A (Re-)introduction to Machine Learning

(c) CS U of Toronto

Adam A Miller

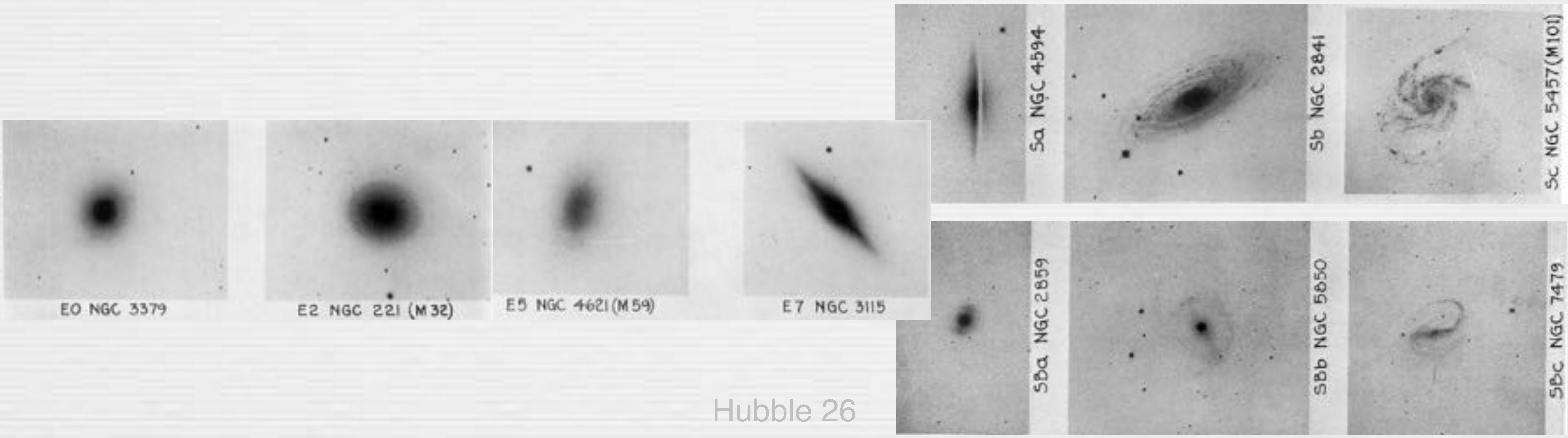
Northwestern/Adler Planetarium

2017 LSSTC DSFP
23 Jan 2016

CIERA
Center for Interdisciplinary Exploration and Research in Astrophysics

ADLER PLANETARIUM

Classification



Hubble 26

Fundamental problem for (nearly) all subfields of astronomy
- a lot of astro is essentially taxonomy

Classification schemes are (typically) well-argued, BUT
- subjective class boundaries are drawn
- constructed from small samples (then propagated forever)
- developed in low-dimensional spaces

Classification

Machine Learning

(aka - data mining, clustering, pattern recognition, AI (sorta) etc)

Fundamentally concerned with the problem of classification

- methods extend to regression as well

Address many challenges of classical taxonomy-like classification

- class boundaries drawn via (user-specified) optimization criteria
- improve and refine classifications with additional information
- can be constructed & developed in high-dimensional spaces

Examples: SPAM filters, Netflix, self-driving cars, etc

credit: SPAM

credit: Netflix

credit: Google

Classification

Machine Learning

two flavors:

labels are unknown

Unsupervised Learning

In the feature space, the number, shape, & size of data groupings is unknown

Machine aims to cluster sources

No natural metric for measuring quality
i.e. results vary from algorithm to algorithm

Can be very useful for data exploration

labels are partially known

(labels are never fully known…)

Supervised Learning

Portion of data labeled by experts or expensive follow-up

Machine maps features ➤ labels

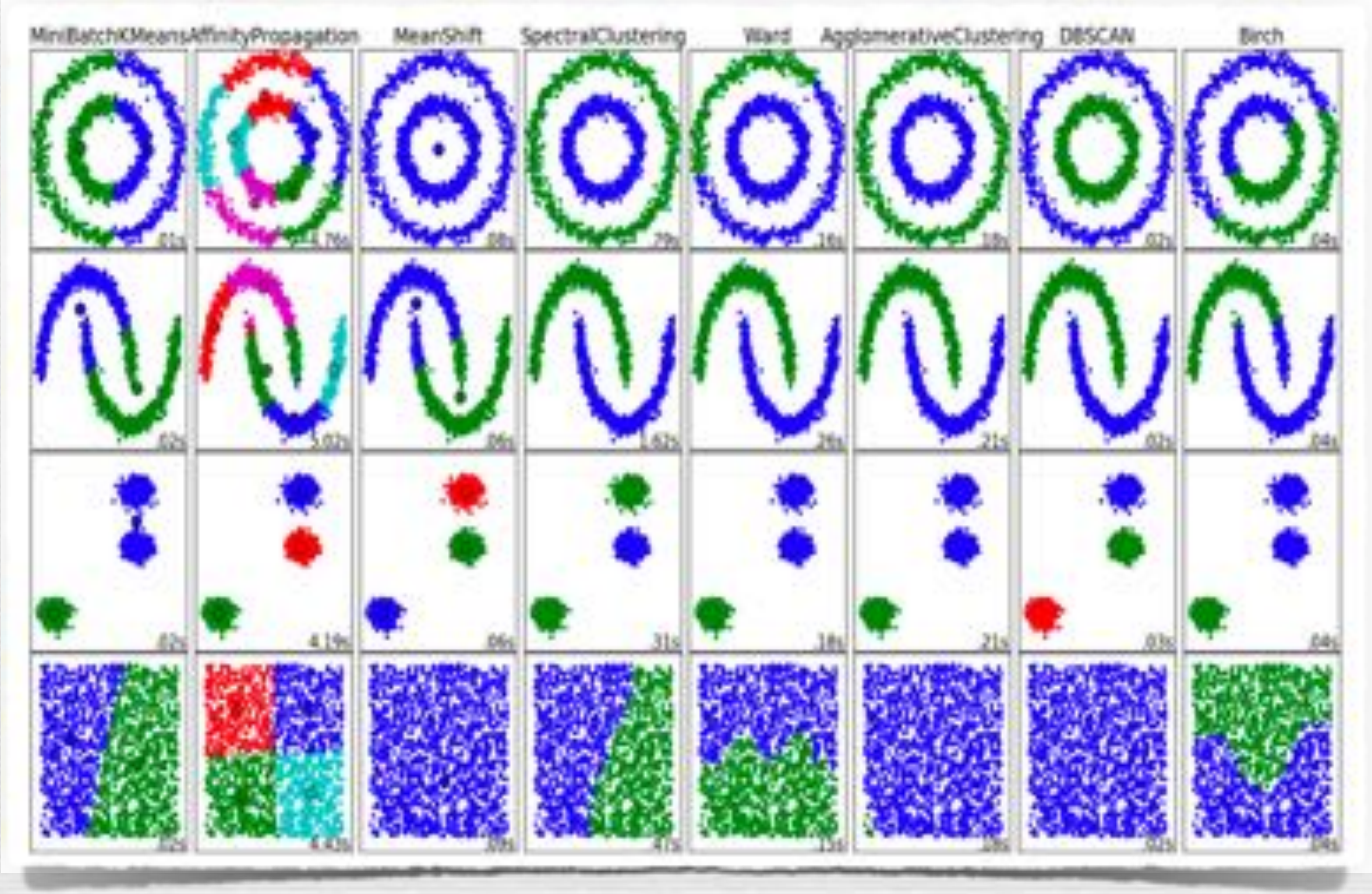
Can optimize accuracy or MSE
results still vary from algorithm to algorithm

Useful for classification & regression

Classification

Machine Learning

Unsupervised

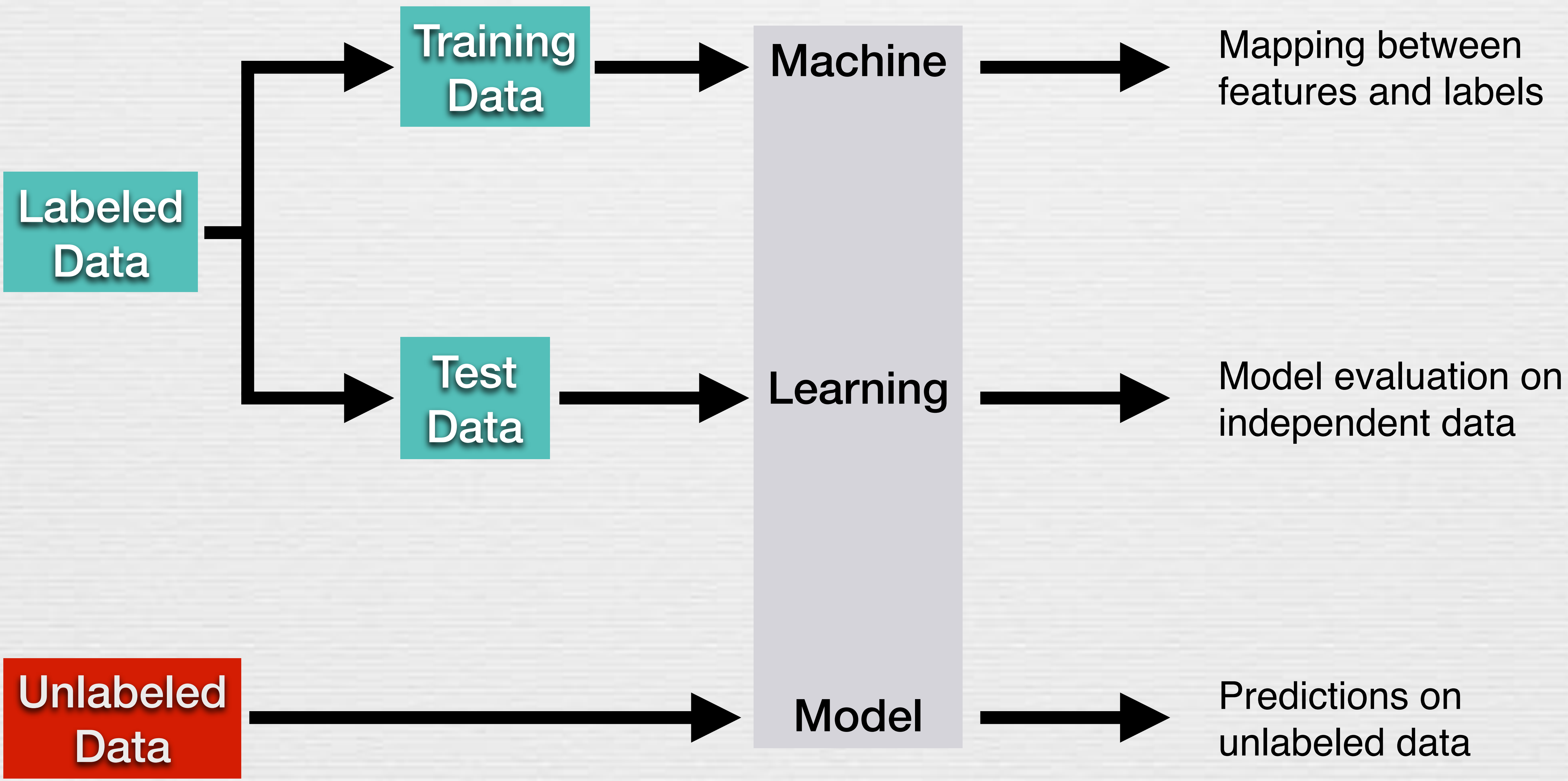


credit: scikit-learn

Classification

Machine Learning

Supervised



sklearn Makes ML "Easy"

4 lines to construct a complex model

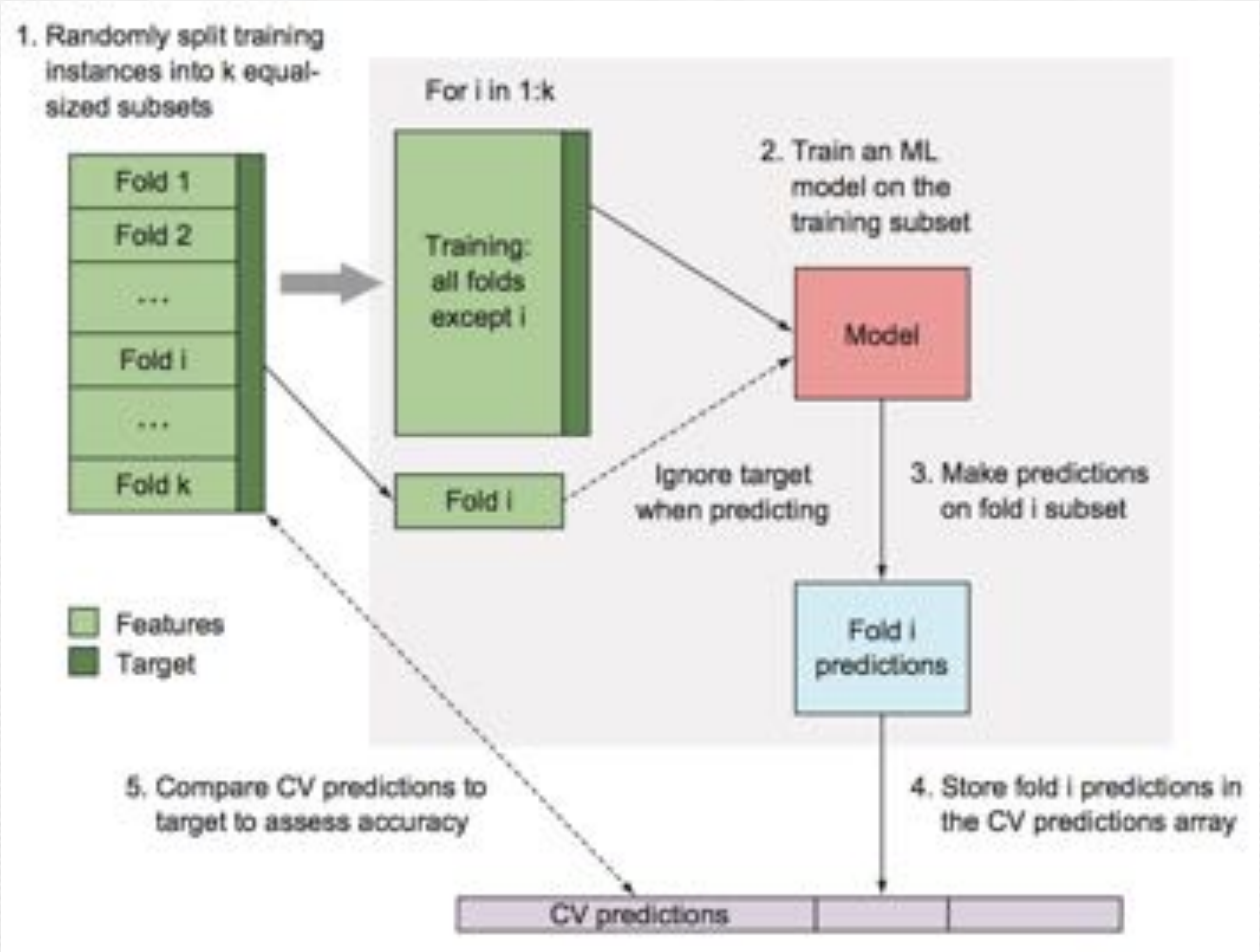
```
from sklearn import datasets
from sklearn.ensemble import RandomForestClassifier
iris = datasets.load_iris()
RFclf = RandomForestClassifier().fit(iris.data, iris.target)
```

sklearn is great,

but be weary of too good to be true

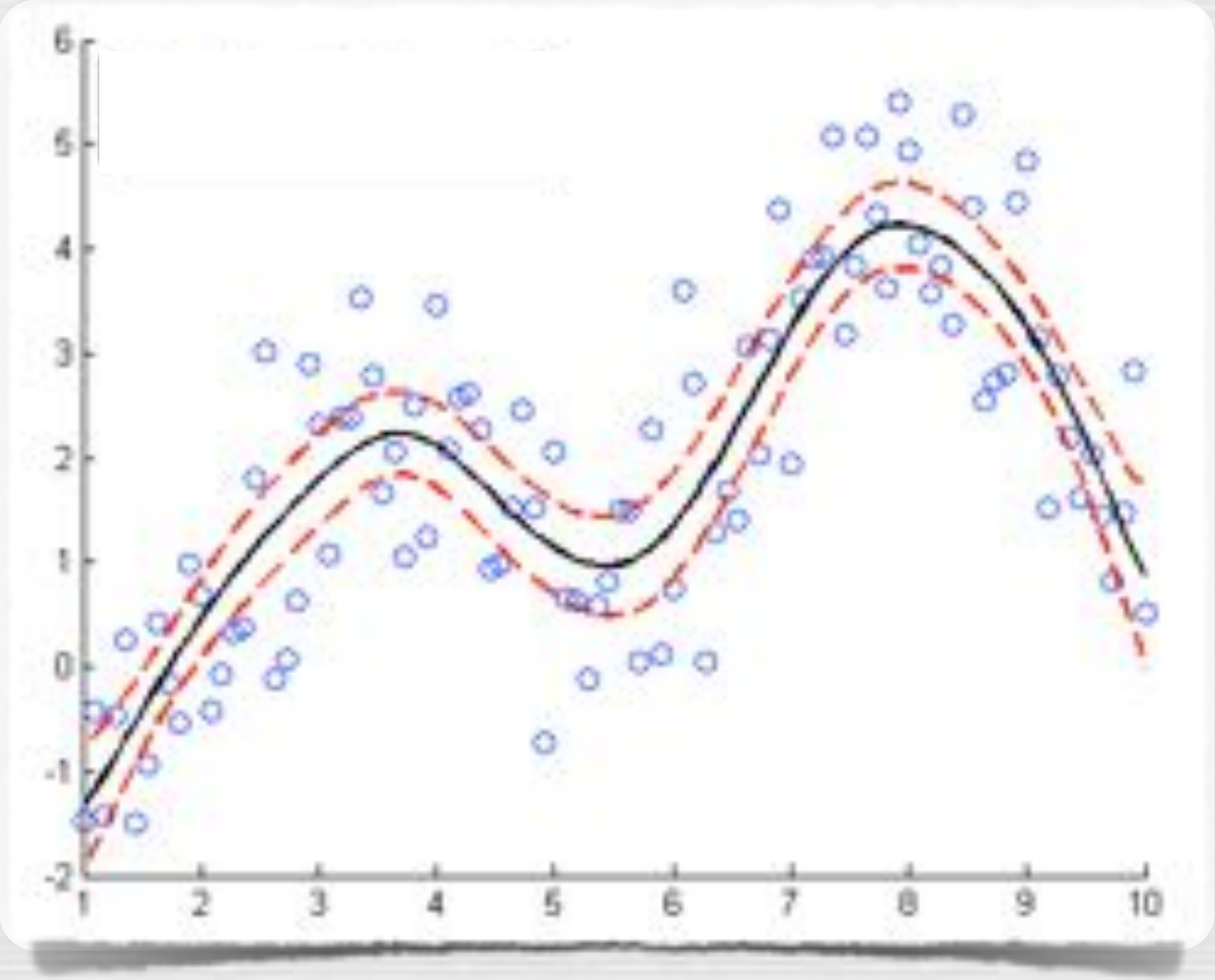
Concepts Worth "Stealing" From ML

- Evaluate algorithms with independent test sets



Concepts Worth “Stealing” From ML

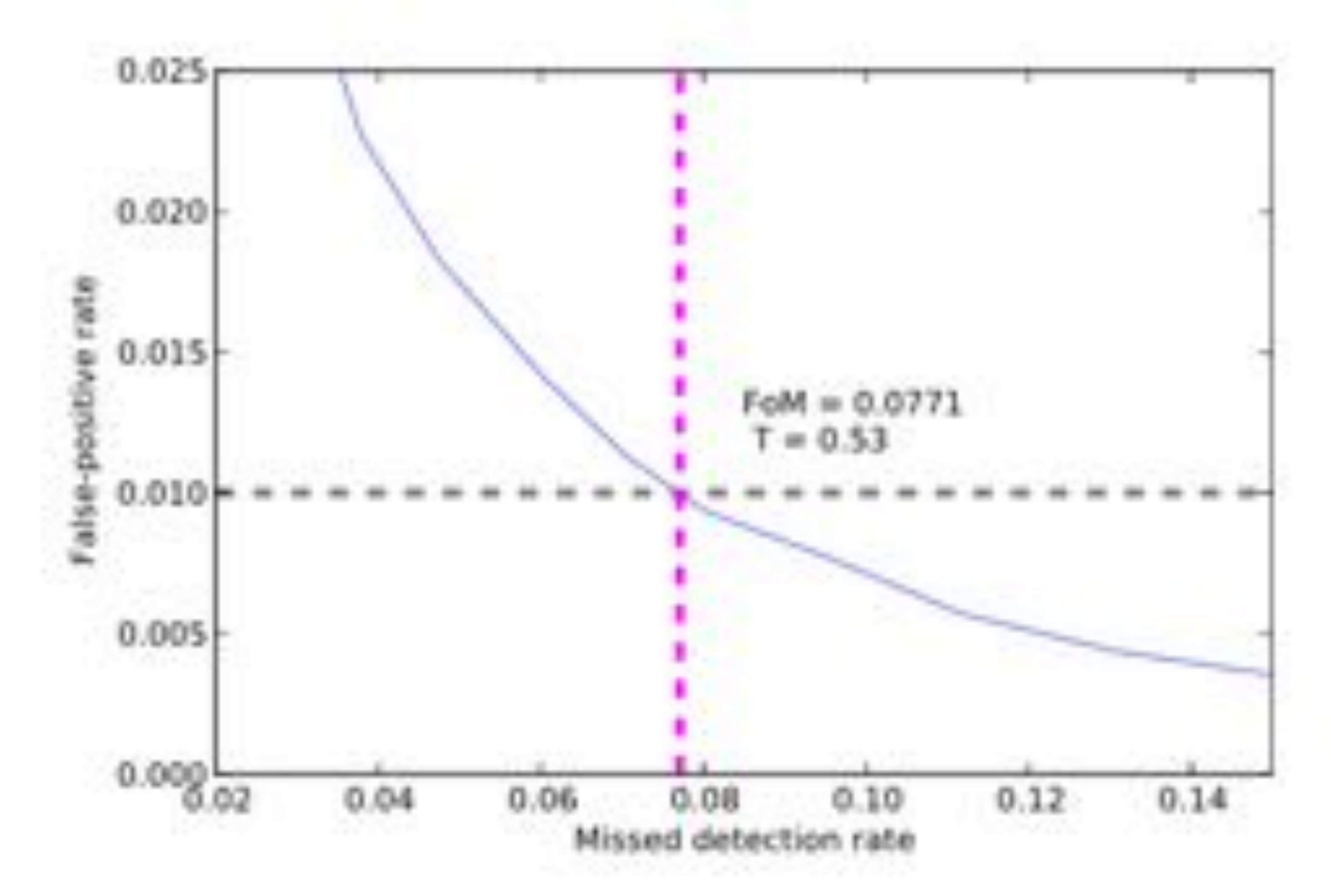
- Evaluate algorithms with independent test sets
- Embrace flexibility, allow data to drive models



credit: blogs.mathworks.com

Concepts Worth "Stealing" From ML

- Evaluate algorithms with independent test sets
- Embrace flexibility, allow data to drive models
- Set decision boundaries to optimize desired outcome



Brink+13

ML Model Selection

Brief review/introduction of terminology

True Positive (TP)	+ classified as +
False Positive (FP)	— classified as +
True Negative (TN)	— classified as —
False Negative (FN)	+ classified as —

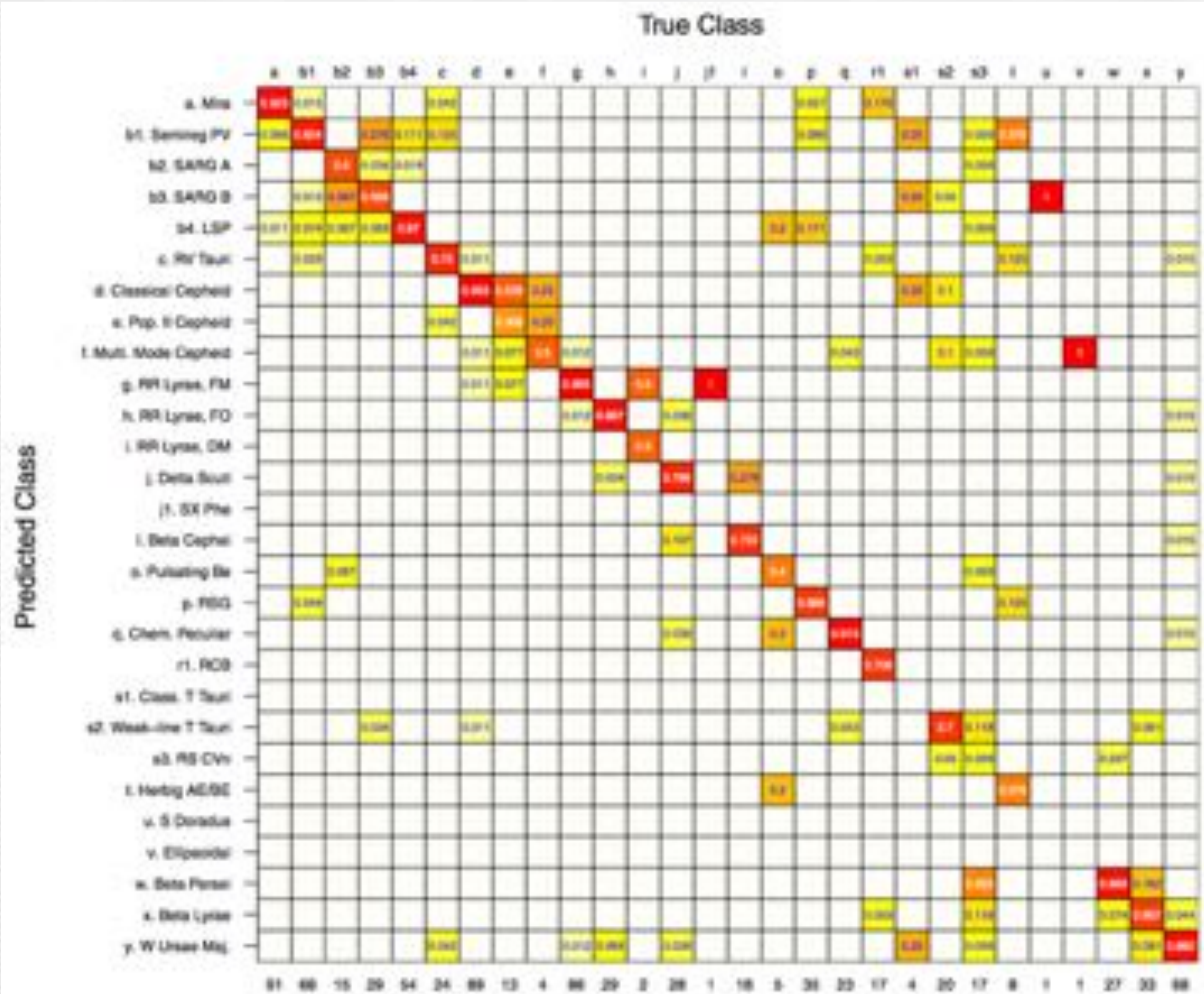
ML Model Selection

Confusion Matrix

		Predicted Class	
		+	—
True Class	+	TP	FN
	—	FP	TN

ML Model Selection

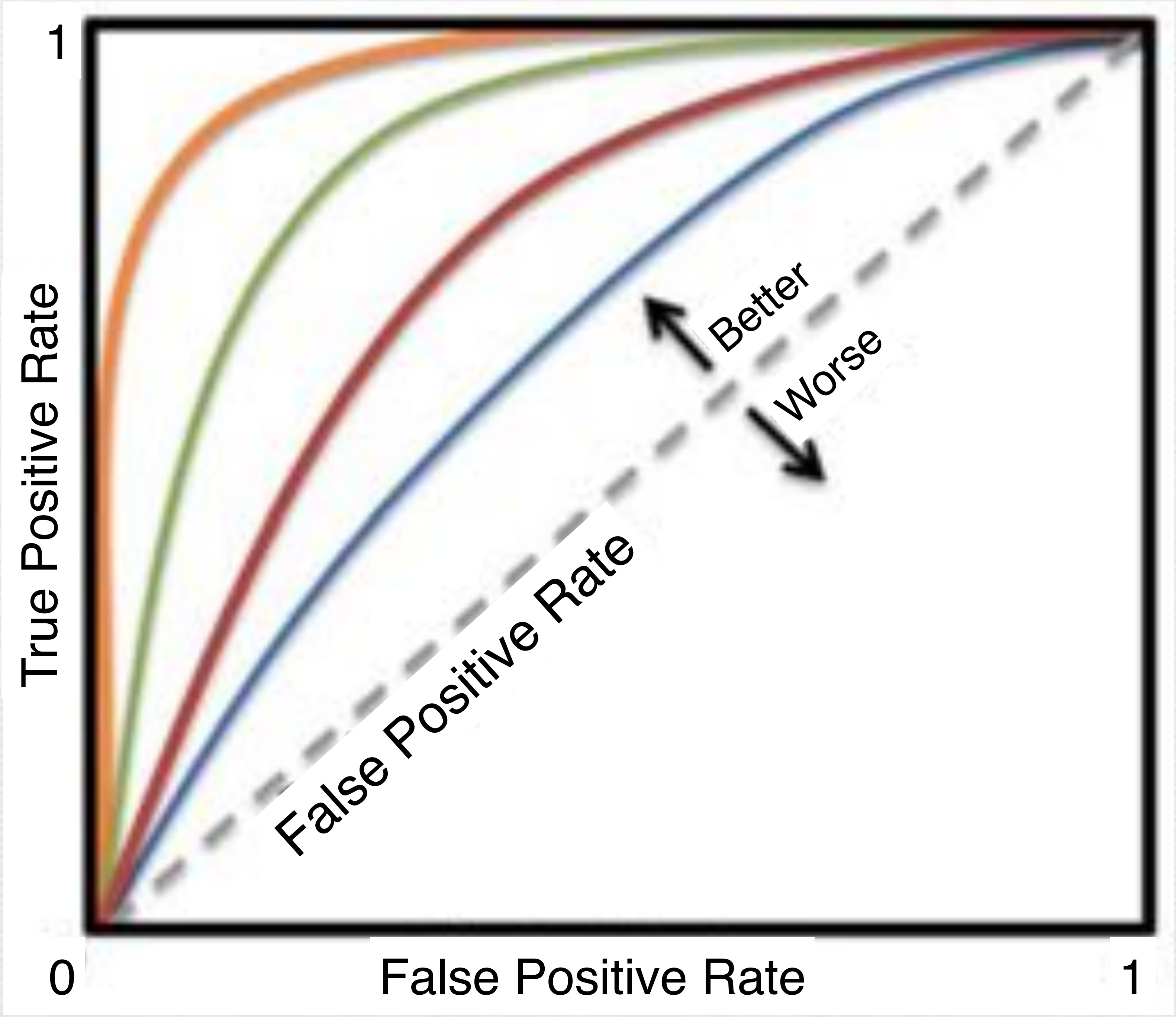
Confusion Matrix



Richards+12

ML Model Selection

True Positive Rate (TPR)	TP / (TP + FN)
False Positive Rate (FPR)	FP / (TN + FP)

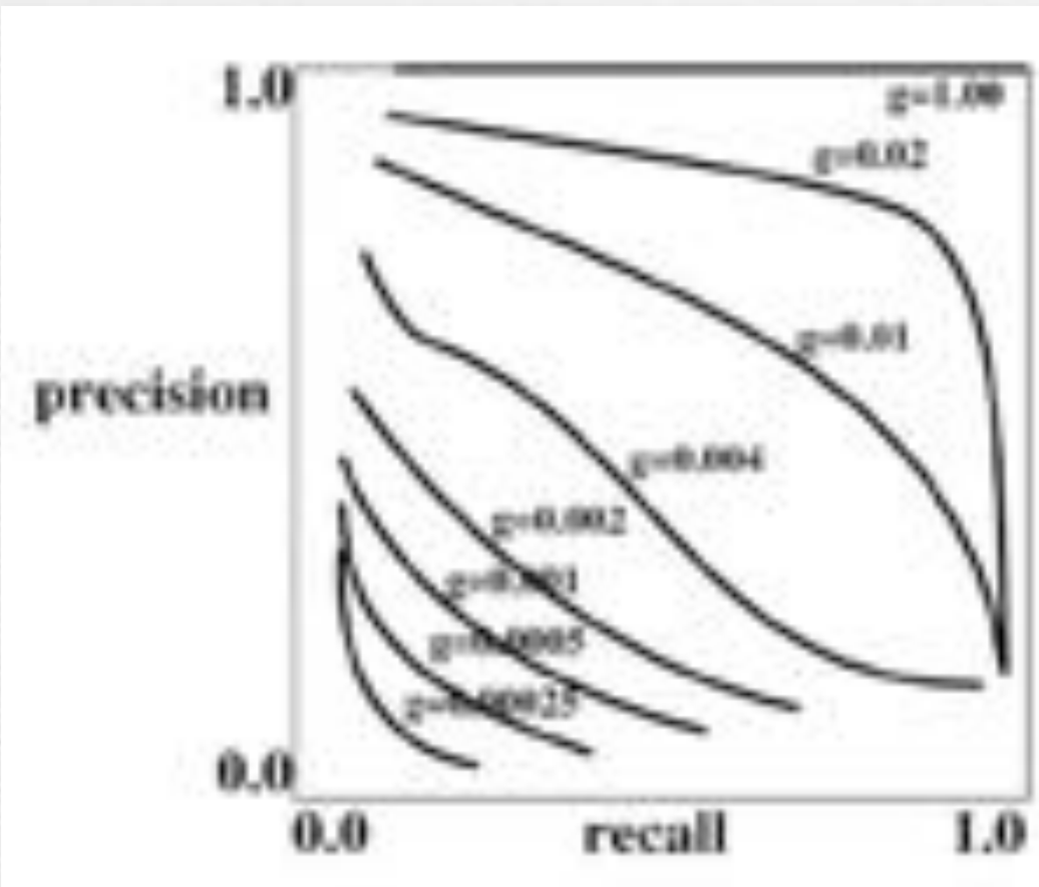


ROC Curve

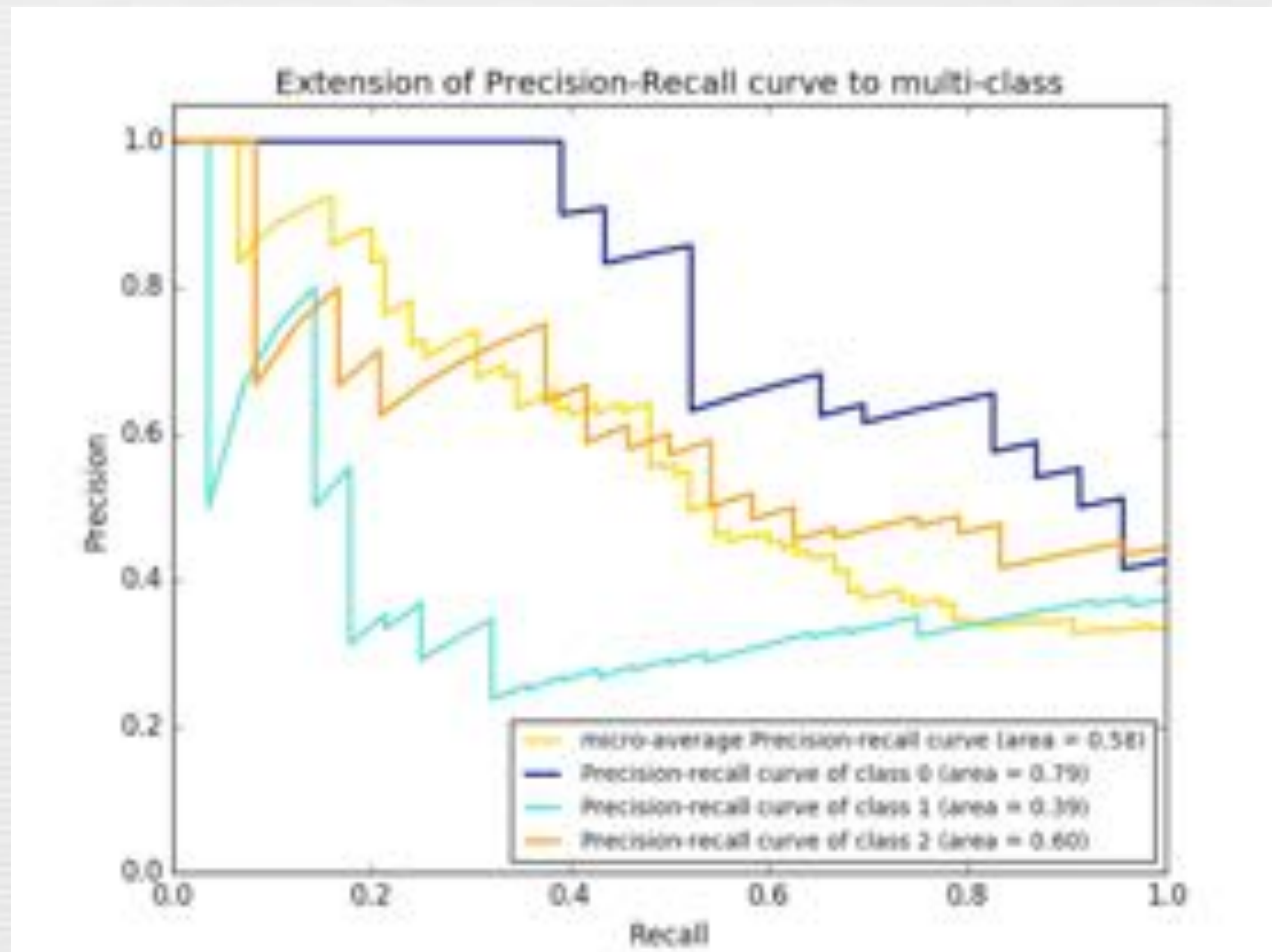
Zahiri+13

ML Model Selection

Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F_1	2 * (P * R)/(P + R)



Huijsmans & Sebe 05



credit: sklearn

Conclusions

Data-driven solutions are a necessity for wide-field surveys

- ML is particularly useful for engineering solutions
- e.g. real-bogus for transients

Off-the-shelf ML algorithms are rarely plug+play for astro

- nasty systematics (heteroskedastic errors & targeting bias)
- e.g., small calibration errors in SDSS for EMP discovery
- e.g., SDSS LRG bias for star-galaxy separation

Principles (sometimes algorithms) of ML are very useful

- when data leads theory, allow data to drive the models
- test the utility of everything with independent observations
- make informed thresholding decisions
- e.g., The Cannon - measuring ages for >10k giants