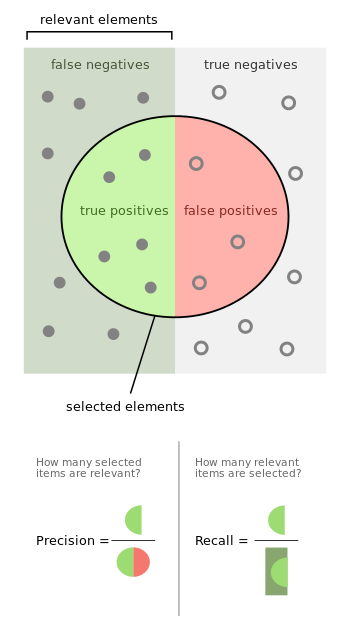
**FINAL Project Proposal and Outline:**

**Machine Inception Learning Algorithm Network (MILAN)**

**Synopsis**

* Use multiple datasets, clean them up and run them thru multiple different machine learning algorithms and compare the results and come up with a “which algorithm(s) work best for which type(s) of data” - in a sense we implicitly set a hypothesis without setting one and prove or disprove it or at least give it some direction…
* Use Supervised Learning Models
* Apply algorithms twice:
  1. using sklearn
  2. using pyspark
* Data Sets (we will drop 4 for a total of 8):
  1. Titanic Passengers
  2. WikiStats
  3. Weblogs
  4. TextCorpus
  5. Wine Quality
  6. Stock Analysis
  7. Sample Stocks
  8. Sentiment Analysis
  9. San Francisco Parking
  10. Enron
  11. Space Data Set
  12. Health: Indians Diabetes
* Visualization of results in Tableau
* Potential algorithms:
  1. RF
  2. XGBoost
  3. AdaBoost
  4. Ensemble
  5. Logistic regression
  6. Linear regression
  7. K-means clustering
  8. Random forest
  9. Naive Bayes Classifier
  10. K-nearest neighbors
  11. Another reference for this: <https://www.kdnuggets.com/2016/08/10-algorithms-machine-learning-engineers.html>

**Processes:**

* We will use multiple datasets to train and test and use the Performance measure to discuss the improvement. We may can use Cohen’s Kappa score, accuracy, F1 score, and or confusion matrix. We will SVM for supervised learning.
* **Cohen's kappa** coefficient (κ) is a **statistic** which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance.
* Precision and recall
  1. In [statistical](https://en.wikipedia.org/wiki/Statistics) analysis of [binary classification](https://en.wikipedia.org/wiki/Binary_classification), the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the [precision](https://en.wikipedia.org/wiki/Precision_(information_retrieval)) p and the [recall](https://en.wikipedia.org/wiki/Recall_(information_retrieval)) r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the [harmonic average](https://en.wikipedia.org/wiki/Harmonic_mean) of the [precision and recall](https://en.wikipedia.org/wiki/Precision_and_recall), where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.
* [](https://en.wikipedia.org/wiki/File:Precisionrecall.svg)
* Confusion Matrix:
  1. In the field of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and specifically the problem of [statistical classification](https://en.wikipedia.org/wiki/Statistical_classification), a **confusion matrix**, also known as an error matrix,[[4]](https://en.wikipedia.org/wiki/Confusion_matrix#cite_note-4) is a specific table layout that allows visualization of the performance of an algorithm, typically a [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) one (in [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning) it is usually called a **matching matrix**). Each row of the [Matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)) represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa).[[2]](https://en.wikipedia.org/wiki/Confusion_matrix#cite_note-Powers2011-2) The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).
  2. It is a special kind of [contingency table](https://en.wikipedia.org/wiki/Contingency_table), with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).
* <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>
* <http://www.cs.cornell.edu/courses/cs678/2006sp/performance_measures.4up.pdf>
* <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>

Technology Stack:

* Application/Analysis Language/Env: Python / Jupyter Notebook
* Libraries: Spark, Scikit
* Source Data: CSV Files
* Backend DB Results storage: Snowflake DB on AWS
* Visualization / Front End: Tableau

Workflow:

1. Develop Python scripts with multiple algorithms
2. Upload CSV data into Python Data Dictionaries
3. Run Data sets against algorithms
4. Store Results in SnowFlake
5. Display via Tableau
6. Writeup 1-2-page analysis and results observed coupled with a final, edited version of this document

Team:

* Aref Erfani
* Khapil Danger
* Filis Cobas
* Andrew Abraham

Assignments:

* Python Scripts: Aref, Andrew, Filis, Kapil
* DB Environment: Kapil, Aref
* Tableau: Kapil, Andrew, Filis
* Written Analysis, Ideation and Presentation: Aref