The Hidden Threats of Decay in AI

June 5, 2020 | Symposium on DS and Statistics

Celeste Fralick, Ph.D.

Chief Data Scientist, Senior Principal Engineer

Office of the CTO

McAfee, LLC

Celeste_fralick@mcafee.com

v2

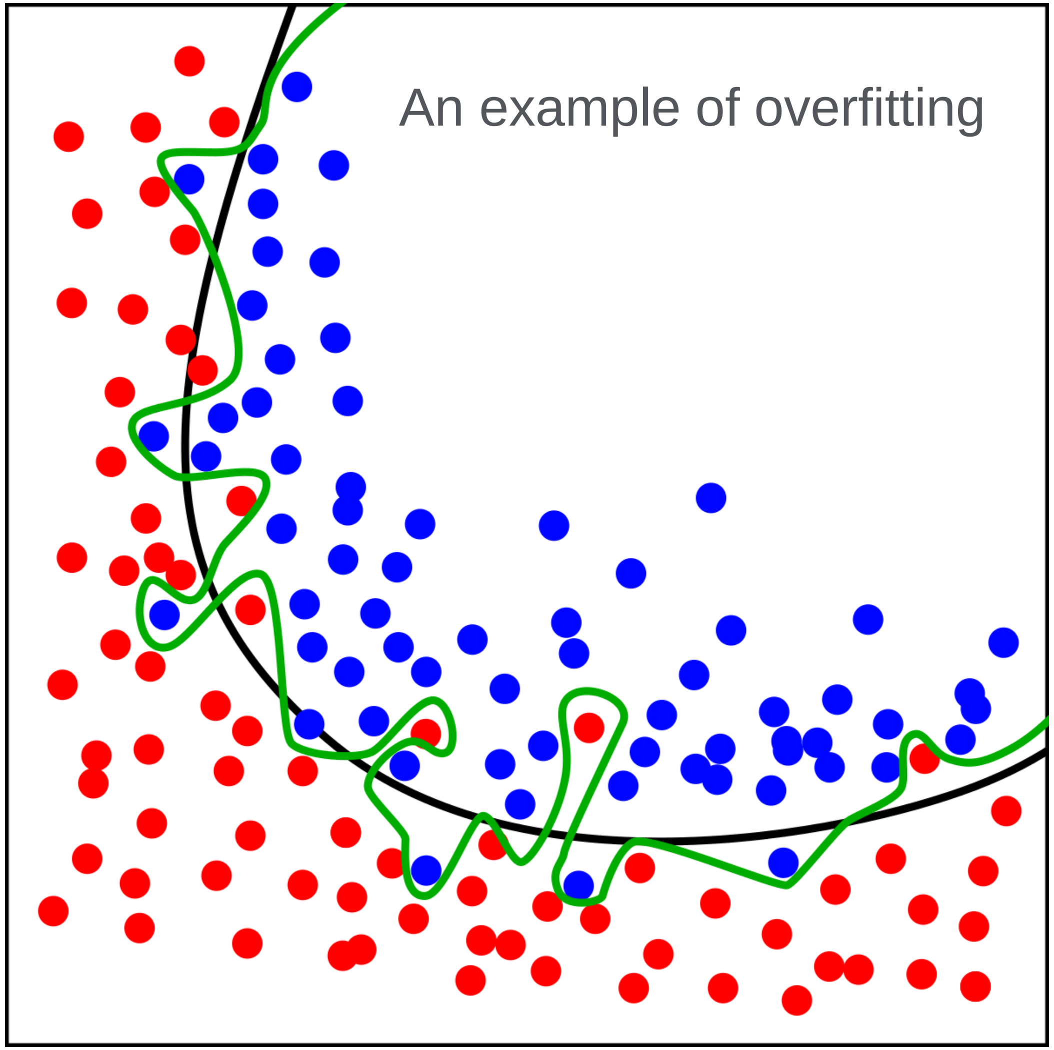
Model Reliability Begins During Development

Do you *really* have a clean data pipeline?

- How do you know? Where did the data come from?
- Is it balanced? Sparse? Dense? Why or why not?
- Have there been data manipulations that impact a sample: filters, caches, down-sampling, etc?
- Is the sample & pipeline repeatable and reproducible?

"Who, What, When, Why, Where and How"

Model Reliability Begins During Development



Check for overfitting *please*

- Training dataset: data used to fit the model
- Validation dataset: data used to validate the generalization ability of the model or for early stopping, *during the training process*.
- Testing dataset: data used to for other purposes other than training and validating.

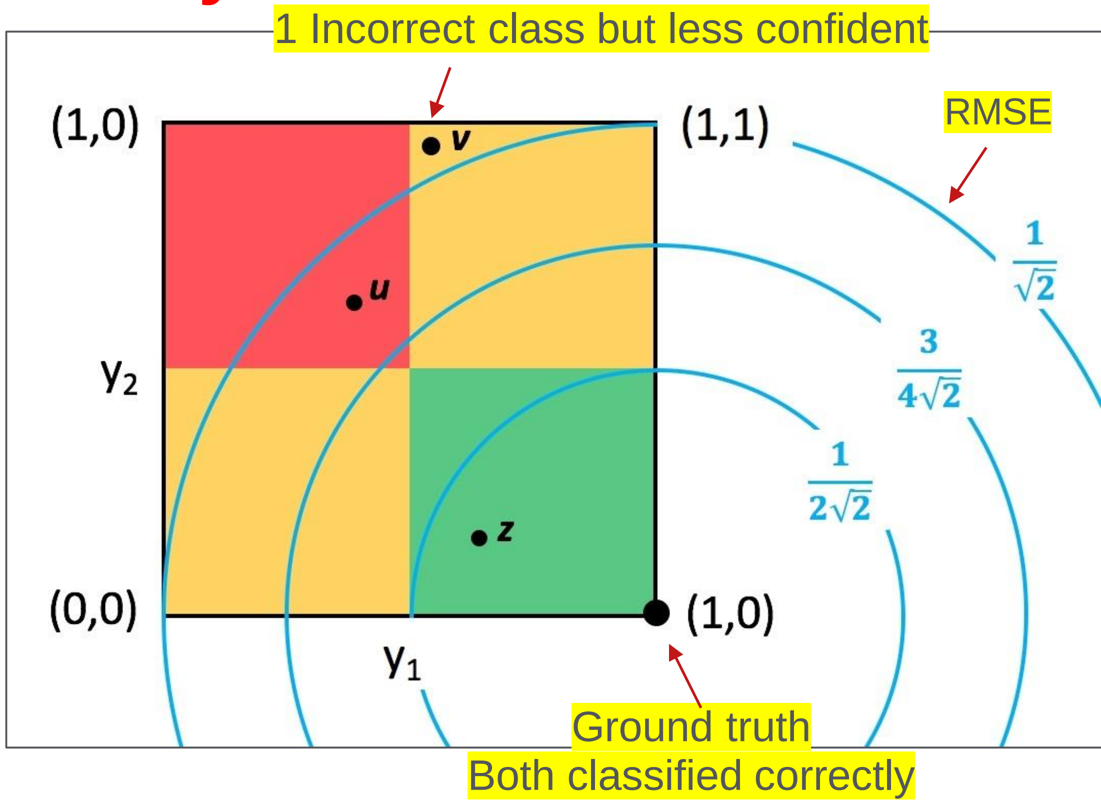
https://stats.stackexchange.com/questions/401696/validation-accuracy-vs-testing-accuracy

Model Reliability Begins During Development

Optimize loss function & target model stability

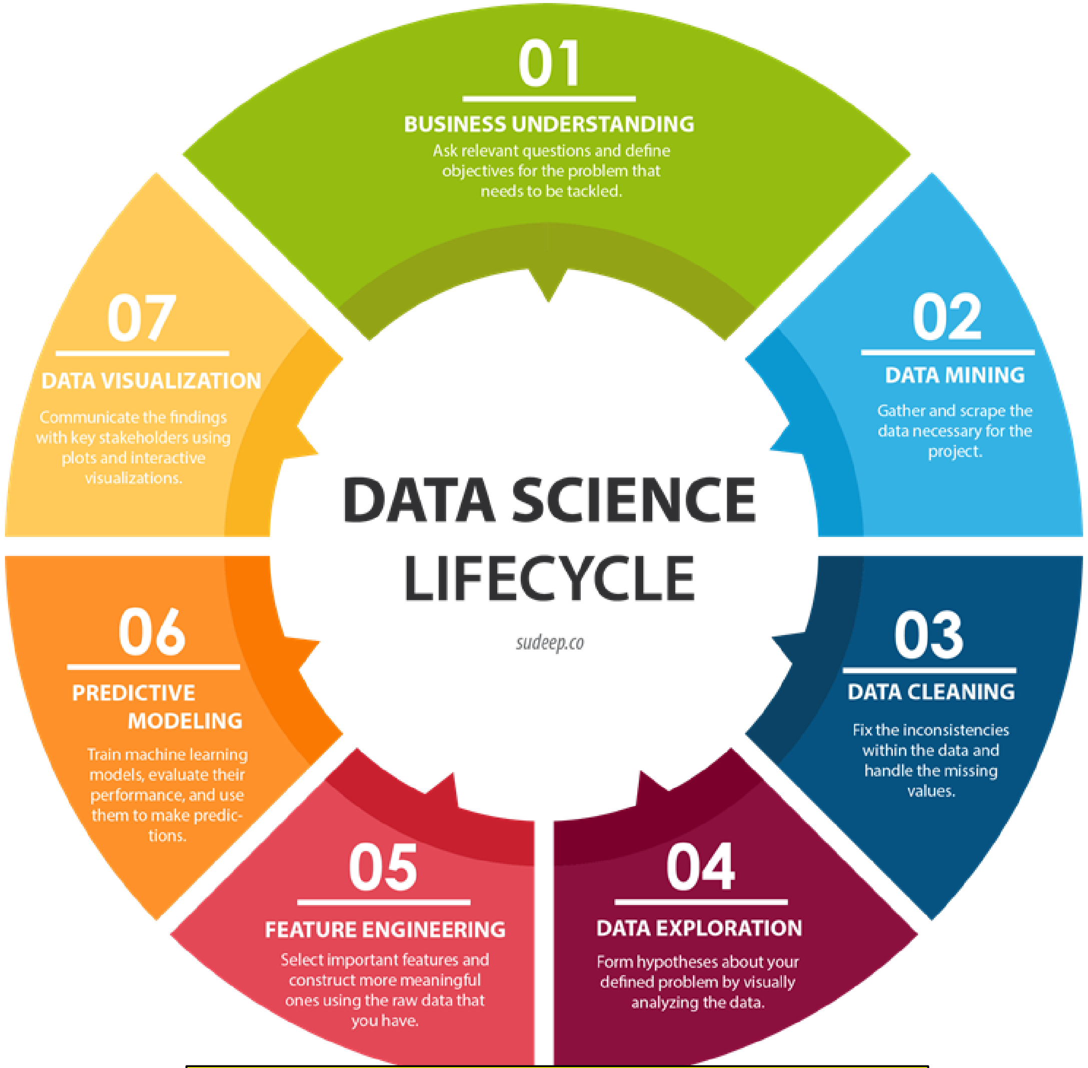
- **3 Models & at least two error rates: ROC and non-ROC (RMSE) (Caruana 2004)**
  - Confidence in predictions
- **Comparing models:**

1. **Covariance Inflation Covariance (CIC): covariance input vs predictor response *(Tibshirani 1999)***
2. **Perturbing training data to overcome local maxima *(Elidan 2002)***
3. **Dual perturb and combine algorithm perturbs test examples w/ random noise *(Geurts 2001)***



- 2 Samples with w/ground truth y1=1, y2=0
- Green square: both classified correctly
- Yellow square: at least 1 classified incorrectly
- Red square: both classified incorrectly
- 3 Models u (least), v, (better) z (best)

https://www.quora.com/How-is-root-mean-square-error-RMSE-and-classification-related,



Where's RISK?

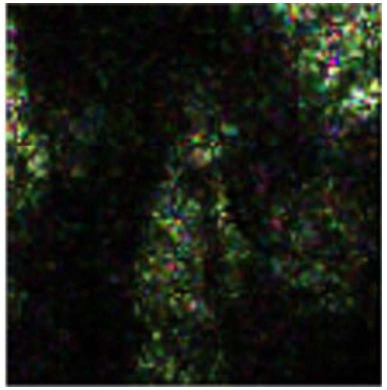
Post-release Analytic Review

- What's changed? Why?
- What are the critical metrics/time?
- Check distributions, error rates
- Has the ground truth evolved?
- Have labels or data evolved?
- How do you monitor change?

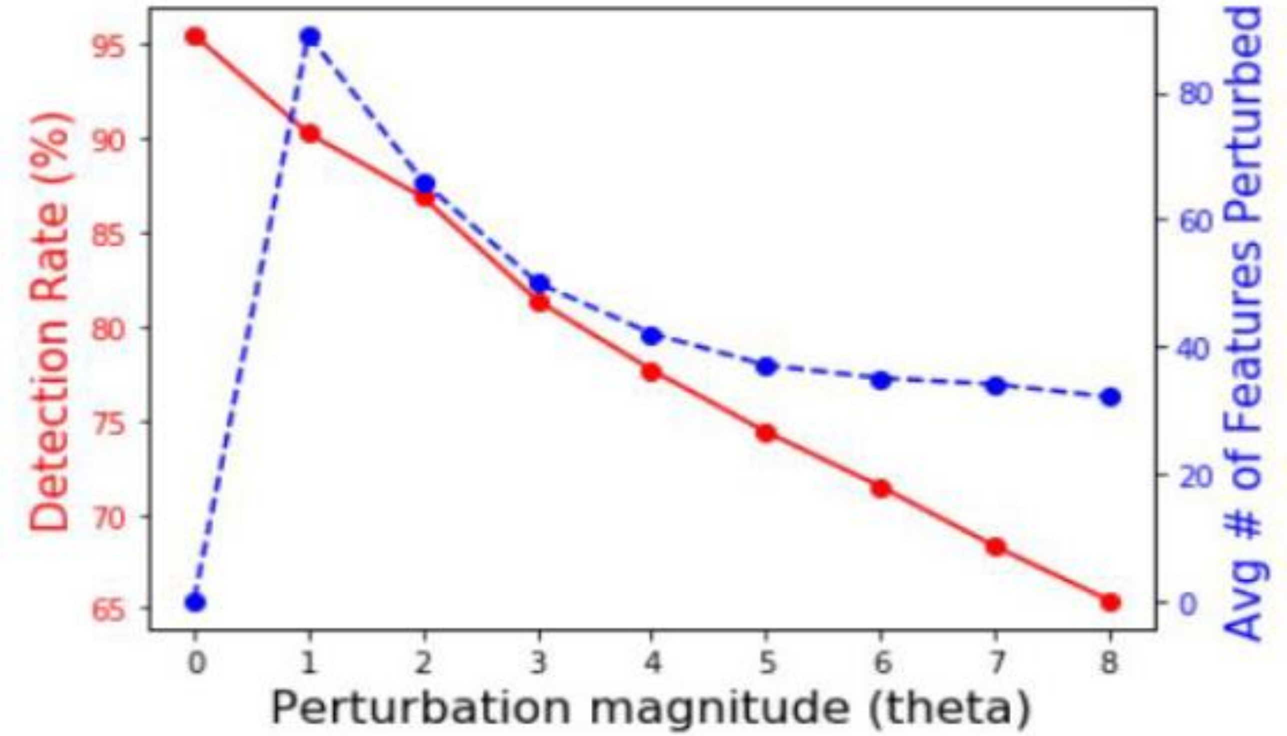
Adversarial Machine Learning: Protect Against Model Hacking

99.68% penguin

93.07% frying pan



Perturbation

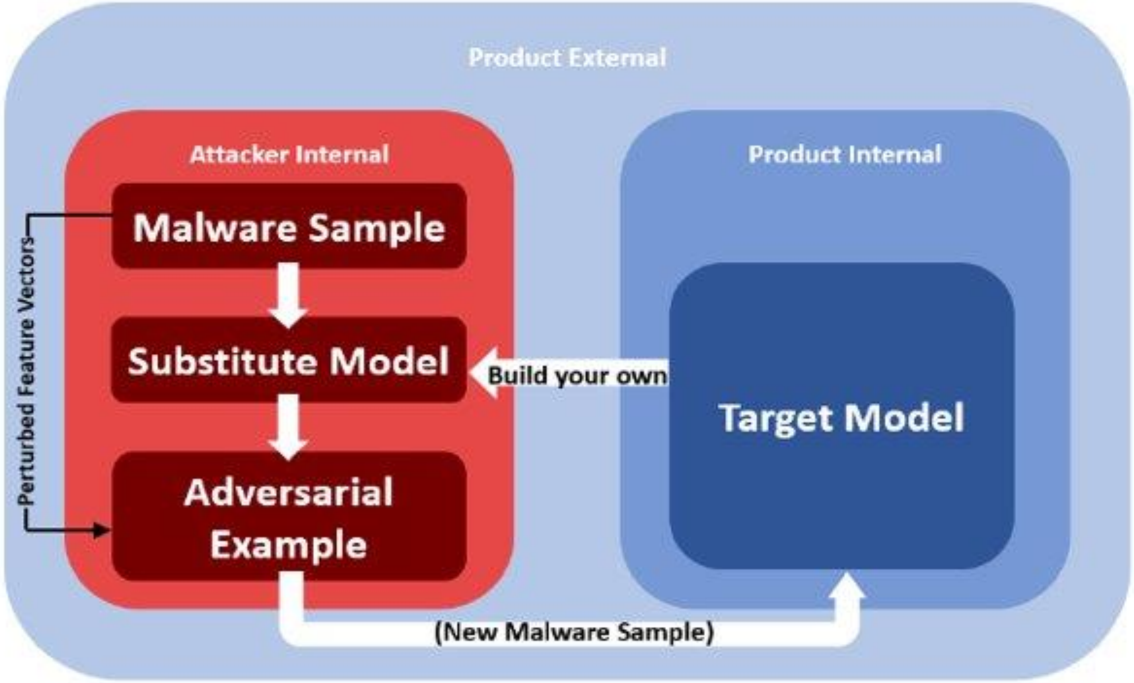


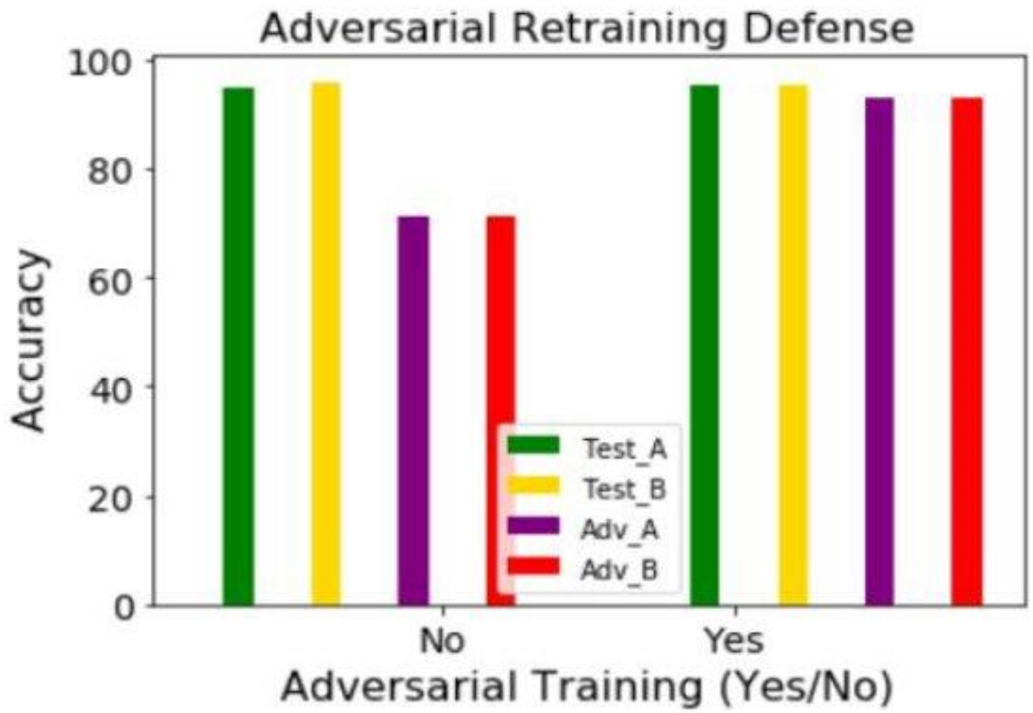
Source Machine Learning Technique \ Target Machine Learning Technique	DNN	LR	SVM	DT	kNN	Ens.
DNN	38.27	23.02	64.32	79.31	8.36	20.72
LR	6.31	91.64	91.43	87.42	11.29	44.14
SVM	2.51	35.56	100.0	80.03	5.19	15.67
DT	0.82	12.22	8.85	89.29	3.31	5.11
KNN	11.75	42.89	82.16	82.95	41.65	31.92

[Papernot 2016]

Adversarial Machine Learning: Model Hacking

- Attack > Detect > Protect
- White/gray/black box testing with prioritized vulnerable features
- Retraining increases test accuracy from 73% >92% by detecting evasion attack without changing original detection rate.
- What happens if FP/FN ↑ ?





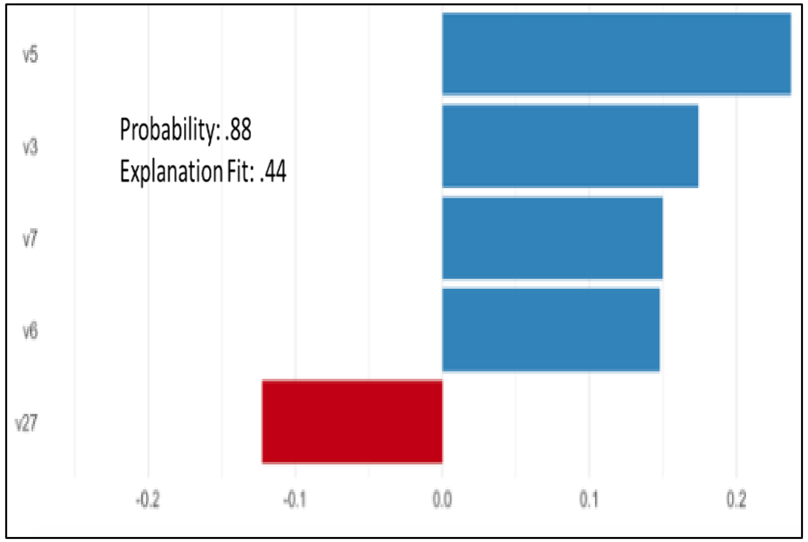
Explainability (XAI) Techniques Can Help Understand Decay

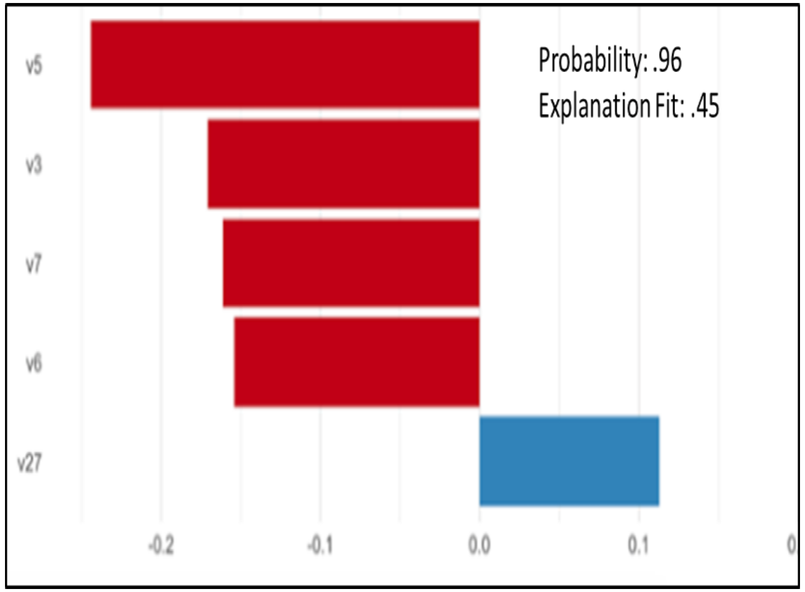
2 unknown cases, but why?

Good label

Bad label

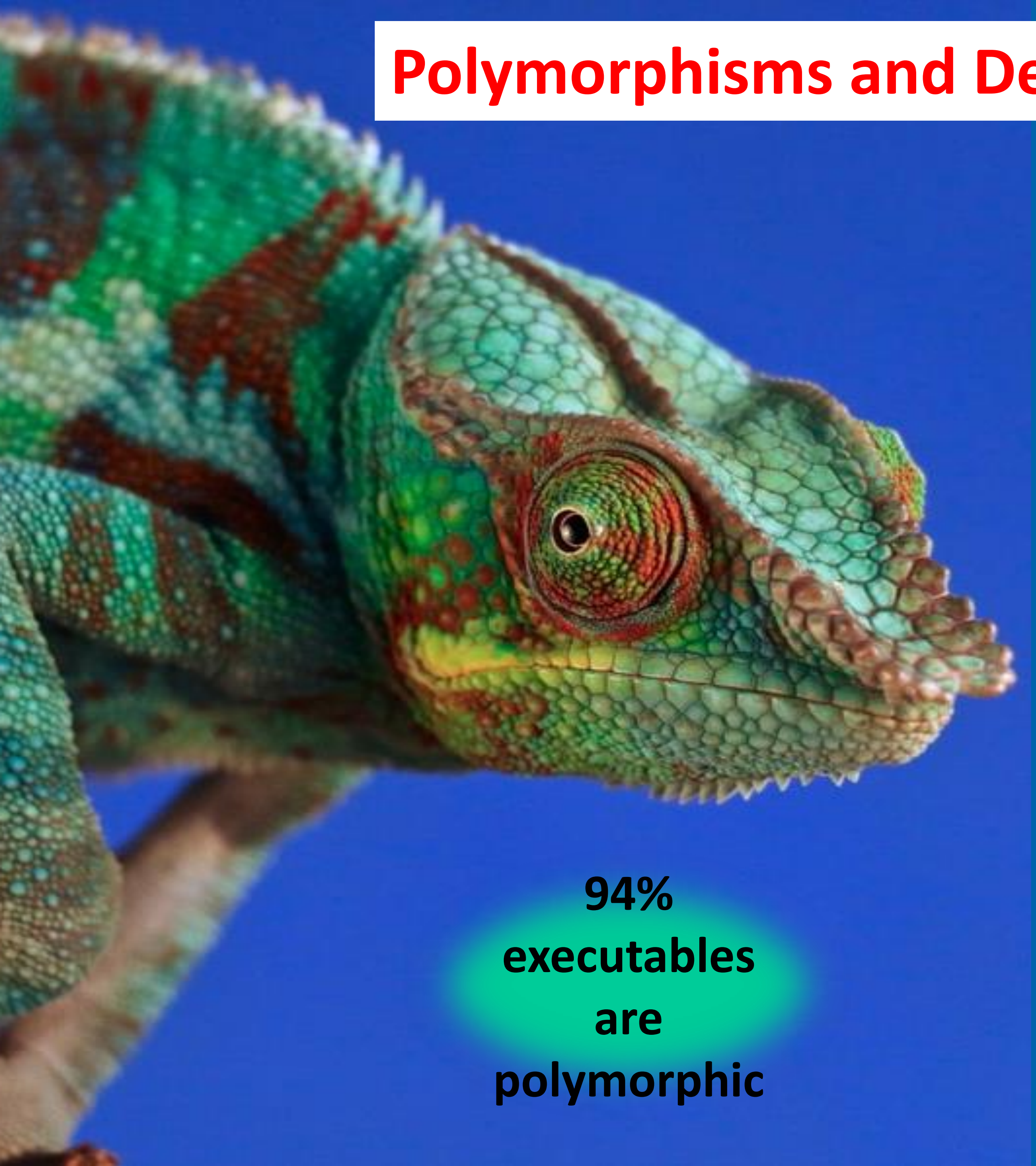
X= feature weight
Y= feature





- Use ML to assess magnitude & direction
- XAI algorithms: LIME, Grad-CAM, SHAP, etc.
- Applicable to intrusion detection, malware
- Understanding unknowns, assessing changes in model field reliability

Polymorphisms and Decay Threaten Model Integrity



- **Concept drift (changing labels/time)**
  - Loss of predictability
  - “No change in distribution”
  - Telemetry / internal
- **Data decay (new data variety/time)**
  - New data, new architecture
  - New categories> labels aren’t changing
- **Learning rate appropriate?**

Minimizing Threats to Decay

1. Monitor and feedback > model reliability
2. Pipeline integrity: "5 W's & 1 H"
3. Incorporate risk management in data science life cycle & AFTER
4. Adversarial Machine Learning (AML/model hacking):
   - Analytic defense: Adversarial Retraining using Examples, Distillation, Feature Squeezing, Noise Addition, RONI, FGSM
   - XAI and vulnerability propensity
5. **Explainability (XAI)**
   - Apply XAI techniques post-dev: LIME, Grad-CAM, SHAP

Combination will minimize threats!

- Thank you.