GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LESSON 1: DATA EXPLORATION

WELCOME!

Instructors:

Jonathan Haski

Classes:

Tuesday, 6:30-9:30

Thursday, 6:30-9:30

Outside of classwork: 5-10 hours a week.

Office Hours: Lets Decide

Happy Hour Next Week?

Contact:

jhaski@gmail.com

I. GOALS OF THE COURSE
II. WHAT IS DATA SCIENCE?
III. THE DATA MINING WORKFLOW

LAB:
IV. PYTHON SETUP
V. WORKING IN UNIX AND PYTHON

I. GOALS OF THE COURSE

Insight in how to find, read, and understand data

Gain techniques for manipulating data

Learn machine learning algorithms commonly used in practice

GOALS OF THE COURSE

Insight in how to find, read, and understand data

Gain techniques for manipulating data

Learn machine learning algorithms commonly used in practice **Lots and**

lots of practice!

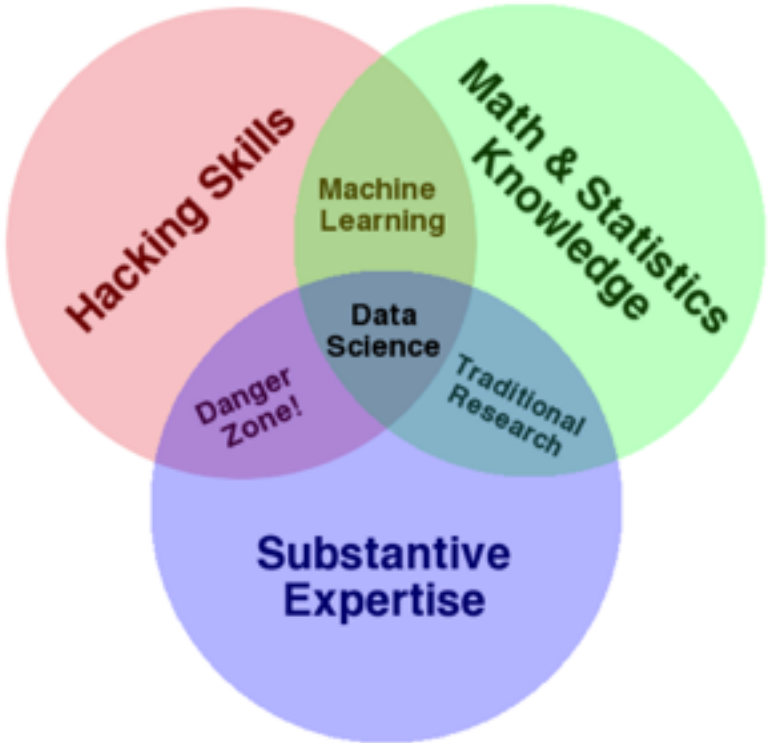
I. WHAT IS DATA SCIENCE?

A set of tools and techniques used to extract useful information from data.

A set of tools and techniques used to extract useful information from data.

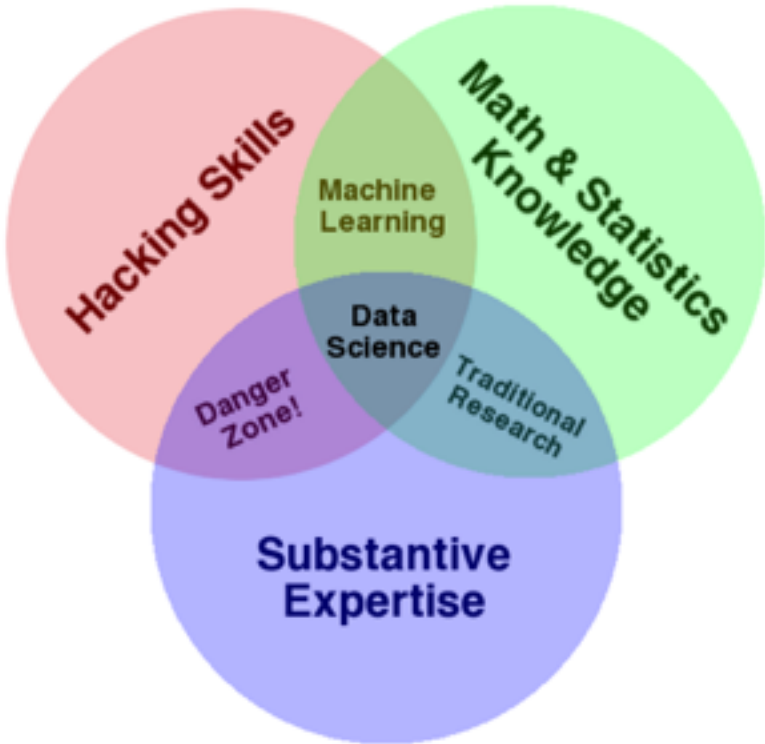
An interdisciplinary, problem-oriented subject.

THE QUALITIES OF A DATA SCIENTIST



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

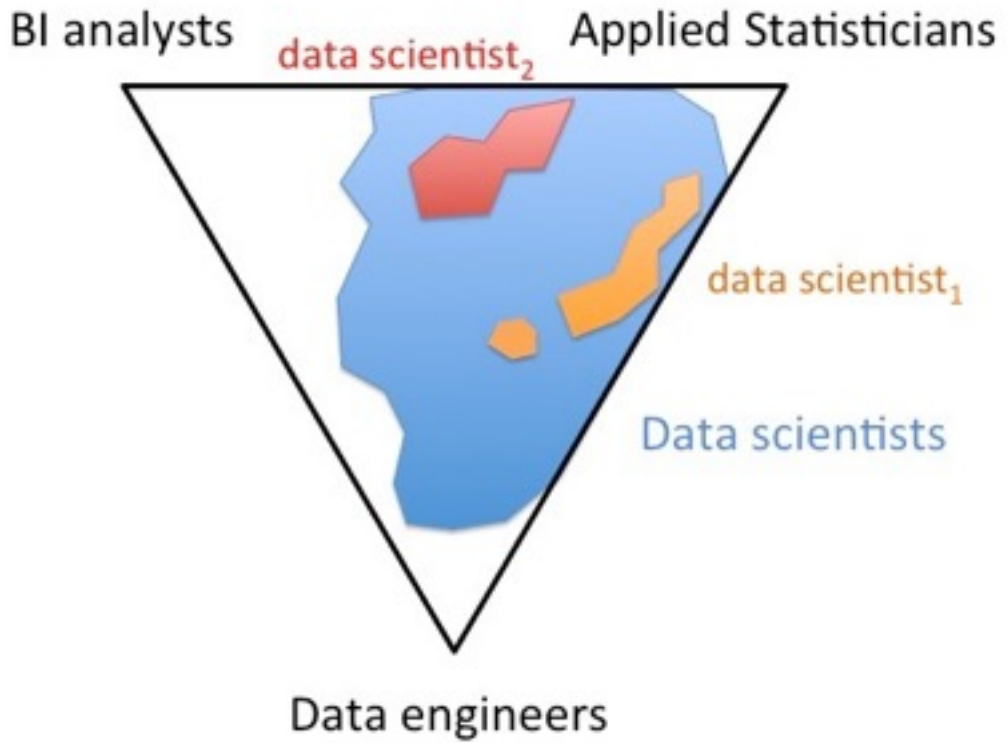
THE QUALITIES OF A DATA SCIENTIST



ONE MORE THING!
Communication skills

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

THE QUALITIES OF A DATA SCIENTIST



source: http://www.p-value.info/2012/12/what-is-data-scientist.html

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.

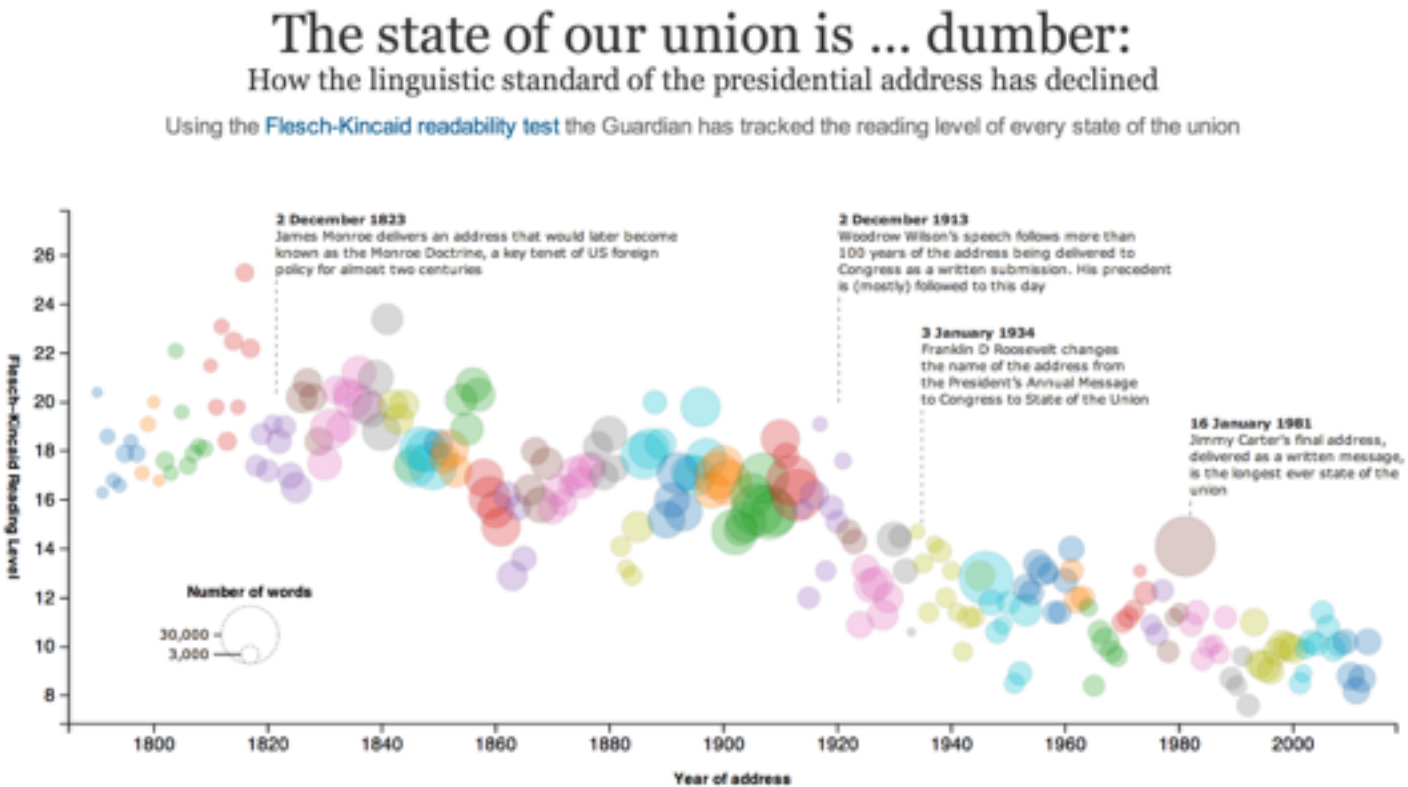
A set of tools and techniques used to extract useful information from data.

An interdisciplinary, problem-solving oriented subject.

The application of scientific techniques to practical problems.

A rapidly growing field.

WHO USES DATA SCIENCE?


The state of our union is ... dumber:
How the linguistic standard of the presidential address has declined
Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every state of the union
2 December 1823
James Monroe delivers an address that would later become known as the Monroe Doctrine, a key tenet of US foreign policy for almost two centuries
2 December 1913
Woodrow Wilson's speech follows more than 100 years of the address being delivered to Congress as a written submission. His precedent is (mostly) followed to this day
3 January 1934
Franklin D Roosevelt changes the name of the address from the President's Annual Message to Congress to State of the Union
16 January 1981
Jimmy Carter's final address, delivered as a written message, is the longest ever state of the union
Number of words
30,000
3,000
Flesch-Kincaid Reading Level
26
24
22
20
18
16
14
12
10
8
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Year of address

- Stack Overflow tag recommendation and response time prediction
- Locating ethnic food in ethnic neighborhoods
- Building optimal fantasy football teams
- Recommending new musical artists
- Identifying key areas to get a taxi in NYC
- Finding the right job for you

Michael E. Driscoll
@medriscoll

Following

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

Reply Retweet Favorite More Pocket

- Statistical and machine learning knowledge
- Computer Science experience (Applied Math)
- Academic curiosity
- Product sense
- Storytelling
- Cleverness

REVIEW

1. What are the leading qualities that make up a data scientist?

2. Name an example of a company that uses data science to help improve their product.

REVIEW

1. Creativity, a statistics and engineering background, wit
2. Amazon: Recommendations to get users to continue making purchases

II. THE DATA SCIENCE WORKFLOW

Dataists

1. Obtain
2. Scrub
3. Explore
4. Model
5. Interpret

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

THE DATA SCIENCE WORKFLOW 26

Jeff Hammerbacher: *Chief Scientist, Cloudera*

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

Ted Johnson: *AT&T Research*

1. Assemble an accurate and relevant data set

2. Choose the appropriate algorithm

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

THE DATA SCIENCE WORKFLOW

Ben Fry: *Principal, Fathom*

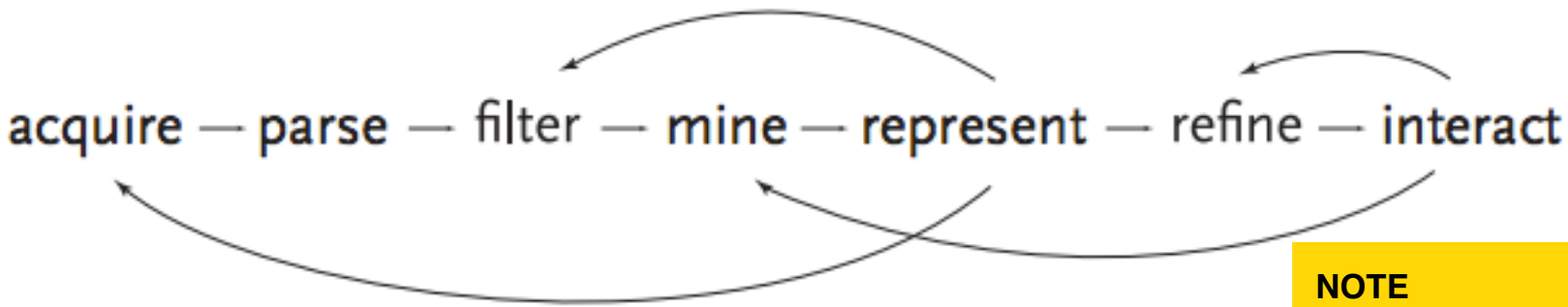
1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

THE DATA SCIENCE WORKFLOW

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

Start with a flexible Question



NOTE

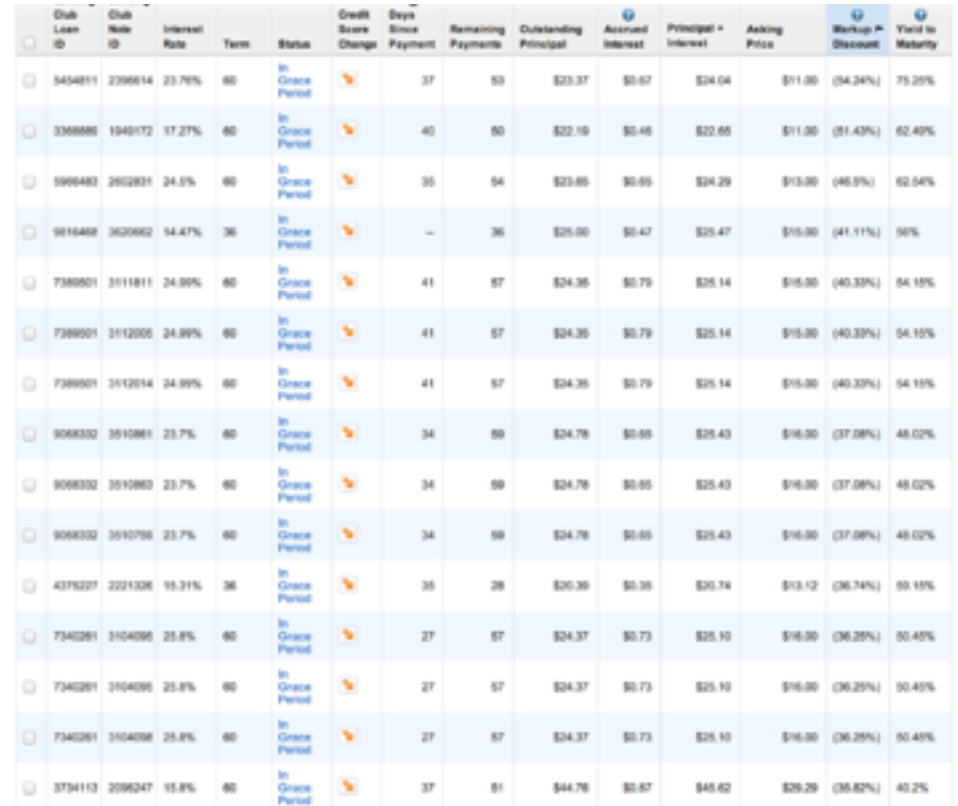
This diagram illustrates the *iterative* nature of problem solving

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

LEVERAGING DATA SCIENCE

LendingClub

Asking a Question
and
Finding the Data



	Loan ID	Note ID	Interest Rate	Term	Status	Credit Score Change	Days Since Payment	Remaining Payments	Outstanding Principal	Accrued Interest	Principal + Interest	Asking Price	Markup / Discount	Yield to Maturity
☐	5454811	2396614	23.76%	60	In Grace Period	↘	37	53	$23.37	$0.67	$24.04	$11.00	(54.24%)	75.25%
☐	3368886	1940172	17.27%	60	In Grace Period	↘	40	50	$22.19	$0.46	$22.65	$11.00	(51.43%)	62.46%
☐	5969483	2802831	24.5%	60	In Grace Period	↘	35	54	$23.65	$0.65	$24.29	$13.00	(46.5%)	62.04%
☐	5816468	3620662	14.47%	36	In Grace Period	↘	--	36	$25.00	$0.47	$25.47	$15.00	(41.11%)	56%
☐	7389501	3111811	24.99%	60	In Grace Period	↘	41	57	$24.35	$0.79	$25.14	$15.00	(40.33%)	54.15%
☐	7389501	3112005	24.99%	60	In Grace Period	↘	41	57	$24.35	$0.79	$25.14	$15.00	(40.33%)	54.15%
☐	7389501	3112014	24.99%	60	In Grace Period	↘	41	57	$24.35	$0.79	$25.14	$15.00	(40.33%)	54.15%
☐	9068332	3510861	23.7%	60	In Grace Period	↘	34	59	$24.78	$0.65	$25.43	$16.00	(37.08%)	48.02%
☐	9068332	3510863	23.7%	60	In Grace Period	↘	34	59	$24.78	$0.65	$25.43	$16.00	(37.08%)	48.02%
☐	9068332	3510759	23.7%	60	In Grace Period	↘	34	59	$24.78	$0.65	$25.43	$16.00	(37.08%)	48.02%
☐	4375227	2221326	15.31%	36	In Grace Period	↘	35	28	$20.39	$0.35	$20.74	$13.12	(36.74%)	59.15%
☐	7340261	3104096	25.8%	60	In Grace Period	↘	27	57	$24.37	$0.73	$25.10	$16.00	(36.25%)	50.45%
☐	7340261	3104095	25.8%	60	In Grace Period	↘	27	57	$24.37	$0.73	$25.10	$16.00	(36.25%)	50.45%
☐	7340261	3104098	25.8%	60	In Grace Period	↘	27	57	$24.37	$0.73	$25.10	$16.00	(36.25%)	50.45%
☐	3734113	2096247	15.8%	60	In Grace Period	↘	37	51	$44.76	$0.87	$45.62	$29.29	(35.82%)	40.2%

LEVERAGING DATA SCIENCE

Acquiring the Data

LEVERAGING DATA SCIENCE

	score	date
0	None	None
1	790-794	July 20, 2011
2	750-754	August 08, 2011
3	745-749	September 08, 2011
4	745-749	October 08, 2011
5	735-739	November 08, 2011
6	715-719	December 27, 2011
7	695-699	January 23, 2012
8	700-704	February 17, 2012
9	690-694	March 18, 2012

Cleaning Data

	score	date
1	792	2011-07-20 00:00:00
2	752	2011-08-08 00:00:00
3	747	2011-09-08 00:00:00
4	747	2011-10-08 00:00:00
5	737	2011-11-08 00:00:00
6	717	2011-12-27 00:00:00
7	697	2012-01-23 00:00:00
8	702	2012-02-17 00:00:00
9	692	2012-03-18 00:00:00

LEVERAGING DATA SCIENCE

Decision Tree Model

```
1) root 651 319 No (0.50998464 0.49001536)
  2) days_last_collection< 36.56405 92  15 No (0.83695652 0.16304348) *
  3) days_last_collection>=36.56405 559 255 Yes (0.45617174 0.54382826)
    6) bankrupt=True 27   0 No (1.00000000 0.00000000) *
    7) bankrupt=False 532 228 Yes (0.42857143 0.57142857)
     14) always_current=False 124  52 No (0.58064516 0.41935484)
       28) days_last_collection< 84.56405 79  22 No (0.72151899 0.27848101) *
       29) days_last_collection>=84.56405 45  15 Yes (0.33333333 0.66666667) *
     15) always_current=True 408 156 Yes (0.38235294 0.61764706)
       30) payments_bin=[0, 1),[1, 2),[3, 4),[5, 6) 58  22 No (0.62068966 0.379310
         60) days_last_collection< 70.56405 38   9 No (0.76315789 0.23684211) *
         61) days_last_collection>=70.56405 20   7 Yes (0.35000000 0.65000000) *
       31) payments_bin=[10, 20),[2, 3),[20, 30),[30, 60),[4, 5),[6, 7),[7, 8),[8, 9),[9, 10) 350 120 Yes
(0.34285714 0.65714286)
         62) days_since_payment>=36.5 146  64 Yes (0.43835616 0.56164384)
          124) collections_count< 4.5 94  46 No (0.51063830 0.48936170)
            248) loangrade=A,C 22   6 No (0.72727273 0.27272727) *
            249) loangrade=B,D,E,F,G 72  32 Yes (0.44444444 0.55555556)
              498) days_since_payment>=40.5 25  10 No (0.60000000 0.40000000) *
              499) days_since_payment< 40.5 47  17 Yes (0.36170213 0.63829787)
                998) remaining_pay< 51.5 35  16 Yes (0.45714286 0.54285714)
                 1996) loanrate>=18.24 17   5 No (0.70588235 0.29411765) *
                 1997) loanrate< 18.24 18   4 Yes (0.22222222 0.77777778) *
                999) remaining_pay>=51.5 12   1 Yes (0.08333333 0.91666667) *
          125) collections_count>=4.5 52  16 Yes (0.30769231 0.69230769) *
         63) days_since_payment< 36.5 204  56 Yes (0.27450980 0.72549020) *
```

Analyze and Apply

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?
3. Examine data to find common distributions and correlations

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?
3. Examine data to find common distributions and correlations
4. Extract new meaning to predict if a user would purchase again or not

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
    1. How many times did a user share through Facebook within a week? A month?
    2. How often did they open up our emails?
3. Examine data to find common distributions and correlations
4. Extract new meaning to predict if user would purchase again
5. Share results (and probably also go back to the drawing board)

REVIEW

1. What seem to be the most common practical steps in the data science workflow?

2. Is the workflow straightforward? Why or why not?

REVIEW

1. Collect data, explore data, create a model, share the results

2. Usually, no! There will always be a need to collect more data and improve the original model.

IV. PYTHON SETUP

High variety of languages used in practice

Statistics: “Python,” R, Matlab, Julia, Fortran, STATA
Scripting: Python, Ruby, Scala, Java
Data Querying: SQL, Hive, Pig

Python is an open source project which is maintained by a large and very active community.

Python is an open source project which is maintained by a large and very active community.

It was originally created by Guido Van Rossum in the 1990s, who currently holds the title of Benevolent Dictator For Life (BDFL).

The presence of a BDFL means that Python has a *unified design philosophy*.

This design philosophy emphasizes *readability* and *ease of use*, and is codified in PEP8 (the Python style guide) and PEP20 (the Zen of Python).

NOTE

PEPs (or Python Enhancement Proposals) are the public design specs that the language follows.

Python in nature is not a statistical language, though used in a variety of ways: web applications, server maintenance, reading and writing text files:

web development https://www.djangoproject.com/
systems admin http://docs.fabfile.org/en/1.6/
(etc) https://github.com/languages/Python

Python evolved alongside Bioinformatics and Data Analysis, introducing stats packages (numpy, scipy) and machine learning packages (scikit-learn, NLTK)

STRENGTHS & WEAKNESSES

ADVANTAGES

- VERY FAST COMPARED TO R
- USEFUL ACROSS PLATFORMS
- EASY TO INTEGRATE
- COMMON OOP TECHNIQUES
- GREAT DOCUMENTATION SUPPORT

DISADVANTAGES

- NO GREAT VISUALIZATION PACKAGES (YET!)
- NATURAL DISPLAY IS LESS READABLE
- LESS NEWBIE FRIENDLY
- LACK OF PARALLEL PROCESSING

III. PYTHON DATA STRUCTURES

The most basic data structure is the None type. This is the equivalent of NULL in other languages.
There are four numeric types: **int, float, bool, complex.**

```
>>> type(1)
<type 'int'>
>>> type(2.5)
<type 'float'>
>>> type(True)
<type 'bool'>
>>> type(2+3j)
<type 'complex'>
```

The next basic data type is the Python list.

A list is an *ordered* collection of elements, and these elements can be of arbitrary type. Lists are mutable, meaning they can be changed in-place.

```
>>> k = [1, 'b', True]
>>> k[2]
True
>>> k[1] = 'a'
>>> k
[1, 'a', True]
```

Likewise, **tuples are immutable arrays of arbitrary elements.**

```
>>> x = (1, 'a', 2.5)
>>> x
(1, 'a', 2.5)
>>> x[0]
1
>>> x[0] = 'b'
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: 'tuple' object does not support item assignment
```

Tuples are frequently used behind the scenes in a special type of variable assignment called tuple packing/unpacking.

The string type in Python represents an immutable ordered array of characters (note there is no char type).

Strings support slicing and indexing operations like arrays, and have many other string-specific functions as well.

String processing is one area where Python excels.

Associative arrays (or hash tables) are implemented in Python as the dictionary type.

```
>>> this_class = {'subject': 'data science', 'instructor': 'jason', 'time': 1800, 'is_cool': True}
>>> this_class['subject']
'data science'
>>> this_class['is_cool']
True
```

Dictionaries are unordered collections of key-value pairs, and *dictionary keys must be immutable*.

Another basic Python data type is the set. Sets are unordered mutable collections of distinct elements.

```
>>> y = set([1,1,2,3,5,8])
>>> y
set([8, 1, 2, 3, 5])
```

These are particularly useful for *checking membership* of an element and for ensuring element *uniqueness*.

LAB: UNIX AND PYTHON

IN CLASS WORK

1. Change our python script to also return minimum, maximum, and average age, and click through rate (clicks/impressions)

2. Homework: Update the script to write a new file instead of using standard out and save it to output

DISCUSSION

FOR NEXT TIME:

1. Finish the homework
2. Review Python