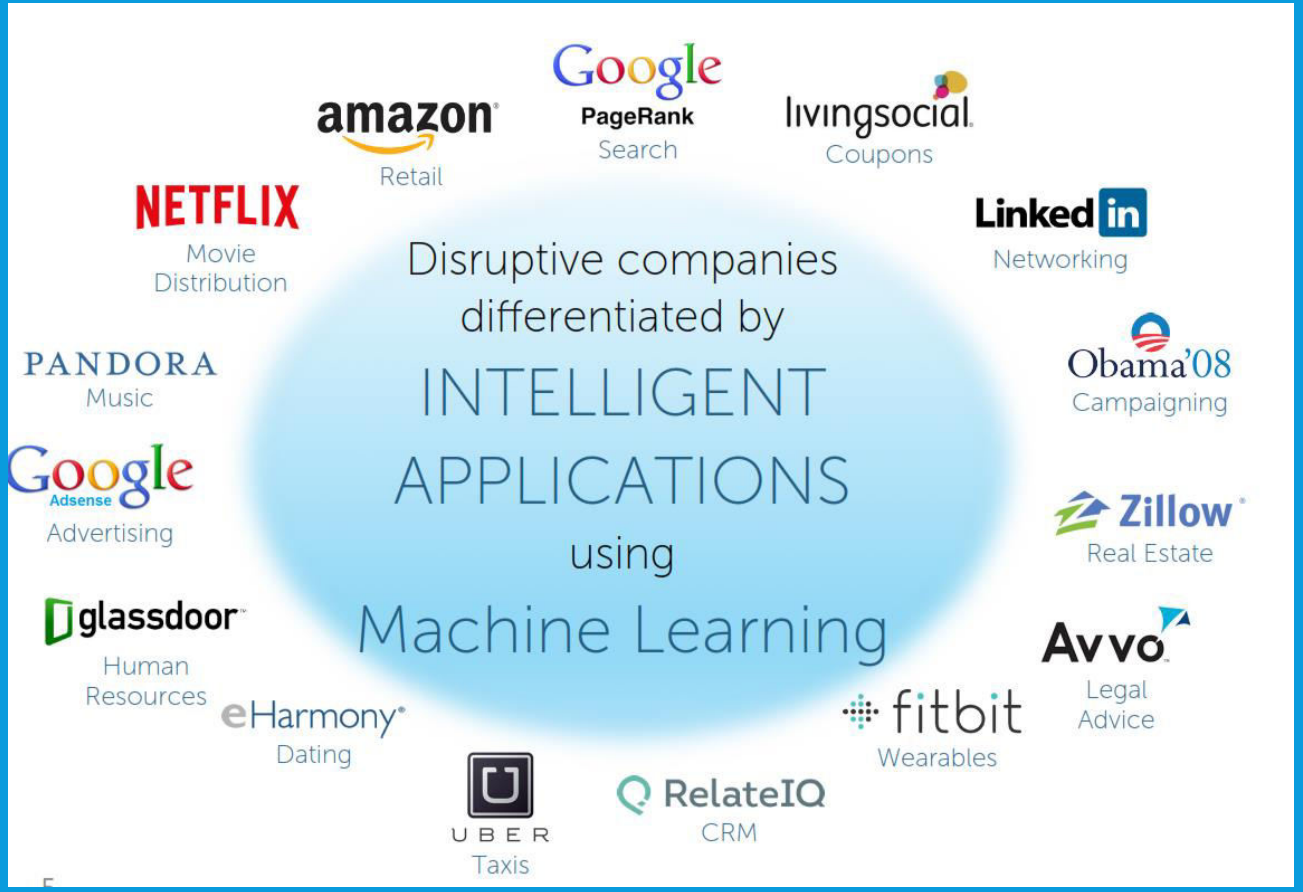
DEVELOPING POWERFUL PREDICTIVE MODELS

Kyu Cho

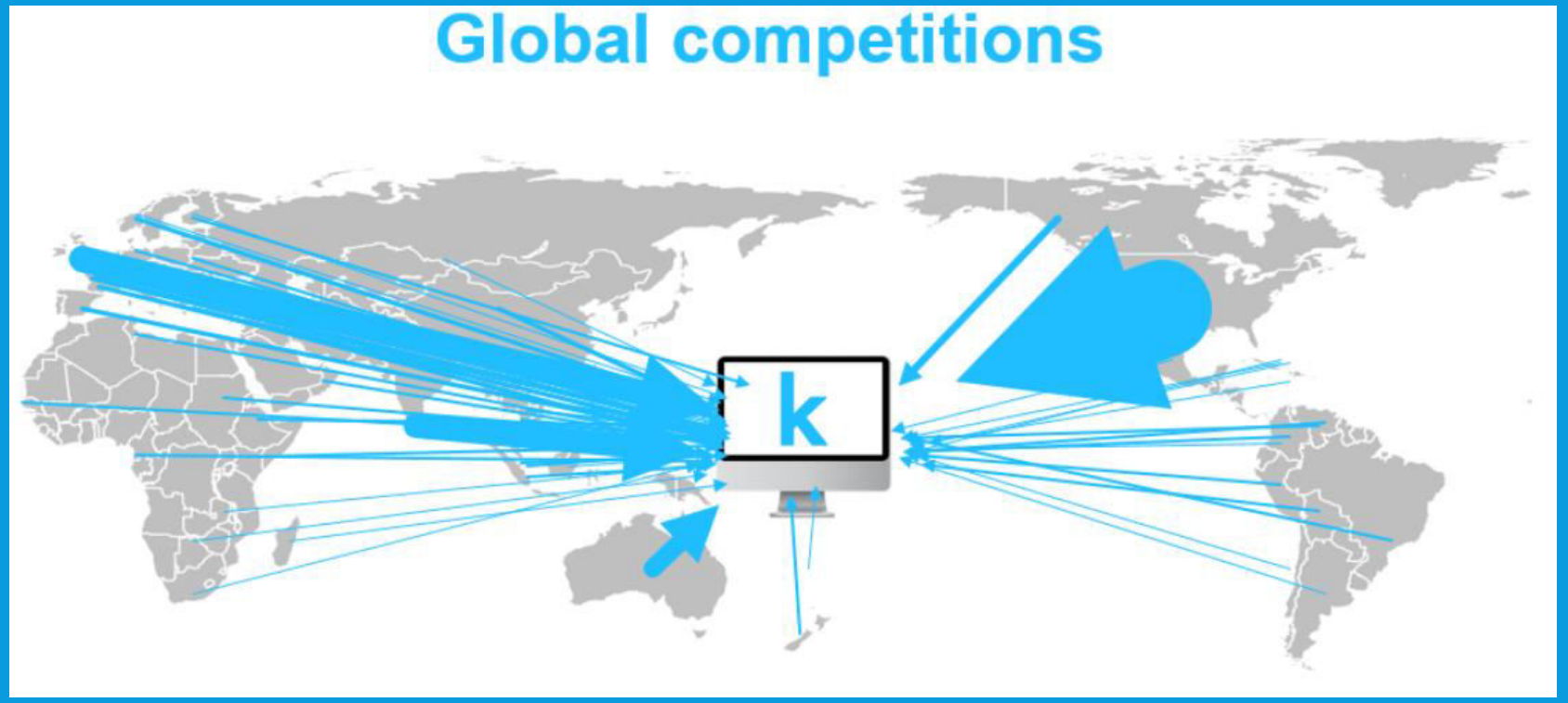
MACHINE LEARNING IN REAL LIFE AND CLOUD COMPUTING

- Internet Search
- Digital Ads
- Recommender System
- Image /speech recognition
- Airline Route Planning
- Healthcare
- Gaming
- Self Driving Cars

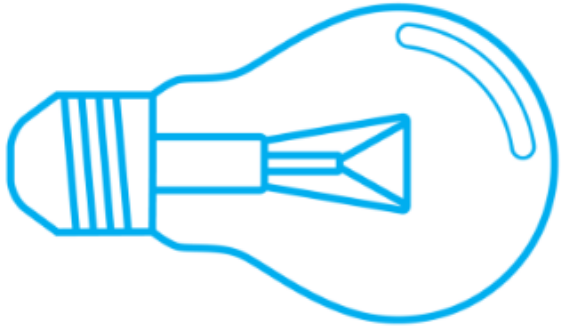


PRACTICE WITH REAL DATA IN KAGGLE

- Kaggle is a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.

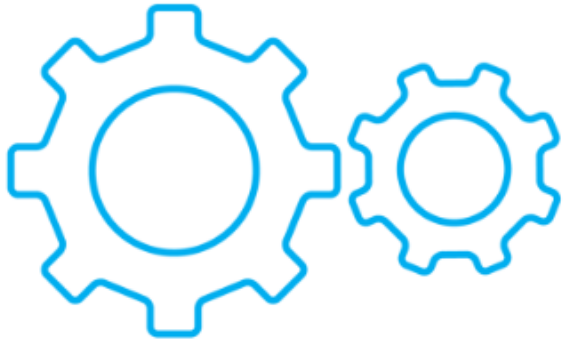


HOW KAGGLE WORKS



1

Users create predictive models,



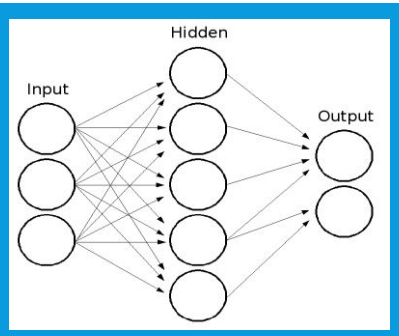
2

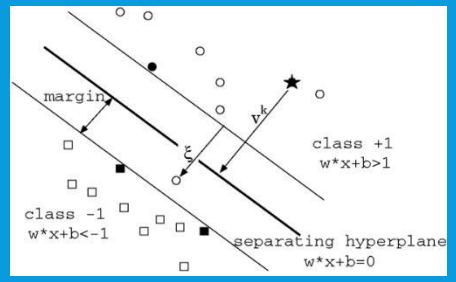
submit these to Kaggle,

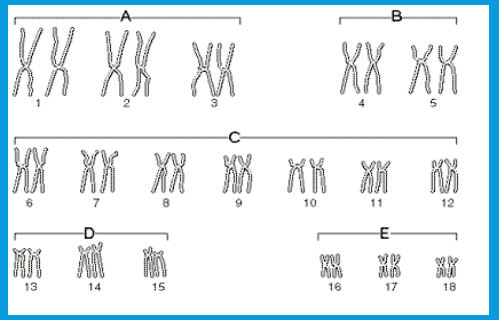
3

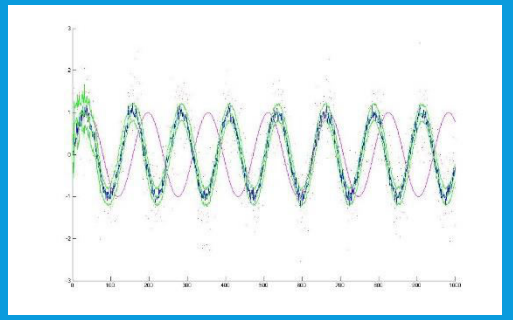
and are scored on their accuracy.

USERS APPLY DIFFERENT TECHNIQUES









- neural networks
- logistic regression
- support vector machine
- decision trees
- ensemble methods
- adaBoost
- Bayesian networks

- genetic algorithms
- random forest
- Monte Carlo methods
- principal component analysis
- Kalman filter
- evolutionary fuzzy modeling

COMPETITIONS ARE JUDGED BASED ON PREDICTIVE ACCURACY

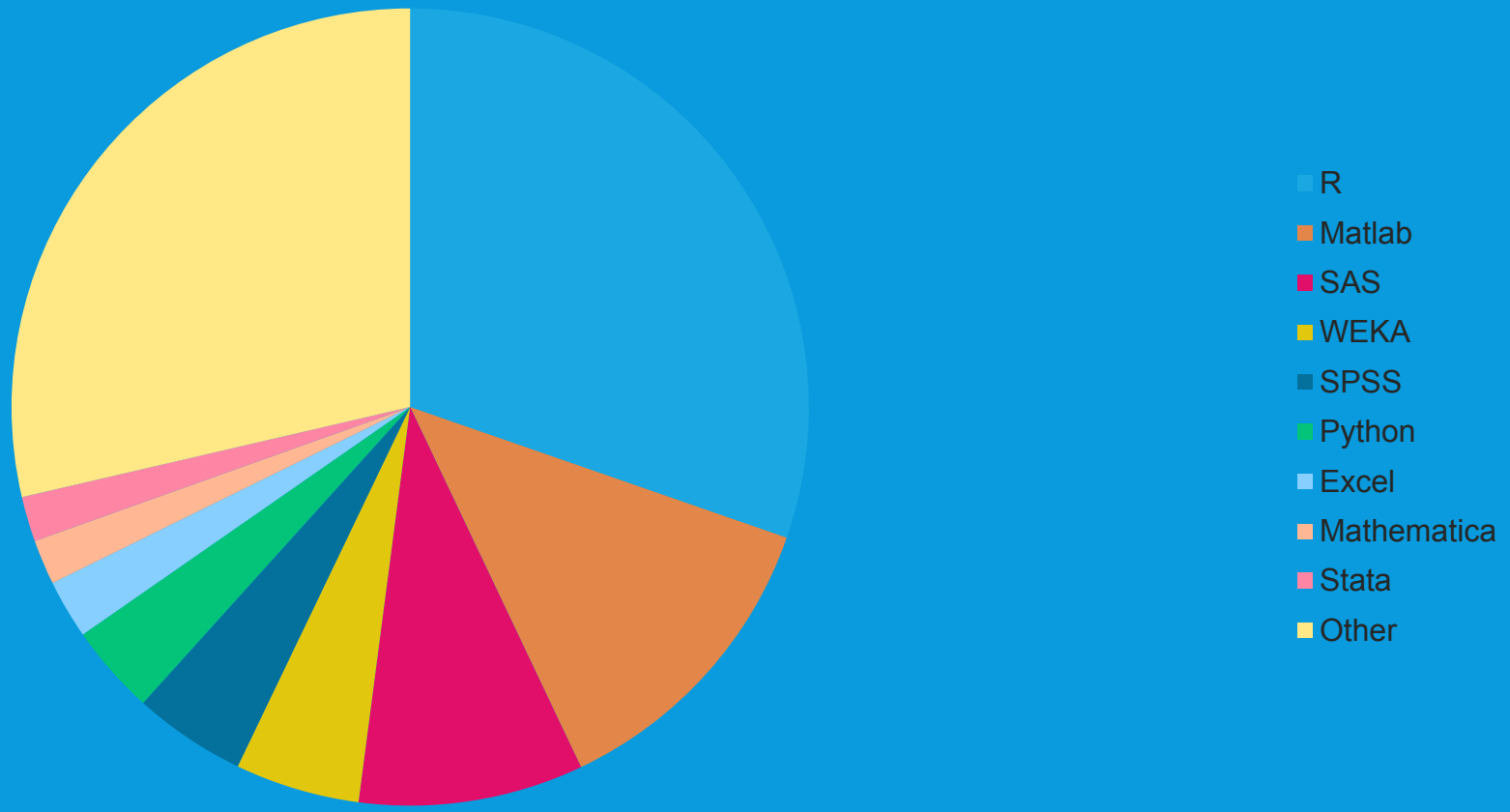
#	Team Name	RMSE	Entries	Latest Submission
1	PEW *	0.640871	130	6:00pm, Monday 1 November 2010
2	UriB *	0.646554	118	9:33am, Saturday 30 October 2010
3	Just For Fun *	0.649665	11	2:34am, Thursday 2 September 2010
4	Old Dogs With New Tricks *	0.649922	87	7:49am, Tuesday 2 November 2010
5	JohnL *	0.652753	11	10:10am, Thursday 7 October 2010
6	PunyPetunias *	0.65485	52	12:04pm, Tuesday 21 September 2010
7	ulvund *	0.655488	52	8:59pm, Thursday 28 October 2010
8	Diogo *	0.655815	85	5:57pm, Monday 1 November 2010
9	Jasonb *	0.656661	50	9:43am, Saturday 23 October 2010
10	ChessMaster *	0.65683	44	6:53pm, Friday 17 September 2010

COMPETITION MECHANICS

Training dataset			
Age		*Income*	*Default*
58	$	95,824.00	TRUE
73	$	20,708.00	FALSE
59	$	82,152.00	FALSE
66	$	25,334.00	FALSE
39	$	35,952.00	FALSE
78	$	51,754.00	FALSE
76	$	76,479.00	TRUE
71	$	96,614.00	TRUE
22	$	27,701.00	FALSE
57	$	35,841.00	FALSE

Test dataset			
Age		*Income*	*Default*
73	$	53,445.00	
61	$	36,679.00	
47	$	90,422.00	
44	$	79,040.00	
46	$	67,104.00	
30	$	69,992.00	
75	$	78,139.00	
28	$	66,058.00	
24	$	75,240.00	
54	$	89,503.00	

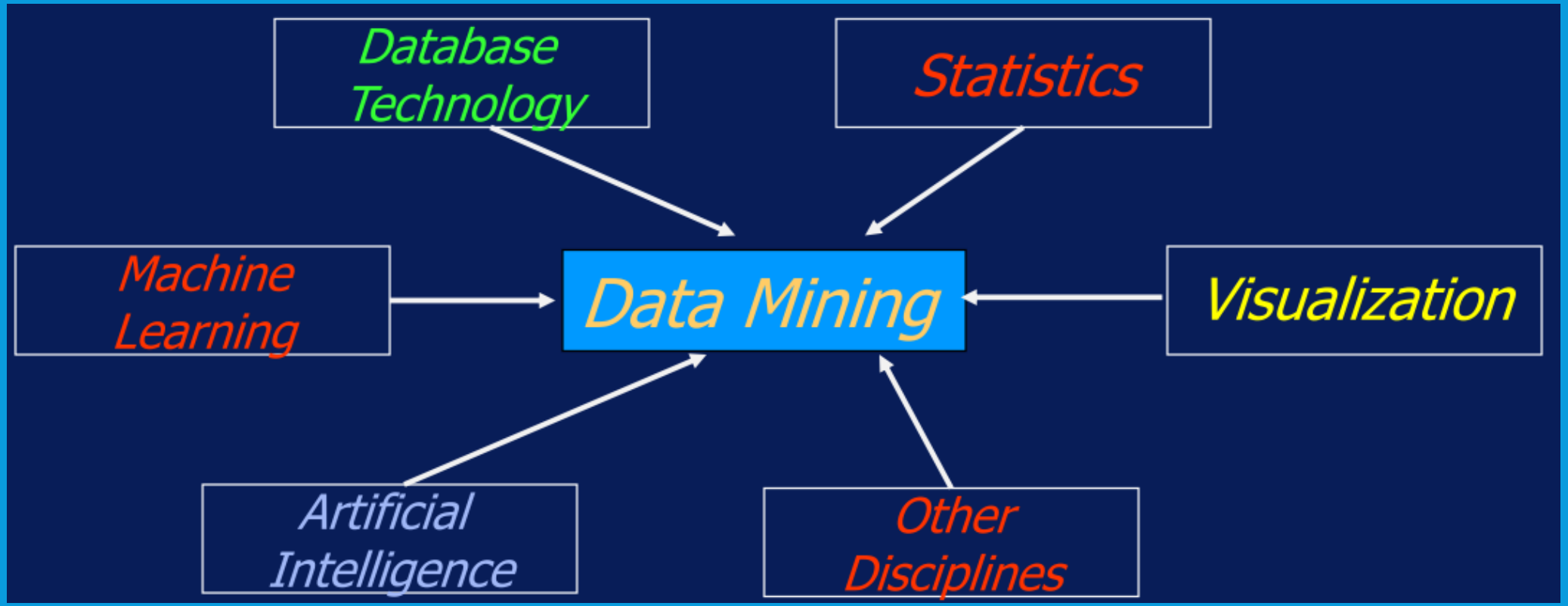
TOOLS IN KAGGLE



OVERALL GOAL

- Goal:
- To transform data into insight for making better decisions.

MULTIDISCIPLINARY FILED



7 STEP PROCESS

- Preprocessing
- Exploration
- Feature Engineering
- Feature Selection
- Building Models
- Ensembling
- Validation

PREPROCESSING

Goal:

To clean data

1) Imputing missing variables

    - Mean, Median, Prediction, Ignore

2) Remove outliers

3) Normalization / Standardization

4) Text variable cleaning

5) Dummify categorical variables

6) Oversampling for rate event

7) Remove highly correlated variables

8) Remove nearly zero variance variables

EXPLORATION

Goal:

To understand the relationship between variables

1) Hypothesis generation
2) Find out Independent Variables and dependent variables
3) Understand the distribution of numerical variables and freq table
4) Scatter Plot, Histogram, Correlation analysis
5) Chi-square test (Categorical vs Categorical)
6) Z-test / T-test, ANOVA (Categorical vs Continuous)

FEATURE ENGINEERING

Goal:

To create meaningful new variables

1) PCA

FEATURE SELECTION

Goal:

To select only influential variables to reduce the noise and computational complexity.

1) Dimensionality reduction

BUILDING MODELS

Goal:

To find the best fit models

1) Regression

- Linear regression, Regularizations: Ridge(L2), Lasso(L1)

2) Classification

- Logistic regression, SVM, Decision Tree

3) Clustering

- K nearest neighbors, Latent Dirichlet allocation(LDA)

200+ more various models exist

ENSEMBLING

Goal:

To enhance the models

1) Bagging (Mean of multiple predictions based model)
2) Boosting (Weight based model)
3) Stacking (Prediction as new variables)

VALIDATION

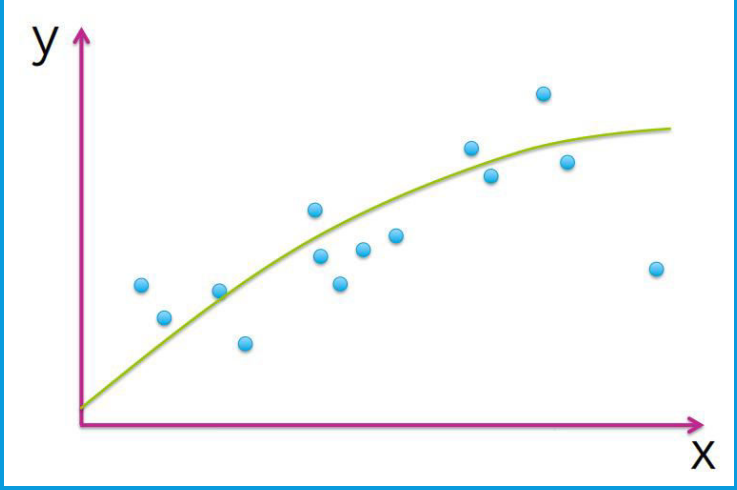
Goal:

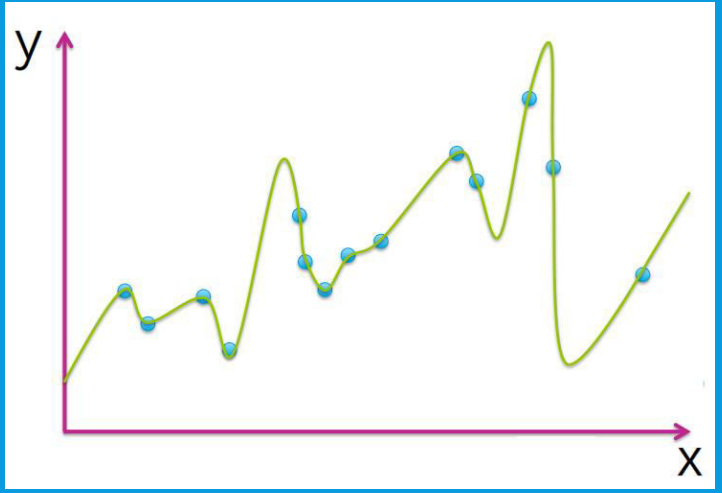
Is to making sure the model is not overfitting and validating errors.

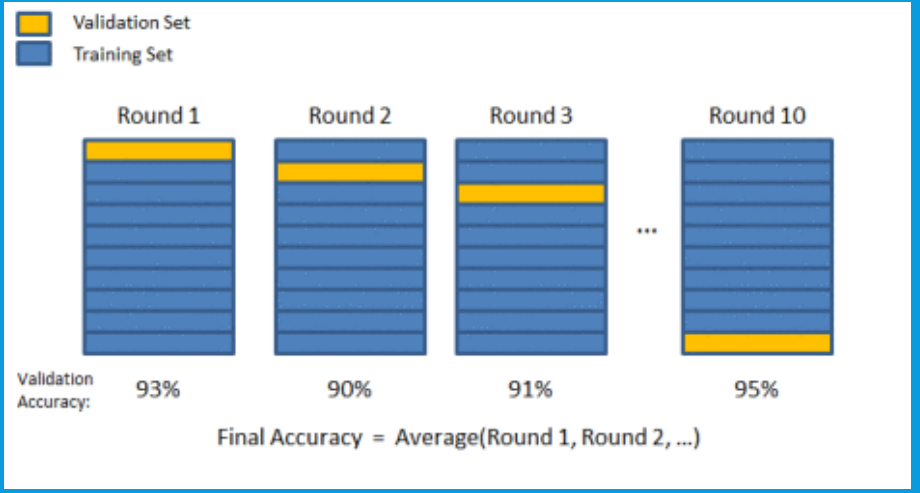
1) K-fold
2) Leave one out
3) Root Mean Square Error (RMSE) – Linear Regression
4) Accuracy - Classification
5) Sensitivity (ratio of positive cases), Specificity (ratio of negative cases)
6) Area Under the Curve (AUC) – ex) Logistic Regression

VALIDATION CONT.

K-fold, leave one out, overfitting, RMSE



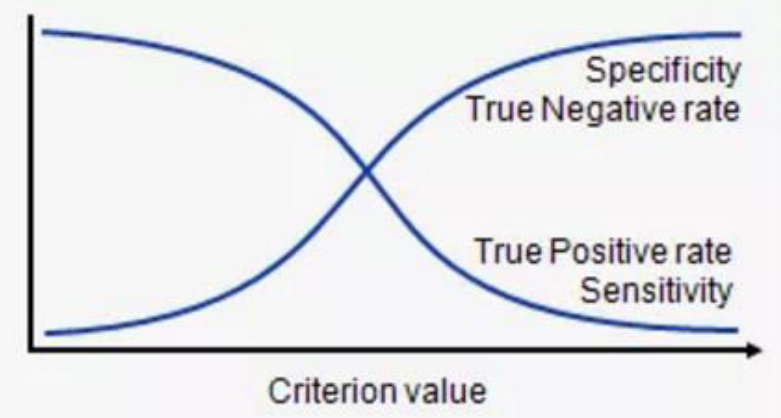


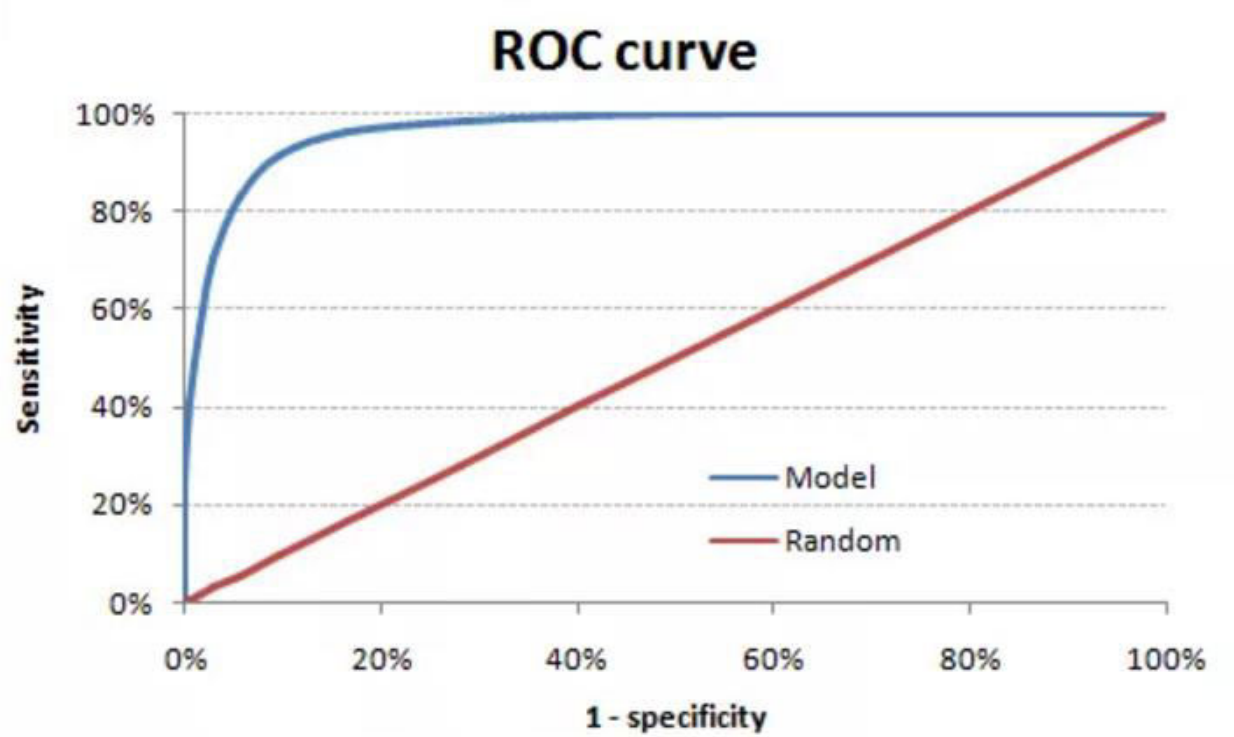


VALIDATION CONT.

Accuracy, Sensitivity, Specificity, AUC

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		





REPEAT

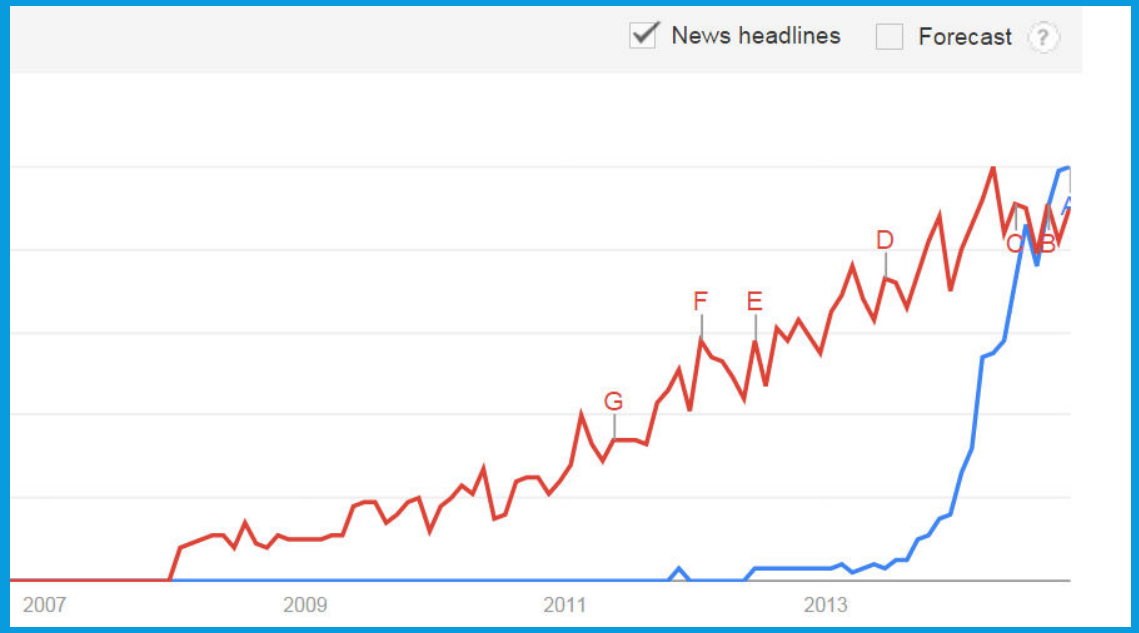
Repeat the whole process until your model is accurate enough but 100% accuracy isn't feasible.

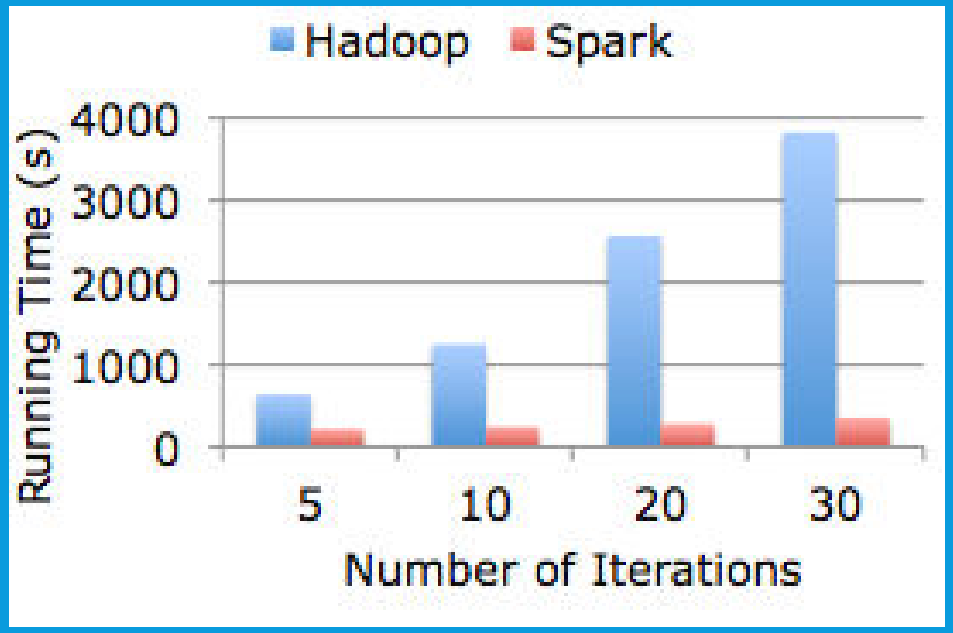
CLOUD COMPUTING - HADOOP

CLOUD COMPUTING – Spark

Spark is a choice for the future big data applications that possibly would require lower latency queries, iterative computation and real time processing on data.

1. Leveraging the memory of the Hadoop cluster. Lower latency computation.
2. Streaming, real time batch processing/modification, machine learning all in the same cluster





MAPREDUCE VS SPARK