CLOUDERA

NetHope Summit 2019

MEET THE TEAM

Ifigeneia (Ifi) Derekli

Senior Solutions Engineer
Security SME Lead

Tim Spann

https://www.datainmotion.dev/

Field Engineer - Data in Motion @PaasDev

Ian Brooks

Senior Solutions Engineer

WHY ARE WE HERE?

- To introduce and demonstrate the **data lifecycle**
- Highlights:
  - It's a team sport
  - It is iterative
  - Open source, free tools can be used to do great things
- Disclaimer:
  - We are data specialists, not disaster-relief specialists
  - This is just a simple demo to showcase the data lifecycle. Not a solution proposal

DATA SCIENCE LIFECYCLE

Teams & Responsibilities

Business Analyst

Jobs role includes representing the interest of the business and defining the problem

Data Engineer

Job role includes discovering, collecting, and cleaning the data

Data Scientist

Job role includes feature engineering, feature selection, ML model building, validation, and evaluation

IT / System Admin

Jobs role includes managing the IT infrastructure, managing deployed ML models

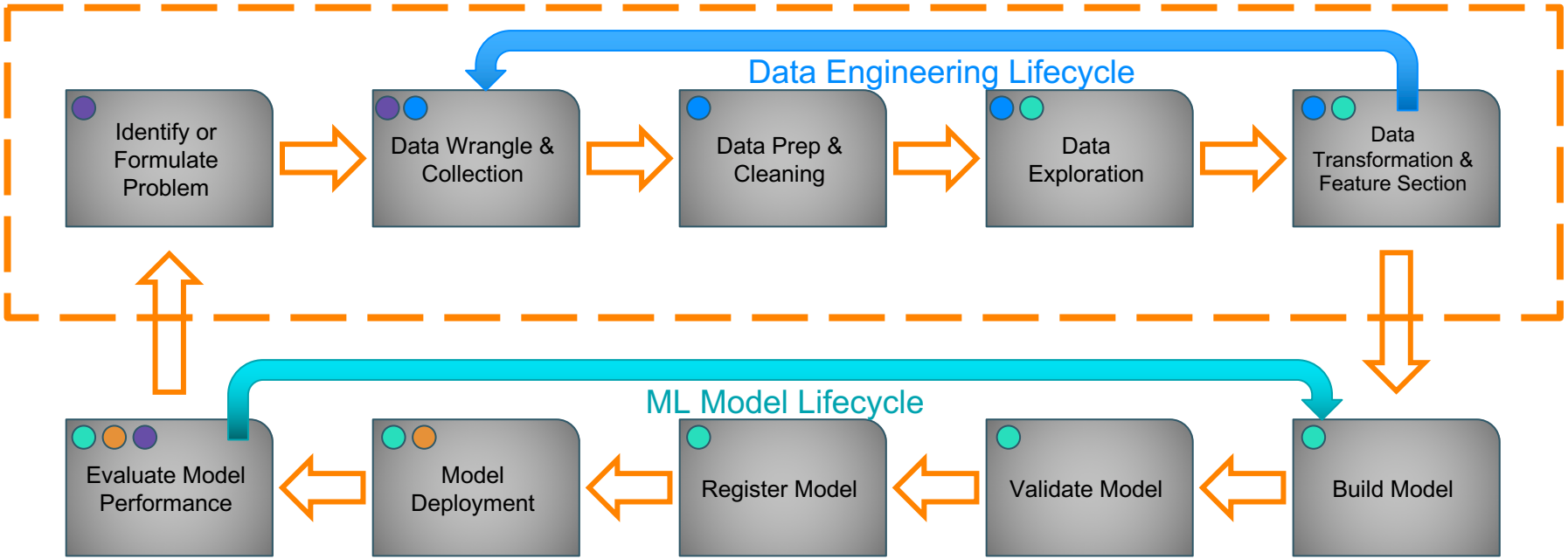
DATA SCIENCE LIFECYCLE


Business Analyst


Data Engineer


Data Scientist


IT / System Admin



FORMULATE THE PROBLEM - Iteration #1


Business Analyst

Identify or Formulate Problem

Hurricanes in Puerto Rico - we know there will be a future disaster

Goal #1: Respond to disaster faster and more effectively

- Identify location of need to send help
- Prioritize based on severity and volume

ASSUMPTIONS:

- People speak English
- Mostly everyone uses a cellphone and social media

DATA COLLECTION AND PREPARATION - Iteration #1

Step 1: Identify relevant data sources

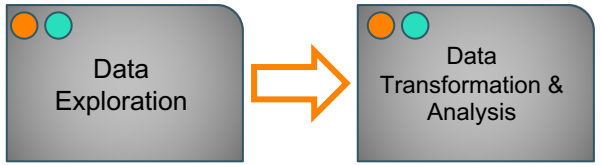
- Twitter public data (2017 Hurricane Maria data, for demo purposes)
- Demographics
- Public services data (hospitals, airpots, stores, etc)

Step 2: Ingest data to a central location

- Apache NiFi used for real time ingestion of Twitter data
- Spark used for batch ingestion of simple CSVs

DATA EXPLORATION & OUTCOMES - Iteration #1


Data Scientist



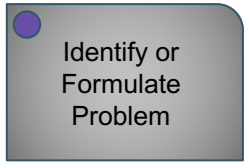
Data Exploration conclusions:

- Volume of tweets directly related to power outages
  - In correlation with density of population, temporary WiFi towers can be put in place at those locations
- NLP and Picture recognition model can be used to categorize need
  - Tree down?
  - Flooding?
  - Power outage?

Outcome:

- Real-time understanding of areas of need

FORMULATE THE PROBLEM - Iteration #2

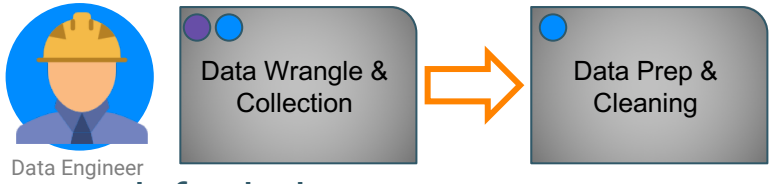


We identified a way to understand better and faster where help is needed. Let's take it a step further.

Goal #2: Improve communication between people on the ground and decision makers

- Provide an avenue for people in need to ask for help directly
- Provide an avenue for Disaster Relief Orgs to share helpful information

DATA COLLECTION AND PREPARATION - Iteration #2



Step 1: Provide a website/phone-app for people to ask for help

- Apache NiFi used to serve webapp and collect data
- Data collected: location, what's the problem, pictures (our personal inputs, for demo purposes)

Step 2: Provide a website to send information back

- Nearest WiFi hotspot / hospital / etc
- Where is water / food / medics available now? In an hour? Tomorrow?

DATA EXPLORATION & OUTCOMES - Iteration #2

Outcomes:

- More informed decisions on where to send resources and help by correlating crowdsourced data from webapp and social media
- Real-time communication of where to find help with people in need

THANK YOU

CLOUDERA