GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

Srikanta Patra
Data Science Instructor, General Assembly


INTRODUCTION

WELCOME

Srikanta Patra

https://www.linkedin.com/in/srikanta-patra

srikantapatra@gmail.com

- Twenty years experience in Software Development, Business Intelligence, Data Warehousing, Data Science.
- @General Assembly: Learnt Data Science Immersive and working as a Data Science coach and consultant.
- Freelance Data consultant.
- Love mountains and outback. Had lots of fun trekking Himalayas for a week ….☺

GENERAL ASSEMBLY

IRi Growth delivered.

ca technologies

ABOUT YOU

About you and what brings you here.

LEARNING OBJECTIVES

- What is Data Science
- Popular tools & resources to visualize, analyze, & model data.
- Problems that Data Science can Address.
- Create a custom learning plan to build your data science skills after this workshop!

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 opens up a door to a variety of opportunities
  - Data science has been dubbed the “Sexiest job of the 21st century”!

INTRODUCTION

WHAT IS DATA SCIENCE AND WHAT CAN IT DO?

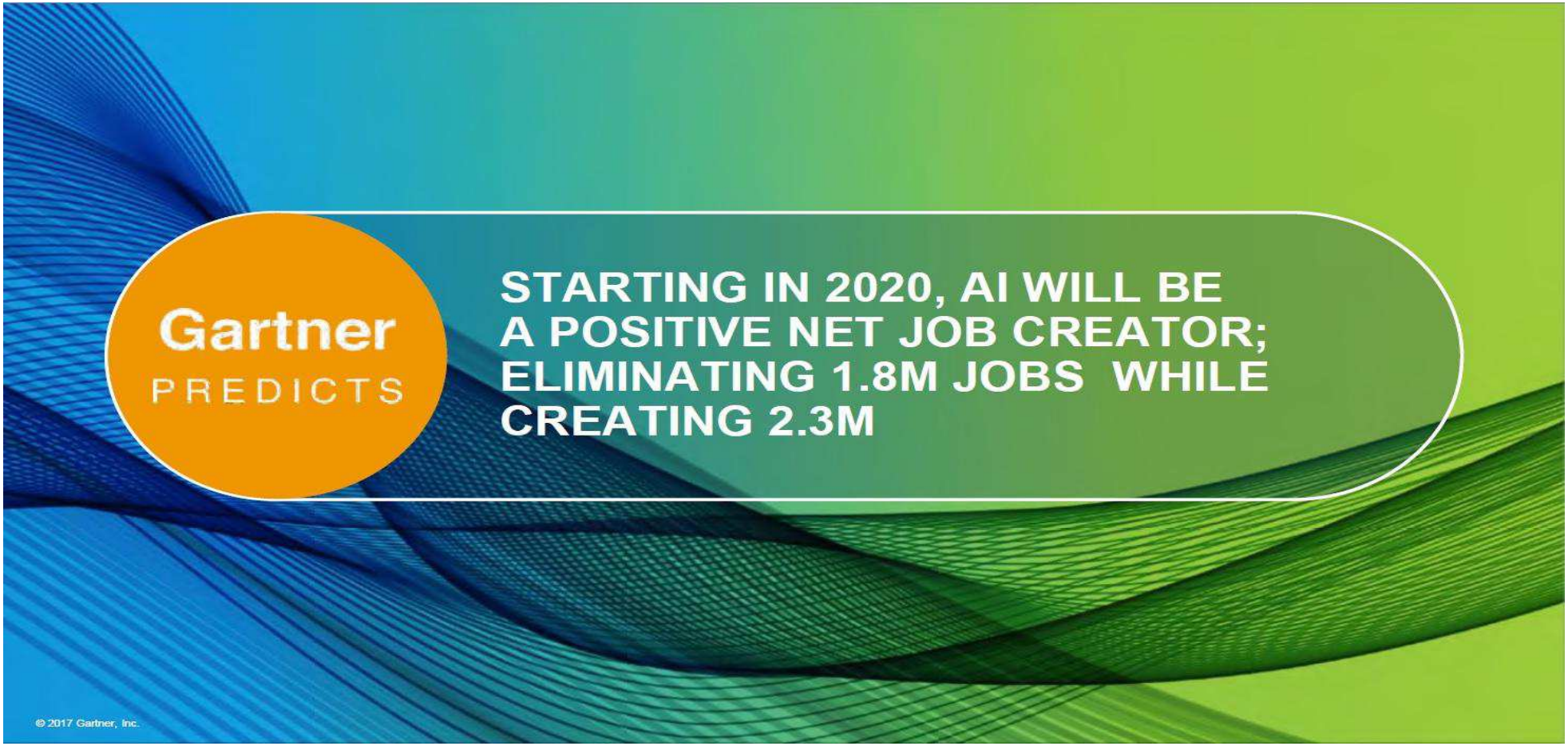
WHAT IS DATA SCIENCE?

THE SEXIEST JOB OF THE 21ST CENTURY

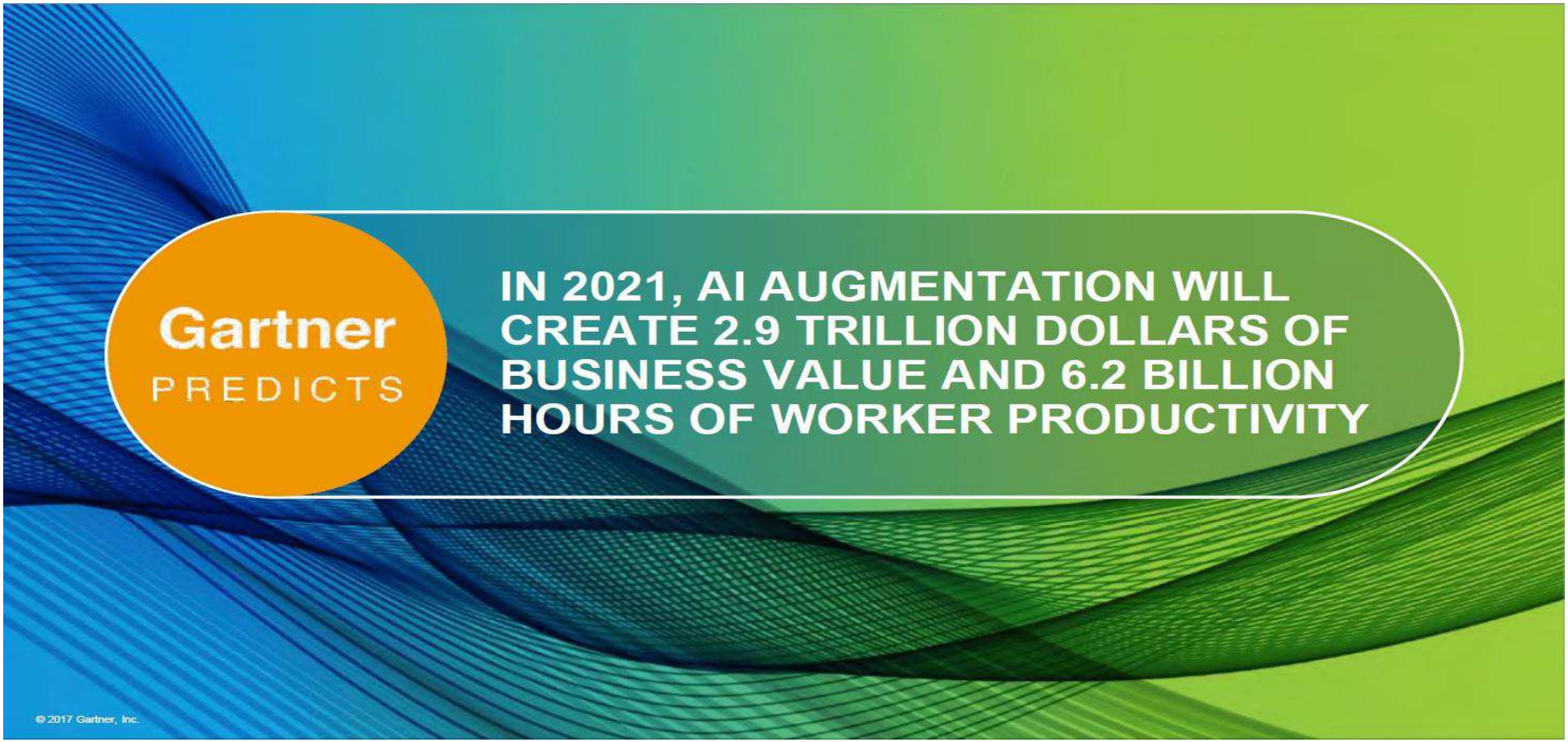
- **Data Science**: A set of tools and techniques used to extract useful information from data.

  - An interdisciplinary, problem-solving oriented subject.
  - The application of scientific techniques to practical problems.
  - A rapidly growing field.

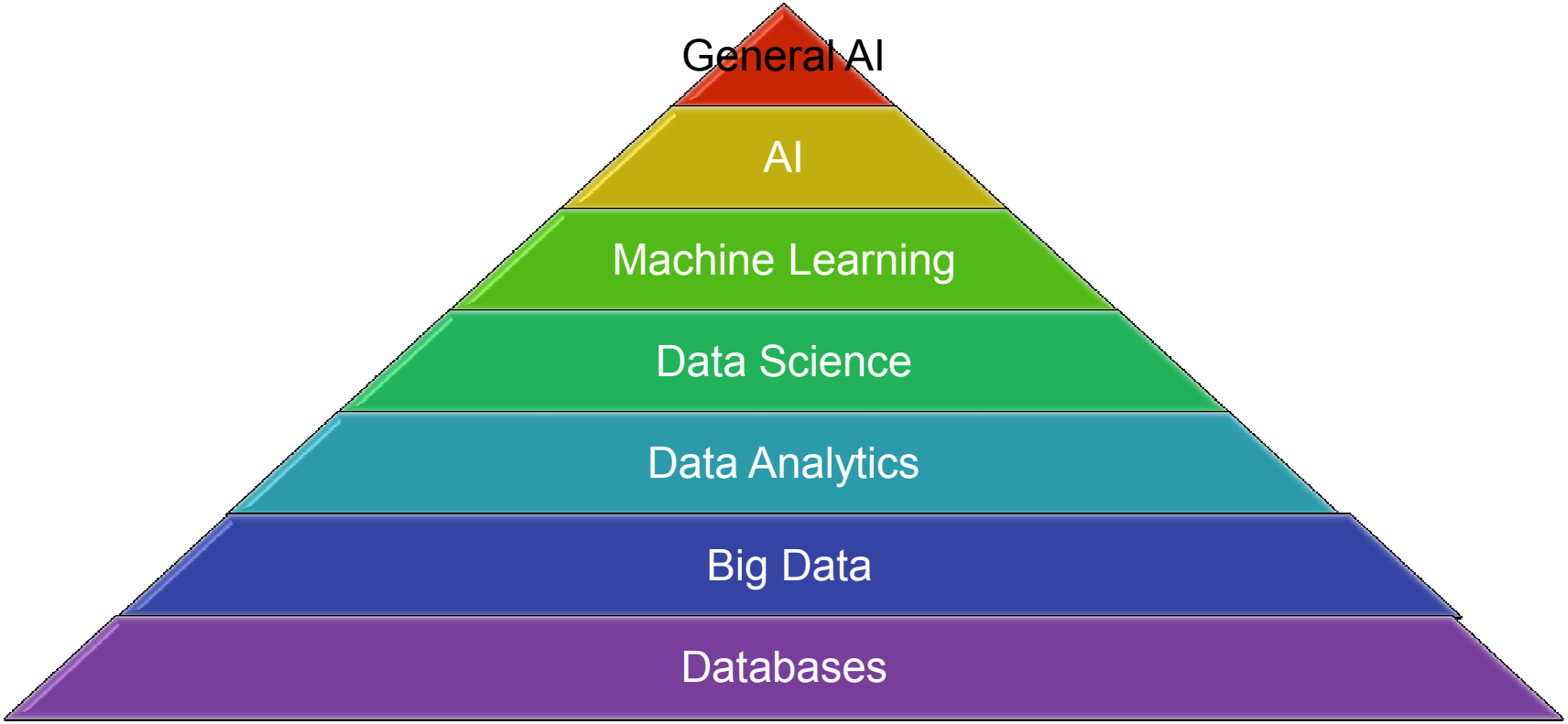
GARTNER


Gartner
PREDICTS
STARTING IN 2020, AI WILL BE A POSITIVE NET JOB CREATOR; ELIMINATING 1.8M JOBS WHILE CREATING 2.3M
© 2017 Gartner, Inc.

GARTNER

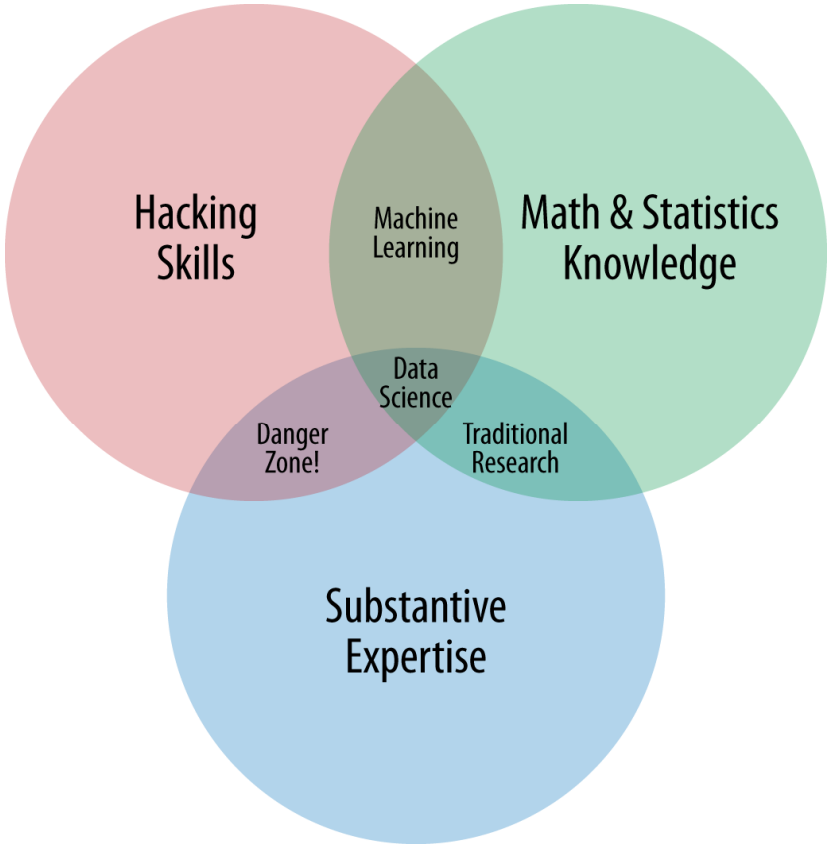

Gartner
PREDICTS
IN 2021, AI AUGMENTATION WILL CREATE 2.9 TRILLION DOLLARS OF BUSINESS VALUE AND 6.2 BILLION HOURS OF WORKER PRODUCTIVITY
© 2017 Gartner, Inc.

EVOLUTION OF DATA

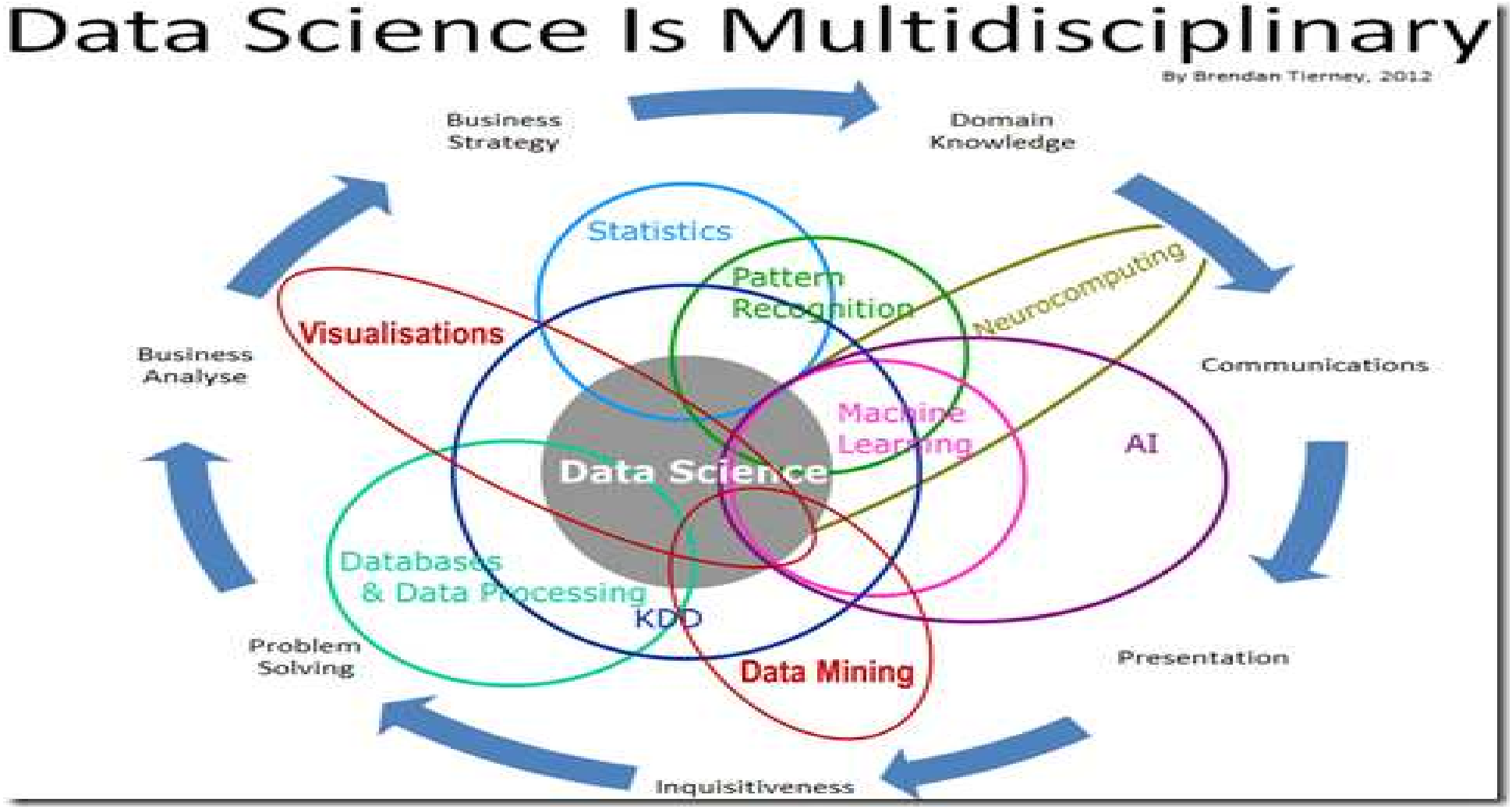


QUALITIES OF A DATA SCIENTIST

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



Data Science Is Multidisciplinary

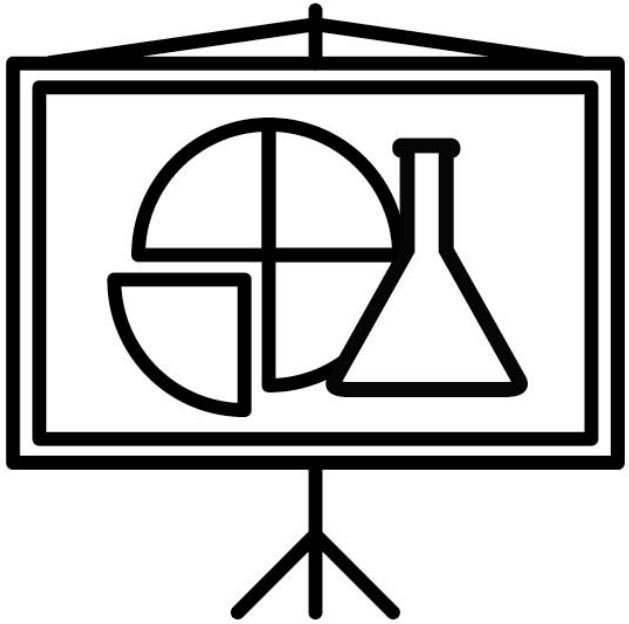


DISCUSSION

LET’S DISCUSS YOUR IDEAS

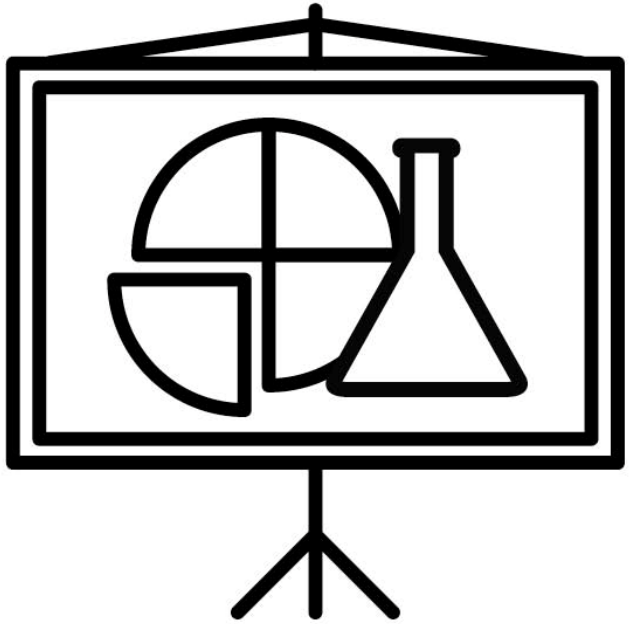
WHAT CAN DATA SCIENCE DO FOR ME?

- Ask good questions:
  - What is required?
  - How are results evaluated? (measures of success)
  - What do we currently know? (existing data)
  - What has happened? (descriptive analytics)
  - What will happen (if)? (predictive analytics)
  - What to do to achieve what we require? (insight)
- Define and test a hypothesis/run experiments.



WHAT CAN DATA SCIENCE DO FOR ME?

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



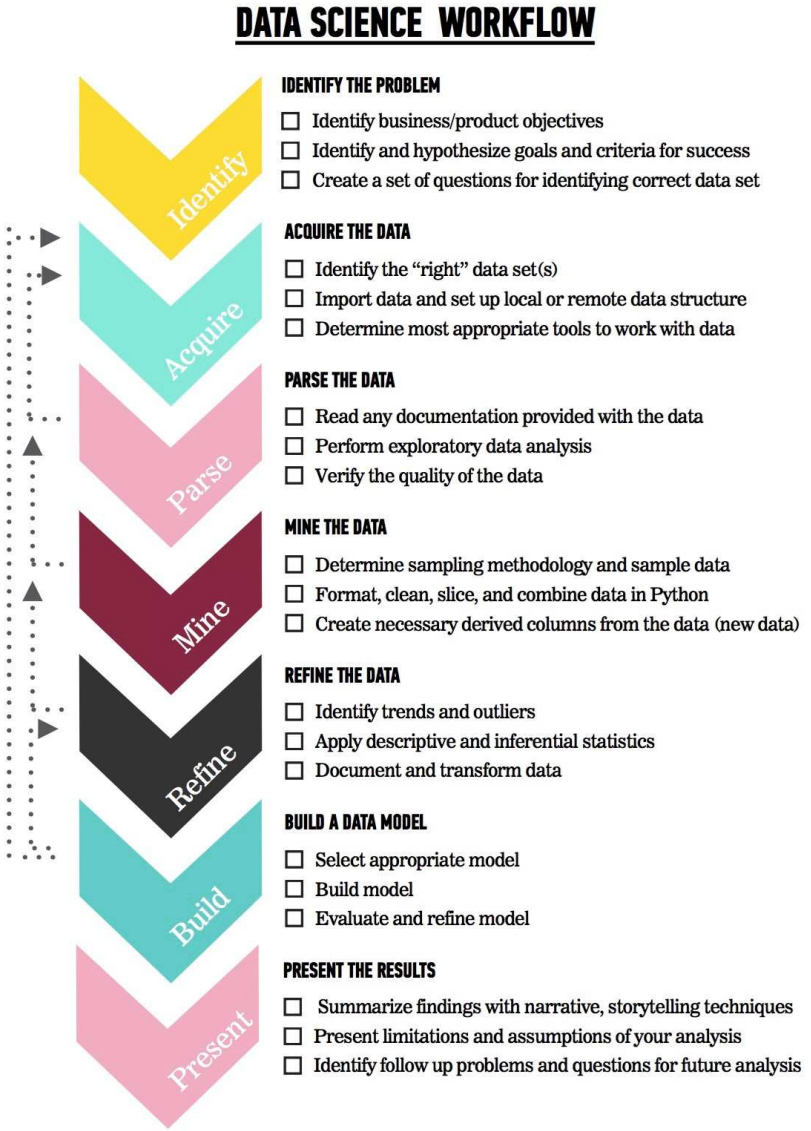
DEMO

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



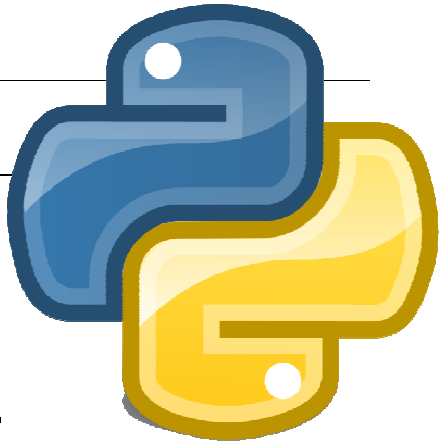
INDICATORS OF NEW ONLINE PURCHASE

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

GUIDED PRACTICE

EXPLORING THE DATA SCIENCE TOOLKIT

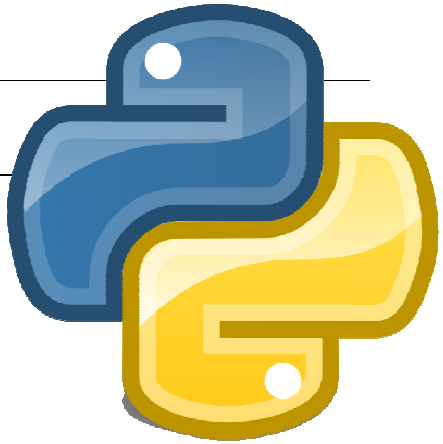
WHY PYTHON?



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!

WHY PYTHON



- High-level, interpreted and general-purpose dynamic programming language that focuses on code readability.
- Fewer steps as compared to Java or C++.
- Founded in 1991 and makes programming easy and fun.
- Supports multiple programming paradigms.
- Involve imperative, object-oriented, and functional programming.

https://pypi.python.org/pypi

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**.

INTRODUCTION

WHAT ARE ALGORITHMS, ANYWAY?

DISCUSSION

ALGORITHMS?

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

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?

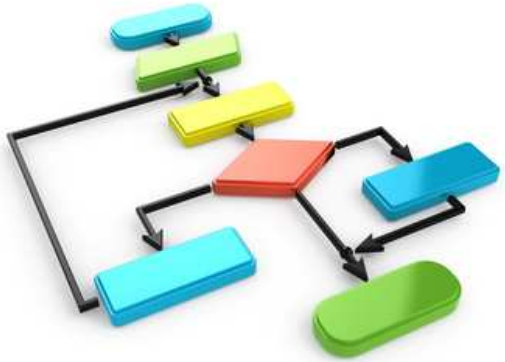
- What if you put on your jacket before you put on your coat?

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.



CRITERIA OF A GOOD ALGORITHM

- An algorithm is an **unambiguous description** that makes clear what has to be implemented.
  - In a recipe, a step such as "*Bake until done*" is ambiguous because it doesn't explain what "*done*" means.
  - In a computational algorithm, a step such as "*Choose a large number*" is vague: what does "large" mean to a computer?

ALGORITHM

CRITERIA OF A GOOD ALGORITHM

- An algorithm expects a defined set of **inputs**.
  - For example, it might require two numbers where both numbers are greater than zero. Or it might require a word, or a list of zero or more numbers.

- An algorithm produces a defined set of **outputs**.
  - It might output the larger of the two numbers, an all-uppercase version of a word, or a sorted version of the list of numbers.

CRITERIA OF A GOOD ALGORITHM

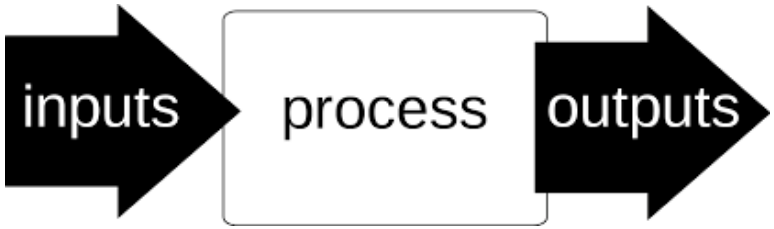
- An algorithm should be **guaranteed to terminate** and **produce a result**, always stopping after a finite time.

- If an algorithm could potentially run forever, it wouldn’t be very useful because you might never get an answer!

ALGORITHM

CRITERIA OF A GOOD ALGORITHM

- We can condense some of this information as follows:



WHAT IS THE ALGORITHM FOR…?

EXAMPLES

1. Making breakfast.
2. Commuting to work.
3. Making a cup of coffee.
4. Brushing teeth.

ALGORITHMS

ALGORITHMS IN ACTION

LET US SEE HOW TO WRITE AN ALGORITHM

- 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.

THINKING LIKE AN ALGORITHM

WHAT IS THE OUTPUT

ALGORITHM

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.

THINKING LIKE AN ALGORITHM

HERE IT IS IN PYTHON

Python

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

DISCUSSION…?

WAS IT A GOOD ALGO

1. Does the algorithm above meet the criteria for a good algorithm?
    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?

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 structure from data
- **Generalization** – making predictions from data

MACHINE LEARNING PROBLEMS

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

- For example, suppose you want to predict whether someone will 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 whether a new person will make a purchase within a week.

MACHINE LEARNING PROBLEMS

Supervised Machine Learning:

- You can 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

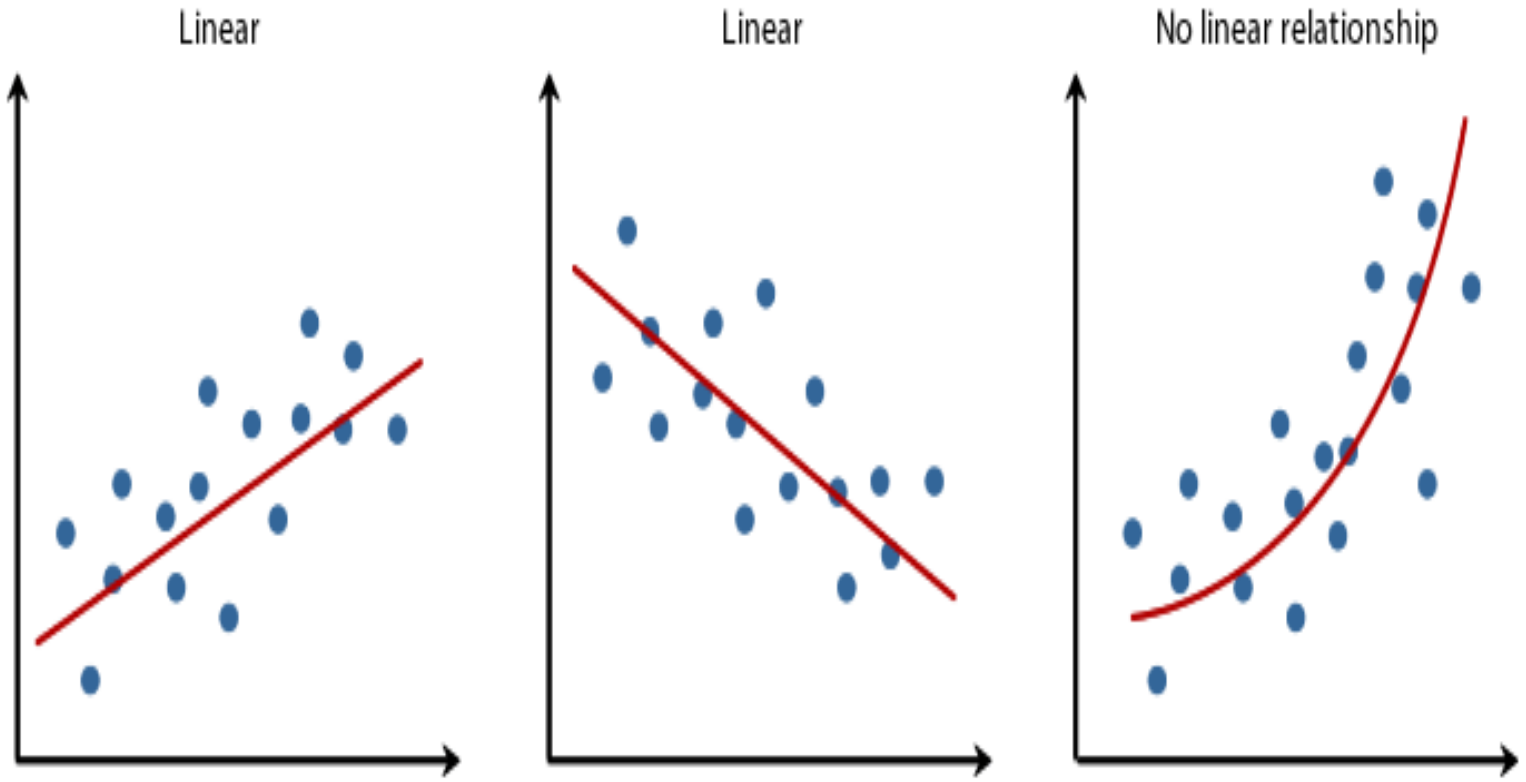
Unsupervised Machine Learning:

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

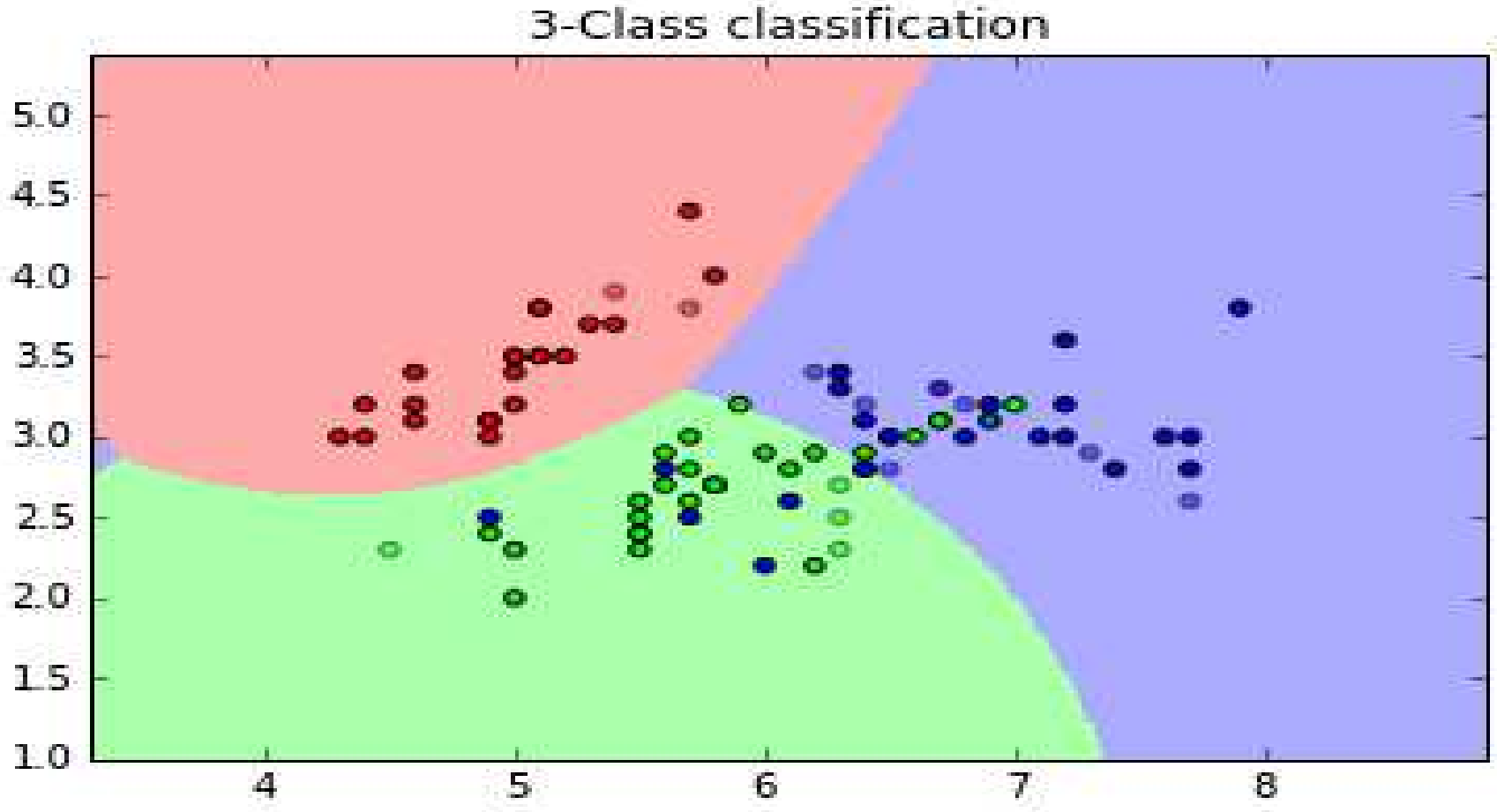
- You can then take action and make an offer or recommend a product specifically to these segments.

- Some unsupervised learning algorithms include:
  - Clustering
  - Anomaly Detection
  - Principal Component Analysis (PCA)

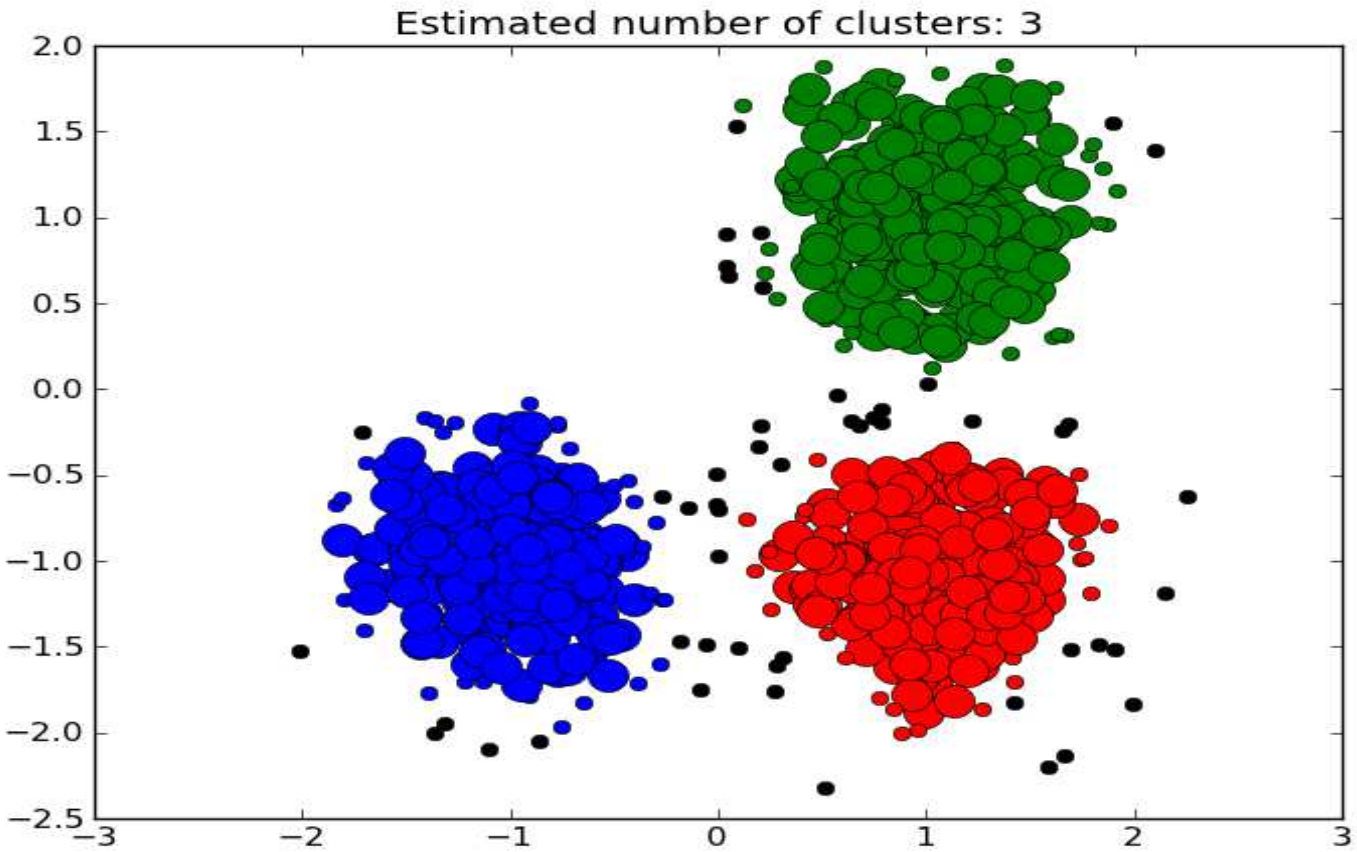
LINEAR REGRESSION



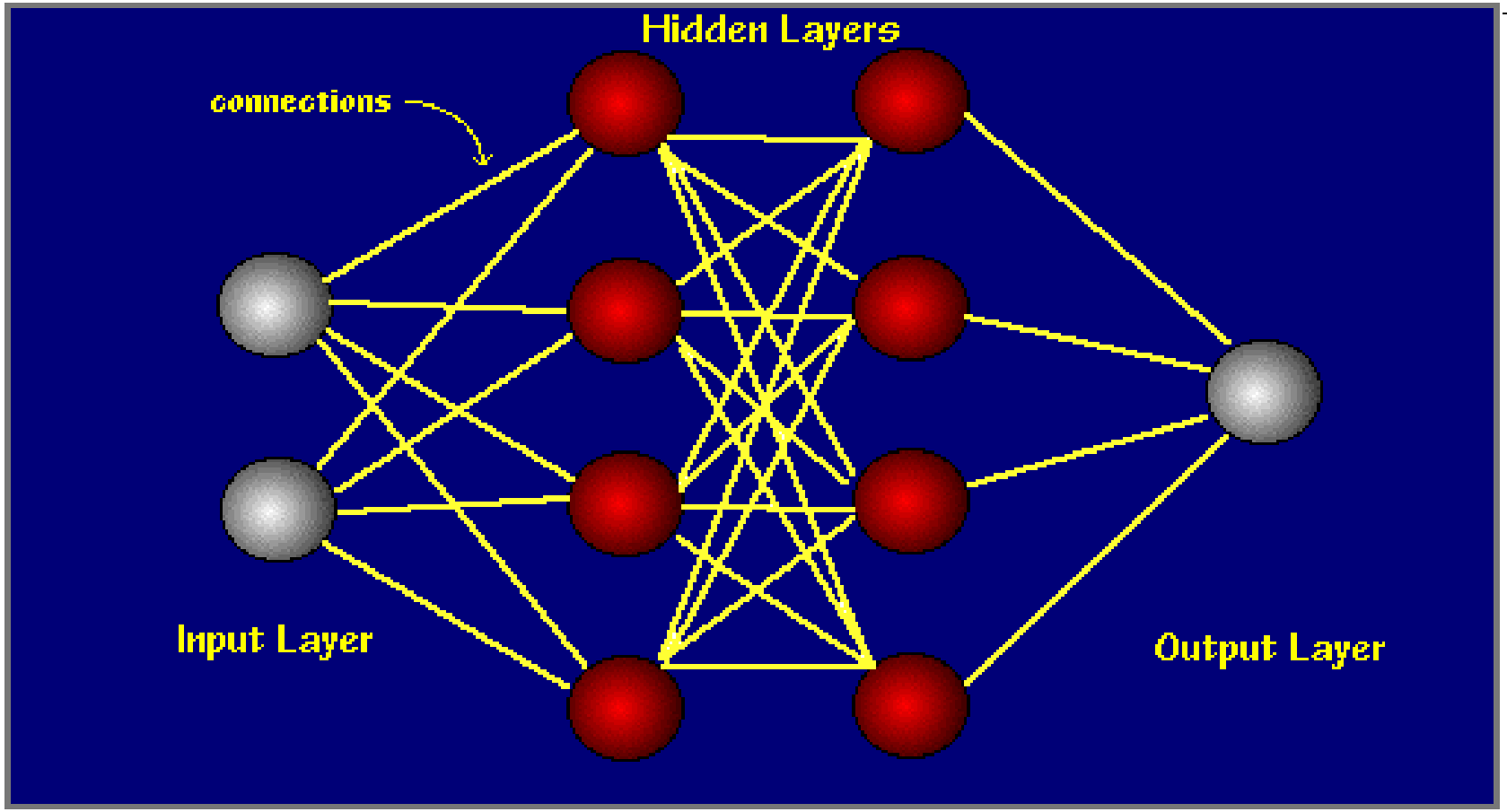
CLASSIFICATION (SUPERVISED LEARNING)



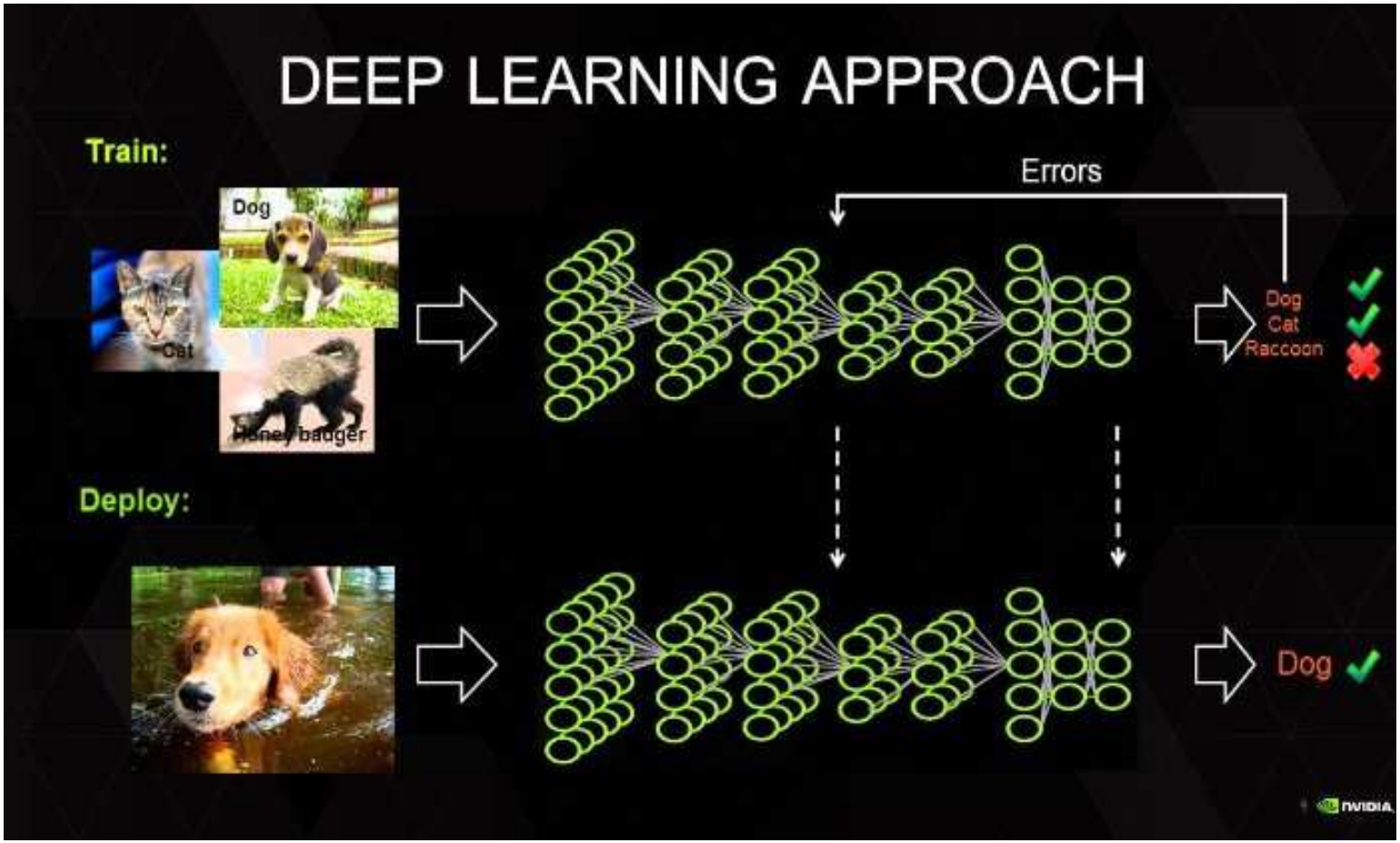
CLUSTERING (UNSUPERVISED LEARNING)



NEURAL NETWORK



DEEP LEARNING



DATA SCIENCE

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)

TAKEAWAYS

WHAT SHOULD YOU DO NEXT?

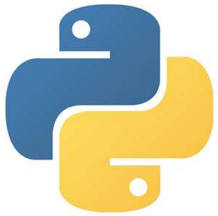
Refer back to your earlier self-assessment:

- Which skills do you want to improve first? Which ones are you most interested in learning about?
- Rank these and identify the top three focus areas.
- For each focus area, identify *at least* one possible resource and a related goal.

TAKEAWAYS

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!

TAKEAWAYS

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)

TAKEAWAYS

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**

TAKEAWAYS

WANT MORE?

General Assembly offers courses in data science!

Check out our:

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

DATA SCIENCE

ADDITIONAL RESOURCES

DATA SCIENCE

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)

DATA SCIENCE

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

DATA SCIENCE

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

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

DATA SCIENCE

Q&A

DATA SCIENCE

EXIT TICKETS

DON’T FORGET TO FILL OUT YOUR EXIT TICKET