In [11]:
```
import pandas as pd
from sklearn.neighbors import KNeighborsClassifier
synth_train = pd.read_csv("synth_train.csv")
synth_valid = pd.read_csv("synth_valid.csv")
synth_test = pd.read_csv("synth_test.csv")
```

Question 1a

In [12]:
```
training_X = synth_train.drop(['y'], axis=1)
training_y = synth_train.y
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(training_X, training_y)
```

Out[12]:
```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkows
ki',
           metric_params=None, n_jobs=1, n_neighbors=5, p=2,
           weights='uniform')
```

Question 1b

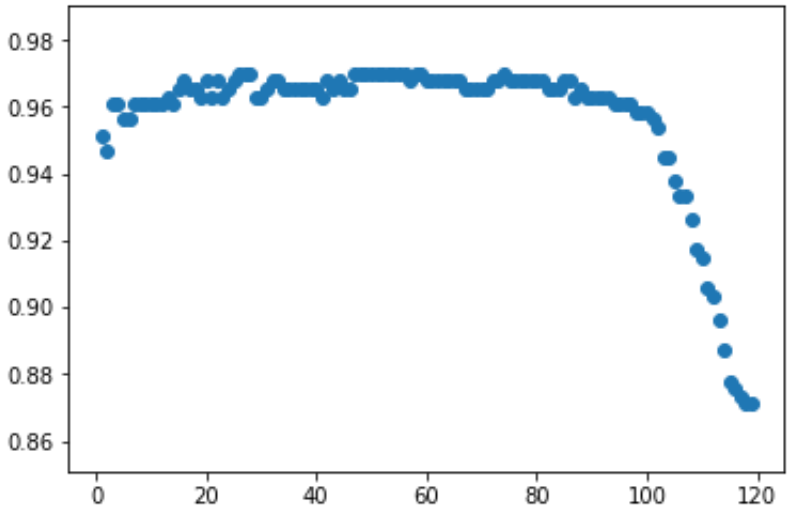
In [10]:
```
import matplotlib.pyplot as plot

accuracy = []
k_values = range(1, 120)

validation_X = synth_valid.drop(['y'], axis=1)
validation_y = synth_valid.y

for n in k_values:
    knn = KNeighborsClassifier(n_neighbors=n)
    knn.fit(training_X, training_y)
    accuracy.append(knn.score(validation_X, validation_y))

plot.scatter(x=k_values, y=accuracy)
plot.show()
max_accuracy = max(accuracy)
best_k = k_values[accuracy.index(max_accuracy)]
print "Best k: %i"%(best_k)
```




```
Best k: 26
```

Best value of k is 26 (may be tied with others. If tied, it is the lowest value of k among the tied.)

Question 1c

In [13]:
```
test_X = synth_test.drop(['y'], axis=1)
test_y = synth_test.y

training_X = synth_train.drop(['y'], axis=1)
training_y = synth_train.y

knn = KNeighborsClassifier(n_neighbors=best_k)
knn.fit(training_X, training_y)
knn.score(test_X, test_y)
```

Out[13]:
```
0.978494623655914
```

97.8495%