Automation Logic
Better business through automation

Maartens Lourens

Machine Learning Engineer - Automation Logic

Twitter: @thundercomb
Github: @thundercomb

Automation Logic
Better business through automation

A Pragmatic Introduction to Machine Learning for Engineers

Automation Logic
Better business through automation

Overview:

1. Introduction
2. The Problem
3. The Solution

Automation Logic
Better business through automation

Talk takeaways:

1. Anyone can take ML for a spin
2. Deep math knowledge not required

Automation Logic
Better business through automation

Introduction

Automation Logic
Better business through automation

RNNs and Me

Automation Logic
Better business through automation

Andrej Karpathy:

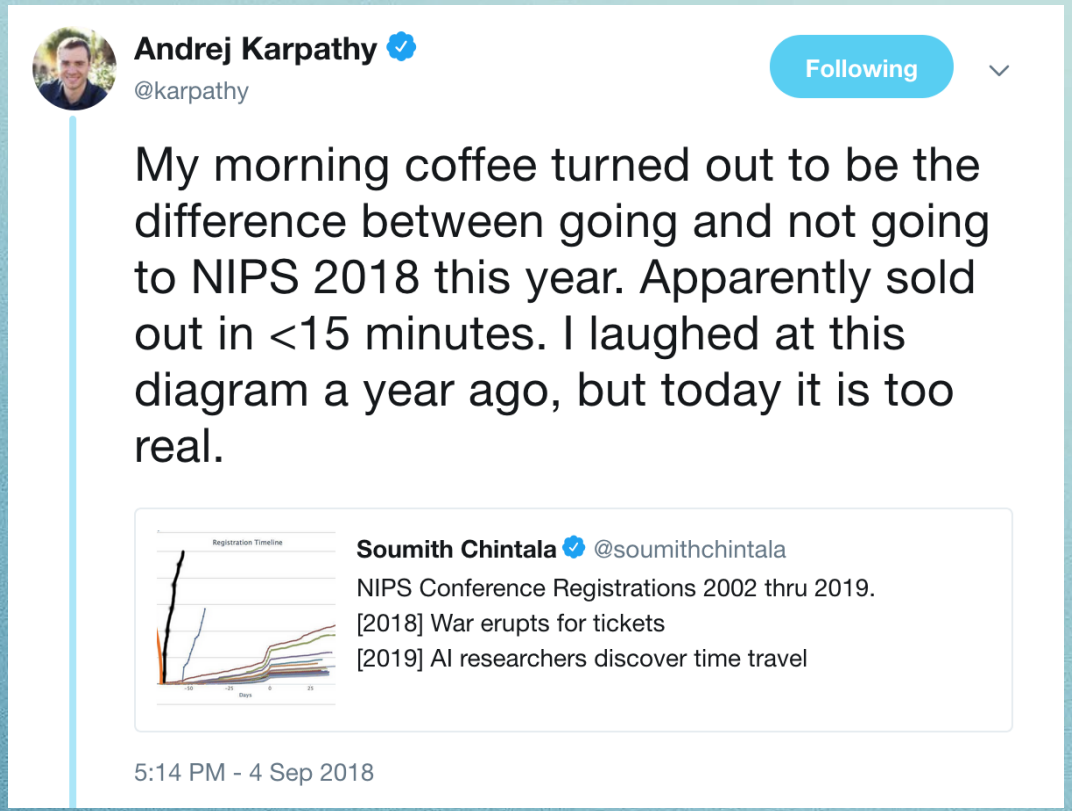
The Unreasonable Effectiveness of Recurrent Neural Networks

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Automation Logic
Better business through automation

AlphaGo & DeepMind

Automation Logic
Better business through automation



Automation Logic
Better business through automation

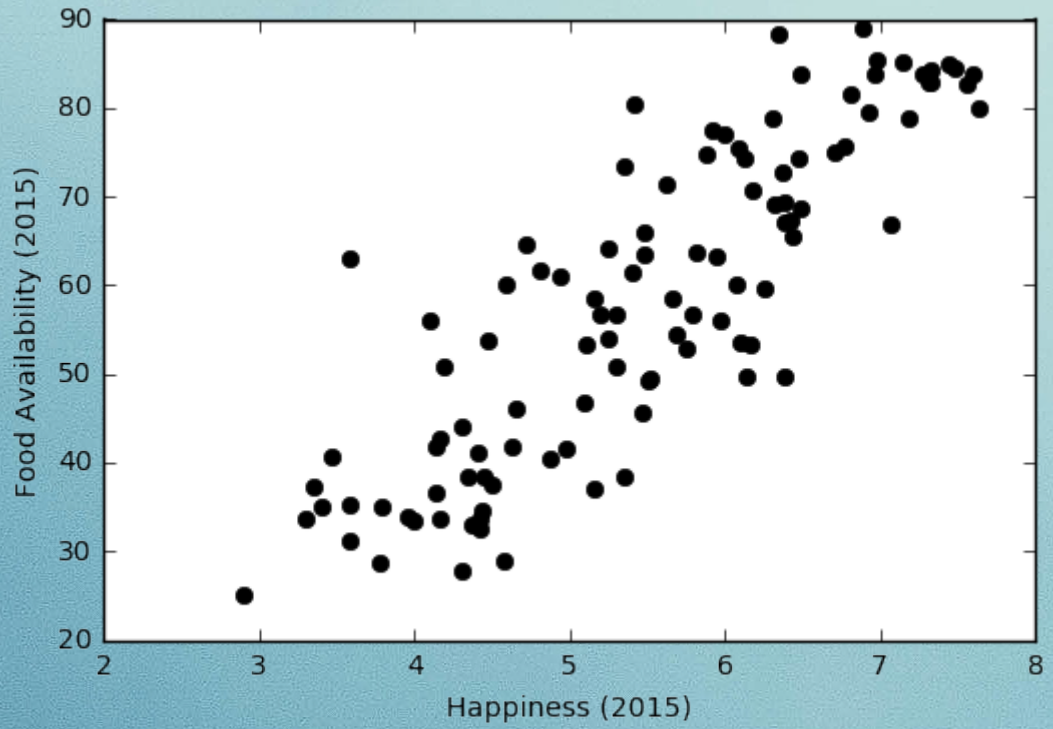
Classical Machine Learning
vs
Deep Learning

Automation Logic
Better business through automation

So… What is Machine Learning?

Automation Logic
Better business through automation

Learning from Data


Food Availability (2015)
Happiness (2015)
90
80
70
60
50
40
30
20
2
3
4
5
6
7
8

Automation Logic
Better business through automation

But is it Machine Learning?

Automation Logic
Better business through automation

Data science produces **insights**
Machine learning produces **predictions**
Artificial intelligence produces **actions**

Credit: David Robinson at http://varianceexplained.org/r/ds-ml-ai/

Automation Logic
Better business through automation

The Problem

System.log

Jul 3 03:42:41 Maartenss-MacBook-Pro parsecd[307]: BUG in libdispatch client: dispatch_mig_server: mach_msg() failed (ipc/send) msg too small - 0x10000008
Jul 3 03:42:41 Maartenss-MacBook-Pro systemstats[50]: assertion failed: 17E199: systemstats + 689866 [D1E75C38-62CE-3D77-9ED3-5F6D38EF0676]: 0x5
Jul 3 03:42:41 Maartenss-MacBook-Pro systemstats[50]: assertion failed: 17E199: systemstats + 914800 [D1E75C38-62CE-3D77-9ED3-5F6D38EF0676]: 0x40
Jul 3 03:42:41 Maartenss-MacBook-Pro com.apple.xpc.launchd[1] (com.apple.WebKit.Networking.1E6606D2-EA49-4EA8-B73B-1A3DE74D35DE[318]): Service exited with abnormal code: 1

Wifi.log

Tue Jul 3 03:42:41.149 <kernel> Creating all peerManager reporters
Tue Jul 3 03:42:41.184 <airportd[153]> _initLocaleManager: Started locale manager
Tue Jul 3 03:42:41.194 <airportd[153]> airportdProcessDLILEvent: en0 attached (down)
Tue Jul 3 03:42:41.219 <kernel> wl0: setAWDL_PEER_TRAFFIC_REGISTRATION: active 0, roam_off: 0, err 0 roam_start_set 0 forced_roam_set 0
Tue Jul 3 03:42:41.269 <kernel> AirPort_Brcm43xx::syncPowerState: WWEN[disabled]

Automation Logic
Better business through automation

Jul 3 05:11:21 --- last message repeated 1 time ---

Automation Logic
Better business through automation

$ grep -l “\-\-\- last message repeated 1 time \-\-\-” /var/log/*.log

/var/log/system.log

Automation Logic
Better business through automation

Jul 3 05:11:21 --- last message repeated 13 times ---

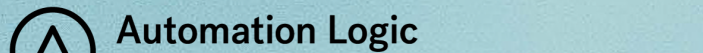
Automation Logic
Better business through automation

Classification

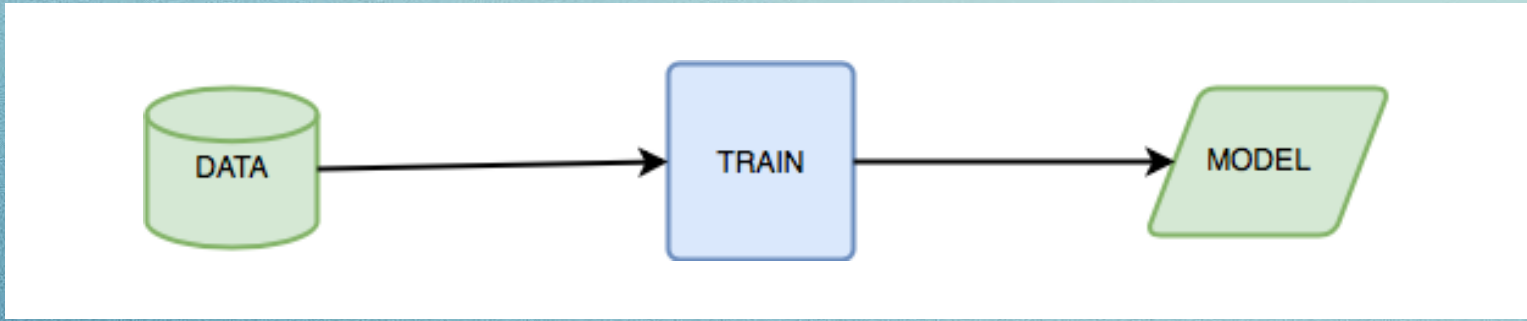
is a

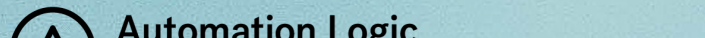
Supervised Machine Learning

problem

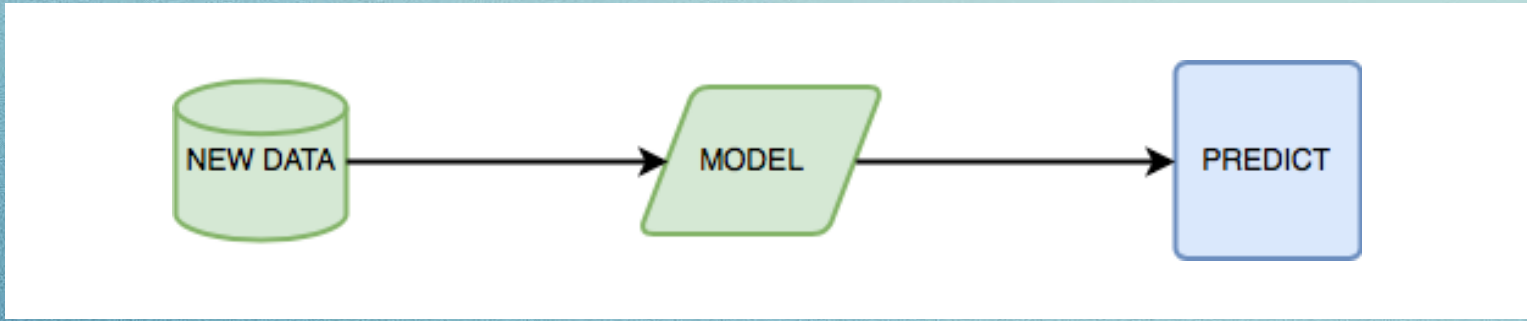

Automation Logic
Better business through automation

Training


DATA
TRAIN
MODEL


Automation Logic
Better business through automation

Predicting


NEW DATA
MODEL
PREDICT

Automation Logic
Better business through automation

The Solution

Automation Logic
Better business through automation

Python & Scikit-Learn

Automation Logic
Better business through automation

```
import os
import glob
import shutil
import numpy as np
import pandas as pd

from sklearn import preprocessing
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer

from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import SGDClassifier
from sklearn import svm, naive_bayes, linear_model, tree, ensemble, neighbors,
                    semi_supervised, neural_network, discriminant_analysis
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
```

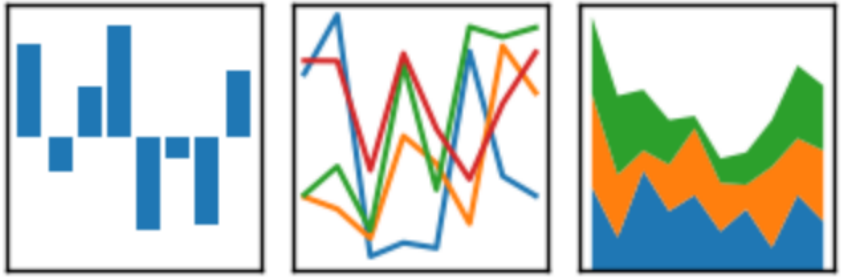
Automation Logic
Better business through automation

```
def copy_data(src_file_path, dst_file_path):
    if not os.path.exists(dst_file_path):
        os.mkdir(dst_file_path)
    for logfile in glob.glob(src_file_path + "/*.log"):
        if os.stat(logfile)[6] > 10000:
            logfile_name = logfile.split('/')[-1]
            shutil.copyfile(logfile, dst_file_path + "/" + logfile_name)
```

Automation Logic
Better business through automation

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

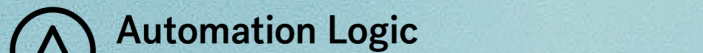


Automation Logic
Better business through automation

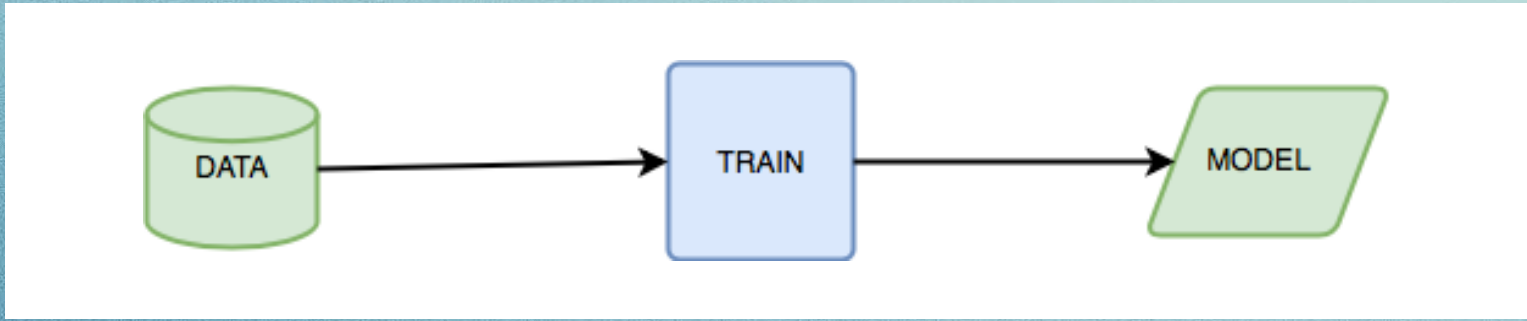
```
def read_data(logfile_path):
    log_collection = pd.DataFrame()
    logs = pd.DataFrame()
    logfiles = glob.glob(logfile_path + "/*.log") # Get list of log files
    for logfile in logfiles:
        logs = pd.read_csv(logfile, sep="\n", header=None, names=['data'])
        logs['type'] = logfile.split('/')[-1]
        # Add log file data and type to log collection
        log_collection = log_collection.append(logs)

    # Remove empty lines
    log_collection = log_collection.dropna()

    return log_collection
```

Automation Logic
Better business through automation

```
source_data_dir = "/var/log"
data_dir = "data"

copy_data(source_data_dir, data_dir)
log_collection = read_data(data_dir)
```


Automation Logic
Better business through automation

Training

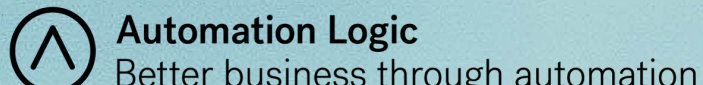

DATA
TRAIN
MODEL

Automation Logic
Better business through automation

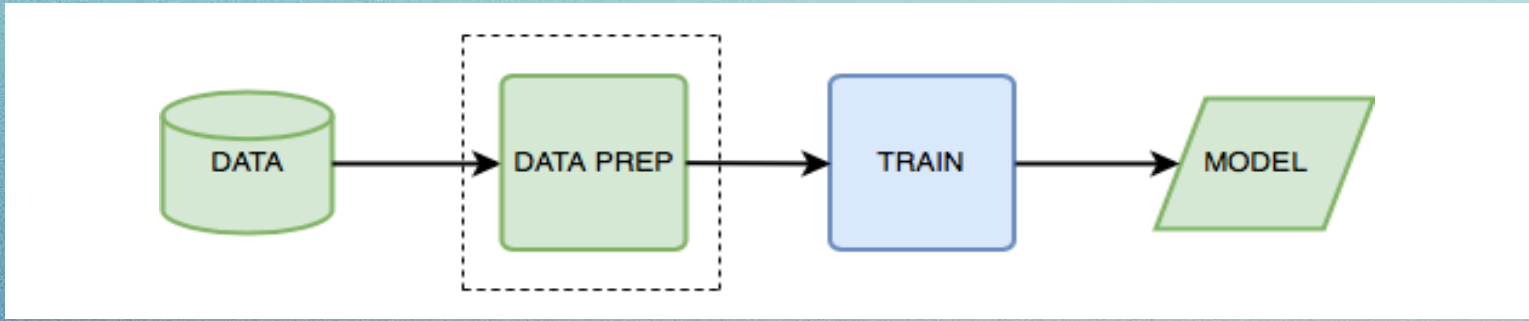
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(log_collection['data'],
                                                    log_collection['type'],
                                                    test_size=0.2,
                                                    random_state=42)
```

Automation Logic
Better business through automation

```
def train(algorithm, X_train, y_train):
    model = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()),
                      (‘clf', algorithm)])
```


Automation Logic
Better business through automation

Training


DATA
DATA PREP
TRAIN
MODEL

Automation Logic
Better business through automation

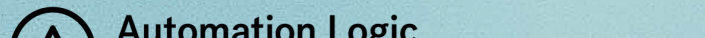
```
def train(algorithm, X_train, y_train):
    model = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()),
                      (‘clf', algorithm)])
    model.fit(X_train, y_train)
    return model
```

Automation Logic
Better business through automation

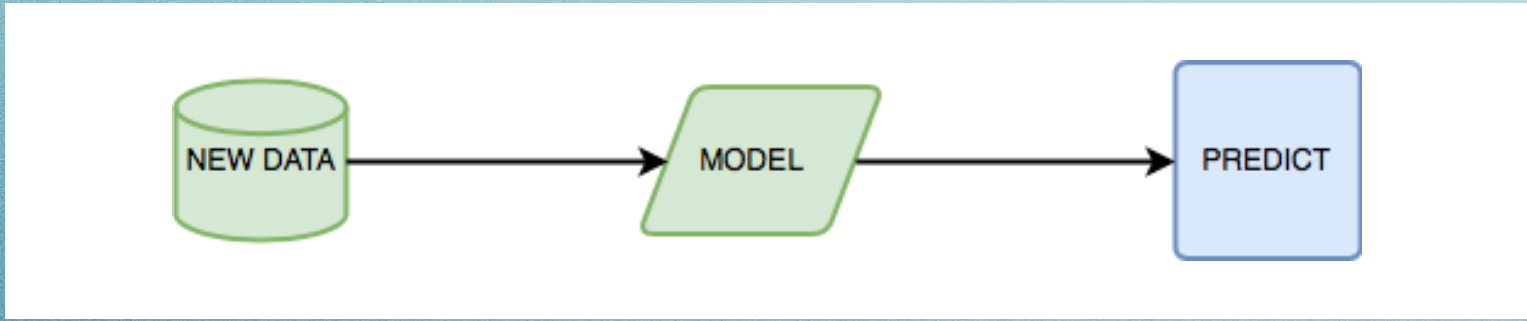
```
algorithms = [
    linear_model.SGDClassifier(loss='hinge', penalty='l2', alpha=1e-3,
random_state=42, max_iter=5, tol=None),
    naive_bayes.MultinomialNB(),
    naive_bayes.BernoulliNB(),
    tree.DecisionTreeClassifier(max_depth=1000),
    tree.ExtraTreeClassifier(),
    ensemble.ExtraTreesClassifier(),
    svm.LinearSVC(),
    neighbors.NearestCentroid(),
    ensemble.RandomForestClassifier(),
    linear_model.RidgeClassifier(),
]
```

Automation Logic
Better business through automation

```
for algorithm in algorithms:
    model = train(algorithm, X_train, y_train)
```


Automation Logic
Better business through automation

Predicting


NEW DATA
MODEL
PREDICT

Automation Logic
Better business through automation

```
for algorithm in algorithms:
    model = train(algorithm, X_train, y_train)
    predictions = model.predict(X_test)
```

Automation Logic
Better business through automation

True & False Positives
True & False Negatives

Automation Logic
Better business through automation

The Boy Who Cried Wolf

Automation Logic
Better business through automation

True Positive:	*Reality:*	A wolf threatens
	Shepherd:	“Wolf!”
True Negative:	*Reality:*	No wolf threatens
	Shepherd:	Quiet.
False Positive:	*Reality:*	No wolf threatens
	Shepherd:	“Wolf!”
False Negative:	*Reality:*	A wolf threatens
	Shepherd:	Quiet.

Credit:
https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative

Automation Logic
Better business through automation

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Recall = TP / (TP + FN)

Precision = TP / (TP + FP)

F1 Score = 2 * ((Precision * Recall) / (Precision + Recall))

Automation Logic
Better business through automation

```
def report(classifier, actual, predictions):
    print("\033[1m" + classifier + "\033[0m\033[50m\n")

    actual = np.array(actual)

    print(confusion_matrix(actual, predictions))
    print
    print(classification_report(actual, predictions))
    print("Accuracy: " + str(round(accuracy_score(actual, predictions),2)))
    print
```

Automation Logic
Better business through automation

```
for algorithm in algorithms:
    model = train(algorithm, X_train, y_train)
    predictions = model.predict(X_test)
    report((str(algorithm).split('(')[0]), y_test, predictions)
```

SGDClassifier

```
[[2536    0    0    0    0    0    0    0    0    0]
 [   0  131    0    0    0    0    0    0    0    0]
 [   0    0 1569  172    0    0    0    0    0    0]
 [   0    0   39 1277    0    0    0    3    0    0]
 [   0    0    0    0 1091    0    0    0    0    0]
 [   0    0    0    0    0    0  116    0    0    0]
 [   0    0    0    0    0    0  948    0    0    1]
 [   0    0    0    0    0    0    0  919    0    1]
 [   0    0    0    0    0    0    0    0  761    1]
 [   0    0    0    0    3    0    1    0    4 1714]]
```

	precision	recall	f1-score	support
corecaptured.log	1.00	1.00	1.00	2536
fsck_apfs.log	1.00	1.00	1.00	131
install.log	0.98	0.90	0.94	1741
system.log	0.88	0.97	0.92	1319
wifi-08-23-2018__12:54:38.121.log	1.00	1.00	1.00	1091
wifi-08-24-2018__09:09:14.458.log	0.00	0.00	0.00	116
wifi-08-24-2018__12:47:32.191.log	0.89	1.00	0.94	949
wifi-08-28-2018__14:27:47.184.log	1.00	1.00	1.00	920
wifi-09-03-2018__12:45:21.309.log	0.99	1.00	1.00	762
wifi.log	1.00	1.00	1.00	1722
avg / total	**0.96**	**0.97**	**0.97**	11287

Accuracy: 0.97

MultinomialNB

```
[[2536    0    0    0    0    0    0    0    0    0]
 [   0  131    0    0    0    0    0    0    0    0]
 [ 161    0 1577    1    0    0    0    0    0    2]
 [ 111    0   90 1111    0    0    0    3    0    4]
 [   0    0    0    0 1088    0    0    2    0    1]
 [   0    0    0    0    0    0  114    1    0    1]
 [   0    0    0    0    9    0  938    1    0    1]
 [   0    0    0    0   23    0    0  895    0    2]
 [   0    0    0    0   14    0    0    1  746    1]
 [   0    0    0    0    2    0    0    0    1 1719]]
```

	precision	recall	f1-score	support
corecaptured.log	0.90	1.00	0.95	2536
fsck_apfs.log	1.00	1.00	1.00	131
install.log	0.95	0.91	0.93	1741
system.log	1.00	0.84	0.91	1319
wifi-08-23-2018__12:54:38.121.log	0.96	1.00	0.98	1091
wifi-08-24-2018__09:09:14.458.log	0.00	0.00	0.00	116
wifi-08-24-2018__12:47:32.191.log	0.89	0.99	0.94	949
wifi-08-28-2018__14:27:47.184.log	0.99	0.97	0.98	920
wifi-09-03-2018__12:45:21.309.log	1.00	0.98	0.99	762
wifi.log	0.99	1.00	1.00	1722
avg / total	**0.94**	**0.95**	**0.95**	11287

Accuracy: 0.95

BernoulliNB

```
[[2536    0    0    0    0    0    0    0    0    0]
 [   0  131    0    0    0    0    0    0    0    0]
 [ 293    1 1429   18    0    0    0    0    0    0]
 [ 227    0   26 1066    0    0    0    0    0    0]
 [   0    0    0    0 1091    0    0    0    0    0]
 [   0    0    0    0    0    3  113    0    0    0]
 [   0    0    0    0    0   10  938    0    0    1]
 [   0    0    0    0    0    0    0  918    0    2]
 [   0    0    0    0    0    0    0    0  761    1]
 [   0    0    0    0    2    0    0    0    1 1719]]
```

	precision	recall	f1-score	support
corecaptured.log	0.83	1.00	0.91	2536
fsck_apfs.log	0.99	1.00	1.00	131
install.log	0.98	0.82	0.89	1741
system.log	0.98	0.81	0.89	1319
wifi-08-23-2018__12:54:38.121.log	1.00	1.00	1.00	1091
wifi-08-24-2018__09:09:14.458.log	0.23	0.03	0.05	116
wifi-08-24-2018__12:47:32.191.log	0.89	0.99	0.94	949
wifi-08-28-2018__14:27:47.184.log	1.00	1.00	1.00	920
wifi-09-03-2018__12:45:21.309.log	1.00	1.00	1.00	762
wifi.log	1.00	1.00	1.00	1722
avg / total	**0.94**	**0.94**	**0.93**	11287

Accuracy: 0.94

DecisionTreeClassifier

```
[[2534    0    0    2    0    0    0    0    0    0]
 [   0  131    0    0    0    0    0    0    0    0]
 [   0    0 1737    4    0    0    0    0    0    0]
 [   1    0  127 1191    0    0    0    0    0    0]
 [   0    0    0    0 1090    0    0    0    0    1]
 [   0    0    0    0    0   43   73    0    0    0]
 [   0    0    0    0    0  112  836    0    0    1]
 [   0    0    0    0    0    0    0  919    0    1]
 [   0    0    0    0    0    0    0    0  761    1]
 [   0    0    0    0    3    0    1    4    2 1712]]
```

	precision	recall	f1-score	support
corecaptured.log	1.00	1.00	1.00	2536
fsck_apfs.log	1.00	1.00	1.00	131
install.log	0.93	1.00	0.96	1741
system.log	0.99	0.90	0.95	1319
wifi-08-23-2018__12:54:38.121.log	1.00	1.00	1.00	1091
wifi-08-24-2018__09:09:14.458.log	0.28	0.37	0.32	116
wifi-08-24-2018__12:47:32.191.log	0.92	0.88	0.90	949
wifi-08-28-2018__14:27:47.184.log	1.00	1.00	1.00	920
wifi-09-03-2018__12:45:21.309.log	1.00	1.00	1.00	762
wifi.log	1.00	0.99	1.00	1722
avg / total	**0.97**	**0.97**	**0.97**	11287

Accuracy: 0.97

ExtraTreeClassifier

```
[[2531    0    0    0    2    0    0    0    0    3]
 [   0  129    0    0    1    0    0    0    0    1]
 [   1    0 1727   10    0    0    0    1    0    2]
 [   2    0  139 1177    0    0    0    0    0    1]
 [   2    0    0    0 1064    0   10    1    3   11]
 [   0    0    0    0    0   43   72    1    0    0]
 [   1    0    0    0   15  164  752    4    7    6]
 [   1    0    0    0    0    0    6  860   48    5]
 [   0    1    1    0    5    0    9    4  733    9]
 [  11    2    1    0   12    2    5    8    3 1678]]
```

	precision	recall	f1-score	support
corecaptured.log	0.99	1.00	1.00	2536
fsck_apfs.log	0.98	0.98	0.98	131
install.log	0.92	0.99	0.96	1741
system.log	0.99	0.89	0.94	1319
wifi-08-23-2018__12:54:38.121.log	0.97	0.98	0.97	1091
wifi-08-24-2018__09:09:14.458.log	0.21	0.37	0.26	116
wifi-08-24-2018__12:47:32.191.log	0.88	0.79	0.83	949
wifi-08-28-2018__14:27:47.184.log	0.98	0.93	0.96	920
wifi-09-03-2018__12:45:21.309.log	0.92	0.96	0.94	762
wifi.log	0.98	0.97	0.98	1722
avg / total	**0.95**	**0.95**	**0.95**	11287

Accuracy: 0.95

ExtraTreesClassifier

```
[[2536    0    0    0    0    0    0    0    0    0]
 [   0  131    0    0    0    0    0    0    0    0]
 [   0    0 1739    2    0    0    0    0    0    0]
 [   0    0  126 1192    0    0    0    0    0    1]
 [   0    0    0    0 1090    0    0    0    0    1]
 [   0    0    0    0    0   39   77    0    0    0]
 [   0    0    0    0    0  108  840    0    0    1]
 [   0    0    0    0    0    0    0  919    0    1]
 [   0    0    0    0    0    0    0    0  760    2]
 [   0    0    0    0    3    0    0    4    2 1713]]
```

	precision	recall	f1-score	support
corecaptured.log	1.00	1.00	1.00	2536
fsck_apfs.log	1.00	1.00	1.00	131
install.log	0.93	1.00	0.96	1741
system.log	1.00	0.90	0.95	1319
wifi-08-23-2018__12:54:38.121.log	1.00	1.00	1.00	1091
wifi-08-24-2018__09:09:14.458.log	0.27	0.34	0.30	116
wifi-08-24-2018__12:47:32.191.log	0.92	0.89	0.90	949
wifi-08-28-2018__14:27:47.184.log	1.00	1.00	1.00	920
wifi-09-03-2018__12:45:21.309.log	1.00	1.00	1.00	762
wifi.log	1.00	0.99	1.00	1722
avg / total	**0.97**	**0.97**	**0.97**	11287

Accuracy: 0.97

LinearSVC

```
[[2536    0    0    0    0    0    0    0    0    0]
 [   0  131    0    0    0    0    0    0    0    0]
 [   0    0 1737    3    0    0    0    0    0    1]
 [   0    0  125 1194    0    0    0    0    0    0]
 [   0    0    0    0 1090    0    0    0    0    1]
 [   0    0    0    0    0   33   82    0    0    1]
 [   0    0    0    0    0   87  861    0    0    1]
 [   0    0    0    0    0    0    0  919    0    1]
 [   0    0    0    0    0    0    0    0  761    1]
 [   0    0    0    0    3    0    0    0    2 1717]]
```

	precision	recall	f1-score	support
corecaptured.log	1.00	1.00	1.00	2536
fsck_apfs.log	1.00	1.00	1.00	131
install.log	0.93	1.00	0.96	1741
system.log	1.00	0.91	0.95	1319
wifi-08-23-2018__12:54:38.121.log	1.00	1.00	1.00	1091
wifi-08-24-2018__09:09:14.458.log	0.28	0.28	0.28	116
wifi-08-24-2018__12:47:32.191.log	0.91	0.91	0.91	949
wifi-08-28-2018__14:27:47.184.log	1.00	1.00	1.00	920
wifi-09-03-2018__12:45:21.309.log	1.00	1.00	1.00	762
wifi.log	1.00	1.00	1.00	1722
avg / total	**0.97**	**0.97**	**0.97**	11287

Accuracy: 0.97

NearestCentroid

```
[[2004    0   16  312   13    3   34   30   15  109]
 [   0  131    0    0    0    0    0    0    0    0]
 [   0    0 1387  354    0    0    0    0    0    0]
 [   0    0  100 1216    0    0    0    3    0    0]
 [   0    0    0    0 1091    0    0    0    0    0]
 [   0    0    0    0    0   75   41    0    0    0]
 [   0    0    0    0    0  105  844    0    0    0]
 [   0    0    0    0    0    0    0  920    0    0]
 [   0    0    0    2    0    0    0    0  760    0]
 [   0    0    0    9   12    2   16   14   10 1659]]
```

	precision	recall	f1-score	support
corecaptured.log	1.00	0.79	0.88	2536
fsck_apfs.log	1.00	1.00	1.00	131
install.log	0.92	0.80	0.86	1741
system.log	0.64	0.92	0.76	1319
wifi-08-23-2018__12:54:38.121.log	0.98	1.00	0.99	1091
wifi-08-24-2018__09:09:14.458.log	0.41	0.65	0.50	116
wifi-08-24-2018__12:47:32.191.log	0.90	0.89	0.90	949
wifi-08-28-2018__14:27:47.184.log	0.95	1.00	0.98	920
wifi-09-03-2018__12:45:21.309.log	0.97	1.00	0.98	762
wifi.log	0.94	0.96	0.95	1722
avg / total	**0.91**	**0.89**	**0.90**	11287

Accuracy: 0.89

RandomForestClassifier

```
[[2536    0    0    0    0    0    0    0    0    0]
 [   0  131    0    0    0    0    0    0    0    0]
 [   0    0 1738    3    0    0    0    0    0    0]
 [   0    0  125 1194    0    0    0    0    0    0]
 [   0    0    0    0 1090    0    0    0    0    1]
 [   0    0    0    0    0   30   86    0    0    0]
 [   0    0    0    0    0   95  854    0    0    0]
 [   0    0    0    0    0    0    0  918    0    2]
 [   0    0    0    0    0    0    0    0  761    1]
 [   0    0    0    0    4    0    0    3    2 1713]]
```

	precision	recall	f1-score	support
corecaptured.log	1.00	1.00	1.00	2536
fsck_apfs.log	1.00	1.00	1.00	131
install.log	0.93	1.00	0.96	1741
system.log	1.00	0.91	0.95	1319
wifi-08-23-2018__12:54:38.121.log	1.00	1.00	1.00	1091
wifi-08-24-2018__09:09:14.458.log	0.24	0.26	0.25	116
wifi-08-24-2018__12:47:32.191.log	0.91	0.90	0.90	949
wifi-08-28-2018__14:27:47.184.log	1.00	1.00	1.00	920
wifi-09-03-2018__12:45:21.309.log	1.00	1.00	1.00	762
wifi.log	1.00	0.99	1.00	1722
avg / total	**0.97**	**0.97**	**0.97**	11287

Accuracy: 0.97

RidgeClassifier

```
[[2536    0    0    0    0    0    0    0    0    0]
 [   0  131    0    0    0    0    0    0    0    0]
 [   0    0 1739    1    0    0    0    0    0    1]
 [   0    0  126 1190    0    0    0    3    0    0]
 [   0    0    0    0 1091    0    0    0    0    0]
 [   0    0    0    0    1    9  105    0    0    1]
 [   0    0    0    0    1   24  923    0    0    1]
 [   0    0    0    0    1    0    0  918    0    1]
 [   0    0    0    0    0    0    0    0  761    1]
 [   0    0    0    0    2    0    0    0    1 1719]]
```

	precision	recall	f1-score	support
corecaptured.log	1.00	1.00	1.00	2536
fsck_apfs.log	1.00	1.00	1.00	131
install.log	0.93	1.00	0.96	1741
system.log	1.00	0.90	0.95	1319
wifi-08-23-2018__12:54:38.121.log	1.00	1.00	1.00	1091
wifi-08-24-2018__09:09:14.458.log	0.27	0.08	0.12	116
wifi-08-24-2018__12:47:32.191.log	0.90	0.97	0.93	949
wifi-08-28-2018__14:27:47.184.log	1.00	1.00	1.00	920
wifi-09-03-2018__12:45:21.309.log	1.00	1.00	1.00	762
wifi.log	1.00	1.00	1.00	1722
avg / total	**0.97**	**0.98**	**0.97**	11287

Accuracy: 0.98

WINNER!

Automation Logic
Better business through automation

What have I learned?

1. Get to know your **data**
2. Understand the **algorithm**

Automation Logic
Better business through automation

Code:

https://github.com/automationlogic/log-analysis

Automation Logic
Better business through automation

Thank You!

Questions?

Twitter: @thundercomb

Code: https://github.com/automationlogic/log-analysis