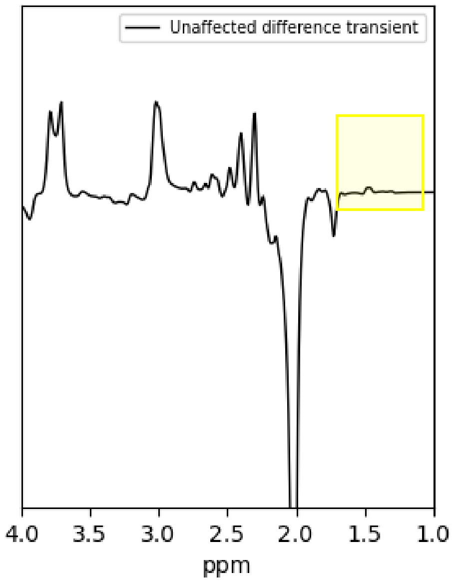


https://www.educative.io/answers/what-is-feature-extraction

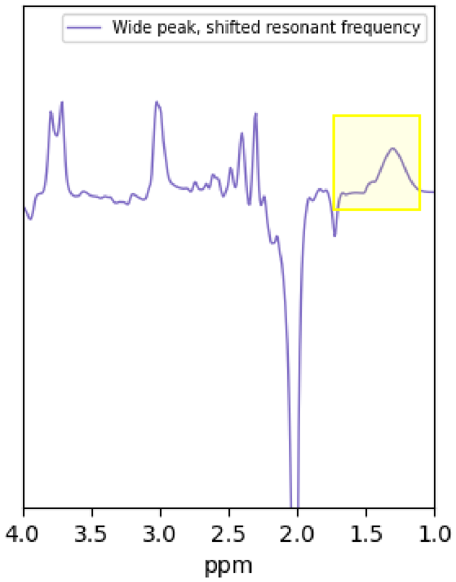
UNIVERSITY OF CALGARY

#3 Data Science: Data Bias Mitigation

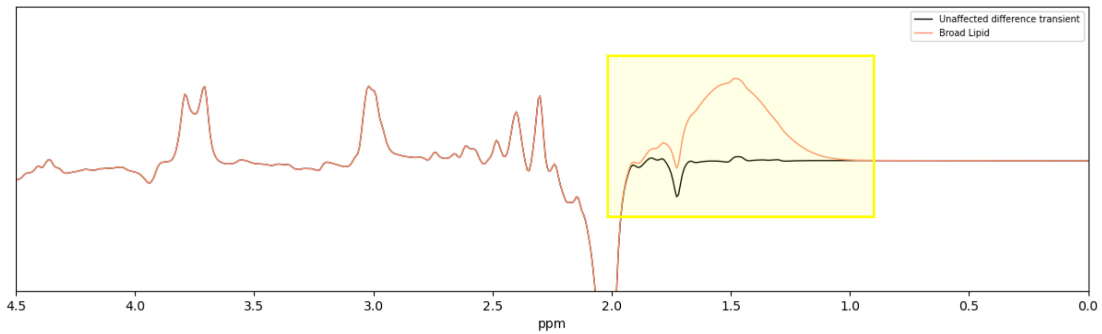
- **My Challenge**: Contaminated samples are unusable – discard them



Artifact not present



Artifact does not affect other peaks - <u>keep</u>



Artifact begins to affect other peaks - <u>discarded</u>

- **Recommendation**: Use your EDA to apply different processing techniques
- **Coding Recommendation**: Matplotlib/Seaborn

#4 Python Programming: Coding Practices

- If code is being reused, write FUNCTIONS

```
# Global and Local variables in a Function
# Declaring a global variable
x=5

# Defining the function
def FunctionLocalGlobal():
    # Creating a local variable
    y=x+10
    print('value of x is:',x)
    print('value of y is:',y)

# Calling the function
FunctionLocalGlobal()
```

```
value of x is: 5
value of y is: 15
```

https://thinkingneuron.com/user-defined-functions-in-python/

- ✓ Reduces code length
- ✓ Reduces typing errors (typos)

- ➢ Be aware of default values for non-user defined functions

#4 Python Programming: Coding Practices

- If processing large amounts of data, consider
  - GPU vs. CPU operations
  - Freeing memory (garbage collection)
  - External libraries
    (such as for profiling: py-spy, scalene)

Processes | Performance | App history | Startup | Users | Details | Services

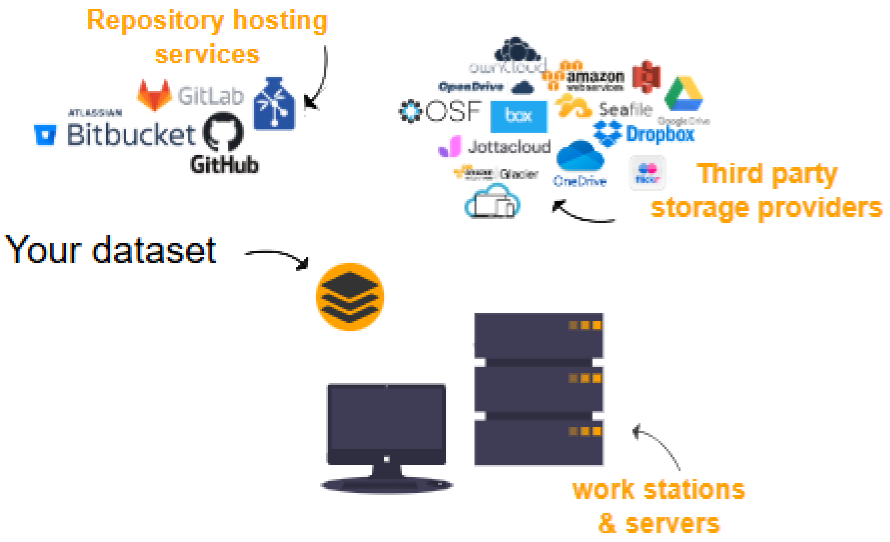
Name	14% CPU	98% Memory	17% Disk	0% Network
Microsoft Visual Studio 2013 (32...	0,5%	206,3 MB	1,5 MB/s	0 Mbps
Google Chrome (32 bit)	0,1%	178,6 MB	7,7 MB/s	0 Mbps
Google Chrome (32 bit)	7,6%	114,6 MB	0,3 MB/s	0 Mbps
Google Chrome (32 bit)	0%	95,4 MB	0,8 MB/s	0 Mbps
Google Chrome (32 bit)	0%	81,0 MB	0 MB/s	0 Mbps
Google Chrome (32 bit)	0,1%	69,5 MB	4,0 MB/s	0 Mbps
Google Chrome (32 bit)	0%	67,8 MB	0,1 MB/s	0 Mbps
Skype (32 bit)	0,5%	56,2 MB	0,7 MB/s	0 Mbps
RamMap - physical memory an...	0%	54,1 MB	0,1 MB/s	0 Mbps
Google Chrome (32 bit)	0,1%	46,3 MB	0,1 MB/s	0 Mbps
Google Chrome (32 bit)	0,2%	46,3 MB	0 MB/s	0 Mbps
Google Chrome (32 bit)	0%	38,9 MB	0 MB/s	0 Mbps
Google Chrome (32 bit)	0%	37,7 MB	0,1 MB/s	0 Mbps
Steam Client Bootstrapper (32 b...	1,1%	27,7 MB	0,1 MB/s	0 Mbps

https://superuser.com/questions/707132/100-of-ram-is-used-when-the-task-manager-shows-a-cumulative-usage-of-what-sho

UNIVERSITY OF CALGARY

#4 Python Programming: Coding Practices

- **My Challenge**: Poorly documenting code and functions



- **Recommendation:** Properly document and write reusable code
- **Coding Recommendation**: Datalad

UNIVERSITY OF CALGARY

https://handbook.datalad.org/en/latest/basics/101-138-sharethirdparty.html

#5 Machine Learning: Selecting a Library

- Which library is better?

TensorFlow

PYTORCH

#5 Machine Learning: Selecting a Library

PyTorch

Pros

- Can be quicker to edit models (experimentation)
- Efficient memory usage

Cons

- Visualization is not built-in
- Newer (2017)

```python
class NeuralNet(nn.Module):
    def __init__(self, num_of_class):
        super(NeuralNet, self).__init__()
        self.layer1 = nn.Sequential(
            nn.Conv2d(1, 16, kernel_size=5, stride=1, padding=2),
            nn.BatchNorm2d(16),
            nn.ReLU(),
            nn.MaxPool2d(kernel_size=2, stride=2))
        self.layer2 = nn.Sequential(
            nn.Conv2d(16, 32, kernel_size=5, stride=1, padding=2),
            nn.BatchNorm2d(32),
            nn.ReLU(),
            nn.MaxPool2d(kernel_size=2, stride=2))
        self.fc = nn.Linear(7 * 7 * 32, num_of_class)

    def forward(self, x):
        out = self.layer1(x)
        out = self.layer2(out)
        out = out.reshape(out.size(0), -1)
        out = self.fc(out)
        return out
```


UNIVERSITY OF CALGARY

https://builtin.com/data-science/pytorch-vs-tensorflow

#5 Machine Learning: Selecting a Library

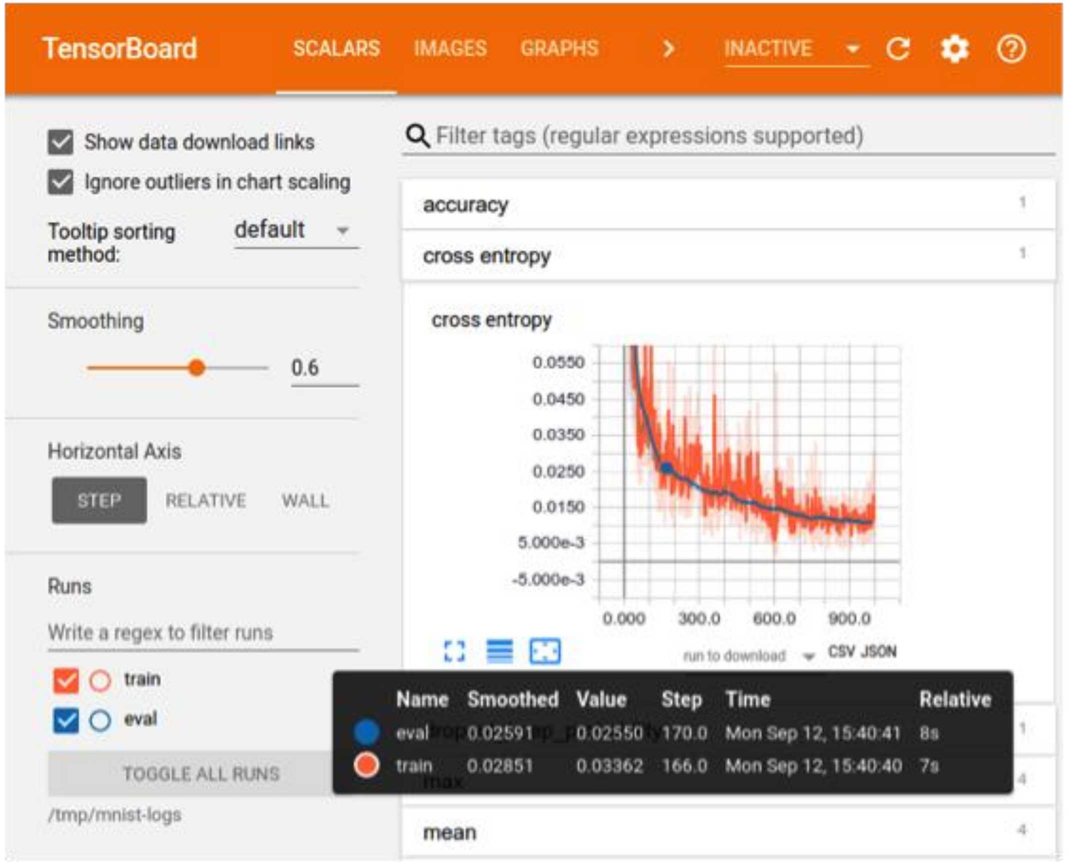
TensorFlow (Keras)

<u>Pros</u>

- Built-in visualization
- Production-ready

<u>Cons</u>

- Harder to make quick changes
- Slower implementation (distributed training)



https://builtin.com/data-science/pytorch-vs-tensorflow

#5 Machine Learning: Selecting a Library

- **My challenge**: Implementing a custom neural network layer

https://www.shutterstock.com/image-photo/brute-force-forcing-ball-into-triangular-97401623

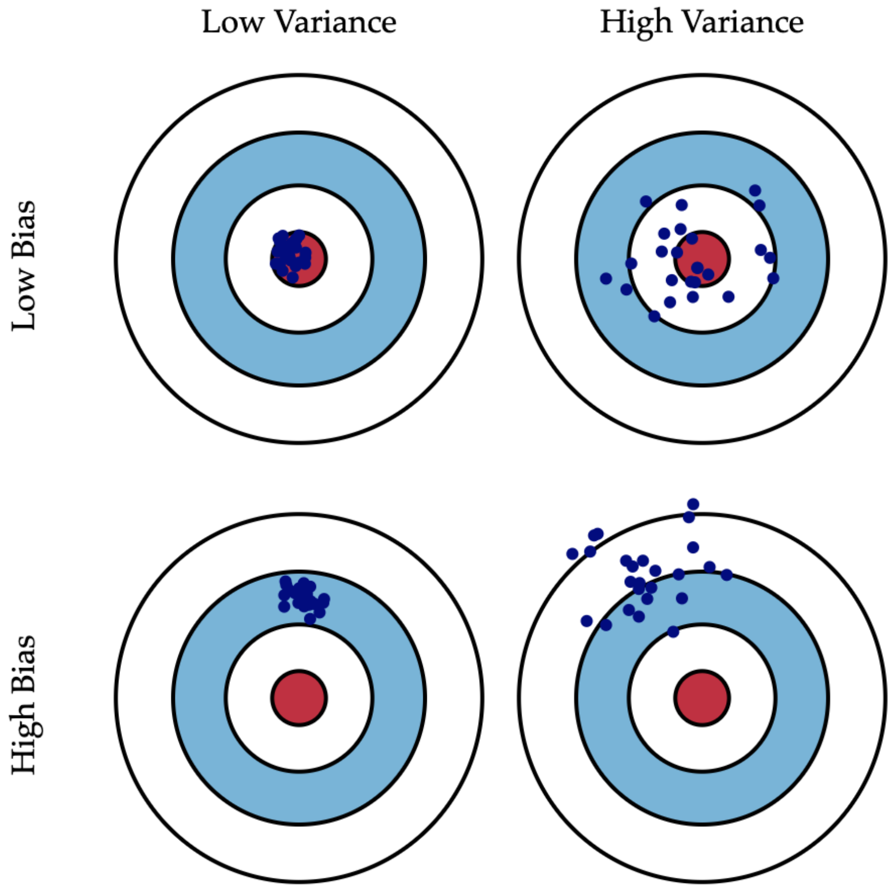
- **Recommendation:** Custom layers CAN be easier to implement in PyTorch
- **Coding Recommendation**: ____________________

#6 Machine Learning: Model Bias and Variance

- What is the difference between model bias and variance?

#6 Machine Learning: Model Bias and Variance

- What is the difference between model bias and variance?
  - **Bias** (model validity): error tends to one direction over another, also refers to fairness of a model
  - **Variance** (model reliability): oscillation of expected value that any individual sample is likely to cause

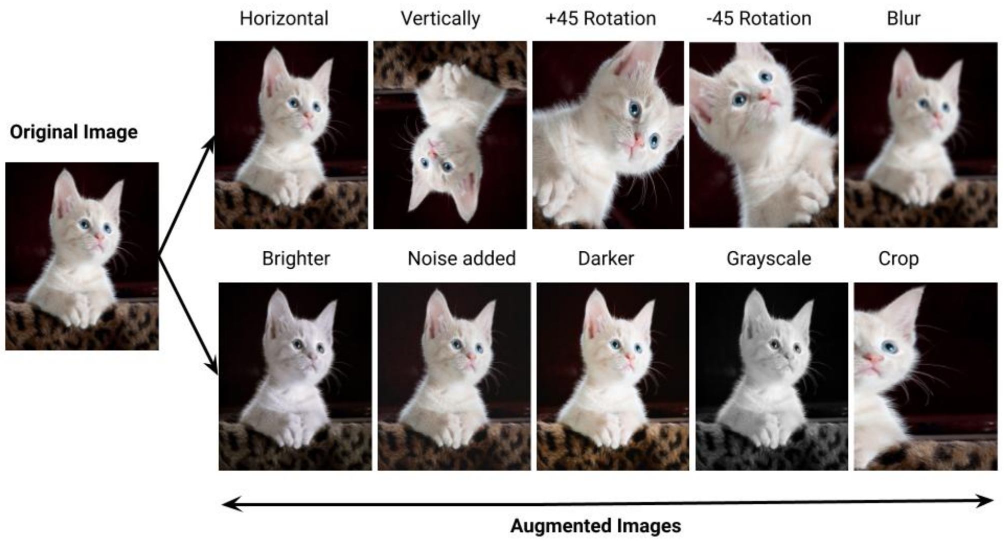


10.1148/ryai.220010

https://cristianefragata.medium.com/machine-learning-bias-and-variance-26b6ee572af

#6 Machine Learning: Model Bias and Variance

- Strategies
  - **Dropout**: increase robustness while reducing node interdependence to reduce variance
  - **Early stopping**: stop training when generalization error is minimized to reduce mathematical bias
  - **Data augmentation**: geometric transformations (need consideration for how will it interfere with the application) to reduce fairness bias

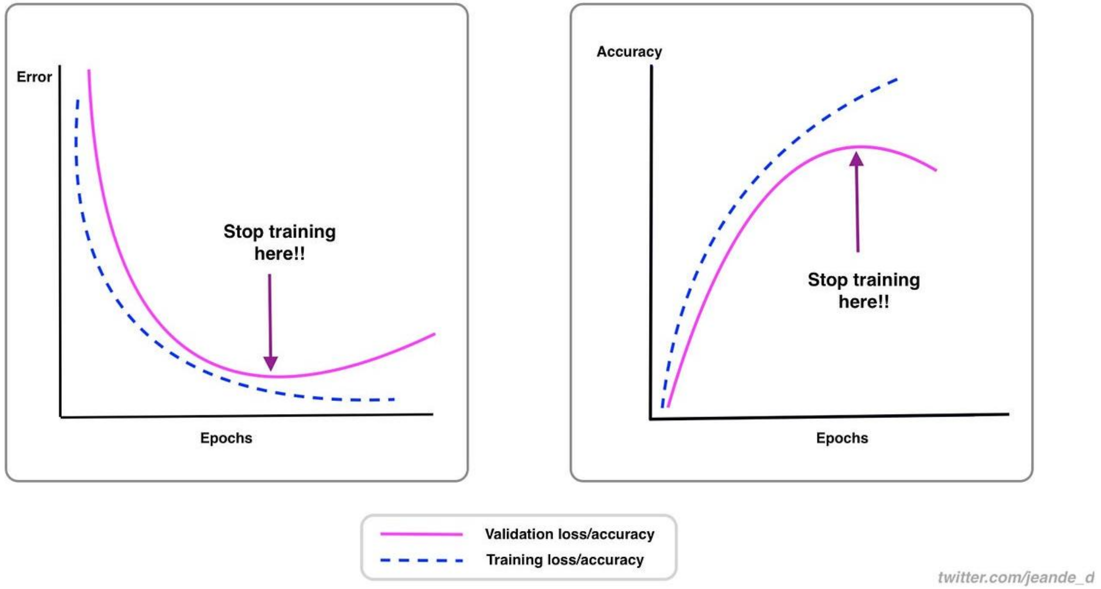


https://ubiai.tools/what-are-the-advantages-anddisadvantages-of-data-augmentation-2023-update/

UNIVERSITY OF CALGARY

#6 Machine Learning: Model Bias and Variance

- **My Challenge**: Implementing custom early stopping with unrealistic threshold



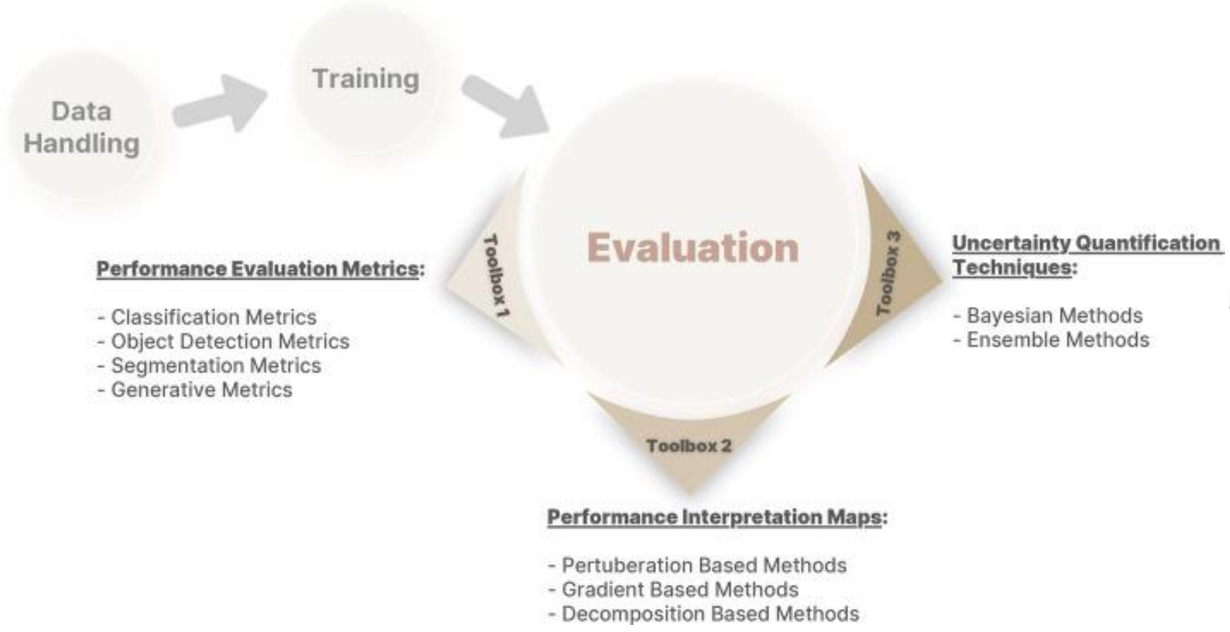
https://jeande.medium.com/early-stopping-explained-62eebce1127e

- **Recommendation:** Pay attention to intermediate errors/results
- **Coding Recommendation**: Scikit-learn

UNIVERSITY OF CALGARY

#7 Data Science: Metric Bias

- **Performance metrics** considerations
  - Balance explainability and generalizability through multi-metric use (e.g. confusion matrix)
  - What is the model truly learning?



Predicted Class

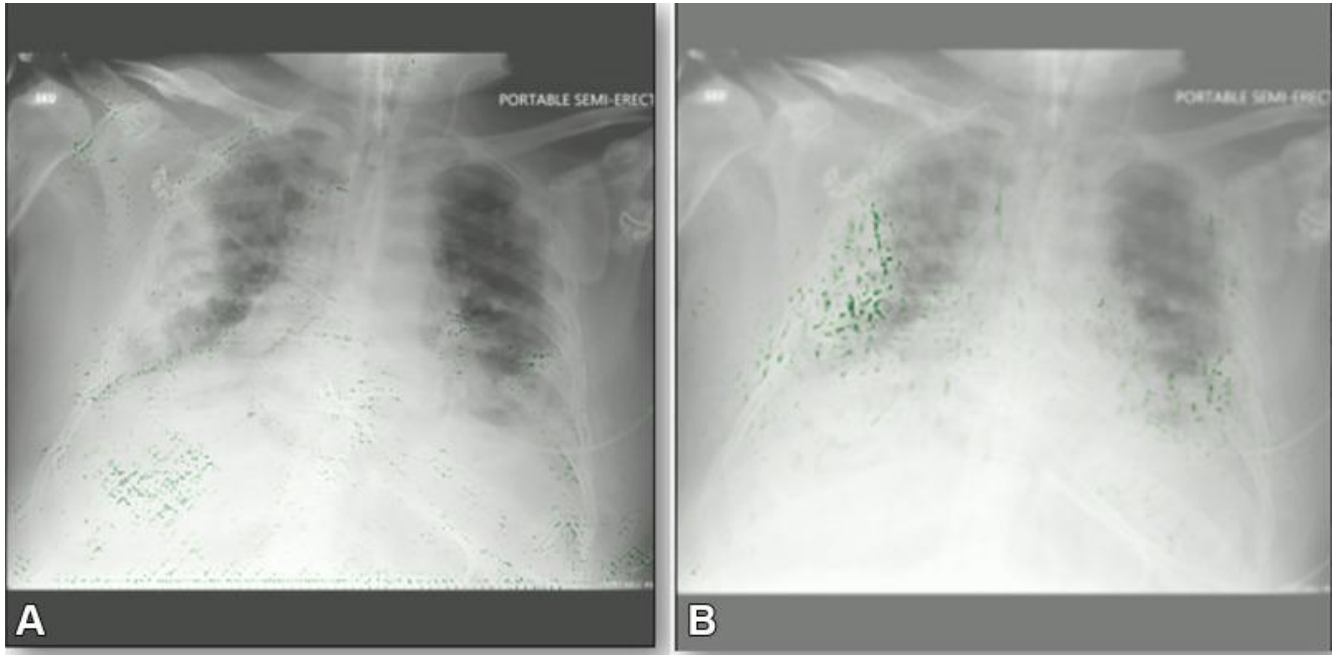
Actual Class	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN)	Sensitivity $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP)	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TP+FP)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

https://pmc-ncbi-nlm-nih-gov.ezproxy.lib.ucalgary.ca/articles/PMC9530766/

#7 Data Science: Metric Bias

- **Interpretive maps** – should be assessed based on utility, sensitivity to weight randomization, repeatability (intra-architecture), and reproducibility (inter-architecture)

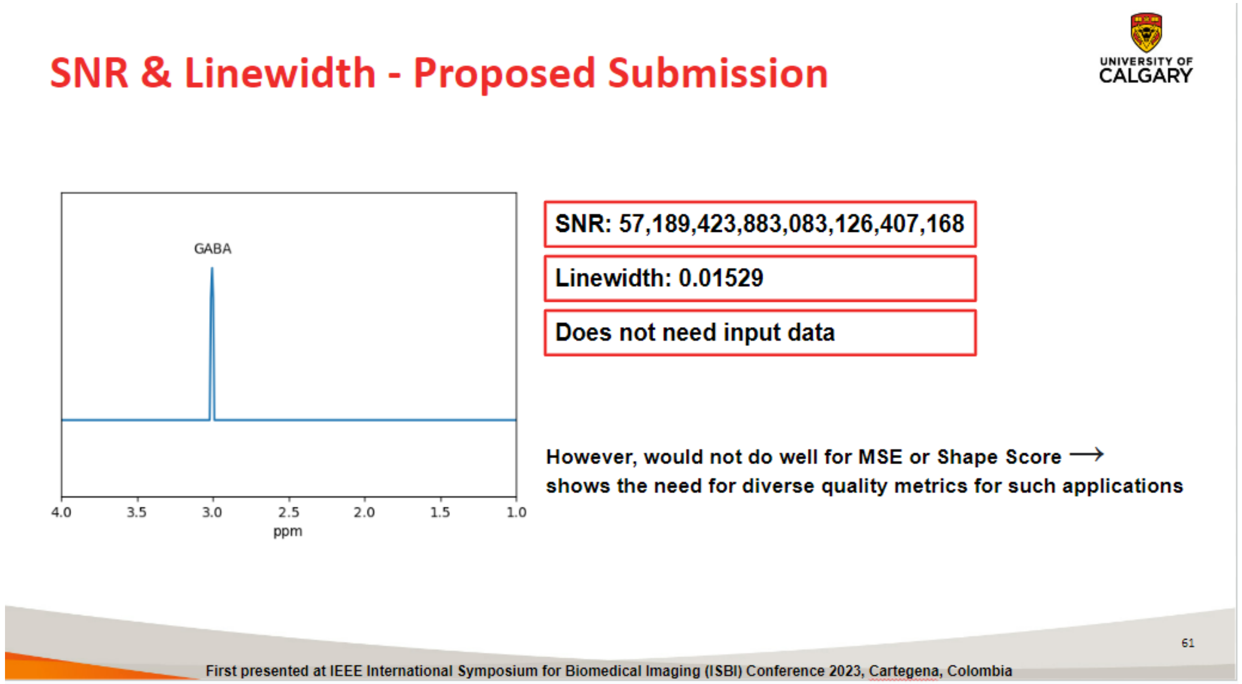
- **Uncertainty quantification** – calibrated confidence (predicted output vs. actual probability)



https://pmc-ncbi-nlm-nih-gov.ezproxy.lib.ucalgary.ca/articles/PMC9530766/

#7 Data Science: Metric Bias

- **My Challenge**: ML model misinterpreting domain specific optimization criteria

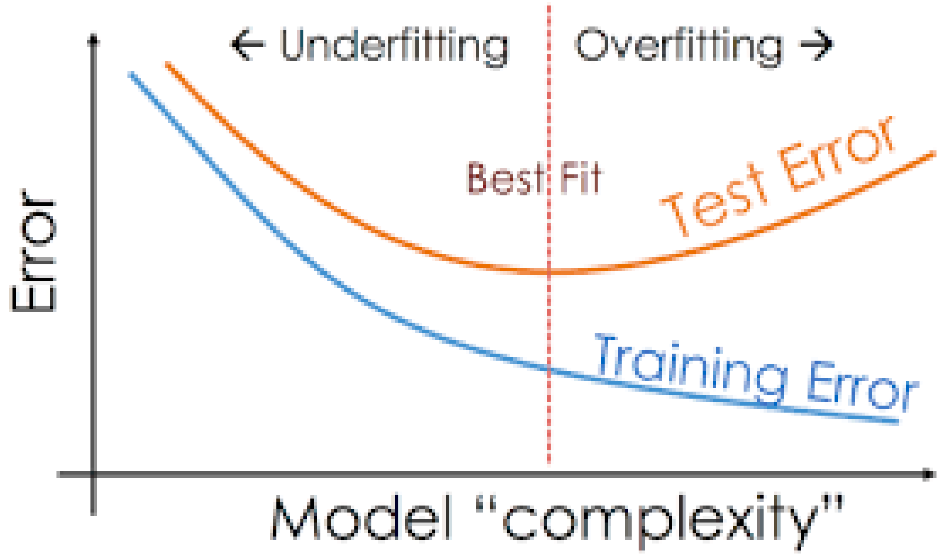


https://link.springer.com/article/10.1007/s10334-024-01156-9

- **Recommendation:** Evaluate the right balance of domain and ML specific metrics
- **Coding Recommendation**: GradCAM

#8 Machine Learning: Tracking Performance

- What can cause a model to overfit vs. underfit?
  - High bias (underfit) vs. High variance (overfit)
  - Model capacity (extent to learn data representation)
  - Dataset
    - Size
    - Feature variance
  - Regularization

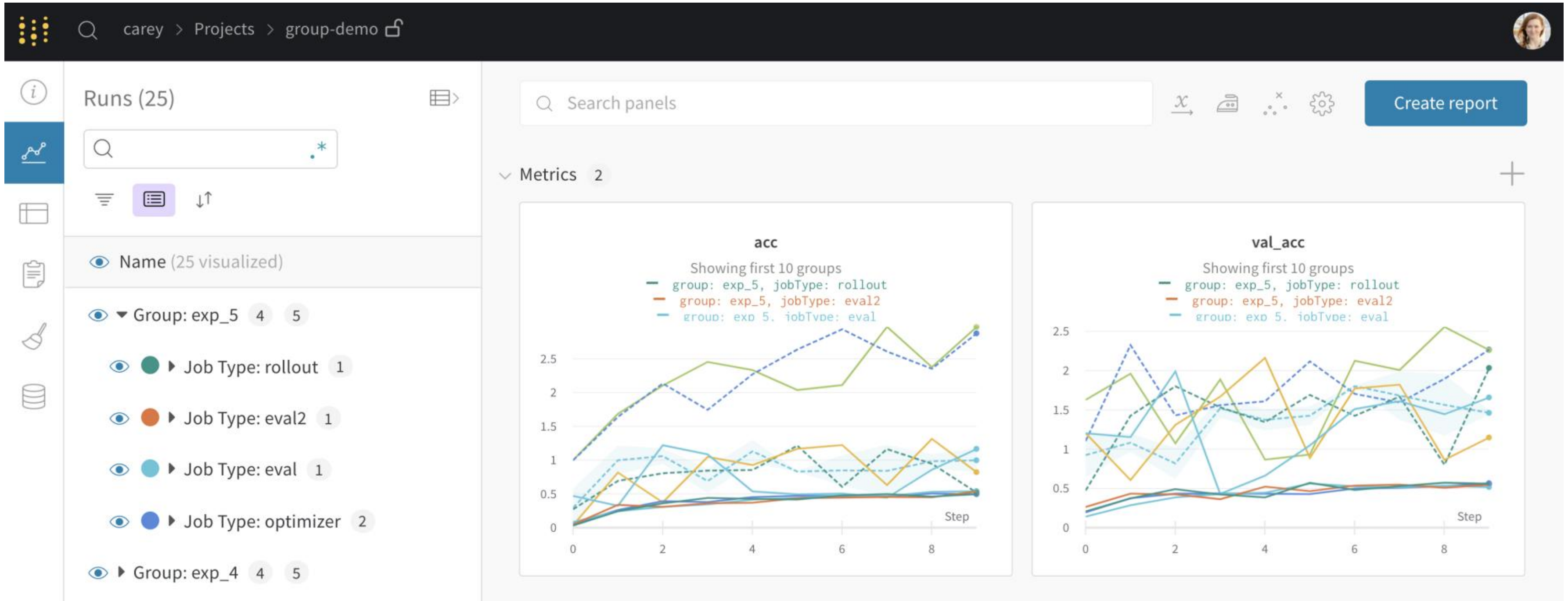


https://www.analyticsvidhya.com/blog/2020/02/underfitting-overfitting-best-fitting-machine-learning/

UNIVERSITY OF CALGARY

#8 Machine Learning: Tracking Performance

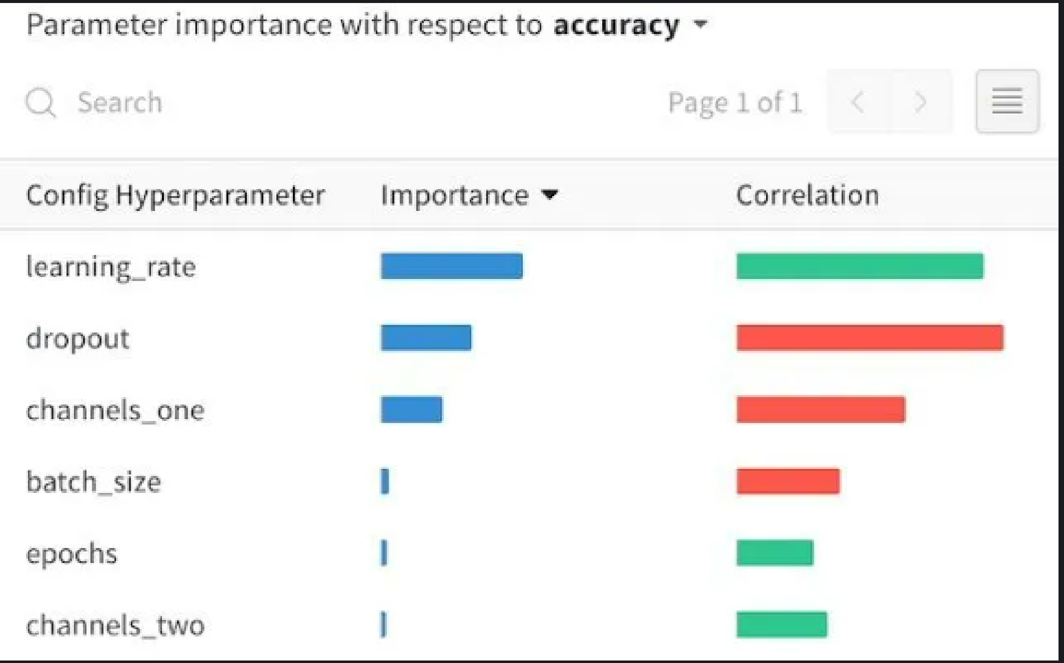
- Wandb
  - Platform to track and compare machine learning model iterations



https://docs.wandb.ai/guides/runs/grouping

UNIVERSITY OF CALGARY

#8 Machine Learning: Tracking Performance



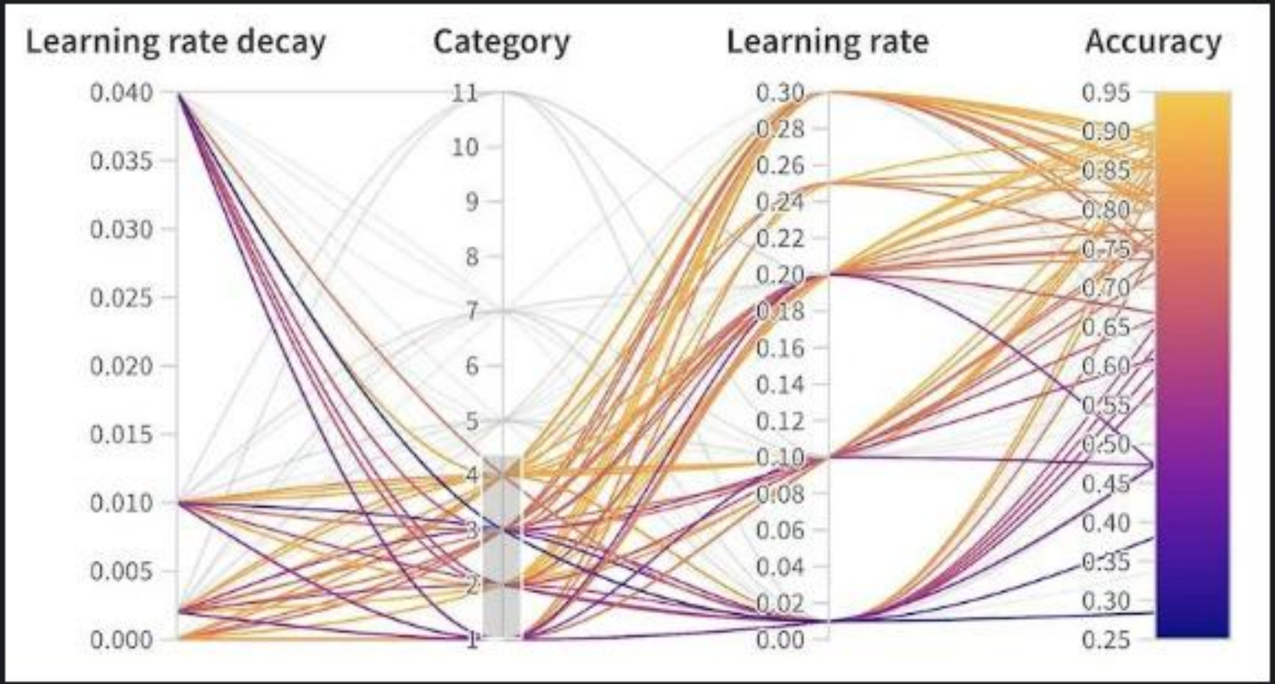
Parameter importance

Visualize which hyperparameters affect the metrics you care about. Weights & Biases comes with default visualizations that make it easy to get started without writing custom code to compare machine learning experiments.

https://wandb.ai/site/sweeps

UNIVERSITY OF CALGARY

#8 Machine Learning: Tracking Performance



Bayesian optimization

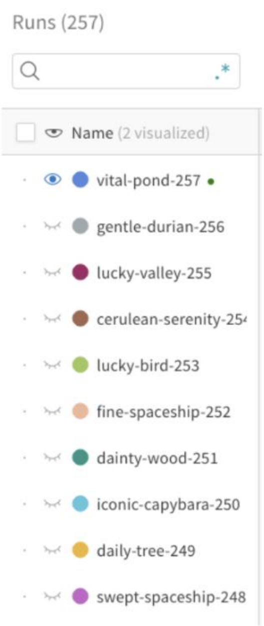
Use our transparent implementations of popular algorithms, or customize your own logic for sweeps.

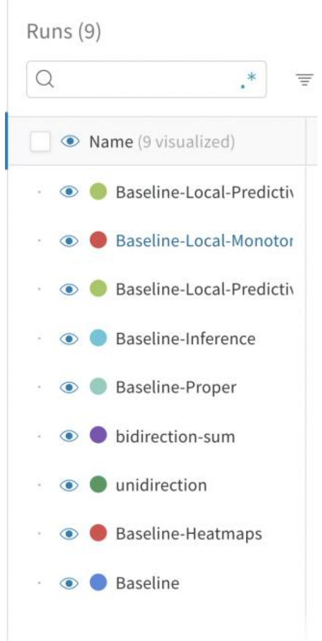
https://wandb.ai/site/sweeps

UNIVERSITY OF CALGARY

#8 Machine Learning: Tracking Performance

- **My Challenge**: Redundant hyperparameter combination testing





https://docs.wandb.ai/guides/app/features/runs-table

https://docs.wandb.ai/guides/integrations/add-wandb-to-any-library

- **Recommendation:** Document naming conventions
- **Coding Recommendation**: wandb

UNIVERSITY OF CALGARY

#9 Machine Learning: Allocating Resources

- What is a TPU or NPU?

#9 Machine Learning: Allocating Resources

- What is a TPU or NPU?

  - Tensor Processing Unit / Neural Processing Unit are ASICS aimed at accelerating AI applications

  - Ideal processing unit types have been identified for different DL tasks

https://arxiv.org/pdf/1907.10701

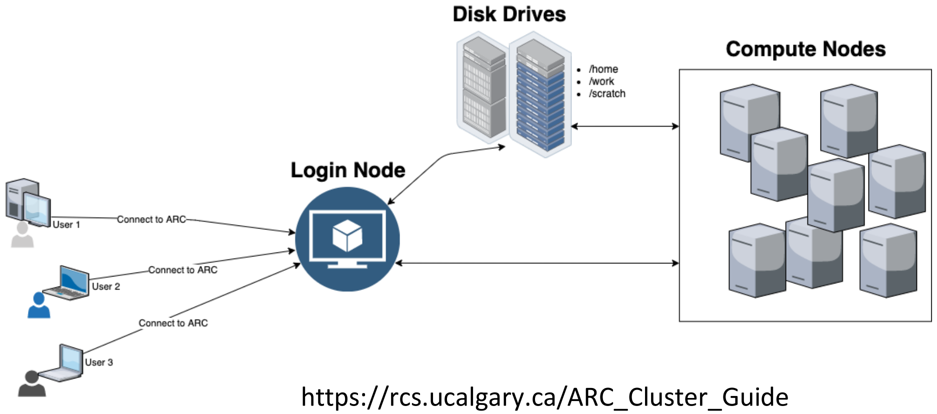
https://premioinc.com/blogs/blog/what-is-the-difference-between-cpu-vs-gpu-vs-tpu-complete-overview

UNIVERSITY OF CALGARY

#9 Machine Learning: Allocating Resources

- High Performance Computing (HPC) can provide more computational resources (storage, processing) than your local computer

UCalgary ARC cluster

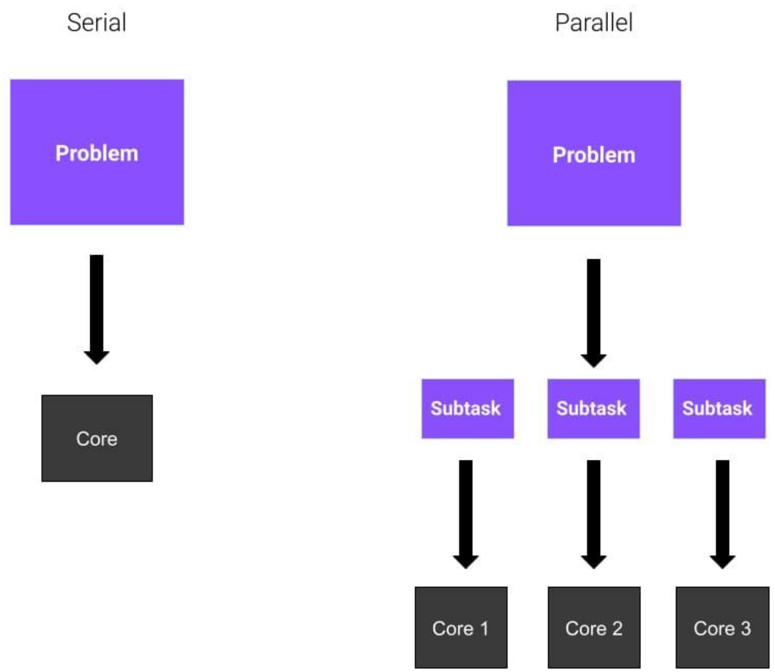


https://rcs.ucalgary.ca/ARC_Cluster_Guide

- Submitting a 'job' to the cluster with SLURM
- Benefits of running multiple iterations in parallel
- For more general information: https://rcs.ucalgary.ca/ARC_Cluster_Guide
- For more information on (past) summer school sessions: https://rcs.ucalgary.ca/RCS_Summer_School_2024

#9 Machine Learning: Allocating Resources

- **My Challenge**: Running models sequentially locally



- **Recommendation:** Learning to use UCalgary's HPC ARC cluster
- **Coding recommendation:** SLURM

UNIVERSITY OF CALGARY

https://www.sitepoint.com/python-multiprocessing-parallel-programming/

#10 Communicating your Findings

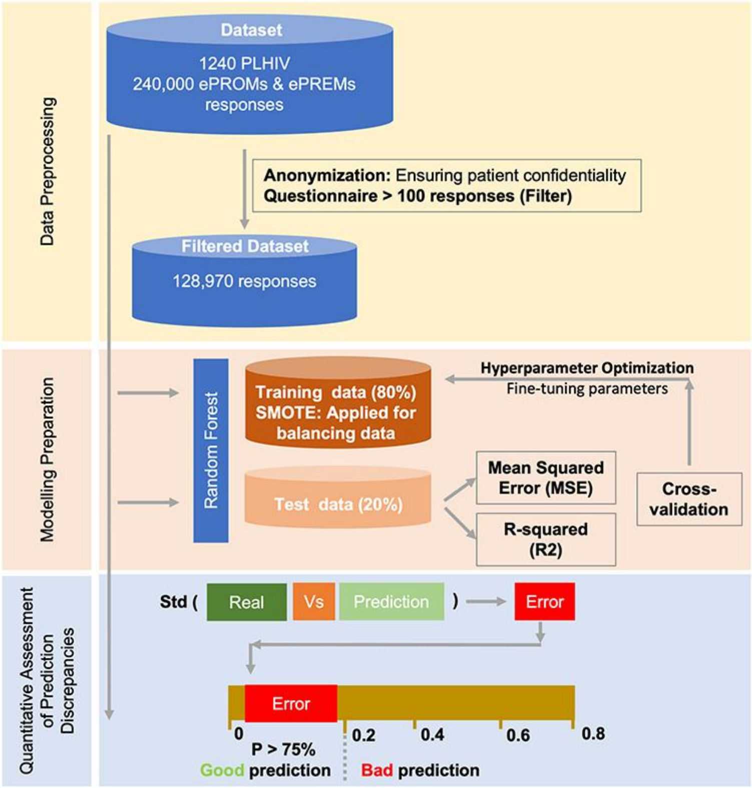
- When visualizing data, consider:
  - Vocabulary - consistent with domain standards expected in the field
  - Scalability - overestimation or underestimation of performance
  - Consistency – colors, naming, marker shape, etc.
  - Labels and caption – self-contained
  - Accuracy

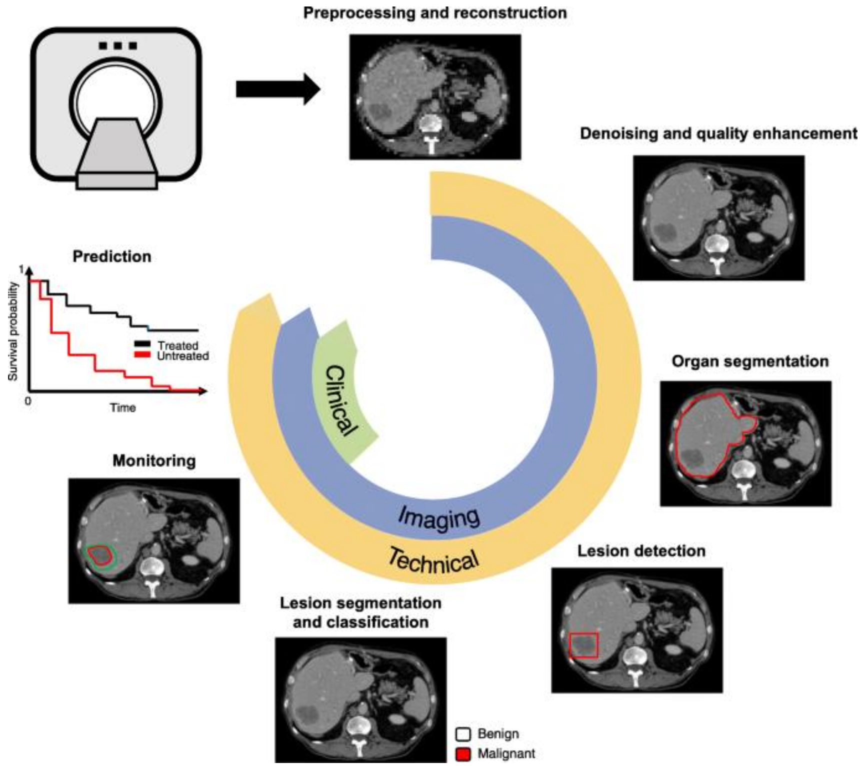
https://chartexpo.com/blog/data-visualization-guide

UNIVERSITY OF CALGARY

#10 Communicating your Findings

- **My Challenge**: Creating comprehensive figures based on target audience





- **Recommendation:** Getting feedback from those WITH and WITHOUT domain knowledge
- **Coding Recommendation:** Draw.io, VENNGAGE

UNIVERSITY OF CALGARY

https://www.dovepress.com/developing-a-prototype-machine-learning-model-to-predict-quality-of-li-peer-reviewed-fulltext-article-IPRP

https://insightsimaging.springeropen.com/articles/10.1186/s13244-019-0832-5

Presentation Takeaways

- VISUALIZE your data throughout the process
- Consider all the pros and cons of PARAMETERS and PERFORMANCE METRICS prior to training
- Write code in a way that is EFFICIENT and USER-FRIENDLY and ALLOCATES appropriate resources

UNIVERSITY OF CALGARY

Thank you!

UNIVERSITY OF CALGARY