the rise of the machines

A primer to machine learning and predictive analytics using Azure ML

Beat Schwegler

head in the cloud feet on the ground

Twitter: @cloudbeatsch
Blog: http://cloudbeatsch.com


Email

If you like **Francis International Airport**, we recommend **Naked Lunch**.

Naked Lunch
409 FOLLOWERS

You listened to **Engineers**. Check out **I Like Trains**.

We Saw The Deep
I Like Trains

You listened to **Francis International Airport**. Here's an album you might like.

Everyone Is Having Fun
Jack Beauregard

People who listen to **William Fitzsimmons** are also listening to **Joseph Arthur**.

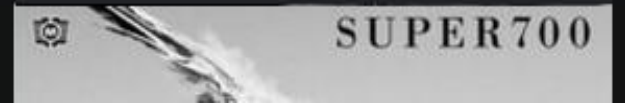
Joseph Arthur
13,486 FOLLOWERS

Like **Headphone** and **Balthazar**? Check out **dEUS**.

If you like **Most Unpleasant Men**, we recommend **The Secret Love Parade**.

If you like **Collapse Under The Empire**, we recommend **ef**.

You listened to **Francis International Airport**. Check out **Super700**.



predictive analytics
predictions based on models

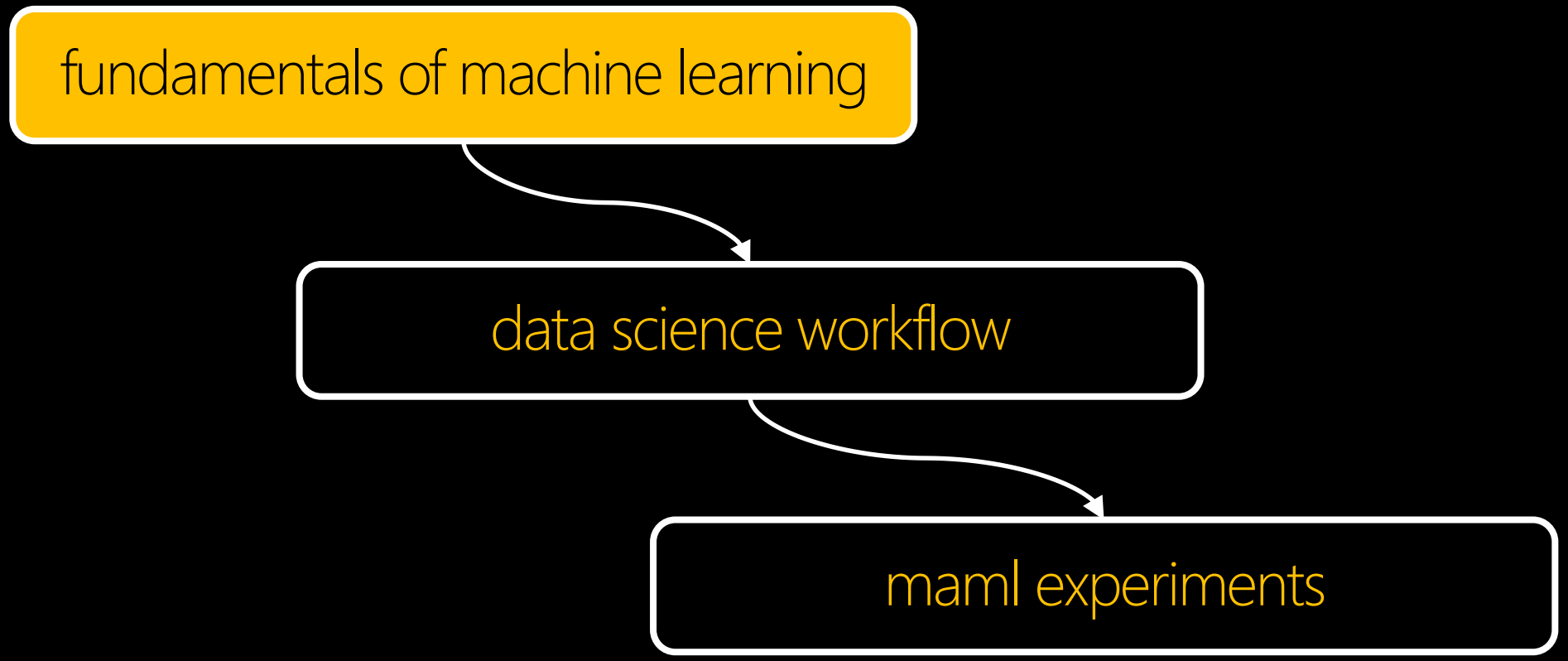
New York, NY (NYC) to Los Angeles, CA (LAX)

Mon, 12/03 - Fri, 12/21 · 1 adult · Economy · Change search

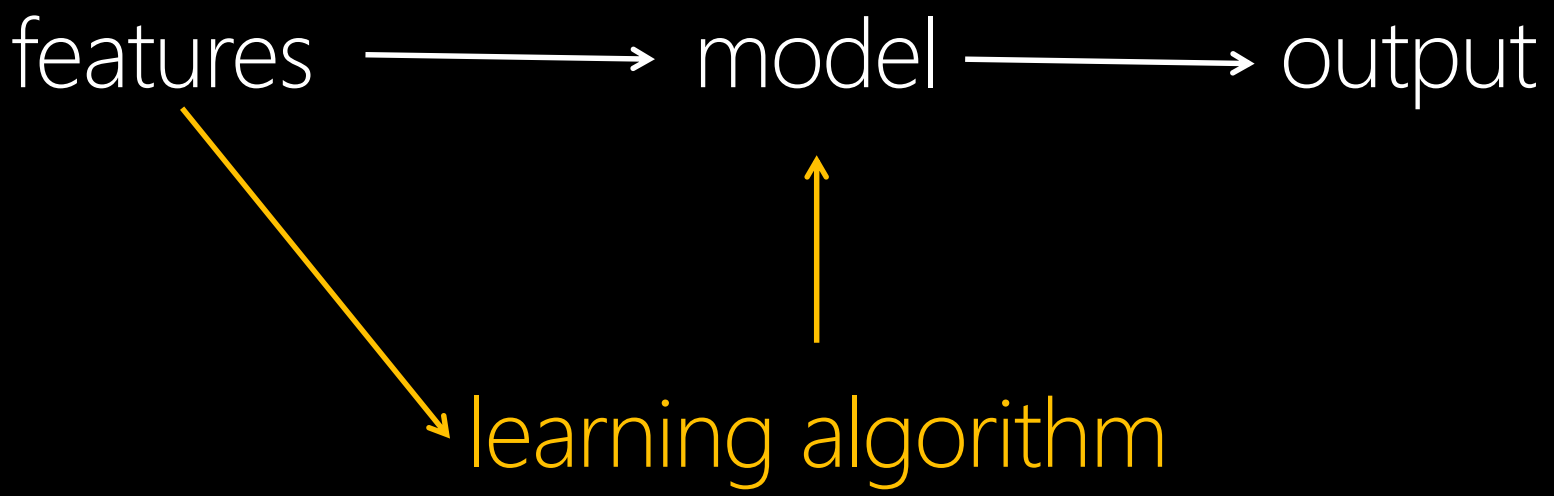
Tip: Buy · Fares rising $50+ · 80%+ Confidence

predictive maintenance
fix it before it breaks

predictive assistant
make predictions personal


fundamentals of machine learning
data science workflow
maml experiments

machine learning
algorithms and systems that improve
their performance with experience


features
model
output
learning algorithm

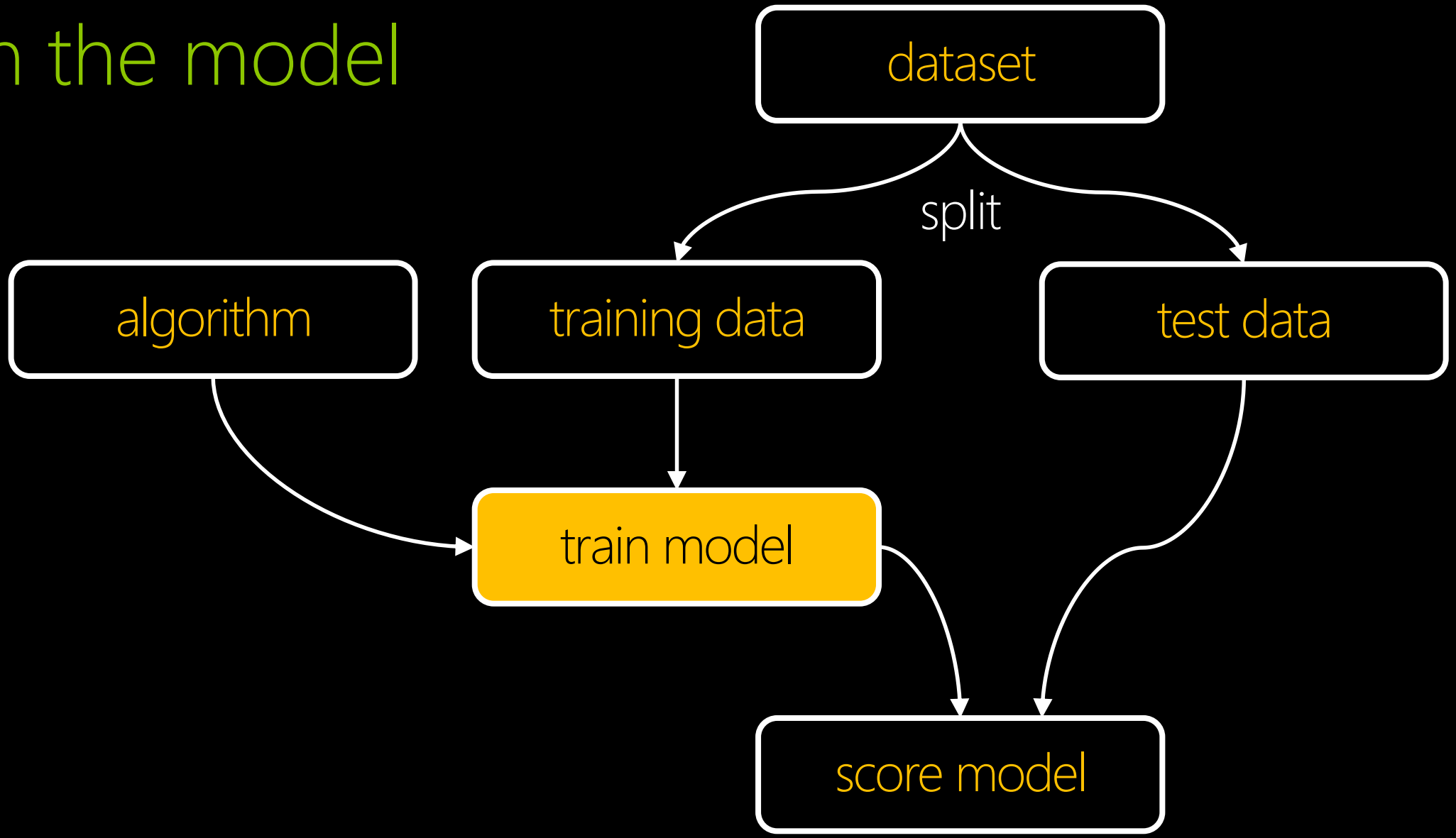
features describe the domain
e.g. income, age, education, …

labels augment the learning data
e.g. this photo contains a man

unsupervised learning
discovering clusters and associations using unlabeled data

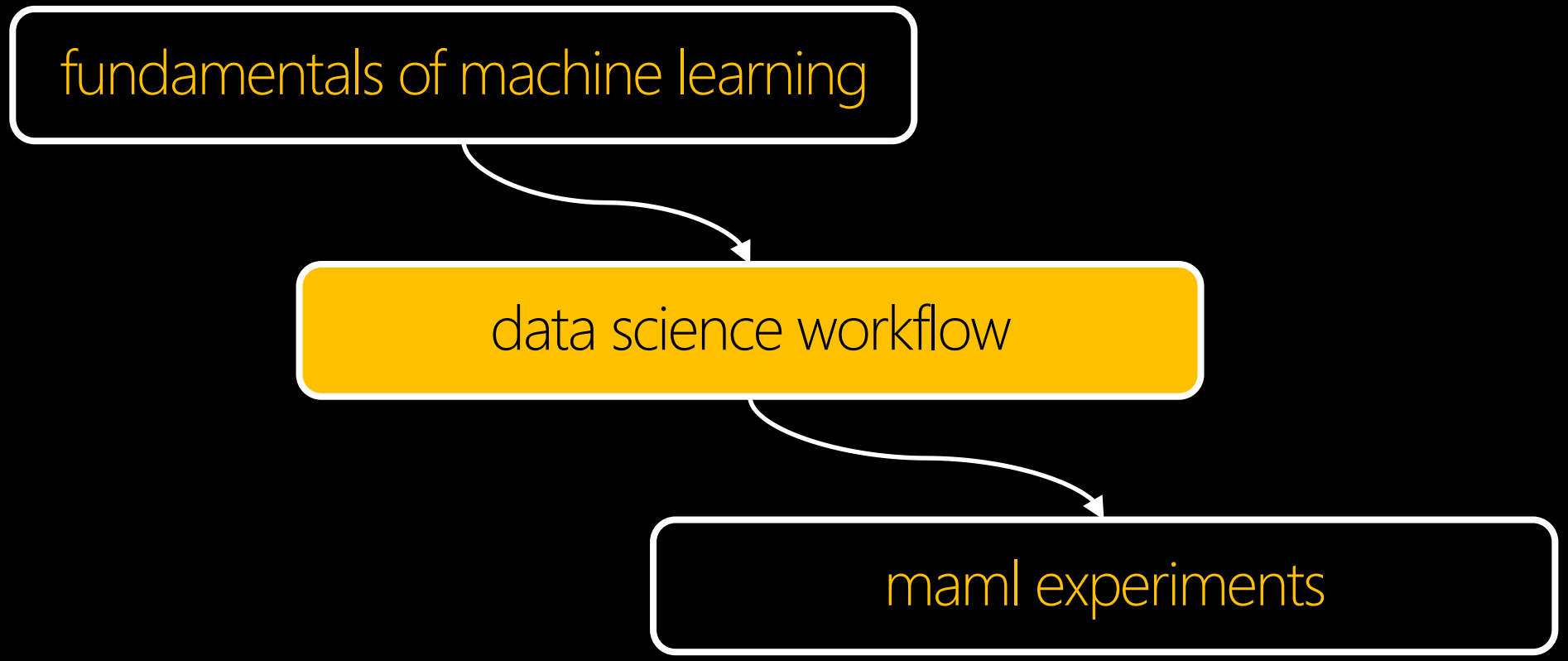
supervised learning
using labeled data to train the model

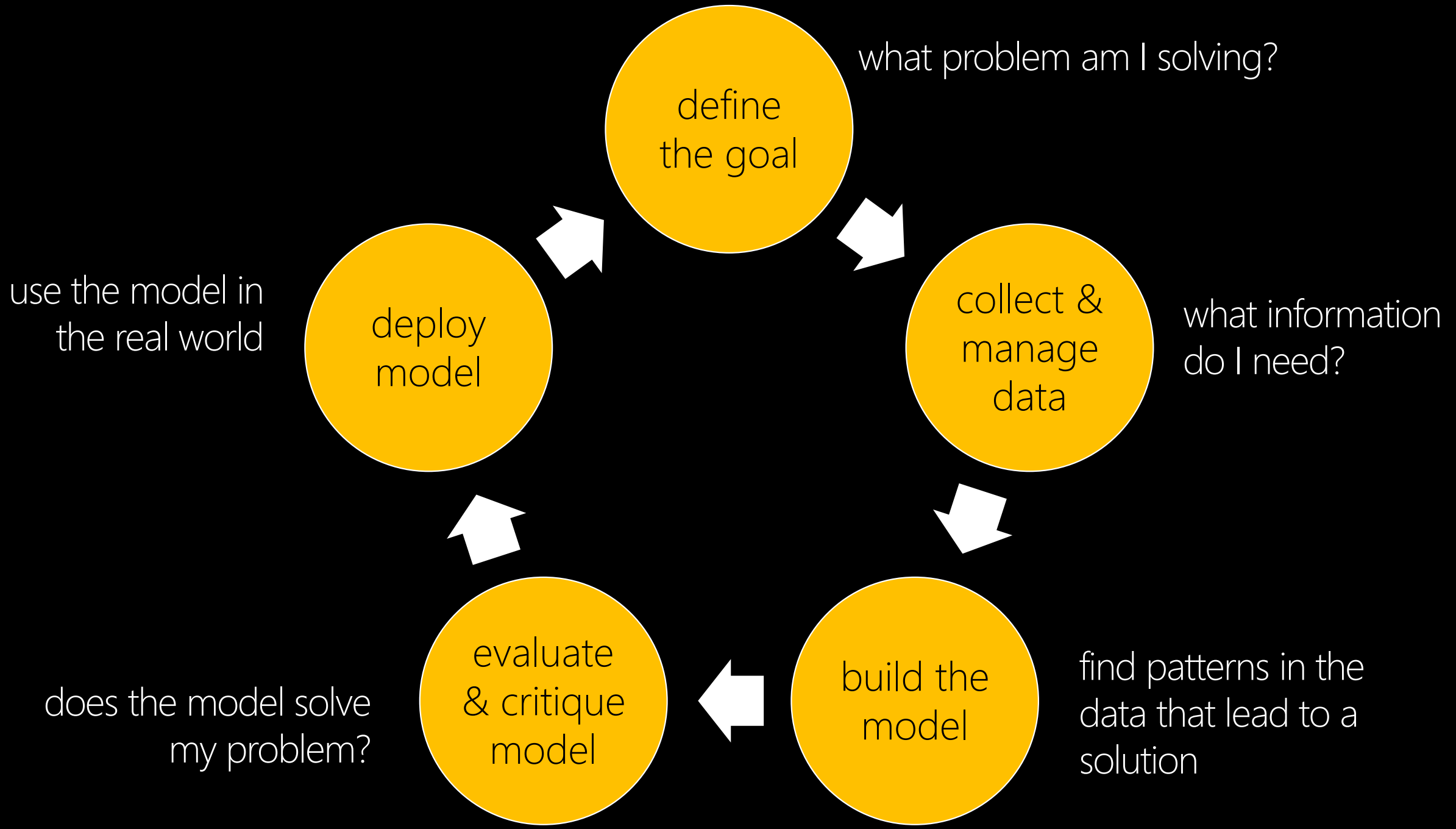
train the model



different tasks require different algorithms
finding similar companies k-means clustering
hourly bike rental prediction regression tree
credit risk prediction decision tree

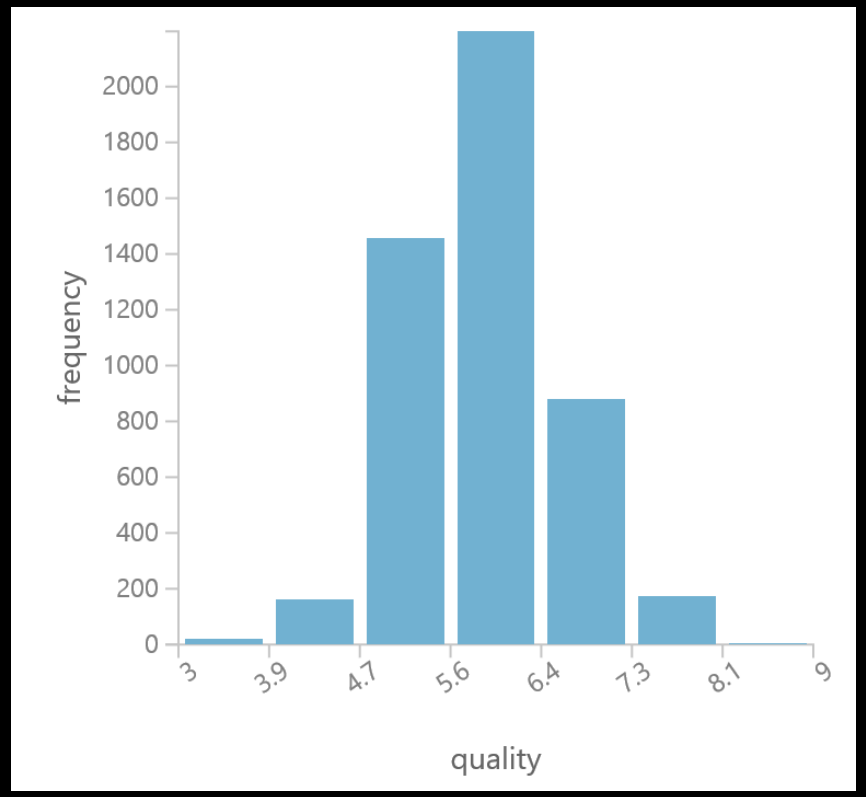
machine learning tools
open source: R, mahout, python, weka …
commercial: Azure ML, SAS, SPSS, MATLAB …

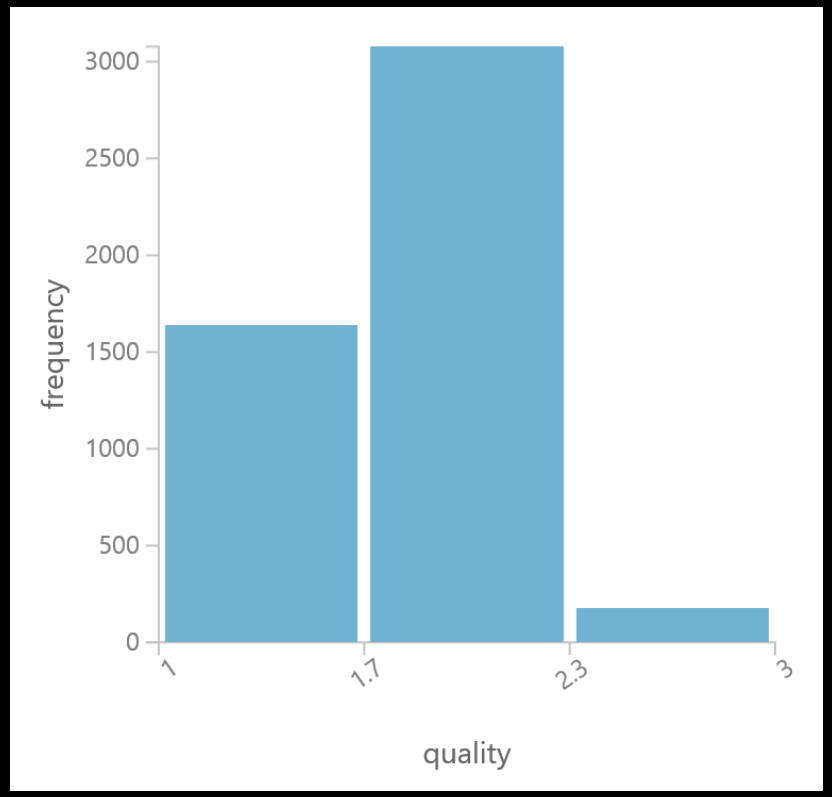

fundamentals of machine learning
data science workflow
maml experiments


define the goal
what problem am I solving?
collect & manage data
what information do I need?
build the model
find patterns in the data that lead to a solution
evaluate & critique model
does the model solve my problem?
deploy model
use the model in the real world

ask the right question!
when will the clutch fail?
versus
what is the probability that the clutch fails within the next 3 months?

feature construction
transform e.g. cumulative time dependent
clipping outliers e.g. faulty measurements
normalize e.g. avoid feature dominance
quantize e.g. create categories (great/bad)



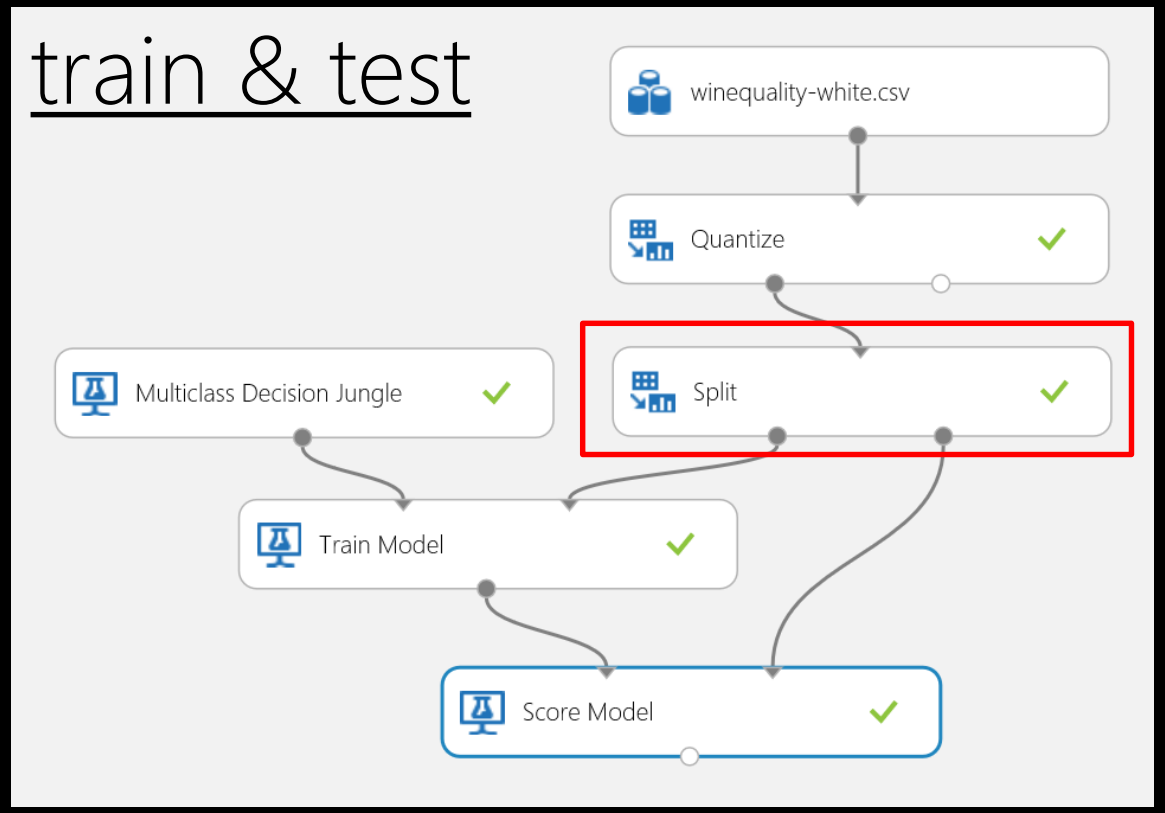


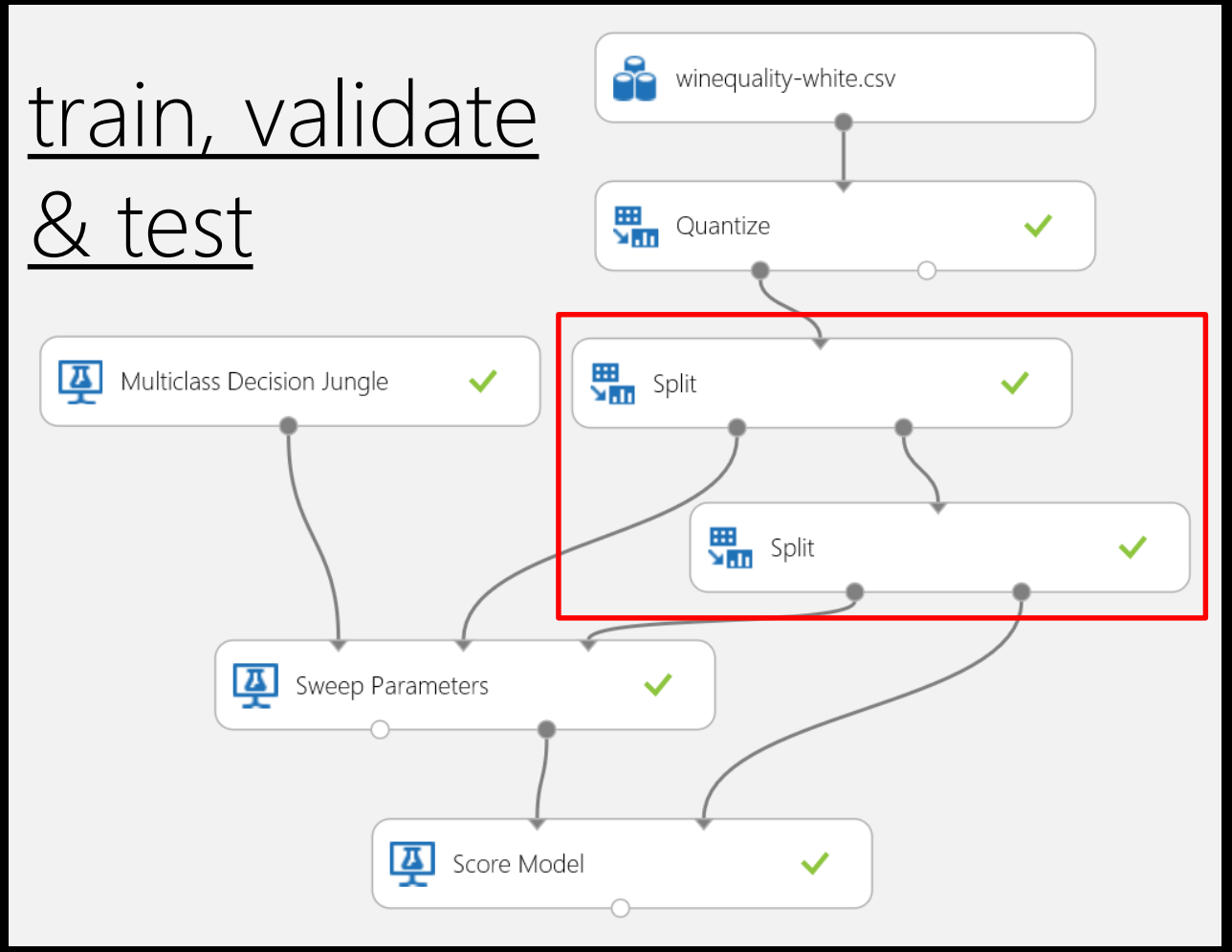
e.g. quantize wine quality ratings into bad/good/great buckets

train & test the model
better data often beats better algorithms
more training data != better model
does the algorithm correlate features?
does it support online learning?

deciding on algorithms
linear regression to predict
decision trees
    to classify classification tree (2 & multi)
    to predict regression tree

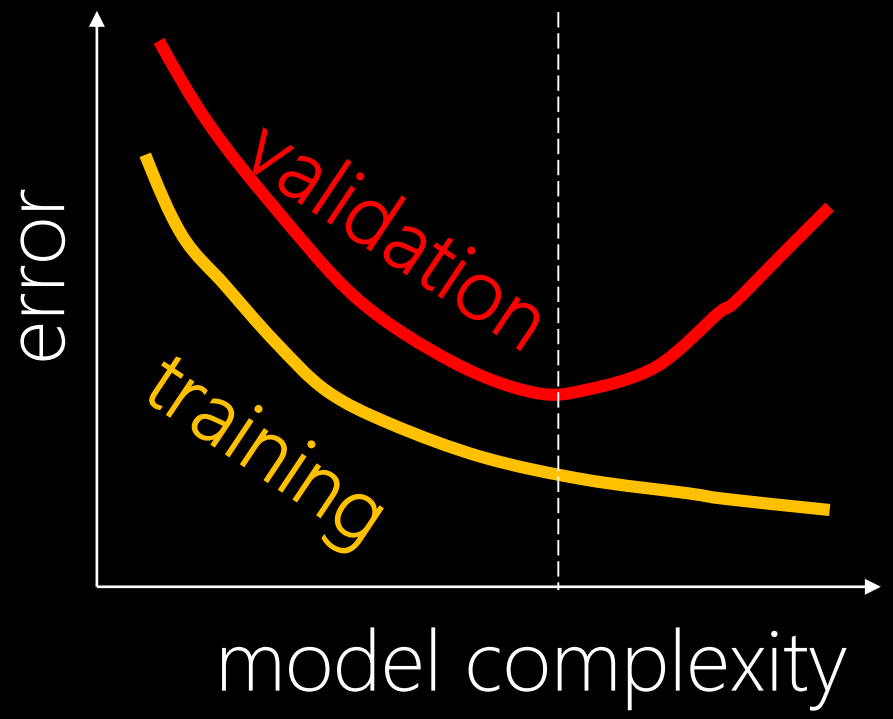
2-way partitioning versus 3-way partitioning



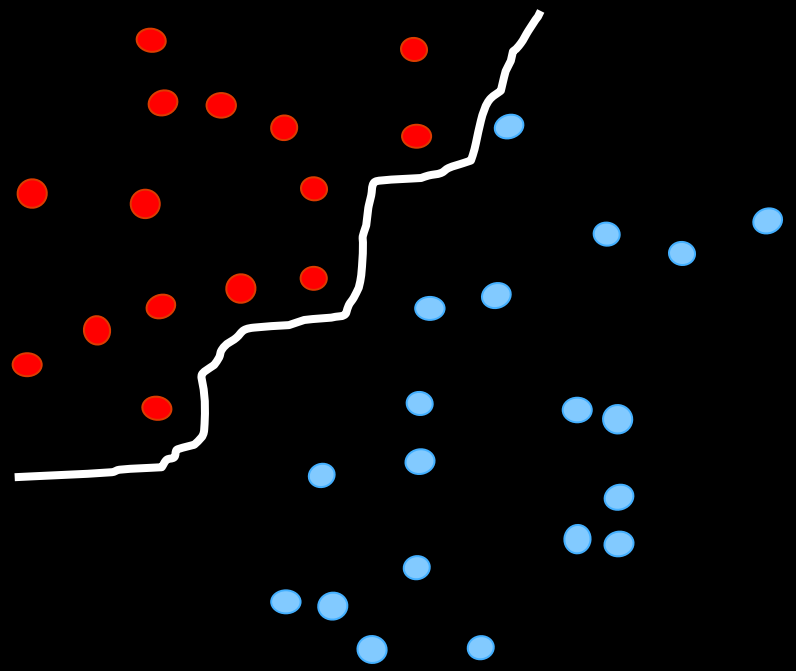


cross validation - divide dataset into n folds

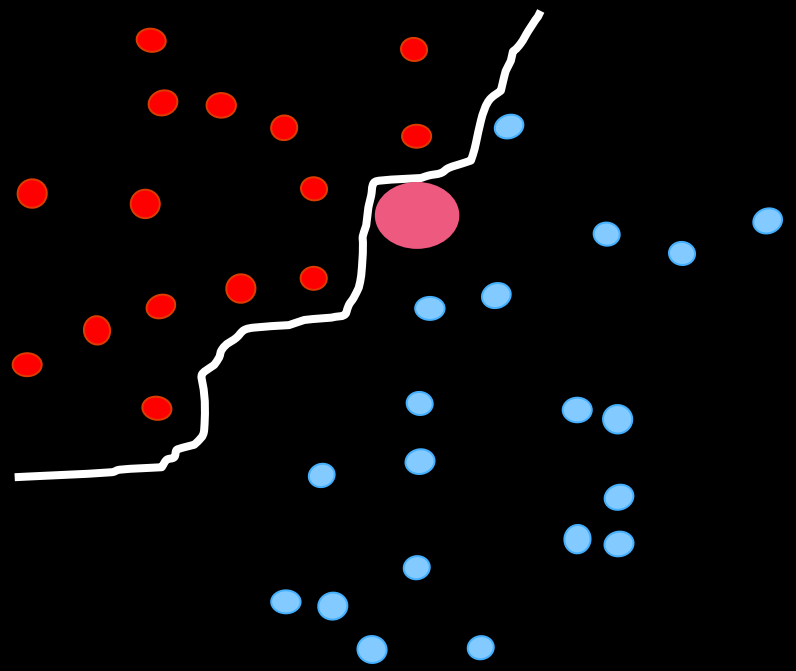
trade off between training error & validation performance



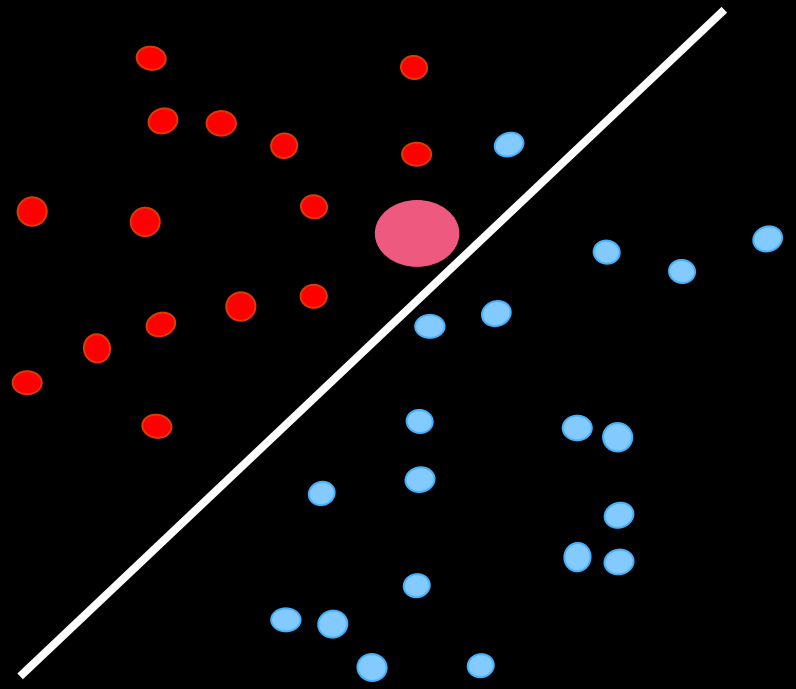
over fitting and generalization



over fitting and generalization



over fitting and generalization



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
621	129	0.733	0.790	0.5	0.758

False Positive	True Negative	Recall	F1 Score
165	185	0.828	0.809

accuracy: $\frac{TP+TN}{P+N}$ → 1 means all predictions are correct (over fitting)

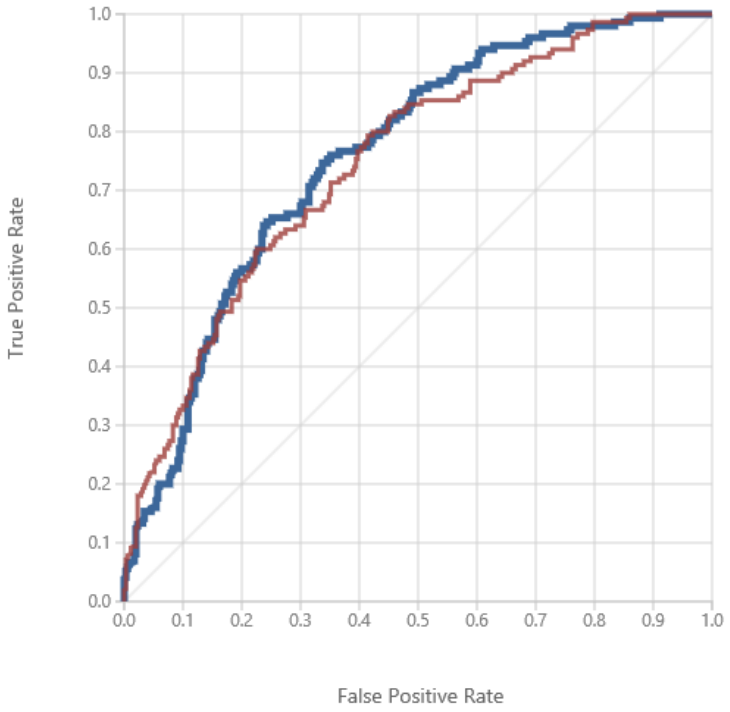
precision: $\frac{TP}{TP+FP}$ → 1 means all P are P (but increased # of FN)

recall: $\frac{TP}{TP+FN}$ → 1 means no FN (just another name for TP rate)

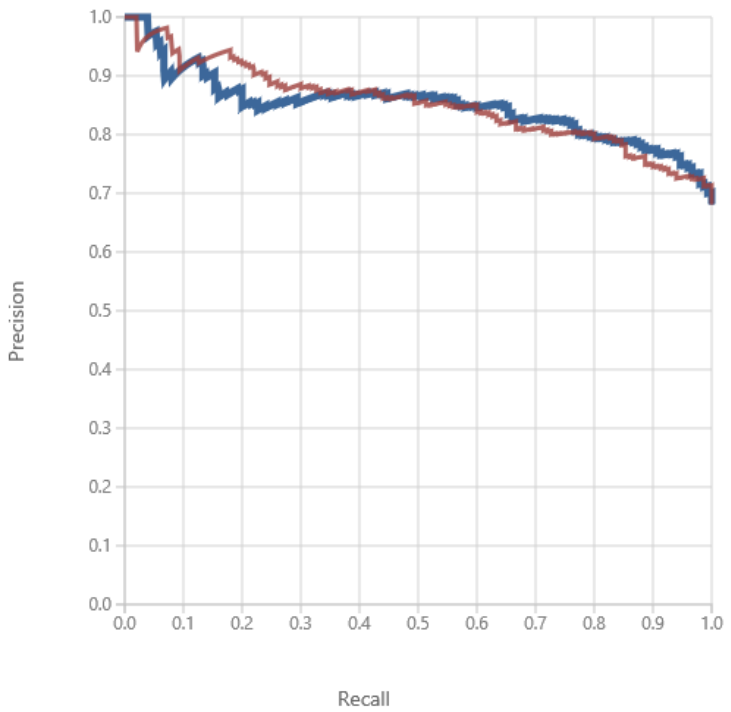
F1 score: $\frac{precision+recall}{2}$ → the closer to 1, the better

evaluate the model

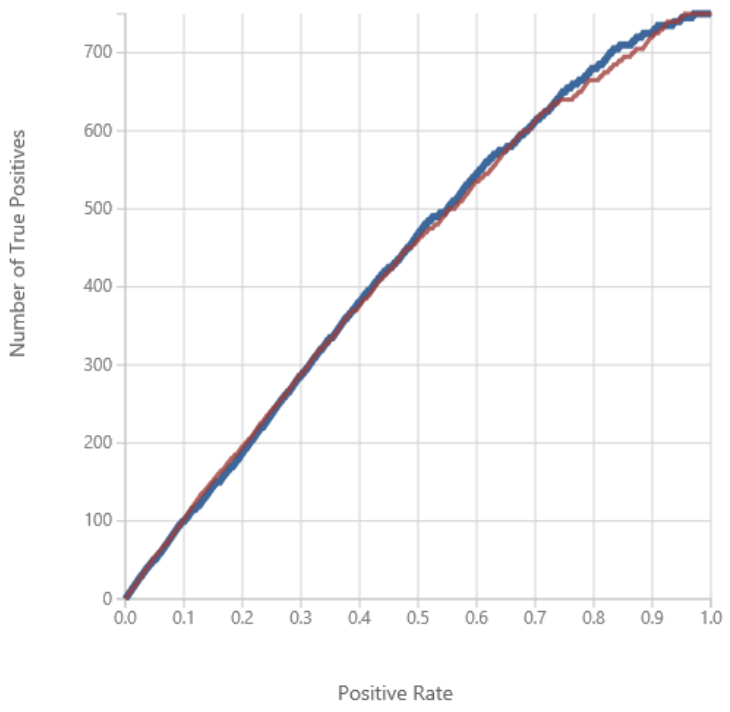
ROC (Receiver Operating Characteristic)



Precision/Recall



Lift



True Positive	False Negative	Accuracy	Precision
621	129	0.733	0.790

False Positive	True Negative	Recall	F1 Score
165	185	0.828	0.809

Threshold

0.5

AUC

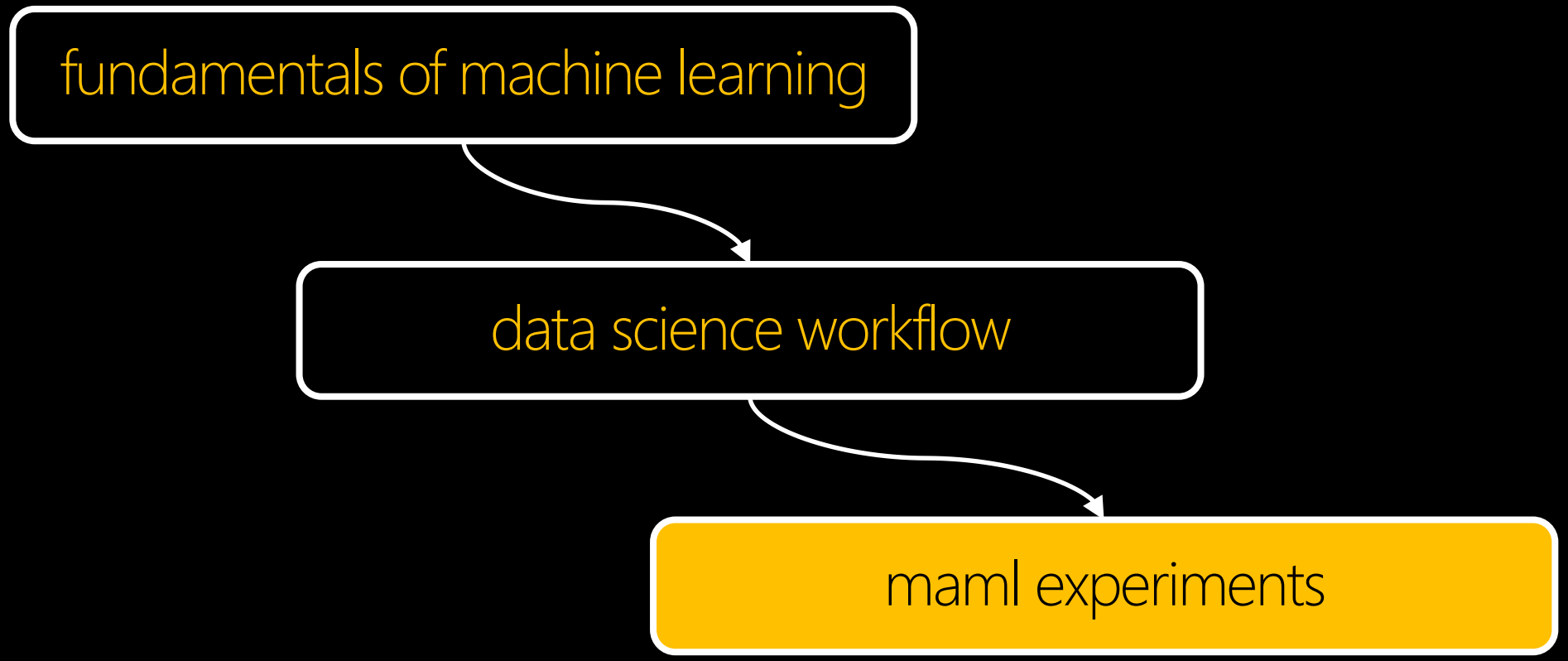
0.758

put the model in production as a Web Service

```
{
    "Id": "wineTypeScore",
    "Instance":
     {
         "FeatureVector":
         {
             "pH": "3.5",
             "alcohol": "10.7"
         },
         "GlobalParameters": {}
     }
}
```

predict

Scored Labels	Scored Probabilities
white	0.999882
red	0.242678
white	0.533091
white	0.971664
white	0.971664
white	0.533091
white	0.6598
white	0.999882
red	0.242678
white	0.888038


fundamentals of machine learning
data science workflow
maml experiments

References to the wine dataset:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties.
In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Available at:

[@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016
[Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf
[bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

in closing…