MMCI MACHINE LEARNING Lab

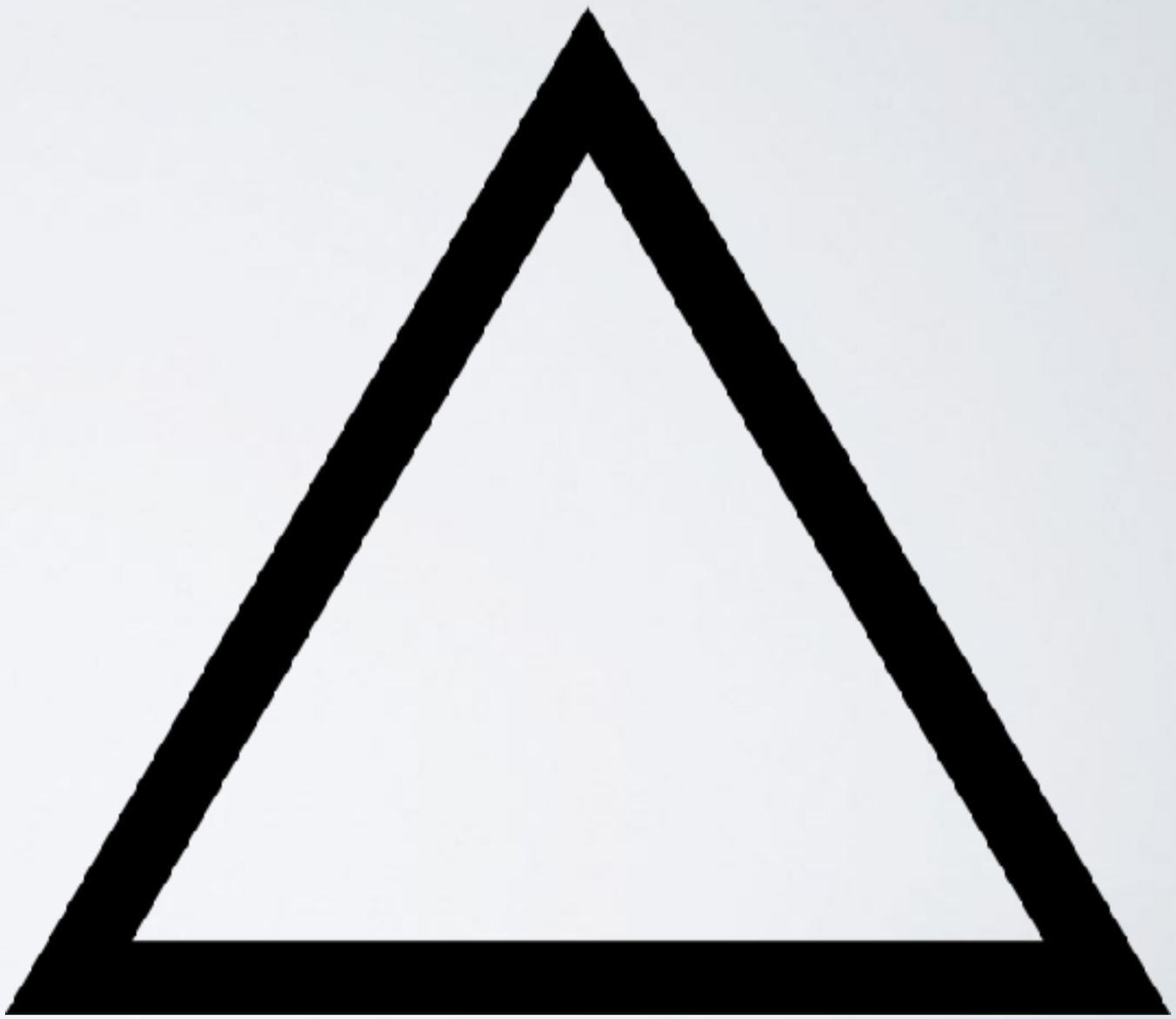
MACHINE LEARNING LAB

- Founded by economist Anthony Goldbloom (2010)
- Platform to host predictive modeling competitions for companies (e.g. NASA)
- Sold to google
- kaggle
- Note for several projects:
  - Data (size)
  - Kernels
  - Evaluation

kaggle

MACHINE LEARNING LAB

- 2nd Annual MMCi Predictive Modeling Competition
- Analytic:
  - Cohort(s)
  - Endpoint(s)
  - Predictor(s)
- Project:
  - Data Storage
  - Data Compute
  - Results



DATA DELIVERY (IDEAL)

- PROBLEM: Create algorithm that allows us to predict the number of repositories for an MMCi cohort.

- List the "points of the triangle" (cohort, endpoint, predictors).

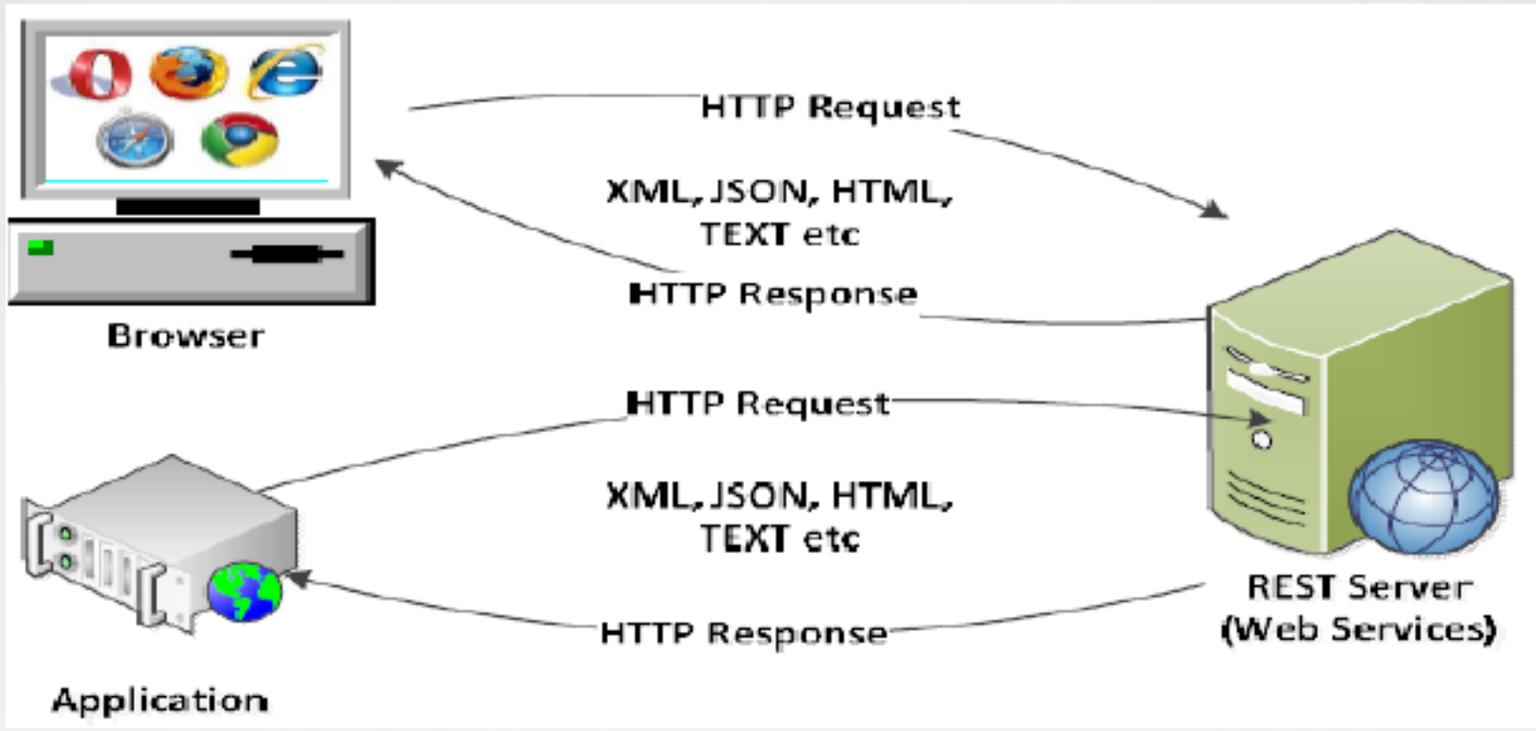
- GitHub API:

  - Cohort: https://api.github.com/repos/Duke-Translational-Bioinformatics/mmci-practical-datascience/forks

  - Endpoint: https://api.github.com/users/benneely

  - Predictors: https://api.github.com/users/benneely

WEB SERVICE IN ACTION



DATA DELIVERY (LAB)

- Problem: Create an algorithm that allows Duke Endocrinology department to predict readmissions.

- Cohort: <fill in>

- Endpoint: <fill in>

- Predictors: <fill in>

STORAGE/COMPUTE/RESULTS

- Microsoft Azure Machine Learning Studio (google)

- Set a bookmark here.

- Interactively Explore Interface.

DATA STORAGE

- Data Sets
- + New
- From Local File
- Choose File from disk
- Add description
- OK

DATA EXPLORATION

- Open in Notebook
- Adult Census Income:
    - Problem: Create Algorithm that predicts adult income using the 2010 Census Data.
- Questions:
    - “Triangle”: Cohort / Endpoints / Predictors?
    - Missingness?
    - Data Representation?
    - Clinical Relevance?
    - What Kind of models can I use? What makes sense?

DATA EXPLORATION

- Open in Notebook
- Problem: Create an algorithm that allows Duke Endocrinology department to predict readmissions.
    - Questions:
    - "Triangle": Cohort / Endpoints / Predictors?
    - Missingness?
    - Data Representation?
    - Clinical Relevance?
    - What is the sample size?
    - What is the event rate?

MODELING PIPELINE (DEMO)

sklearn.pipeline.Pipeline

class sklearn.pipeline. **Pipeline** (*steps*, *memory=None*) [source]

Pipeline of transforms with a final estimator.

Sequentially apply a list of transforms and a final estimator. Intermediate steps of the pipeline must be 'transforms', that is, they must implement fit and transform methods. The final estimator only needs to implement fit. The transformers in the pipeline can be cached using `memory` argument.

The purpose of the pipeline is to assemble several steps that can be cross-validated together while setting different parameters. For this, it enables setting parameters of the various steps using their names and the parameter name separated by a '__', as in the example below. A step's estimator may be replaced entirely by setting the parameter with its name to another estimator, or a transformer removed by setting to None.

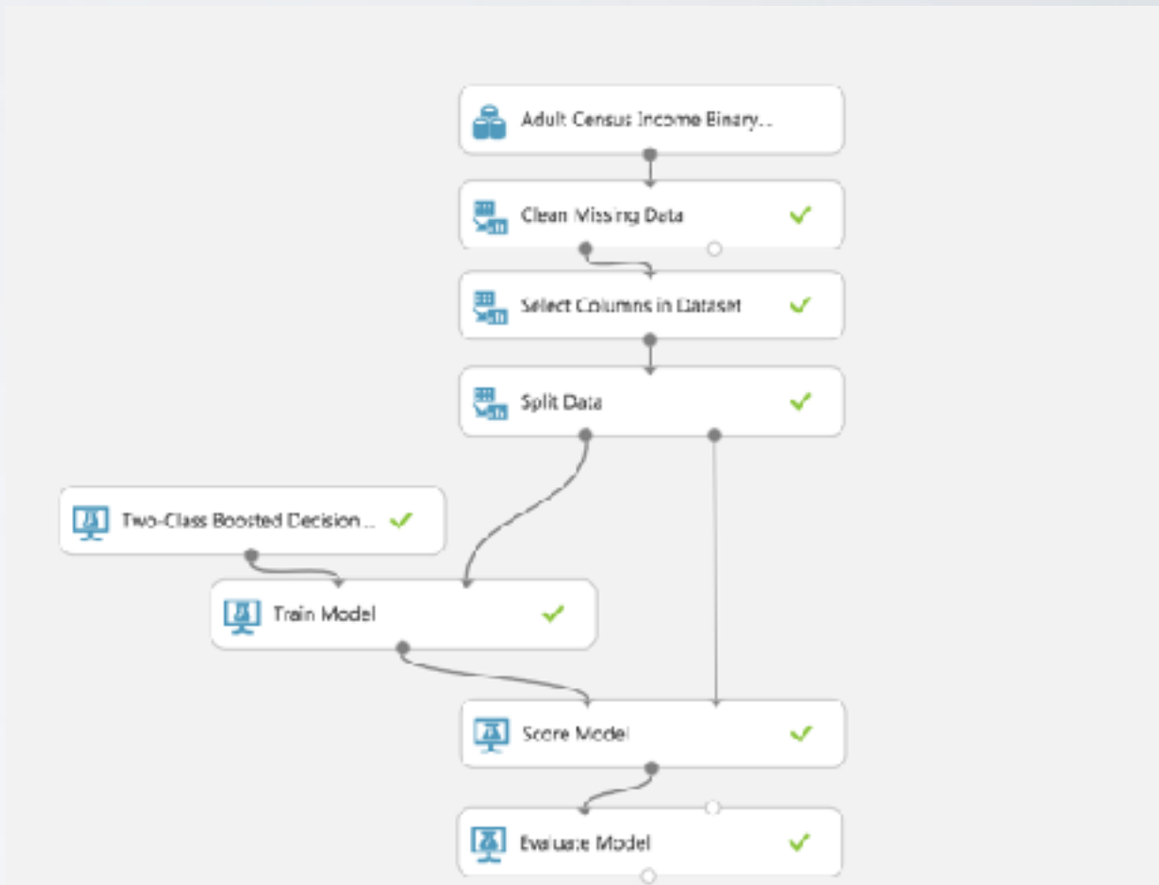
Read more in the User Guide.

Parameters: **steps** : list

List of (name, transform) tuples (implementing fit/transform) that are chained, in the order in which they are chained, with the last object an estimator.

memory : None, str or object with the joblib.Memory interface, optional

Used to cache the fitted transformers of the pipeline. By default, no caching is performed. If a string is given, it is the path to the caching directory. Enabling caching triggers a



EXPERIMENTS

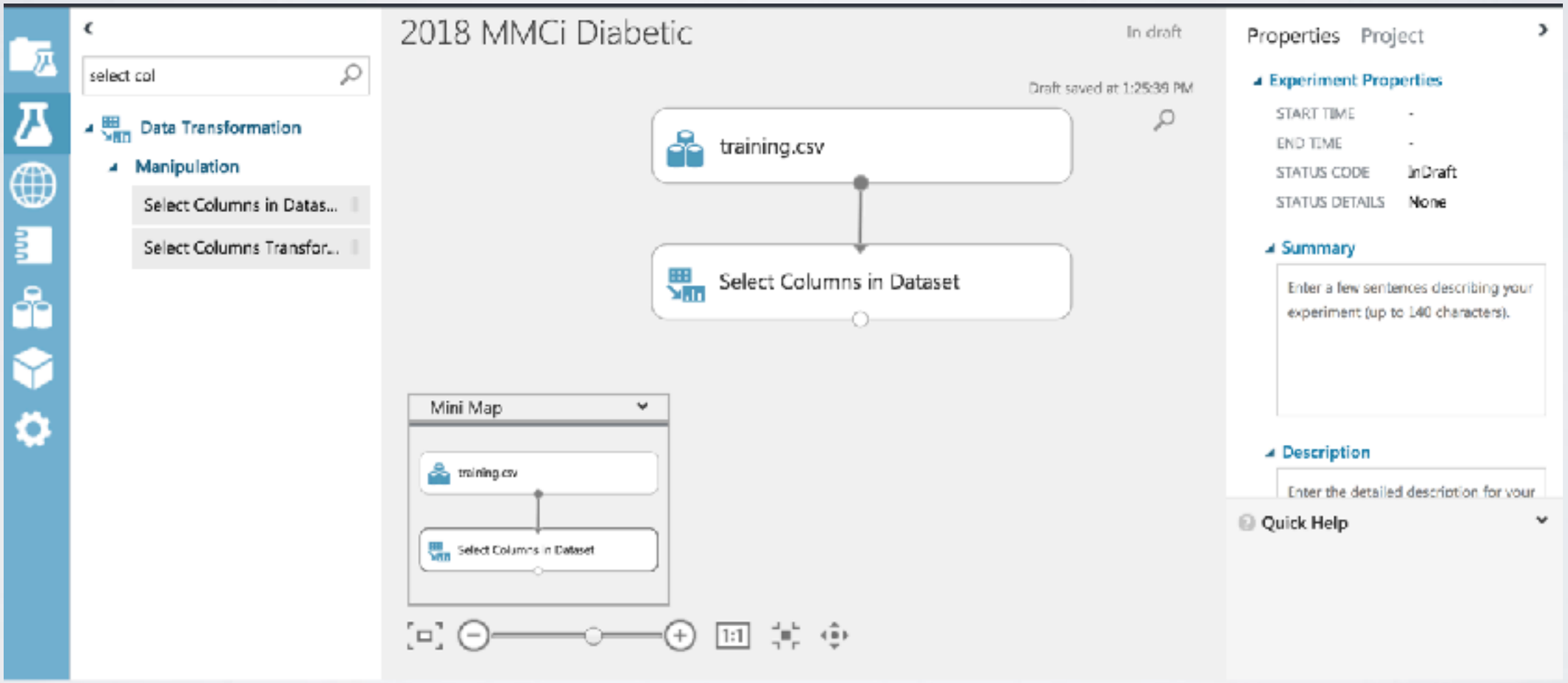
- + New

- Blank Experiment

EXPERIMENTS

EXPERIMENTS

Define Graph: (1) Load Data (2) Pare Down Columns **Node**

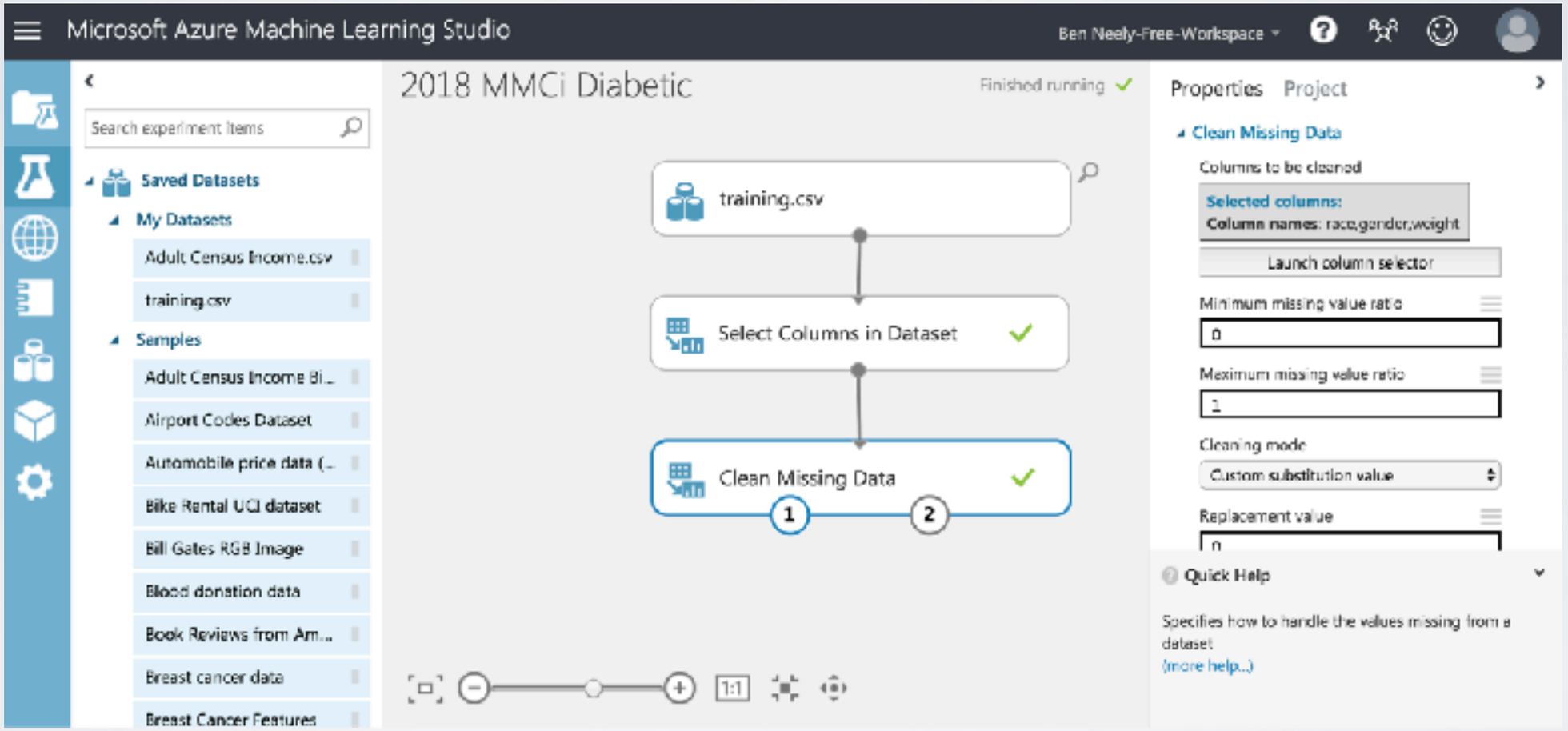


DATA EXPLORATION (TAKE 2)

- Open your experiment
- Remember the Problem: Create an algorithm that allows Duke Endocrinology department to predict readmissions.
- Use Azure Machine Learning Studio to complete:
  - Write down variables you choose for your models
  - Make note of:
    - Missingness
    - Data Representation
    - Don't forget you need to think about the "triangle"!

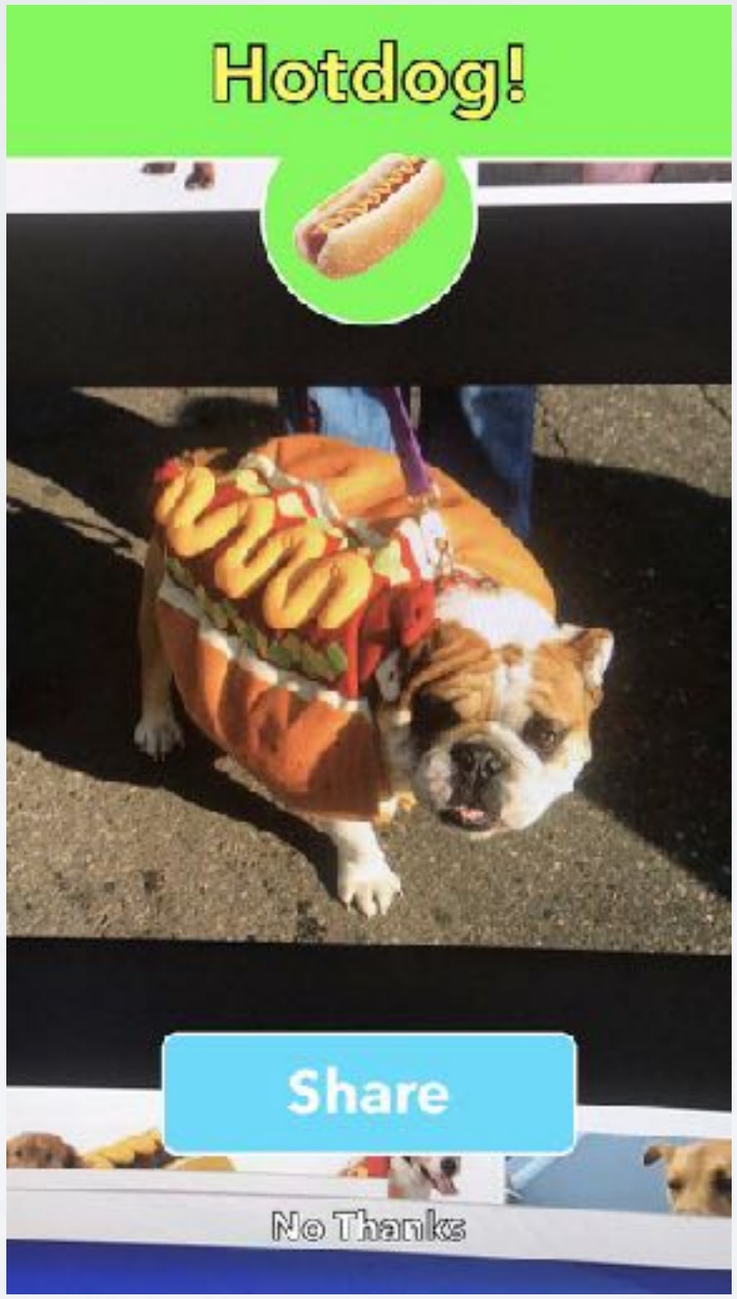
EXPERIMENTS

Missing Data Treatment / Imputation



- different strategies for different columns?
- what is the right amount of missing ness?
- are these default methods enough?

VALIDATION

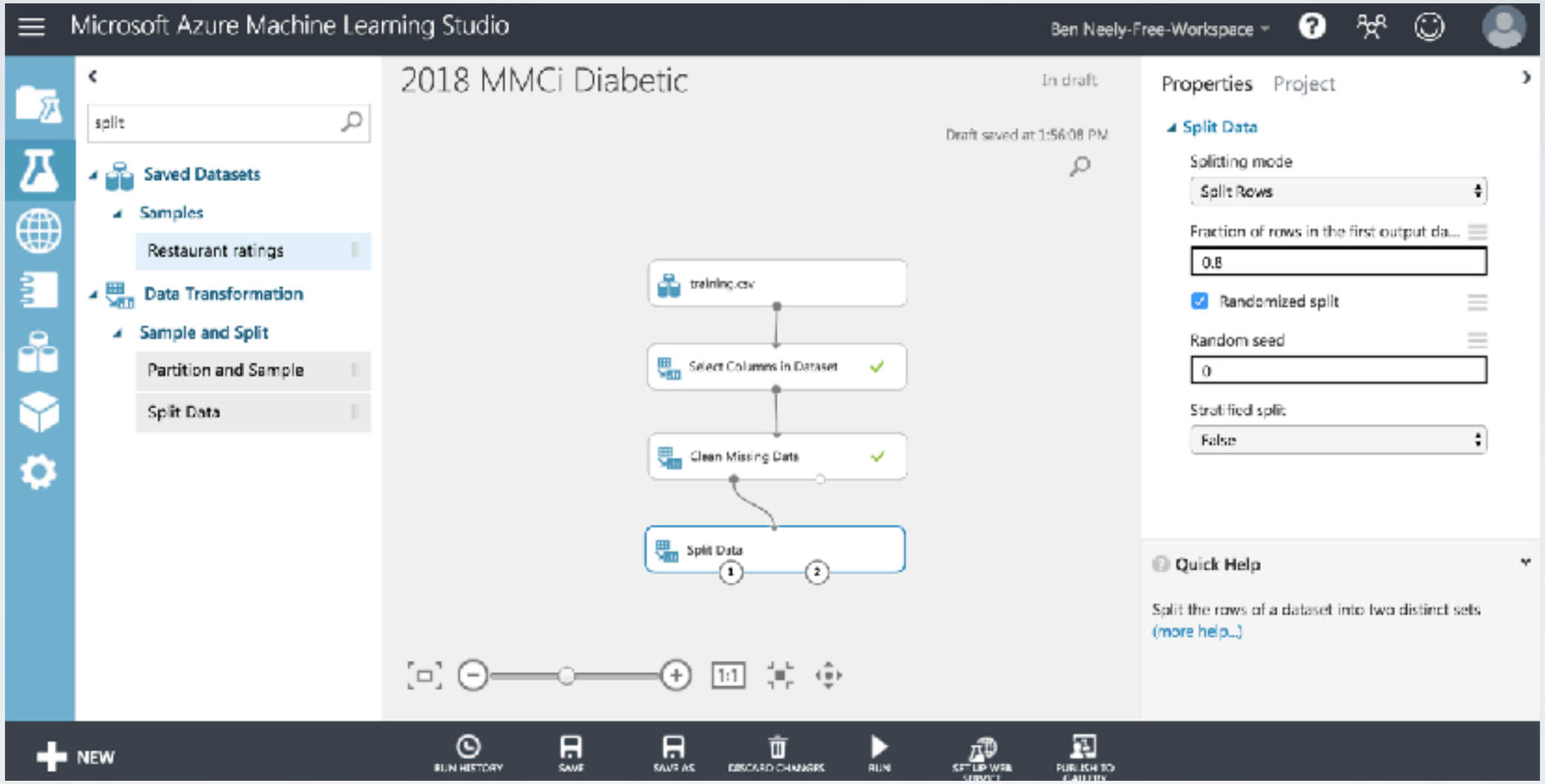


VALIDATION

- Internal
  - Cross-Validation
    - Split-Sample
- External
  - Geographic / Temporal
  - Fully Independent

EXPERIMENTS

Split Sample Data



TRAINING

Regression

Two-Class Logistic Regression

Two-Class Bayes Point Machine

Classification

Two-Class Decision Jungle

Ordinal Regression

Poisson Regression

Two-Class Averaged Perceptron

Linear Regression

Clustering

Two-Class Boosted Decision Tree

Decision Forest Regression

Anomaly Detection

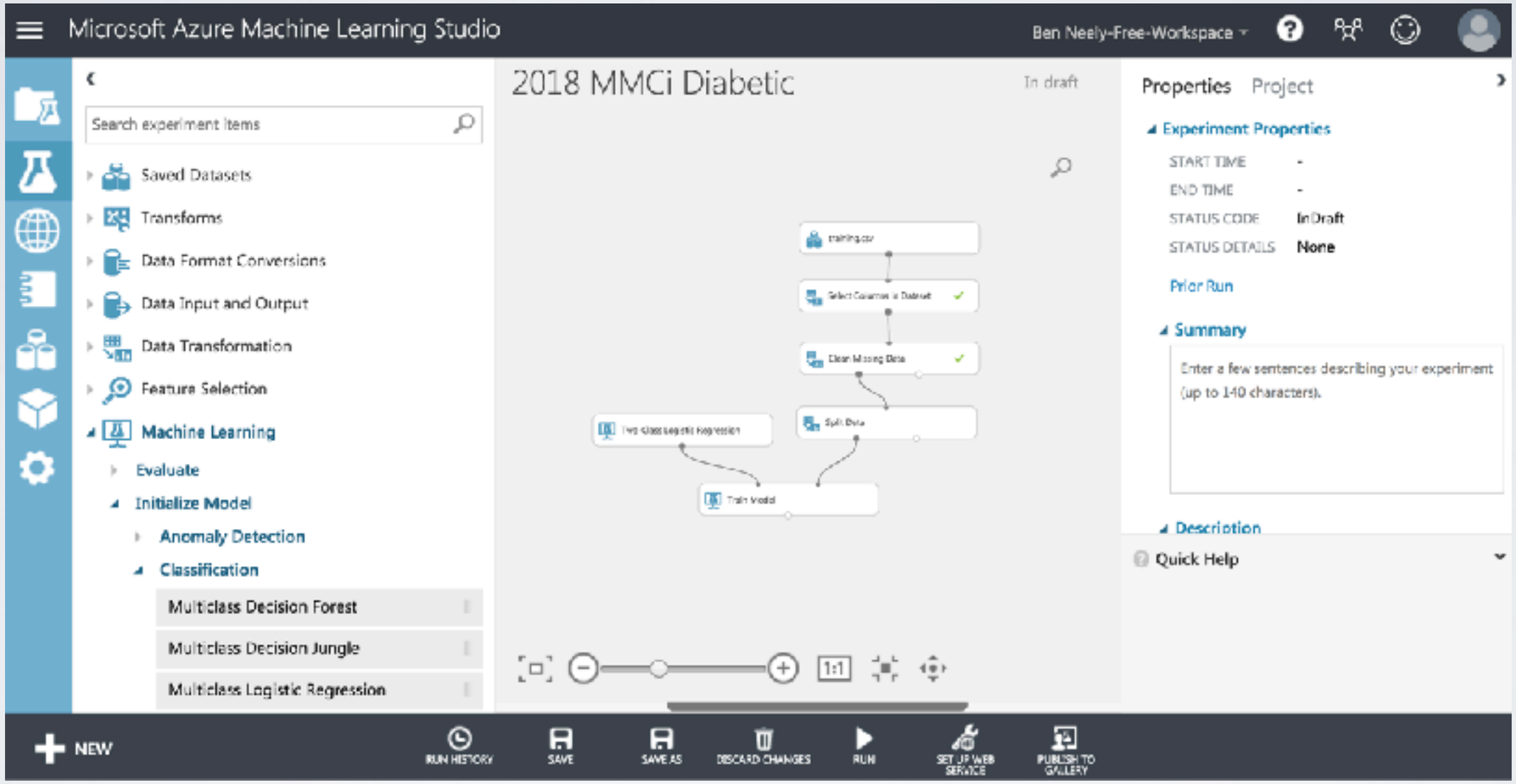
Two-Class Support Vector Machine

Two-Class Locally-Deep Support Vec...

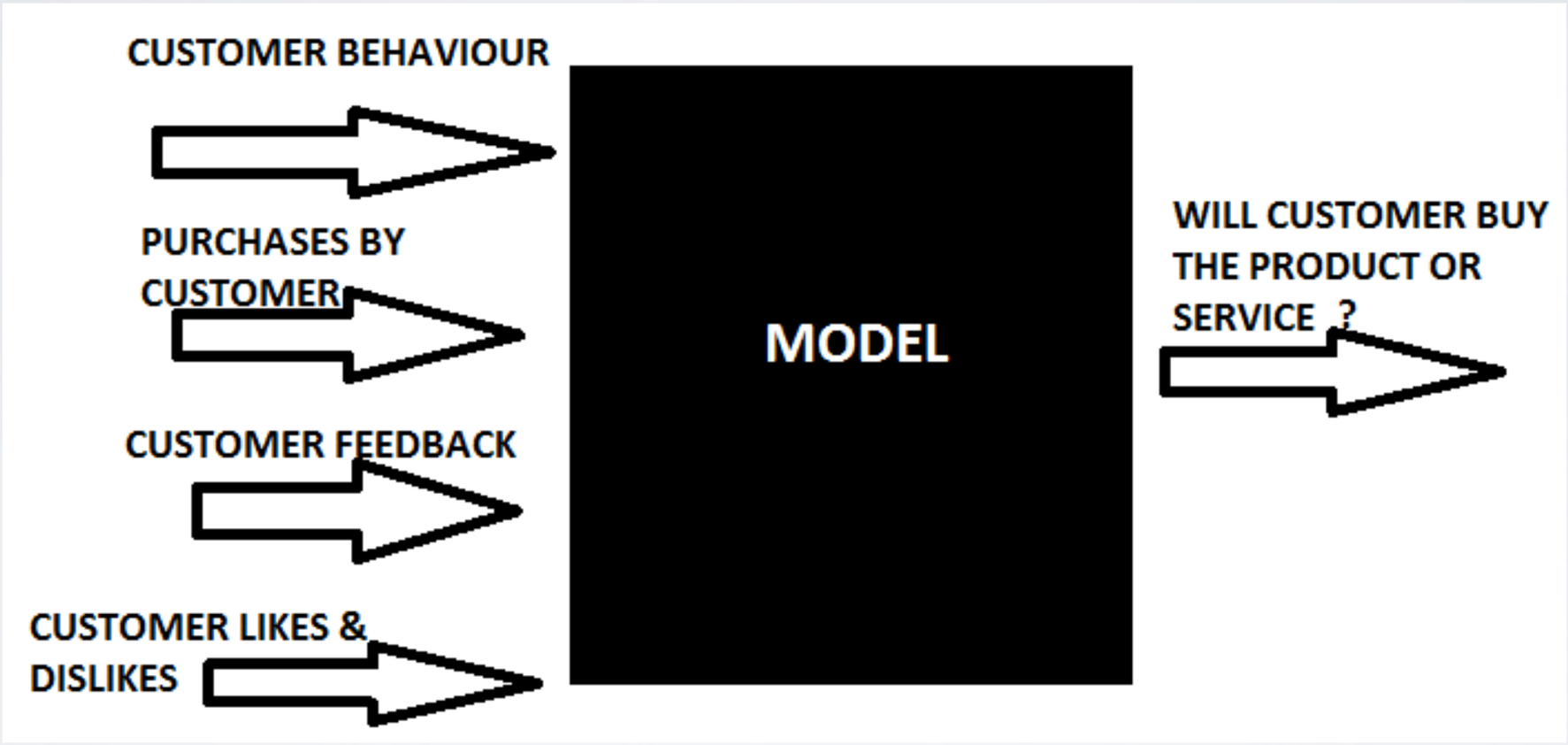
Two-Class Neural Network

EXPERIMENTS

Training

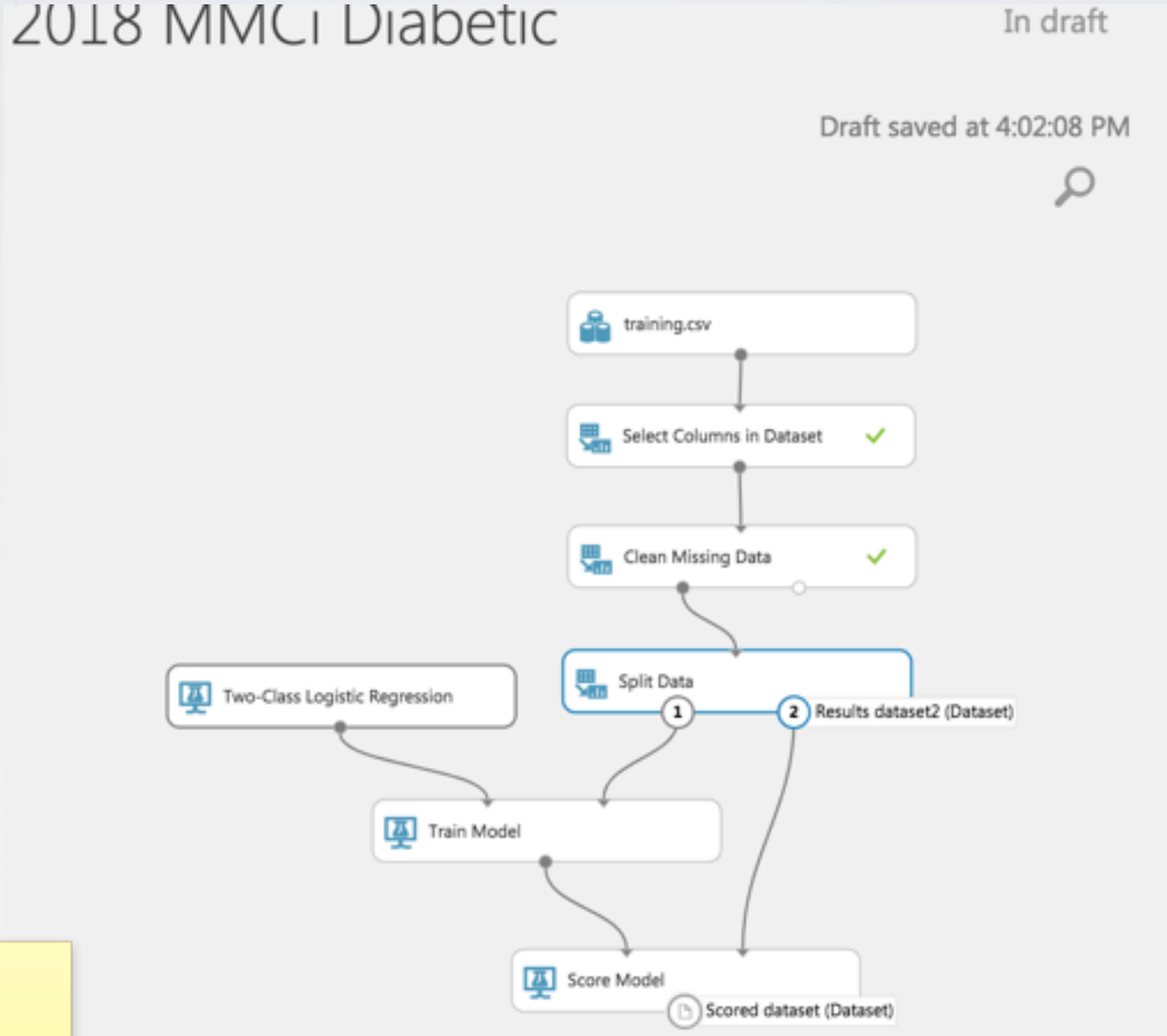


SCORE / PREDICT



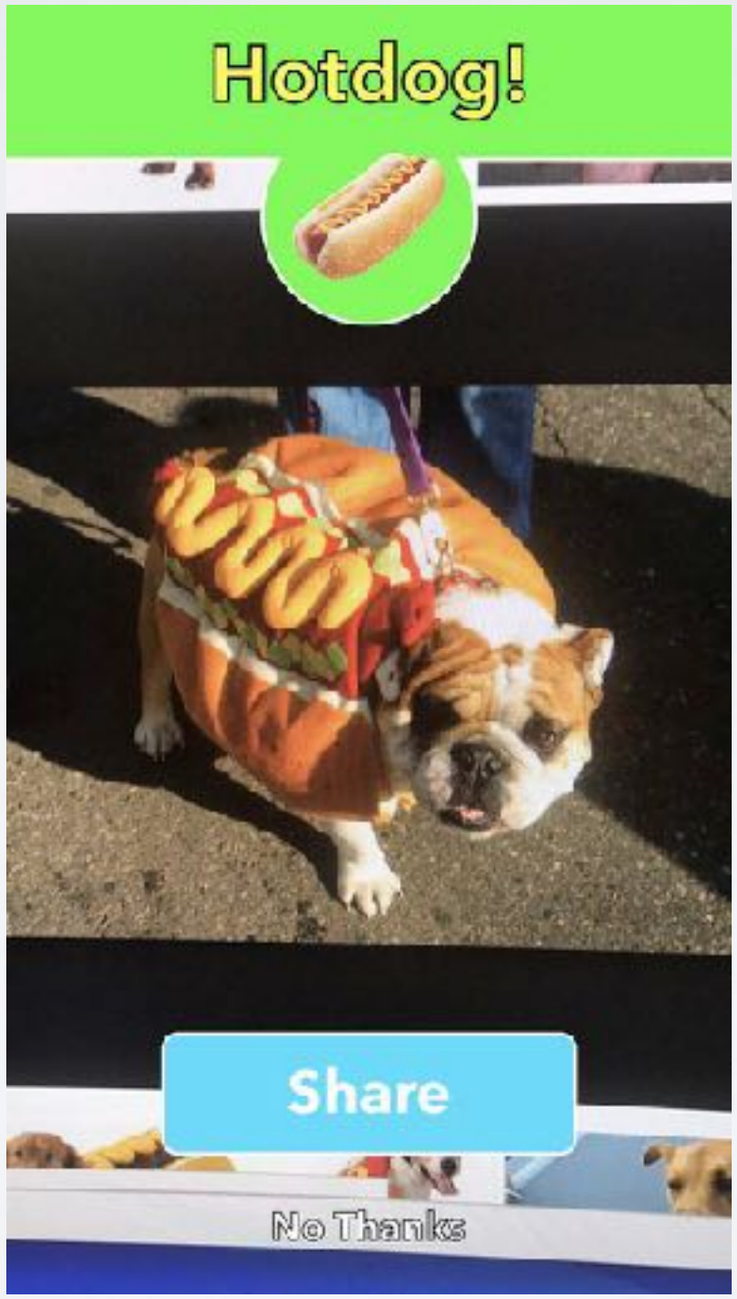
EXPERIMENTS

Scoring



- run experiment here and you can inspect the “black box”

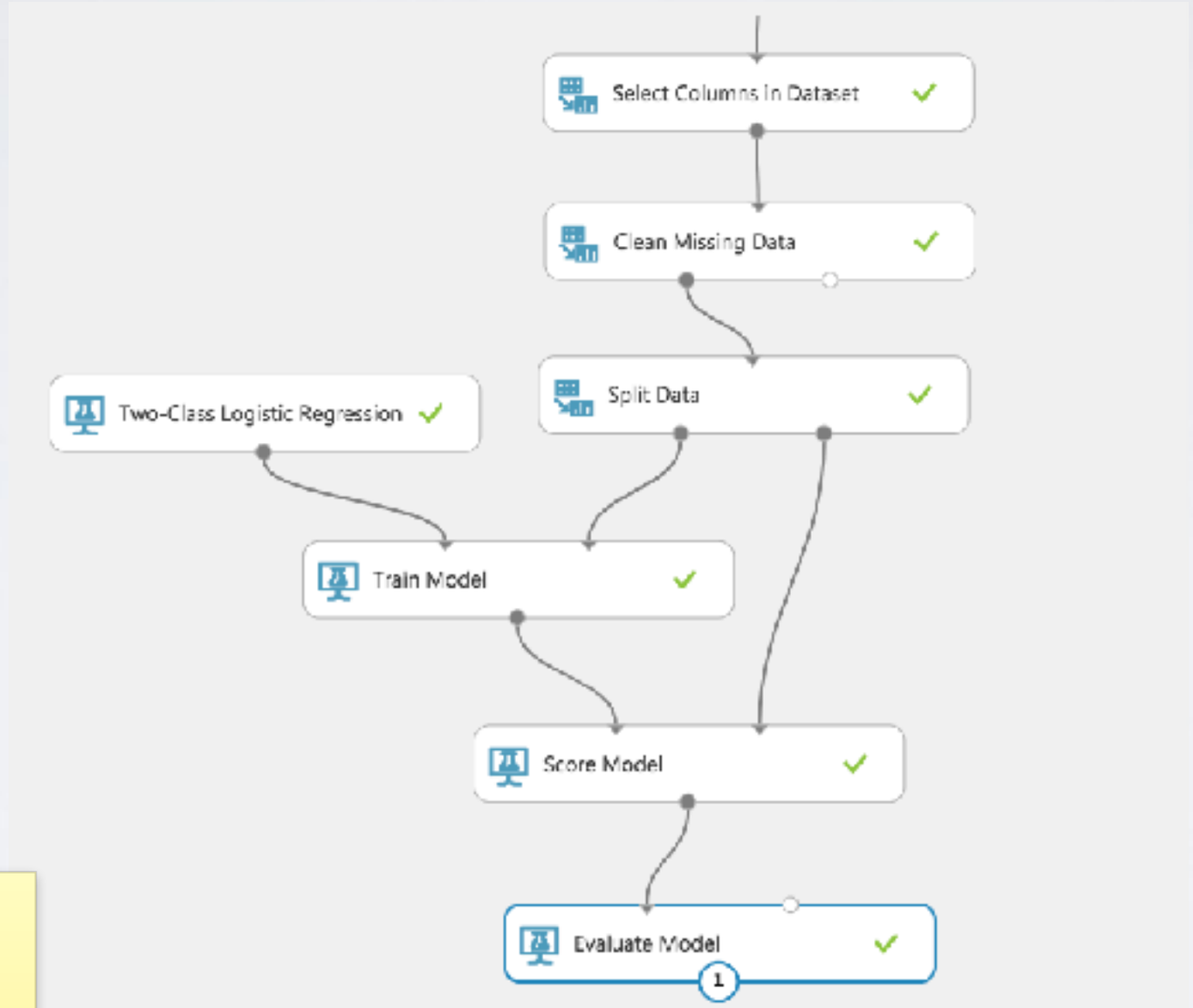
EVALUATION



- Favorite two models:
- perfect sensitivity
- perfect specificity

EXPERIMENTS

Evaluation

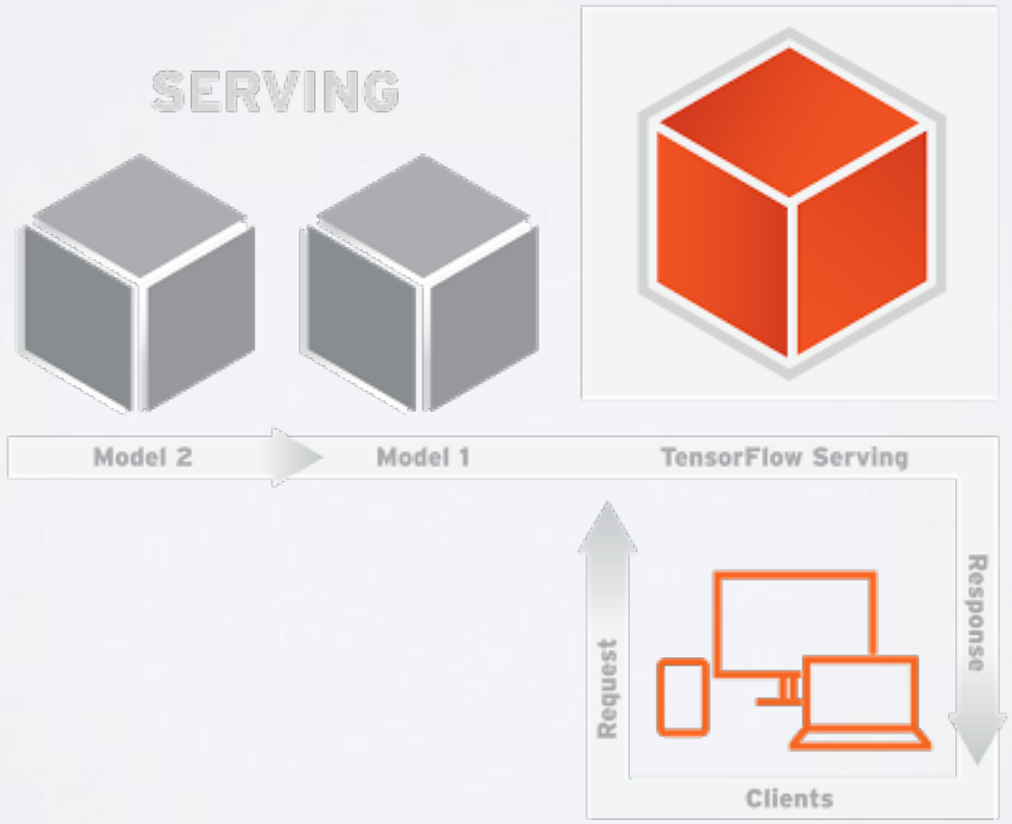


- self reporting AUC (shout outs)
- is this the end of the story?

ISOLATED

DEPLOYMENT


CONTINUOUS TRAINING PIPELINE
Data
Learner
Model 2
Model 1

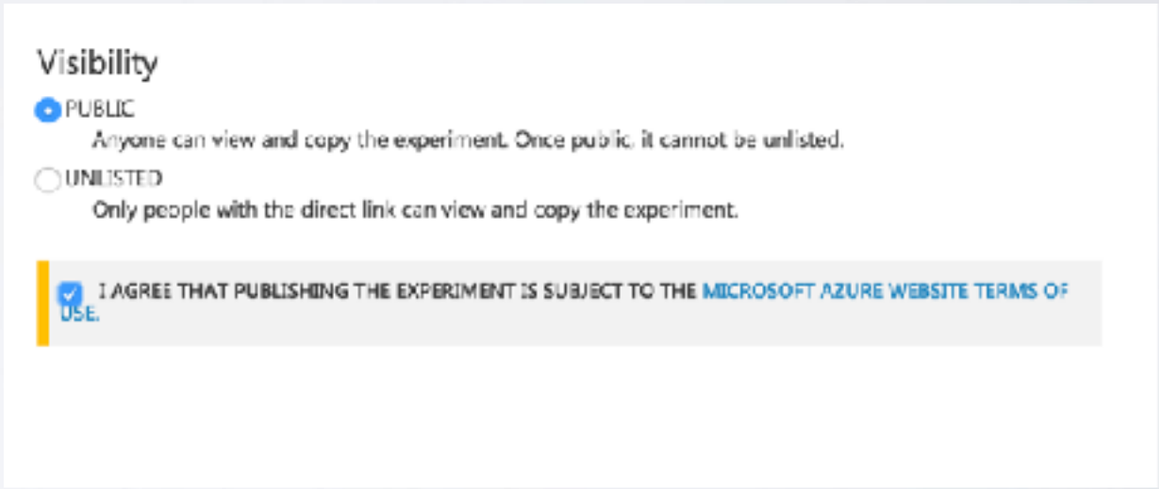

SERVING
Model 2
Model 1
TensorFlow Serving
Request
Response
Clients

DEPLOYMENT (ML STUDIO)

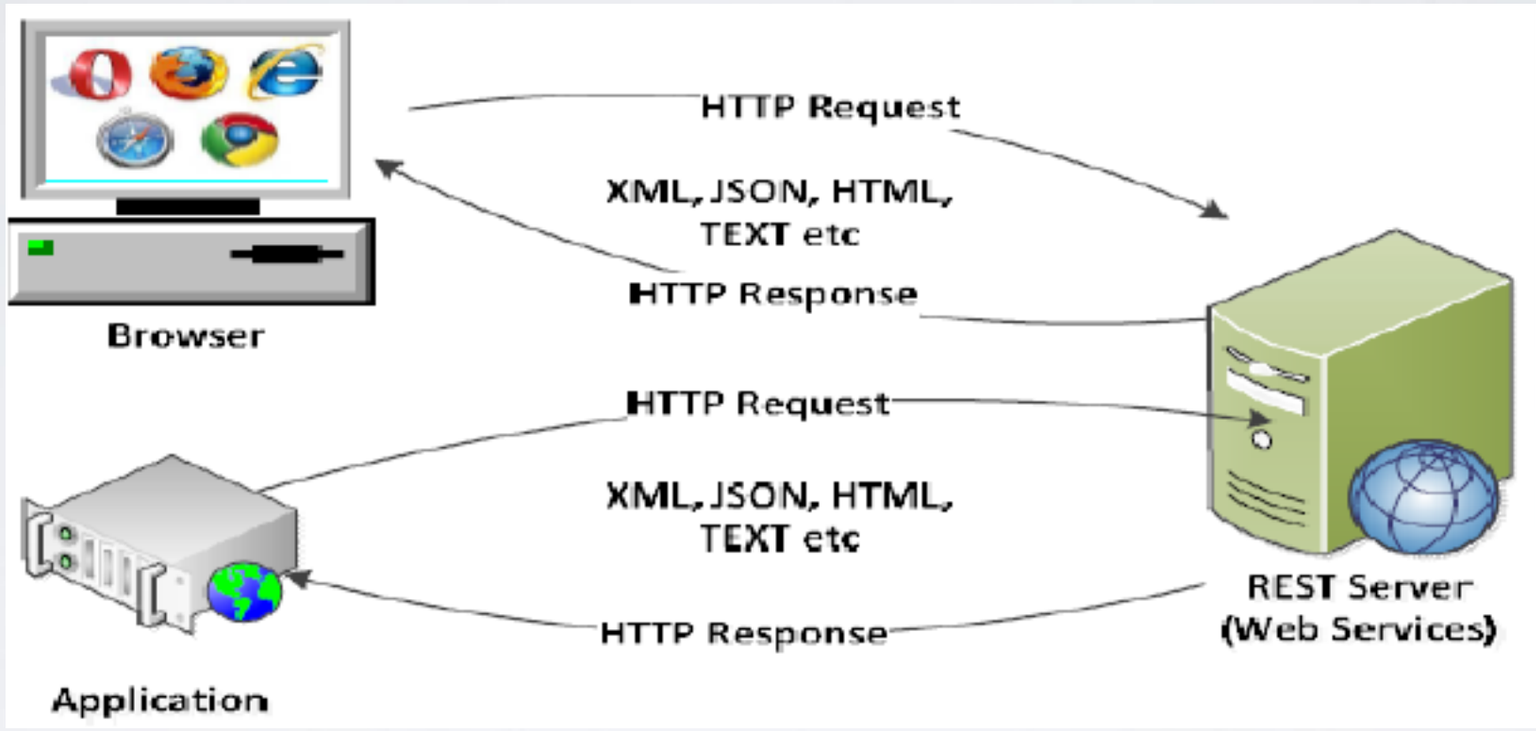
- Within Experiment:
  - click Save
  - click Run (play)
  - Hover Over **Set-Up Web Services:**
    - Predictive Web Service [Recommended]

WEB SERVICE IN ACTION

- Click on Experiments
- Choose Experiment that is the Web Service
- Click Run
- Publish to Gallery
  - Give Good Metadata (naming)
  - Don't need to upload image
  - See Screen shot to right for choosing visibility
  - Deploy Web Service



WEB SERVICE IN ACTION



TEST YOUR WEB SERVICE!

- Click Web Services
- Open your web service (click hyperlink under NAME)
- Test (Blue Box) - Fill In Data - Click check to score.
- Click Test (Not Blue Box) - Do same as above.
- Discuss experience.
- Potential Use cases?

POST INPUTS

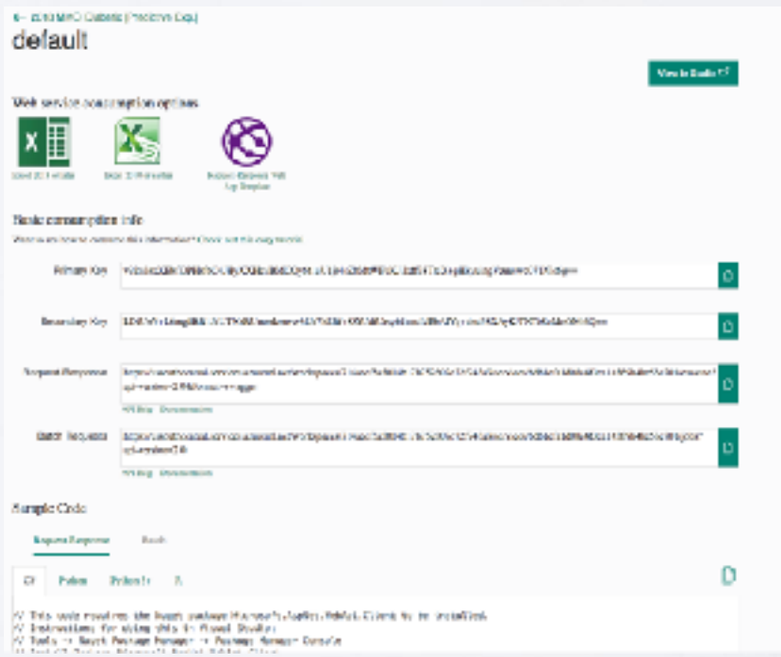
- RACE: Caucasian
- GENDER:Male
- AGE:[30, 40)
- WEIGHT: [150-175) (uh oh, what are the units????)
- NUM_LAB_PROCEDURES:33

HEALTH SYSTEM DESIGN

- How should this tooling be implemented (bed-side tools?)
- Cadence of predictions?
- Are the algorithms proprietary?
- Do patients who have contributed to the algorithm deserve a cut?
- Do the algorithms contain PHI (can we legally make them publicly available if we want to?)

BAKE OFF

- Web Services
- Click hyperlink under "NAME"
- Next to General, click New Web Services Experience
- Click Consume
- You Should see:



BAKE OFF

- Go to Google Sheets Link and enter requested data.