Python for Data Science

Arjun Ray

INTRODUCTION

General Info

- bathrooms are back past the elevators and to the left
- this workshop runs from 10 AM - 3 PM
- 30 min lunch at 12 PM
- free  in the kitchen
-  @ GA-guest – password : yellowpencil
- technical issues?!

About Me – Arjun Ray

PAST
ex-academic; worked on Neuroscience and Genomics

CURRENT
full stack programmer
teacher
musician

FAVORITE LANGUAGES
Python, Javascript, Ruby, R, SQL, Lisp, C

About You

during this workshop, **you** will learn:

- how to plan a data science project from start to finish
- how to use python inside of Jupyter Notebook
- how to use python packages such as pandas, numpy, scikit-learn to create predictive models of data and visualize results
- how to read documentation for python packages
- how to think about a data science project from start to finish

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

- understand what makes a Data Scientist
- understand where Data Science is used
- understand how Data Science is used

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

- understand why Python is ideal for doing Data Science
- learn about scientific packages commonly used in python
- use a few packages

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

- define what a statistical model is
- understand how different models are classified
- run a few models and visualize results

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

- understand how to approach a data science problem from start to finish in a stepwise manner

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

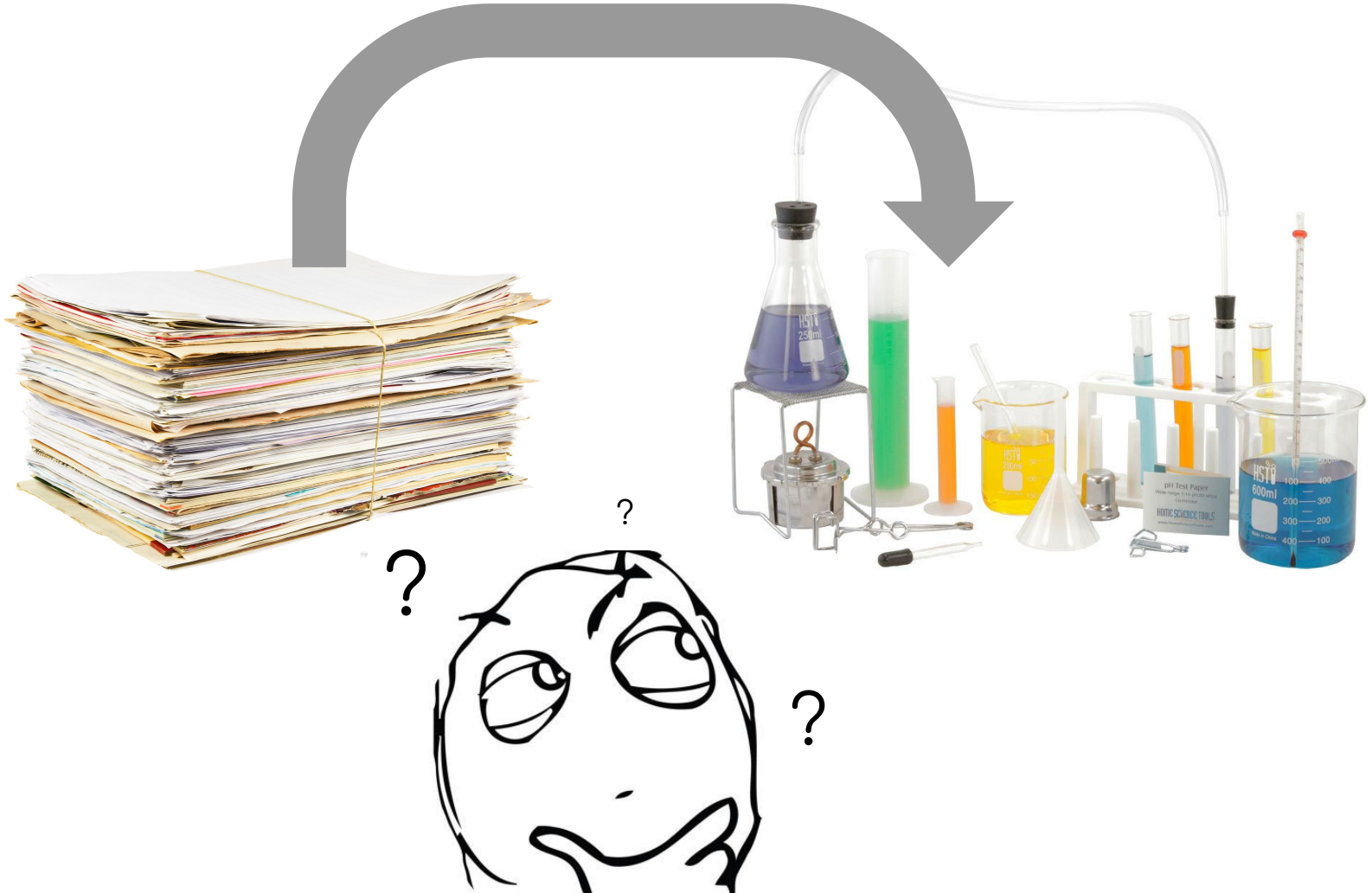
- recap important concepts from workshop
- learn about resources, classes, bootcamps to further your Data Science ability

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

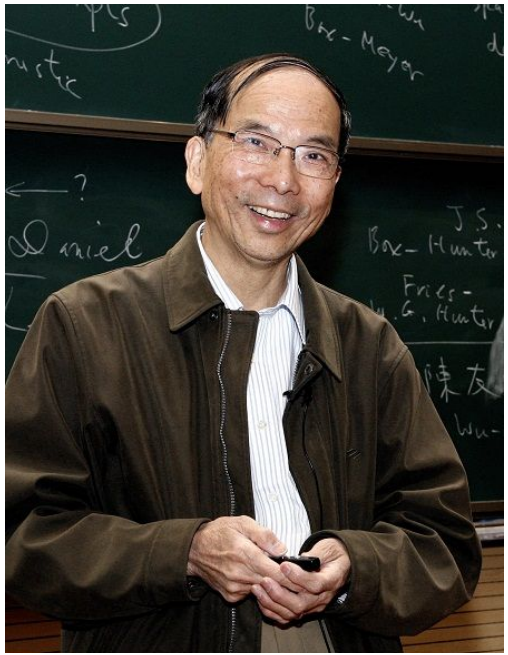
- allow students to ask any questions they want
- filling out Exit Ticket (student feedback form)


?
?
?
pH Test Paper
HOME SCIENCE TOOLS
600ml

WHAT IS DATA SCIENCE?

Statistics = Data Science?

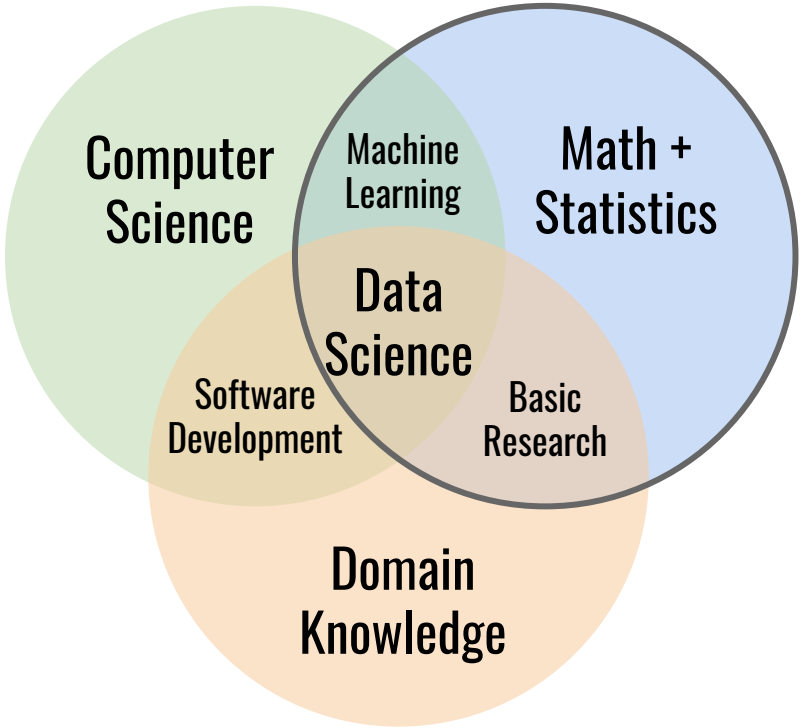
- Dr. Wu gave a lecture in 1997 that first described Data Science
- Statistician = Data Scientist



Chien-Fu Jeff Wu
Data Scientist ~~Statistician~~

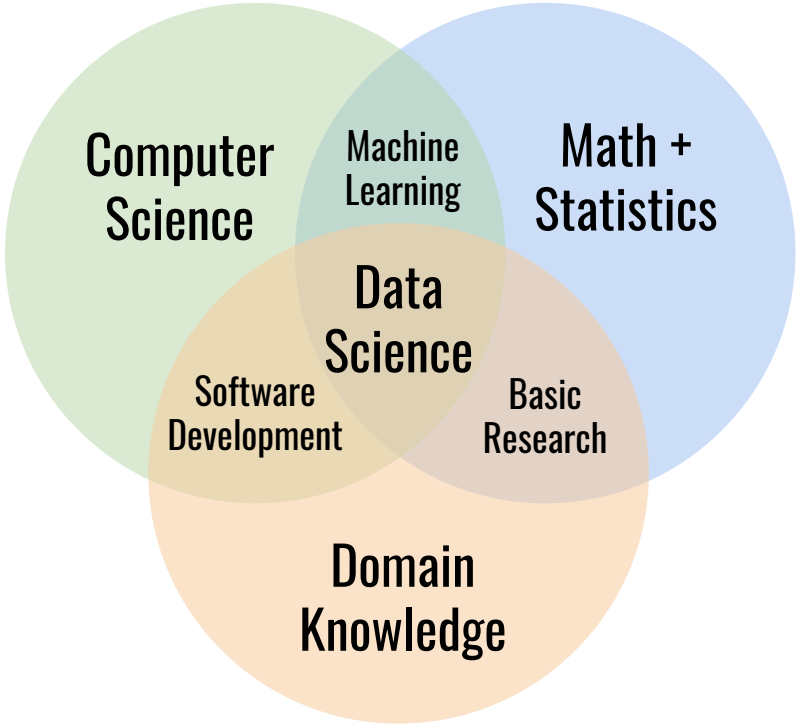
Modern Data Science

- a Data Scientist is the unicorn → highly desirable job category
- Dr. Wu's definition of a Data Scientist is outdated

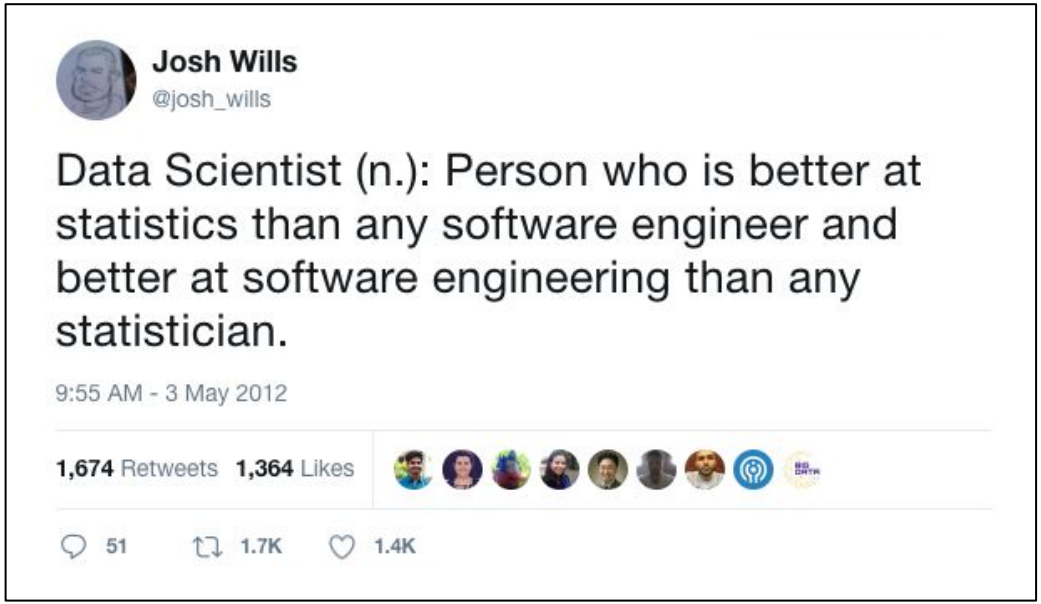


Modern Data Science

- a Data Scientist is the unicorn → highly desirable job category
- **Jobs**
  2018 data science jobs in USA
  > 490,000 (McKinsey & Co.)
- **Job openings**
  up to 1.5 million (McKinsey & Co.)
- **Salary:**
  $116,000 - $163,500 in 2017 (Forbes)



So What is a Data Scientist?



So What is a Data Science?

- a set of tools and techniques used to extract useful information from data
- an interdisciplinary, problem-solving-oriented subject
- the application of scientific techniques to practical problems

which industries use Data Science?


travel

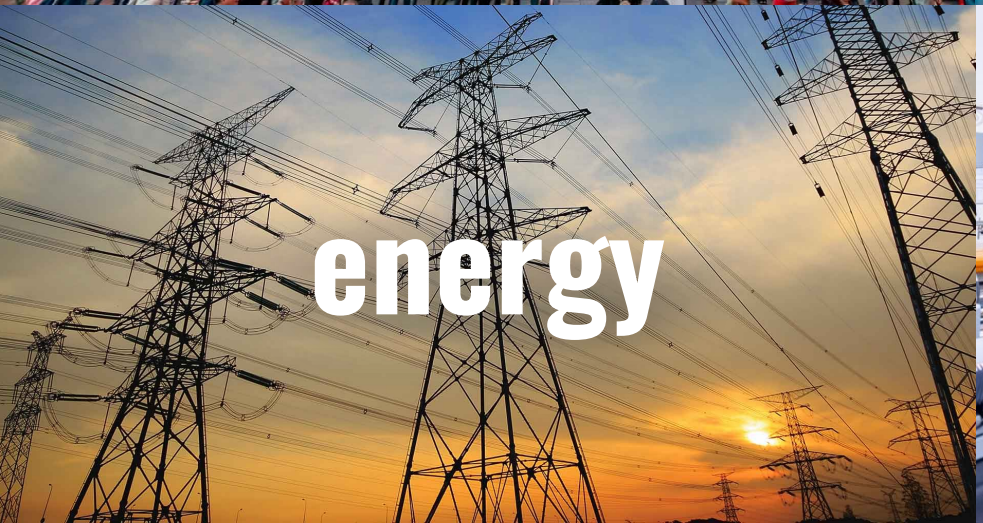

government


gaming


finance

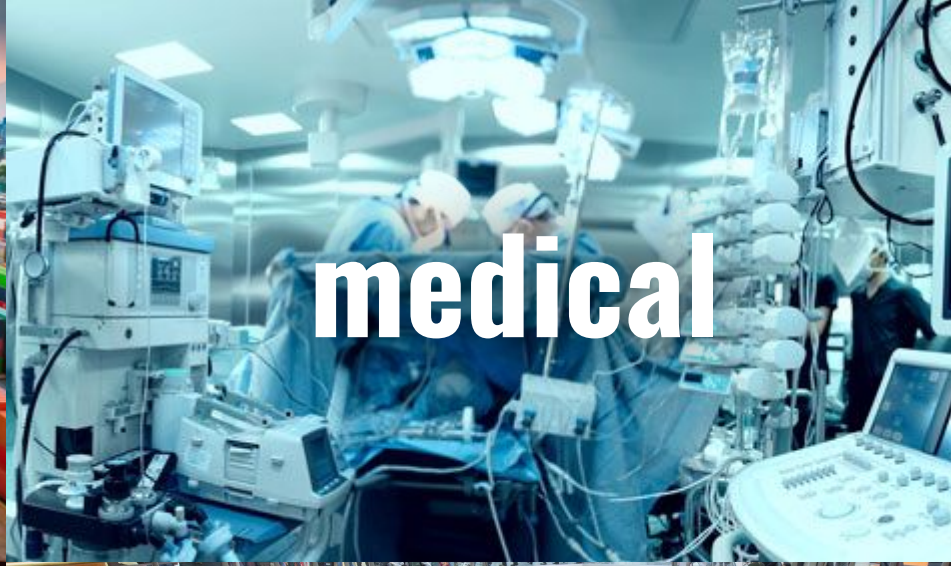

marketing


biotech


energy


manufacturing


shipping


medical

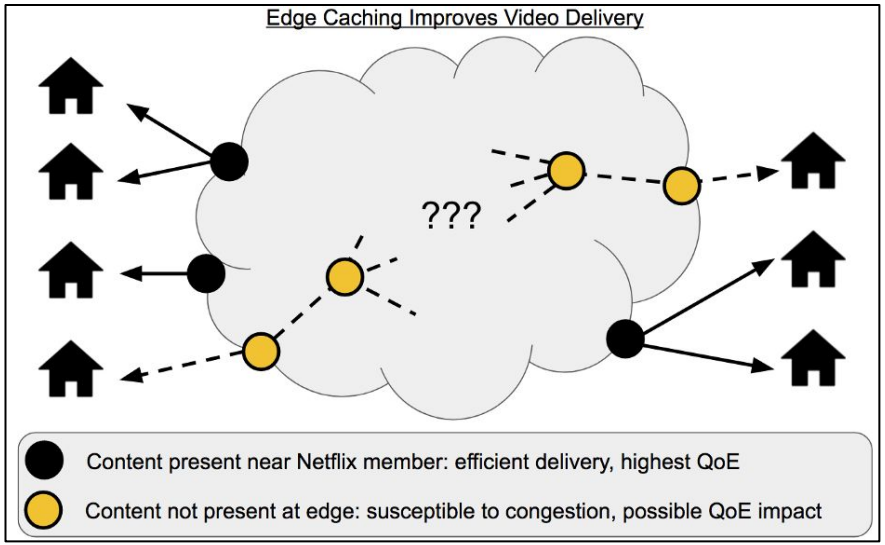

legal


retail

which industries don't use Data Science?

How Data Science is Used – Netflix

- Content Delivery Network (CDN)
- recommended Movies/Shows based on user behavior
- viewership categorization for current programming used to determine new content



How Data Science is Used – IBM Watson

- answering questions posed in natural language
  - beat best human Jeopardy players in 2011
  - clinical decision support systems
  - teaching assistance

How Data Science is Used – DataKind

"Mobile phones, sensors, and new software have created an abundance of data that can be mined, understood, and harnessed to gain new insights about our world and transform almost every sector. The same algorithms and techniques that companies use to boost profits can be leveraged by mission-driven organizations to improve the world, from battling hunger to advocating for child well-being and more."

– DataKind

How Data Science is Used – DataKind

- predicting wheat rust in Ethiopia from satellite imagery
- how to reduce traffic fatalities to zero in cities
- using satellite imagery to find villages in need in Kenya and Uganda
- Using Open Data to Uncover Potential Corruption

How Data Science is Used – Elevators

- ThyssenKrupp Elevator uses thousands of sensors (IoT) to pre-emptively repair elevators
- cloud connected elevators using Microsoft Azure Intelligent Systems Service (Azure ISS).
- ‘smart’ elevators teach technicians how to fix them, using probabilistic error reporting

Data Science vs Data Analytics

DATA SCIENCE	DATA ANALYTICS
predictive / prescriptive	descriptive
why did it happen?	what happened?
visualizing and communicating findings	
programming	using pre-built tools


from watson.di import ContainerAware
from watson.events import types
from watson.framework import events
from watson.http.messages import Response, Request
from watson.common.imports import get_qualified_name
from watson.common.contextmanagers import suppress
ACCEPTABLE_RETURN_TYPES = (str, int, float, bool)
class Base(ContainerAware, metaclass=abc.ABCMeta):
"""The base class for all controllers.
Attributes:
__action__ (string): The last action that was called on the
"""
def execute(self, **kwargs):
method = self.get_execute_method(**kwargs)
self.__action__ = method
return method(**kwargs) or {}

WHY PYTHON?

The Koan of Python

1. open your shell:
   - terminal
   - Anaconda command prompt
2. ➥ `jupyter notebook`
3. if a jupyter notebook session does not open in your browser automatically, copy the url in the shell and open it in your browser
4. click the new dropdown on the top right and select a Python 2 notebook
5. In the code block 

➥ `import this`

Python is

“a programming language that lets you work quickly and integrate systems more effectively”

– python

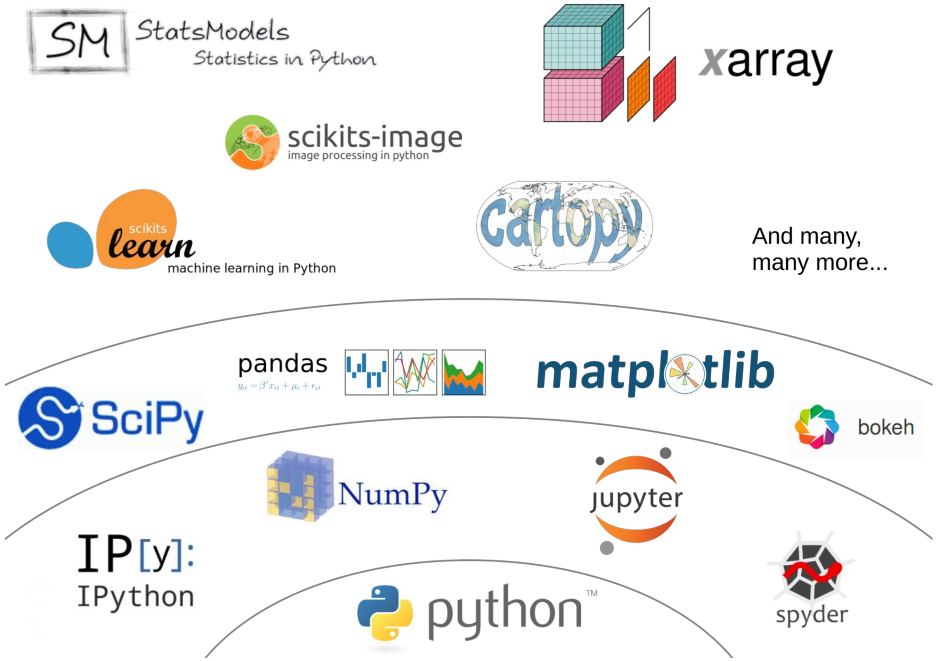
- high-level programming language for general-purpose programming
- everyone’s second favorite language

Python for Data Science

- **huge** dynamic community of developers and users
- competes well with other languages in fields such as
  - basic regression and classification
  - machine learning
  - computer vision
  - natural language processing
- general-purpose language makes python extendable outside of Data Science

Python Scientific Ecosystem

- there are **many** packages
- we will be focusing on:
  - Jupyter Notebook (IPython)
  - NumPy
  - pandas
  - matplotlib
  - scikit-learn
  - StatsModels

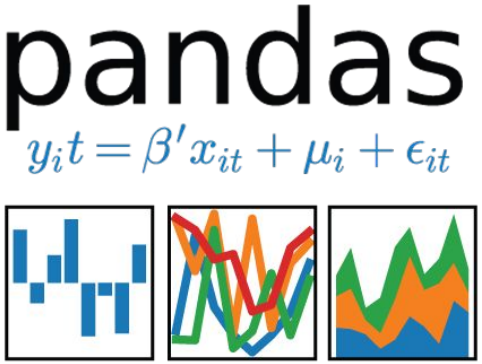


Basic Python Quiz

1. break into groups of 2-3
2. work together to define one the following python terms (**5 min**):
    a. String
    b. List
    c. Function
    d. Boolean
    e. Dictionary
    f. Object
    g. Module
    h. Class
3. present your definition to the class

download and extract code demos
(see link for instructions)
https://goo.gl/3ZfQcR

pandas



- ➤ allows manipulation of excel table-like data*
- ➤ sophisticated methods for complex data restructuring
- ➤ basically, its' python data manipulation on steroids

1. in your shell, use `cd` to navigate to to the extracted files folder
2. ➥ `jupyter notebook`
3. in **jupyter notebook**, open "demos/pandas.ipynb"

WATCH / CODE ALONG

numpy

- **fast** multi-dimensional arrays
- useful linear algebra, Fourier transform, and random number capabilities

1. in **jupyter notebook**, open “demos/numpy.ipynb”

WATCH / CODE ALONG

matplotlib

- 2D plotting library for producing publication / presentation quality figures
- MATLAB-like interface

1. in **jupyter notebook**, open “demos/matplotlib.ipynb”

WATCH / CODE ALONG

estimator
multiple
values
statement
when
random
default
mixture
residual
examples
equal
binary
population
different
across
names
options
set
data
how
using
intercepts
describes
estimated
title
they
carlo
logistic
symbol
monte
starting
first
multilevel
where
integration
latent
used
errors
covariates
format
referred
missing
type
efa
test
list
available
all
variances
new
covariance
covariate
ordered
specified
muth
sets
must
observed
slope
parameter
between
option
if
survival
more
file
command
contains
likelihood
commands
outcomes
these
overall
means
found
because
label
name
will
free
part
specify
value
continuous
models
each
standard
variables
factor
number
chapter
fixed
have
parameters
select
example
second
numerical
algorithm
measured
information
other
outcome
intercept
censored
sample
maximum
modeling
estimation
from
classes
factors
dependent
analysis
above
explanation
conjunction
count
zero
within
group
class
growth
time

MODELING

To Build a Model

"all models are wrong but some are useful"

– George Box, statistician

- models make a useful approximation of real world phenomena by attempting to explain / structure the variability in the data
- statistical models leverage the power of probability and mathematics to make informative predictions (inferences) based on pre-existing data
- models can be trained; they can learn your data

What Makes a Model?

based on some value(s) of feature(s) $x \rightarrow$ predict outcome y

x (input)

- independent variable
- also called:
  - predictor, regressor, explanatory variable, feature

y (output)

- dependent variable
- also called
  - response, regressand, explained variable, outcome, label

What Makes a Model?

$$y = f(x_1, x_2, x_3 \ldots)$$

x (input)

- independent variable
- also called:
  - predictor, regressor, explanatory variable, feature

y (output)

- dependent variable
- also called
  - response, regressand, explained variable, outcome, label

Classes of Models Based on Outcome

	CONTINUOUS	CATEGORICAL
SUPERVISED	regression	classification
UNSUPERVISED	dimensionality reduction	clustering

- continuous variable: one that can take on infinitely many, uncountable values
- categorical variable: one that can take on one of a limited, and usually fixed, number of possible values

Classes of Models Based on Outcome

	CONTINUOUS	CATEGORICAL
SUPERVISED	regression	classification
UNSUPERVISED	dimensionality reduction	clustering

- supervised learning: training a model where the outputs exist for sets of inputs
- unsupervised learning: training a model where outputs do not exist for the inputs

Classes of Models Based on Outcome

CONTINUOUS	CATEGORICAL
regression	classification
dimensionality reduction	clustering
quantitative	*qualitative*

Classes of Models Based on Outcome

CONTINUOUS	CATEGORICAL
regression	classification
dimensionality reduction	clustering
quantitative	*qualitative*

based on hours worked vs sales for the past year, we want to predict sales for any value of hours worked

Classes of Models Based on Outcome

CONTINUOUS	CATEGORICAL
regression	classification
dimensionality reduction	clustering
quantitative	*qualitative*

based on patient health care history data, we want to predict whether a patient will live past 80

Classes of Models Based on Outcome

making predictions

SUPERVISED	regression	classification
UNSUPERVISED	dimensionality reduction	clustering

extracting structure

Classes of Models Based on Outcome

making predictions → *generalizations*

SUPERVISED	regression	classification
UNSUPERVISED	dimensionality reduction	clustering

extracting structure → *representations*

Classes of Models Based on Outcome

making predictions → *generalizations*

SUPERVISED	regression	classification
UNSUPERVISED	dimensionality reduction	clustering

extracting structure → *representations*

training a model to predict acceptance into a graduate school based on ranking of undergraduate school using previous admissions data

Classes of Models Based on Outcome

making predictions → *generalizations*

SUPERVISED	regression	classification
UNSUPERVISED	dimensionality reduction	clustering

extracting structure → *representations*

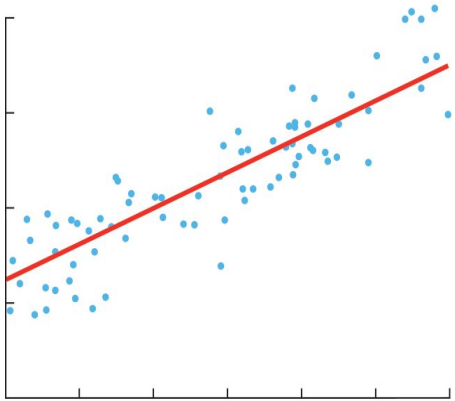
training a model to determine whether or not a picture contains a cat by feeding it labeled pictures of cats and non-cats

Model Type Quiz

1. break into groups of 2-3
2. figure out which of the four model types each model below is (**5 min**):
    a. we have a huge customer data set and we want to identify sub-groups of customers that may have similar purchase, useage, or other tendencies
    b. we want to train a computer vision algorithm to detect the species of plant from a silhouette of a leaf
    c. we want to estimate future rain accumulation per storm based on this past years data about storm rain accumulation

Regression

SUPERVISED CONTINUOUS



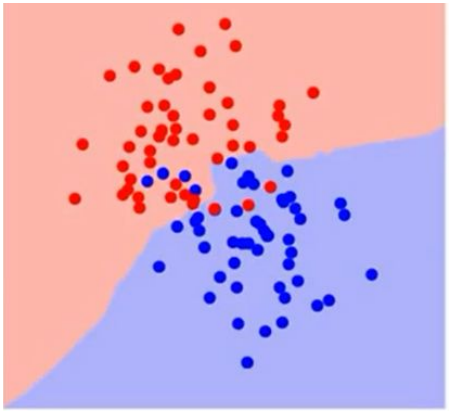
- build a model that best fits data
- predict a continuous value
- types:
  - Linear Regression, Logistic Regression...

1. in **jupyter notebook**, open “demos/regression.ipynb”

WATCH / CODE ALONG

Classification

SUPERVISED CATEGORICAL



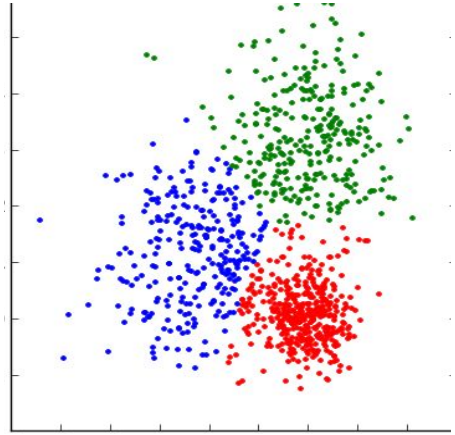
- map feature per input to categorical target classes
- types:
  - K-Nearest Neighbors, Naïve Bayes, Random Forest...

WATCH VIDEO ON KNN

K-NEAREST NEIGHBORS

Clustering

UNSUPERVISED CATEGORICAL



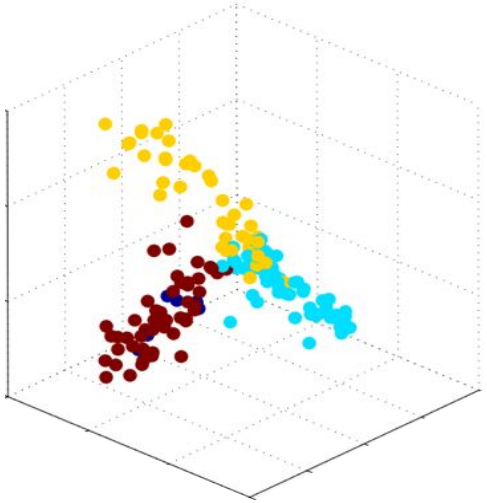
- Similar to classification, except no labels/classes provided
- finds common threads that a human couldn't see
- types:
  - K-Means, Density-Based, etc.

1. in **jupyter notebook**, open "demos/clustering.ipynb"

WATCH / CODE ALONG

Dimensionality Reduction

UNSUPERVISED CONTINUOUS



- not all features contribute meaningfully to data variability
- collapses features, reduces complexity and computation
- types:
  - Principal Component Analysis (PCA)...

WATCH VIDEO ON PCA

Most people will choose the top position

Why & How ?


too many models?

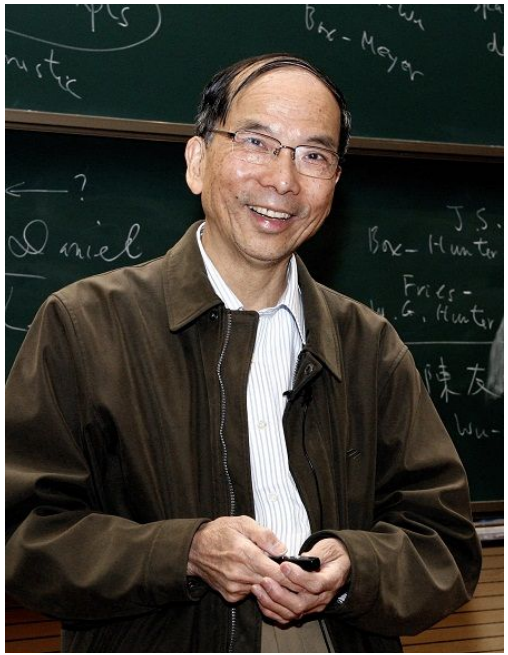
Model Selection

- it takes training, experience and talent to choose the best model(s) for your problem and set the up correctly
- sometimes the hardest step is to figure out the question you want to ask first!
- a model is pretty much useless if the question is badly formed

DATA SCIENCE PIPELINE

Data Science Workflow ...

- according to Dr. Wu in 1997 ...



Chien-Fu Jeff Wu
Data Scientist ~~Statistician~~

… A Modern Version

- a more granular approach …

Data Science Workflow

1. **IDENTIFY** the problem
2. **ACQUIRE** the data
3. **PARSE** the data
4. **MINE** the data
5. **REFINE** the data
6. **BUILD** a data model
7. **PRESENT** the results

Data Science Workflow

IDENTIFY THE PROBLEM

- [ ] Identify business/product objective
- [ ] Identify and hypothesize goals and criteria for success
- [ ] Create a set of questions for identifying correct data set

Data Science Workflow

ACQUIRE THE DATA

- [ ] Identify the “right” data set(s)
- [ ] Import data and set up local or remote data structure
- [ ] Determine most appropriate tools to work with data

Data Science Workflow

PARSE THE DATA

- [ ] Read any documentation provided with the data
- [ ] Perform exploratory data analysis
- [ ] Verify the quality of the data

Data Science Workflow

MINE THE DATA

- [ ] Determine sampling methodology and sample data
- [ ] Format, clean, slice, and combine data in Python
- [ ] Create necessary derived columns from the data (new data)

Data Science Workflow

REFINE THE DATA

- [ ] Identify trends and outliers
- [ ] Apply descriptive and inferential statistic
- [ ] Document and transform data

Data Science Workflow

BUILD A DATA MODEL

- [ ] Select appropriate model
- [ ] Build model
- [ ] Evaluate and refine model

Data Science Workflow

PRESENT THE RESULTS

- [ ] Summarize findings with narrative, storytelling techniques
- [ ] Present limitations and assumptions of your analysis
- [ ] Identify follow up problems and questions for future analysis

RECAP & FUTURE GROWTH

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

- understand what makes a Data Scientist
- understand where Data Science is used
- understand how Data Science is used

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

- understand why Python is ideal for doing Data Science
- learn about scientific packages commonly used in python
- use a few packages

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

- define what a statistical model is
- understand how different models are classified
- run a few models and visualize results

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

- understand how to approach a data science problem from start to finish in a stepwise manner

Agenda

1. WHAT IS DATA SCIENCE?
2. WHY PYTHON?
3. MODELING
4. DATA SCIENCE PIPELINE
5. RECAP & FUTURE GROWTH
6. Q&A

OBJECTIVES

- recap important concepts from workshop
- learn about resources, classes, bootcamps to further your Data Science ability

Future Growth

- demand for Data Scientists outpaces supply
- there are many great resources for delving deeper:
  - DIY
    - Kaggle competitions
    - play with public data
  - Bootcamps
    - Data Science Immersive (General Assembly)
    - Data Science Part-Time (General Assembly)
  - online resources
    - Khan Academy, Udemy, Codecademy

Thanks To ...

TIM HOGAN
Data Science Instructor
General Assembly
Boston

JONATHAN TABLADA
Marketing Lead
General Assembly
Boston

ALL OF YOU
this was my first time teaching this workshop!

Q&A

please fill out an Exit Ticket
(see link for instructions)
https://goo.gl/3ZfQcR