# Handwritten Digit Recognition

Knowledge Discovery and Data Mining

Claudio Aracena | Rafael Possas | Tinju Abraham | Tengfei Shan

THE UNIVERSITY OF
SYDNEY

- 80 different people / Each wrote 0-9 digits twice (accurately and inaccurately)
- 1593 rows with 256 columns each
- Scanned and stretched in a rectangular box of 16x16 gray scale of 256 values.
- Fixed threshold (127) applied on each pixel scaling into a boolean value 0/1
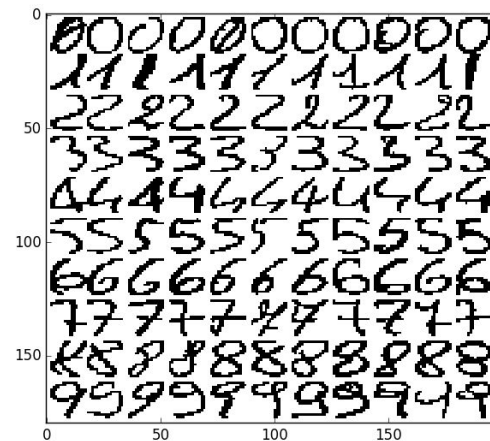
- 9 Different algorithms / 169 combinations
- Execution time with and without multiprocessing: 151.2 secs and 470.9 secs
- Accuracy range: 70.34% - 95.92%
- Top 2 Algorithms: SVM and Random Forest
- Tools: Python, Scikit-learn, Pandas, Matplotlib, Joblib and cProfile
- Top 10 first results were achieved using SVM with different parameters

Algorithm	Accuracy	Parameters
Support Vector Machine	95.92%	C=10, gamma=0.03, kernel= rbf
Random Forest	94.96%	n_estimators=300, max_features=log2, max_depth=None
Logistic Regression	91.58%	C=0.1, multi_class=ovr
K-Nearest Neighbors	90.82%	n_neighbors=7, p=3, algorithm=brute
Linear Discriminant Analysis	88.65%	solver=eigen, n_components=8
Multinomial Naïve Bayes	84.58%	alpha=1
Gaussian Naïve Bayes	79.21%	
AdaBoost	71.68%	n_estimators=10, learning_rate=0.3
Decision Trees	70.34%	max_depth=None, max_features=sqrt

- PCA for Dimensionality Reduction and Data Compression
- MNIST - 99.44% accuracy: 60.000 / 10.000 samples training/test set - 28x28 rectangle
- SVM = Memory Efficiency (Subset of training Points), Versatility (Different Kernels)
- SVM poor generalization when hyperparameters are not tuned

- Random Forests still way below SVM