# Lecture 6: Playing with Spam Dataset

อ.ปรัชญ์ ปิยะวงศ์วิศาล

# Today’s Topics

Lab Exercise

- Exploring spam dataset with Pandas

# Importing data with Pandas

```
import pandas as pd
messages = pd.read_csv('./SMSSpamCollection',
                                    sep = '\t',
                                    names = ["label","message"])
```

*messages* <= Read from tab-separated data file

data has 2 columns: **label** and **message**

# Quick glance into the data

```
print(messages)                    # print all rows

print(message.head())              # print the first 5 rows

message.describe()                 # quick statistics
```

# Grouping data by column values

**messages.groupby('label').describe()**

Show quick statistics for “ham” and “spam” classes

# Create a new column

```
messages['length'] = messages['message'].map(lambda text: len(text))
print(messages.head())
```

Create a new column named **length** where **length = len(message)**

# Closer look at length

```
# what's the length of the longest message?
messages.length.describe()
print(messages.message[messages.length > 900])
messages.length.plot(bins=20, kind='hist')                    # look at the histogram


# spam vs ham length
message.groupby('label').length.describe()
messages.hist(column='length', by='label', bins=50)
```

# Data Processing – Split messages into words

Say, we want to split this message:

**“Free entry in 2 a wkly comp to win FA Cup”**

into a list of words (tokens):

**[Free, entry, in, 2, a, wkly, comp, to, win, FA, Cup]**

How do we do it the easy way?

# Solution - Split messages into words with TextBlob

1. Create split function:

```
def split_words(message):
    return TextBlob(message).words
```

2. Apply **split_words()** to messages:

```
messages.message.head().apply(split_into_tokens)
```

# Next week

More on TextBlob, NLP...