

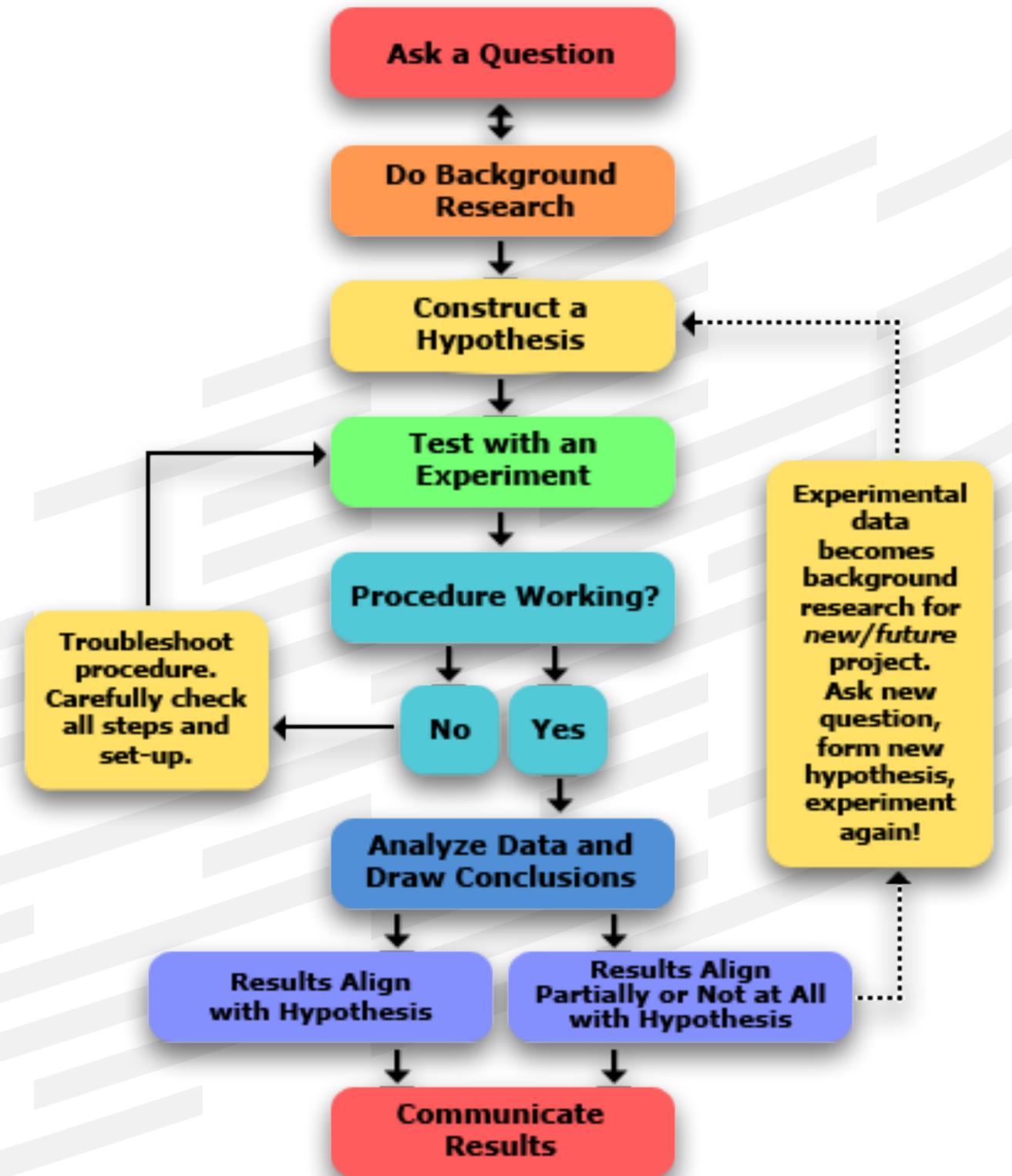
# End-to-End Data Science



**Billy OKAL**  
**Data Science Africa (Accra) 2019**

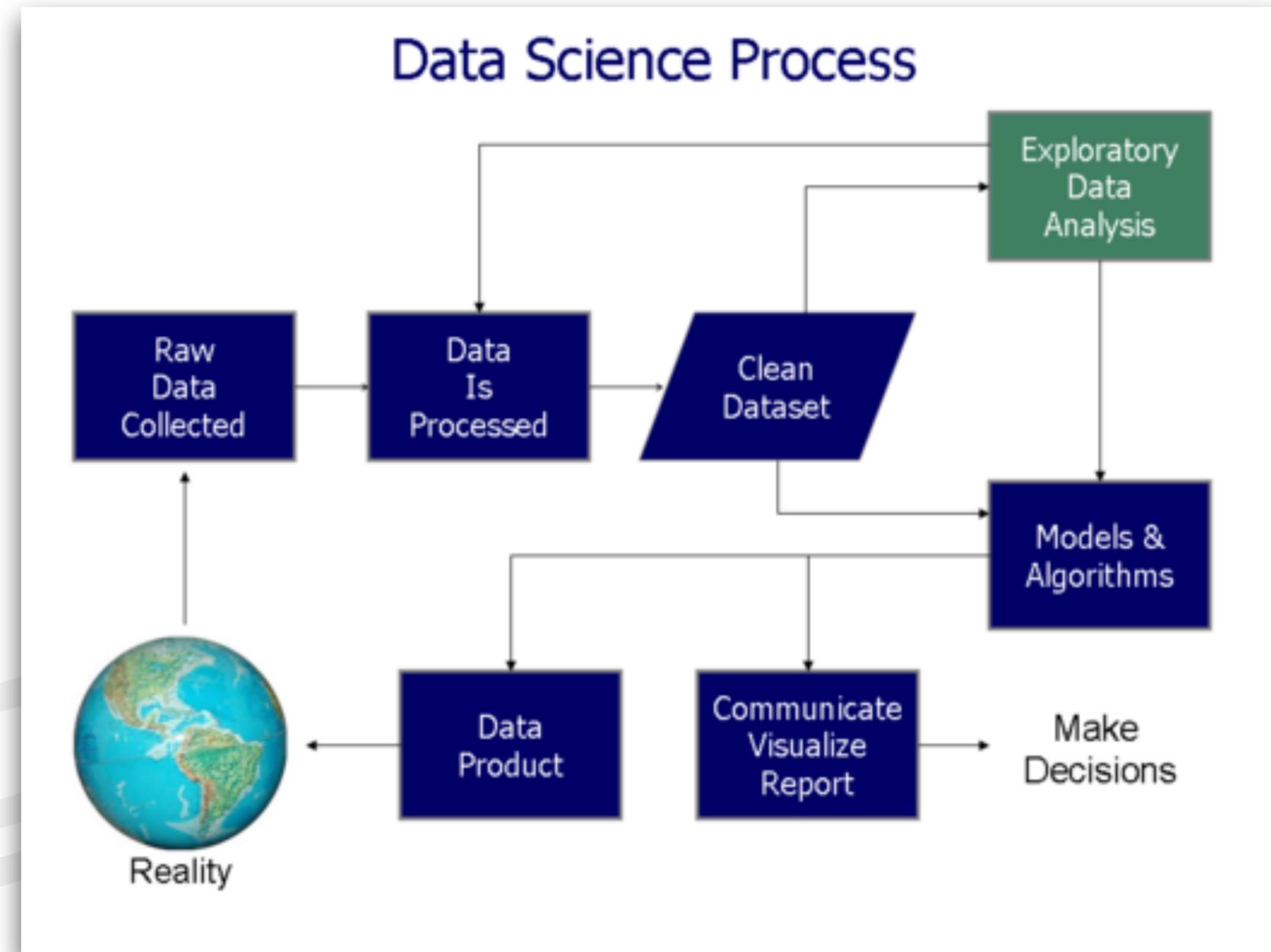
# The Scientific Method

- Data is collected after a question is posed, implications?
- Hypothesize first
- Cleaning can be reduced to filtering for what we want, because we know what we want.



# Data Science Process

- Data is collected often **before** the question is posed, implications?
- Cleaning is not just filtering for what we want, because **we don't know what we want yet**
- Hypothesis comes **after** data? chicken and egg again
- Maintenance is a first class task
- Updates to deployments



# end to end idiom

## Definition of *end to end*

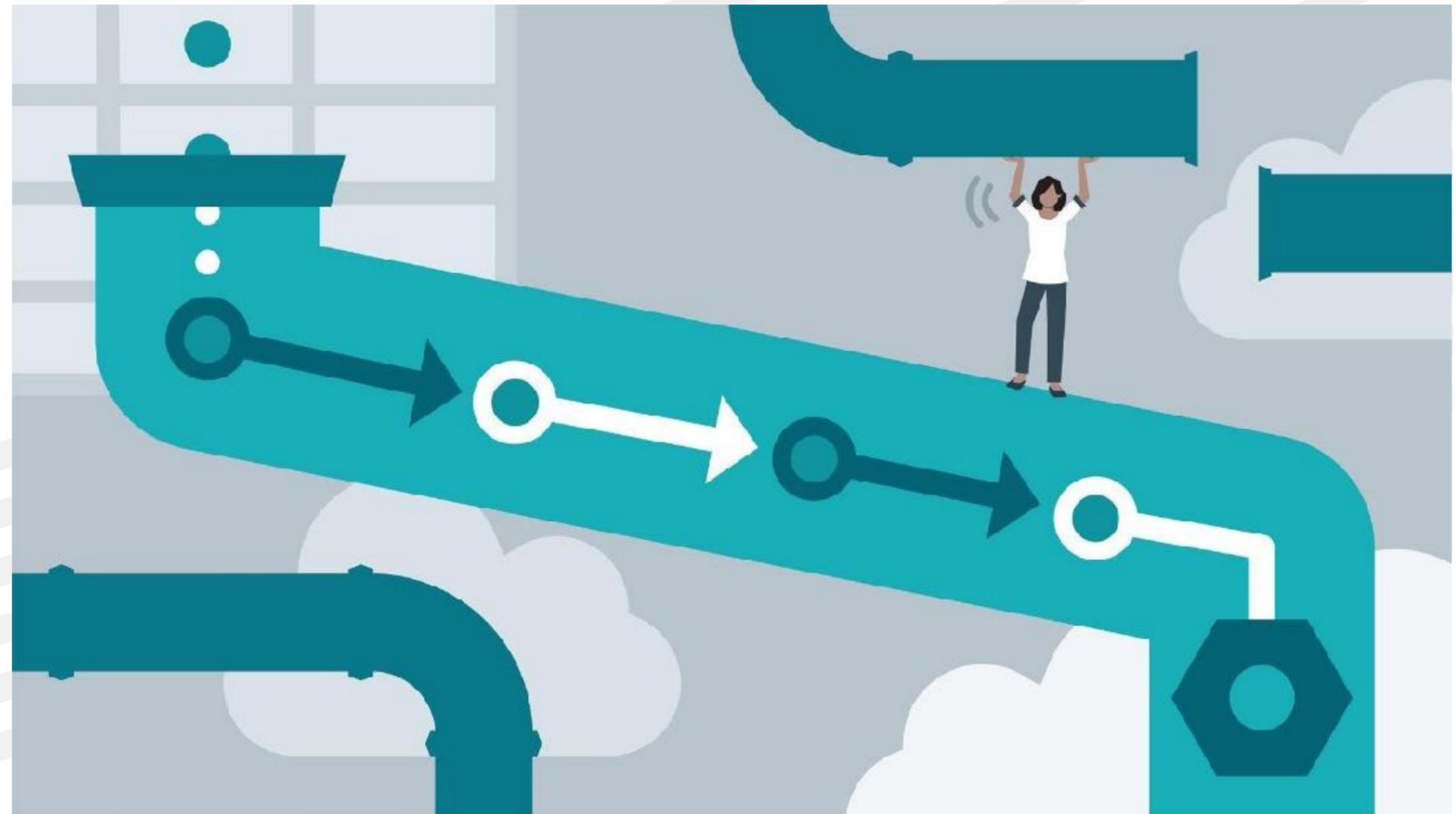
: with ends touching each other

// Put the two small tables *end to end*.

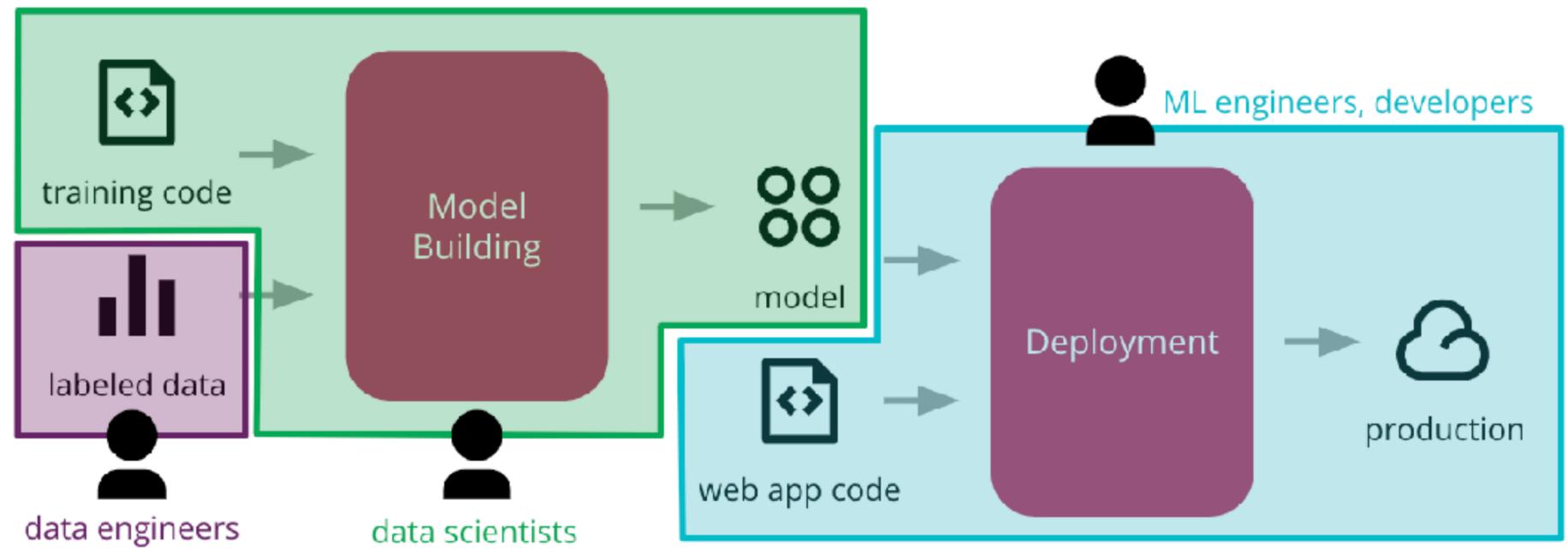
// railroad cars lined up *end to end*

# What are the Ends?

- Closed loop
- Ends of each section touching
- No leaking of spurious assumptions in between
- Biases come in at controlled locations
- What skillsets are needed?

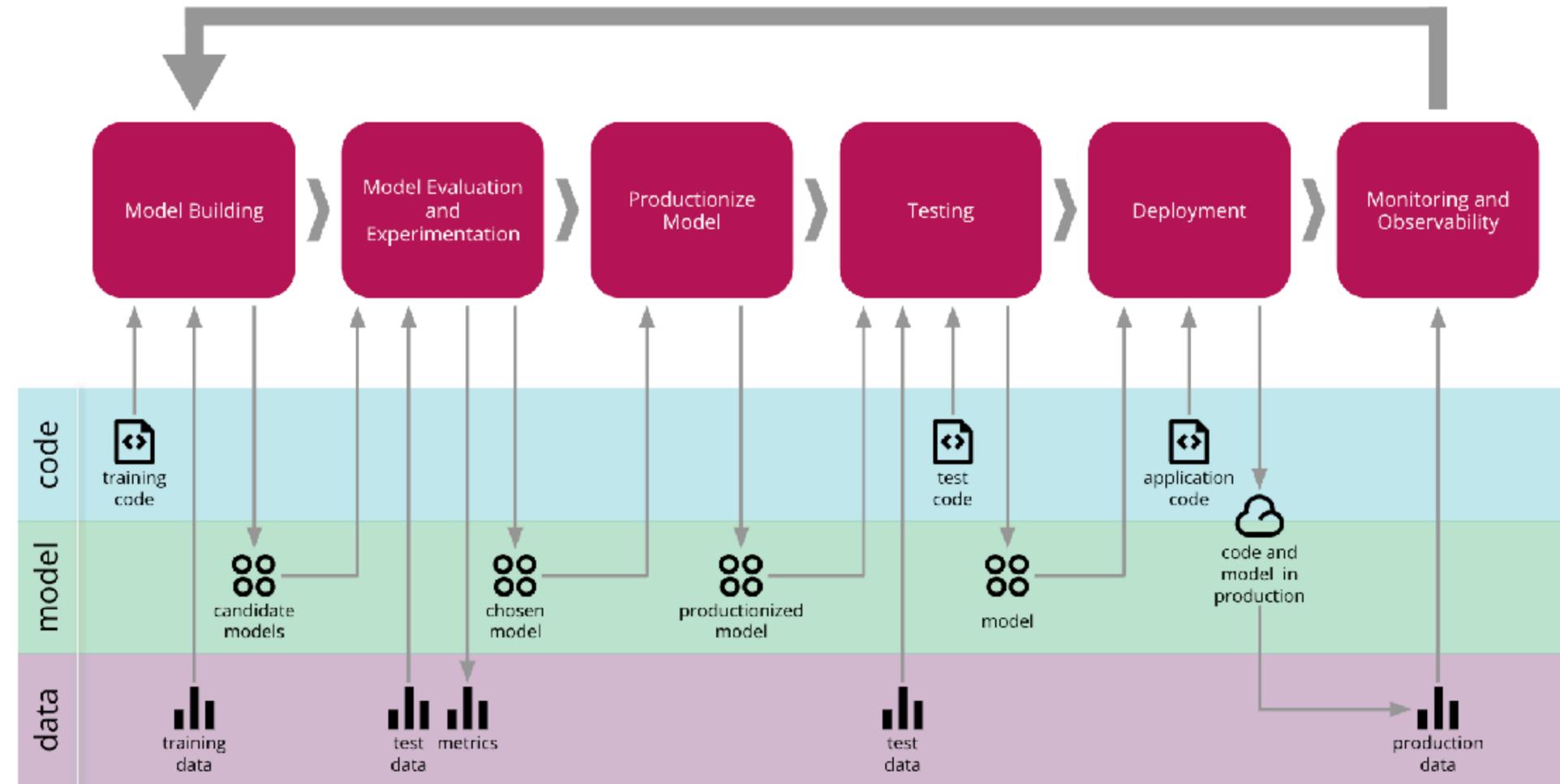


# Pipeline backbone



## Systems

- Mature, production software at all stages
- Build flexibility for fast prototyping



# End-to-End DS Stories

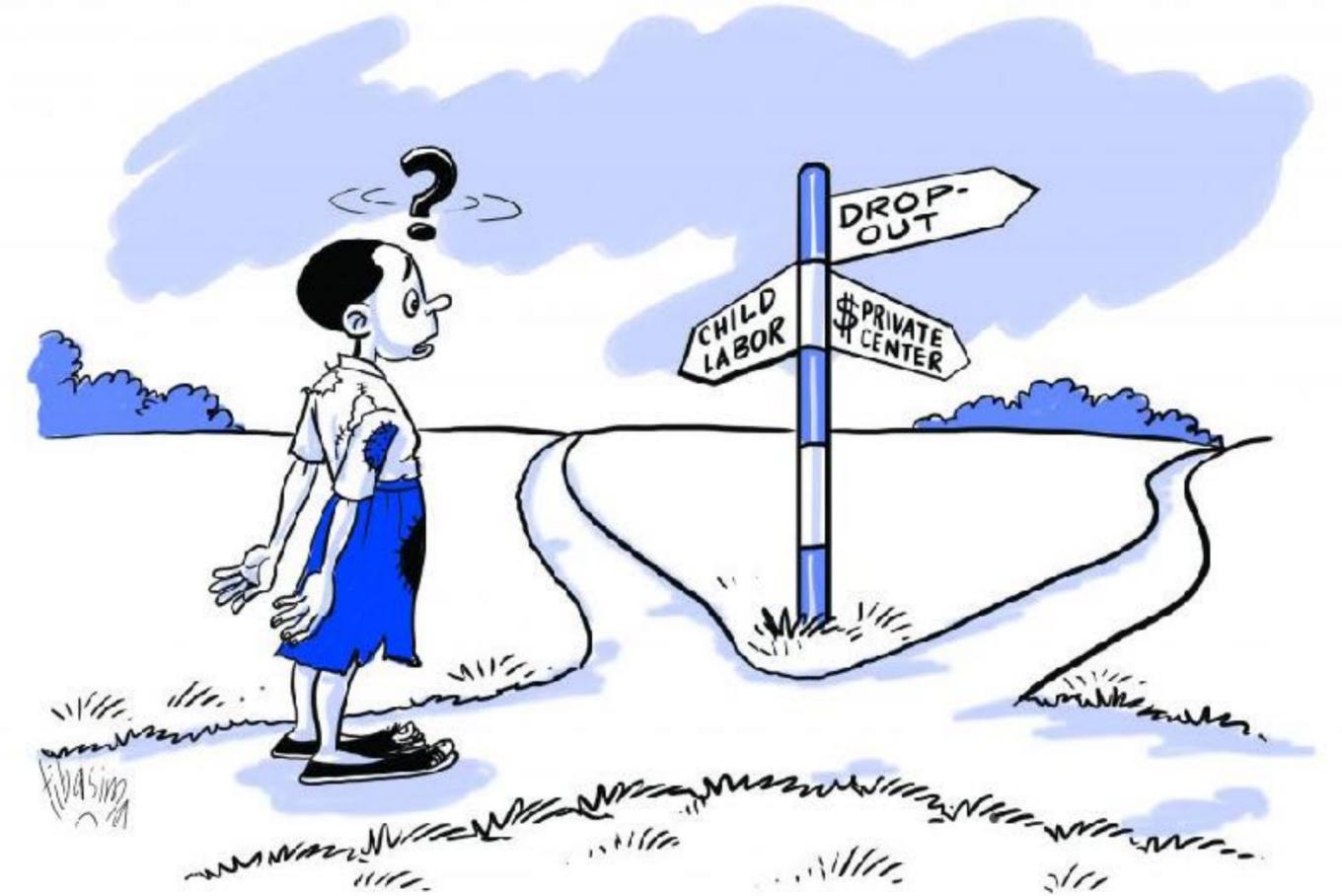


SCHOOL

Kibasing

# End-to-End DS Stories — School Dropouts

- Study reasons for school dropouts, esp. amongst girls
- Built tooling (web and mobile application) for client to use the analysis and prediction (ensemble methods)



## An Ensemble Predictive Model Based Prototype for Student Drop-out in Secondary Schools

Neema Mduma <sup>1\*</sup>, Khamisi Kalegele <sup>2</sup>, Dina Machuve <sup>1</sup>

<sup>1</sup> *The Nelson Mandela African Institution of Science and Technology, TANZANIA*

<sup>2</sup> *Tanzania Commission for Science and Technology, TANZANIA*

\*Corresponding Author: [mduman@nm-aist.ac.tz](mailto:mduman@nm-aist.ac.tz)

**Citation:** Mduma, N., Kalegele, K. and Machuve, D. (2019). An Ensemble Predictive Model Based Prototype for Student Drop-out in Secondary Schools. *Journal of Information Systems Engineering & Management*, 4(3), em0094. <https://doi.org/10.29333/jisem/5893>

A screenshot of the BakiShule web application interface. The form is titled 'Student information' and includes several input fields: 'Select the Gender' (with 'Female' selected), 'Student age' (with '20' entered), 'Household meals per day' (with '2' selected), 'Reading book status' (with 'Student read a book with parents' selected), 'Parent checking exercise status' (with 'Does not check the exercise book' selected), and 'Parent and teacher relation' (partially visible).



# End-to-End DS Stories — Wildlife Conservation and IoT

Beginning of a DS project

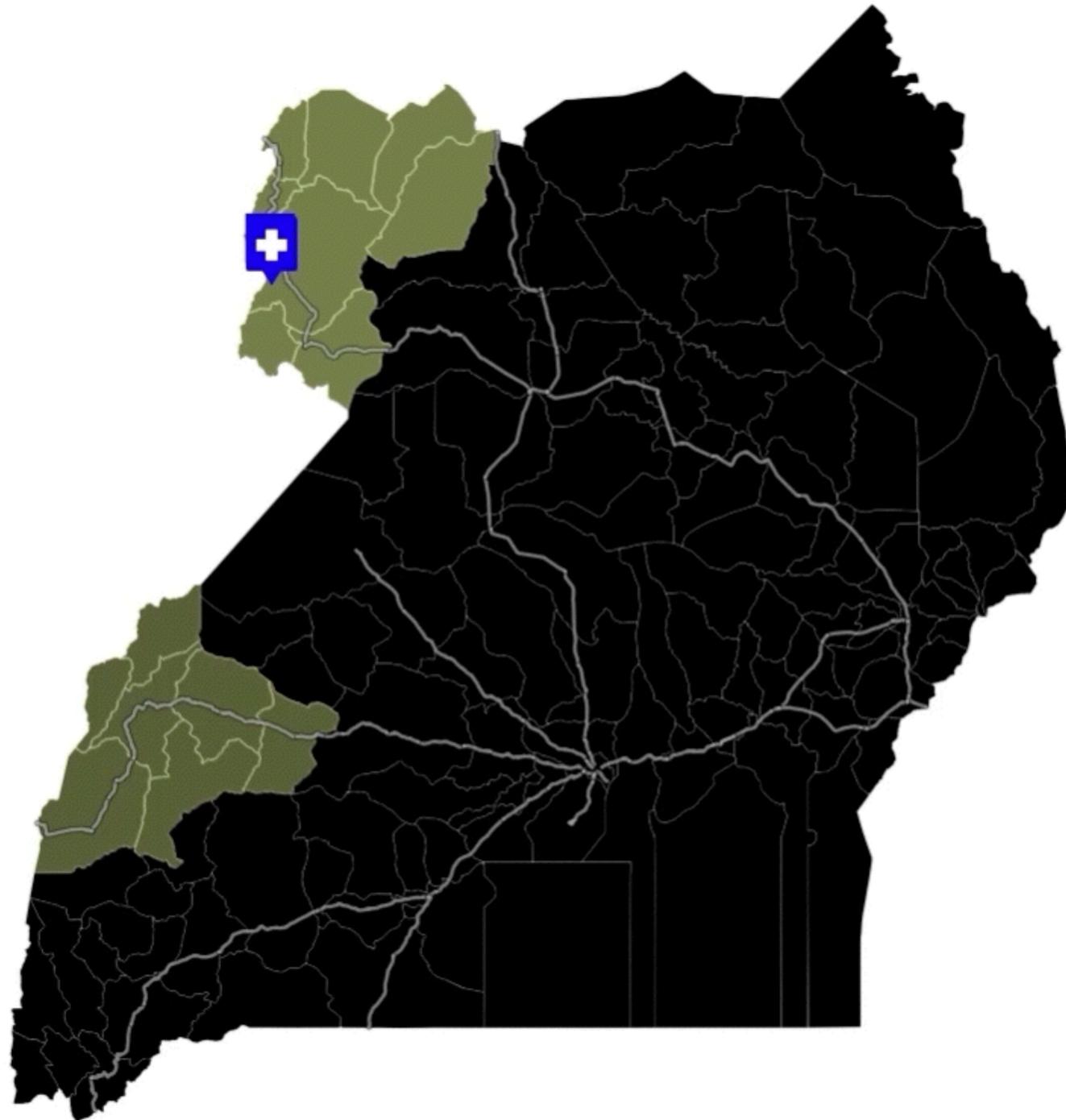
Students building IoT instruments for use in data collection



Hunting for data

Relationship with stakeholders

# End-to-End DS Stories — Ambulance Service Monitoring



2018-02-24 00:42:07

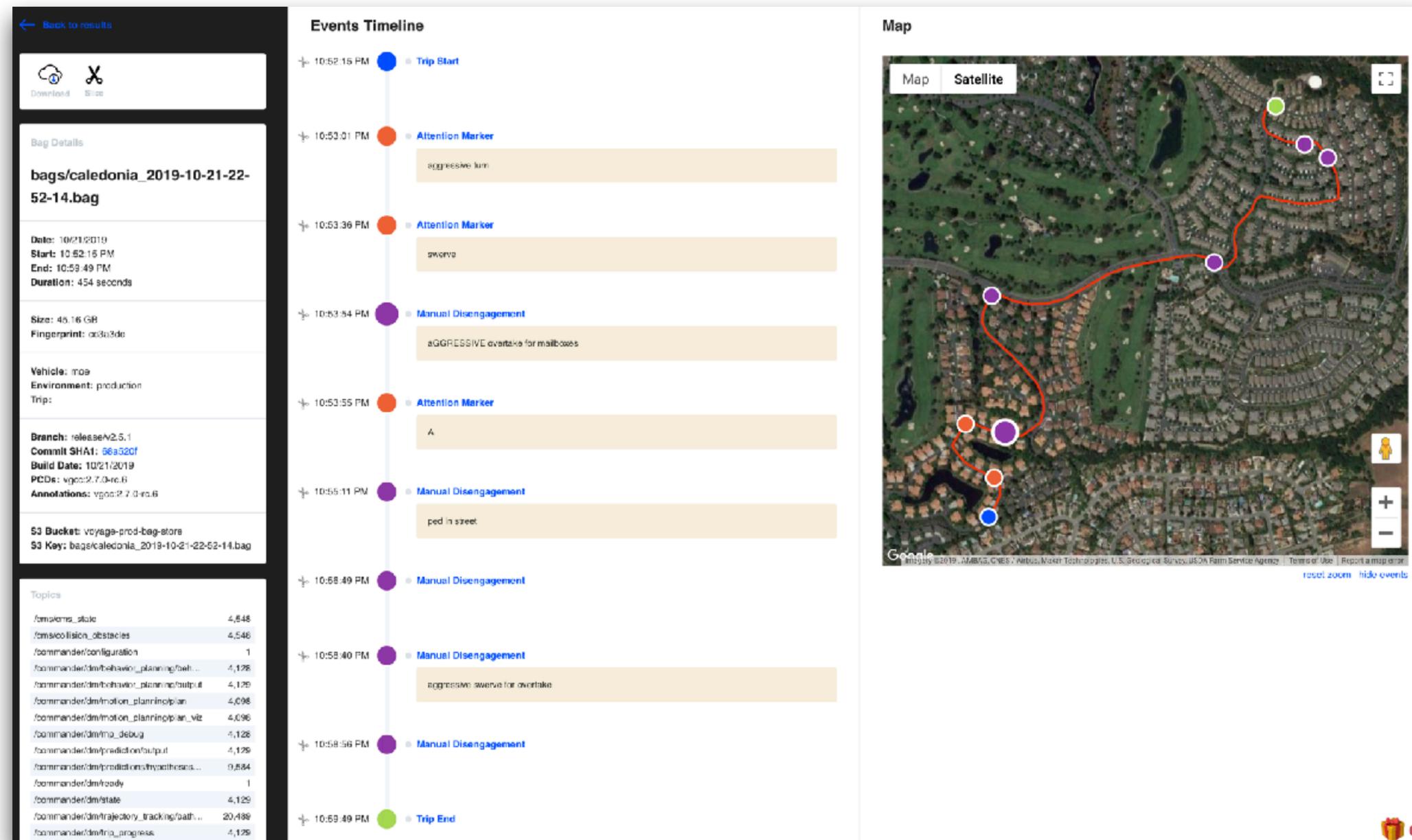




# AV Data Science

Mountains of happenstance data from regular logging during operations

Alerts for possible targets/labels of interesting events



**Service Notifier** APP 17:46

**Near-Intervention Event Identified (418561)**

Date: Monday, October 21st 15:18:10

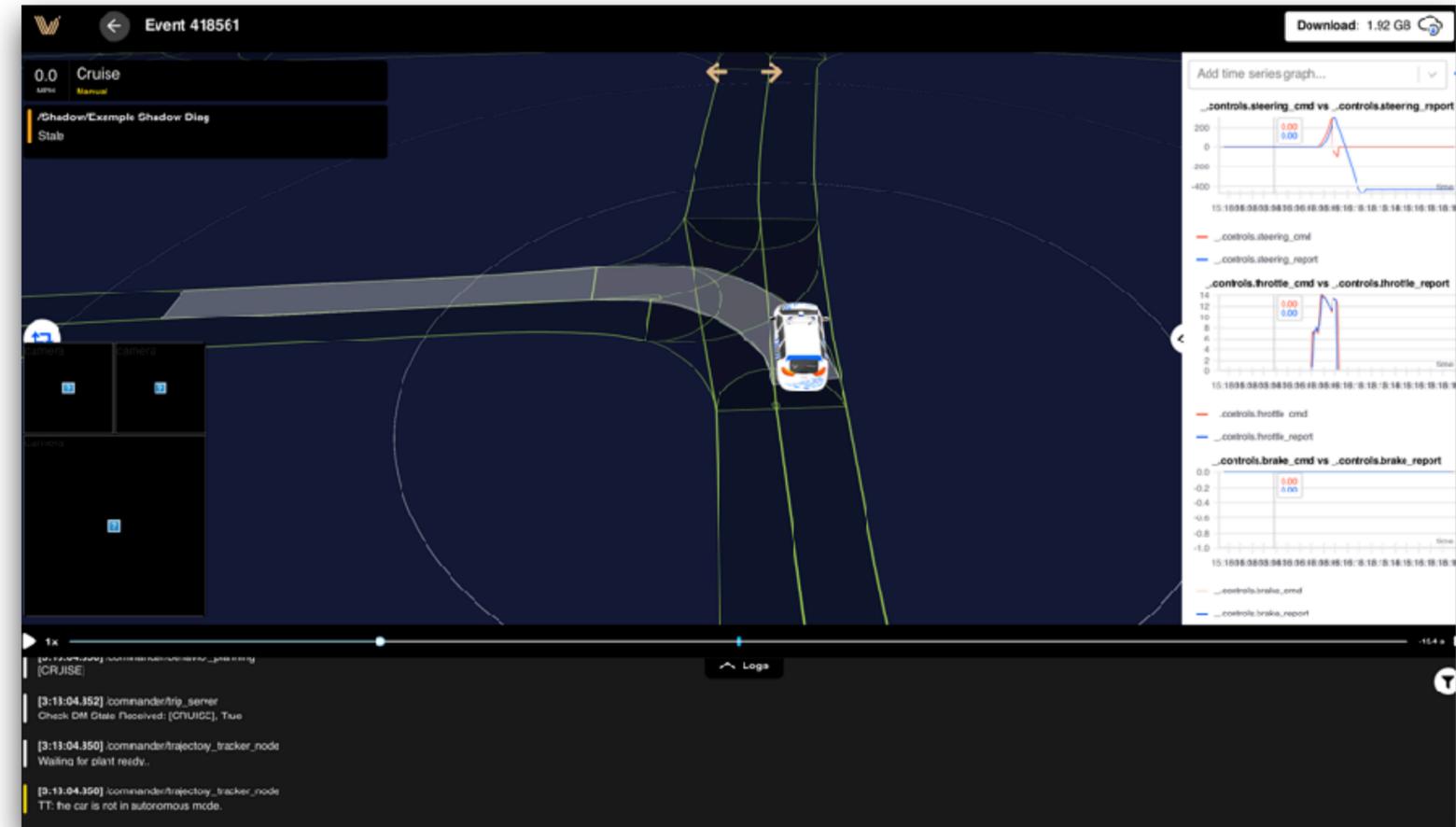
Comment: "actor started to overtake ego as ego was deciding which route to take around Myras villa #near-intervention"

[See Event Details](#)

# AV Data Science

## Variety

- Logs
- Metrics
- Raw sensor data
- Diagnostics data
- Decisions (predictions, plans, commands)



The screenshot shows a detailed view of an event (ID: 418561) in a data analysis dashboard. The main content area is divided into several sections:

- Comment:** actor started to overtake ego as ego was deciding which route to take around Myras villa (near-intervention)
- near-intervention:** A tag indicating the event type.
- Bag Context:** A timeline of events:
  - 00:17:44 PM: Trip Start
  - 00:18:10 PM: **NO** Manual Disengagement (actor started to overtake ego as ego was deciding which route to take around Myras villa (near-intervention))
  - 00:18:59 PM: **ALL** Attention Marker (small desk for actor in opp lane)
- Location:** A satellite map view showing the vehicle's position on a residential street.
- Details:** Metadata including ID (418561), Created (October 21st 2019, 3:18:10 pm), Deployment (vgll), Bag (24188), and Metadata UUID (801684cd-4f8b-4d0b-a506-406518183069).
- Build Info:** Commander (releaseV2.5.1), Commander Timestamp (October 21st 2019, 5:02:54 am), Annotations (vgll:2.5.6), PCDS (vgll:2.5.6), and Vehicle (milhouse).
- Buttons:** 'Open in Triage' and 'Submit for Annotations score'.
- Bag Slices:** A summary of the bag containing 418561 topics and 1.92 GB of data.

# Summary

- Data comes first, hypothesis second
- Iterative process, relationships, closed loop (with respect to stakeholders)
  - Do not use enhanced interrogation on data, go back to stakeholders, rethink hypotheses
- Communication of results, positive and negative. Good visualization, understandable by users
- Plan for and account for maintenance — no one-off deployments in the real world
- Take feedback



# We are hiring

[voyage.auto/careers](https://voyage.auto/careers)