GENERAL ASSEMBLY

DATA SCIENCE 101

Sri Kanajan

LEARNING OBJECTIVES

- Explain the field of data science, defining common roles & trends.
- Explore popular tools & resources to visualize, analyze, & model data.
- Recognize the types of problems that can be solved by data science.
- Apply the data science workflow to provide real world recommendations.

DATA SCIENCE 101

PRE-WORK

PRE-WORK REVIEW

- Bring a laptop with Anaconda installed. Scroll to your operating system version and click on the install button for Anaconda with Python 2.7.

- We will be using Jupyter Notebooks as the main IDE for the workshop. If you have installed Anaconda, then you are ready to go!

DATA SCIENCE 101

OPENING

ABOUT YOU

- Before we dive in, let’s talk a bit about you!

- Name
- Why are you doing this class and what specifically do you expect out of this class?

ABOUT ME

- Welcome to Data Science 101!

- Here's a bit about me:

Pivoted out of a 12 year systems engineering and management career with little technical skills and reinvented myself from scratch to become a data scientist. Now a senior data scientist in the banking industry.

OUR EXPECTATIONS

- You're ready to take charge of your learning experience.
- You're curious and excited about data science!
- You've installed Anaconda with Python 2.7.

THE BIG PICTURE

- What we'll cover:
  - Why data science & what it can do for me?
  - Data science skills
  - Explore the Data Science Toolkit
  - Analyse data
  - Algorithms in action

THE BIG PICTURE

- Why this topic matters:
  - Data science is a sought-after skill
  - Using Python due to its increased popularity and simplicity

- Why this topic rocks:
  - Data science is now revolutionizing everything around us due to proliferation of machine data and cloud storage and computing

INTRODUCTION

WHAT IS DATA SCIENCE AND WHAT CAN IT DO FOR ME?

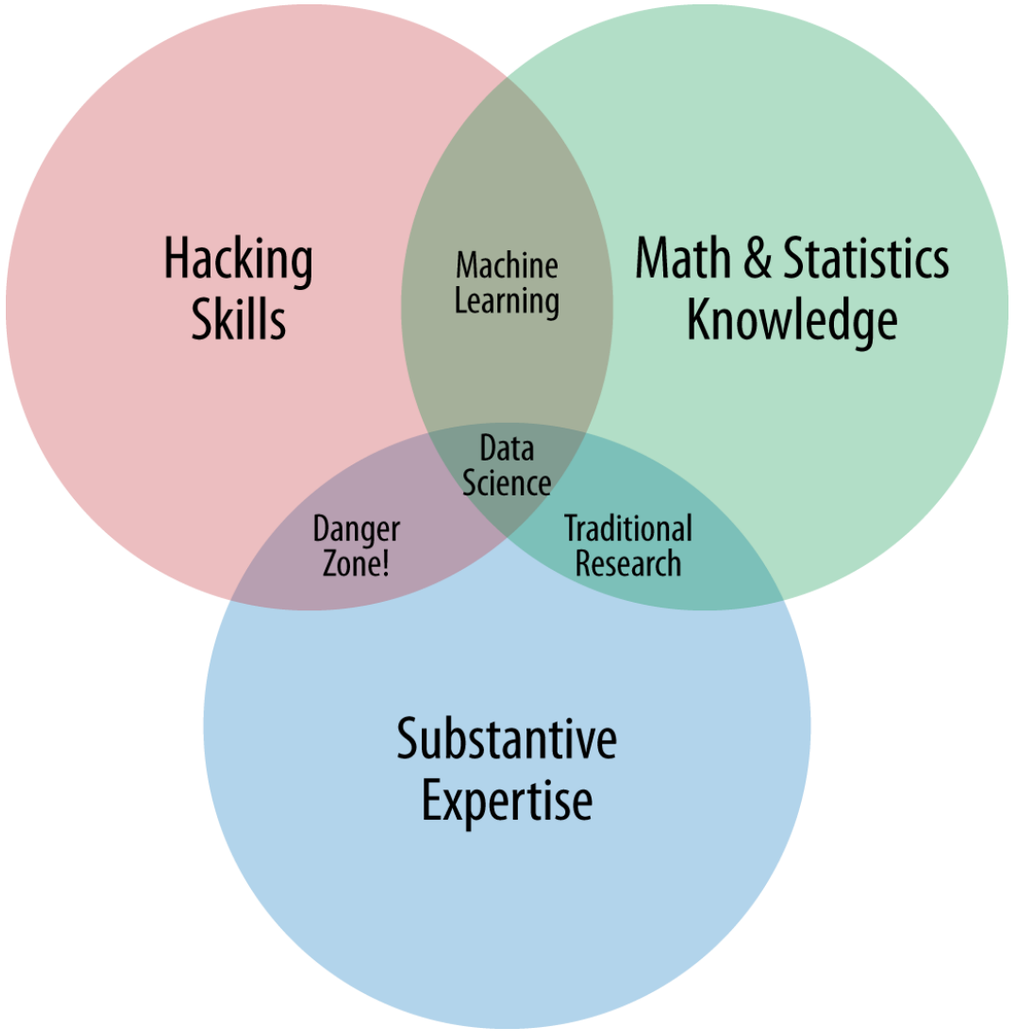
WHAT IS DATA SCIENCE?

THE SEXIEST JOB OF THE 21ST

- **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.
  - A rapidly growing field.

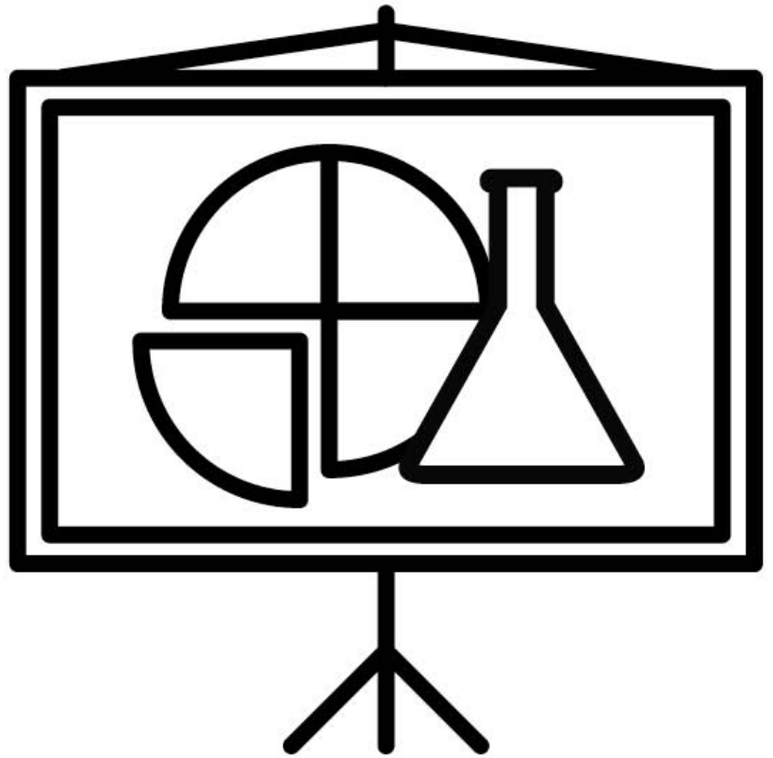
QUALITIES OF A DATA SCIENTIST

- Programming skills
- Math and Statistics knowledge
- Business acumen (substantive expertise)
- Plus: Communication skills



WHAT DOES DATA SCIENCE INVOLVE

- Scrape, munge, & sample business relevant data.
- Manipulate, sanitize, and wrangle data.
- Visualize data.
- Understand data relationships.
- Tell the machine how to learn and predict from data.
- Create data products that deliver actionable insight.
- Tell relevant business stories from data.



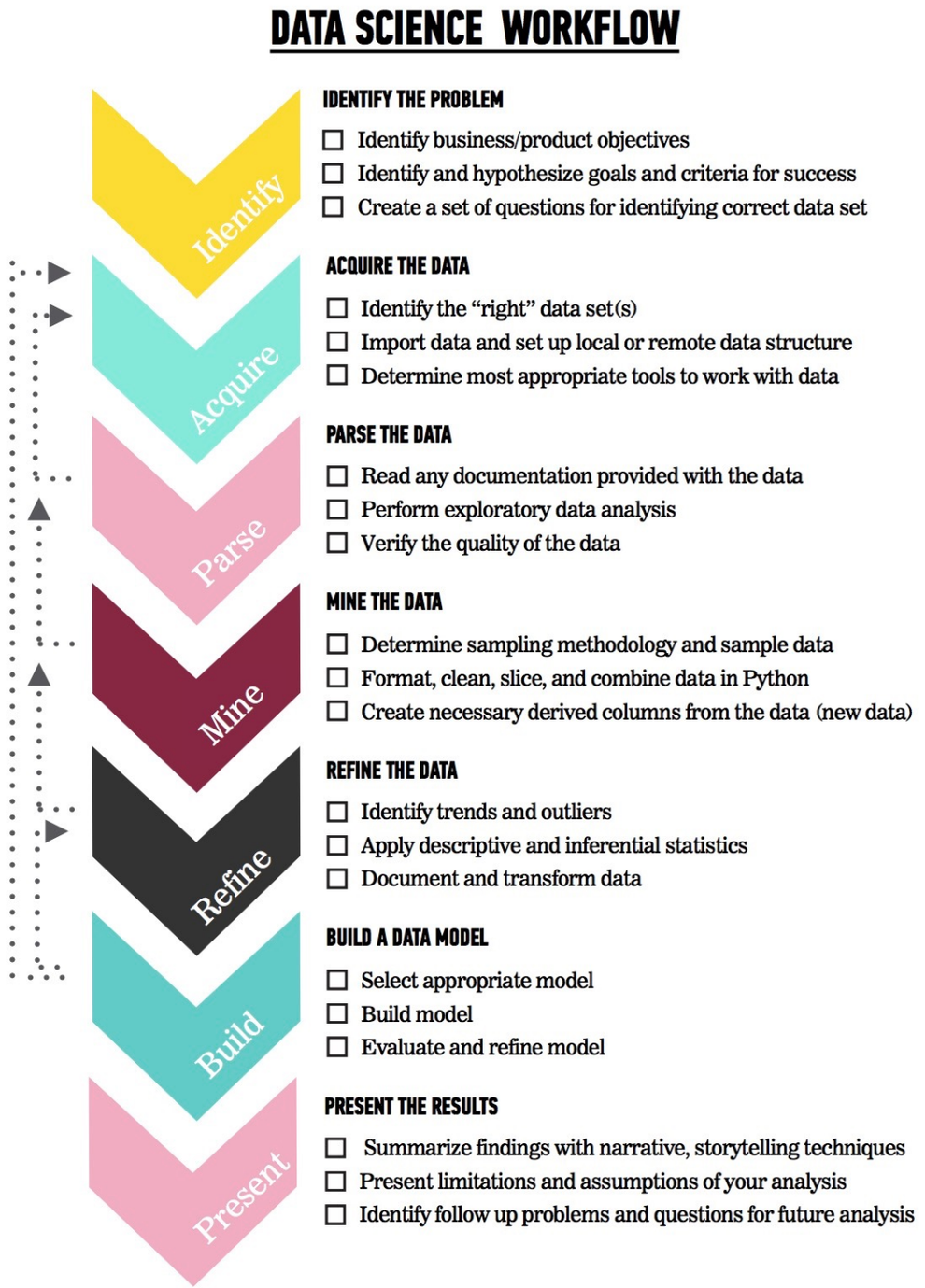
DEMO

VISUALIZING THE DATA SCIENCE WORKFLOW

THE DATA SCIENCE WORKFLOW

MAIN PHASES

- Identify the problem
- Acquire the data
- Parse the data
- Mine the data
- Refine the data
- Build a data model
- Present the results



GUIDED PRACTICE

EXPLORING THE DATA SCIENCE TOOLKIT

WHY PYTHON?

Python is:

- Great for rapid prototyping and full-stack commercial applications.
- A **modern**, elegant, object-oriented language.
- Highly **expressive**, i.e., you can be more **productive**.
- Well documented and has an established and **growing community.**
- Comes with **"batteries included**" **-** in other words, Python has libraries that will help you do a ton of different tasks!

PACKAGES

- Libraries of code written to solve particular set of problems
- Can be installed with: `conda install <package name>`

- Ever used Excel? How would you like working with data structured in a similar way, but without the irritation of formatting, long formula, and better graphics?
  - Try **pandas!**

- Does your application require the use of advanced mathematical functions or numerical operations with arrays, vectors or matrices?
  - Try **SciPy** (scientific Python).
  - Try **NumPy** (numerical Python).

PACKAGES

- Are you interested in using Python in a data science workflow and exploit the use of machine learning in your applications
  - Look no further than **Scikit-learn**.

- Are you tired of the boring-looking charts produced with Excel? Are you bored of looking for the right menu to move a label in your plot?
  - Take a look at the visuals offered by **matplotlib**.

- Is your boss asking about significance testing and confidence intervals? Are you interested in descriptive statistics, statistical tests, plotting functions, and result statistics?
  - Well, **statsmodels** offers you that and more.

- All the data you require is available freely on the web but there is no download button and *you* need to scrape the website?
  - You can extract data from HTML using **Beautiful soup**.

INSTRUCTIONS (GITHUB)

- We recommend using a Jupyter notebook for this practice.

To get a hold of the starter code, you'll need to download these materials.

1. Visit this page: https://github.com/kskk02/data-science-101-cwe-materials
2. Click on the "Clone or Download" button, and click "Download ZIP"
3. Unzip the file downloaded in a **known** location in your file system
4. Open Jupyter: Open a terminal
    - **Mac**: Using spotlight search for "Terminal"
    - **Windows**: Click the "Start" button and type "cmd"
    - In the terminal type: `` `jupyter notebook` ``
5. Navigate to the folder where you have saved the file in step 1
6. Open the file from the Jupyter interface
7. Voilà, you are ready to type the commands we will cover below

- In this guided practice we are using a sample dataset, demonstrate how to carry out descriptive analytics using the **pandas** library we introduced above.

INTRODUCTION

ALGORITHMS IN PYTHON

ACTIVITY: WHAT COMES TO MIND WHEN YOU HEAR THE

DIRECTIONS

1. What do you think when you hear the word “algorithm”?
2. Can you give an example?
3. Do you use any algorithms in your every-day-life?

DELIVERABLE

Discussion with the class

ALGORITHM

A SET OF STEPS TO ACCOMPLISH A TASK

- Algorithms need to have their steps in the right order.

- When you write an algorithm, the order of the instructions is very important.

ALGORITHM

A SET OF STEPS TO ACCOMPLISH A TASK

- Would you put on your shoes before you put on your socks?

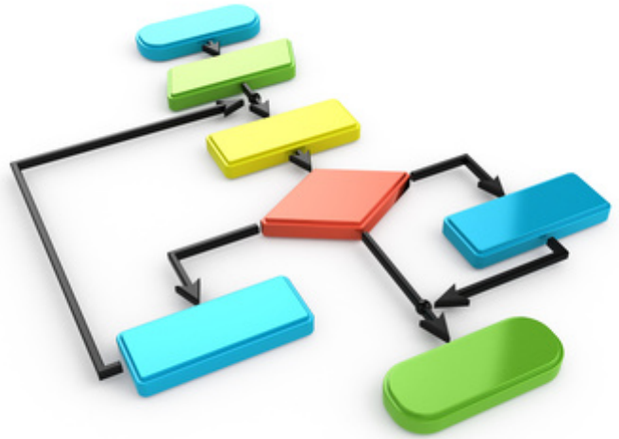
- What if you put on your jacket before you put on your coat?

ALGORITHM

COMPUTER SCIENCE

- Algorithms are a formal way of describing precisely defined instructions.

- Computers are very good at carrying out series of precisely defined instructions.



DEMO

ALGORITHMS IN ACTION

LET US SEE HOW TO WRITE AN

- We will use Python to write our algorithm

Example:

- **Problem**: Given a list of positive numbers, return the largest number on the list.
- **Inputs**: A list ***L*** of positive numbers.
- The list must contain at least one number.

WHAT IS THE OUTPUT

- **Output**: A number ***n***, which will be the largest number of the list.

WHAT IS THE OUTPUT

- **A**LGORITHM

1. Set the variable `max` to 0.
2. For each number `x` in the list `L`, compare it to `max`.
   - If `x` is larger, set `max` to `x`.
3. `max` is now set to the largest number in the list.

HERE IT IS IN PYTHON

Python

```python
def find_max(L):
    max = 0
    for x in L:
        if x > max:
            max = x
    return max
```

ACTIVITY: DISCUSSION…?

DIRECTIONS

1. Does the algorithm meet the criteria below?
    1. It is unambiguous?
    2. Does it have defined inputs and outputs?
    3. Is it guaranteed to terminate?
    4. Does it produce the correct results?
    5. When will the algorithm fail?

DELIVERABLE

Discuss in your group and we will compare with the entire class afterwards.

INDEPENDENT PRACTICE

ANALYZE SOME DATA!

INSTRUCTIONS (DROPBOX)

- We recommend using a Jupyter notebook for this practice.

The Dropbox link provided has a Zip file with the materials for the class.

1. Unzip the file downloaded in a known location in your file system
2. Locate the file called DataScience101_Part1_GuidedPractice.ipynb
3. Open Jupyter: Open a terminal
    - **Mac**: Using spotlight search for "Terminal"
    - **Windows**: Click the "Start" button and type "cmd"
    - In the terminal type: `` `jupyter notebook` ``
4. Navigate to the folder where you have saved the file in step 1
5. Open the file from the Jupyter interface
6. Voilà, you are ready to type the commands we will cover below

- In this guided practice we are using a sample dataset, demonstrate how to carry out descriptive analytics using the **pandas** library we introduced above.

NOW YOU TRY!

FLOWERS AND MORE

- You are a business intelligence manager at a fast moving startup that deals with flowers.

- You need to analyze some data for iris flowers of three different species.

- The business has received a sample data set with typical measures for the following three species for iris flowers…

IRIS DATA SET

- Famous data set analyzed by Ronald Fisher

- 50 samples of 3 different flower types:
  - Setosa
  - Virginica
  - Varsicolor

- 4 features:
  - Sepal: length and width
  - Petal: length and width

- Let us use Python to review some analytics that will help us differentiate these three species.

INSTRUCTIONS

- We recommend using a Jupyter notebook for this practice.

From the materials downloaded:

1. Unzip the file downloaded in a known location in your file system
2. Locate the file called DataScience101_Part1_IndPractice.ipynb
3. Open Jupyter: Open a terminal
    - **Mac**: Using spotlight search for "Terminal"
    - **Windows**: Click the "Start" button and type "cmd"
    - In the terminal type: `` `jupyter notebook` ``
4. Navigate to the folder where you have saved the file in step 1
5. Open the file from the Jupyter interface
6. Voilà, you are ready to type the commands we will cover below

- In this guided practice we are using a sample dataset, demonstrate how to carry out descriptive analytics using the **pandas** library we introduced above.

MACHINE LEARNING

ALGORITHMS IN THE CONTEXT OF MACHINE LEARNING

- Machine learning is a branch of artificial intelligence. It is concerned with the construction and study of systems that can learn from data.

- The core of machine learning deals with representation and generalization.

- **Representation** – extracting a representation of the system that generated the data
- **Generalization** – making predictions from the data

MACHINE LEARNING PROBLEMS

- **Supervised Machine Learning**: Making predictions (generalization)

- For example, suppose you want to predict whether someone will make make a purchase the week after they visit your site.

- You have a set of data on previous customers, including age, interests, previous purchases, time of visit, etc.

- You know whether previous customers made a purchase within a week of their last visit.

- So, the problem is combining all the existing data into a model that can predict

MACHINE LEARNING PROBLEMS

- **Supervised Machine Learning:**

- You can then take action and send a reminder or offer a discount.

- Amazon, Netflix, and others do this based on the history of their existing customers.

- Some examples of supervised learning algorithms include:
  - linear regression
  - decision trees
  - neural networks

MACHINE LEARNING PROBLEMS

- **Unsupervised Machine Learning**: Extracting structure (representation)

- For example, suppose you want to understand your customer base so that you can produce appropriate segments that you can target with your next marketing campaign.

- You have a set of data about your customers, including age, location, previous purchases, time of visit, etc.

- But what characteristics should you use?

MACHINE LEARNING PROBLEMS

- **Unsupervised Machine Learning:**

- Based on these attributes you can find similarities and differences that provide groupings (segments) of customers.

- You can then take action and make an offer or recommend a product specifically to these segments.

- Some unsupervised learning algorithms include:
  - clustering
  - anomaly detection

GUIDED PRACTICE

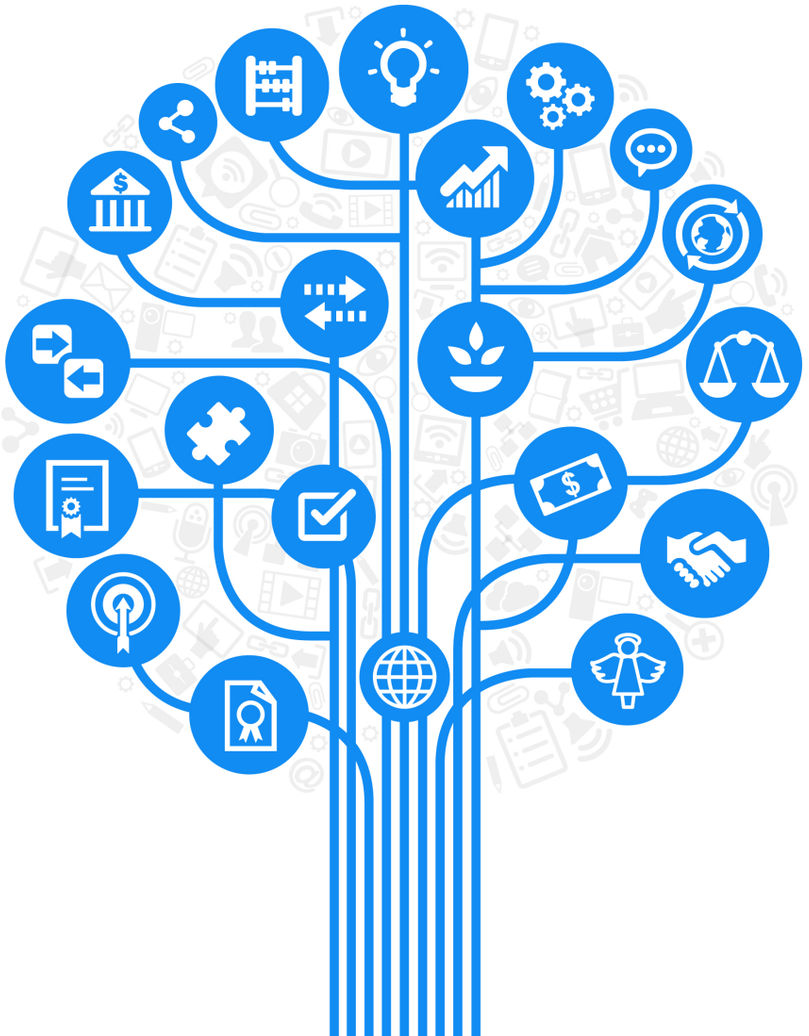
THINKING LOGICALLY

THINKING LOGICALLY

- We just reviewed types of machine learning models at a high level.

- We mentioned decision trees in the context of supervised learning.

    - Do you remember what a supervised learning model is again?
    - Why are they called "trees"?



LET’S APPLY OUR KNOWLEDGE

- During a doctor's examination some patients show the following characteristics:
  - X1: temperature
  - X2: coughing
  - X3: reddening throat

- The doctor has the following outcomes for the patients:
  - Y = {W1, W2, W3, W4, W5}
    - W1: cold
    - W2: tonsilitis
    - W3: flu
    - W4: pneumonia
    - W5: healthy

EXAMPLE

The doctor is required to find a diagnosis based on the symptoms presented by the patient.

In data science terms, the doctor requires a model where `Y` (the diagnosis) depends on `X` (the symptoms). The rules below illustrate such a model:

1. If `X1 < 98` , `Y`= is healthy.

2. If `X1` has values between [98, 102] and `X3`="there is no reddening of throat", then `Y`=cold;

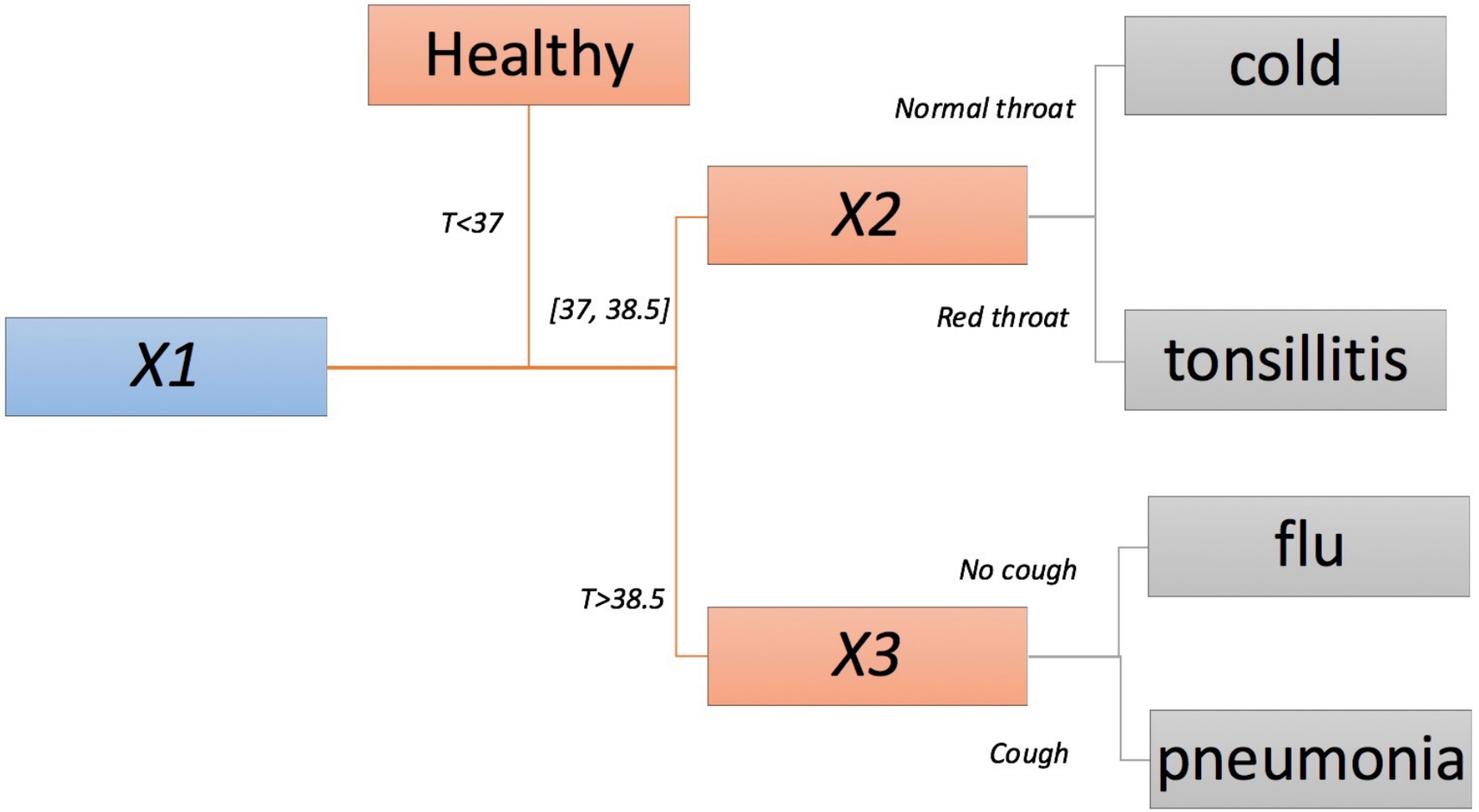
3. If `X1` has values between [98, 102] and `X3=`"there is reddening of throat", then `Y`=tonsillitis;

4. If `X1  99` and `X2=`"there is no cough", then `Y`=flu;

5. If `X1  99` and `X2=`"there is cough", then `Y`=pneumonia;

EXAMPLE

Any new (unseen) patient can now be diagnosed using these rules.



INDEPENDENT PRACTICE

DATA SCIENCE: CASE STUDY

INSTRUCTIONS

- We recommend using a Jupyter notebook for this practice.

From the materials:

1. Unzip the file downloaded to a known location in your file system
2. Locate the file called DataScience101_Part2_DecisionTree.ipynb and the Iris dataset.
3. Open Jupyter: Open a terminal
    - **Mac**: Using spotlight search for "Terminal"
    - **Windows**: Click the "Start" button and type "`cmd`"
    - In the terminal type: `` `jupyter notebook` ``
4. Navigate to the folder where you have saved the file in step 1
5. Open the file from the Jupyter interface
6. Voilà, you are ready to follow this practice

- In this independent practice we are using the Iris data set to see how Python can help us construct a decision tree like the one we have discussed.

CONCLUSION

REVIEW & RECAP

- In this workshop, we've covered the following topics:
  - Why data science?
  - What can data science do for me?
  - What is the data science workflow?
  - How to analyze and visualize data using Python
  - Define the role of algorithms and their relationship with machine learning
  - Demonstrate how these concepts can be applied to make predictions

TAKEAWAYS

LEARNING PLAN

Evaluate your data science skills! How confident are you with:

- Programming skills (Python or R)
- Knowledgable in algebra and statistics (analyzing and modeling data)
- Business acumen (how to work with stakeholders)
- Industry expertise (for the type of field you're working within)
- Communication skills (visualize data, tell stories)

WHAT SHOULD YOU DO NEXT?

Refer back to your earlier self-assessment:

1. Which skills do you want to improve first? Which ones are you most interested in learning about?

2. Rank these and identify the top three focus areas.

3. For each focus area, identify *at least* one possible resource and a related goal.

WHAT SHOULD YOU DO NEXT?

Want to be a better programmer?

Work on these:

- Continue learning Python syntax on sites like Codecademy or Code School.
- Already know R? Work on comparing the two.
- Interested in other frameworks? Try Spark!

WHAT SHOULD YOU DO NEXT?

Want to brush-up on your math and statistics skills?

Have a look at these:

- **Data Analysis with Open Source Tools**, P. K. Jannert
- **Pattern Recognition and Machine Learning**, C. Bishop
- **Data Science and Analytics with Python**, J Rogel-Salazar
- An Introduction to Statistical Learning with Applications in R (free PDF)
- Elements of Statistical Learning (free PDF)

WHAT SHOULD YOU DO NEXT?

Concerned about business acumen & communication skills?

Have a look at these:

- **Data Science for Business**, F. Provost and T. Fawcett
- **Storytelling with Data: A Data Visualization Guide for Business Professionals, C. Nussbaumer Knaflic**

WANT MORE?

General Assembly offers courses in data science!

Check out our:

- Part-time Data Science Course
- Data Science Immersive Course

ADDITIONAL RESOURCES

BOOKS

- **Data Analysis with Open Source Tools**, P. K. Jannert
- **Data Science for Business**, F. Provost and T. Fawcett
- **Pattern Recognition and Machine Learning**, C. Bishop
- **Data Science and Analytics with Python**, J Rogel-Salazar
- An Introduction to Statistical Learning with Applications in R (free PDF)
- Elements of Statistical Learning (free PDF)
- Think Stats (free PDF or HTML)
- Mining of Massive Datasets (free PDF)

MOOCS

- Andrew Ng’s Machine Learning Class on Coursera link
- MIT’s Artificial Intelligence course link
- Johns Hopkins' Data Analysis Methods link
- Cal Tech’s Learning from Data course link

AGGREGATORS

- DataTau: Like Hacker News, but for data
- MachineLearning on reddit: Very active subreddit
- Quora's Machine Learning section: Lots of interesting Q&A
- Quora's Data Science topic FAQ
- KDnuggets: Data mining news, jobs, classes and more

DATA SCIENCE 101

SOCIAL

- Hillary Mason (@hmason): Data Scientist in Residence at Accel and Scientist Emeritus at bitly.
- Dj Patil (@dpatil): VP of Product at RelateIQ.
- Jeff Hammerbacher (@hackingdata): Founder and Chief Scientist at Cloudera and Assistant Professor at the Icahn School of Medicine at Mount Sinai.
- J Rogel-Salazar (@quantum_tunnel): Data scientist at IBM and GA instructor
- Peter Skomoroch (@peteskomoroch): Equity Partner at Data Collective, former Principal Data Scientist at LinkedIn.
- Drew Conway (@drewconway): Head of Data at Project Florida

Q&A

EXIT TICKETS

DON’T FORGET TO FILL OUT YOUR EXIT TICKET