Week 5

Machine Learning and Big Data - DATA622

CUNY School of Professional Studies

No free lunch Theorem

Bias-free learning is futile

CUNY School of Professional Studies

TANSTAAFL

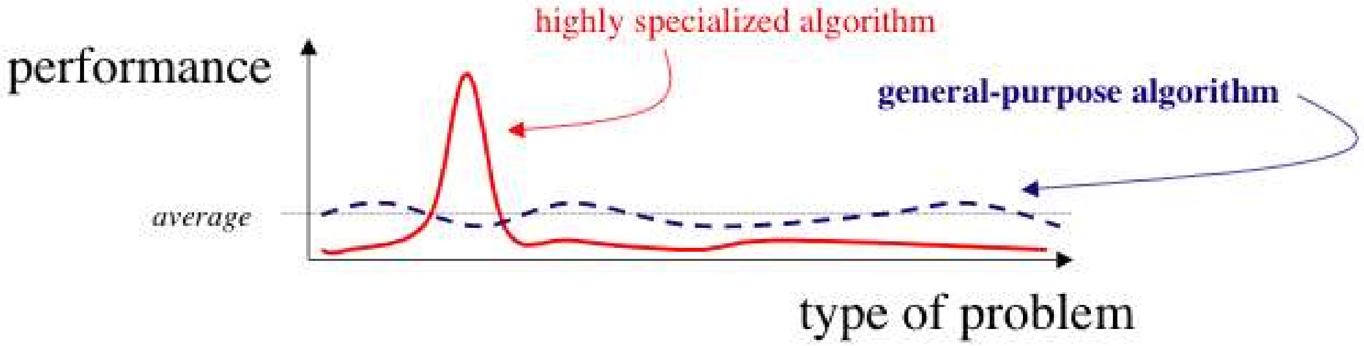
- No-Free-Lunch Theorem states:
  - No single classifier works the best for all possible problems
  - We need to make assumptions to generalize (we need bias)

TANSTAAFL

- Theorem:

  The average performance of any pair of algorithms across all possible problems is identical.

- If an algorithm achieves superior results on some problems, then it must pay with inferiority on the other problems



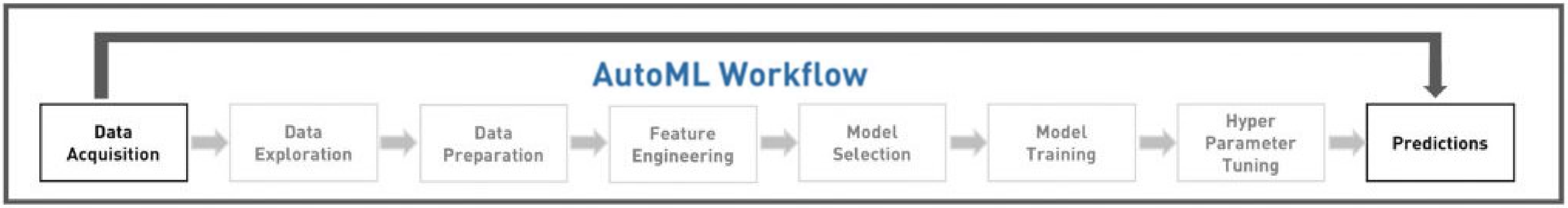
Source: https://medium.com/@LeonFedden/the-no-free-lunch-theorem-62ae2c3ed10c

Automated ML (AutoML)

Does AutoML break TANSAAFL ?

CUNY School of Professional Studies

AutoML



Let’s consider the no-free lunch theorem again.

If we automate the ML Lifecycle with AutoML, does that mean no-free lunch doesn’t apply?

No, no-free lunch theorem still applies because

- AutoML is just optimizing the process – it is not doing creative work
- AutoML will give you faster results – but not necessarily optimal

Bias

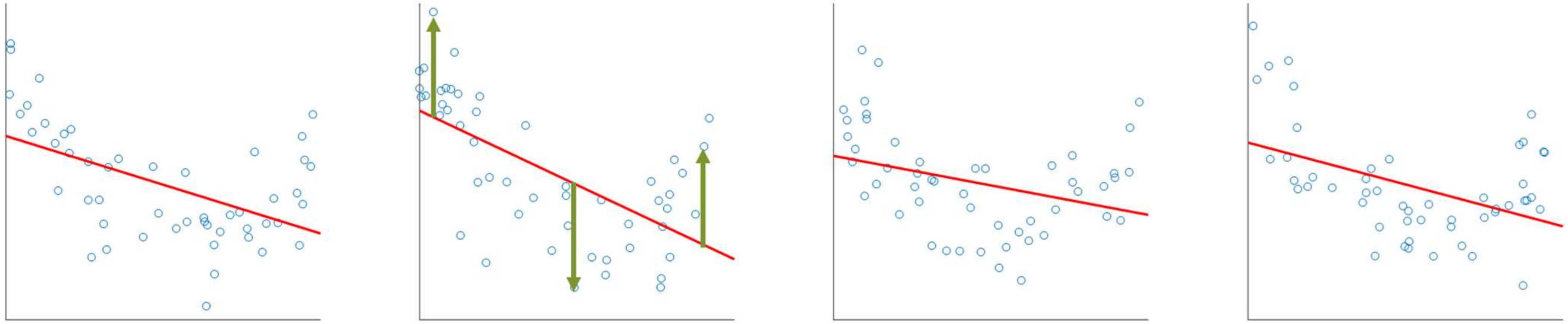
Complexity of models

Bias

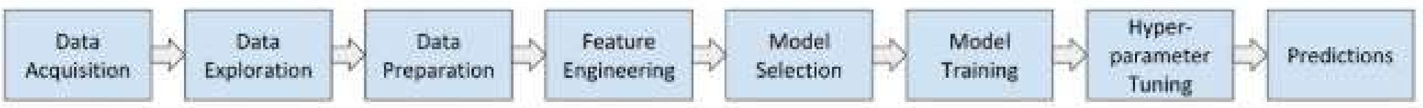
- Dictionary definition:
  - Bias: Prejudice in favor of or against one thing
  - Prejudice: Preconceived opinion
- Statistics definition:
  - Bias: The tendency of a statistic to overestimate or underestimate the population parameter you're trying to measure
  - Bias: The approach that leads to a systematic difference between the true parameters of a population and the statistics used to estimate those parameters
- Machine Learning definition:
  - Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
  - Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training data.

Bias

- Regardless of training sample, or size of training sample, model will produce consistent errors



How is bias introduced?



At each step of the ML Lifecycle you must make decisions which will impact the

- Data Acquisition
  Data is collected from various sources like files, databases, etc. and merged into one medium. What datasets you select will impact your model.
- Data Exploration
  Understanding and exploring data in order to identify the .
- Data Preparation
  To use data for training, some data processing needs to be done. This includes data cleaning for duplicates, processing missing data, removing noisy data, etc.
- Feature Engineering
  The goal of feature engineering is to convert categorical and ordinal data into numerical features. Includes feature creation, and overlaps with data preparation.
- Model Selection
  Selecting the correct model. Research needs to be done to finalize which model will work best for the dataset. At this step, the model is trained, interpreted, and evaluated for best performance. Key impact to bias.
- Model Training
  At this step, the model is trained, interpreted, and evaluated for best performance. Key impact to bias.
- Hyperparameter Tuning
  (Optional) Hyperparameter tuning to improve the performance of models by fine-tuning the parameters electing the correct model.
- Prediction (Serving)
  Make predictions of the unseen data. How you serve the model, monitor the model, and the decay of the model will impact its accuracy.

Variance

Generalizing the model

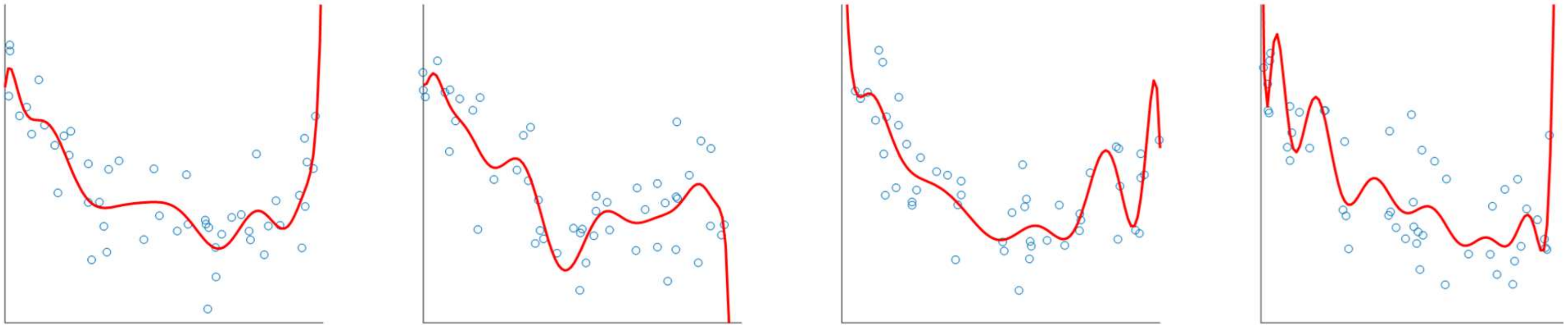
CUNY | School of Professional Studies

Variance

- Inability to perfectly estimate parameters from limited data
- Dictionary definition:
  - Bias: Prejudice in favor of or against one thing
  - Prejudice: Preconceived opinion
- Statistics definition:
  - Bias: The tendency of a statistic to overestimate or underestimate the population parameter you’re trying to measure
  - Bias: The approach that leads to a systematic difference between the true parameters of a population and the statistics used to estimate those parameters
- Societal bias is different to machine learning/statistical bias
- Bias is not intrinsically bad - if it is suitable for the problem domain
- Societal bias must be minimized. Machine Learning cannot exist without bias.

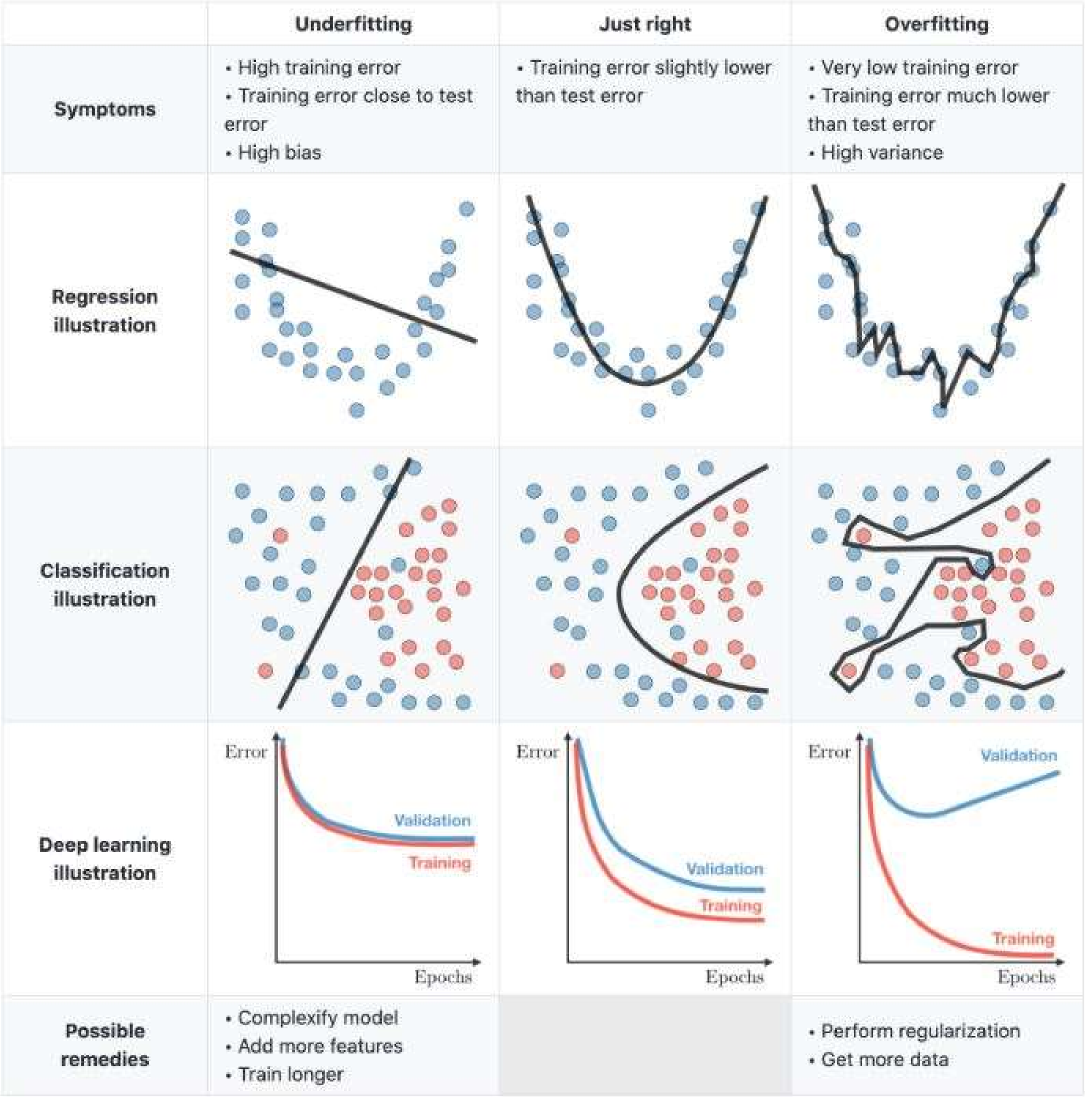
Variance

- Different samples of training data yield different model fits/results



Over & under-fitting

- Reducing Underfitting
  1. Increase model complexity
  2. Increase number of features
  3. Remove noise from the data
  4. Increase the number of epochs / increase the duration of training
- Reducing overfitting
  1. Increase training data.
  2. Reduce model complexity
  3. Early stopping during the training phase
  4. Adding noise to the data
  5. Regularization
  6. Use dropout for neural networks to tackle overfitting.



	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data

Bias-Variance Trade-off

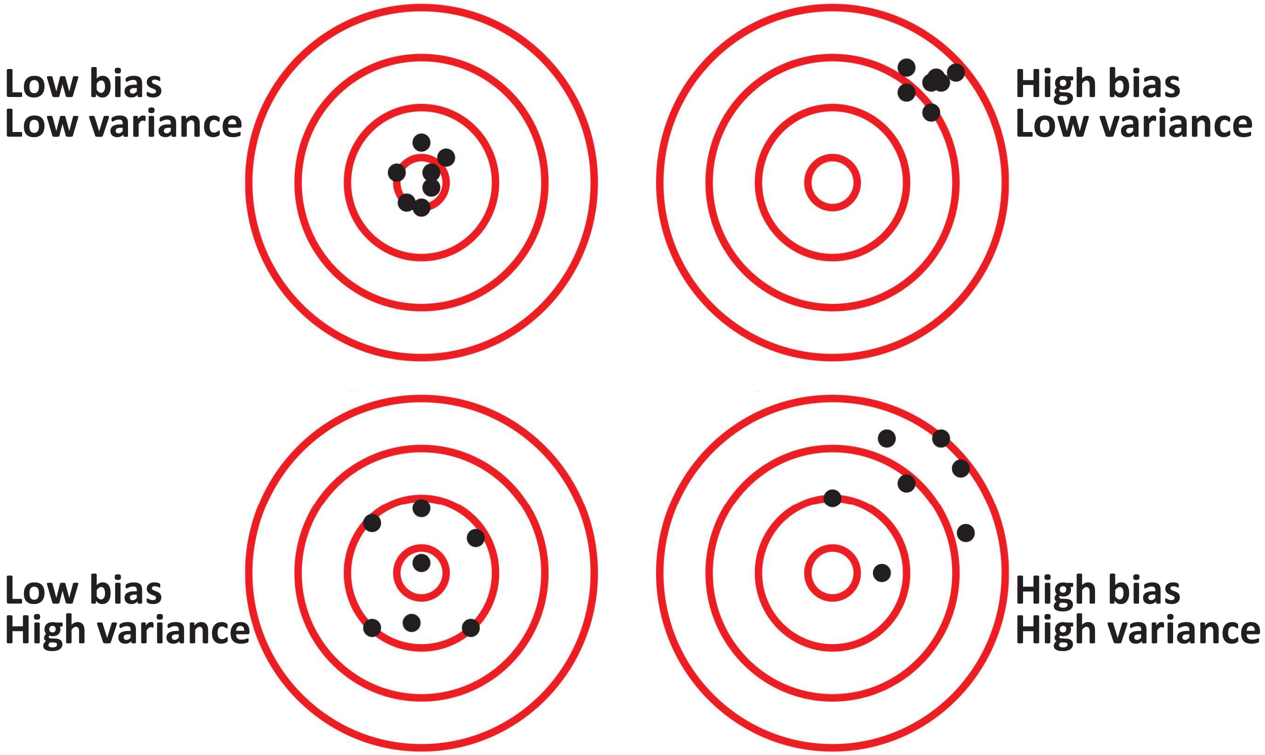
Bias-free learning is futile

Bias-Variance Trade-off

- Models of higher complexity have lower bias but higher variance.
- Once model complexity passes a certain threshold, models “overfit” with the variance term dominating the test error
- From this point onward, increasing model complexity will only decrease performance (i.e., increase test error)
- Once we pass a certain threshold, “larger models are worse.”
- Bias-variance trade off Is revealed via test set not training set

Bias-variance tradeoff using targets

- The goal would be for the points to be near the center of the target.
- To the extent which the points are far from the center, they suffer from bias.
- The variance is the dispersion. The bias and variance combine to form mean squared error



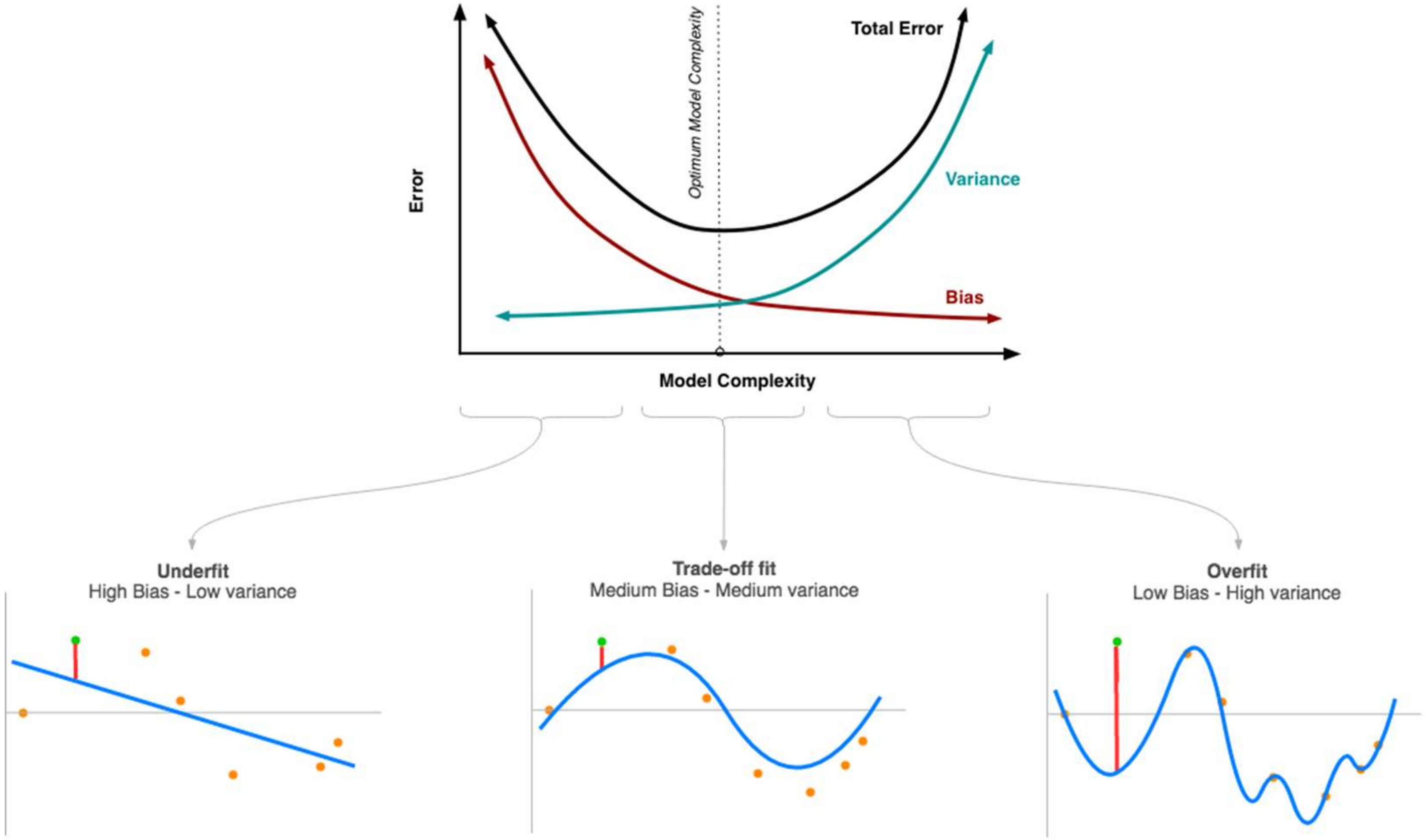
Low bias
Low variance

High bias
Low variance

Low bias
High variance

High bias
High variance

Bias-Variance Trade-off



Source http://www.ebc.cat/author/eduard-bonadagmail-com/)

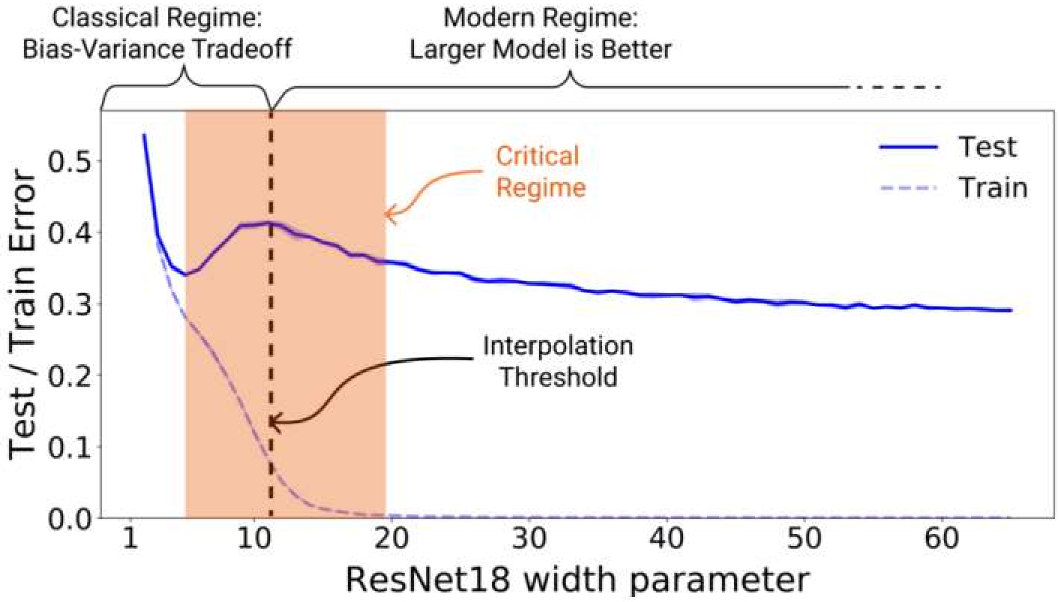
Double-descent

Deep learning: where the classical bias-variance trade-off breaks down

CUNY School of Professional Studies

Double-descent

- Observed in neural networks, where the bias-variance tradeoff breaks down
- Test error keeps decreasing as we over-parametrize the network or add more training examples



Double-descent

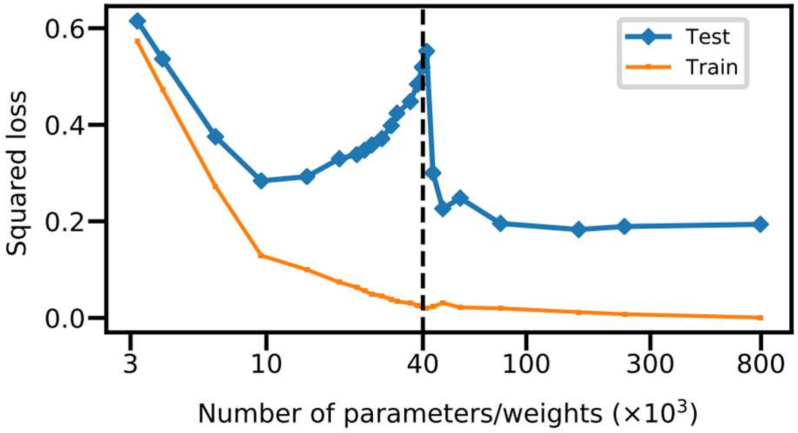


Figure 4: **Double descent risk curve for fully connected neural network on MNIST.** Training and test risks of network with a single layer of H hidden units, learned on a subset of MNIST ($n = 4 \cdot 10^3$, $d = 784$, $K = 10$ classes). The number of parameters is $(d+1) \cdot H + (H+1) \cdot K$. The interpolation threshold (black dotted line) is observed at $n \cdot K$.

Reconciling modern machine learning practice and the bias-variance trade-off, Mikhail Belkin, et al https://arxiv.org/pdf/1812.11118.pdf

Double-descent

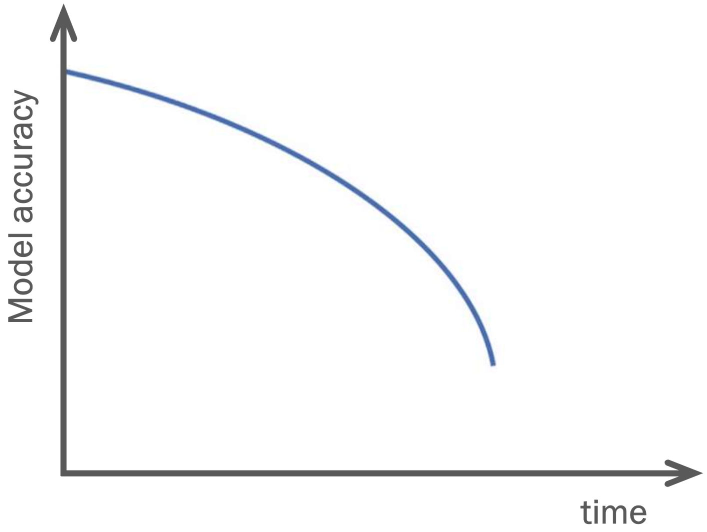
- Double descent provides an indication that even though models that pass through every training data point are indeed overfitted
- The structure of the resulting network forces the interpolation to be smooth and results in superior generalization to unseen data
- As long as we keep increasing the model complexity, test error keep decreasing
- After certain complexity, the testing error start to be smaller than the sweet spot that we get within the under-parameterization regime

Data Drift

Just as data has “shape” – that shape will change (reflecting the underlying environment)

Model Decay

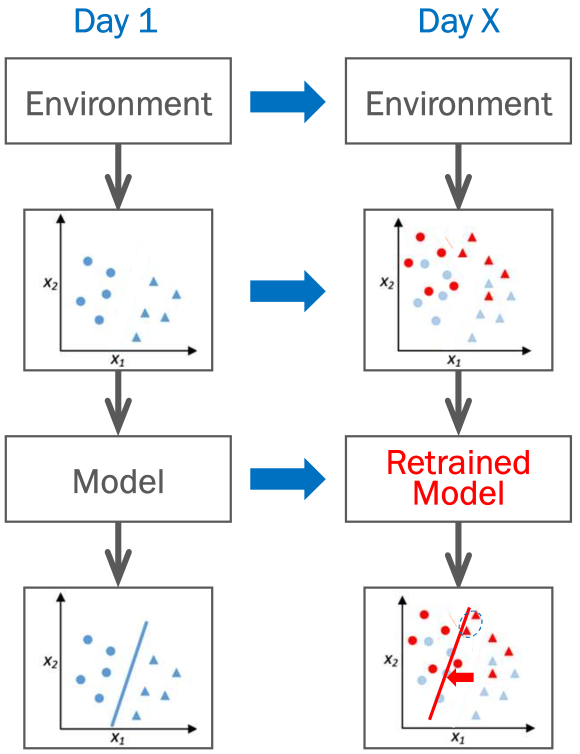
- When a model is put into production model accuracy starts to degrade
- You will never achieve the accuracy in production that you did in training
- The accuracy and quality of models will continue to decay over time
- This is completely different to traditional software – because machine learning is data driven



Source: TLL
Unsupervised/clustering models are more complex.

Non-stationarity

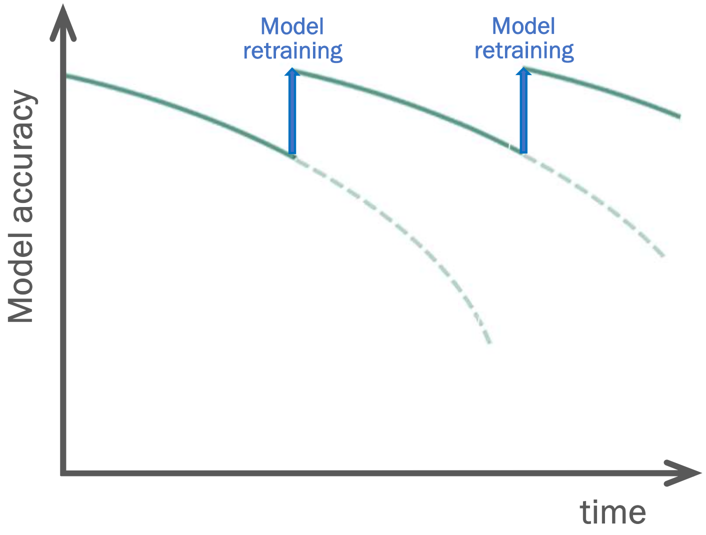
- Concept drift refers to changes in the environment mechanism (“concept”) that the model is trying to predict
- Concepts – relationships between input & output data - evolve and are not stationary
- Regime change refers to a major shift in concept e.g. due to Regulations
- Data drift refers to the changes in the input data (feature dataset) – features are not stationary



Source: TLL
Unsupervised/clustering models are more complex.

Models Require Maintenance

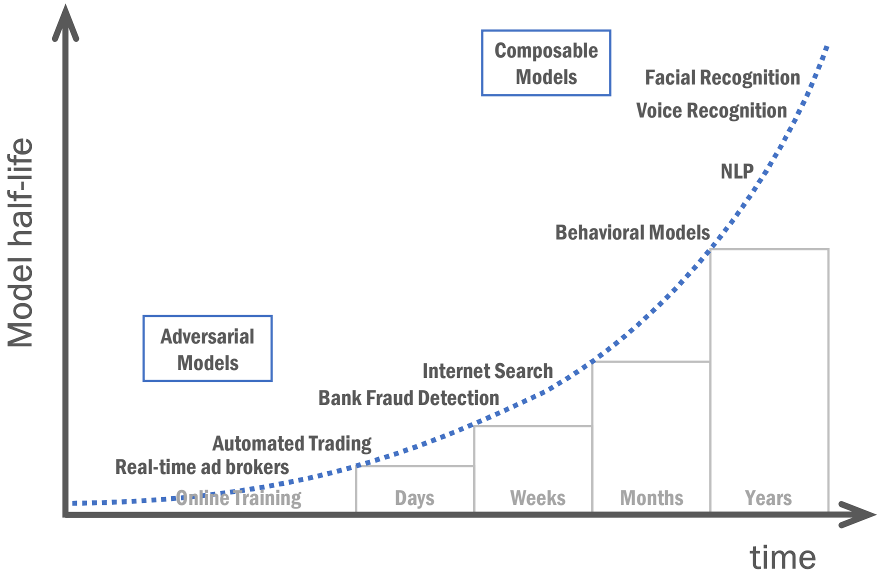
- Models must be retrained when model accuracy degrades
- Retraining is required for the entire life of the model



Source: TLL
Unsupervised/clustering models are more complex.

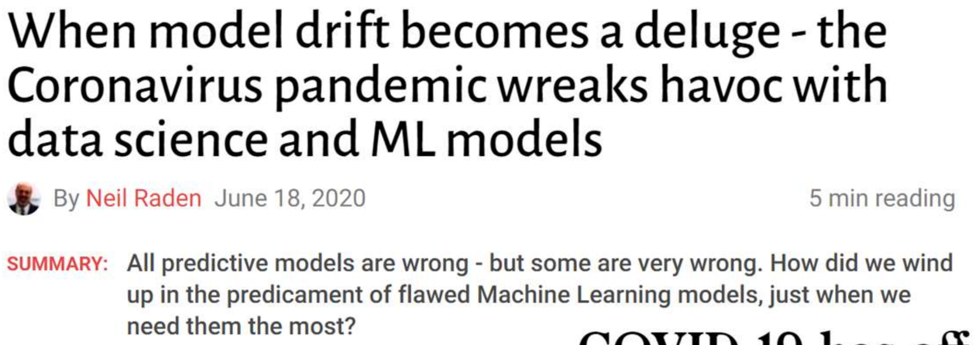
Half-life of Models

- Models have a gradual decline (the half-life) which determines how long the model is effective i.e. how frequently it must be retrained
- The half-life is determined by how quickly the data distribution is shifting ("data drift")



Source: TLL
Unsupervised/clustering models are more complex.

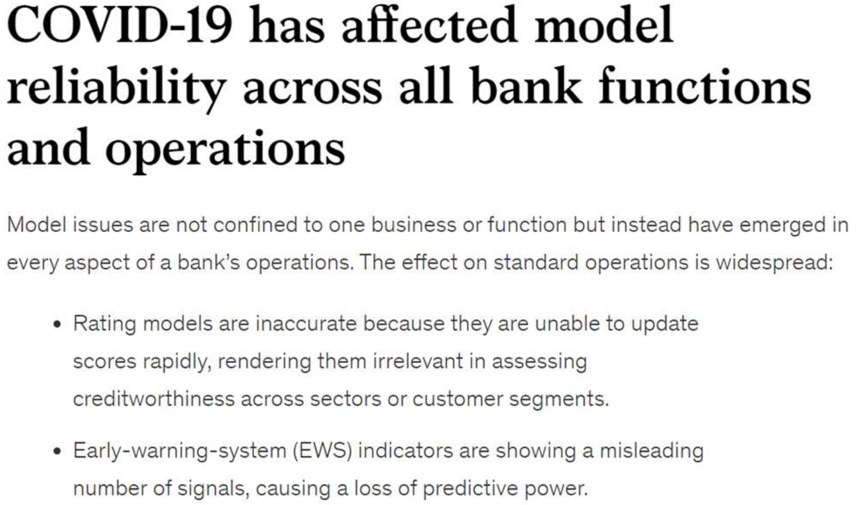
Concept Drift



When model drift becomes a deluge - the Coronavirus pandemic wreaks havoc with data science and ML models

By Neil Raden June 18, 2020 5 min reading

SUMMARY: All predictive models are wrong - but some are very wrong. How did we wind up in the predicament of flawed Machine Learning models, just when we need them the most?



COVID-19 has affected model reliability across all bank functions and operations

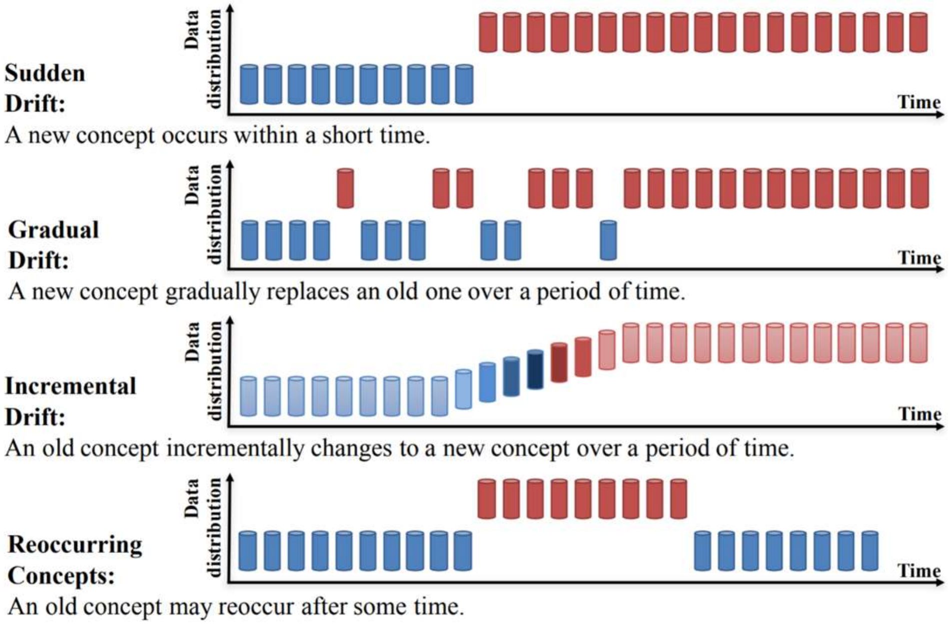
Model issues are not confined to one business or function but instead have emerged in every aspect of a bank's operations. The effect on standard operations is widespread:

- Rating models are inaccurate because they are unable to update scores rapidly, rendering them irrelevant in assessing creditworthiness across sectors or customer segments.
- Early-warning-system (EWS) indicators are showing a misleading number of signals, causing a loss of predictive power.

Source: TLL
Unsupervised/clustering models are more complex.

Half-life of Models

- Concept Drift occurs when the environment changes but the model does not e.g. COVID-19
- Periodic drift reflects periodicity that wasn't captured in training data e.g. seasonality
- Sudden drift may also indicate a catastrophic failure in data processing e.g. data merge
- Its important to track outliers to identify they are part of a new pattern or true one-off events


Sudden Drift:
Data distribution
Time
A new concept occurs within a short time.
Gradual Drift:
Data distribution
Time
A new concept gradually replaces an old one over a period of time.
Incremental Drift:
Data distribution
Time
An old concept incrementally changes to a new concept over a period of time.
Reoccurring Concepts:
Data distribution
Time
An old concept may reoccur after some time.

Source: TLL
Unsupervised/clustering models are more complex.