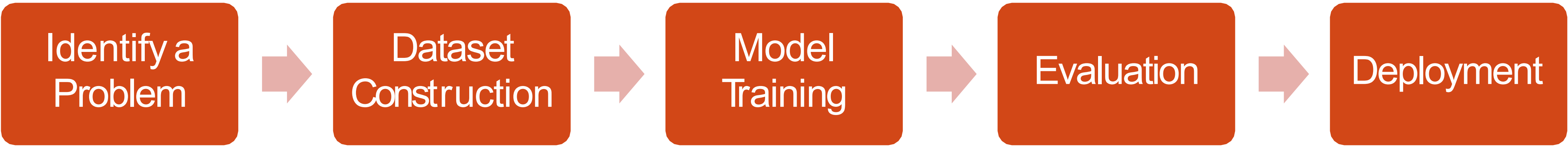
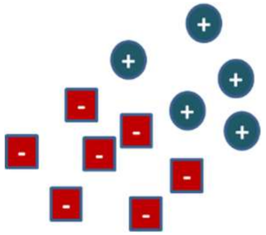
Twelve ways to fool the masses with machine learning

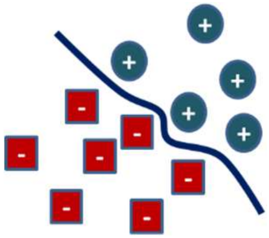
"IF YOU WANT TO TELL PEOPLE THE TRUTH, MAKE THEM LAUGH, OTHERWISE THEY'LL KILL YOU"

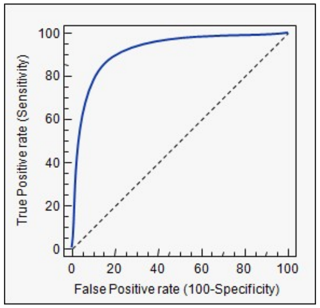
Machine Learning Lifecycle



Identify a Problem → Dataset Construction → Model Training → Evaluation → Deployment



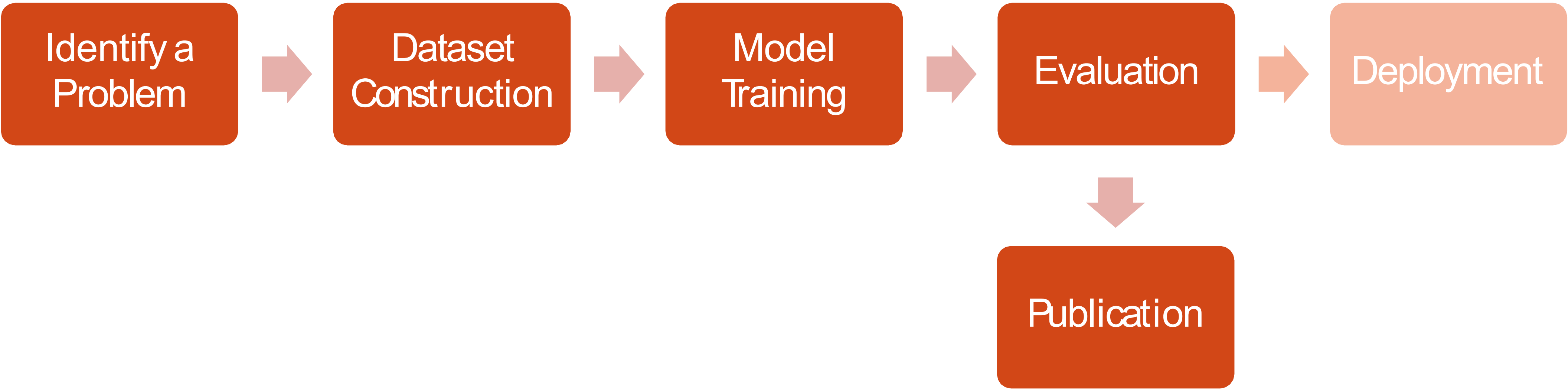




True Positive rate (Sensitivity)
100
80
60
40
20
0
0
20
40
60
80
100
False Positive rate (100-Specificity)

Webserver
Software Package

Machine Learning Lifecycle in Academia



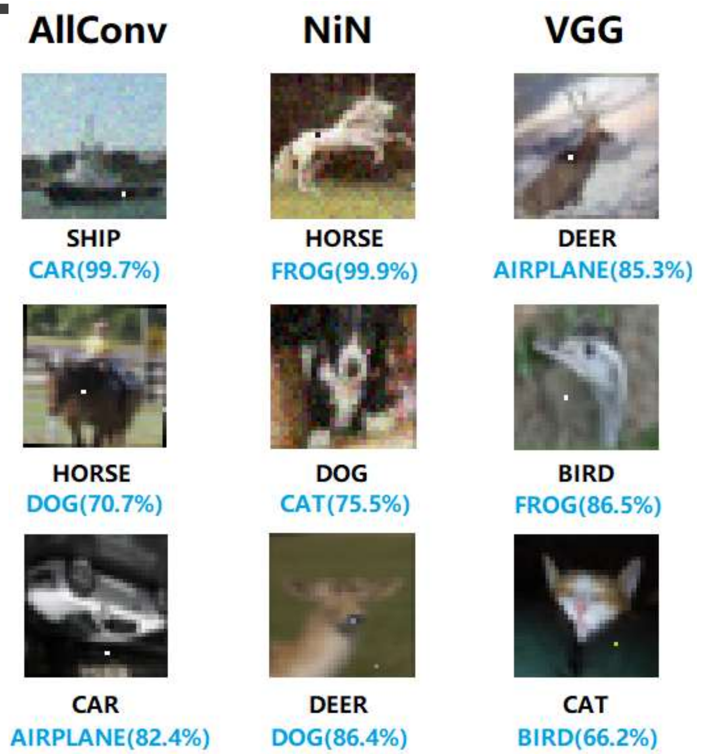
Impacts of Overselling a System

Bad name to the field

Stunted growth of the field

Psychological impact on researchers

One pixel attack for fooling deep neural networks



10 AI Failures in 2017

2017 in Review: 10 AI Failures

At Synced we are naturally fans of machine intelligence, but we also realize some new techniques struggle to perform their tasks effectively, often blundering in ways that humans would not

Artificial ignorance: The 10 biggest AI failures of 2017

From self-driving car accidents to Face ID hacks, artificial intelligence didn't have a flawless year.

By Olivia Krauth | January 4, 2018, 4:00 AM PST

https://syncedreview.com/2017/12/23/2017-in-review-10-ai-failures/
https://www.techrepublic.com/article/the-10-biggest-ai-failures-of-2017/

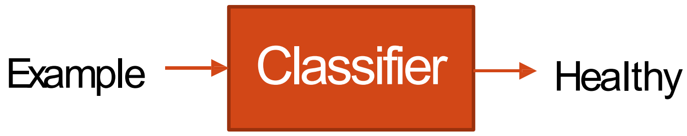
Su, J., Vargas, D. V., & Kouichi, S. (2017). One pixel attack for fooling deep neural networks. arXiv preprint arXiv:1710.08864.

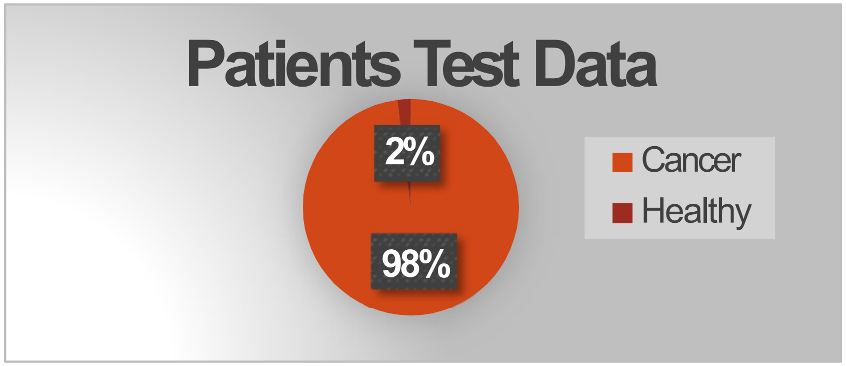
12 ways to oversell your method

1. Use a biased Accuracy Metric

Biased Performance Metric

- Example
  - Use Accuracy for a highly unbalanced dataset.


Example → Classifier → Healthy



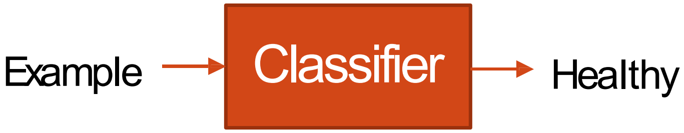
Accuracy= 98/100= 0.98

“Our method shows 98% **<u>accuracy</u>** on test data”

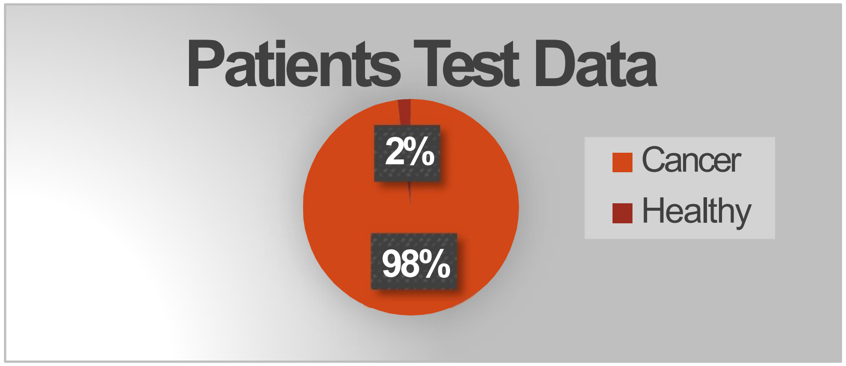
https://www.google.com.pk/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwixmLa1wYreAhXH-6QKHSiVB3sQjRx6BAgBEAU&url=https%3A%2F%2Ftheoldreader.com%2Fprofile%2FBinaryjesus%3Fpage%3D5&psig=AOvVaw0Os7PAOtdBGSvss0g68wNR&ust=1539763766998657

Biased Performance Metric

- Example
  - Use Accuracy for a highly unbalanced dataset.

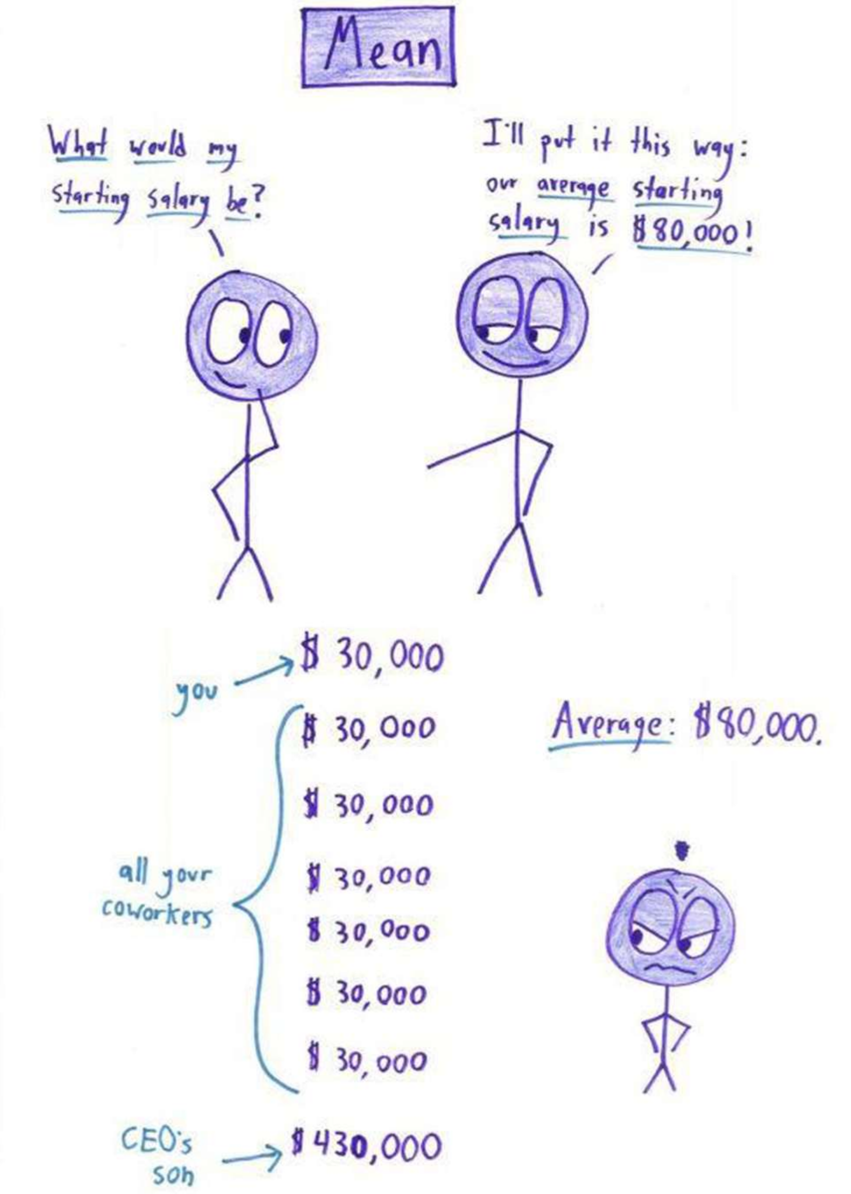


Example → Classifier → Healthy



Accuracy= 98/100= 0.98

"Our method shows 98% **accuracy** on test data"



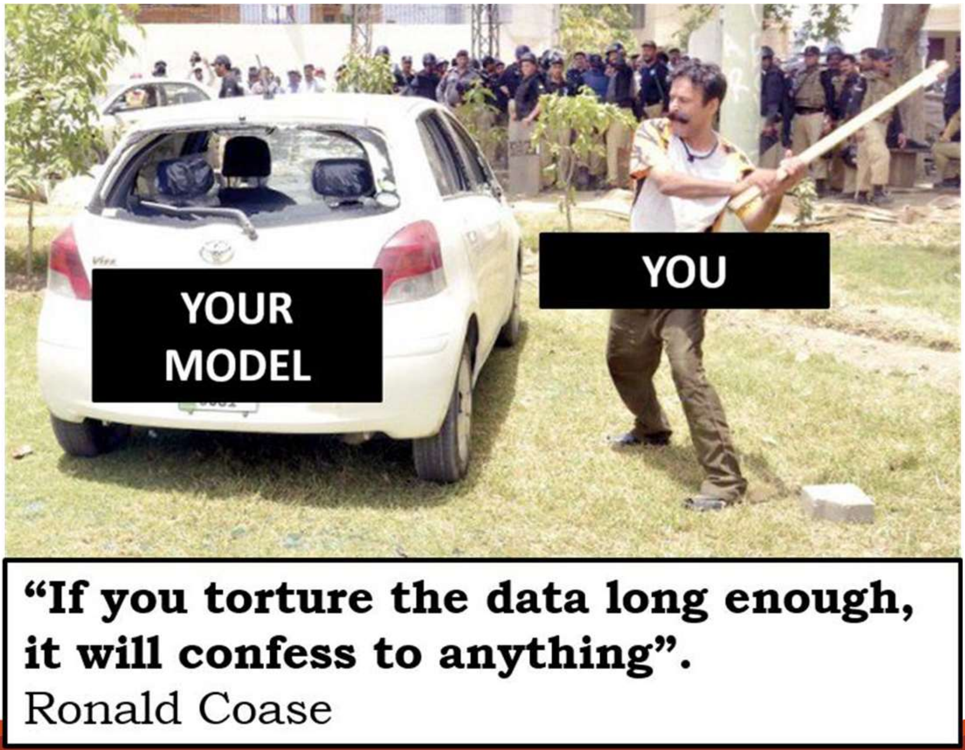
https://www.google.com.pk/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwixmLa1wYreAhXH-6QKHSiVB3sQjRx6BAgBEAU&url=https%3A%2F%2Ftheoldreader.com%2Fprofile%2FBinaryjesus%3Fpage%3D5&psig=AOvVaw0Os7PAOtdBGSvss0g68wNR&ust=1539763766998657

2. Maximize the performance metric without cross-validation

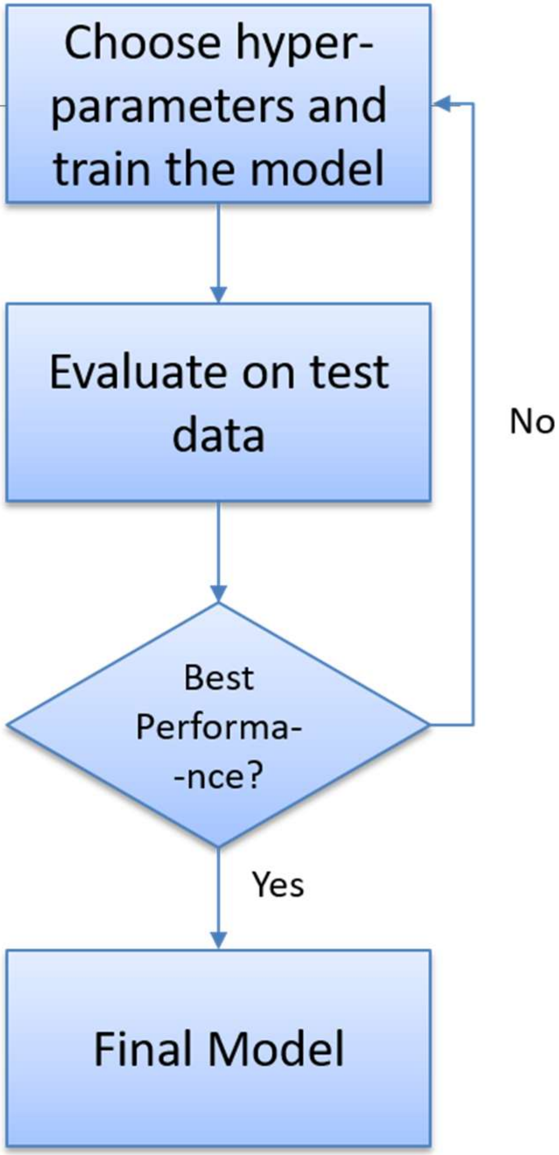
Maximize the performance Metric

Choose hyper-parameters that maximize the performance metric on test data

Forget that you are not allowed direct/indirect use of test data labels while training



"If you torture the data long enough, it will confess to anything".
Ronald Coase



Choose hyper-parameters and train the model

Evaluate on test data

Best Performa- -nce?

No

Yes

Final Model

3. (Indirectly) use labeled information in validation

Use Labels

Present cross-validation results and use labels (directly or indirectly) as features

```
Initialize my_model, results=[]
for fold in Folds:
    my_model.train(fold.train_data)
    p=my_model.evaluate_performance(fold.test_data)
    results.append(p)
final_score=average(results)
```

Report final_score as the average performance of your model

Fold #	Accuracy(%)
1	70
2	95
3	95
4	95
5	95
Average (The result to be reported)	**90**

4. Ignore the fact that examples may not be independent of each other

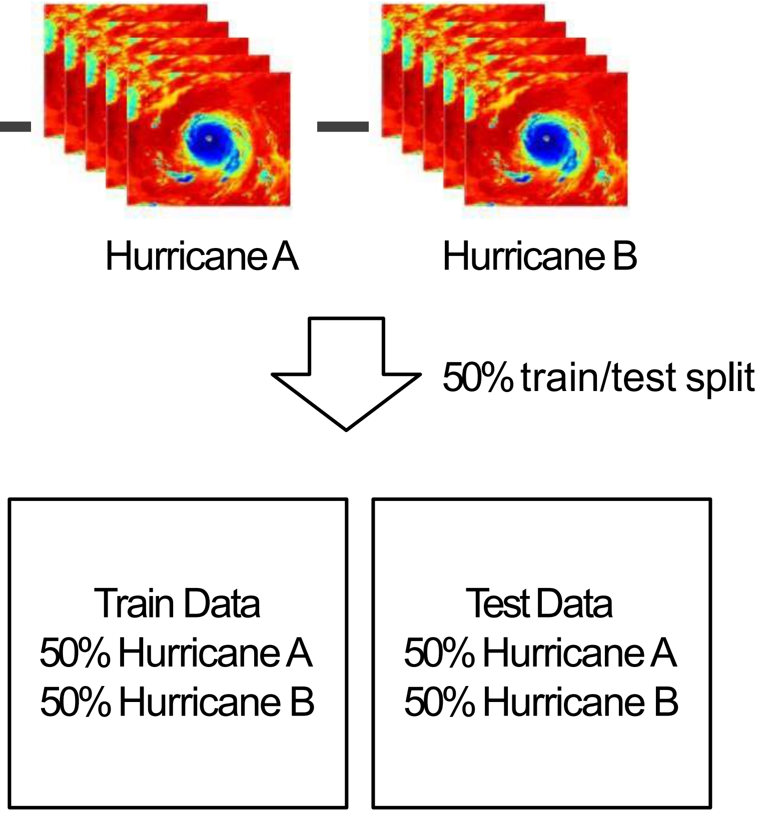
Train/ Test Overlap

There may be groups of closely related examples in the dataset

Random splitting may not ensure train/test disjoint-ness

- A closely related example to a test example may be a part of training

Over-estimation of generalization



Hurricane A

Hurricane B

50% train/test split

Train Data
50% Hurricane A
50% Hurricane B

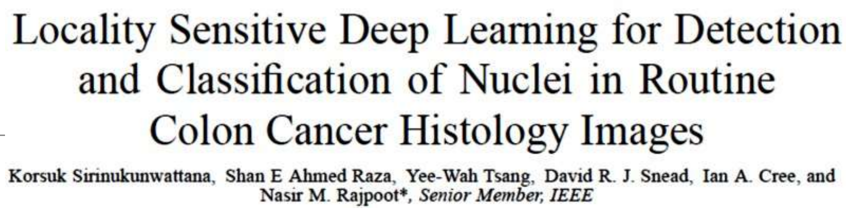
Test Data
50% Hurricane A
50% Hurricane B

5. Do not compare with a simple baseline classifier

Baseline Comparison

Start with the mostcomplex and “in” method

Do not check if the simpler(not-so-in) methods perform at par



Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images

Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David R. J. Snead, Ian A. Cree, and Nasir M. Rajpoot*, *Senior Member, IEEE*

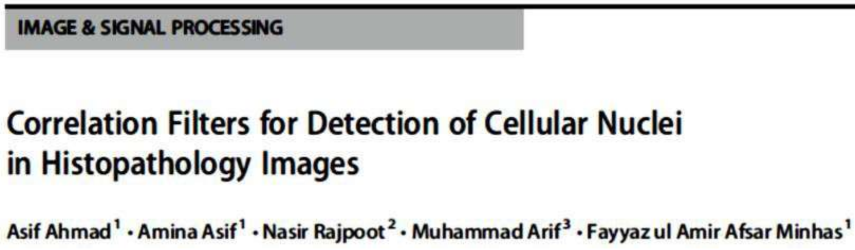


IMAGE & SIGNAL PROCESSING

Correlation Filters for Detection of Cellular Nuclei in Histopathology Images

Asif Ahmad[1] · Amina Asif[1] · Nasir Rajpoot[2] · Muhammad Arif[3] · Fayyaz ul Amir Afsar Minhas[1]

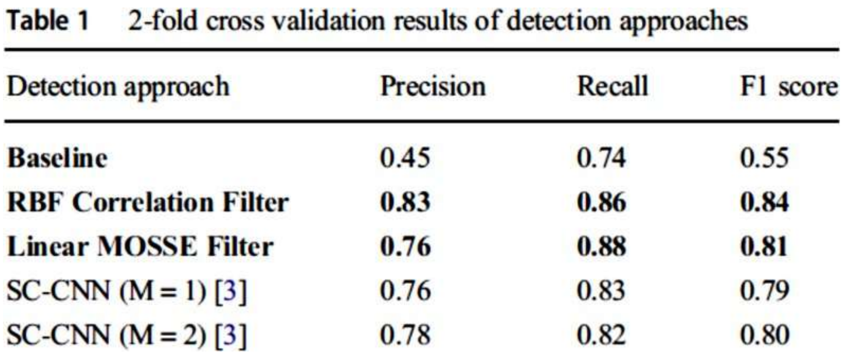
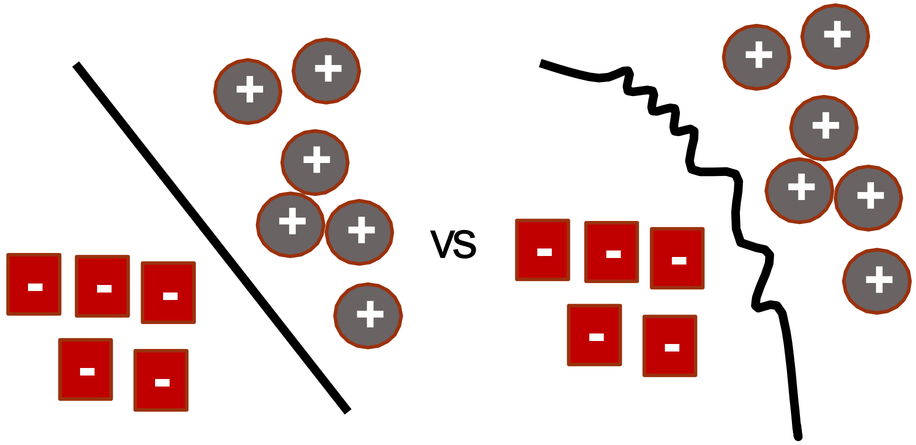


Table 1 2-fold cross validation results of detection approaches

Detection approach	Precision	Recall	F1 score
Baseline	0.45	0.74	0.55
RBF Correlation Filter	**0.83**	**0.86**	**0.84**
Linear MOSSE Filter	**0.76**	**0.88**	**0.81**
SC-CNN (M = 1) [3]	0.76	0.83	0.79
SC-CNN (M = 2) [3]	0.78	0.82	0.80



VS

6. Compare your model with un-optimized versions of other models or ones that have been trained using different data

Not Very Fair Comparison

Use best parameters for your model butforget to optimize other models

Different cross-validation protocols

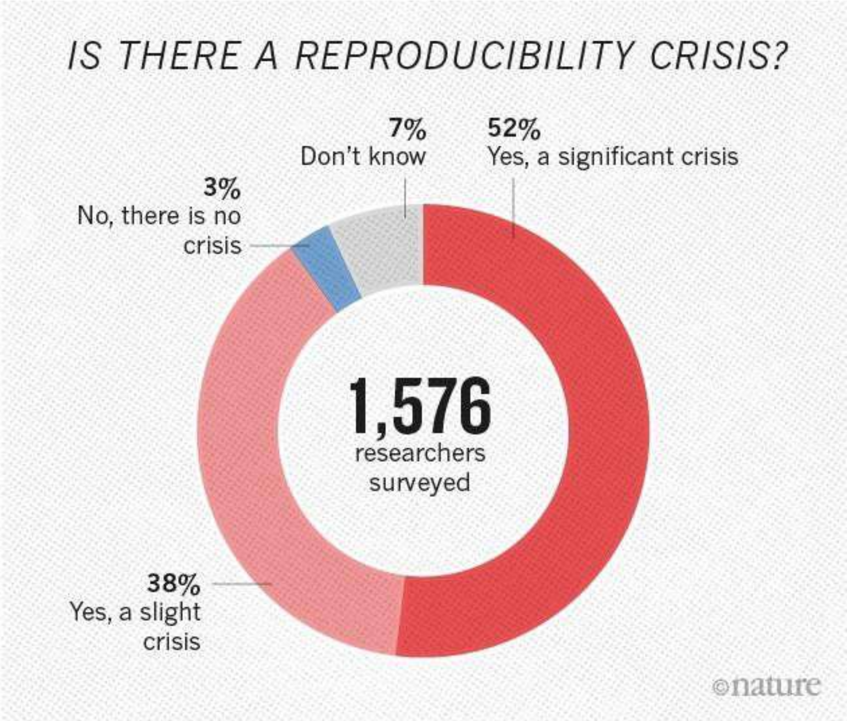
Performance results over different data

7. Present your paper in a way that doesn't allow reproducibility

No Reproducibility

Do not provide detailed performance results, codes or a webserver

Keep the model a “black-box”



Nature, 2016, M. Baker, 1,500 scientists lift the lid on reproducibility

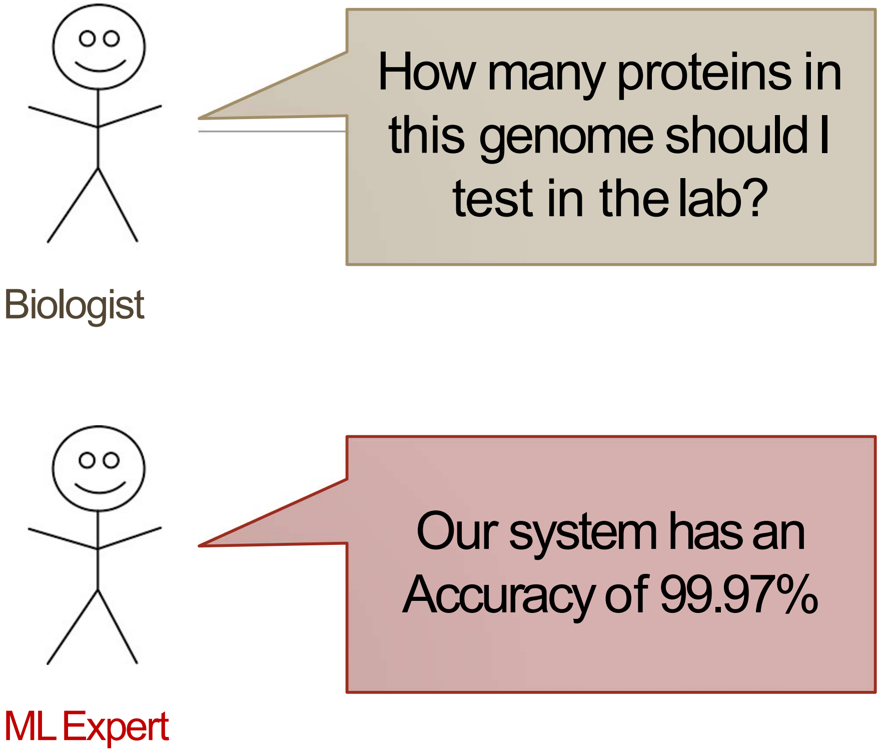
8. Choose a performance metric irrelevant to the problem domain

Irrelevant or uninterpretable metrics

Take problems from other domains

- Biology
- Chemistry
- Physics

Use metrics which the domain experts cannot interpret



Biologist

How many proteins in this genome should I test in the lab?

ML Expert

Our system has an Accuracy of 99.97%

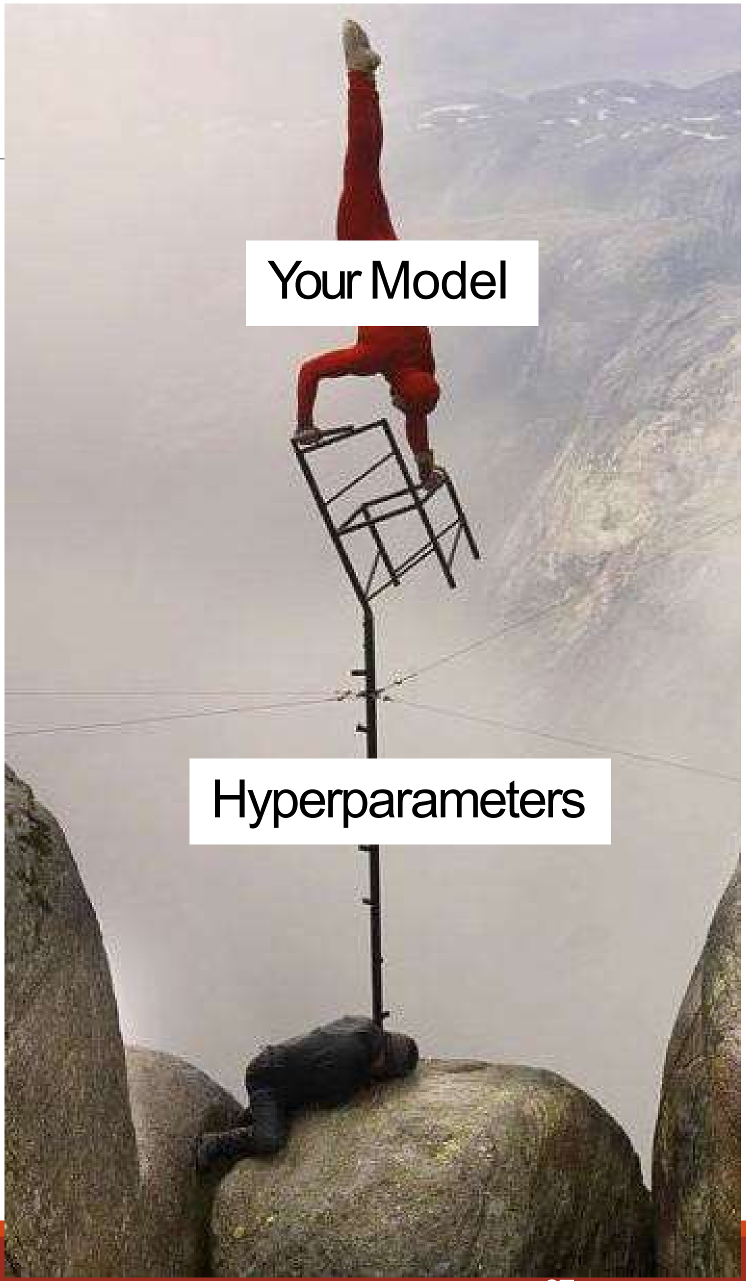
9. Do NOT analyze the sensitivity of your model to changes in data, hyper-parameter values or randomness

Sensitivity Analysis

Do not analyze the sensitivity of your system to experimental conditions

- Minor change in hyperparameters
- Randomness in folds
- Changes in data

Save the seed if the model is too sensitive to randomness



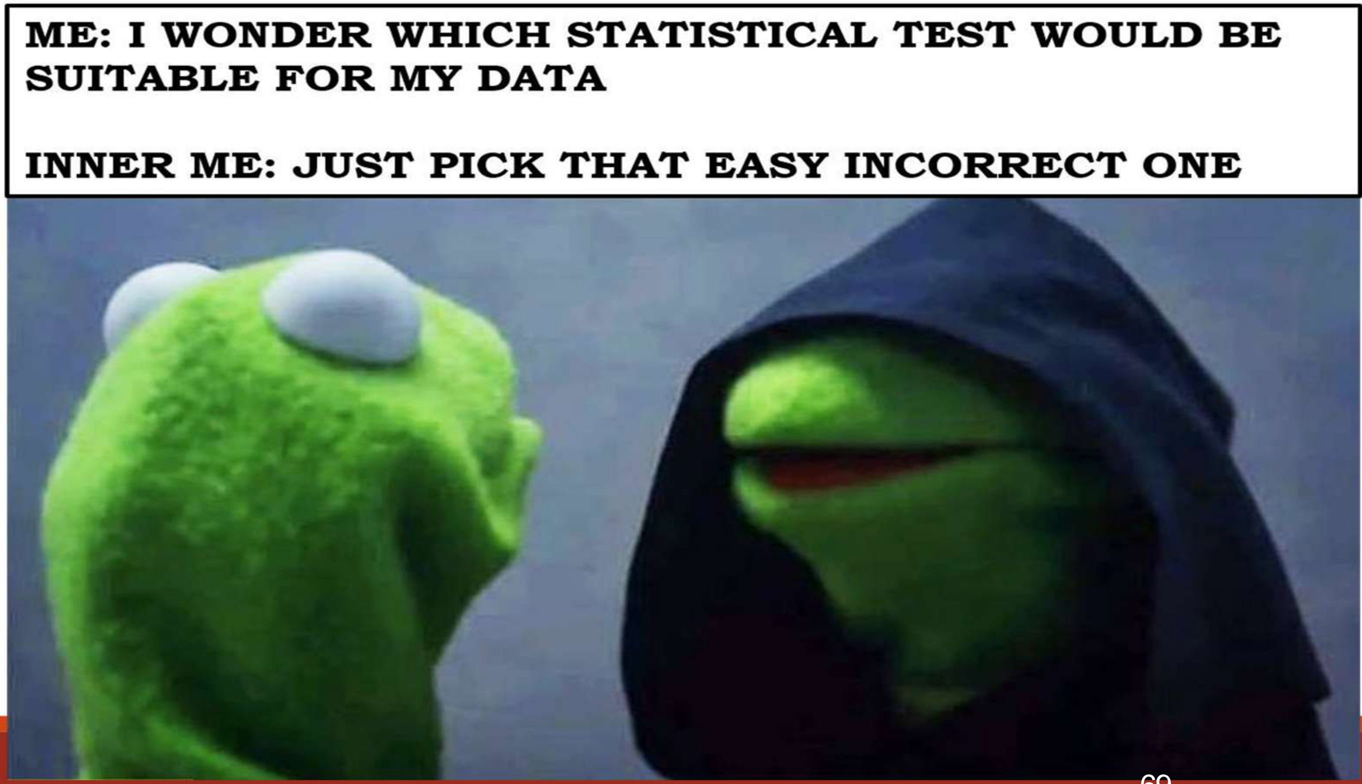
10. Use statistical tests even when their underlying assumptions are not met.

Underlying Assumptions

Most statistical tests are valid only when certain conditions are true

Use statistical tests even if they might not be applicable

"There are three kinds of lies: lies, damned lies, and statistics." (Mark Twain)


ME: I WONDER WHICH STATISTICAL TEST WOULD BE SUITABLE FOR MY DATA
INNER ME: JUST PICK THAT EASY INCORRECT ONE

11. Use buzzwords and pretty plots to whip your readers into submission

Intimidate the Readers

"If you can't convince them, confuse them".

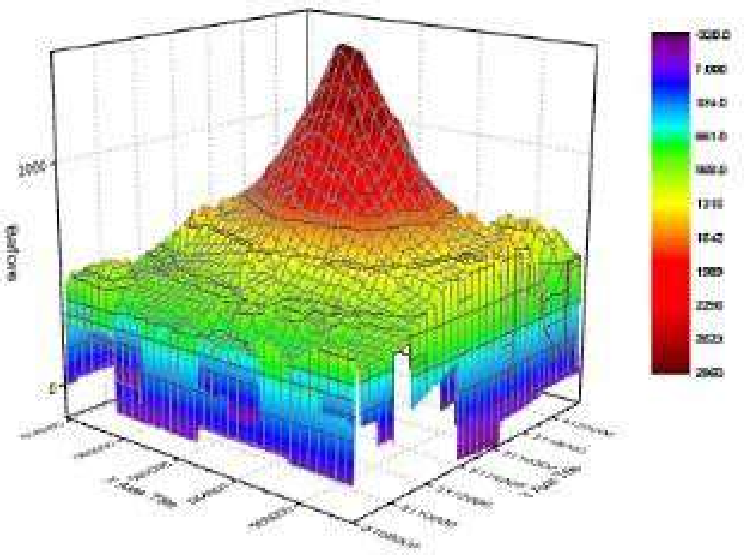
- (Harry S. Truman)

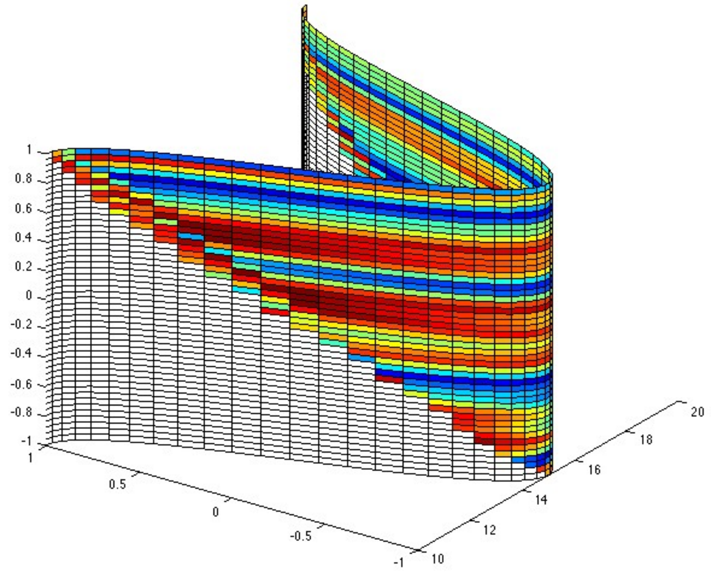
Use buzzwords

Ambiguous or hard to understand terminologie

- Multimodal Hyperspectral Convo-residual Superp Blockchained Deep Learning

Lots of colorful plots





12. Care only about publishing and let go of the concept of generalization and practical use

Impact Factor is all that matters

Focus on publishable ε-improvement

Stay away from new scary problems

Do not consult domain experts

Just publish!!

13. Thirteen is the new twelve

Explainable Model

No need to worry if your model makes sense or not

Interpretablility

Conclusion

When reviewing or supervising research studies, look out forthese tactics

“It’s easier to fool people than to convince them that they have been fooled”

-Mark Twain

Required Reading

https://arxiv.org/ftp/arxiv/papers/1901/1901.01686.pdf