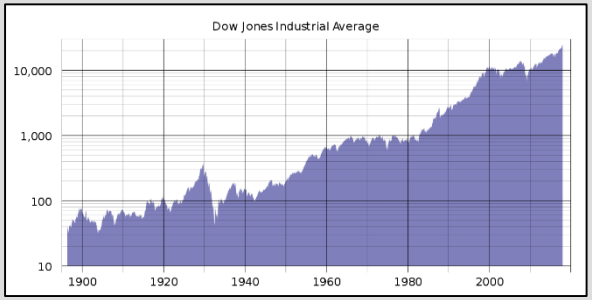
Machine Learning Overview

DATA SCIENCE & ANALYTICS

Topics

- Supervised Methods
- Unsupervised Methods
- ML Workflows
- Metrics

Supervised Methods

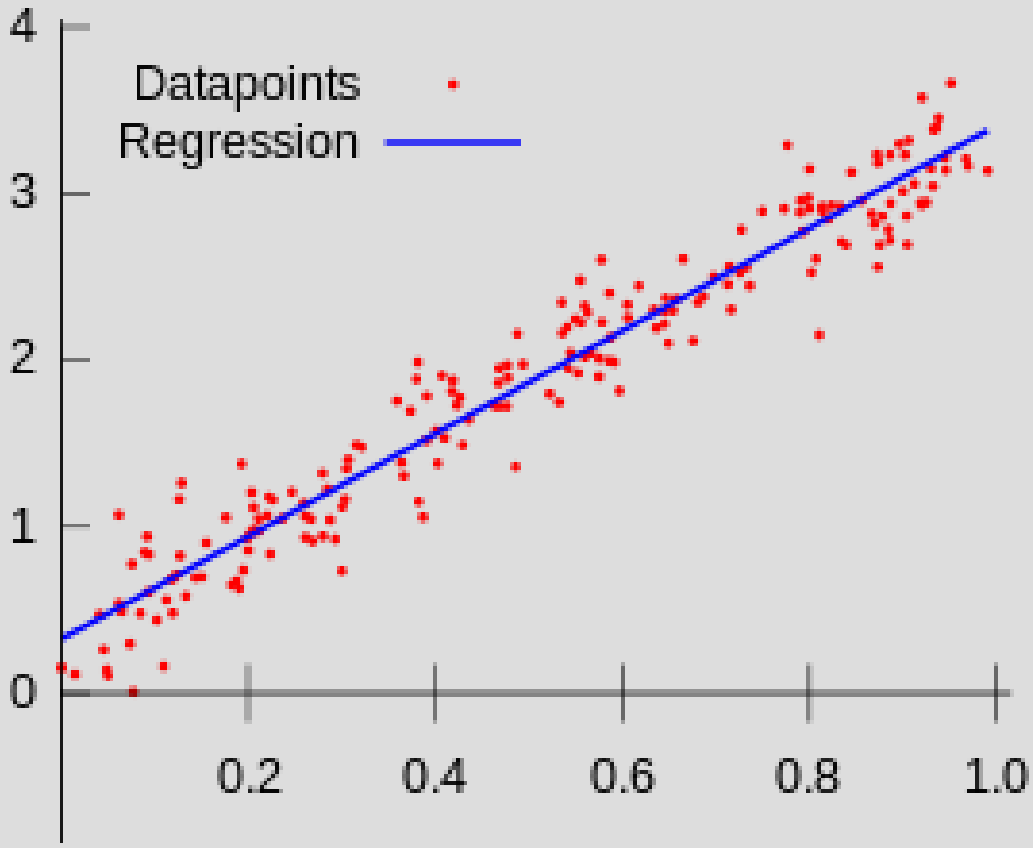


- You can observe
  - Set of data
    - Imagery, stock market, …
  - Labels
    - Class (person, cat), price, etc.
  - Goal
    - ML(new input ; params) => predict value or class
- Techniques
  - Regression, support vector machines, neural networks, minimum risk Bayes decision classifier, decision trees and forests, probabilistic graphical models, perceptron, AND A BUNCH MORE

DATA SCIENCE & ANALYTICS

Example: Regression

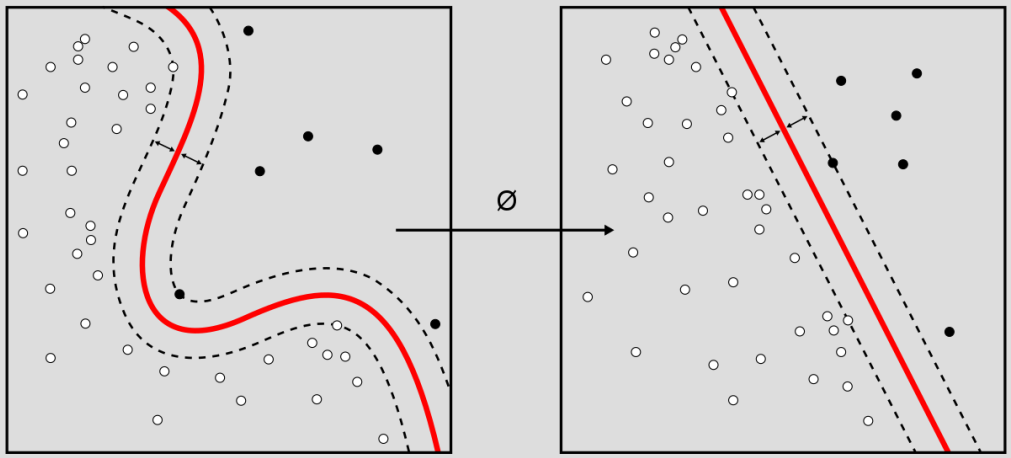
- Dataset
  - x = one feature
  - y = output
- ML training
  - Find best line
    - Slope
    - Intercept
- ML testing
  - Provided a new input, can we predict its value?



DATA SCIENCE & ANALYTICS

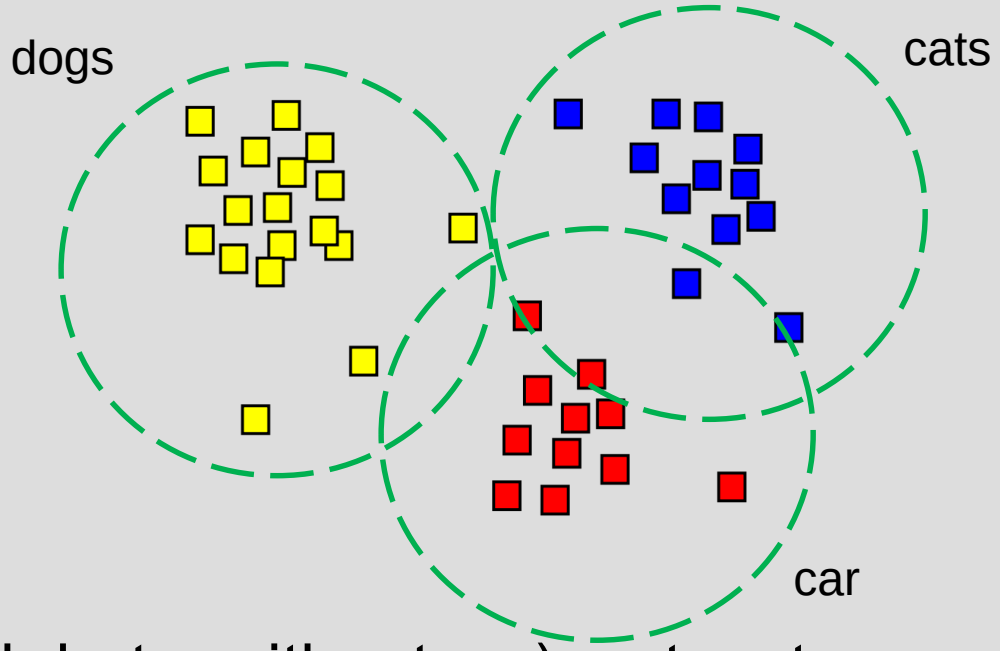
Example: Support Vector Machine

- Dataset
  - x = features
  - y = class label
- ML training
  - Find hyperplane
- ML testing
  - Provided a new input, what class does it belong to?



Unsupervised Methods

- You can observe
  - Set of data
    - Images, cyber, …
  - Labels
    - NOPE
  - Goal = identify *structure*
    - Clusters of people (with hats, without, …), cats, etc.

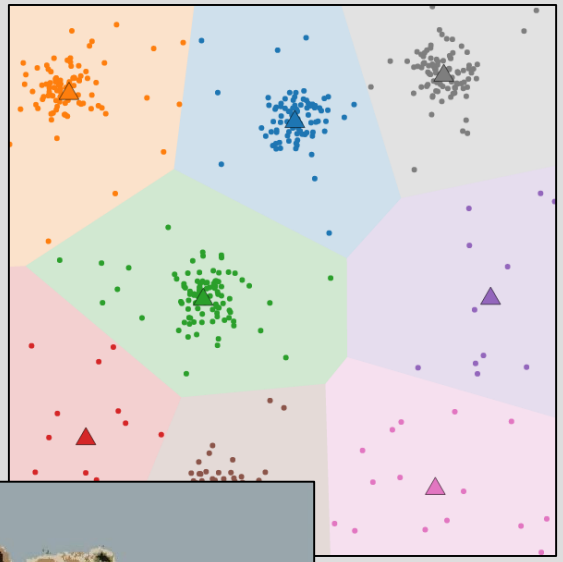


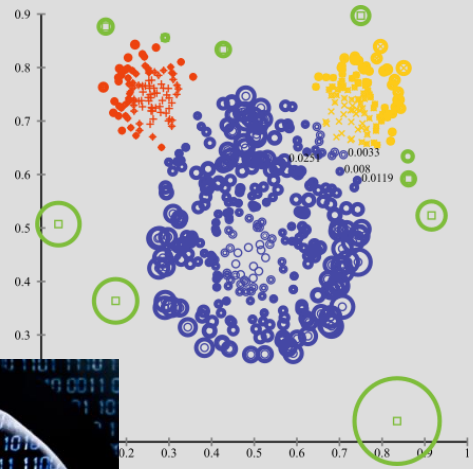
- Techniques
  - K-means, fuzzy c-means, mean shift, DBSCAN, self organizing feature maps (SOFM), neural gas, neural networks, PGMs, AND A BUNCH MORE

DATA SCIENCE & ANALYTICS

Example: k-means

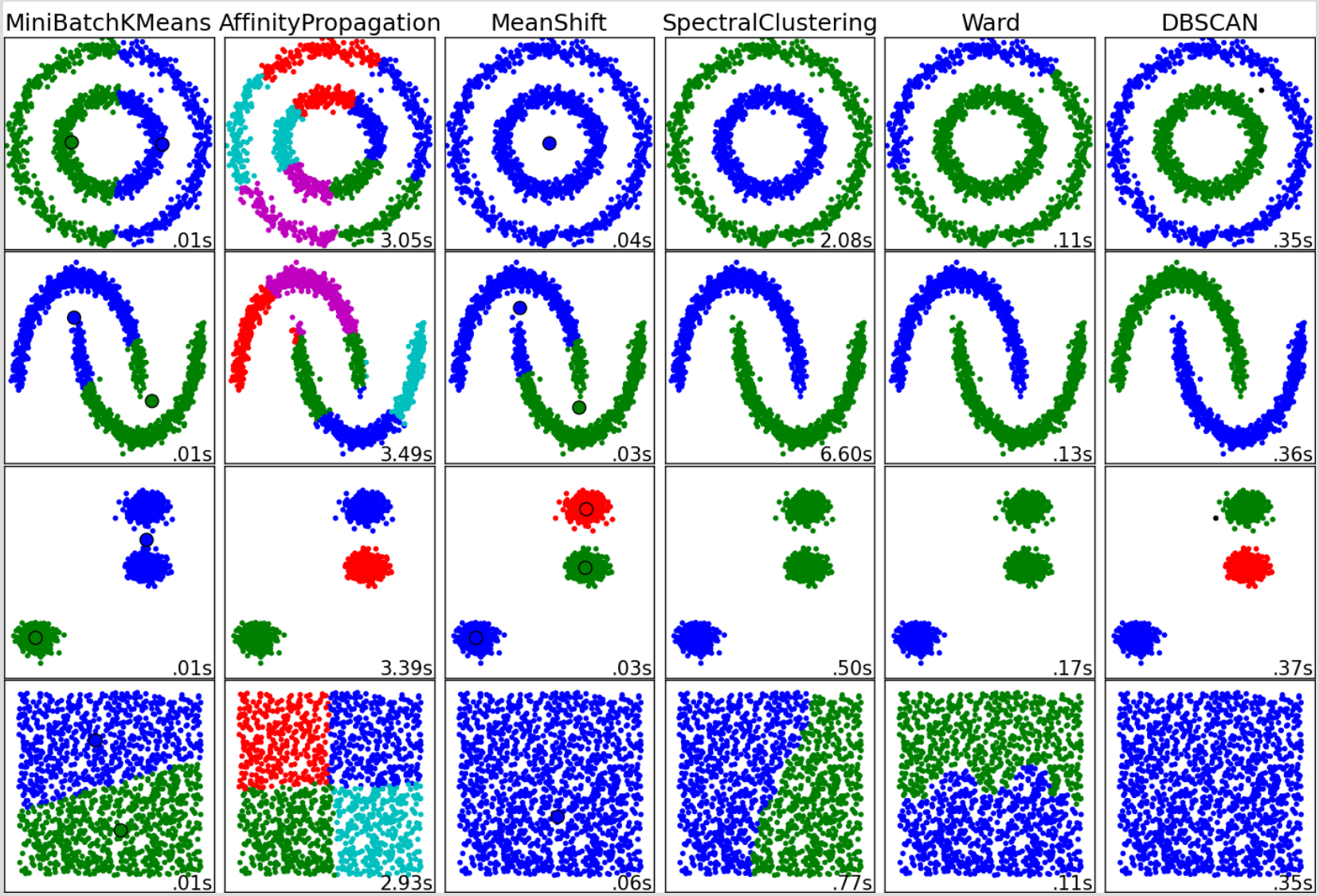
- Dataset
  - x = features
- ML
  - Group data into k different classes
  - Cluster centers
  - Error function
- Examples
  - Image segmentation
  - Anomaly detection and cybersecurity
  - …





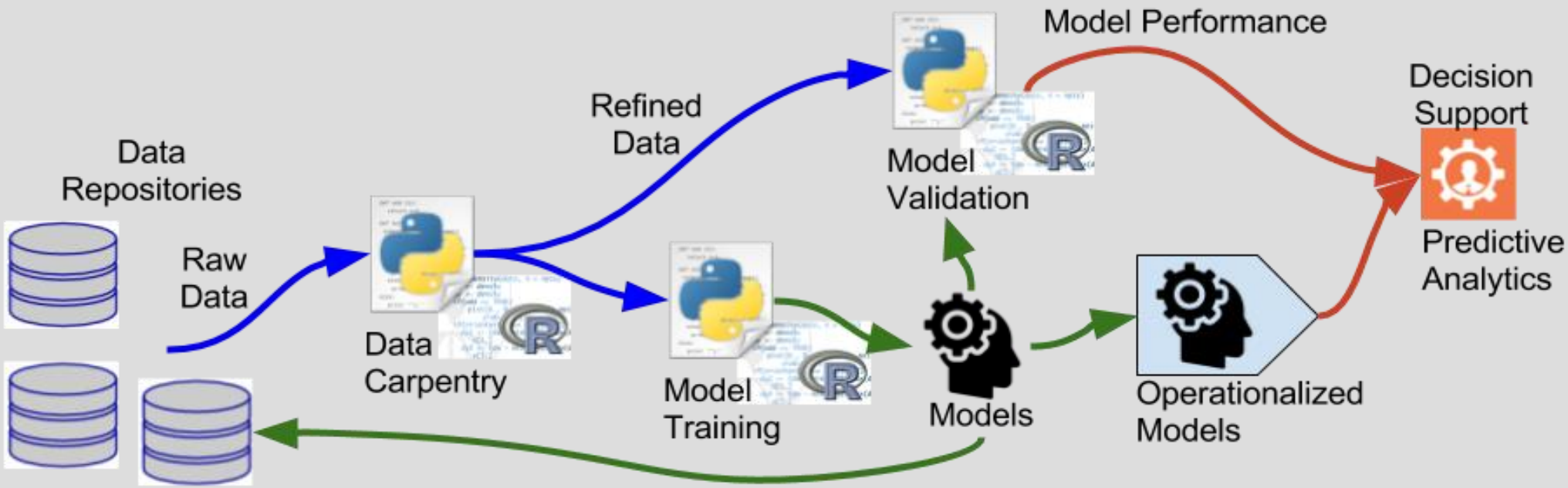
DATA SCIENCE & ANALYTICS

Many Clustering Algorithms



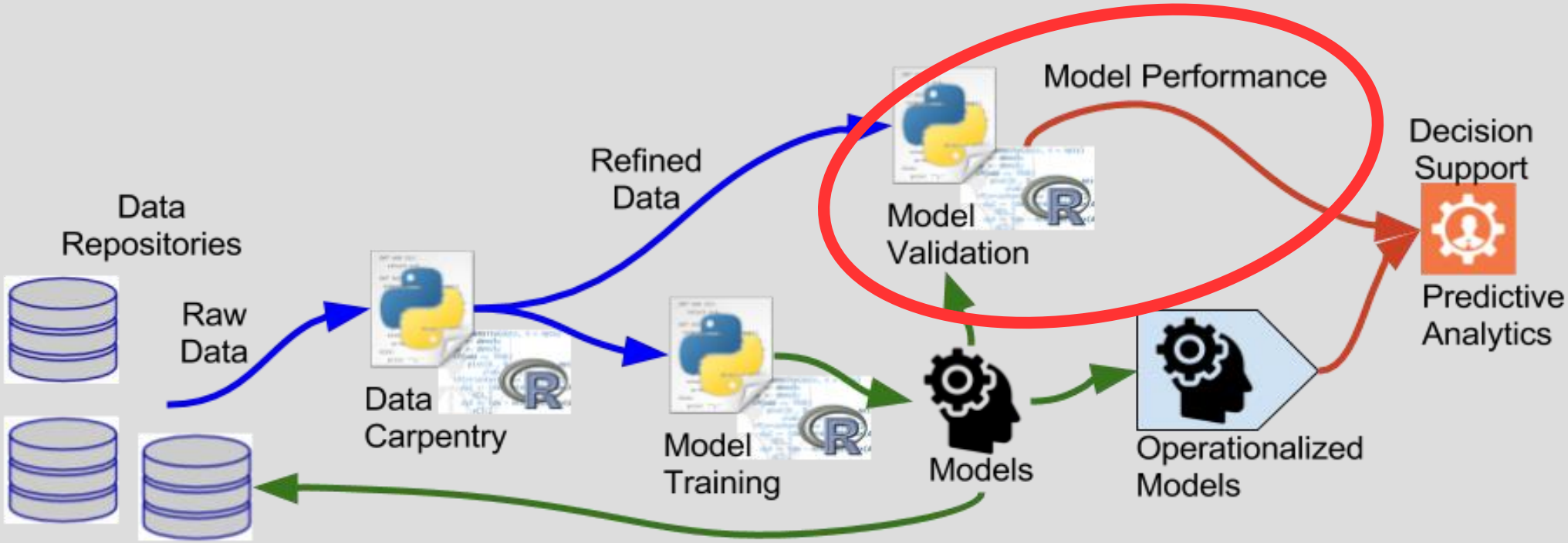
DATA SCIENCE & ANALYTICS

Machine Learning Workflows



DATA SCIENCE & ANALYTICS

Machine Learning Workflows



DATA SCIENCE & ANALYTICS

Regression Metric

- R-Squared (R^2)

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

  - Coefficient of determination
- Adjusted R-Squared

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = R^2 - (1 - R^2)\frac{p}{n-p-1}$$

  - Always less than R^2
  - R^2 can artificially increase with more explanatory variables (independent, predictors, input)
  - Adj. R^2 only increases when the R^2 increase more than likely by random chance

Regression Validation

- Recall : Anscombe's quartet
- Need visualization of data to see that the regression has broken down or is not suitable
- Analysis of residuals (visual and numerical)
  - Random or not?
  - Varied with time?
- Additional Reading
  - https://en.wikipedia.org/wiki/Regression_validation

DATA SCIENCE & ANALYTICS

Classification Metrics

- Consider Two Class Problem (yes/no)
  - When true answer is **yes** and your model says **yes**, that is a True Positive
  - When true answer is **no** and your model says **no**, that is a True Negative
  - When true answer is **yes** and your model says **no**, that is a False Negative
  - When true answer is **no** and your model says **yes**, that is a False Positive

Classification Metrics

- Confusion Matrix

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

DATA SCIENCE & ANALYTICS

Classification Metrics

- Why a 90% “accurate” model is not always good enough
- My favorite professor’s problem
  - Land mine detection algorithms for the US Army
    - Yes: It is a land mine
    - No: It is not a land mine

DATA SCIENCE & ANALYTICS

Classification Metrics

- Why a 90% “accurate” model is not always good enough
- My favorite professor’s problem
  - Land mine detection algorithms for the US Army
    - Yes: It is a land mine
    - No: It is not a land mine
- What is the cost of a False Positive?

Classification Metrics

- Why a 90% “accurate” model is not always good enough
- My favorite professor’s problem
  - Land mine detection algorithms for the US Army
    - Yes: It is a land mine
    - No: It is not a land mine
- What is the cost of a False Positive?
- What is the cost of a False Negative?

Classification Metrics

- Precision: (PPV) positive prediction value
  - How often is a predicted Yes value correct?
- Recall: (TPR) true positive rate
  - How many of the expected Yes are predicted yes?

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Classification Metrics

- F-score (or F_1 Score)
  - Measure of accuracy combining Precision and Recall

$$F_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

DATA SCIENCE & ANALYTICS

Clustering Metrics

- Cluster Validation
  - https://en.wikipedia.org/wiki/Cluster_analysis#Evaluation_and_assessment
- Distance metric driven
  - Ratios of points to centroids or cluster members
  - Euclidean vs Mahalanobis vs other
- Davies-Bouldin index: average ratio of cluster-to-cluster size versus center distance
- Dunn index: ratio between the minimal inter-cluster distance to maximal intra-cluster distance
- Others

DATA SCIENCE & ANALYTICS

Conclusion

- Measures and analyses of machine learning models are critical before operationalizing
  - Fully understand the model
  - Measure the performances against expected and unexpected data
  - Weigh the consequences of erroneous responses

DATA SCIENCE & ANALYTICS

Questions?