NC STATE UNIVERSITY

Overview NLP Data Science Projects

Dr. Melissa G Ngamini

NC STATE Data Science Academy

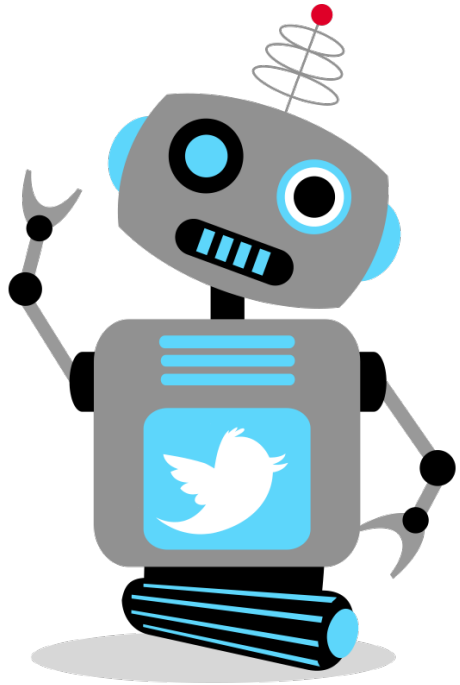
December 1, 2022

Plan of Presentation

Melissa Ngamini, Ph.D

- **Education**:
  - Bachelor of Science in Pure Mathematics
  - Ph.D in Applied Mathematics
- **Academic Career**:
  - Assistant Professor of Mathematics
    - Classification using Support Vector Machine
    - Classification using Random Forest
    - Mathematical Research Experience (Applied Statistics/Data Science).
- **Data Scientist Career**:
  - Lead Data Scientist
    - Agent Assist Problem Statement Identification
    - Audio Transcripts Topic Modelling
    - Topic Classification Chatbot
    - Document Classification Application
  - Senior Data Scientist WARNERMEDIA
    - Identity Graph
    - Streaming Platform Engagement
    - Ad Attribution
    - Twitter Bot & Topic Detection
  - Head of Education
    - Data for Hope Education Program (2022)
    - Custom Project Coordinator (2021)
    - Custom Projects Volunteer (2018 - 2020)

Project: Twitter Bot & Topic Detection



Twitter Bot & Topic Detection

Introduction

Business Goals:

1. Determine the **probability a user account is Blocked/Bot** from its Tweet
2. Create Twitter Bot Network.
3. Identify the top 5 topics the High and Low Probability bots are pushing in a particular hashtag during the Primary Elections in 2020.

Client:

CNN politics

Data:
Pulled from twitter daily in a 8:00 AM to 8:00 PM window..

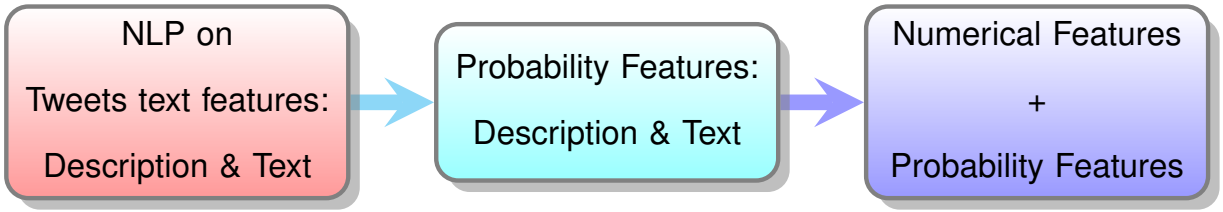
Technology Selection

- **Data Source & Platform**

- **Data Science**

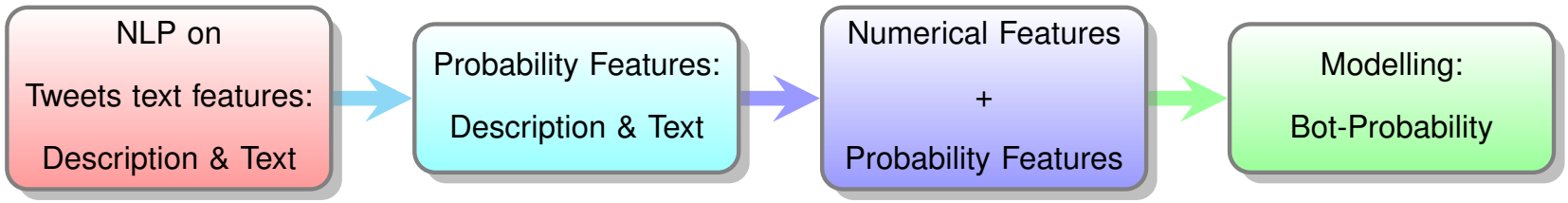
- **Data Visualization**

Features Selection & Modelling



1 Natural Language Processing

- Feature used: *DESCRIPTION* and *TWEET* (text)
- Model:
  - *TF – IDF* Vectorizer to transform text to feature vectors that can be used as input to estimator.
  - Stochastic Gradient Descent to create a new *DESCRIPTION – PROBABILITY* and *TWEET – PROBABILITY*.



2 Supervised Learning

- Features used: All Numerical features
- Model: Gradient Boosting to create the *BOT – PROBABILITY*

Definitions of the Topic of a Group of Twitter Users

1. **Definition:**
   The topic of particular Tweet is the subject matter or issue a group of Twitter Users are pushing through **Replies**, **Quotes** *or* **Retweets**.

2. **Method of Differentiation of Topic:**
   Latent Dirichlet Allocation (LDA) uses Dirichlet priors for the document-topic and word-topic distributions, lending itself to better generalization

3. **Interpretation of Topics from LDA:**
   The topic of an **LDA Component** *is inferred by looking at the set of 10 words that are used the most together in a* **Tweet** *in that component and cross checked against the occurrence of those words across all components.*

How to build a Twitter Bot Network

- **Network**:

  It is a catalog of systems components often called nodes or vertices and the directed interaction between them called edges or links.

- **Twitter Social Network**:
  - Nodes: Twitter User
  - Edges: Interaction
    - source = *SCREEN – NAME*
    - target = *REFERENCED – SCREEN – NAME* (Original, Reply or Quote)

Twitter Social Network Visualization
#IMPEACHMENT 2019-10-04 to 2019-10-13

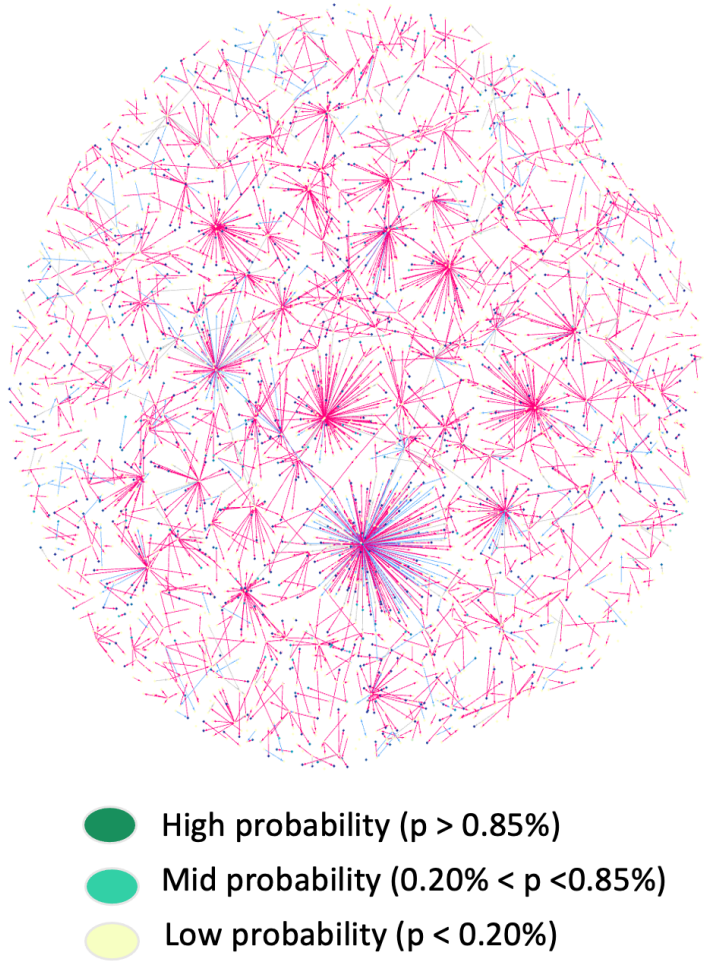


Figure: #Impeachment Network

Topics, High Probability Bots & Influencers in # IMPEACHMENT

Topics in # IMPEACHMENT

- Terms in Topic 0
  - trump, impeachment, new, support, much, sayshummingbird, advice, carter, legend, jimmy
- Terms in Topic 1
  - trump, good, sondland, man, realdonaldtrump, american, ambassador, voter, turkey, anything
- Terms in Topic 2
  - trump, president, decision, sondland, new, ambassador, syrian, inquiry, intention, crystal
- Terms in Topic 3
  - great, unmatched, turkey, anything, wis, mitt, romney, pompous, people, senator
- Terms in Topic 4
  - impeachment, democrat, voter, collus, liar, realdonaldtrumpschiff, scheme, schiff, adam, biden

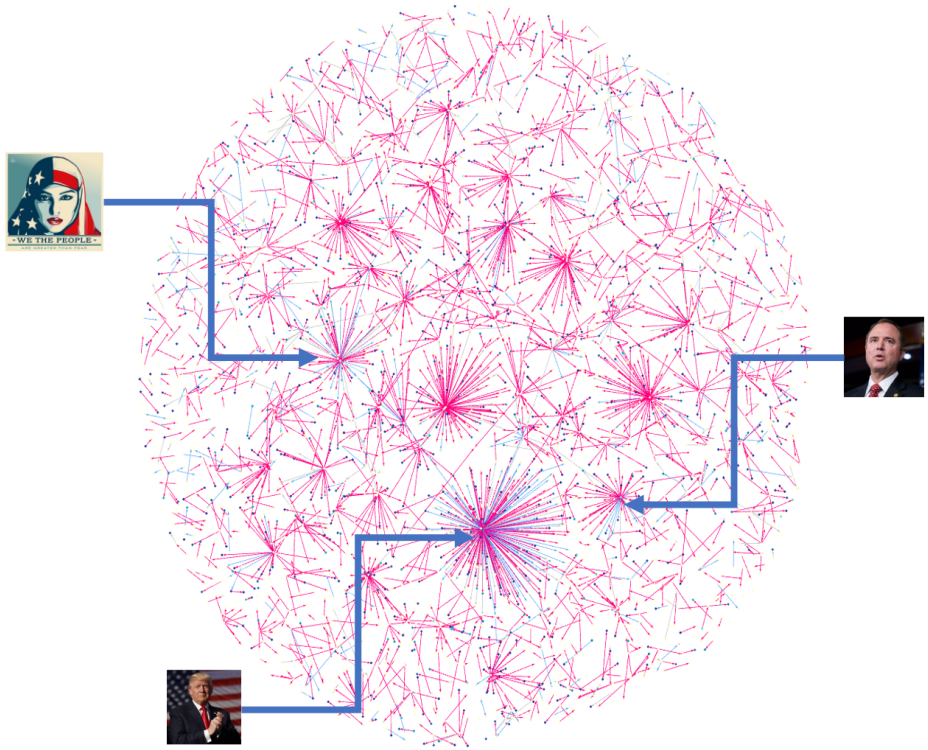


Figure: Topic 0 Top Words

Topic Classifier for Chatbot

Topic Classifier for Chatbot

Introduction

Business Goals:

1. Create a **Topic Classifier model** to predict **Use Case/Topic** for different Utterance
2. Use the model to label the new utterance for new dataset

Client:

 Payroll team

Data:

Utterances from Chatbot

Technology Selection

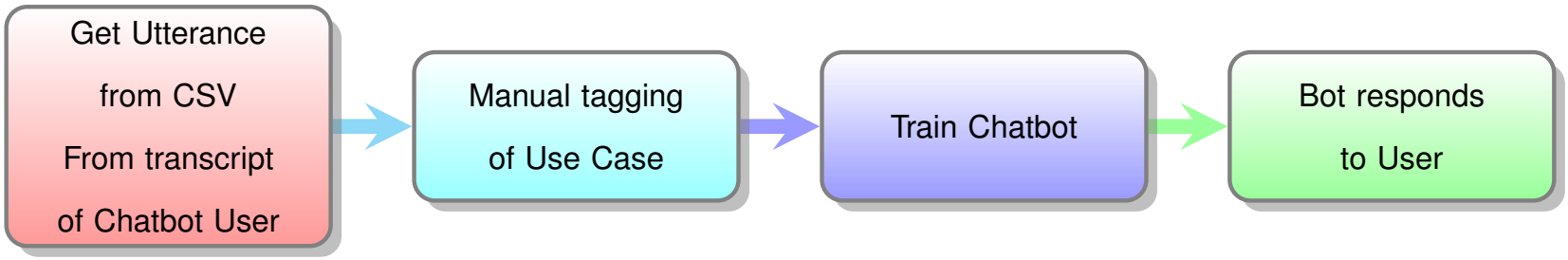
- **Data Science**

- **Web Application**

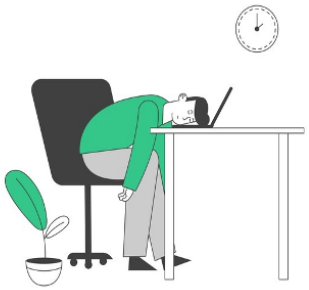
Topic Classification

Topic Classification:
Topic classification is a 'supervised' machine learning technique, one that needs training before being able to automatically analyze texts.

Manual Labeling Solution for Chatbot

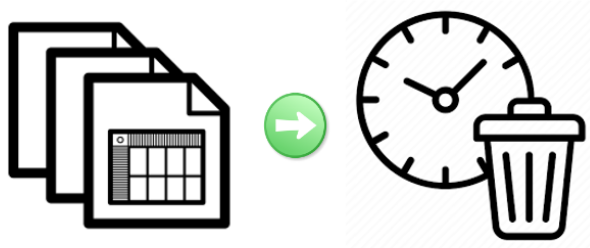


Problem with Manual Labeling



Manual Effort

Typing mistakes



Big dataset

Time consuming

Topic Classification Solution

Natural Language Processing

- Preprocessing Libraries:
  - *NLTK*
  - *SPACY*
- Word Embeddings:
  - *Word2Vec*: Pre-trained word embeddings that are released by Google
  - Domain Trained Embedding *BOW* (Bag Of Words)

Supervised Learning

- Features used: *Utterances*
- Models to obtain the *UseCase*:
  - **(CNN)** Convolutional Neural Network
  - **(BERT)** Bidirectional Encoder Representation from Transformers
  - **MIx-Model** (BERT + CNN)

Topic Classification Solution Models

CNN: Convolutional Neural Network using Domain trained Embedding

- is a class of deep, feed-forward artificial neural networks (where connections between nodes do not form a cycle).Â
- is based on the concept of filtering, where filter weights are learned using stochastic gradient descent (SGD) algorithms.
- context-free models generate a single word embedding representation for each word in the vocabulary independently from the word context.
- Embedding size = 300

BERT: Bidirectional Encoder Representation from Transformers

- uses the transformer architecture in addition to a number of different techniques to train the model
- Pre-Training model + fine-tuning
- Takes advantages of previous models
- is a deeply bidirectional, unsupervised language representation able to create word embeddings that represent the semantic of words in the context they are used.
- Embedding size = 1024

Topic Accuracy and Coverage by Models

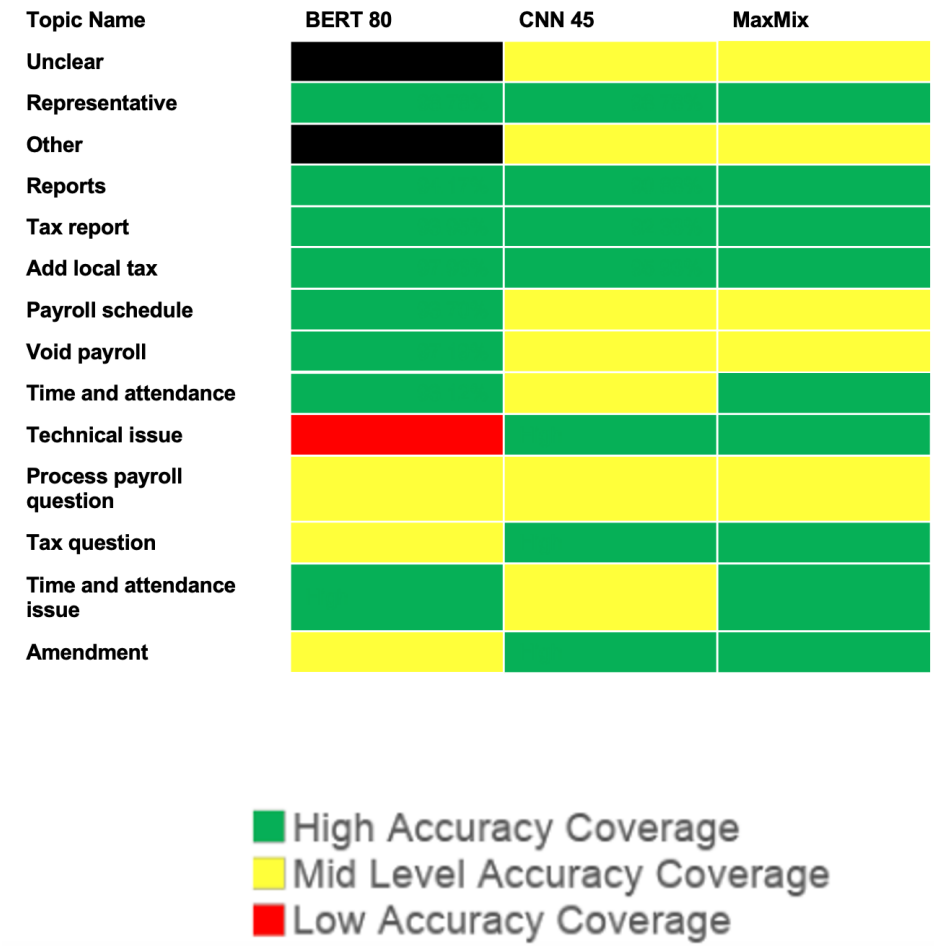
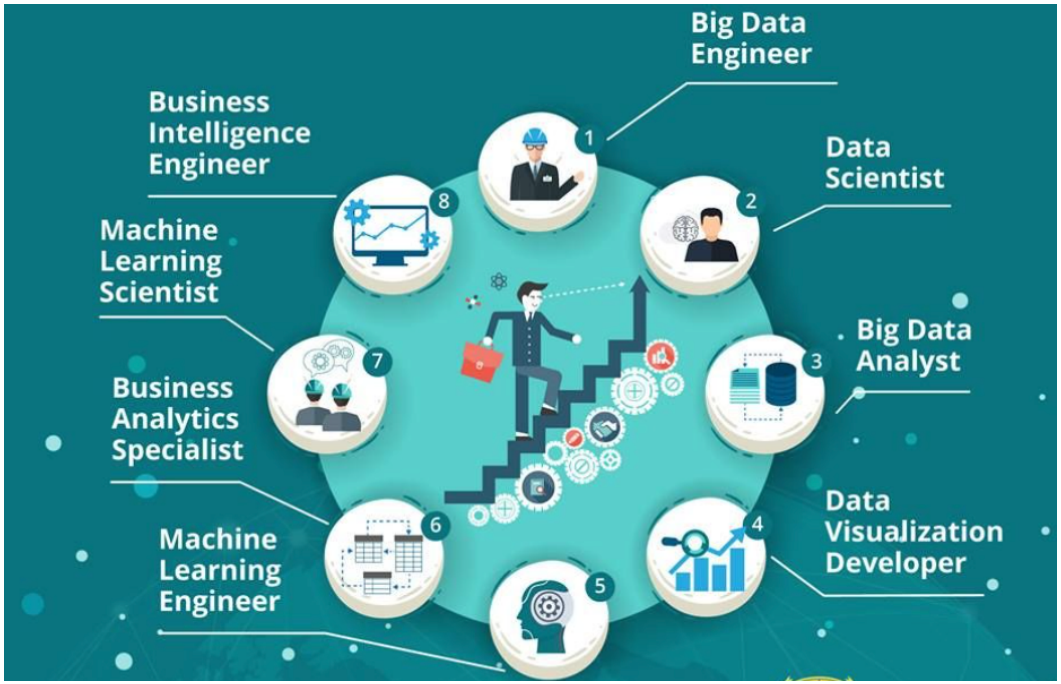


Figure: High Volume Topic Accuracy by Model

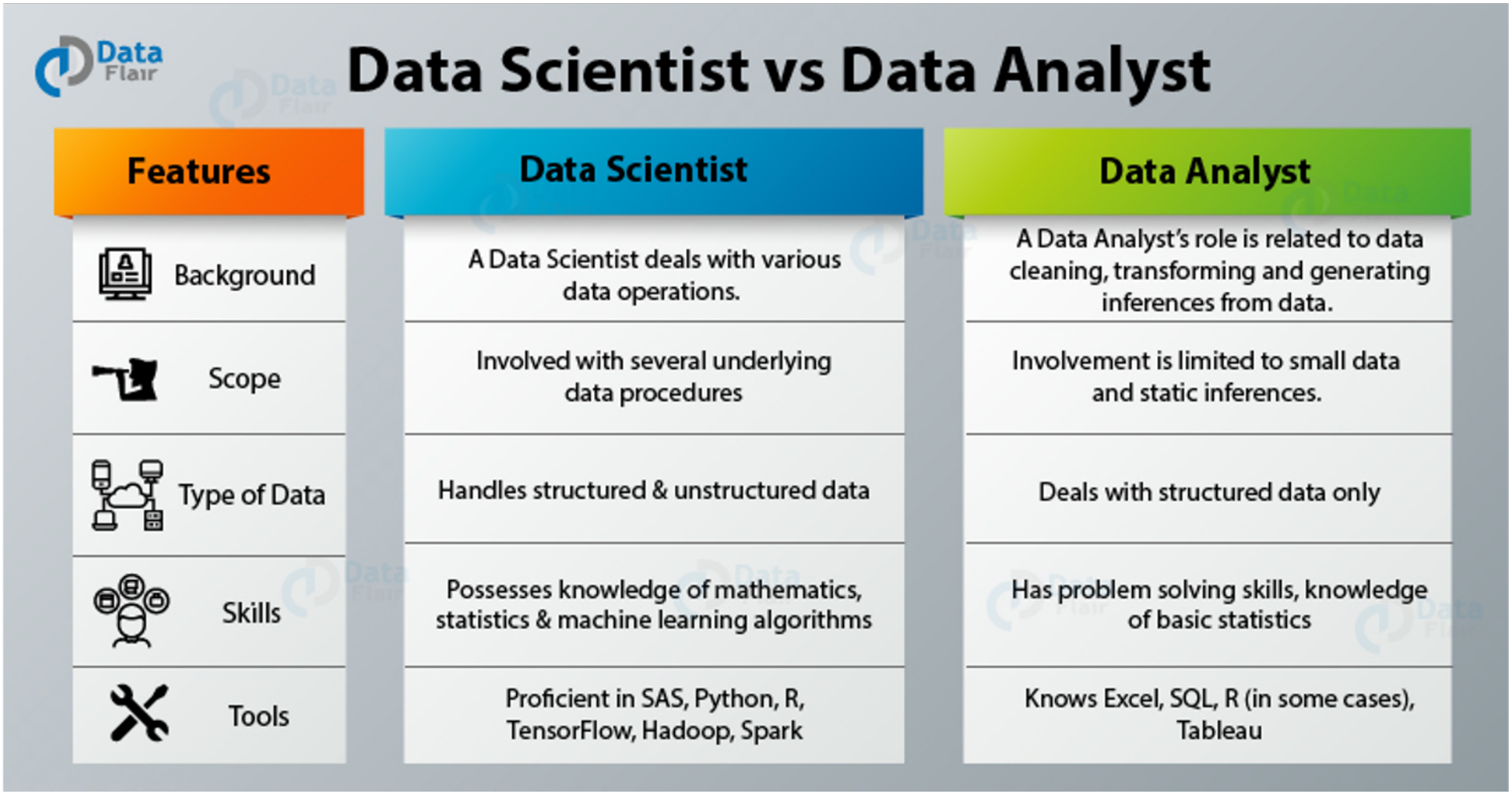
- BERT 80
  - Removed Unclear Other
  - Topic with small distance between each other
  - Semantic understanding needed
- CNN 45
  - 50 Utterances minimum per Topic
  - Topics with large distance between each other
  - Short and to the point sentences
- Accuracy Reports
  - BERT 80: 73%
  - CNN 45: 70%
  - Mix-Model: 84%

How to get a career in Data?

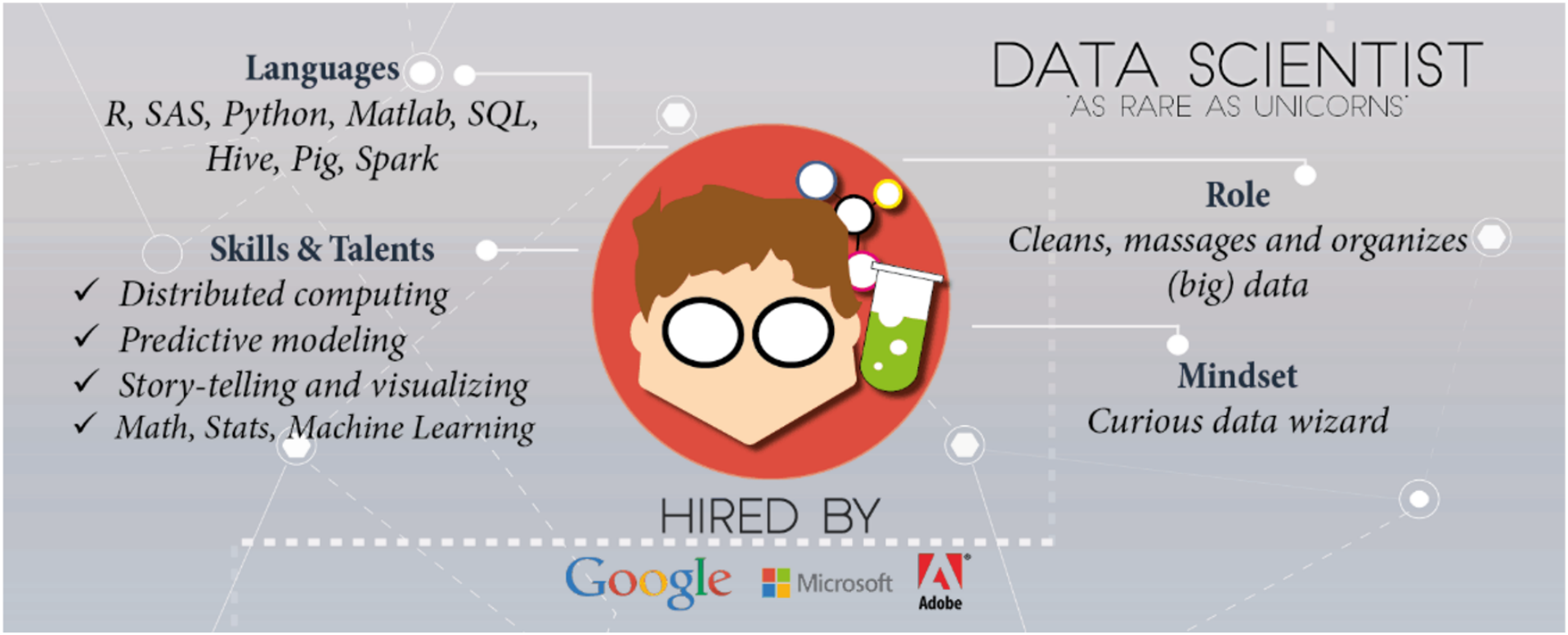


Data Careers

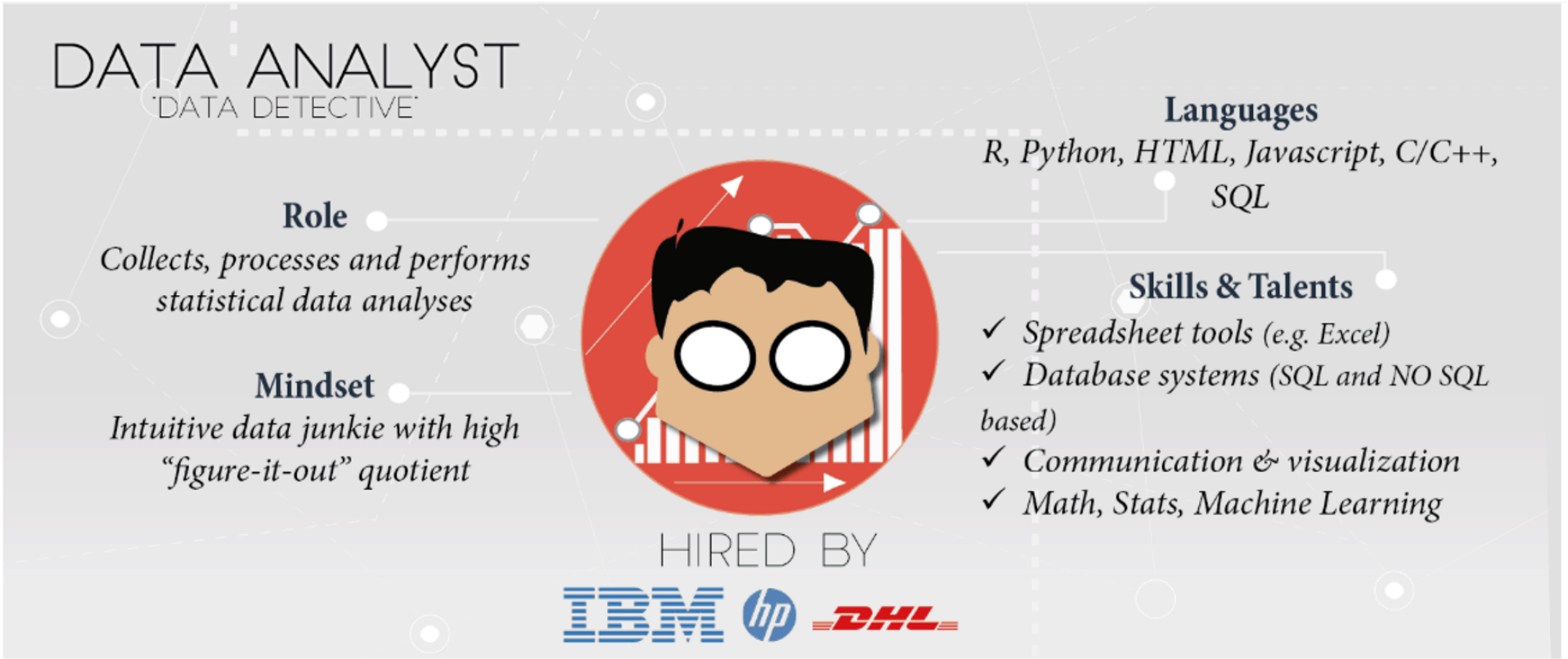
Data Science vs. Data Analytics



What Is a Data Scientist?



What is a Data Analyst?



What is a Data Engineer?

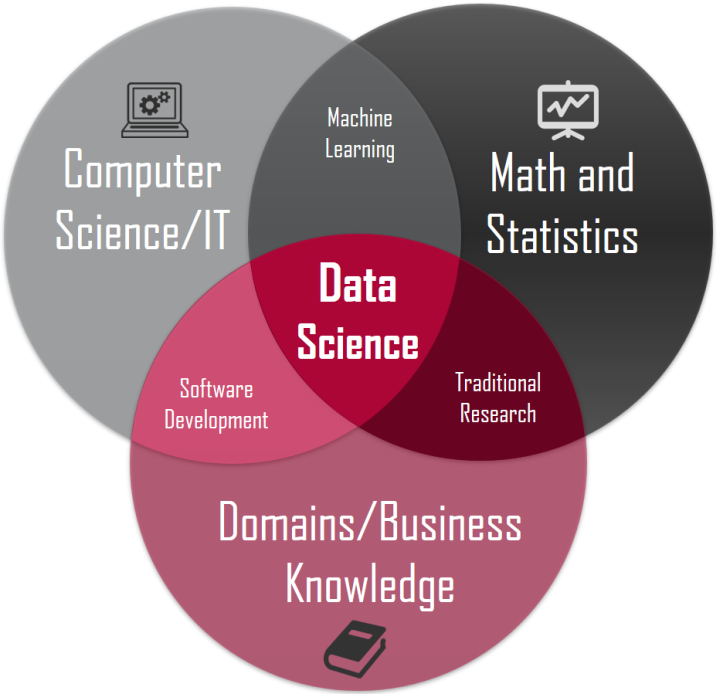
What's next?

1 What do you want to do?

- Business analysis
- Data analysis
- Data visualization
- Financial analysis
- Healthcare analysis
- Recommendation systems
- Image processing
- Natural language processing

2 How do I acquire the skills ?

Should I get a Masters or PhD?

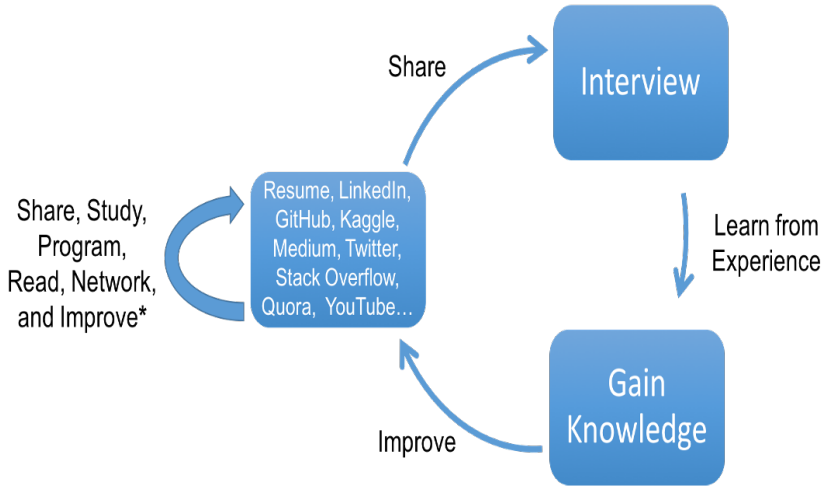


Is there another way ?

codeup
PRINCIPAL ANALYTICS PREP
DIVERGENCE ACADEMY
galvanize
GENERAL ASSEMBLY
METIS
NYC DATA SCIENCE ACADEMY
FLATIRON SCHOOL
IRON HACK
BYTE ACADEMY
BrainStation
coding temple
THINKFUL
INSIGHT
The Data Incubator

Build a Portfolio

1. Include Interesting Projects
2. Work on your Resume
3. Have a Social Media Presence
   - GitHub
   - kaggle
   - tableau
   - LinkedIn
   - Medium
4. Create Website to showcase work



How to get the Job of your Dream?

Figure: Networking

Figure: Conferences

Figure: Volunteer

References

Articles of Importance:

- How to Spot a Twitter Bot
- Q&A: How Pew Research Center identified bots on Twitter
- Bots in the Twittersphere
- How To Implement Intent Recognition With BERT
- The Complete Guide to Building a Chatbot with Deep Learning From Scratch
- Intent Classification and its Significance in Chatbot Development
- How To Go Into Data Science?
- The Ultimate Guide to Getting Started in Data Science
- 5 Best Degrees for Getting into Data Science
- How to Build a Data Science Portfolio
- Advice on Building Data Portfolio Projects

That's all folks!

Questions?