GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 3: MACHINE LEARNING

LAST TIME:

- LINEAR ALGEBRA REVIEW
- PYTHON CONTROL FLOW

QUESTIONS?

I. WHAT IS MACHINE LEARNING?

II. MACHINE LEARNING PROBLEMS

III. PYTHON LIBRARIES

EXERCISES:

III. NUMPY, SCIPY, AND PANDAS

I. WHAT IS MACHINE LEARNING?

WHAT IS MACHINE LEARNING?

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

“The core of machine learning deals with representation and generalization…”

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

“The core of machine learning deals with representation and generalization…”

- representation – extracting structure from data

source: http://en.wikipedia.org/wiki/Machine_learning

WHAT IS MACHINE LEARNING?

from Wikipedia:

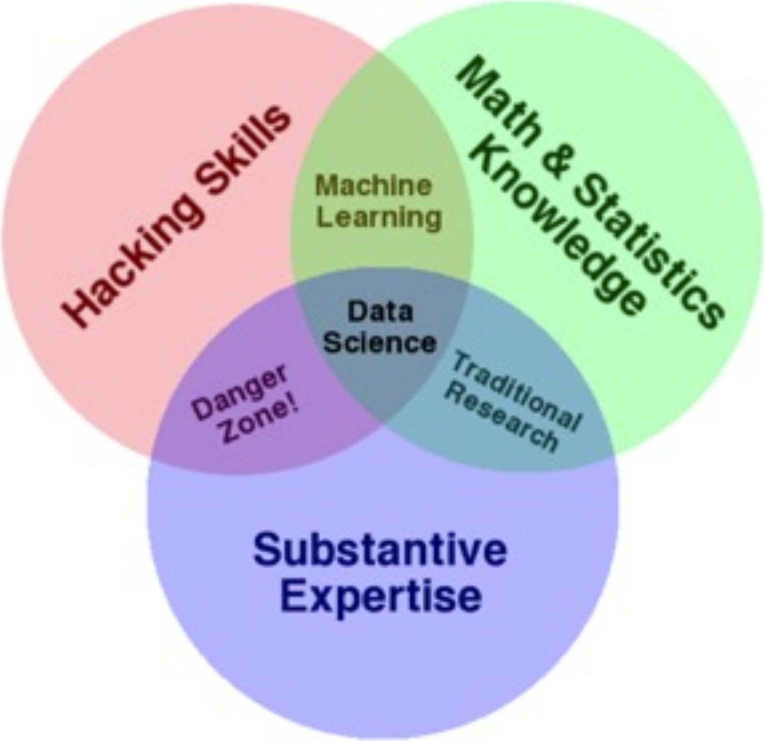
“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data.”

“The core of machine learning deals with representation and generalization…”

- representation – extracting structure from data
- generalization – making predictions from data

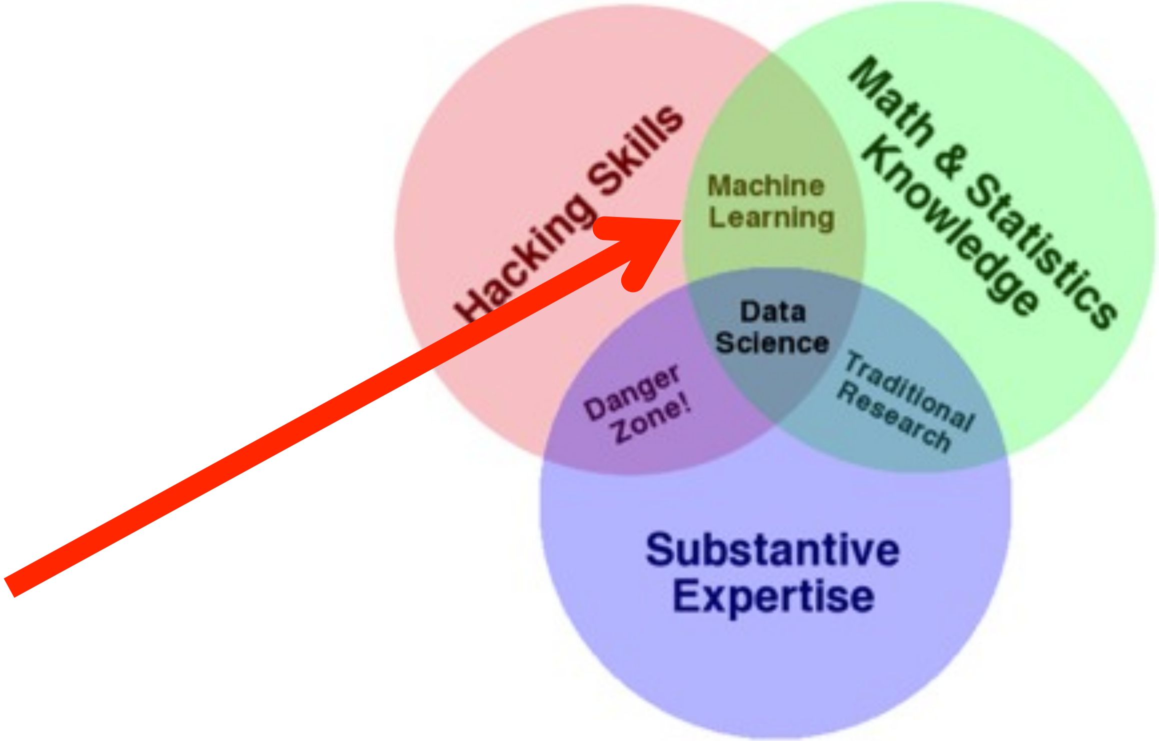
source: http://en.wikipedia.org/wiki/Machine_learning

REMEMBER THIS?



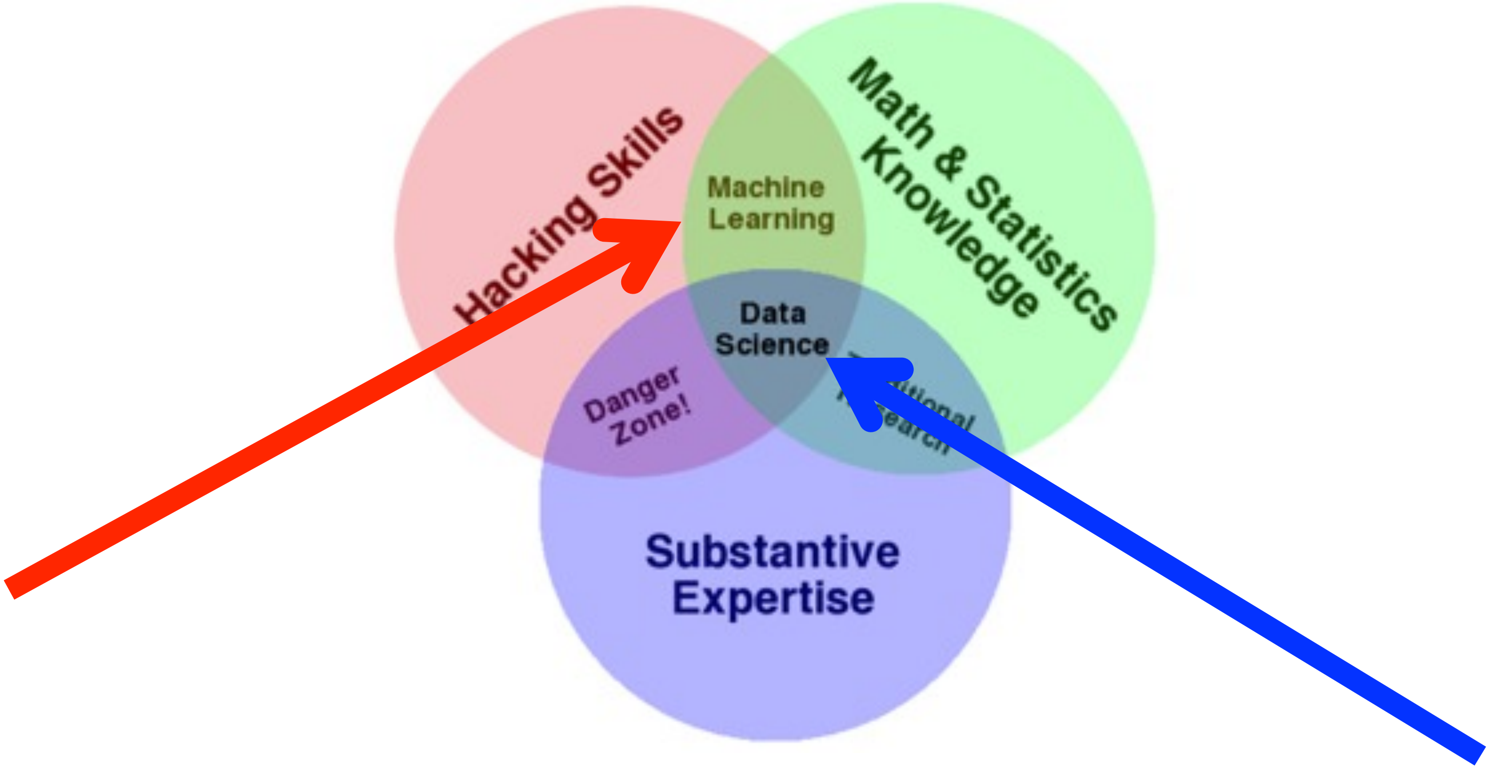
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU ARE HERE



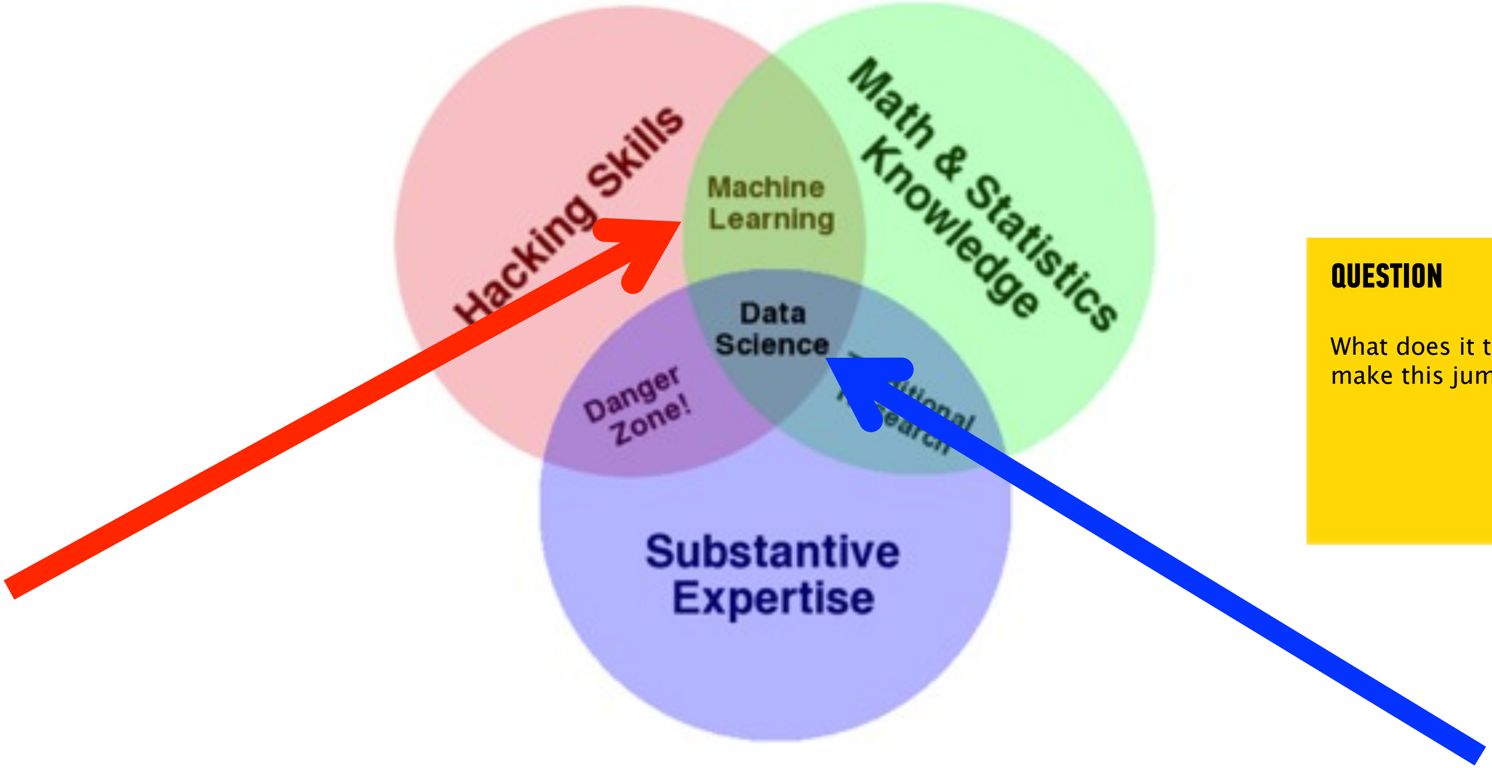
source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU WANT TO GO HERE



source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

YOU WANT TO GO HERE

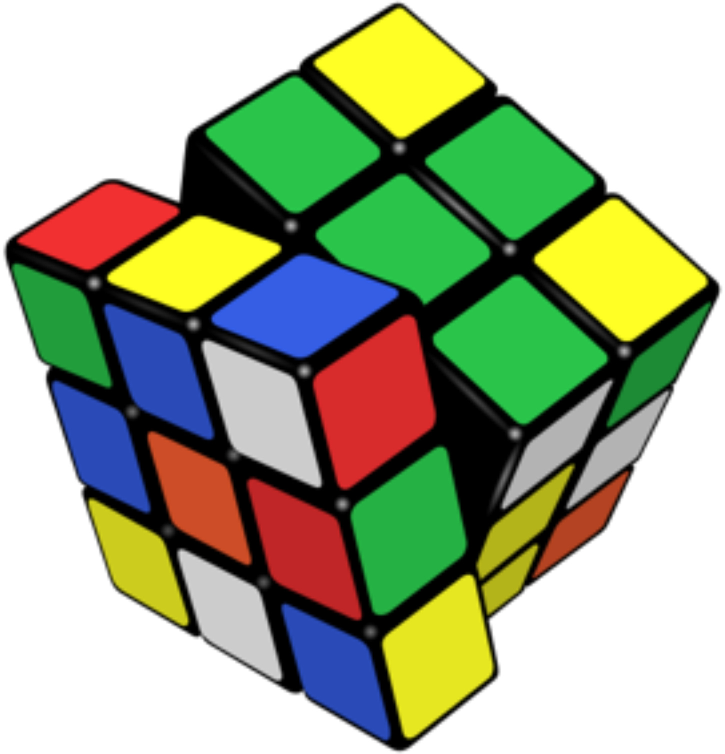


source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

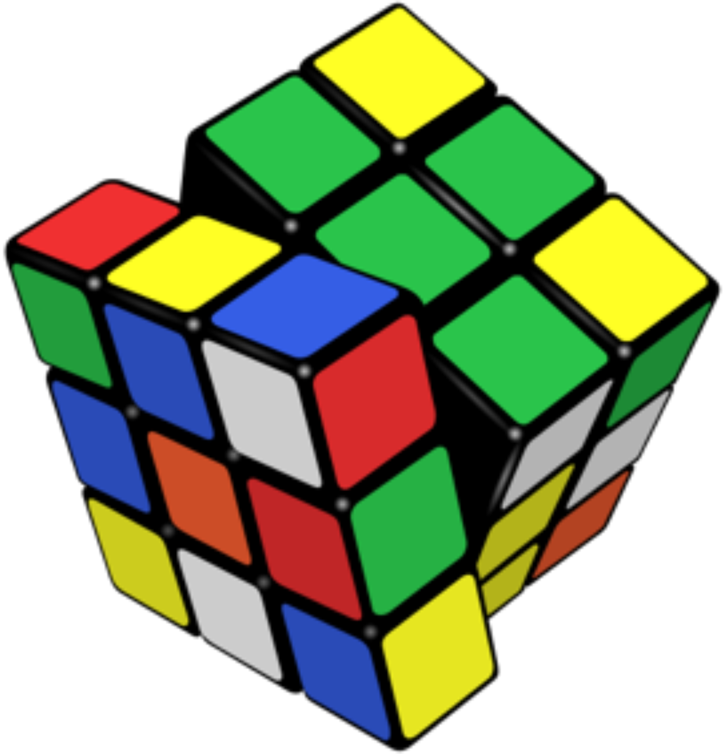
QUESTION

What does it take to make this jump?

ANSWER: PROBLEM SOLVING!



ANSWER: PROBLEM SOLVING!



NOTE

Implementing solutions to ML problems is the focus of this course!

REVIEW

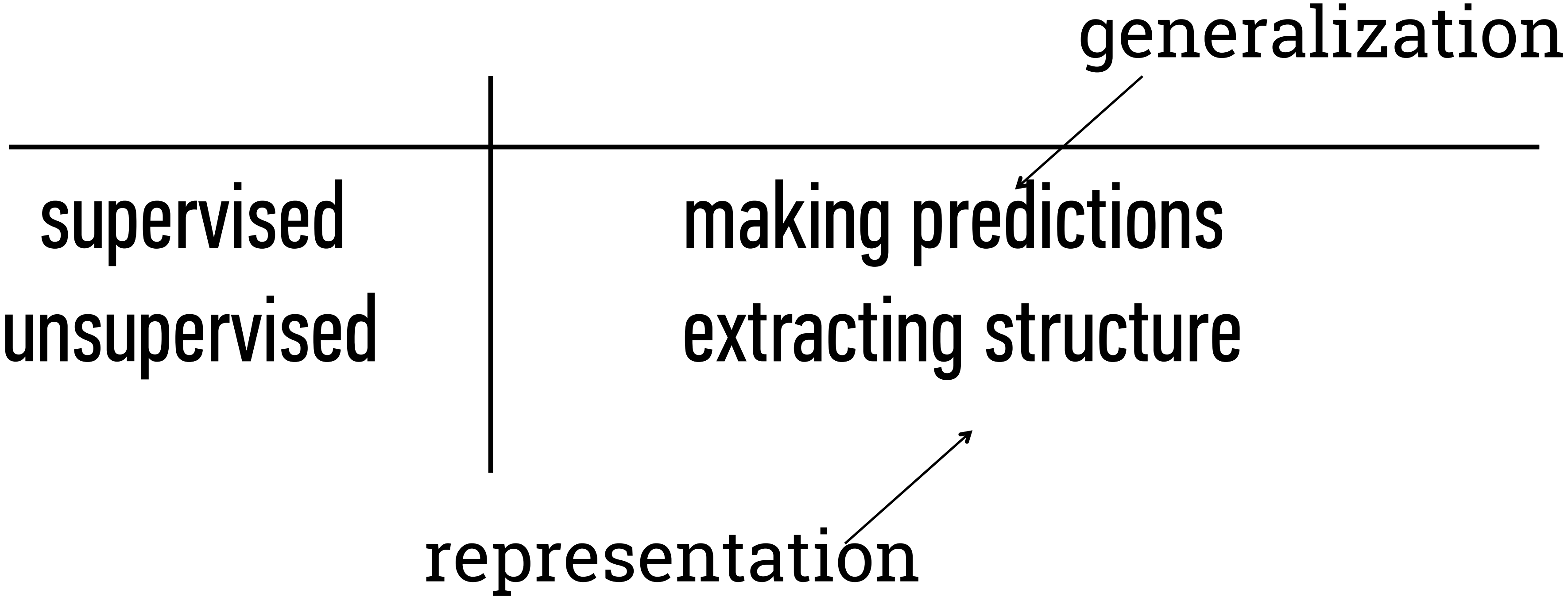
1. What is machine learning?
2. What are the two use cases for machine learning?

II. MACHINE LEARNING PROBLEMS

TYPES OF LEARNING PROBLEMS

supervised	making predictions
unsupervised	extracting structure

REMEMBER WHAT WE SAID BEFORE?



generalization

supervised

unsupervised

making predictions

extracting structure

representation

TYPES OF DATA

	continuous	categorical
	quantitative	qualitative

TYPES OF DATA

continuous	categorical
quantitative	qualitative

NOTE

The space where data live is called the feature space.

Each point in this space is called a record.

TYPES OF ML SOLUTIONS

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

TYPES OF ML SOLUTIONS

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

NOTE

We will implement solutions using models and algorithms.

Each will fall into one of these four buckets.

QUESTION

WHAT IS THE GOAL OF MACHINE LEARNING?

GOALS OF ML

supervised	making predictions
unsupervised	extracting structure

ANSWER

The goal is determined by the type of problem.

QUESTION

HOW DO YOU DETERMINE THE RIGHT APPROACH?

APPROACHES TO ML PROBLEMS

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

ANSWER

The right approach is determined by the desired solution.

APPROACHES TO ML PROBLEMS

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

ANSWER

The
app
det
des

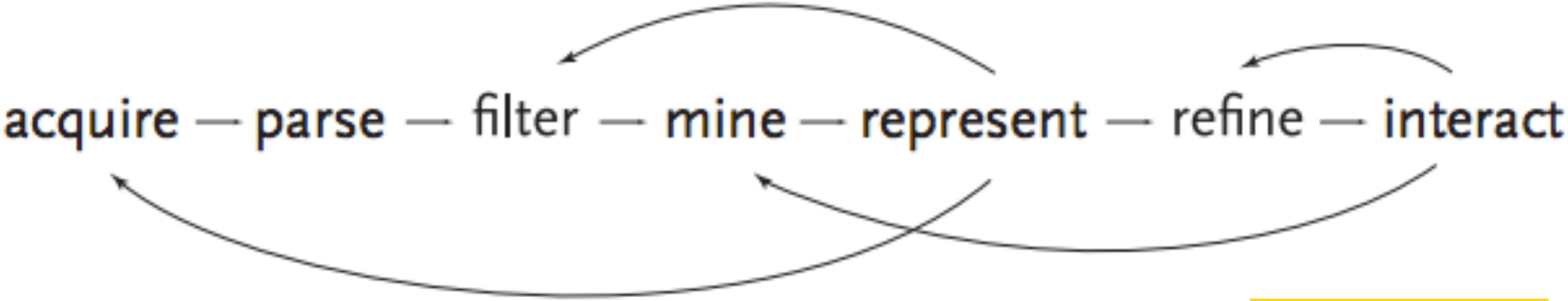
NOTE

All of this depends on your data!

QUESTION

WHAT DO YOU DO WITH YOUR RESULTS?

THE DATA SCIENCE WORKFLOW

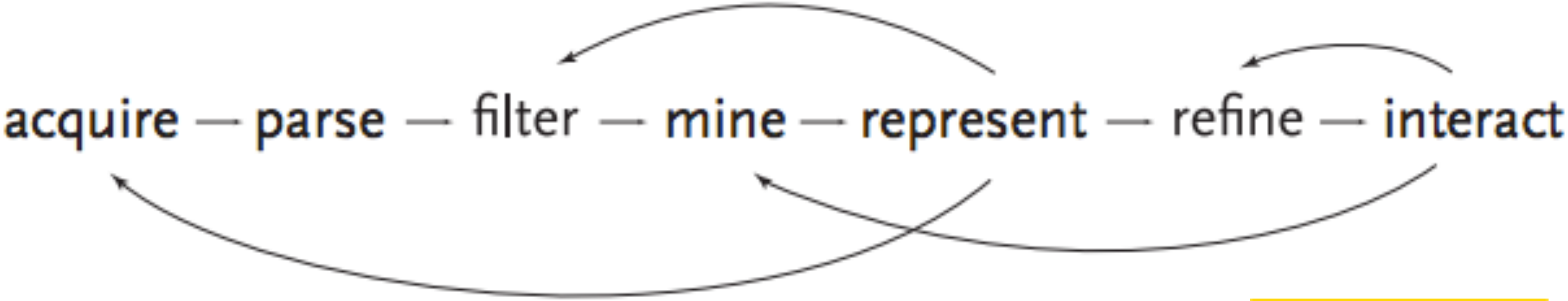

acquire — parse — filter — mine — represent — refine — interact

ANSWER

Interpret them and
react accordingly.

source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW



acquire — parse — filter — mine — represent — refine — interact

source: http://benfry.com/phd/dissertation-110323c.pdf

ANSWER

Int
re

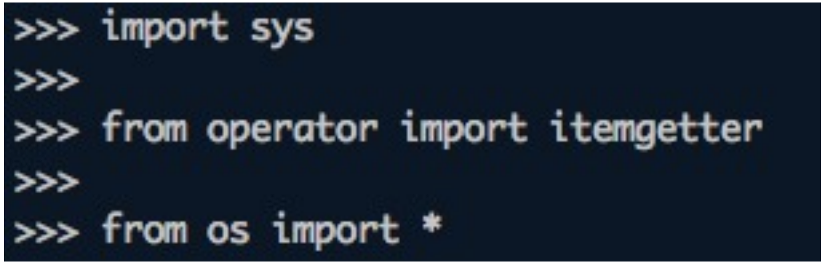
NOTE

This also relies on your problem solving skills!

III. PYTHON LIBRARIES

IMPORT

Python libraries are imported into scripts using the **import statement**.
The import statement can be used in three ways:



```
>>> import sys
>>>
>>> from operator import itemgetter
>>>
>>> from os import *
```

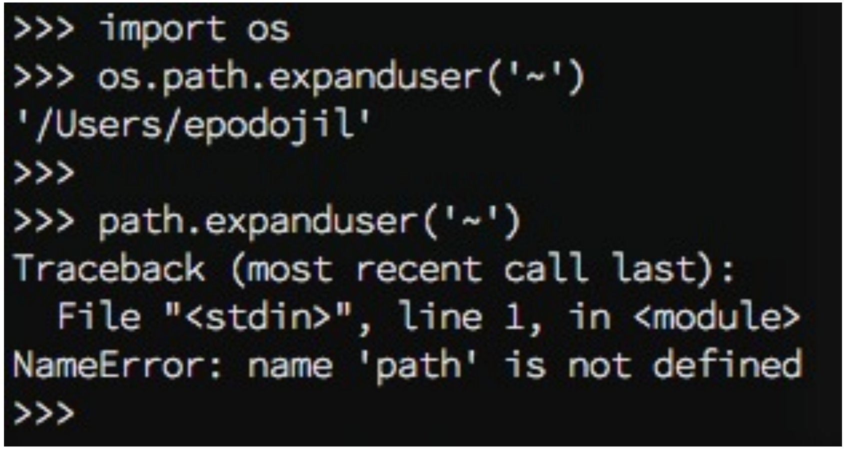
The differences have to do with how each import statement interacts with the local namespace.

NAMESPACES

Python has three types of namespaces:

local, global, and built-in

For our purposes, namespaces are important because they control how imported code can be accessed:

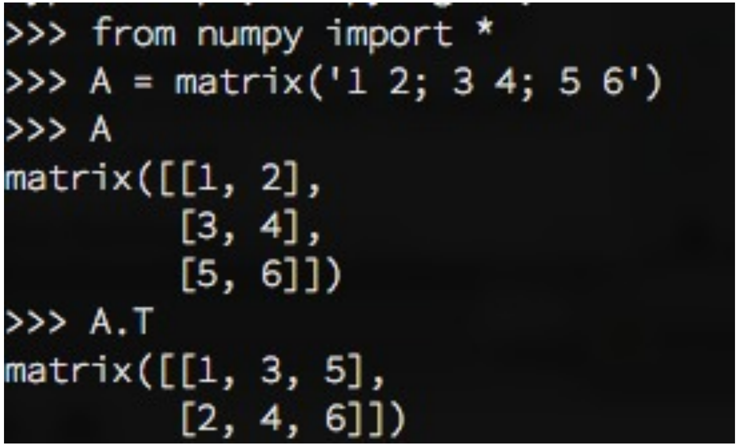


```
>>> import os
>>> os.path.expanduser('~')
'/Users/epodojil'
>>>
>>> path.expanduser('~')
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'path' is not defined
>>>
```

NUMPY

We’ll be using four external libraries that help us structure our data accordingly.

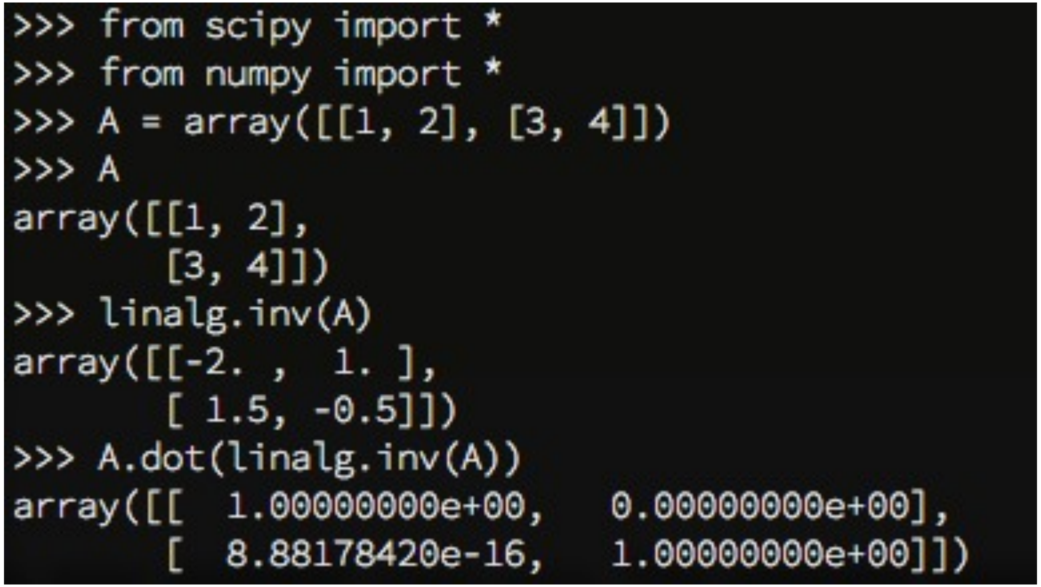
Numpy offers the ability to create arrays (matrices and vectors), as well as some linear algebra functions!



```
>>> from numpy import *
>>> A = matrix('1 2; 3 4; 5 6')
>>> A
matrix([[1, 2],
        [3, 4],
        [5, 6]])
>>> A.T
matrix([[1, 3, 5],
        [2, 4, 6]])
```

SCIPY

Scipy extends numpy by offering additional linear algebra functions, signal processing, Fourier transforms, and other statistics functions



```
>>> from scipy import *
>>> from numpy import *
>>> A = array([[1, 2], [3, 4]])
>>> A
array([[1, 2],
       [3, 4]])
>>> linalg.inv(A)
array([[-2. ,  1. ],
       [ 1.5, -0.5]])
>>> A.dot(linalg.inv(A))
array([[  1.00000000e+00,   0.00000000e+00],
       [  8.88178420e-16,   1.00000000e+00]])
```

PANDAS

PANDAS (python data analysis) provides more rigid data structures more attune to other stats languages, like R or matlab.

R users will find PANDAS to be familiar territory.

SCIKIT-LEARN

Scikit-learn is a library which contains the majority of our machine learning algorithms.

We will be primarily using scikit learn in class to experiment and learn various ML functionality.

There are a lot of other libraries out there that enable you to do some incredibly great things.

We definitely won't explore all of them here, but don't be afraid to use our best friend (Google) to help you find libraries that do things you want to get done.

INTRO TO DATA SCIENCE

LAB: NUMPY

LAB: DATA EXPLORATION

CLASSWORK:

1. Use the pandas library to aggregate NYTimes01-20. We'll want to see clickthrough rate by gender and age.

2. Explore plotting your new aggregated data in various forms to understand the **feature space**, and try using sklearn's linear model function with your aggregate data to predict CTR per age.

DISCUSSION

1. Curate a list of potential final project ideas, as our goal is to answer a question using machine learning. for each question: which “problem” does it fall under?

2. We’ll discuss these in smaller groups first, and share some ideas together as a class.

NEXT CLASS SUBJECT: GETTING DATA. DATABASES AND APIS