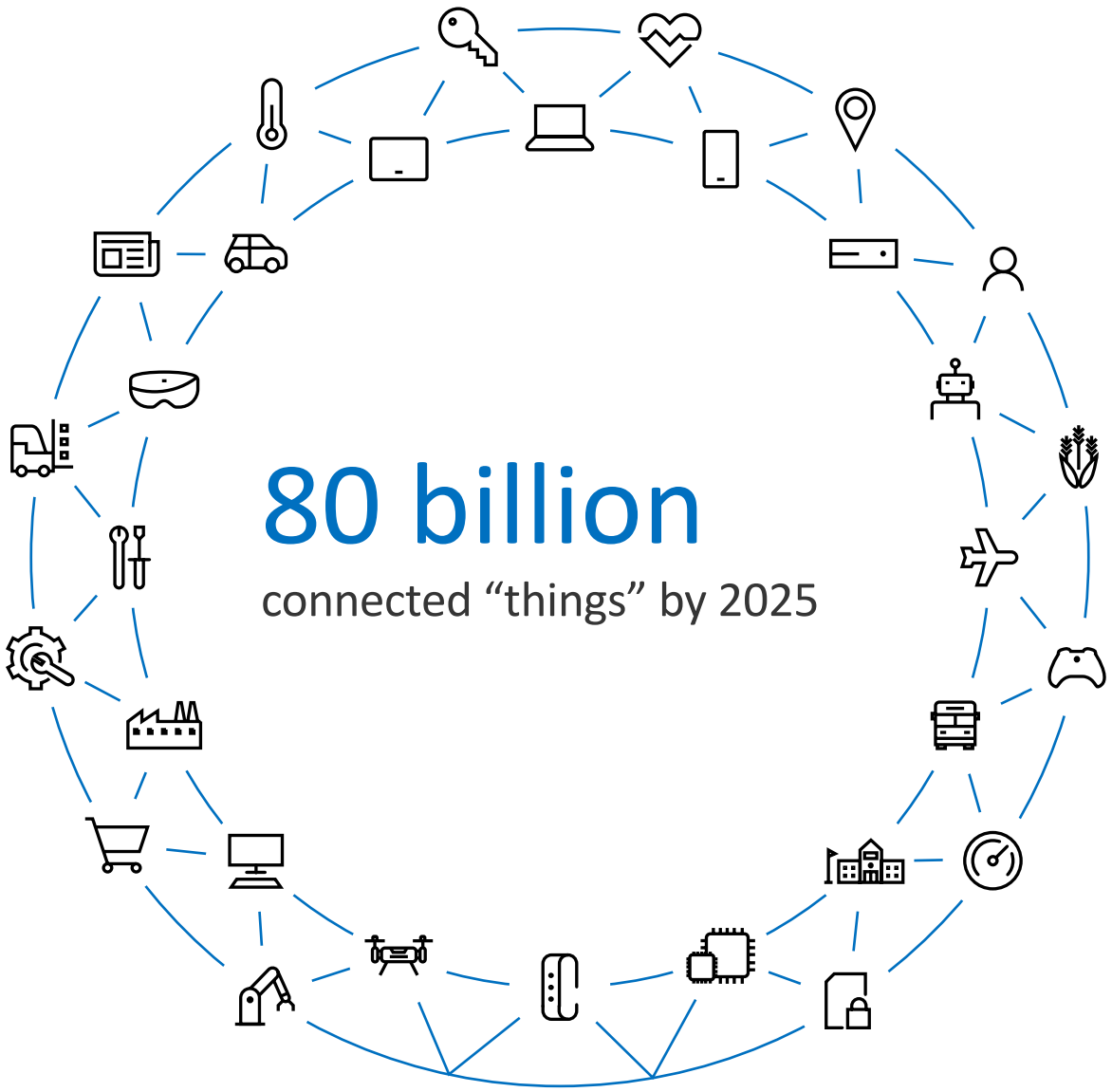
The Importance of Data

Caroline Matthews

Cloud Solution Architect



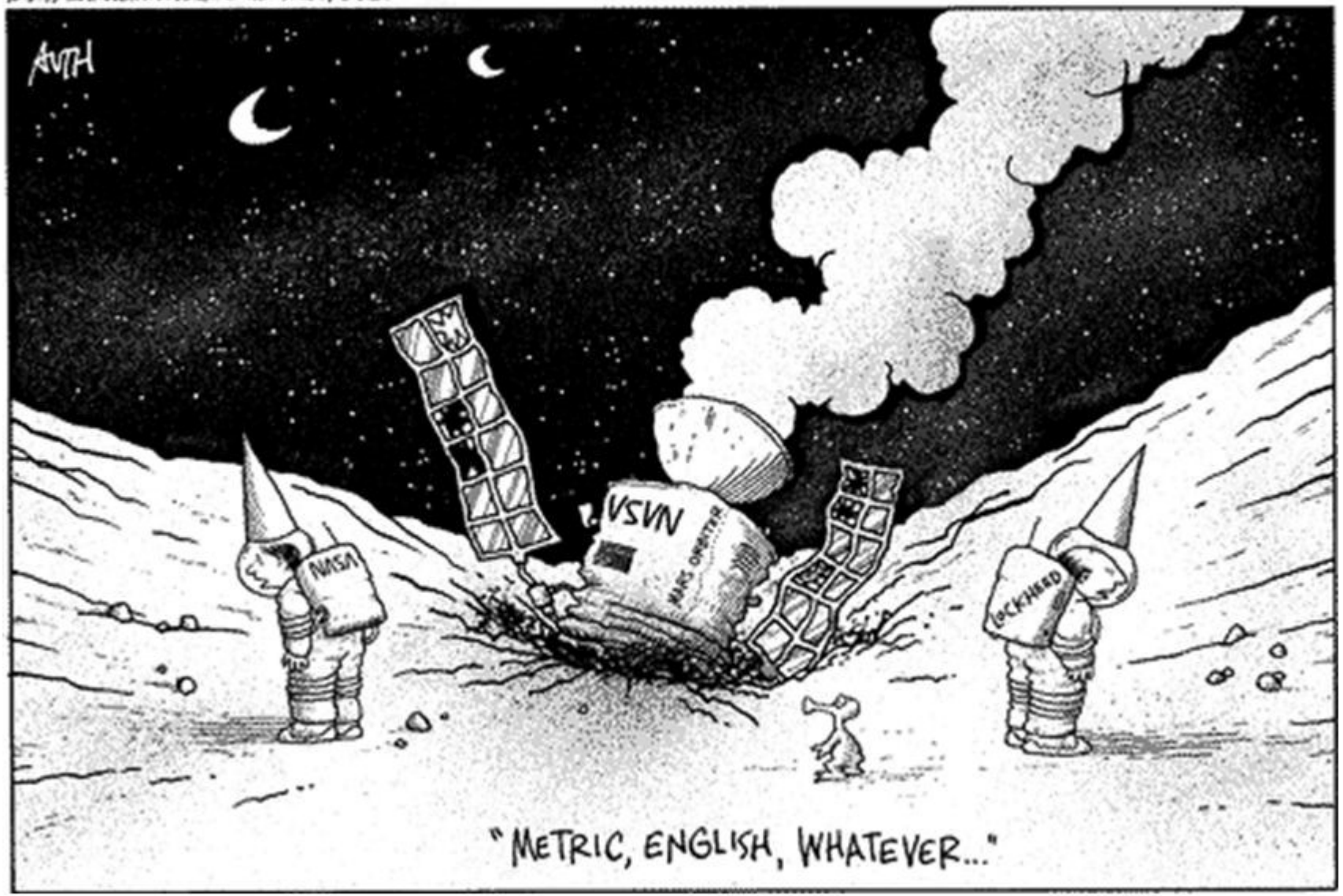
180 zettabytes

of digital data by 2025

One **Zettabyte** is 10^{21} bytes

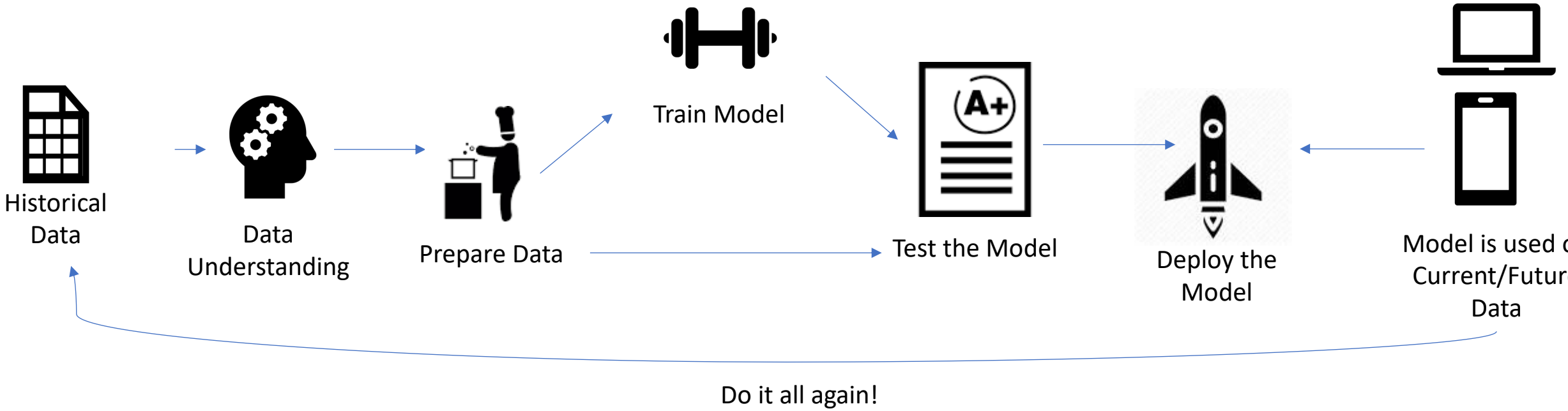
That is 1**,000,000,000,000,000,000,000** bytes

- a thousand Exabytes
- a billion Terabytes
- a trillion Gigabytes


AUTH
NASA
VSVN
MARS ORBITER
LOCKHEED
"METRIC, ENGLISH, WHATEVER..."

So what can we do?

Machine Learning Lifecycle



Explore & Prepare your Data

Look for outliers/missing data

Descriptive Statistics | Visualization

Cleansing techniques

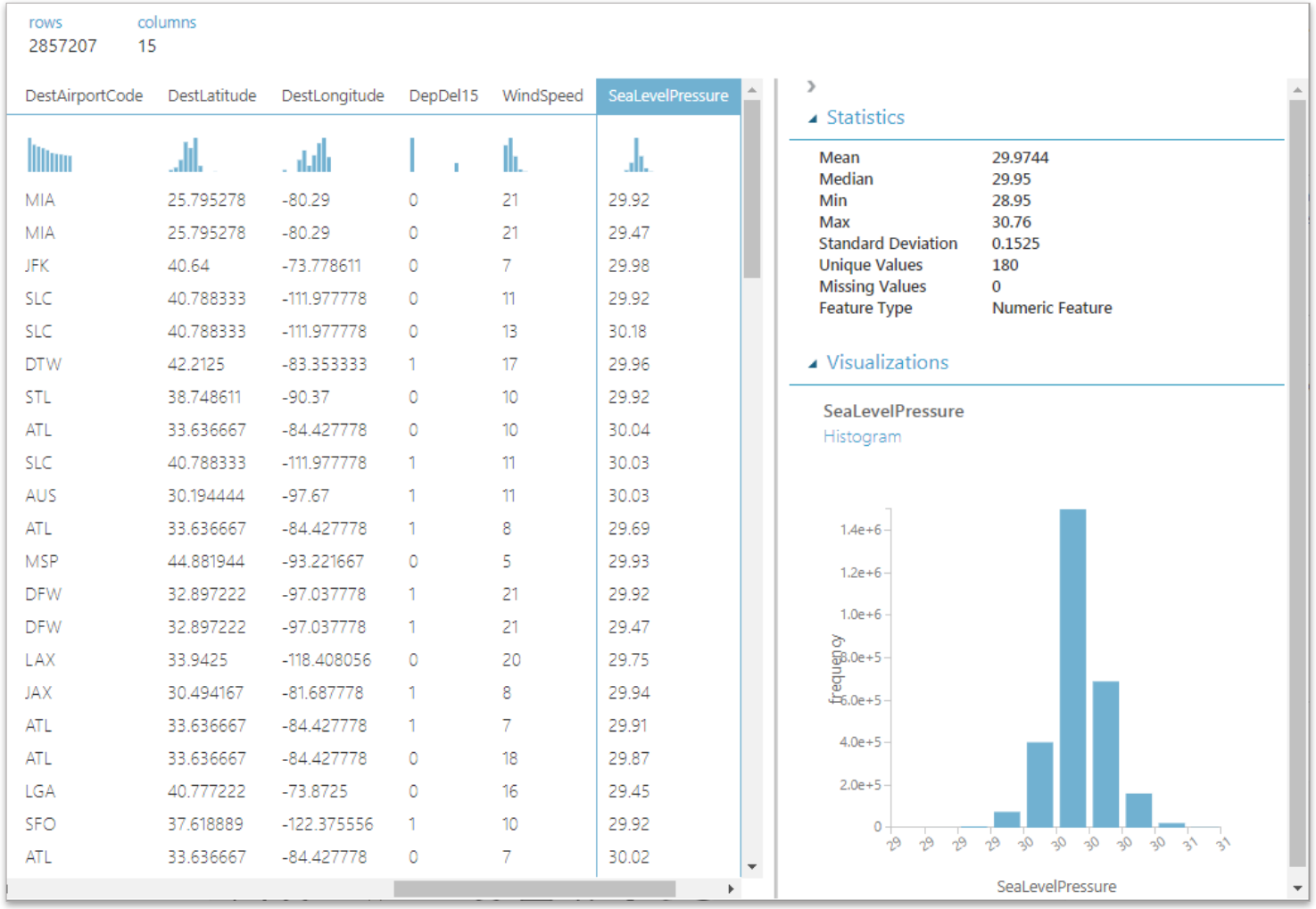
Remove | Substitute | Estimate

Feature Selection

What features are most important?

Feature Engineering

Be creative! What additional features can we calculate?



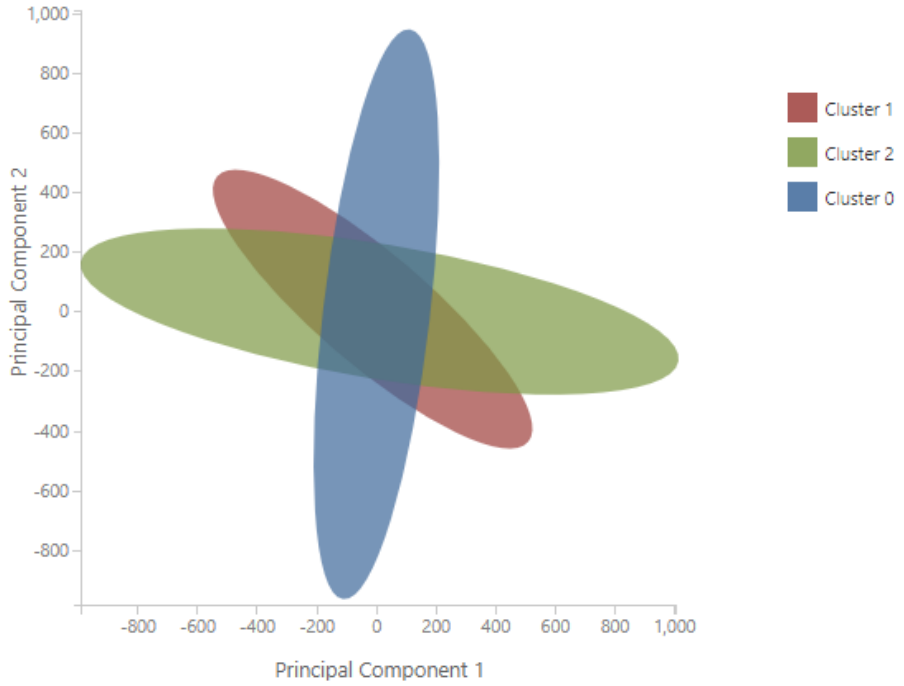
Train Models

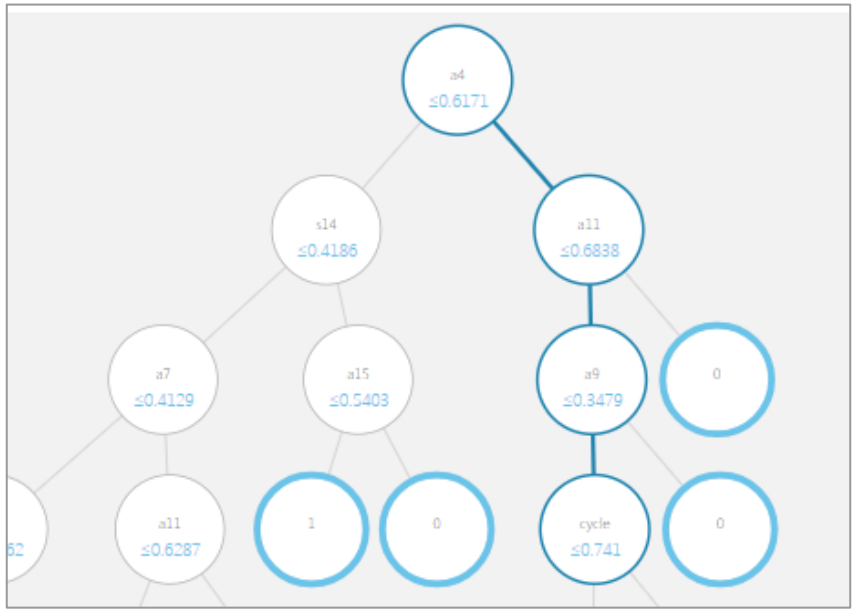
Algorithm Categories

- ✓ Classify – predict yes/no
- ✓ Regression – estimate numerical values
- ✓ Clustering – create similar looking groups of observations

Train with a subset of prepared data

Train many models – Experiment!



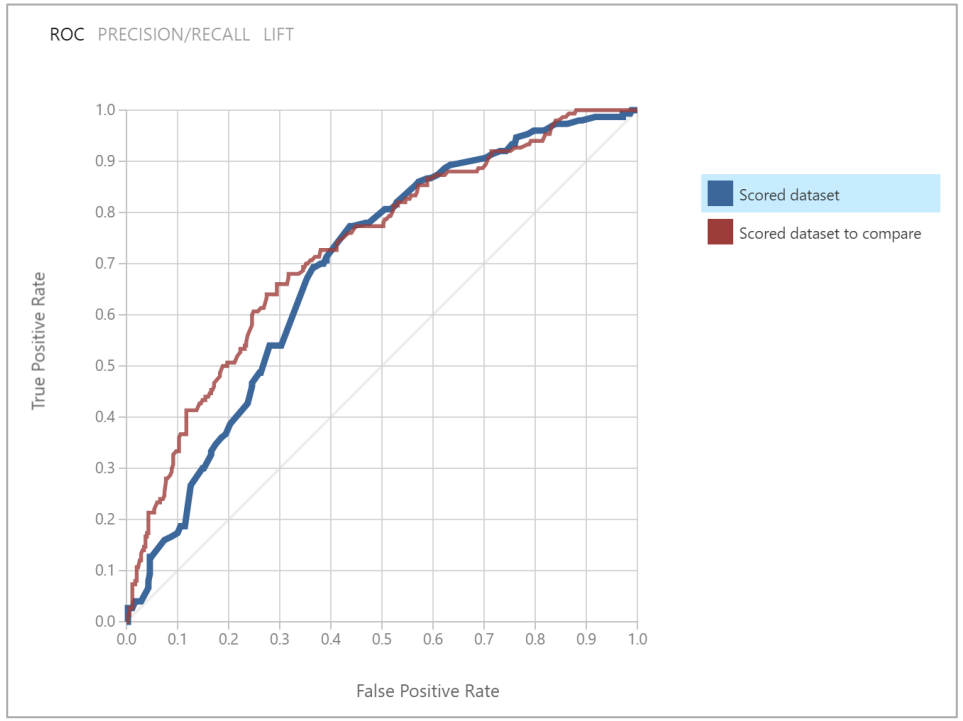


Evaluate Models

Run model with holdout/test data set

Measure (Grade your models!)

Compare Models



Predicted Class

Actual Class

	CCAT	ECAT	GCAT	MCAT
CCAT	89.3%	0.6%	6.9%	3.3%
ECAT	31.4%	42.0%	19.4%	7.2%
GCAT	18.6%	0.4%	79.8%	1.2%
MCAT	16.3%	1.1%	2.4%	80.2%

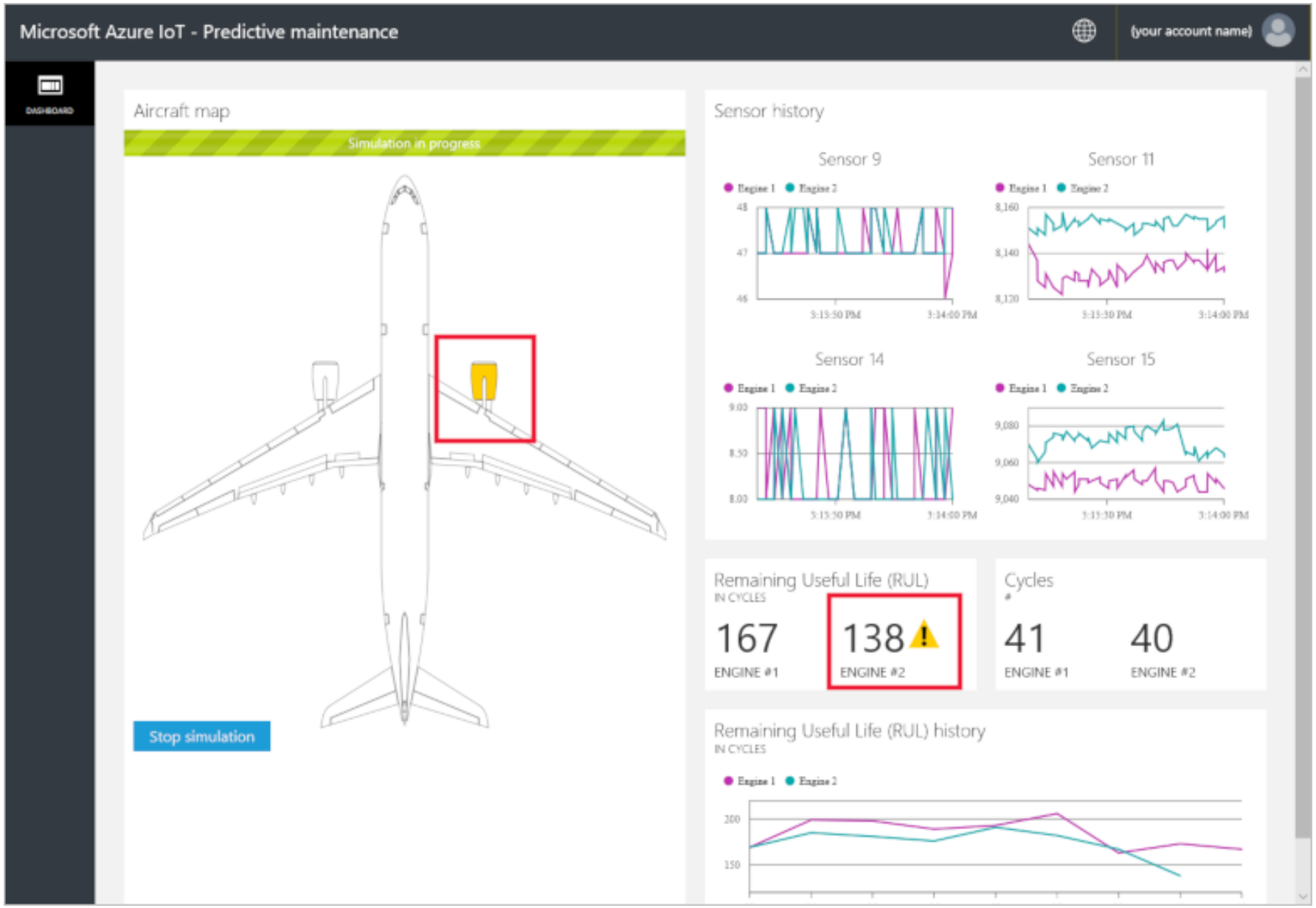
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
10796	166662	0.795	0.553	0.5	0.706

False Positive	True Negative	Recall	F1 Score
8711	670993	0.061	0.110

Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall
(0.900,1.000]	14	7	0.000	0.793	0.000	0.667	0.000	0.793	1.000
(0.800,0.900]	57	26	0.000	0.793	0.001	0.683	0.000	0.793	1.000
(0.700,0.800]	331	123	0.001	0.793	0.005	0.720	0.002	0.793	1.000
(0.600,0.700]	1995	1200	0.004	0.794	0.026	0.639	0.014	0.795	0.998
(0.500,0.600]	8399	7355	0.023	0.795	0.110	0.553	0.061	0.801	0.987
(0.400,0.500]	22006	28264	0.081	0.788	0.265	0.470	0.185	0.816	0.946
(0.300,0.400]	39287	73248	0.213	0.748	0.401	0.395	0.406	0.844	0.838
(0.200,0.300]	49348	146393	0.441	0.635	0.437	0.321	0.684	0.883	0.622
(0.100,0.200]	43247	262108	0.797	0.380	0.383	0.241	0.928	0.926	0.237
(0.000,0.100]	12774	160980	1.000	0.207	0.343	0.207	1.000	1.000	0.000

Deploy Models



Continuously Improve

Gather model results and compare to actual outcomes

Monitor of Time

Feed data back into the process

And repeat!

We are going to need some data!

Data Repos

- ✓Kaggle: https://www.kaggle.com/datasets
- ✓UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets.html
- ✓UN: http://www.un.org/en/databases/index.html
- ✓World Health Organization: http://apps.who.int/gho/data/node.resources
- ✓CDC: https://wonder.cdc.gov/Welcome.html
- ✓Federal Highway Administration: https://nhts.ornl.gov/
- ✓Datahub Collections: https://datahub.io/collections
- ✓Awesome Public Datasets: https://github.com/awesomedata/awesome-public-datasets

Image / NLP Repos

- ✓MS Coco: http://cocodataset.org/#home
- ✓ImageNet: http://www.image-net.org/
- ✓Open Images: https://storage.googleapis.com/openimages/web/index.html
- ✓Twenty Newsgroups (UCI): https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups
- ✓Wikipedia Corpus: https://nlp.cs.nyu.edu/wikipedia-data/
- ✓Spoken Digit: https://github.com/Jakobovski/free-spoken-digit-dataset
- ✓Sentiment Analysis: http://help.sentiment140.com/for-students/