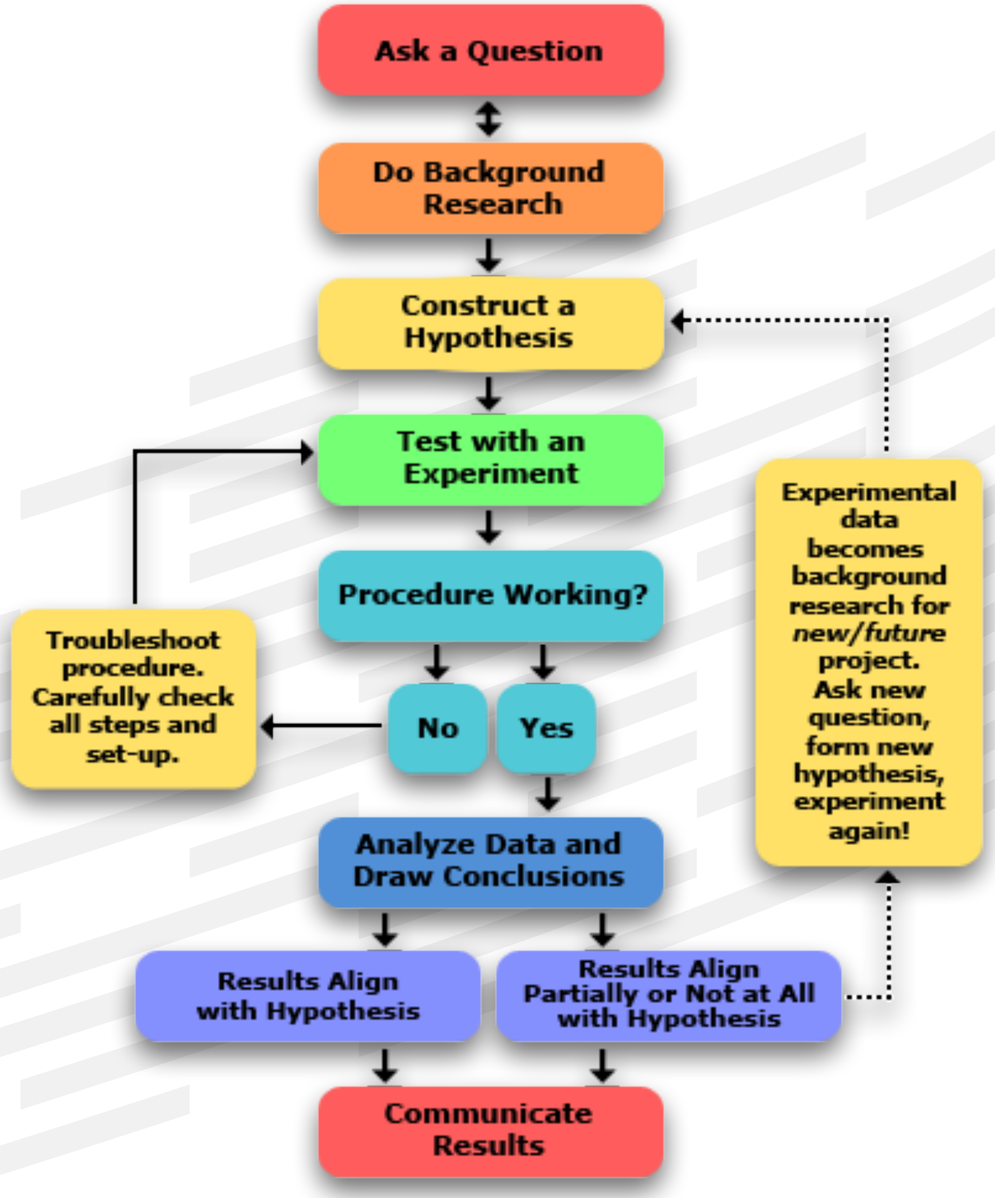
End-to-End Data Science

Billy OKAL
Data Science Africa (Accra) 2019

The Scientific Method

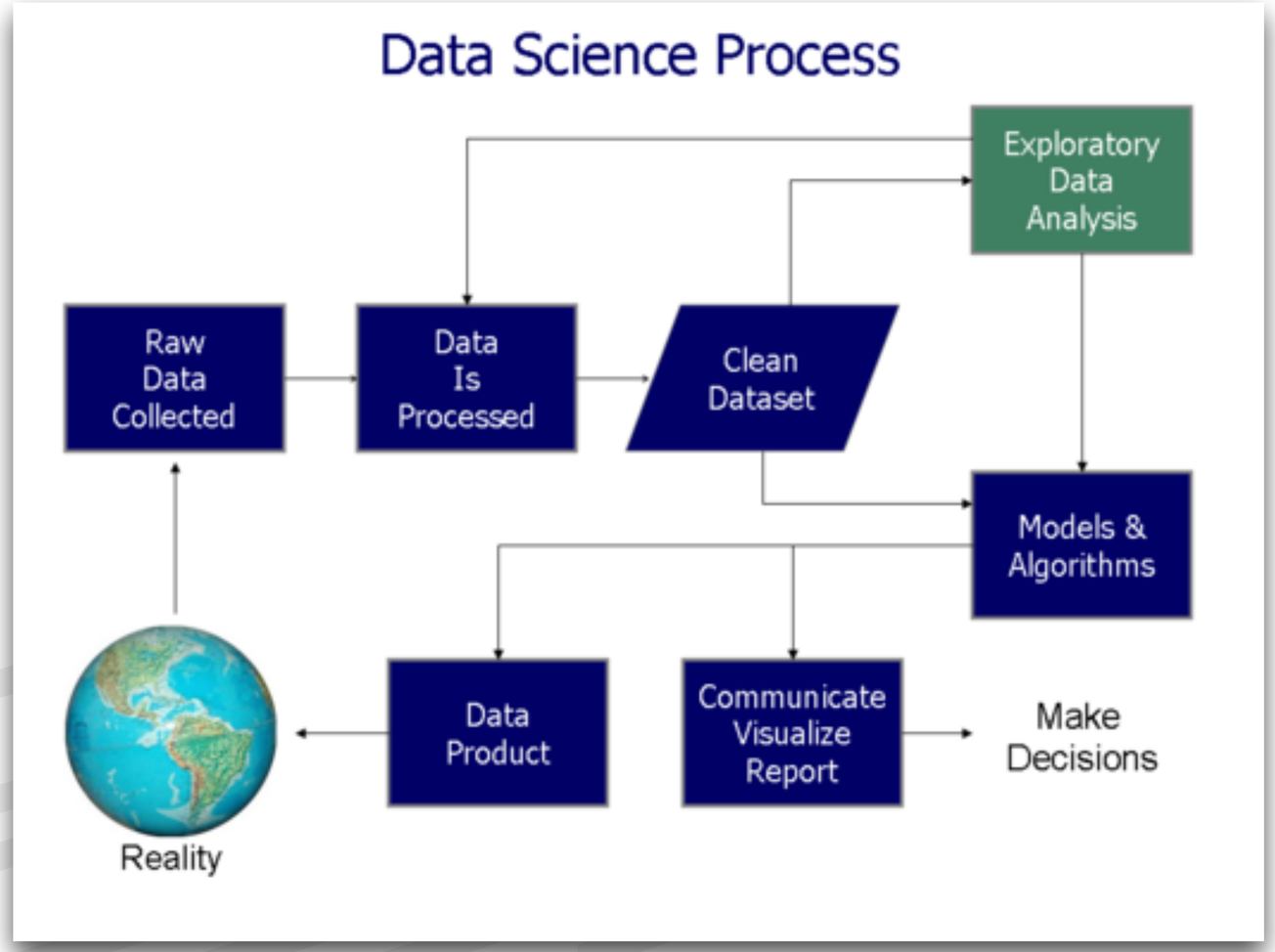
- Data is collected after a question is posed, implications?
  - Hypothesize first
  - Cleaning can be reduced to filtering for what we want, because we know what we want.



https://www.sciencebuddies.org/science-fair-projects/science-fair/steps-of-the-scientific-method

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



https://www.kdnuggets.com/2016/03/data-science-process.html

end to end idiom

Definition of *end to end*

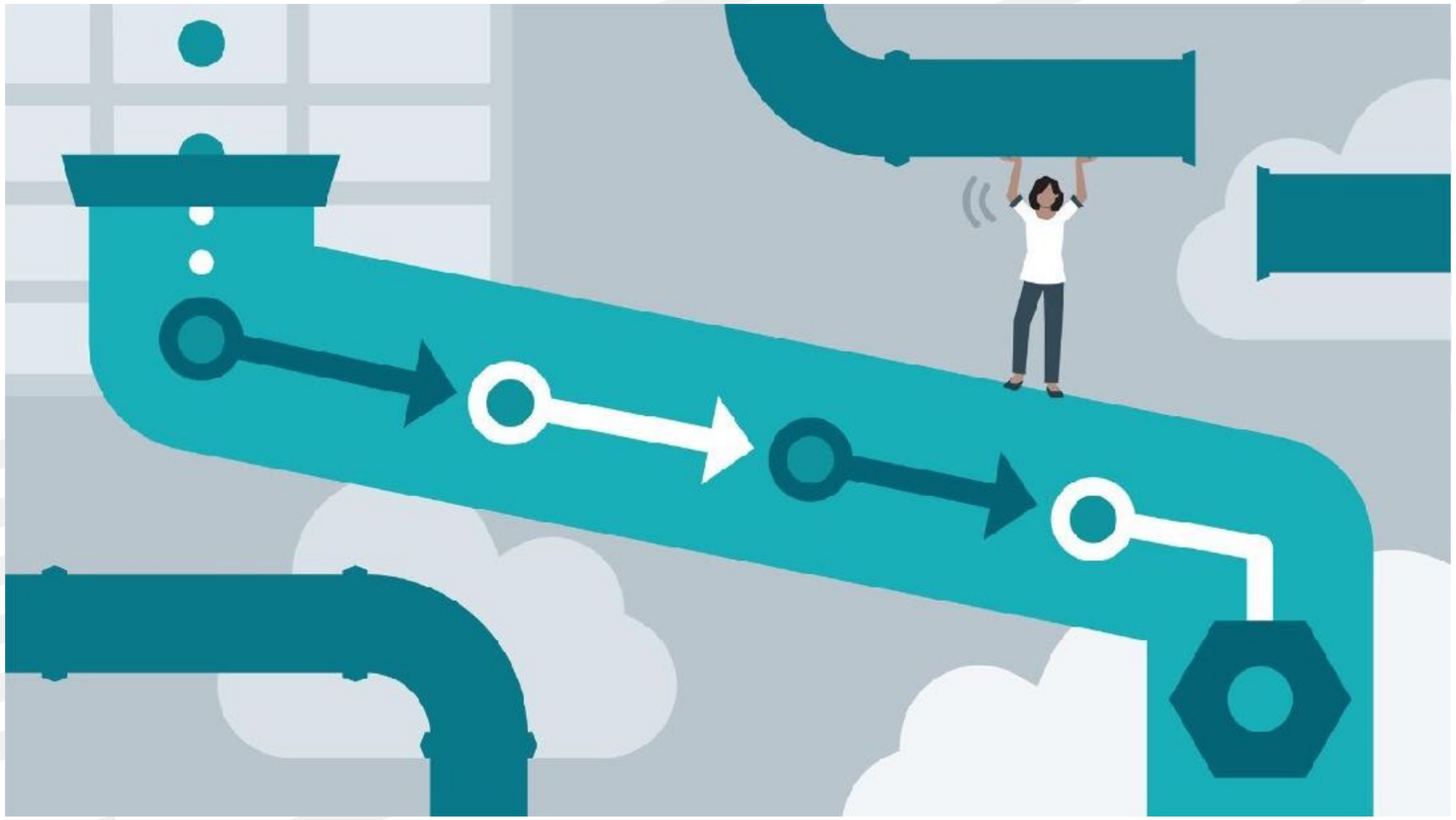
: with ends touching each other

// 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?

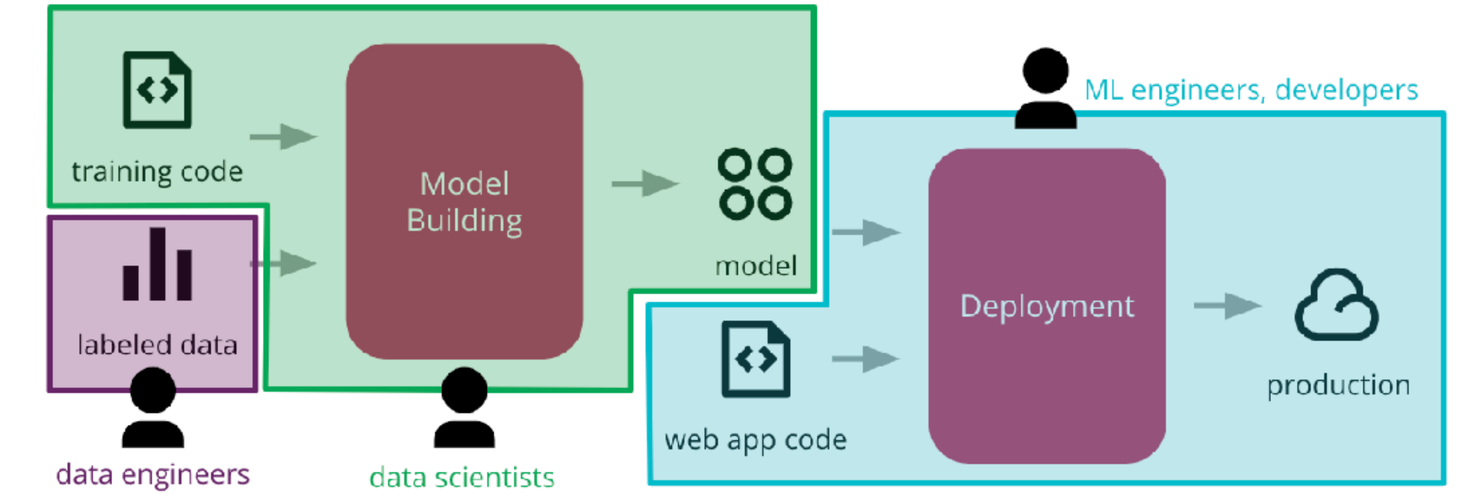


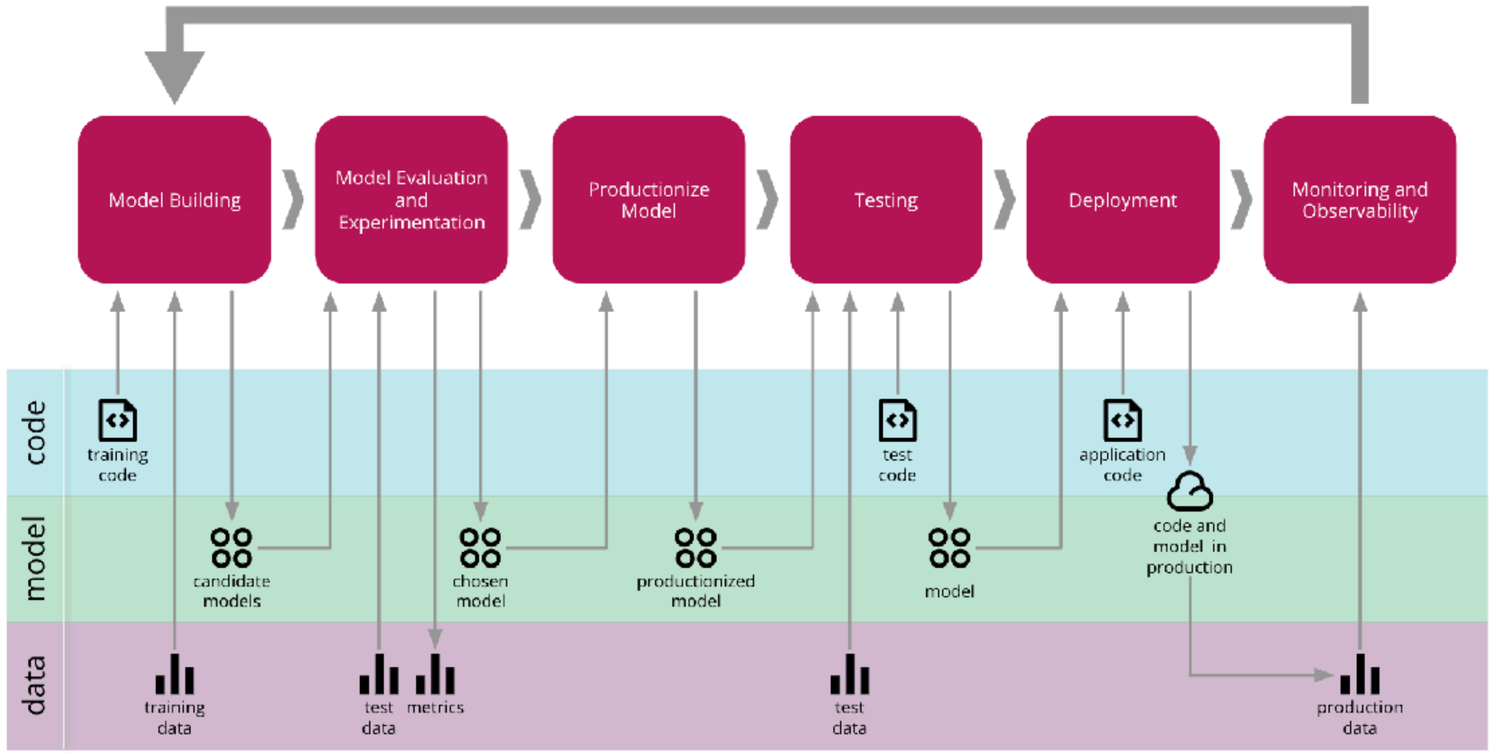
https://www.linkedin.com/learning/data-science-on-google-cloud-platform-building-data-pipelines

Pipeline backbone

Systems

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





https://martinfowler.com/articles/cd4ml.html

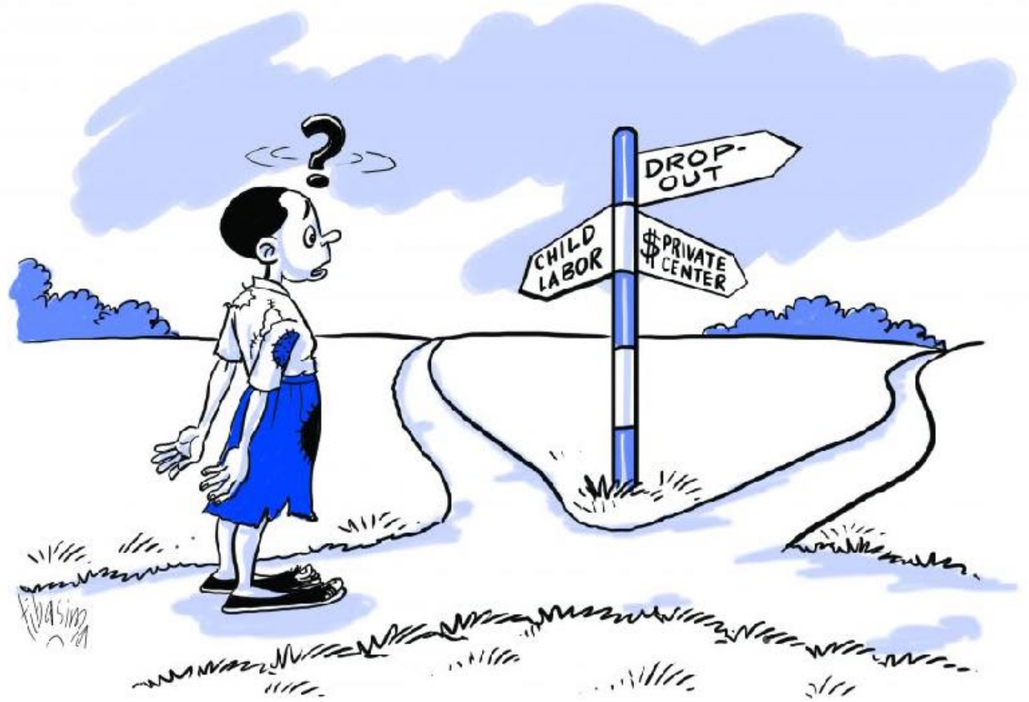
https://www.sysml.cc

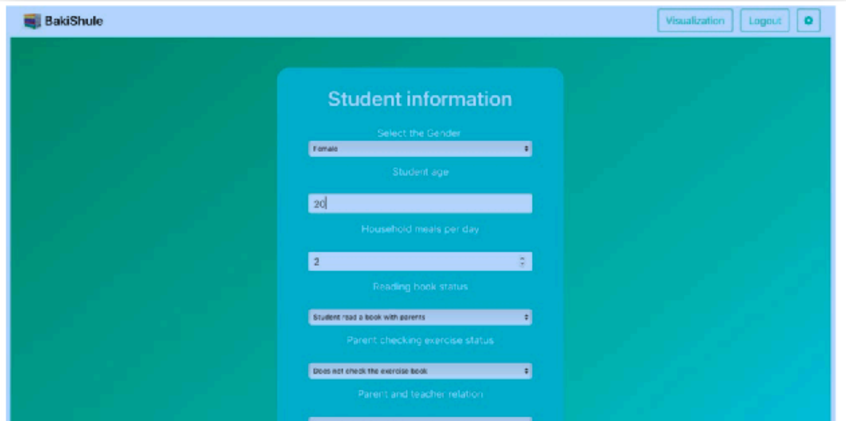
End-to-End DS Stories


SCHOOL

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 [1]*, Khamisi Kalegele [2], Dina Machuve [1]

[1] *The Nelson Mandela African Institution of Science and Technology*, TANZANIA
[2] *Tanzania Commission for Science and Technology*, TANZANIA

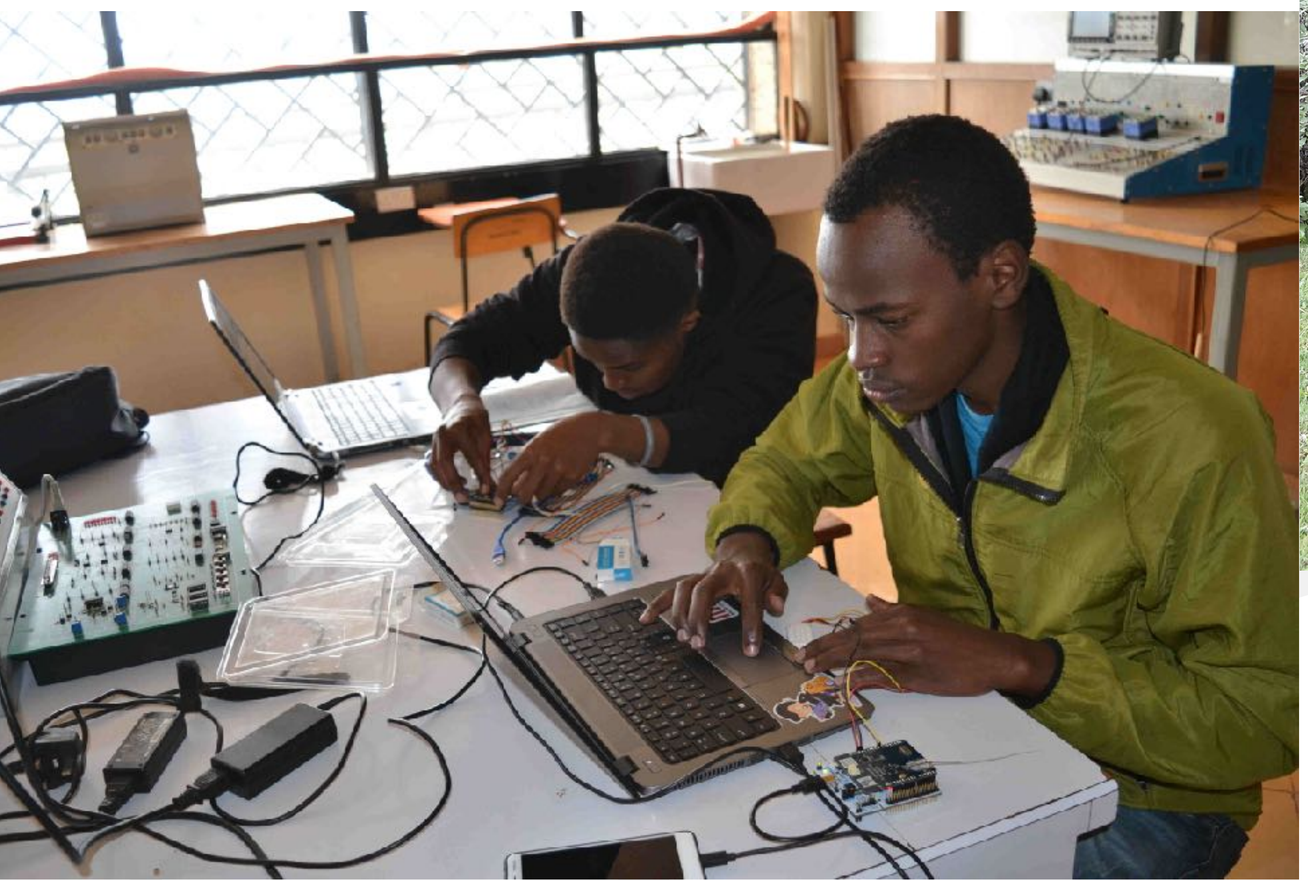
***Corresponding Author:** mduman@nm-aist.ac.tz

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

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

Beginning of a DS project

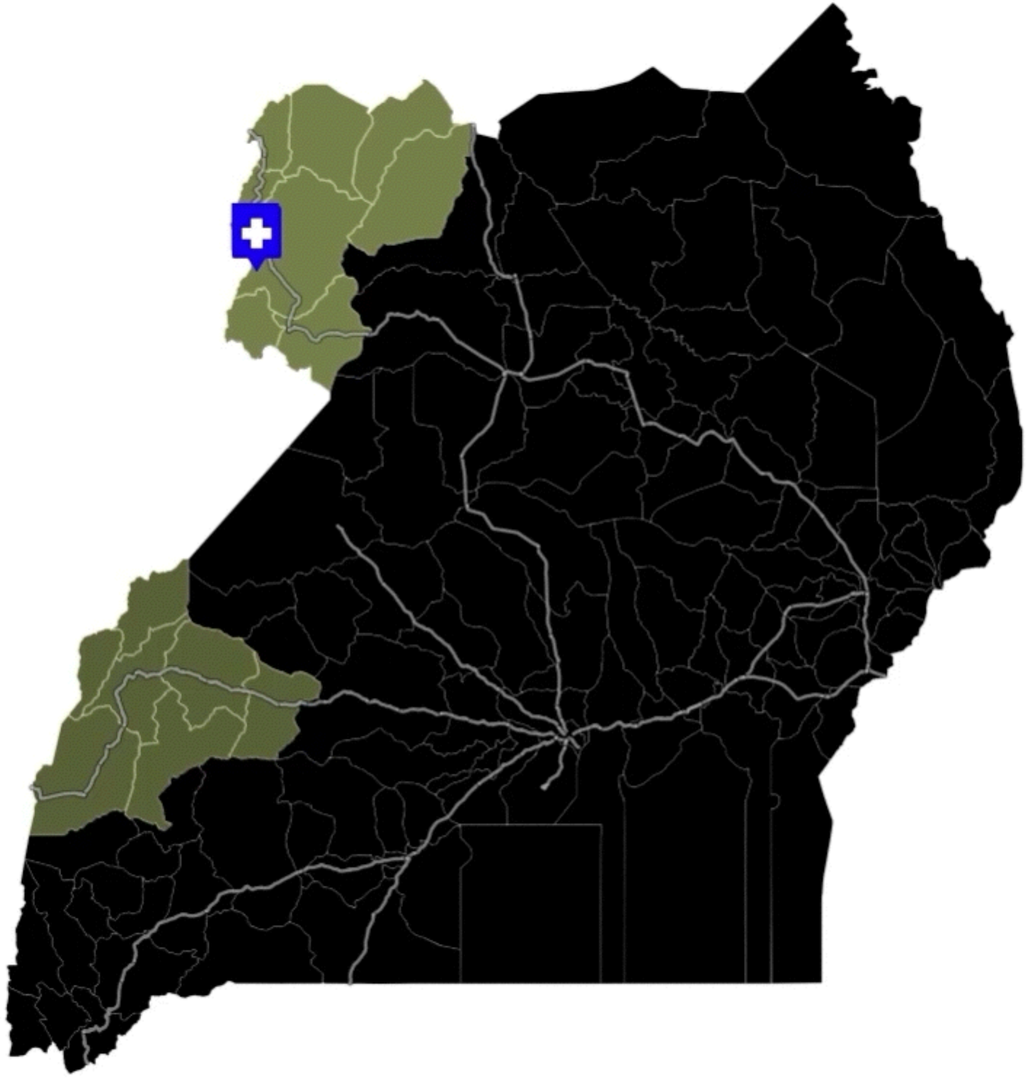
Students building IoT instruments for use in data collection



Hunting for data

Relationship with stakeholders

End-to-End DS Stories — Ambulance Service Monitoring



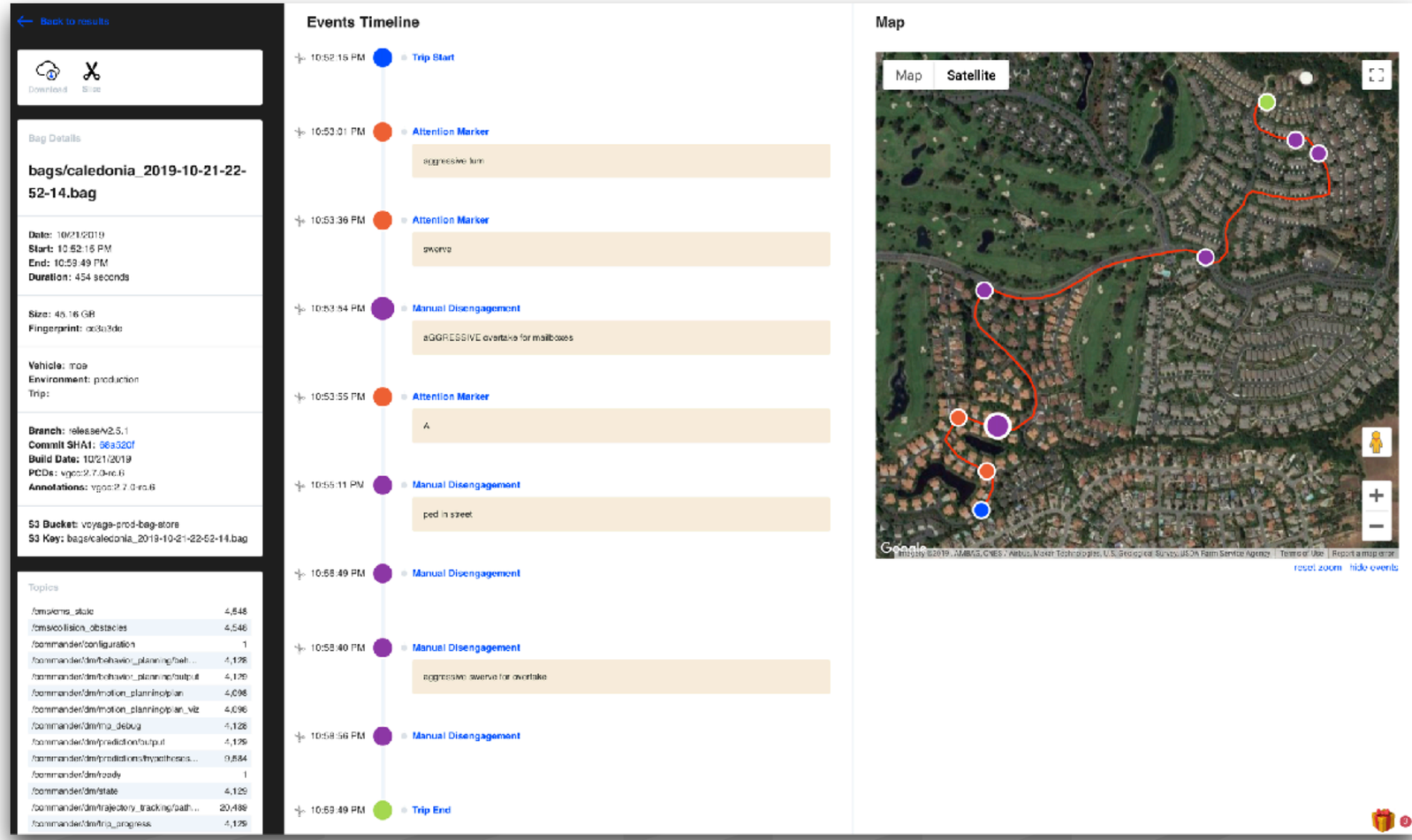
2018-02-24 00:42:07

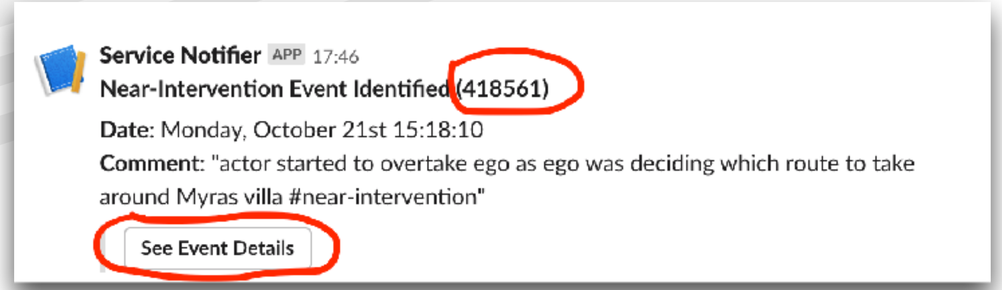

23 GATORS

AV Data Science

Mountains of happenstance data from regular logging during operations

Alerts for possible targets/labels of interesting events

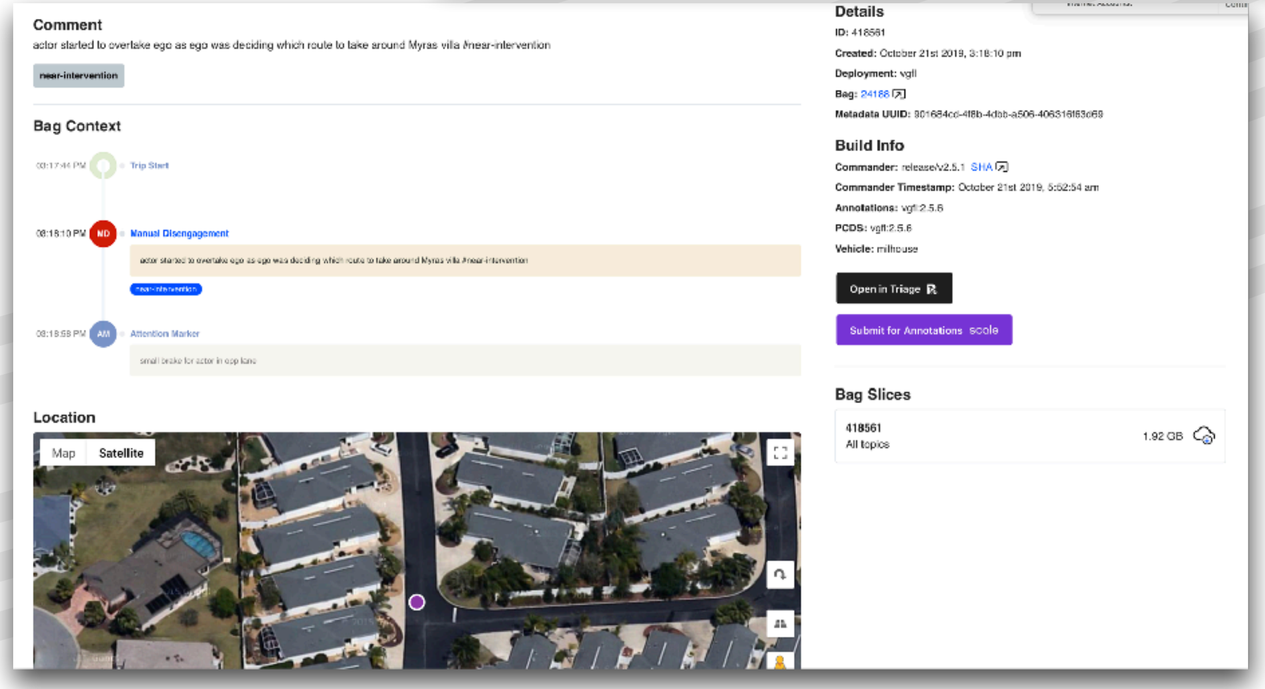




AV Data Science

Variety

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



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

voyage

We are hiring

voyage.auto/careers