Advanced Data Management (CSCI 490/680)

Structured Data

Dr. David Koop

Objects

- `d = dict() # construct an empty dictionary object`
- `l = list() # construct an empty list object`
- `s = set() # construct an empty set object`
- `s = set([1,2,3,4]) # construct a set with 4 numbers`
- Calling methods:
  - `l.append('abc')`
  - `d.update({'a': 'b'})`
  - `s.add(3)`
- The method is tied to the object preceding the dot (e.g. `append` modifies `l` to add `'abc'`)

Python Modules

- Python module: a file containing definitions and statements
- Import statement: like Java, get a module that isn't a Python builtin

```
import collections
d = collections.defaultdict(list)
d[3].append(1)
```

- `import <name> as <shorter-name>`

```
import collections as c
```

- `from <module> import <name>` – don't need to refer to the module

```
from collections import defaultdict
d = defaultdict(list)
d[3].append(1)
```

Other Collections Features

- `collections.defaultdict`: specify a default value for any item in the dictionary (instead of `KeyError`)
- `collections.OrderedDict`: keep entries ordered according to when the key was inserted
  - `dict` objects are ordered in Python 3.7 but `OrderedDict` has some other features (equality comparison, reversed)
- `collections.Counter`: counts hashable objects, has a `most_common` method

None

- Like null in other languages, used as a placeholder when no value exists
- The value returned from a function that doesn't return a value

```
def f(name):
    print("Hello,", name)
v = f("Patricia") # v will have the value None
```

- Also used as a **sentinel** value when you need to create a new object:

```
def add_letters(s, d=None):
    if d is None:
        d = {}
    d.update(count_letters(s))
    return d
```

- Looks like `d={}` would make more sense, but that causes issues

Iterators

- Remember `range, values, keys, items`?
- They return **iterators**: objects that traverse containers, only need to know how to get the next element
- Given iterator `it, next(it)` gives the next element
- `StopIteration` exception if there isn't another element
- Generally, we don't worry about this as the for loop handles everything automatically…but you cannot index or slice an iterator
- `d.values()[0]` will not work!
- If you need to index or slice, construct a list from an iterator
- `list(d.values())[0]` or `list(range(100))[-1]`
- In general, this is slower code so we try to avoid creating lists

List Comprehensions

- Shorthand for transformative or filtering for loops
- ```
  squares = []
  for i in range(10):
      squares.append(i**2)
  ```
- `squares = [i**2 for i in range(10)]`
- Filtering:
- ```
  squares = []
  for i in range(10):
      if i % 3 != 1:
          squares.append(i ** 2)
  ```
- `squares = [i**2 for i in range(10) if i % 3 != 1]`
- if clause **follows** the for clause

Dictionary Comprehensions

- Similar idea, but allow dictionary construction
- Could use lists:
  - `names = dict([(k, v) for k,v in … if …])`
- Native comprehension:
  - ```
    names = {"Al": ["Smith", "Brown"], "Beth":["Jones"]}
    first_counts ={k: len(v) for k,v in names.items()}
    ```
- Could do this with a for loop as well
    ```

Assignment 1

- Using Python for data analysis
- Analyze hurricane data (through 2018)
- Provided `a1.ipynb` file (right-click and download)
- Use basic python (+ collections module) for now to demonstrate language knowledge
- Use Anaconda or hosted Python environment
- Due next Wednesday
- Turn .ipynb file in via Blackboard

Exceptions

- errors but potentially something that can be addressed
- try-except-else-finally:
  - `except` clause runs if exactly the error(s) you wish to address happen
  - `else` clause will run if no exceptions are encountered
  - `finally` always runs (even if the program is about to crash)
- Can have multiple `except` clauses
- can also raise exceptions using the `raise` keyword
- (and define your own)

Classes

- `class ClassName:`

  `    …`
- Everything in the class should be indented until the declaration ends
- `self`: `this` in Java or C++ is `self` in Python
- Every instance method has `self` as its first parameter
- Instance variables are defined **in methods** (usually constructor)
- `__init__`: the constructor, should initialize instance variables
- ```
  def __init__(self):
      self.a = 12
      self.b = 'abc'
  ```
- ```
  def __init__(self, a, b):
      self.a = a
      self.b = b
  ```

Class Example

- ```
  class Rectangle:
      def __init__(self, x, y, w, h):
          self.x = x
          self.y = y
          self.w = w
          self.h = h

      def set_corner(self, x, y):
          self.x = x
          self.y = y

      def set_width(self, w): self.w = w

      def set_height(self, h): self.h = h

      def area(self):
          return self.w * self.h
  ```
  ```

Arrays

What is the difference between an array and a list (or a tuple)?

Arrays

- Usually a fixed size—lists are meant to change size
- Are mutable—tuples are not
- Store only one type of data—lists and tuples can store anything
- Are faster to access and manipulate than lists or tuples
- Can be multidimensional:
  - Can have list of lists or tuple of tuples but no guarantee on shape
  - Multidimensional arrays are rectangles, cubes, etc.

Why NumPy?

- Fast **vectorized** array operations for data munging and cleaning, subsetting and filtering, transformation, and any other kinds of computations
- Common array algorithms like sorting, unique, and set operations
- Efficient descriptive statistics and aggregating/summarizing data
- Data alignment and relational data manipulations for merging and joining together heterogeneous data sets
- Expressing conditional logic as array expressions instead of loops with `if-elif-else` branches
- Group-wise data manipulations (aggregation, transformation, function application).

[W. McKinney, Python for Data Analysis]

```
import numpy as np
```

Textbook's Notebooks

- https://github.com/wesm/pydata-book/
- ch04.ipynb
- Click the raw button and save that file to disk
- …or download/clone the entire repository

Creating arrays

- `data1 = [6, 7.5, 8, 0, 1]`
  `arr1 = np.array(data1)`
- `data2 = [[1,2,3,4],[5,6,7,8]]`
  `arr2 = np.array(data2)`
- Number of dimensions: `arr2.ndim`
- Shape: `arr2.shape`
- Types: `arr1.dtype, arr2.dtype`, can specify explicitly (`np.float64`)

Creating Arrays

- Zeros: `np.zeros(10)`
- Ones: `np.ones((4,5))`
- Empty: `np.empty((2,2))`
- _like versions: pass an existing array and matches shape with specified contents
- Range: `np.arange(15)`

Types

- "But I thought Python wasn't stingy about types…"
- numpy aims for speed
- Able to do array arithmetic
- int16, int32, int64, float32, float64, bool, object
- `astype` method allows you to convert between different types of arrays:

```
arr = np.array([1, 2, 3, 4, 5])
arr.dtype
float_arr = arr.astype(np.float64)
```

numpy data types (dtypes)

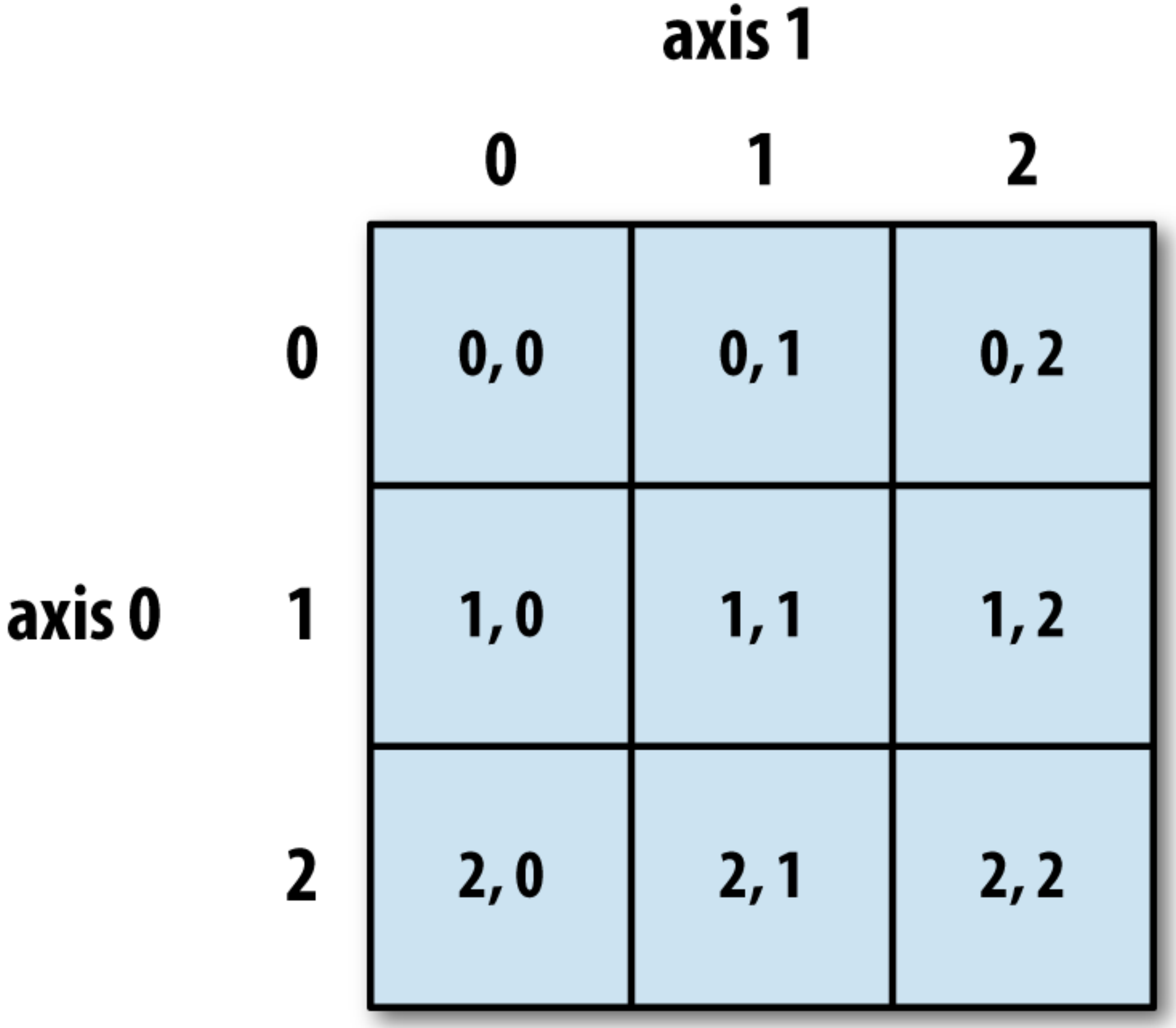
Type	Type code	Description
`int8, uint8`	`i1, u1`	Signed and unsigned 8-bit (1 byte) integer types
`int16, uint16`	`i2, u2`	Signed and unsigned 16-bit integer types
`int32, uint32`	`i4, u4`	Signed and unsigned 32-bit integer types
`int64, uint64`	`i8, u8`	Signed and unsigned 64-bit integer types
`float16`	`f2`	Half-precision floating point
`float32`	`f4 or f`	Standard single-precision floating point; compatible with C float
`float64`	`f8 or d`	Standard double-precision floating point; compatible with C double and Python `float` object
`float128`	`f16 or g`	Extended-precision floating point
`complex64, complex128, complex256`	`c8, c16, c32`	Complex numbers represented by two 32, 64, or 128 floats, respectively
`bool`	?	Boolean type storing `True` and `False` values
`object`	O	Python object type; a value can be any Python object
`string_`	S	Fixed-length ASCII string type (1 byte per character); for example, to create a string dtype with length 10, use `'S10'`
`unicode_`	U	Fixed-length Unicode type (number of bytes platform specific); same specification semantics as `string_` (e.g., `'U10'`)

[W. McKinney, Python for Data Analysis]

Operations

- (Array, Array) Operations (elementwise)
  - Addition, Subtraction, Multiplication
- (Scalar, Array) Operations:
  - Addition, Subtraction, Multiplication, Division, Exponentiation
- Indexing
  - Same as with lists plus shorthand for 2D+
  - ```
    arr = np.array([[1,2],[3,4]])
    arr[1,1]
    ```
    ```

2D Indexing



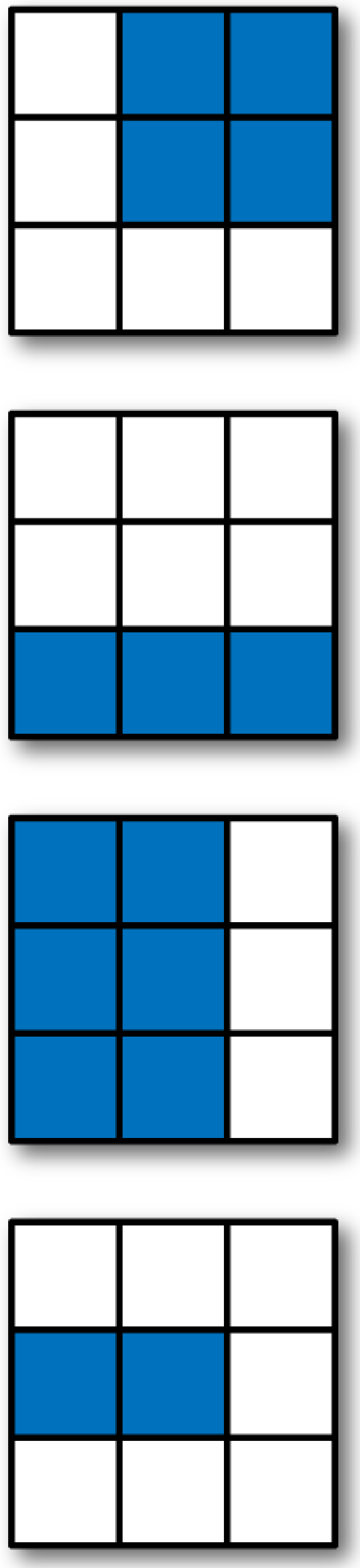
[W. McKinney, Python for Data Analysis]

Slicing

- 1D: Just like with lists except **data is not copied**!
  - `a[2:5] = 3` works with arrays
  - `a.copy()` or `a[2:5].copy()` will copy
- 2D+: comma separated indices as shorthand:
  - `a[1][2]` or `a[1,2]`
  - `a[1]` gives a row
  - `a[:,1]` gives a column

2D Array Slicing

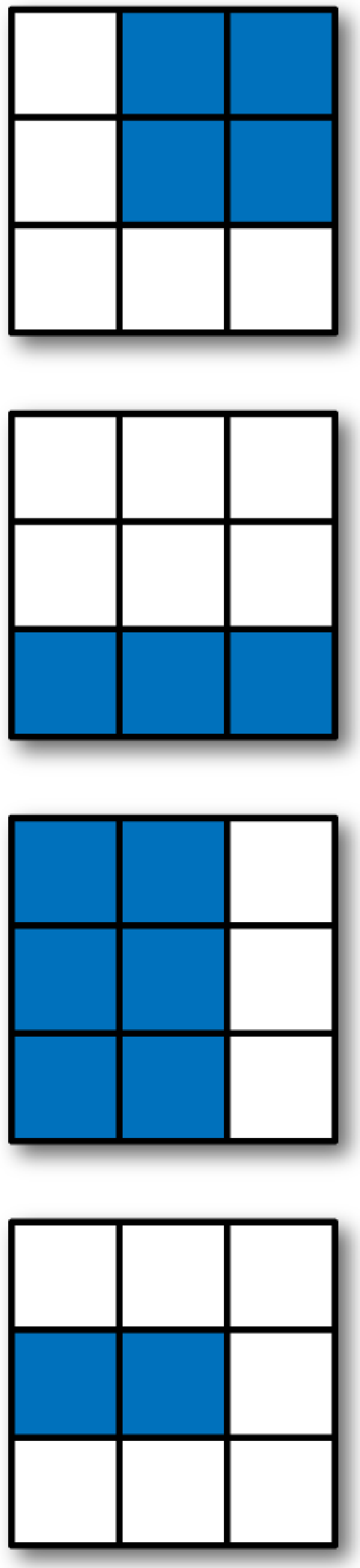
How to obtain the blue slice from array `arr`?



[W. McKinney, Python for Data Analysis]

2D Array Slicing

How to obtain the blue slice from array `arr`?

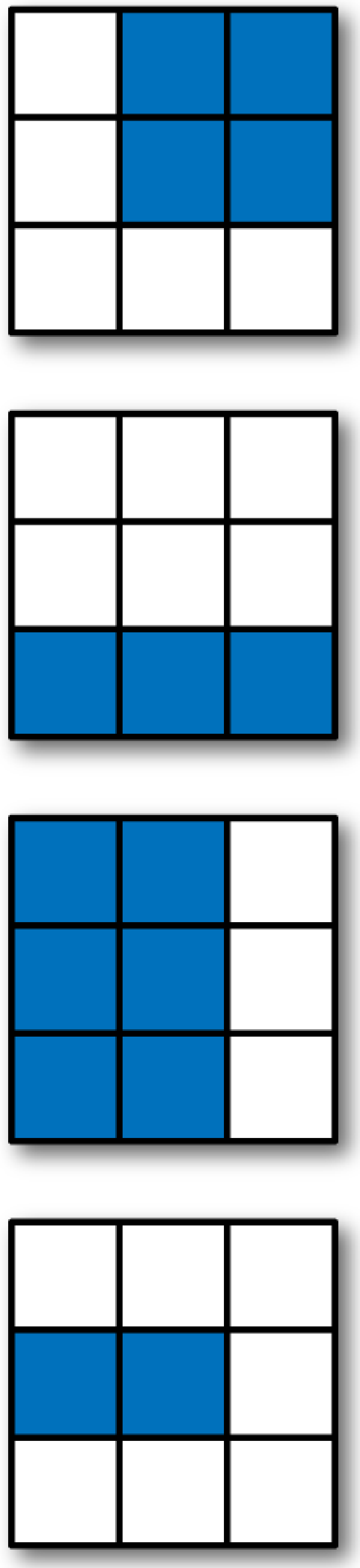


Expression	Shape
`arr[:2, 1:]`	(2, 2)

[W. McKinney, Python for Data Analysis]

2D Array Slicing

How to obtain the blue slice from array `arr`?

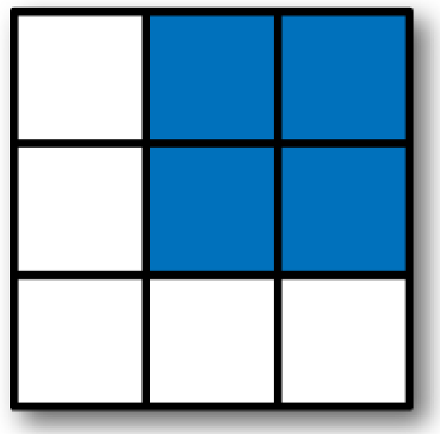


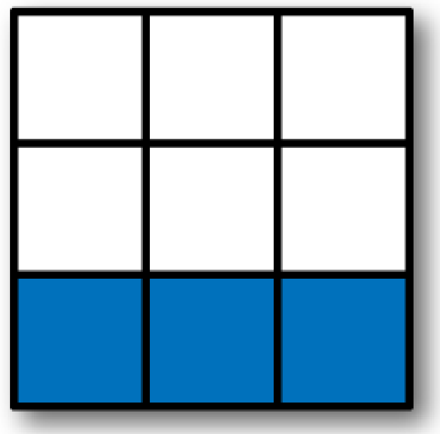
Expression	Shape
`arr[:2, 1:]`	(2, 2)
`arr[2]`	(3,)
`arr[2, :]`	(3,)
`arr[2:, :]`	(1, 3)

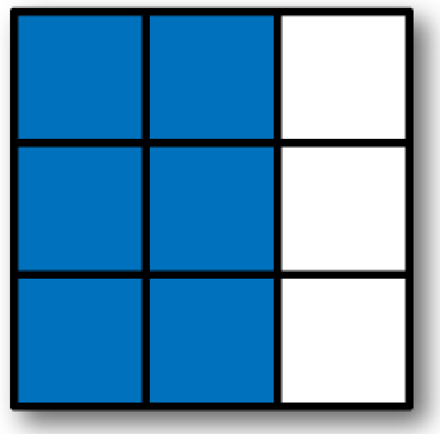
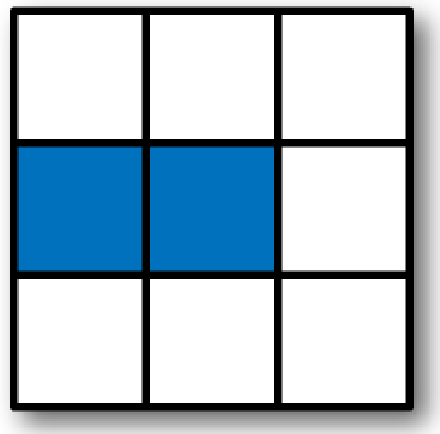
[W. McKinney, Python for Data Analysis]

2D Array Slicing

How to obtain the blue slice from array `arr`?

	Expression	Shape
	`arr[:2, 1:]`	(2, 2)

	`arr[2]`
	`arr[2, :]`
	`arr[2:, :]`
		(3,)
		(3,)
		(1, 3)

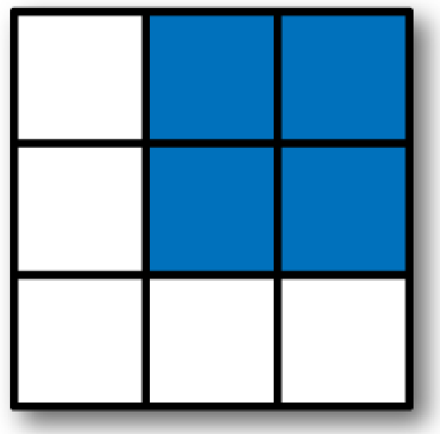
	`arr[:, :2]`	(3, 2)
		

[W. McKinney, Python for Data Analysis]

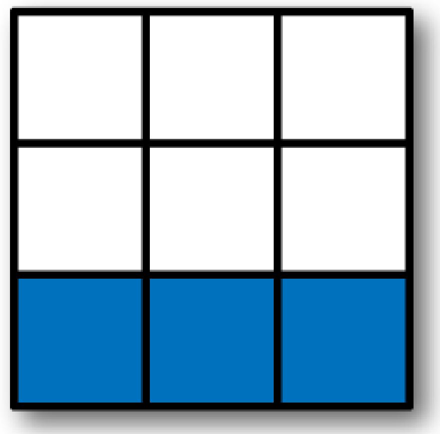
2D Array Slicing

How to obtain the blue slice from array `arr`?

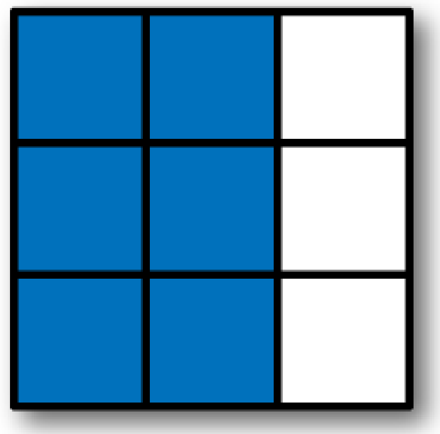
Expression	Shape



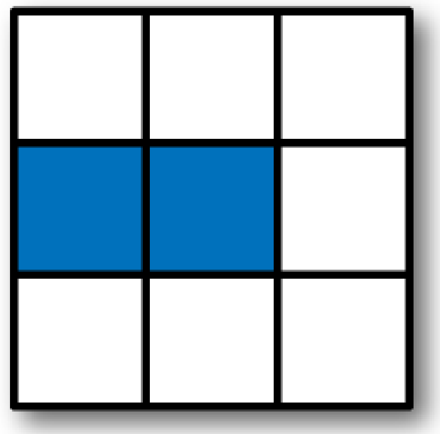
Expression	Shape
`arr[:2, 1:]`	(2, 2)



Expression	Shape
`arr[2]`	(3,)
`arr[2, :]`	(3,)
`arr[2:, :]`	(1, 3)



Expression	Shape
`arr[:, :2]`	(3, 2)



Expression	Shape
`arr[1, :2]`	(2,)
`arr[1:2, :2]`	(1, 2)

[W. McKinney, Python for Data Analysis]

Boolean Indexing

- `names == 'Bob'` gives back booleans that represent the element-wise comparison with the array `names`
- Boolean arrays can be used to index into another array:
  - `data[names == 'Bob']`
- Can even mix and match with integer slicing
- Can do boolean operations (`&`, `|`) between arrays (just like addition, subtraction)
  - `data[(names == 'Bob') | (names == 'Will')]`
- Note: `or` and `and` do not work with arrays
- We can set values too! `data[data < 0] = 0`

Other Operations

- Fancy Indexing: `arr[[1,2,3]]`
- Transposing arrays: `arr.T`
- Reshaping arrays: `arr.reshape((3,5))`
- Unary universal functions (ufuncs): `np.sqrt, np.exp`
- Binary universal functions: `np.add, np.maximum`

Unary Universal Functions

Function	Description
`abs, fabs`	Compute the absolute value element-wise for integer, floating-point, or complex values
`sqrt`	Compute the square root of each element (equivalent to `arr ** 0.5`)
`square`	Compute the square of each element (equivalent to `arr ** 2`)
`exp`	Compute the exponent e^x of each element
`log, log10, log2, log1p`	Natural logarithm (base *e*), log base 10, log base 2, and log(1 + x), respectively
`sign`	Compute the sign of each element: 1 (positive), 0 (zero), or –1 (negative)
`ceil`	Compute the ceiling of each element (i.e., the smallest integer greater than or equal to that number)
`floor`	Compute the floor of each element (i.e., the largest integer less than or equal to each element)
`rint`	Round elements to the nearest integer, preserving the `dtype`
`modf`	Return fractional and integral parts of array as a separate array
`isnan`	Return boolean array indicating whether each value is `NaN` (Not a Number)
`isfinite, isinf`	Return boolean array indicating whether each element is finite (non-`inf`, non-`NaN`) or infinite, respectively
`cos, cosh, sin, sinh, tan, tanh`	Regular and hyperbolic trigonometric functions
`arccos, arccosh, arcsin, arcsinh, arctan, arctanh`	Inverse trigonometric functions
`logical_not`	Compute truth value of `not x` element-wise (equivalent to `~arr`).

[W. McKinney, Python for Data Analysis]

Binary Universal Functions

Function	Description
`add`	Add corresponding elements in arrays
`subtract`	Subtract elements in second array from first array
`multiply`	Multiply array elements
`divide, floor_divide`	Divide or floor divide (truncating the remainder)
`power`	Raise elements in first array to powers indicated in second array
`maximum, fmax`	Element-wise maximum; `fmax` ignores `NaN`
`minimum, fmin`	Element-wise minimum; `fmin` ignores `NaN`
`mod`	Element-wise modulus (remainder of division)
`copysign`	Copy sign of values in second argument to values in first argument
`greater, greater_equal, less, less_equal, equal, not_equal`	Perform element-wise comparison, yielding boolean array (equivalent to infix operators `>, >=, <, <=, ==, !=`)
`logical_and, logical_or, logical_xor`	Compute element-wise truth value of logical operation (equivalent to infix operators `& |, ^`)

[W. McKinney, Python for Data Analysis]

Statistical Methods

Method	Description
`sum`	Sum of all the elements in the array or along an axis; zero-length arrays have sum 0
`mean`	Arithmetic mean; zero-length arrays have `NaN` mean
`std, var`	Standard deviation and variance, respectively, with optional degrees of freedom adjustment (default denominator `n`)
`min, max`	Minimum and maximum
`argmin, argmax`	Indices of minimum and maximum elements, respectively
`cumsum`	Cumulative sum of elements starting from 0
`cumprod`	Cumulative product of elements starting from 1

[W. McKinney, Python for Data Analysis]

More

- Other methods:
  - `any` and `all`
  - `sort`
  - `unique`
- Linear Algebra (`numpy.linalg`)
- Pseudorandom Number Generation (`numpy.random`)

Chicago Food Inspections Exploration

- Based on David Beazley's PyData Chicago talk
- YouTube video: https://www.youtube.com/watch?v=j6VSAsKAj98
- Our in-class exploration:
  - Don't focus on the syntax
  - Focus on:
    - What is information is available
    - **Questions** are interesting about this dataset
    - How to decide on good follow-up questions
    - What the computations mean

Chicago Food Inspections Exploration