The Data Science Life Cycle

and a sane process for navigating it

Why do we need this?

87% of projects don’t make it to production

- Cycle time is too high
- Solutions don’t address needs
- Lack of transparency
- Lack of reproducibility
- Duplicated efforts
- Limited collaboration
- Bad assumptions go unnoticed
- Fragile deployments
- Hidden tech debt from ML

Data Scientists Get

- Faster iteration time
- Less repeated work
- Less frustration from work going to waste
- Improved Collaboration
- Higher experiment success rates (due to issues being caught earlier in the process)
- More engagement from business partners

IT + Business Get

- Greater transparency and trust in results
- More consistency in delivery and results
- Smoother handoffs to put value into production
- Closer collaboration with data scientists (which leads to more value)
- Auditability

Design Goals

Lightweight for easy adoption

Flexible to adapt to differing needs

Minimally opinionated

Minimally Opinionated

Make as few assumptions as possible about

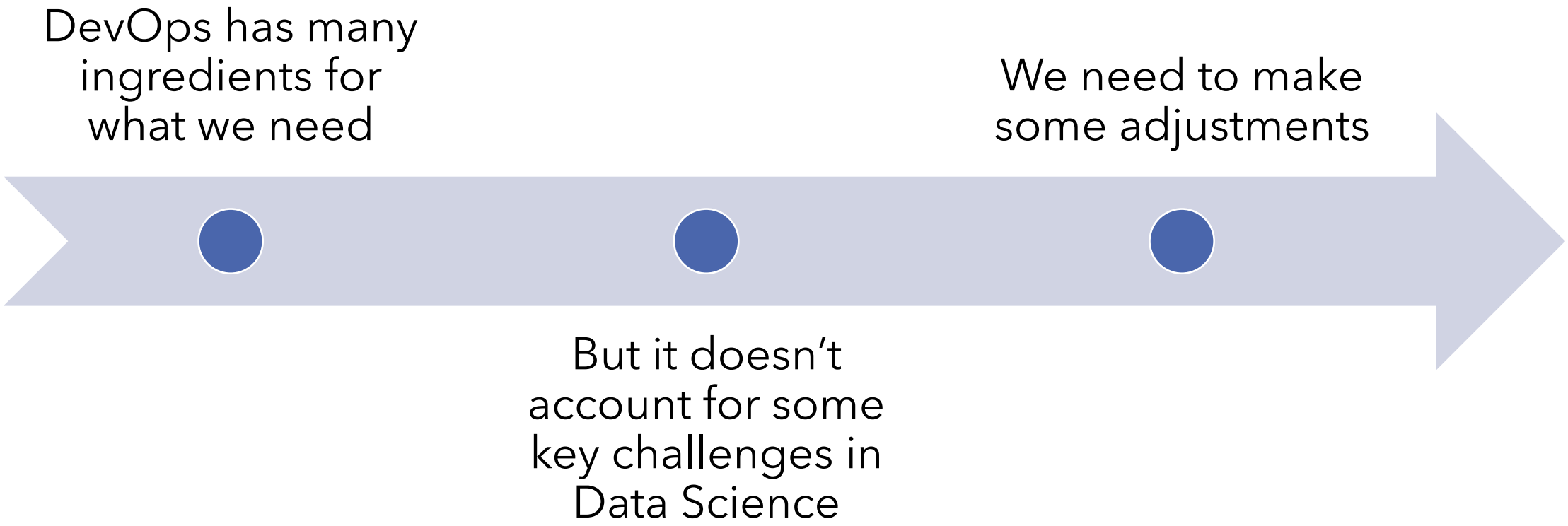
- Organization and team structures
- Type of problem being solved
- Technology Choices

This sounds familiar…

Software engineering faced these challenges before

We solved it with DevOps

So why not just use DevOps?



Engineers learn in order to build, whereas scientists build in order to learn

- Fred Brooks, The Mythical Man Month

Key Differences Between Software and Data Science

Software Engineering

- Outcome is known in advance
- Deterministic behavior
- Code is the only source of changes
- Linear progress

Data Science

- Outcome cannot be known in advanced, only estimated
- Stochastic behavior
- Changes can come from code, models, or data
- Iterative progress

Data Science Outputs

data product

application input

We need a solution that supports us all the way from **Idea** to **Value**

Doesn't This Exist?

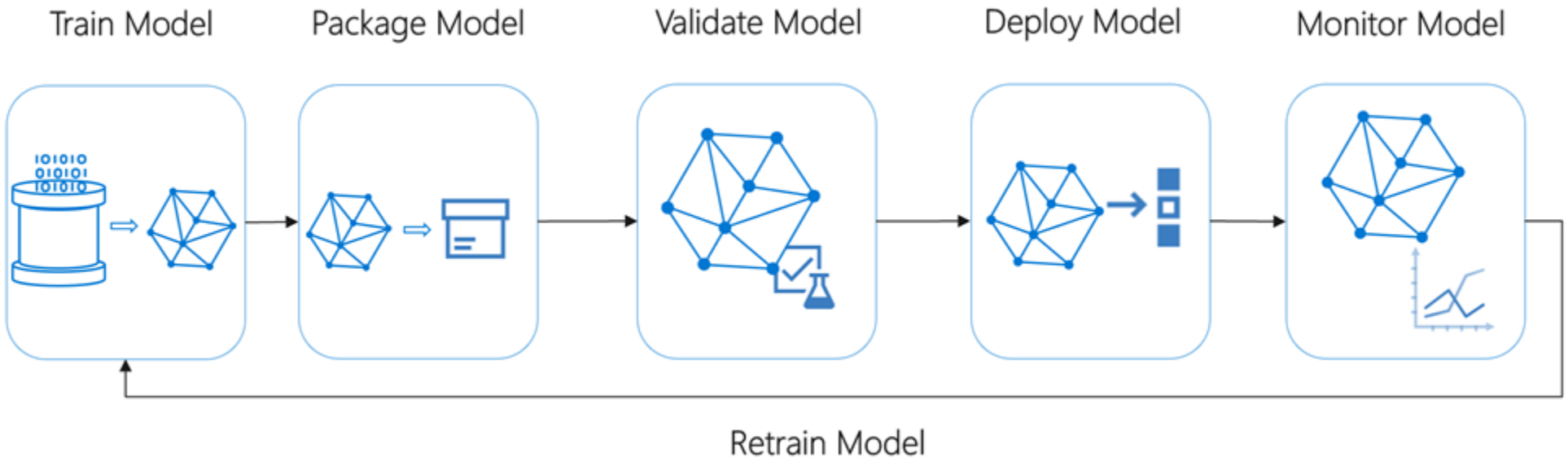
PARTS OF IT DO, BUT IT'S INCOMPLETE AND NOT STANDARDIZED

THE CURRENT TOOLS AVAILABLE EACH ONLY COVER A PART OF THE PROCESS

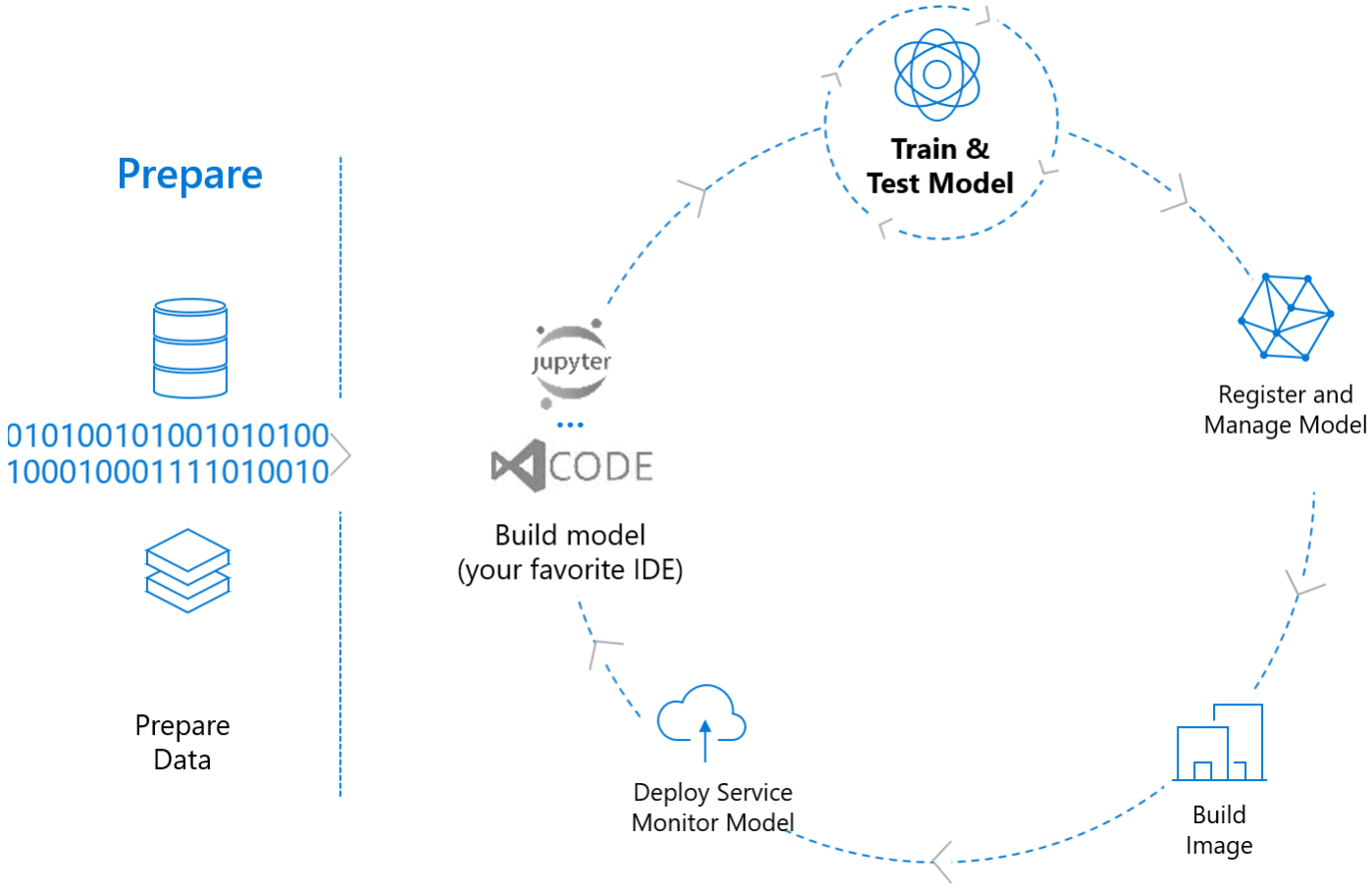
TEAMS AND ORGANIZATIONS HAVE TO FIGURE OUT HOW TO PUT EVERYTHING TOGETHER (WHICH IS HARD)

EVERYONE IS TRYING TO SOLVE THIS FROM SCRATCH

MLOps


Train Model
Package Model
Validate Model
Deploy Model
Monitor Model
Retrain Model

Machine Learning Lifecyle



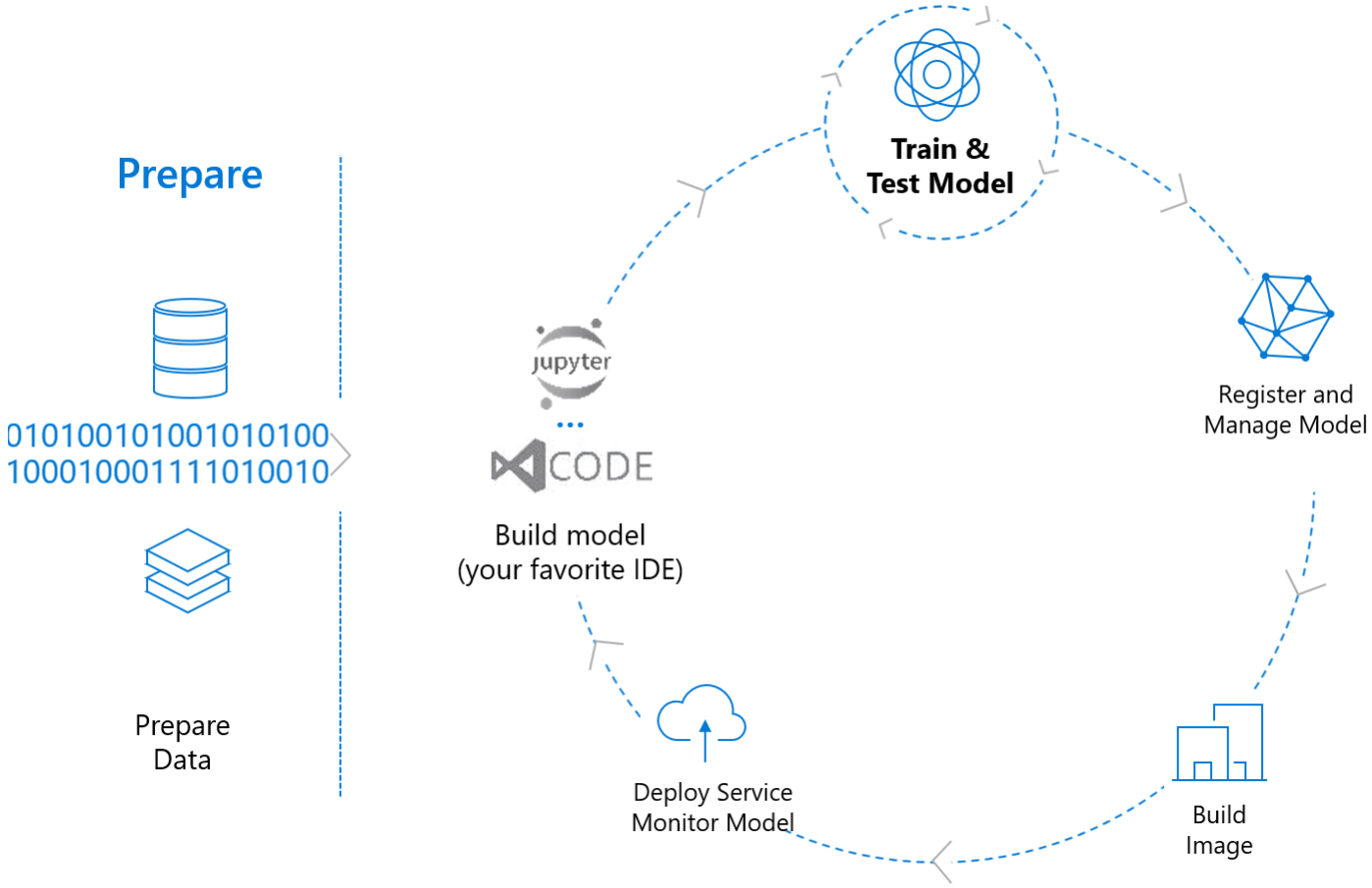
Data Science != Machine Learning

…despite what the pundits may tell you

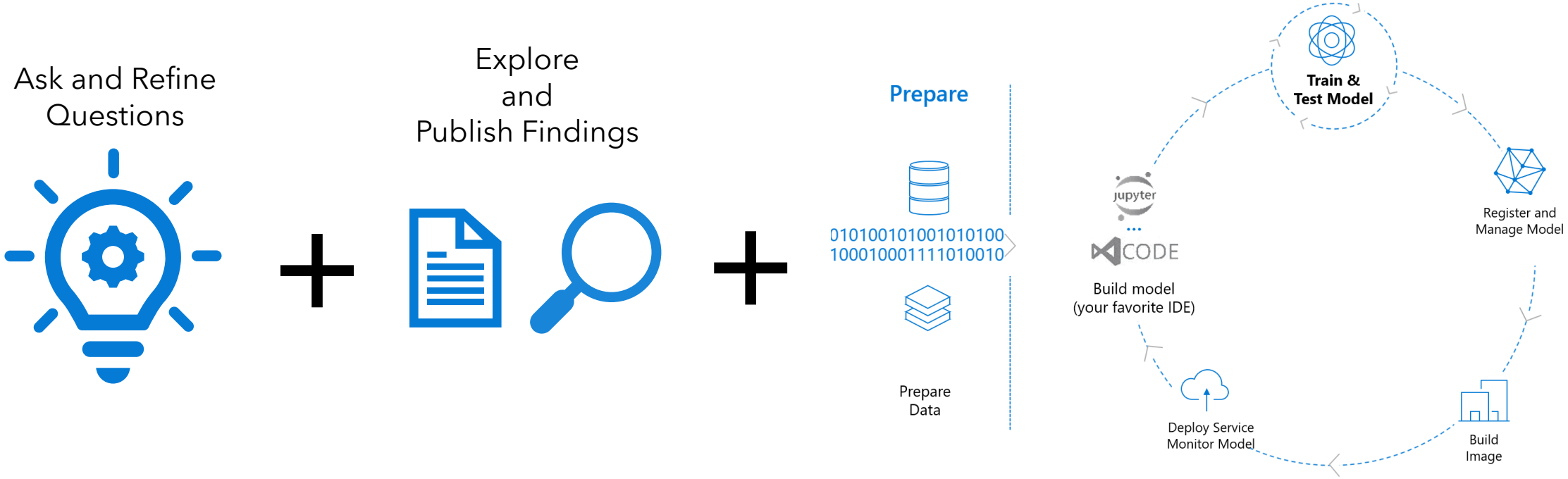
Data Science

Using data to ask and answer questions that are important to your business

Machine Learning Lifecyle



The Data Science Lifecycle



What do we need to track?

Questions, Answers, and Knowledge

Goals, objectives, and metrics

Code and environments

Models

Data

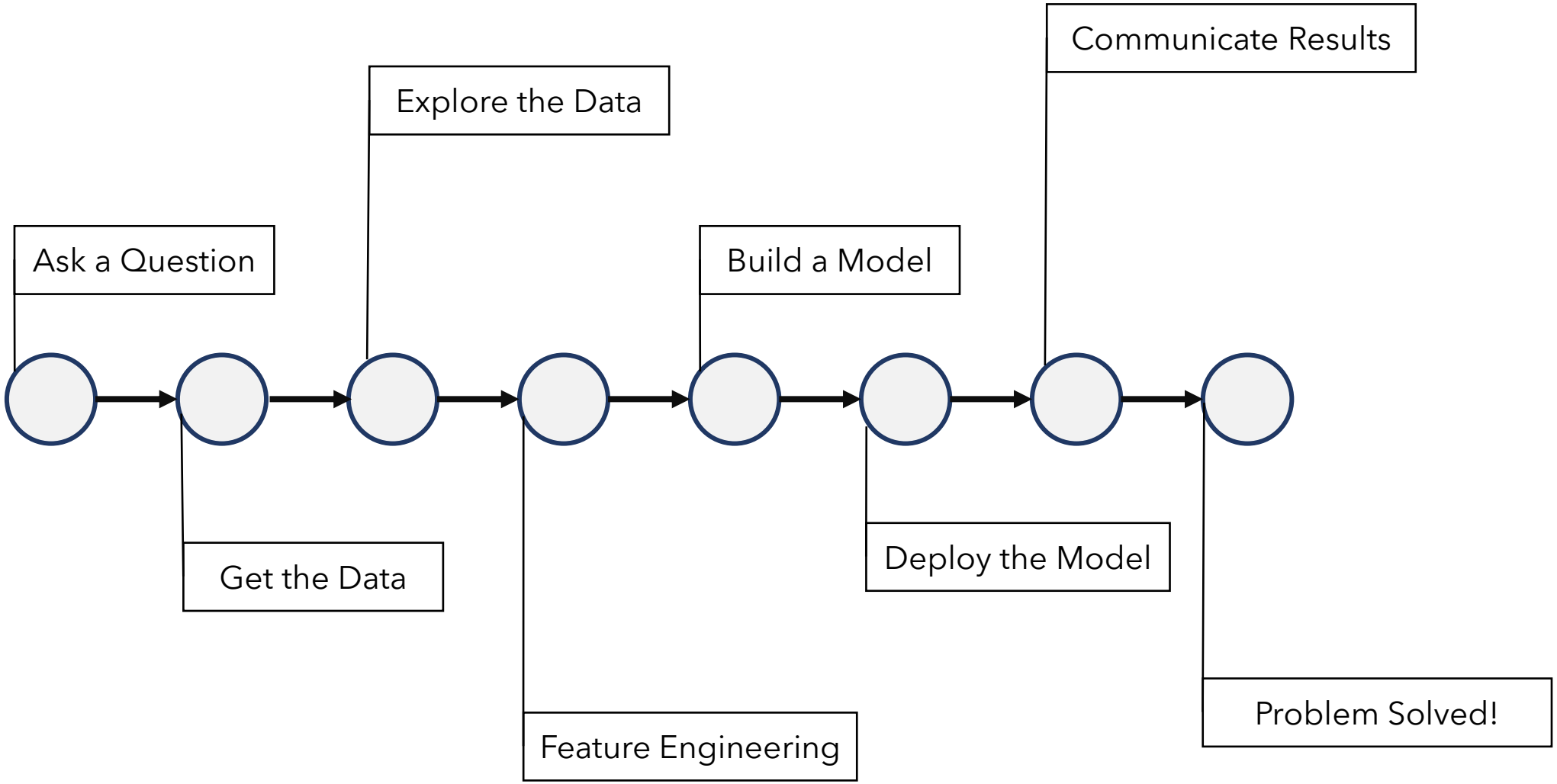
Results

GitHub and AzureML
Better Together

Ideas, Collaboration, and Code

Machine Learning Lifecyle and MLOps

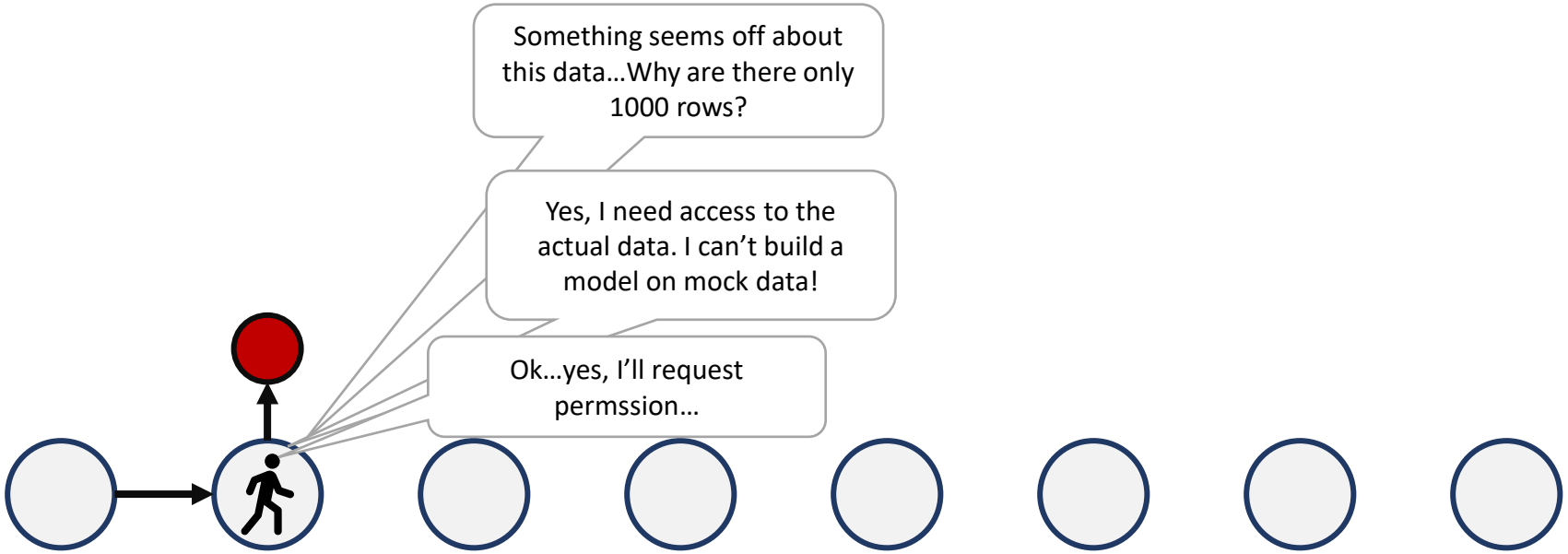
The Data Science Happy Path



Ask a Question

So far, so good…

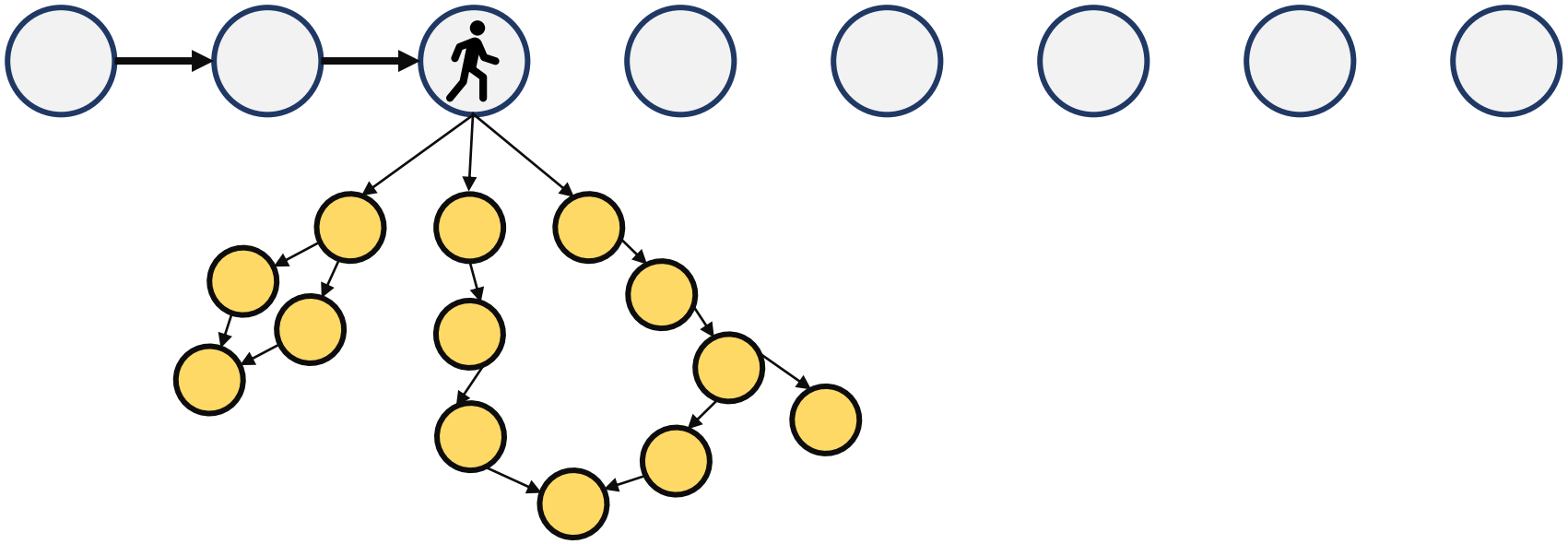
Get the Data


Something seems off about this data…Why are there only 1000 rows?
Yes, I need access to the actual data. I can’t build a model on mock data!
Ok…yes, I’ll request permssion…

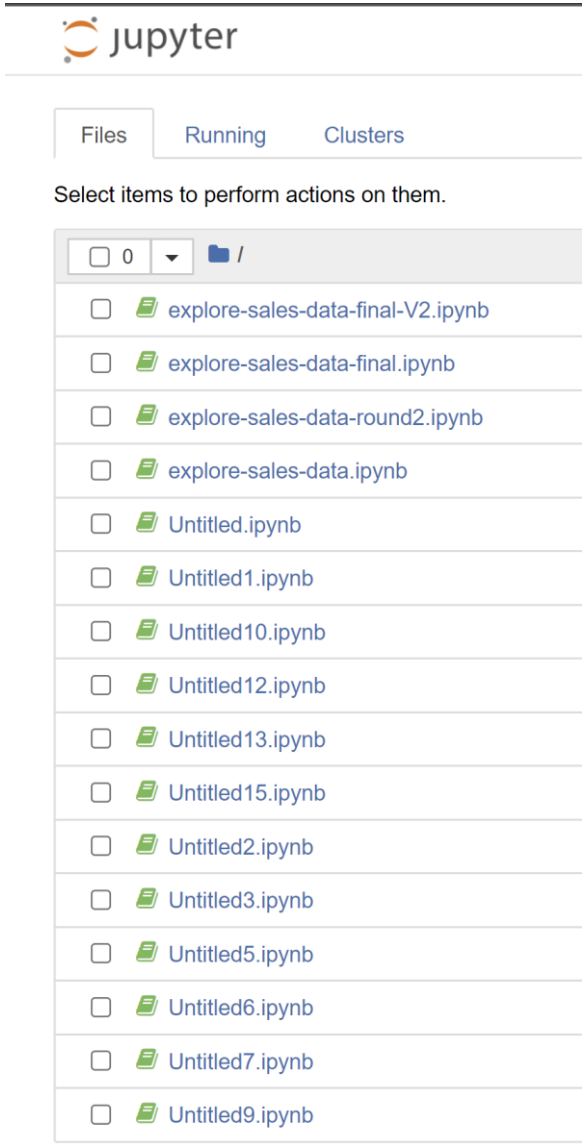
Several weeks later…

Get the Data

Explore the Data

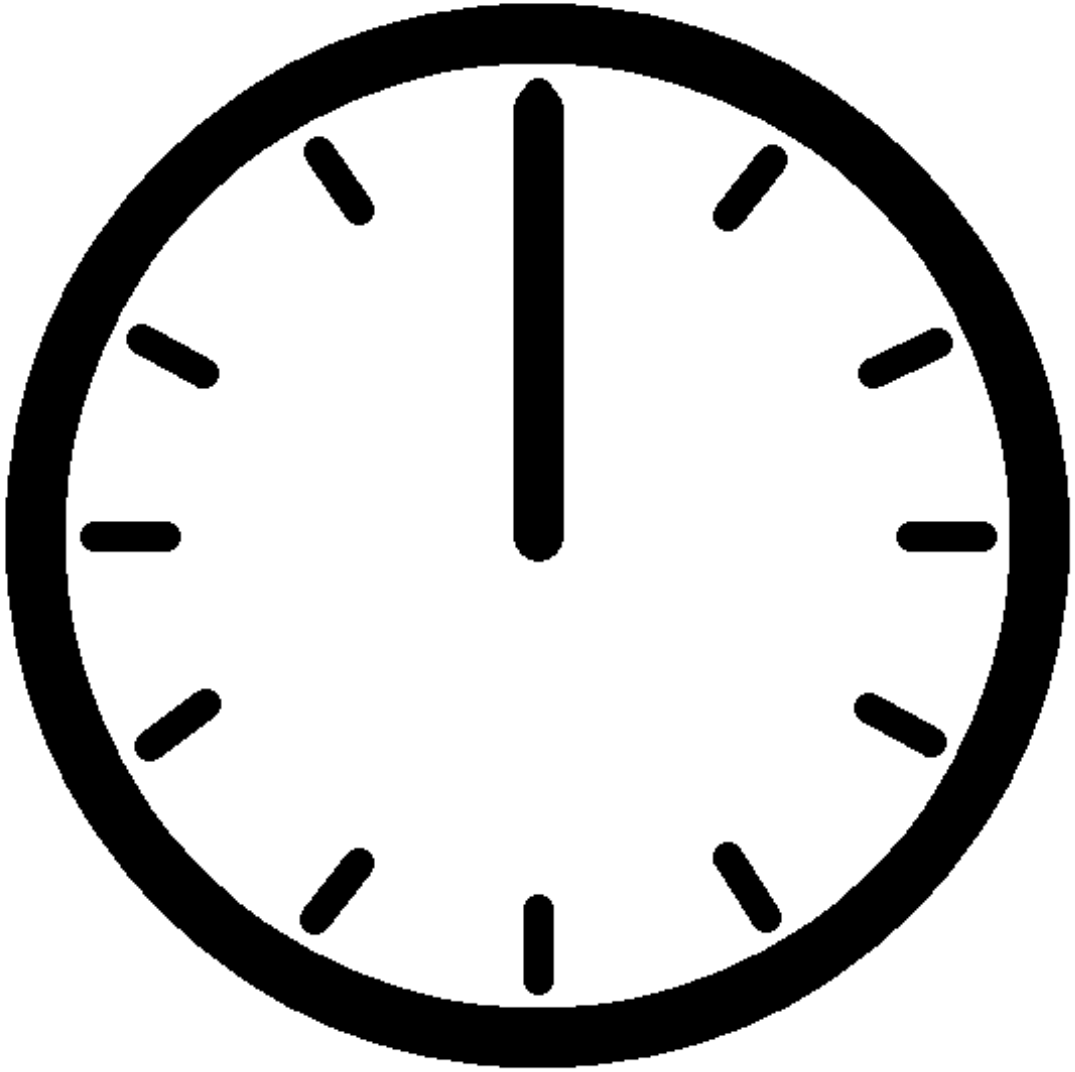


Does this look familiar?

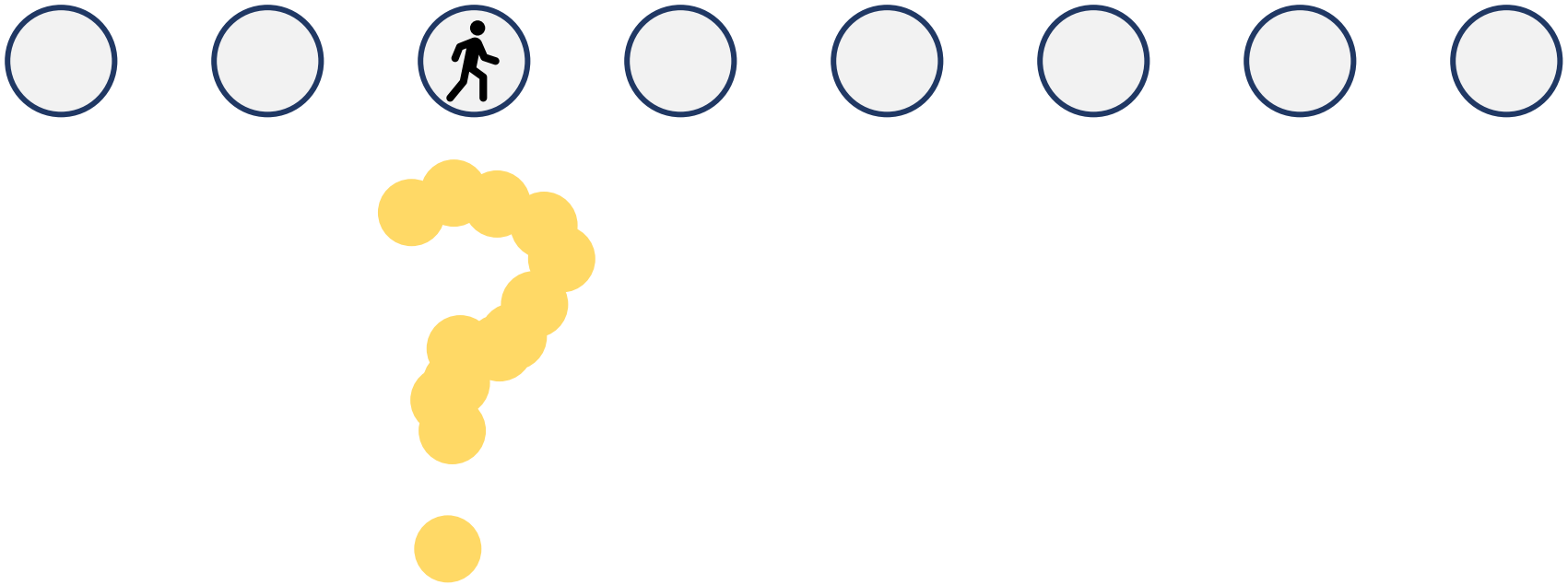


Explore the Data

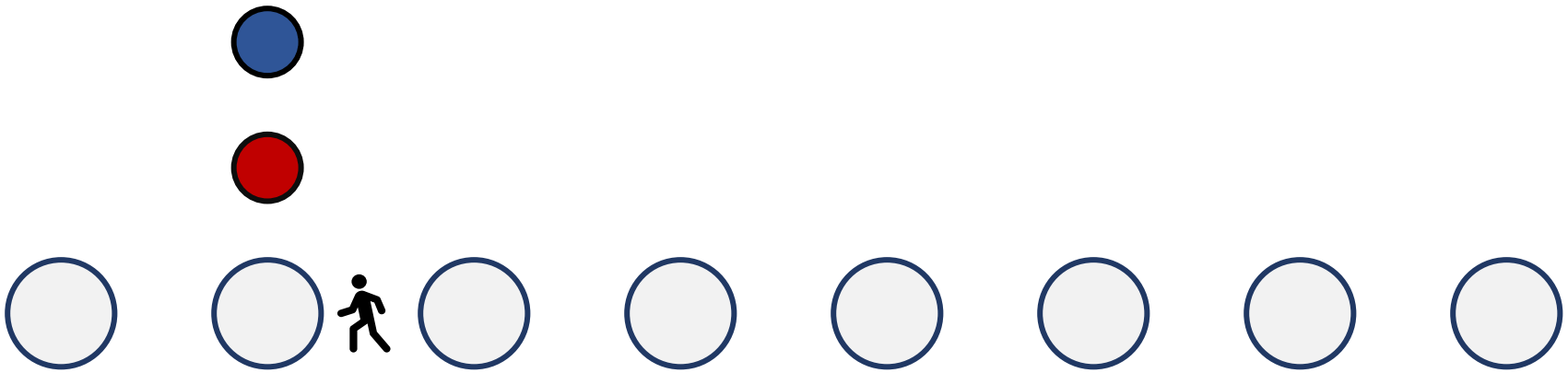
Some time passes….



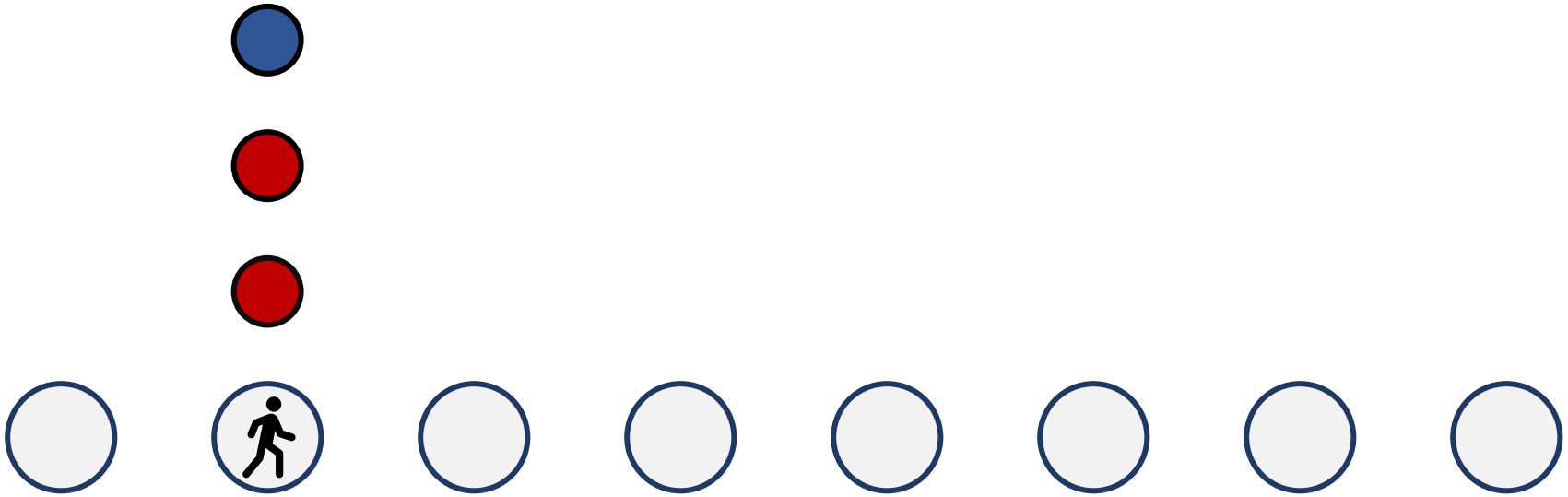
Explore the Data



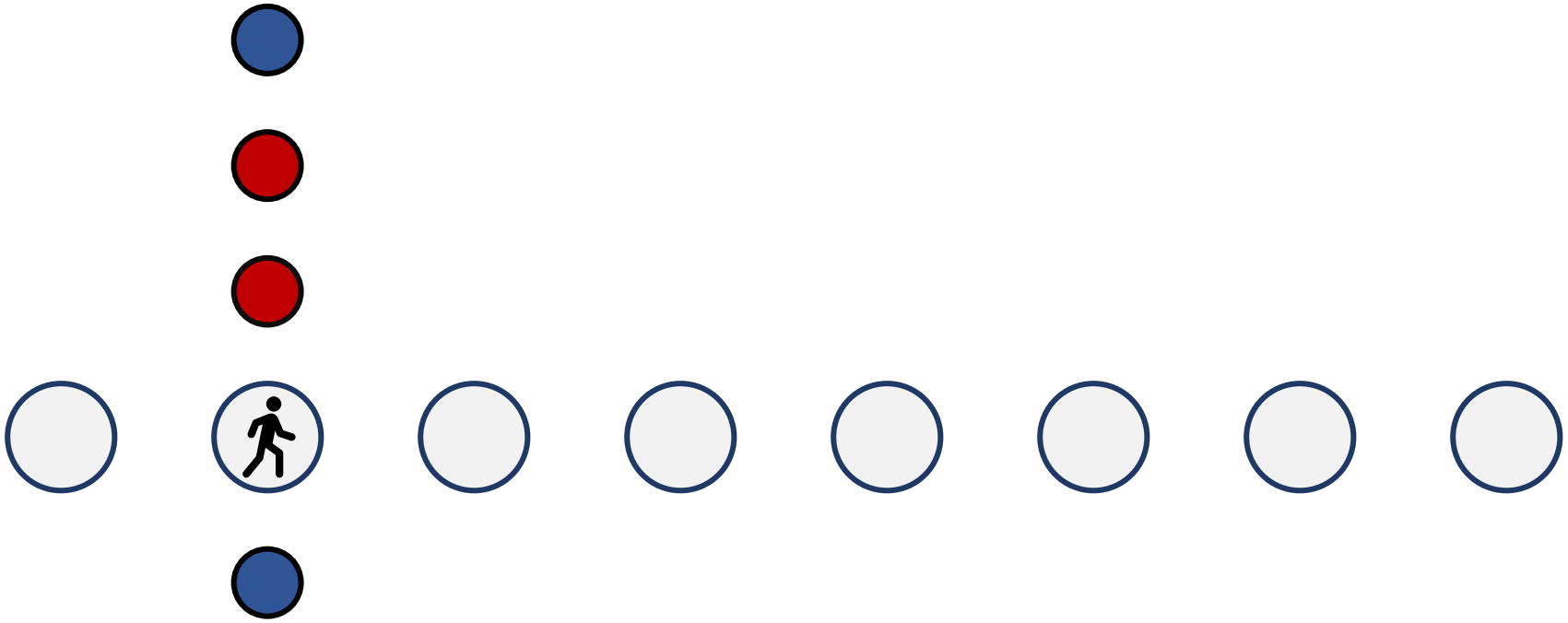
Fix the Data



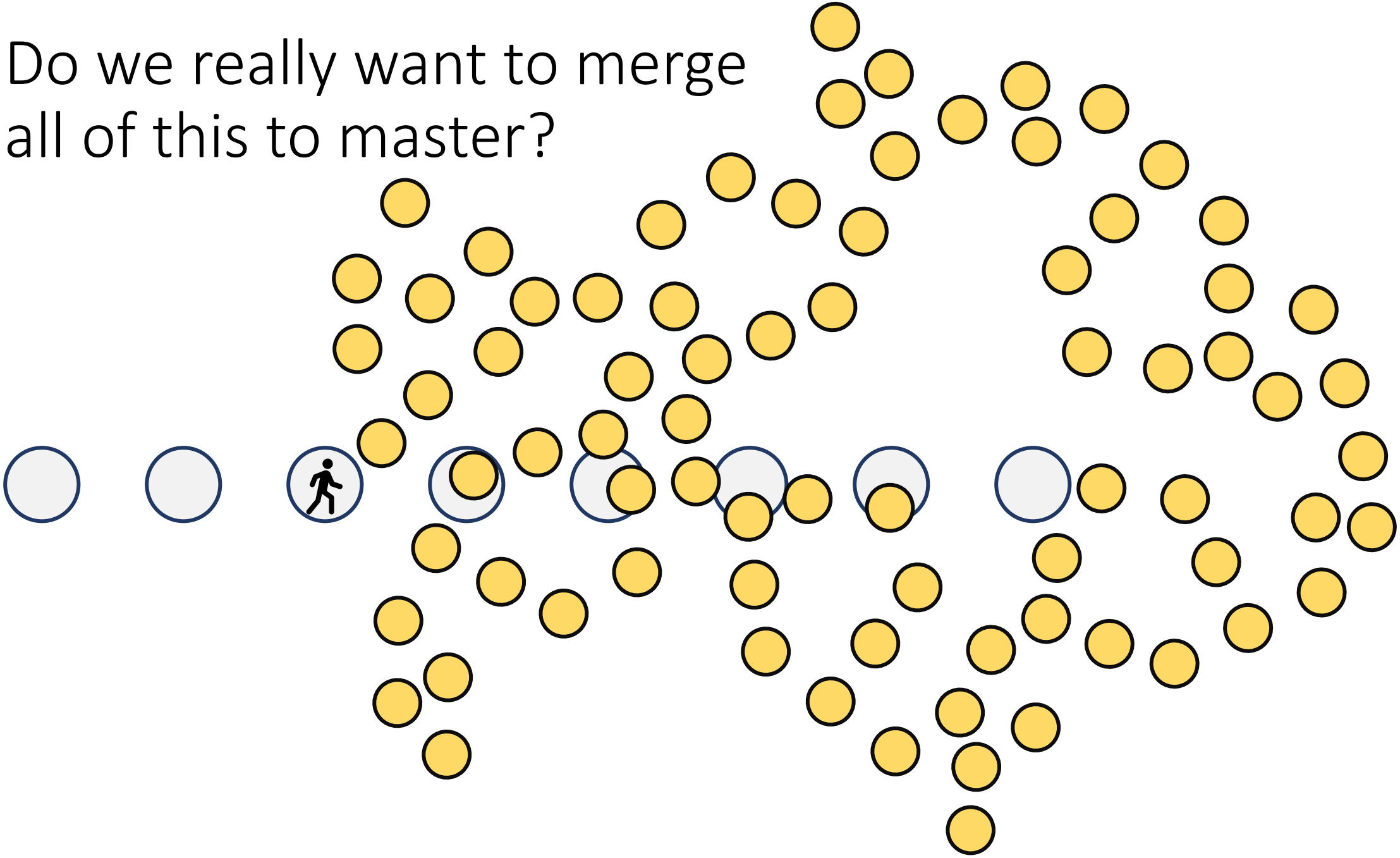
Fix the Data Again



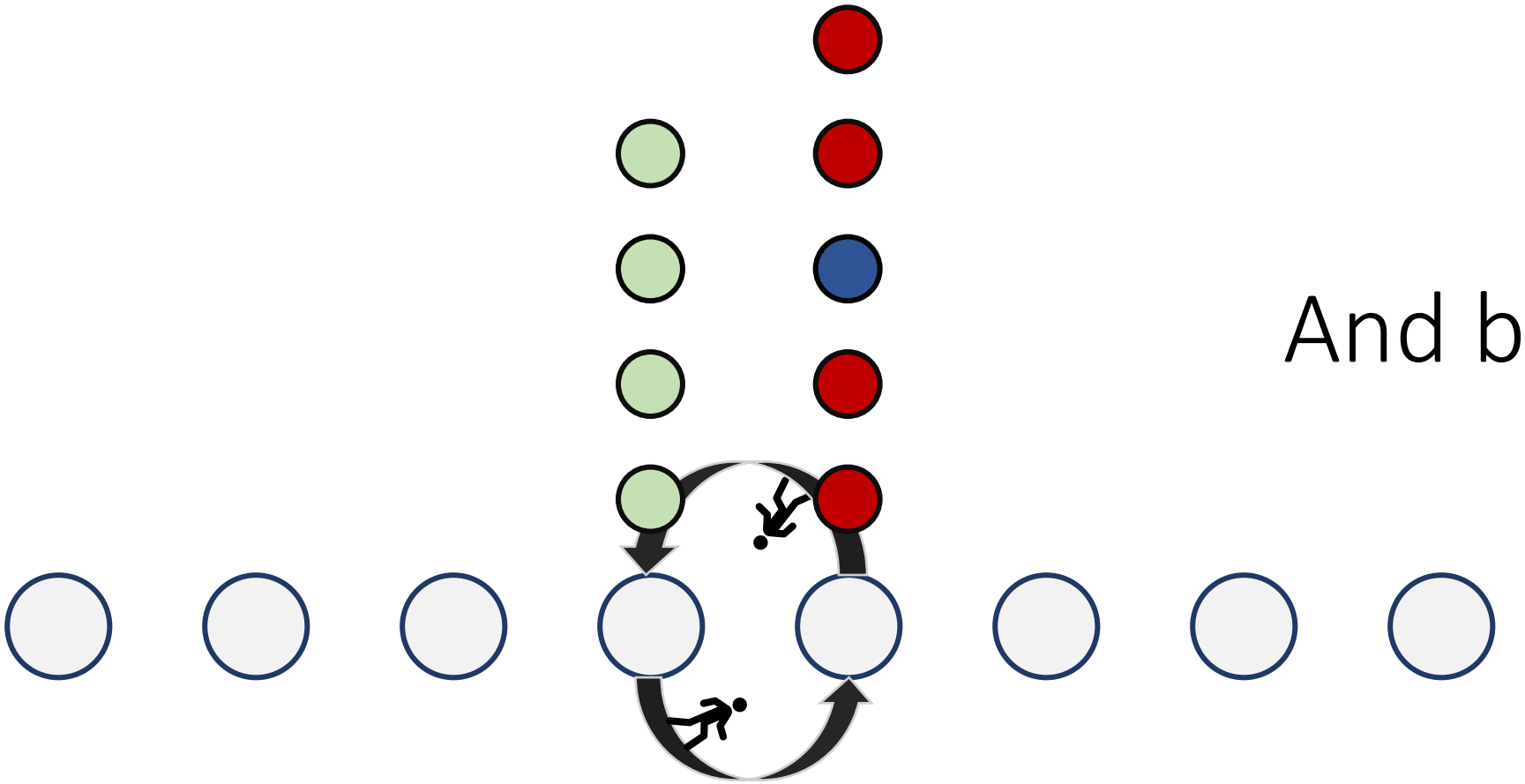
Get More Data



Do we really want to merge all of this to master?



Now it’s time to build features

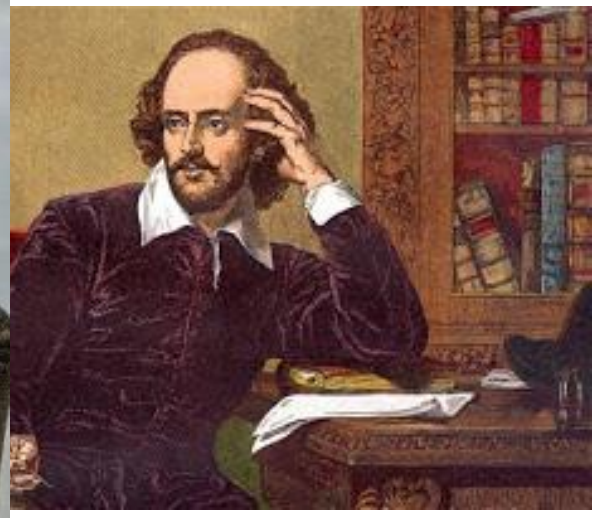


And build models

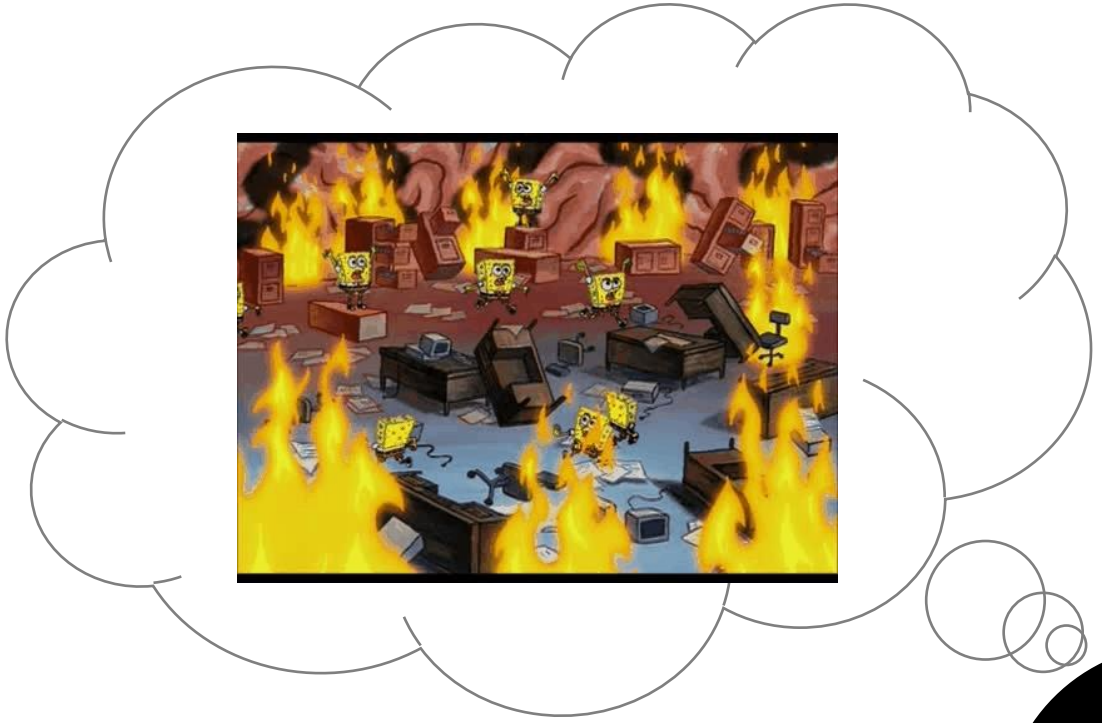
Deploy your model

Communicate Results

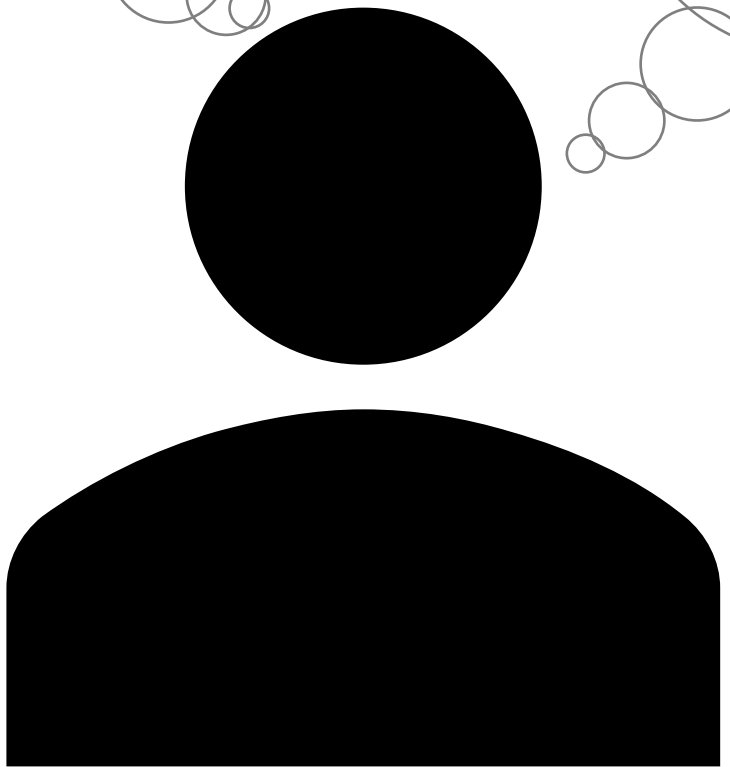
Do you remember
what you did and why?

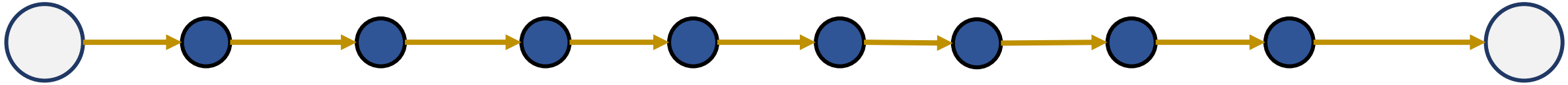


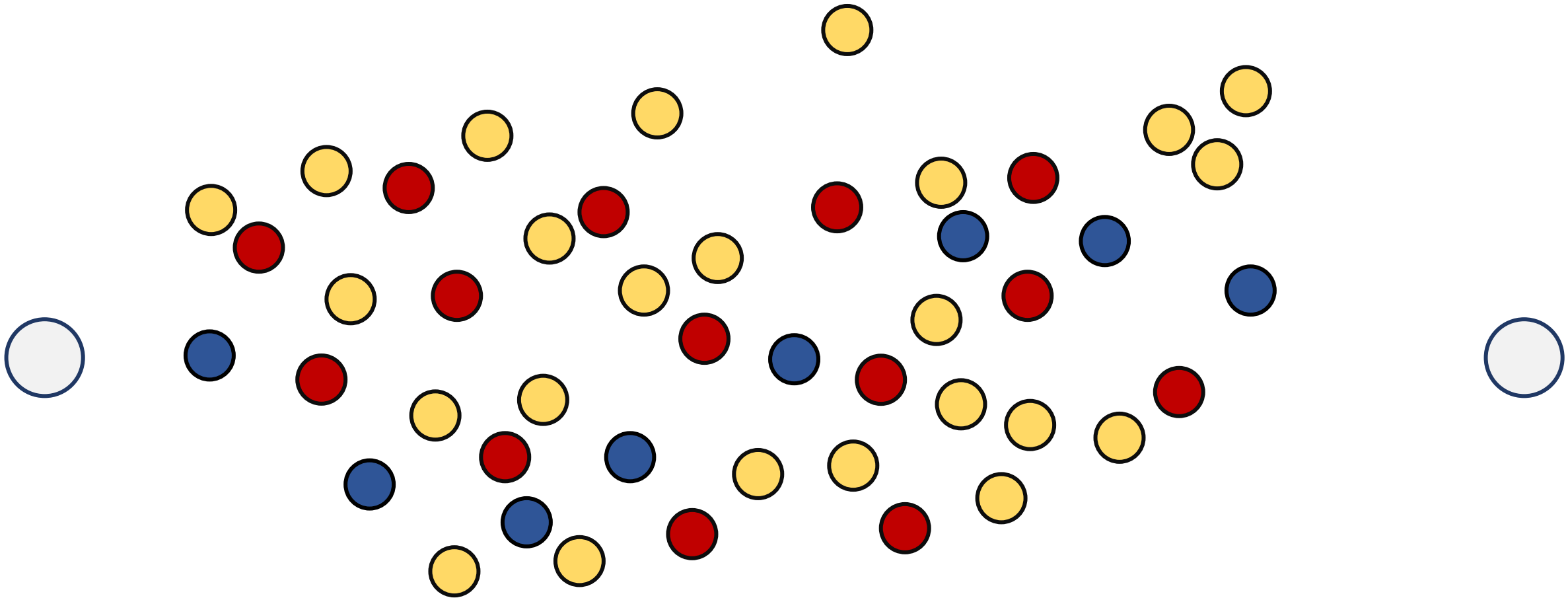
Problem Solved!

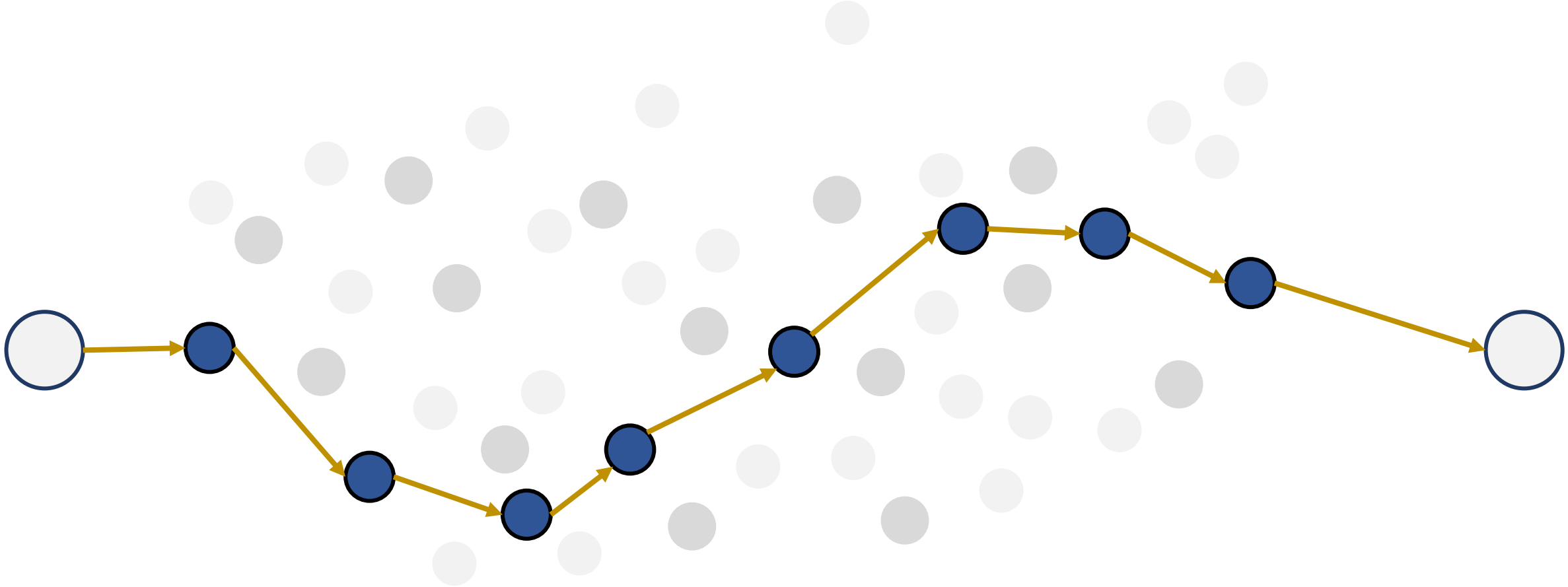


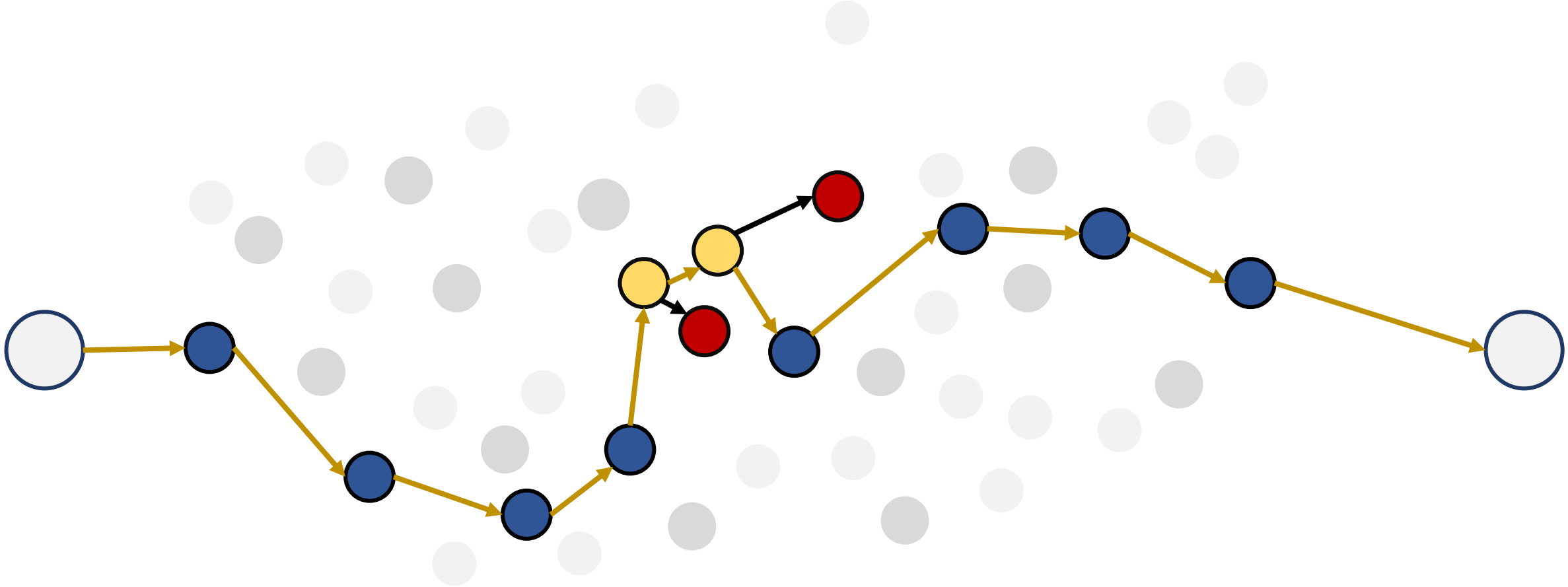

YOU SIT ON A
THRONE OF LIES.











GitHub and Issue-Driven Workflows

Always start with an Issue

A CONVERSATION ABOUT OUR WORK

BE EXPLICIT ABOUT THE WORK TO BE DONE

ASSIGN, LABEL, AND PRIORITIZE WORK

PROMOTES DISCOVERABILITY

Issue Types and Templates

Ask
Describe and Refine Problem Statement and Proposed Solution
Get started

Experiment
Describe experiments and approaches for solving current problem
Get started

Explore
Publish results of exploration to expand knowledge and understanding about the data to solve the problem.
Get started

- [ ] **Get Claims Data** Data
  #5 opened yesterday by charleswm
- [ ] **Build baseline claims model** Experiment
  #4 opened yesterday by charleswm
- [ ] **Create more accurate claim predictions** Ask
  #3 opened yesterday by charleswm

Issue: Ask

- Capture the problem statement
- Define qualitative goals
- Define metrics
- Define success criteria
- Identify required data sources
- Design high-level solution architecture
- Link to related experiments and exploration

charleswm commented now

Problem Statement

Describe the problem you are trying to solve.

Desired Outcome

Describe what outcome will be enabled if successful? What about the existing process will change as a result of your success? How will it change?

Current State

Describe the current state and it's shortcomings. Why does this need to change?

Success Criteria

Impact

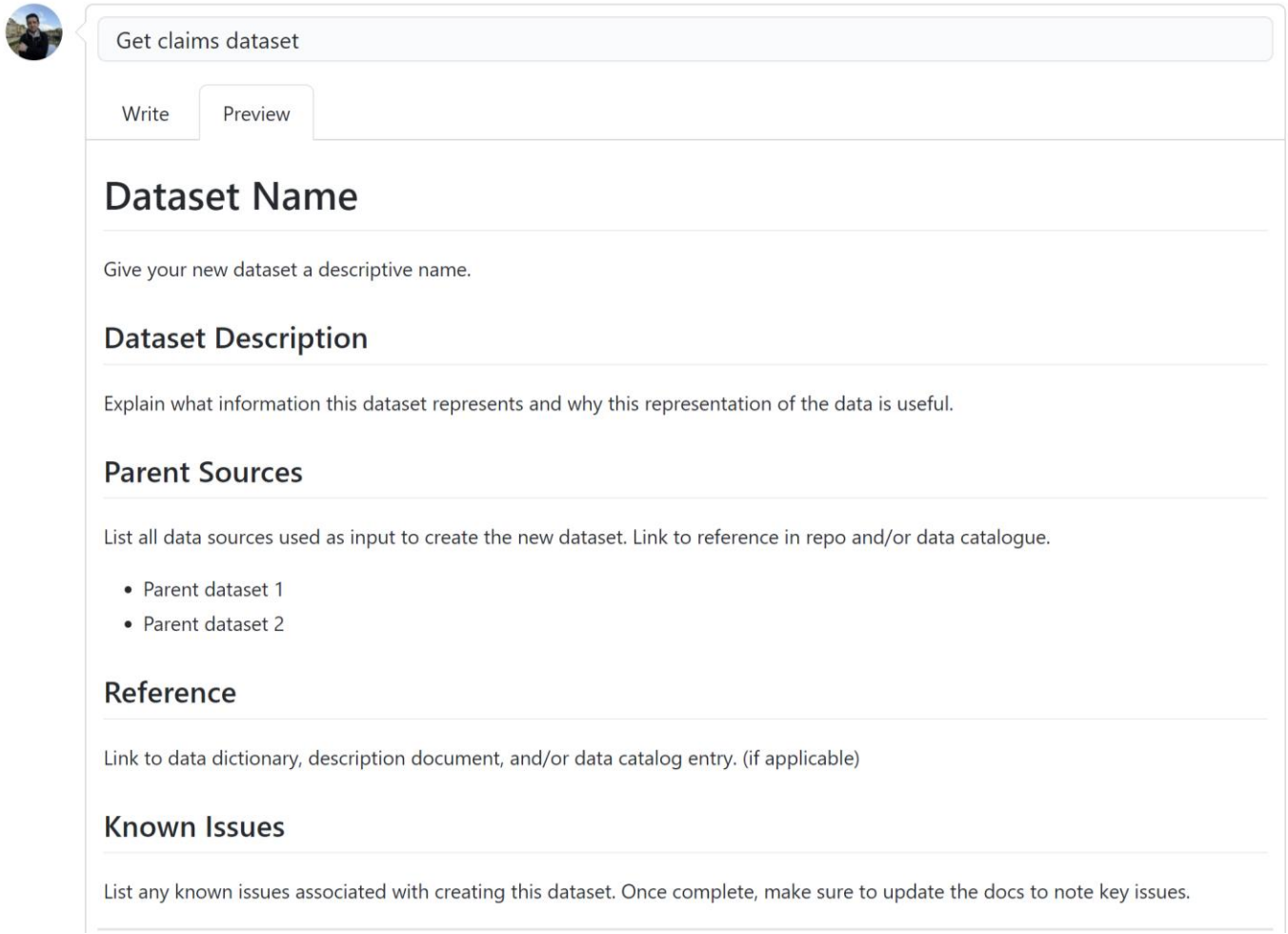
- To the extent possible, estimate the impact of project if successful.
- How much of an improvement do you need for this to be valuable?

Metrics

- Describe the metrics you will use to measure success.
- Be specific in how a metric is defined for the purpose of this problem.
- This may contain a combination of business metrics and technical metrics
- This should also include metrics that we wish to balance against our primary metric.

Issue: Data

- Make data available in environment
- Capture schema (optional)
- Data profile (optional)
- Register data in ml workspace (optional)
- Set up simple testing suite for data and functional tests using pytest and github actions



Get claims dataset

Write | Preview

Dataset Name

Give your new dataset a descriptive name.

Dataset Description

Explain what information this dataset represents and why this representation of the data is useful.

Parent Sources

List all data sources used as input to create the new dataset. Link to reference in repo and/or data catalogue.

- Parent dataset 1
- Parent dataset 2

Reference

Link to data dictionary, description document, and/or data catalog entry. (if applicable)

Known Issues

List any known issues associated with creating this dataset. Once complete, make sure to update the docs to note key issues.

Issue: Explore

- Explore using notebooks
- Summarize findings
- Link to ask issue
- PR and closing pattern

charleswm commented now

Overview

Description of the scope of your exploration including relevant datasets and problem of interest.

Goals

- [ ] Question you plan to answer
- [ ] Question you plan to answer

What was discovered during this exploration?

Describe the overall content of your exploration effort.

For each key insight, briefly describe what you uncovered.

Insight 1

Brief description of insight. Link to artifact.

Insight 2

Brief Description of insight. Link to artifact.

Next Steps

What are the next actions you will take as a result of what you discovered?

Issue: Experiment

- Link to ask issue
- Test different approaches
- Write and run modeling code
- Log experiments using AzureML
- Get peer feedback

Describe experiments and approaches for solving current problem. If this doesn't look right, choose a different type.

Try forecasting model

Write | Preview

Description

Ask Reference: #link-to-ask-issue

What is the goal of this experiment? How are you attempting to solve the problem.

Metrics

What metric are you trying to model? Define this metric if it is not a standard metric.

Assumptions

Describe any assumptions you are making in building this model. Describe what problems may arise if the assumptions are inaccurate.

Results

Experiment History: [Link to experiment tracking source if applicable]
Related PR: [Link to pull request where experiment was tried]

Summarize the results of this experiment (tldr).

diabetes-experiment

Edit table | Refresh | Reset to default view | Include child runs

Run status

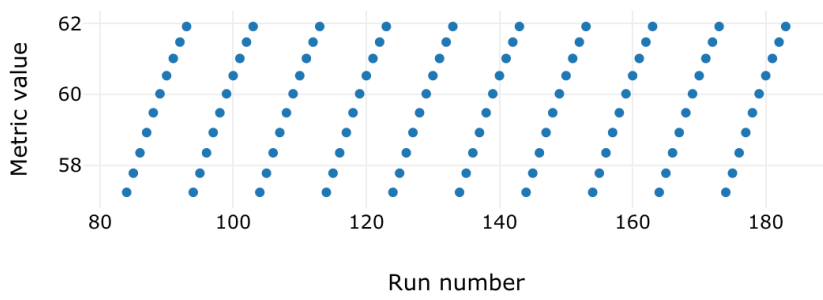
0 Running

183 Completed

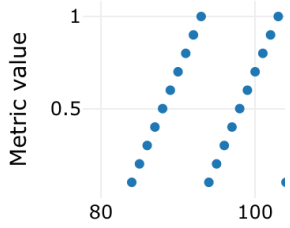
0 Failed

0 Other

rmse



alpha_value

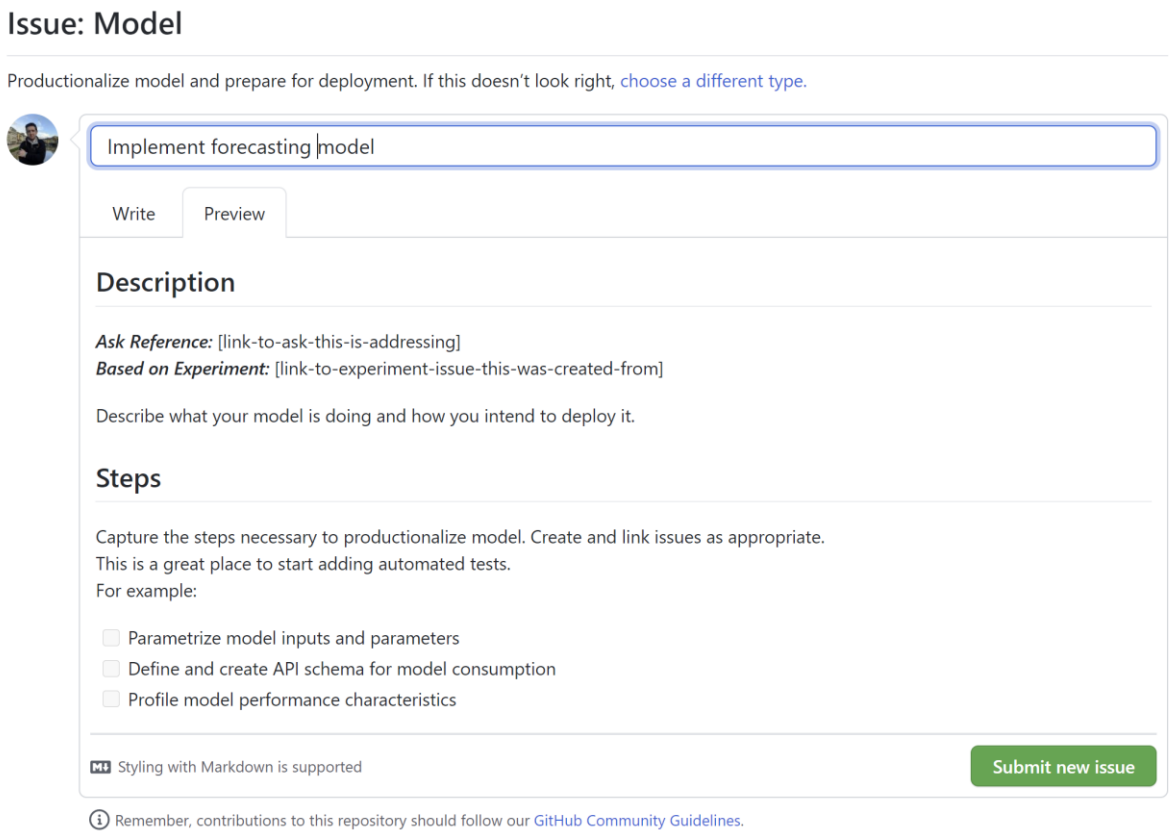


Add filter

Run	Created time	Duration	Status	Compute target	Run type
Run 183	10/29/2019, 9:33:03 AM	3s	Completed	sdk	--
Run 182	10/29/2019, 9:33:00 AM	2s	Completed	sdk	--
Run 181	10/29/2019, 9:32:57 AM	2s	Completed	sdk	--
Run 180	10/29/2019, 9:32:54 AM	2s	Completed	sdk	--
Run 179	10/29/2019, 9:32:51 AM	2s	Completed	sdk	--
Run 178	10/29/2019, 9:32:49 AM	2s	Completed	sdk	--

Issue: Model

- Link to ask issue
- Go from experiment to production
- Train and validate model
- Log runs using AzureML
- Register model in AzureML
- Add logging and tests



Labels

Ask	Define and scope problem and solution
Communicate	Write reports and create dashboards
Data	Get, transform, and validate data
Deploy	Register, package, and deploy model
Experiment	Build features and train models
Explore	Explore and document data to increase understanding

Branching

https://github.com/dslp/dslp/blob/main/branching/branch-types.md

DATA/

EXPERIMENT/

EXPLORE/

MODEL/

Branching

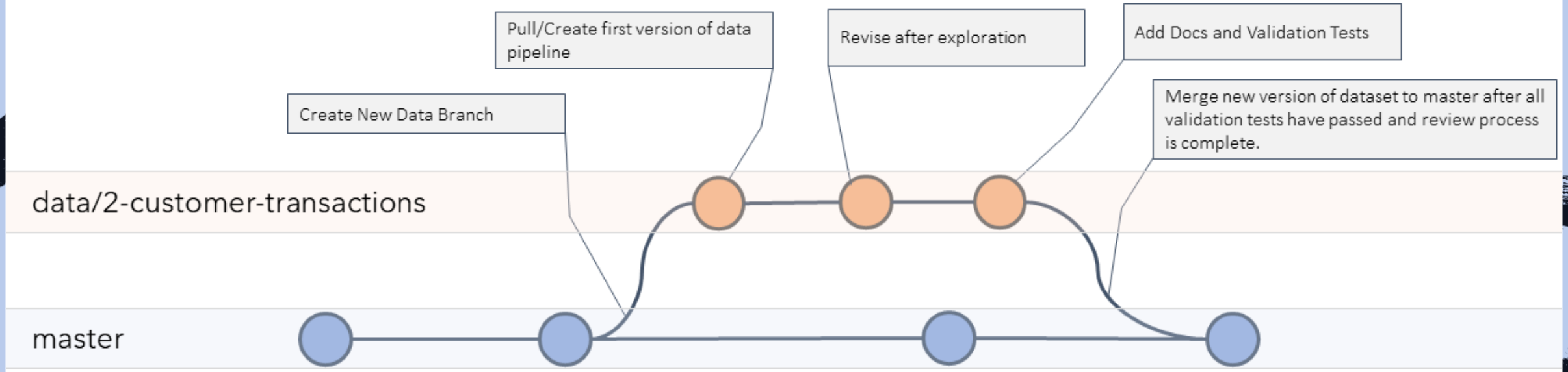
DATA/

- Like a feature branch
- Push code that ingests or creates datasets

Branching

DATA/

Data Branching Pattern



Branching

EXPLORE/

- What does it mean for exploration to be “in production”?
- How do you “test” exploration?
- How do you deprecate old exploration efforts?
- If you don’t deprecate, how do you deal with sprawl?

Branching

EXPLORE/

- Create branches for exploration attempts
- Open a pull request and link to the issue
- Close pull request when finished
- Do not merge to master

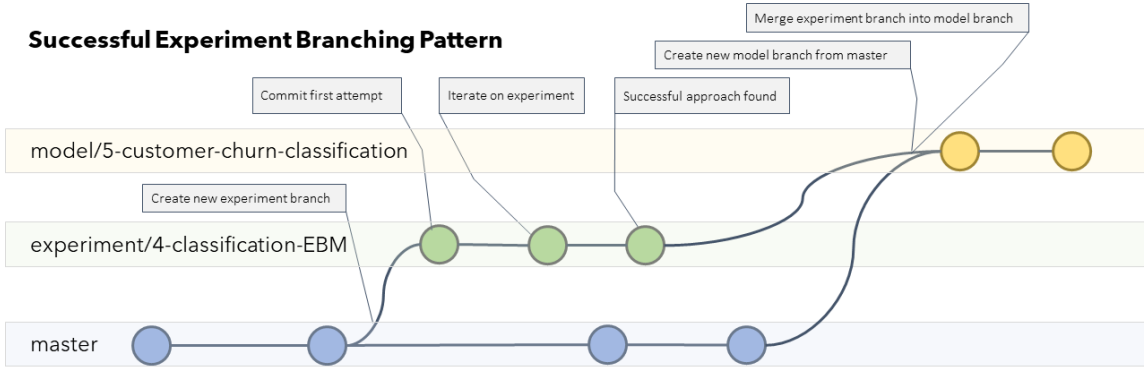
Branching

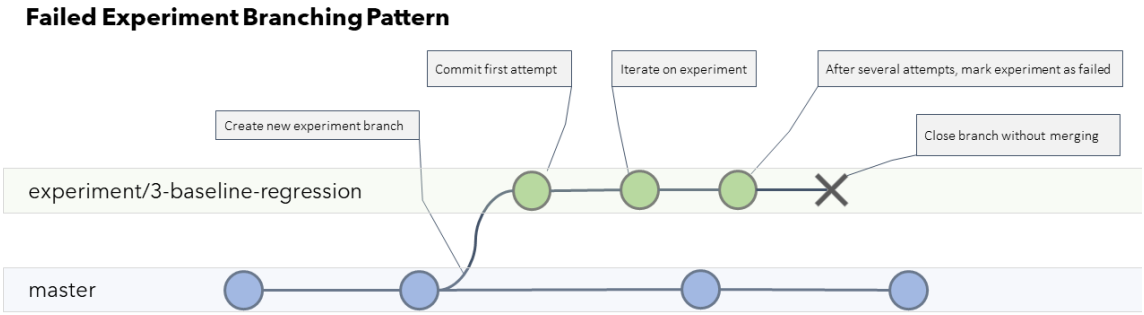
EXPERIMENT/

- Create a separate branch for each experiment attempt
- Test different algorithms, featurization, hyperparams, etc.
- Only successful experiments are merged to master
- Link pull request to ask issue to track all related experiments

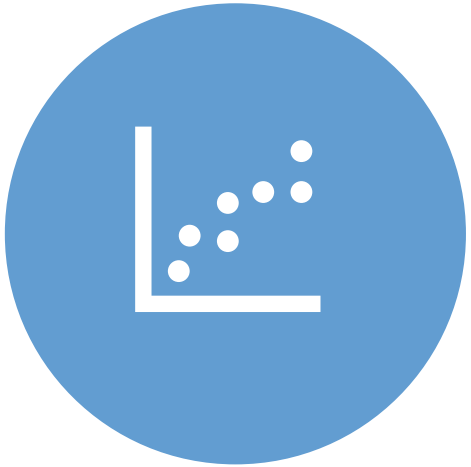
Branching

EXPERIMENT/





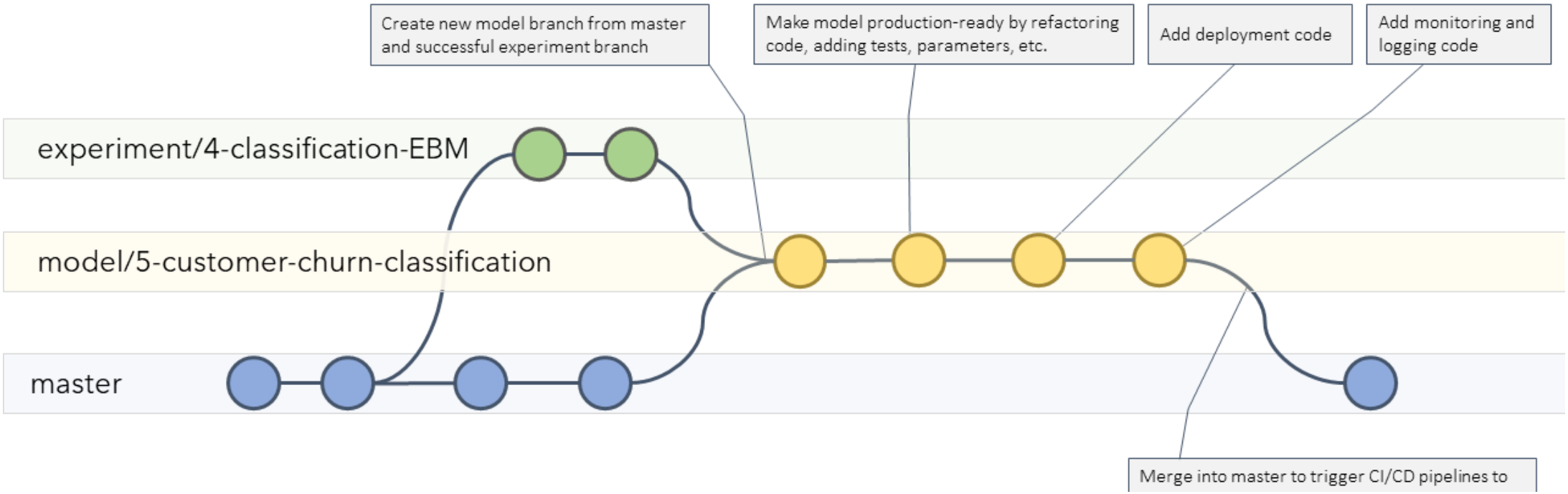
Branching



MODEL/

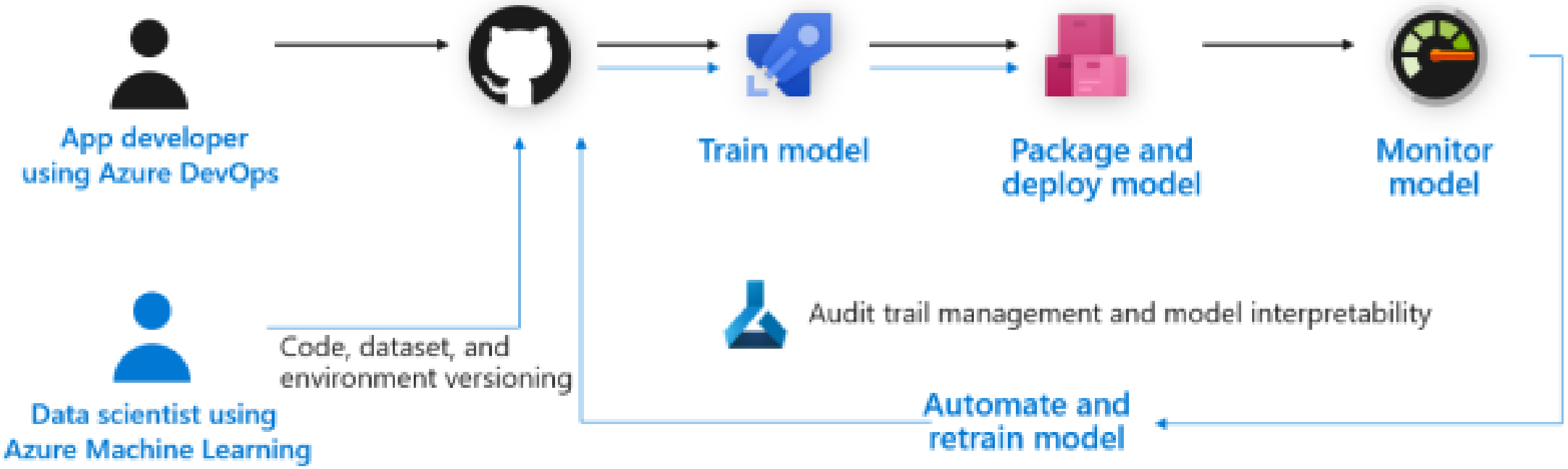
- When an experiment is successful and you want to spend time to productionalize and deploy it, you open a model branch
- Refactor and add logging, tests, etc
- Use CI/CD to automate tests and deployment

Experiment to Model Branching Pattern



Branching

MODEL/



Leverage MLOps once you've built a model