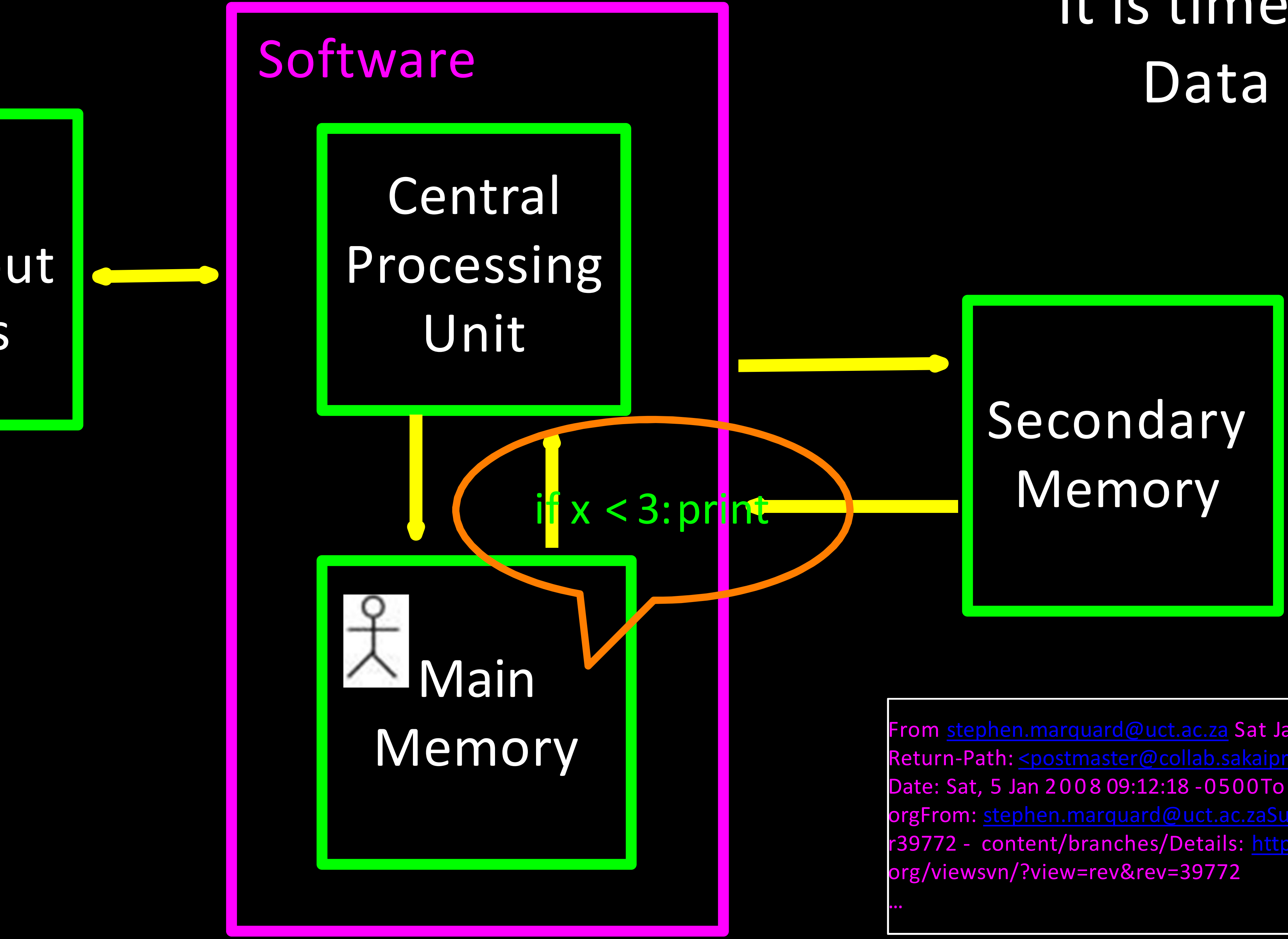
Reading Files

Chapter 7

It is time to go

Data to me



```
From stephen.marquard@uct.ac.za Sat Jan 5 09:14:16 2
Return-Path: <postmaster@collab.sakaiproject.org>
Date: Sat, 5 Jan 2008 09:12:18 -0500To: source@colla
orgFrom: stephen.marquard@uct.ac.zaSubject: [sakai] sv
r39772 - content/branches/Details: http://source.saka
org/viewsvn/?view=rev&rev=39772
…
```

File Processing

le can be thought of as a sequence of lines

```
phen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
ath: <postmaster@collab.sakaiproject.org>
t, 5 Jan 2008 09:12:18 -0500
ce@collab.sakaiproject.org
ephen.marquard@uct.ac.za
 [sakai] svn commit: r39772 - content/branches/

 http://source.sakaiproject.org/viewsvn/?view=rev&rev=3
```

http://www.py4inf.com/code/mbox-short.txt

Opening a File

e we can read the contents of the file, we must tell Py
file we are going to work with and what we will be do
e

done with the open() function

) returns a "file handle" - a variable used to perform o
e file

r to "File -> Open" in a Word Processor

Using open()

e = open(filename, mode)

fhand = open('mbo

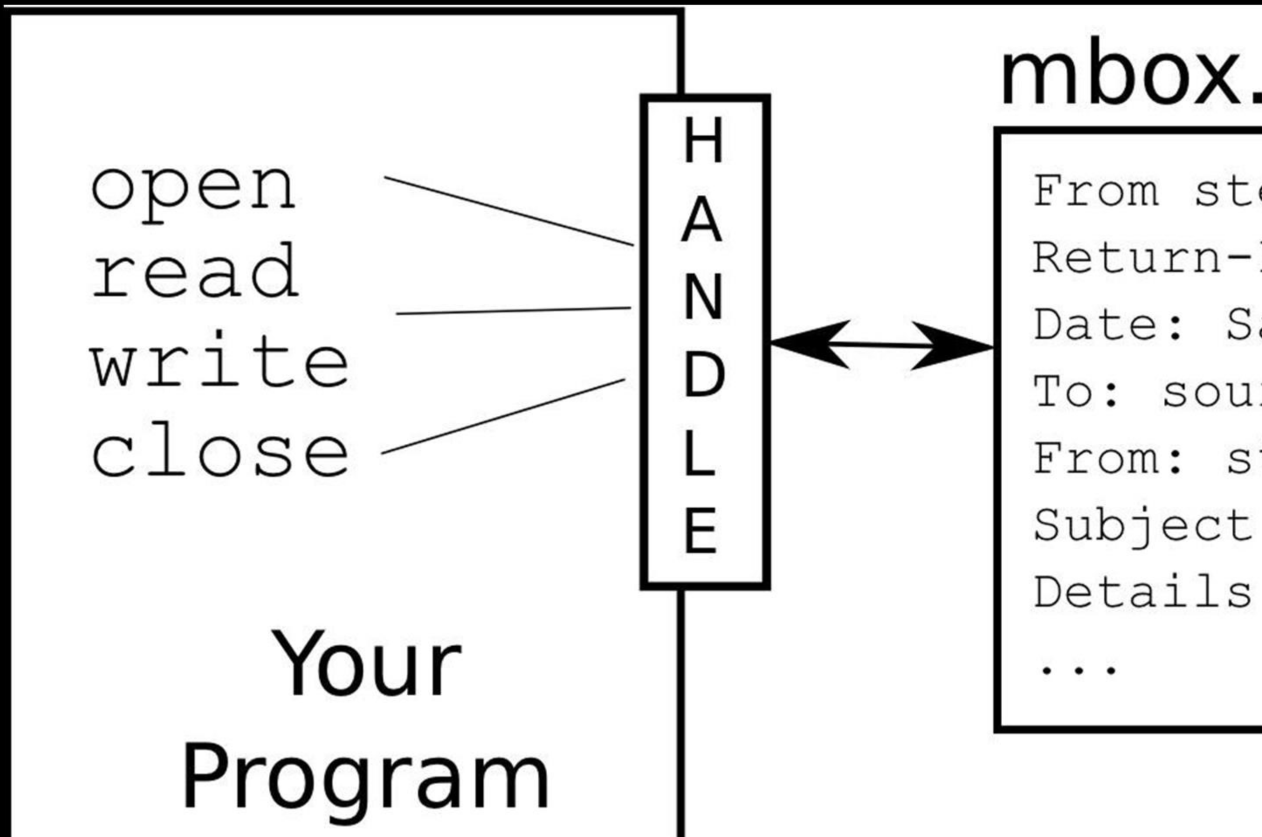
rns a handle use to manipulate the file

ame is a string

e is optional and should be 'r' if we are planning to re
nd 'w' if we are going to write to the file

What is a Handle?

```
= open('mbox.txt')
fhand
 'mbox.txt', mode 'r' at 0x1005088b0>
```



When Files are Missing

```
and = open('stuff.txt')
ack (most recent call last):  File
n>", line 1, in <module>IOError: [Err
h file or directory: 'stuff.txt'
```

The newline Character

se a special character
the “newline” to
te when a line ends

present it as \n in
s

ne is still one character
two

```
>>> stuff = 'Hell
>>> stuff
'Hello\nWorld!'
>>> print stuff
Hello
World!
>>> stuff = 'X\nY
>>> print stuff
X
Y
>>> len(stuff)
3
```

File Processing

file can be thought of as a sequence of lines

```
ephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
-Path: <postmaster@collab.sakaiproject.org>
at, 5 Jan 2008 09:12:18 -0500
rce@collab.sakaiproject.org
tephen.marquard@uct.ac.za
: [sakai] svn commit: r39772 - content/branches/

: http://source.sakaiproject.org/viewsvn/?view=rev&rev=
```

File Processing

file has newlines at the end of each line

```
ephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008\n
Path: <postmaster@collab.sakaiproject.org>\n
at, 5 Jan 2008 09:12:18 -0500\n
rce@collab.sakaiproject.org\n
tephen.marquard@uct.ac.za\n
: [sakai] svn commit: r39772 - content/branches/\n

: http://source.sakaiproject.org/viewsvn/?view=rev&rev=
```

File Handle as a Sequence

ndle open for read can be
as a sequence of strings
ach line in the file is a string
equence

use the for statement to
hrough a sequence

ber - a sequence is an
set

```
xfile = open('mbo
for cheese in xfi
    print cheese
```

Counting Lines in a File

a file read-only

for loop to read each line

the lines and print out
mber of lines

```
fhand = open('mbox.
count = 0
for line in fhand:
    count = count +
print 'Line Count:'
```

```
$ python open.py
Line Count: 132045
```

Reading the *Whole* File

read the whole file
es and all) into a
tring

```
>>> fhand = open('mbox-sh
>>> inp = fhand.read()
>>> print len(inp)
94626
>>> print inp[:20]
From stephen.marquar
```

Searching Through a File

put an if statement in
oop to only print lines
et some criteria

```
fhand = open('mbox-short.t
for line in fhand:
    if line.startswith('Fr
        print line
```

OOPS!

e all these blank
doing here?

```
From: stephen.marquard@u

From: louis@media.berkel

From: zqian@umich.edu

From: rjlowe@iupui.edu
...
```

OOPS!

all these blank
g here?

from the file has a
at the end

statement adds a
to each line

```
From: stephen.marquard@u
\n
From: louis@media.berkel
\n
From: zqian@umich.edu\n
\n
From: rjlowe@iupui.edu\n
\n
...
```

rching Through a File (fix

rip the whitespace
right-hand side of the
g rstrip() from the
ry

ne is considered
nce” and is stripped

```
fhand = open('mbox-short.t
for linein  fhand:
    line = line.rstrip()
    if line.startswith('Fr
        print line
```

```
From: stephen.marquard@u
From: louis@media.berkeley
From: zqian@umich.edu
From: rjlowe@iupui.edu
....
```

Skipping with continue

onveniently
e by using the
statement

```
fhand = open('mbox-short.txt'
for line in fhand:
    line = line.rstrip()
    if not line.startswith('F
        continue
    print line
```

Using in to select lines

n look for a string
here in a line as our
ion criteria

```
fhand = open('mbox-short.
for line in fhand:
    line = line.rstrip()
    if not '@uct.ac.za' i
        continue
    print line
```

```
hen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
ication-Warning: set sender to stephen.marquard@uct.ac.
phen.marquard@uct.ac.za
tephen.marquard@uct.ac.za
d.horwitz@uct.ac.za Fri Jan  4 07:02:32 2008
ication-Warning: set sender to david.horwitz@uct.ac.za
```

Promp
File N

```
nput('Enter the file name:  ')
(fname)

fhand:
startswith('Subject:') :
t = count + 1
 were', count, 'subject lines in', fname
```

```
Enter the file name:  mbox.txt
There were 1797 subject lines in

Enter the file name: mbox-short
There were 27 subject lines in m
txt
```

File

es

```
fname = raw_input('Enter the file name:
try:
    fhand = open(fname)
except:
    print 'File cannot be opened:', fname
    exit()

count = 0
for line in fhand:
    if line.startswith('Subject:') :
        count = count + 1
print 'There were', count, 'subject lines
```

ame: mbox.txt
97 subject lines in mbox.txt

ame: na na boo boo
opened: na na boo boo

Summary

- lary storage
- ng a file - file handle
- ucture - newline character
- g a file line by line with a
o

- Searching for lines
- Reading file names
- Dealing with bad fil