
# DM550/DM857
# Introduction to Programming

Peter Schneider-Kamp

petersk@imada.sdu.dk

http://imada.sdu.dk/~petersk/DM550/

http://imada.sdu.dk/~petersk/DM857/

# SELECTING DATA STRUCTURES

# Reading and Cleaning Words

1. read file given as argument
2. break lines into words
3. strip whitespace & punctuation
4. convert to lower-case letters

- import module sys for command line arguments sys.argv
- Example: `import sys; print(sys.argv)`

- import module string for punctuation
- Example: `import string; print(string.punctuation)`

- use `translate(dict)` to remove punctuation
- Example: `"Hello World!".translate({ord("o"):"",ord("l"):""})`

# Word Frequency in E-Books

1. use program on Project Gutenberg e-book
2. skip over beginning & end of ebook (marked "***")
3. count total number of words
4. count number of times each word is used
5. print 20 most frequently used words

- use Boolean flag to indicate when to start
- use list to gather all words (and count total number)
- use dictionary to count number of times each word is used
- use tuple comparison to sort words

# Markov Analysis

1. generate more meaningful random texts

- word order in texts is not random
- markov analysis maps a finite number of words (prefix) to all possible following words (suffix)

- how to represent the prefixes?

- how to represent the collection of possible suffixes?

- how to represent the mapping from prefixes to suffixes?

# Data Structures

- for mapping, we clearly use a dictionary

- for prefixes, we need to be able to "shift" them (list?)
- we also need to use them as dictionary keys
- thus, we use tuples to present prefixes (+ slicing and "+")

- for suffixes, we need to add elements (list? dictionary?)
- we also need to efficiently generate random word (list?)
- tradeoff space vs time
  - dictionary uses less space and easy to add
  - list uses less time for generating a word
  - can change representation before generation

# Debugging Hard Bugs

- bugs can be hard to find

- four popular strategies
  1. reading: re-read your code, check that it is right!
  2. running: make changes, experiment with outcome
  3. ruminating: take time to think it over (and over)
  4. retreating: revert to a known-to-be-good version

- often combination of these strategies needed
- always good to view debugging as scientific experiment

# Optional Parameters

- have seen functions that take variable length argument list

- also possible to make some parameters optional
- in this case, default value has to be supplied by programmer
- Example:

```
def print_most_common(hist, num = 10):
    t = most_common(hist)
    print "The most common", num, "words are:"
    for n, word in t[:num]:
        print word, "\t", n
print_most_common(freq, 20)
```

# Dictionary Subtraction

1. find all words that do NOT occur in other word list

- to this end, subtract dictionaries from each other
- **Idea:** new dictionary containing with keys only in first dict
- Implementation:

```
def subtract(d1, d2):
    d = {}
    for key in d1:
        if key not in d2:
            d[key] = None
    return d
```

# Random Number Generation

- to work with random numbers, import module random
- Example: import random

- function random() returns random float from 0.0 to < 1.0
- Example: for i in range(10): print(random.random())

- function randint(a, b) returns random integer in range(a,b+1)
- Example: for i in range(10): print(random.randint(1,10))

- function choice(seq) returns random element of a sequence
- Example: random.choice("Slartibartfast")
  random.choice([23, 42, -3.0])

# Random Words

1. choose random word from histogram according to frequency

- how to ensure random choice w.r.t. frequency?
- **Idea 1:** create list with n copies of word with frequency n
- Implementation:

```
def random_word(h):
    t = []
    for word, n in h.items():
        t.extend([word] * n)
    return random.choice(t)
```

- works, but very inefficient!

# Random Words

- **Idea 2:** use list with cumulative sum of frequencies
- Implementation:

```
def random_word(h):
    words = h.keys();  sum = 0;  cum = []
    for word in words:  sum += h[word];  cum.append(sum)
    num = random.randint(1, cum[-1]); low = 0; high = len(cum)-1
    while low < high:
        mid = (low+high) // 2
        if num <= cum[mid]:  high = mid
        elif num > cum[mid]:  low = mid+1
    return words[low]
```

# FILE HANDLING

# Persistence

- persistent = keeping (some) data stored during runs
- transient = beginning from input data each time over

- most programs so far have been transient

- examples of persistent programs:
  - operating systems
  - web servers
  - most app(lication)s on recent Android, iOS, and Mac OS X

- text files are easiest way to save some program state
- alternatively, program states can be saved in databases

# Writing to a File

- we know how to read a file using open(name)
- we can specify read/write mode using open(name, mode)
- Example: f1 = open("anna_karenina.txt", "r")
  f2 = open("myfile.txt", "w")

- use method write(str) of file object to append string to file
- Example: f2.write("This is my first line!\n")
  f2.write("This is my second line!\n")
- each invocation of write(str) will append, not overwrite!

- when you are finished with a file, please close() it
- Example: f1.close()
  f2.close()

# Format Operator

- values need to be converted to a string for use in write(str)
- for single value, the str(object) function can be used
- Example: f.write(str(42))

- alternatively, use *format operator* "%"
- Example: f.write("%d" % 42)
  f.write("The answer is %d, my friend!" % 42)
- first argument *format string*, second argument value
- format sequence %d for integer, %g for float, %s for string

- for multiple values, use tuple as value
- Example: f.write("The %s is %g!" % ("answer", 42.0))

# Directories

- file are organized in *directories*
- every program has a *current directory*
- the current directory is used by default, e.g. for open(name)
- get current directory by importing getcwd() from os module
- Example:
```
import os
print(os.getcwd())
```
- change current working directory by using chdir(path)
- Example:
```
os.chdir("..")
print(os.getcwd())
```
- list contents of a given directory by using os.listdir(path)
- Example:
```
print(os.listdir("dm502"))
```

# Filenames and Paths

- path = directory & file name
- *relative paths* start from current directory
- Example:

```
path1 = "dm536/tools/anna_karenina.txt"
```

- *absolute paths* are independent from current directory
- Example:

```
path2 = "/Users/petersk/sdu/dm536/tools/anna_karenina.py"
```

- can be obtained from relative path using os.path.abspath(path)
- Example:

```
path3 = os.path.abspath(path1)
```

# Operations on Paths

- check whether a directory or file exists using os.path.exists
- Example: os.path.exists(path1) == True
  os.path.exists("no_name") == False

- check whether a path is a directory using os.path.isdir
- Example: os.path.isdir(path1) == False
  os.path.isdir("..") == True

- check whether a path is a file using os.path.isfile
- Example: os.path.isfile(path1) == True
  os.path.isfile("..") == False

# CLASSES & OBJECTS

# User-Defined Types

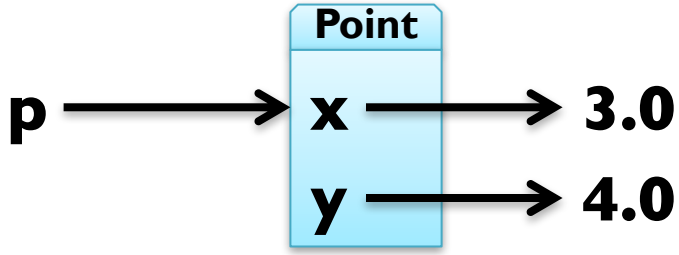
- we want to represent points (x,y) in 2-dimensional space
- which data structure to use?
  - use two variables x and y
  - store coordinates in a list or tuple of length 2
  - create user-defined type
- we can use Python's classes to implement new types
- Example:

```
class Point(object):
    """represents a point in 2-dimensional space"""
print(Point)     # class
p = Point()      # create new instance of class Point
print(p)         # instance
```

# Attributes

- using *dot notation*, you can assign values to instance variables
- Example: p.x = 3.0

  p.y = 4.0



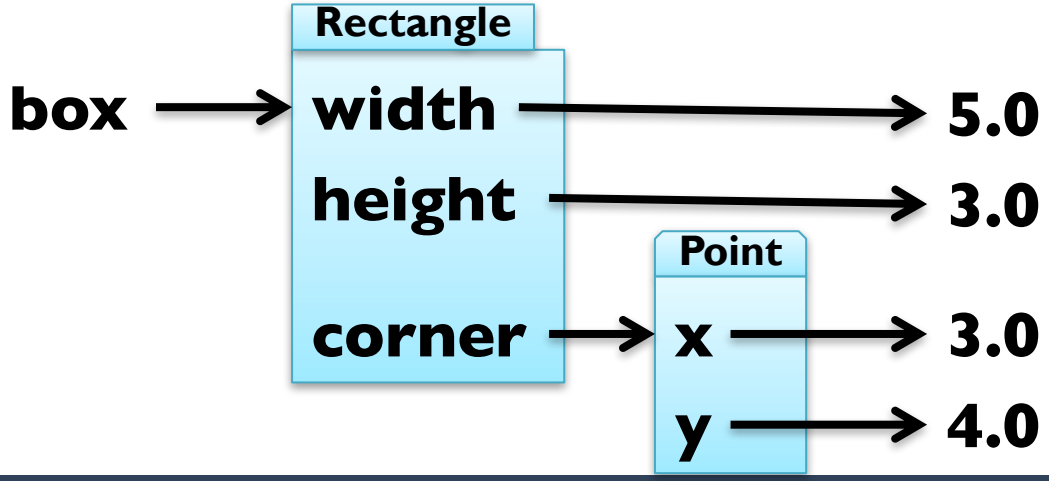

- instance variables are called *attributes*
- attributes can be assigned to and read like any variable
- Example:

```
print("(%g, %g)" % (p.x, p.y))
distance = math.sqrt(p.x**2 + p.y**2)
print (distance, "units from the origin")
```

# Representing a Rectangle

- rectangles can be represented in many ways, e.g.
  - width, height, and one corner or the center
  - two opposing corners
- here we choose width, height and the lower-left corner
- Example:

```
class Rectangle(object):
    "represents a rectangle using attributes width, height, corner"
box = Rectangle()
box.width = 5.0
box.height = 3.0
box.corner = p
```



# Instances as Return Values

- functions can return instances
- Example: find the center point of a rectangle

```
def find_center(box):
    p = Point()
    p.x = box.corner.x + box.width / 2.0
    p.y = box.corner.y + box.height / 2.0
    return p
box = Rectangle()
box.width = 5.0;        box.height = 3.0
box.corner = Point()
box.corner.x = 3.0;     box.corner.y = 4.0
print(find_center(box))
```

# Objects are Mutable

- by assigning to attributes, an object is changed
- Example: update size of rectangle

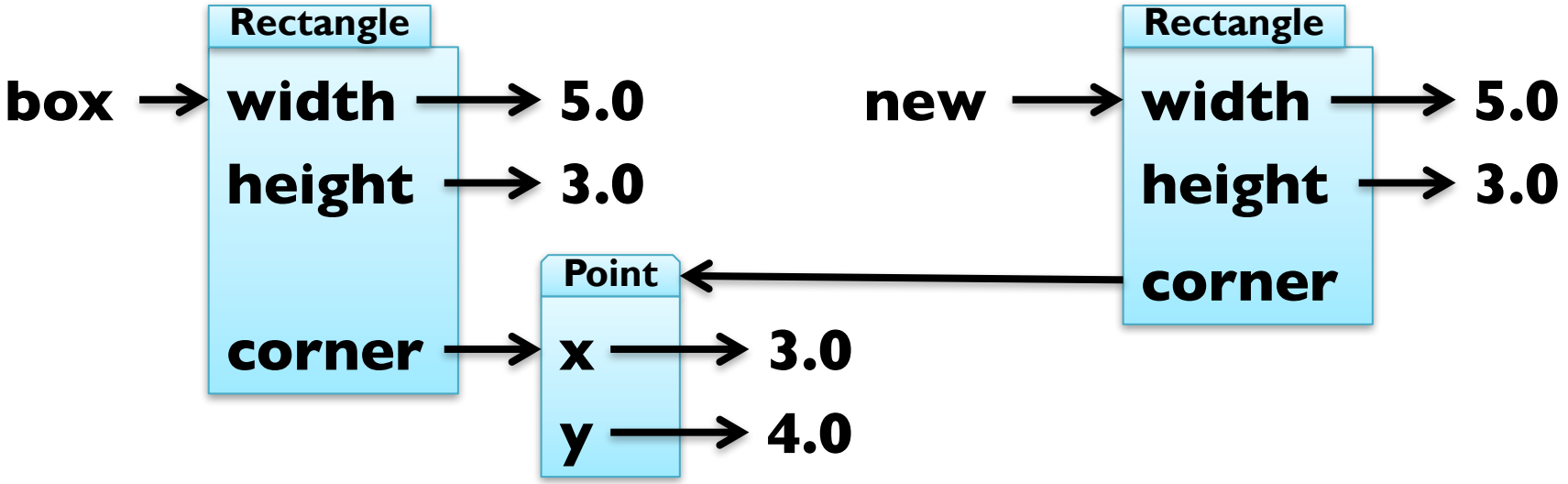
```
box.width = box.width + 5.0
box.height = box.height + 3.0
```

- consequently, also functions can change object arguments
- Example:

```
def double_rectangle(box):
    box.width *= 2
    box.height *= 2
double_rectangle(box)
```

# Copying Objects

- import module copy to make copies of objects
- Example: import copy
  new = copy.copy(box)




- shallow copy, use copy.deepcopy(object) to also copy Point

# Debugging User-Defined Types

- you can obtain type of an instance by using type(object)
- Example: print(type(box))

- you can check if an object has an attribute using hasattr
- Example: hasattr(box, "corner") == True

- you can get a list of all attributes using dir(object)
- Example: dir(box)

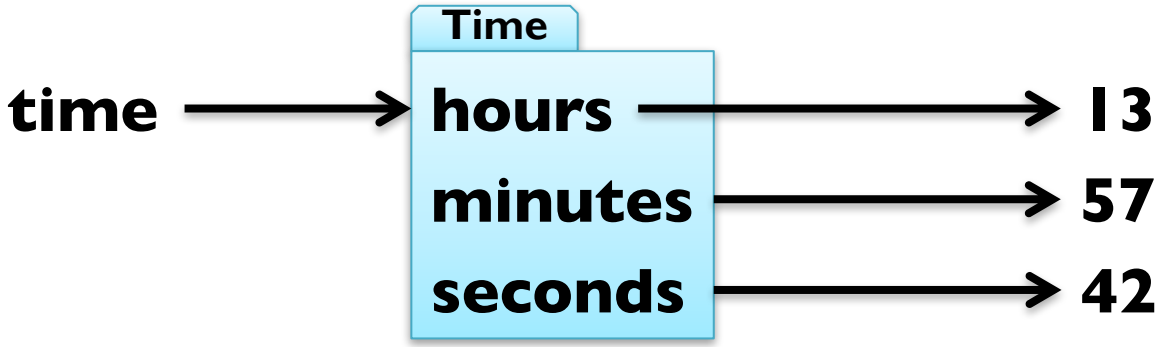
- print __doc__ and __module__ for more information!

# CLASSSES & FUNCTIONS

# Representing Time

- Example: user-defined type for representing time

```
class Time(object):
  """represents time of day using hours, minutes, seconds"""
time = Time()
time.hours = 13
time.minutes = 57
time.seconds = 42
```



# Pure Functions

- pure function = does not modify mutable arguments
- Example: add two times

```
def add_time(t1, t2):
  sum = Time()
  sum.hours = t1.hours + t2.hours
  sum.minutes = t1.minutes + t2.minutes
  sum.seconds = t1.seconds + t2.seconds
  return sum
time = add_time(time, time)
print "%dh %dm %ds" % (time.hours, time.minutes, time.seconds)
```

# Modifiers

- modifiers = functions that modify mutable arguments
- Example: incrementing time

```
def increment(time, seconds):
    time.seconds += seconds
```

```
increment(time, 86400)
print "%dh %dm %ds" % (time.hours, time.minutes, time.seconds)
```

# Modifiers

- modifiers = functions that modify mutable arguments
- Example: incrementing time

```
def increment(time, seconds):
    time.seconds += seconds
    minutes, time.seconds = divmod(time.seconds, 60)
    time.minutes += minutes
    hours, time.minutes = divmod(time.minutes, 60)
    time.hours + = hours
increment(time, 86400)
print "%dh %dm %ds" % (time.hours, time.minutes, time.seconds)
```

- this was *prototype and patch* (or *trial and error*)

# Prototyping vs Planning

- alternative to prototyping is *planned development*
- high-level observation: time representable by just seconds
- Example: refactoring function working with time

```
def time_to_int(time):
    return time.seconds + 60 * (time.minutes + 60 * time.hours)
def int_to_time(seconds):
    time = Time();  minutes, time.seconds = divmod(seconds, 60)
    time.hours, time.minutes = divmod(minutes, 60);   return time
def add_time(t1, t2):
    return int_to_time(time_to_int(t1) + time_to_int(t2))
```

# Prototyping vs Planning

- alternative to protyping is *planned development*
- high-level observation: time representable by just seconds
- Example: refactoring function working with time

```
def time_to_int(time):
    return time.seconds + 60 * (time.minutes + 60 * time.hours)
def int_to_time(seconds):
    time = Time();  minutes, time.seconds = divmod(seconds, 60)
    time.hours, time.minutes = divmod(minutes, 60);   return time
def increment(time, seconds):
    t = int_to_time(seconds + time_to_int(time))
    time.seconds = t.seconds;  time.minutes = t.minutes
    time.hours = t.hours
```

# Prototyping vs Planning

- alternative to protyping is *planned development*
- high-level observation: time representable by just seconds
- Example: refactoring function working with time

```
def time_to_int(time):
    return time.seconds + 60 * (time.minutes + 60 * time.hours)
def int_to_time(seconds):
    time = Time();  minutes, time.seconds = divmod(seconds, 60)
    time.hours, time.minutes = divmod(minutes, 60);   return time
def increment(time, seconds):
    return int_to_time(seconds + time_to_int(time))
```

# Debugging using Invariants

- invariant = requirement that is always true
- assertion = statement of an invariant using assert
- Example: check that time is valid

```
def valid_time(time):
    if time.hours < 0 or time.minutes < 0 or time.seconds < 0:
        return False
    return time.minutes < 60 and time.seconds < 60
def add_time(t1, t2):
    assert valid_time(t1) and valid_time(t2)
    return int_to_time(time_to_int(t1) + time_to_int(t2))
```

- also useful to check before return value