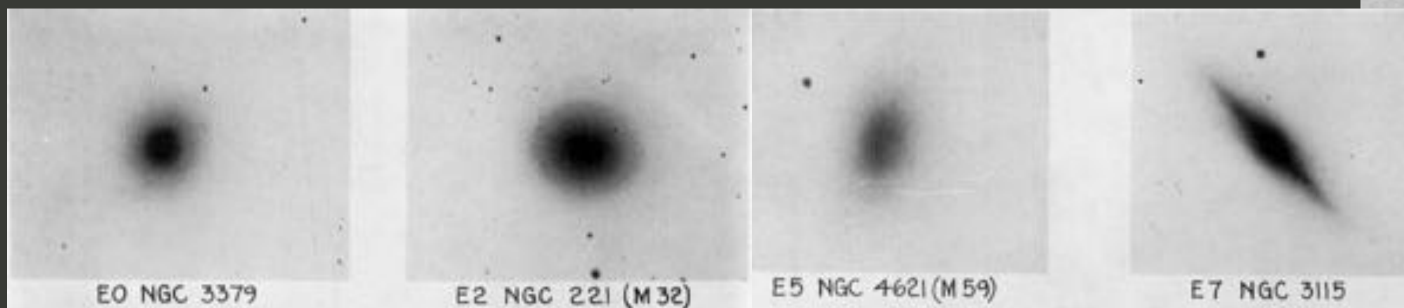
An Introduction to Machine Learning

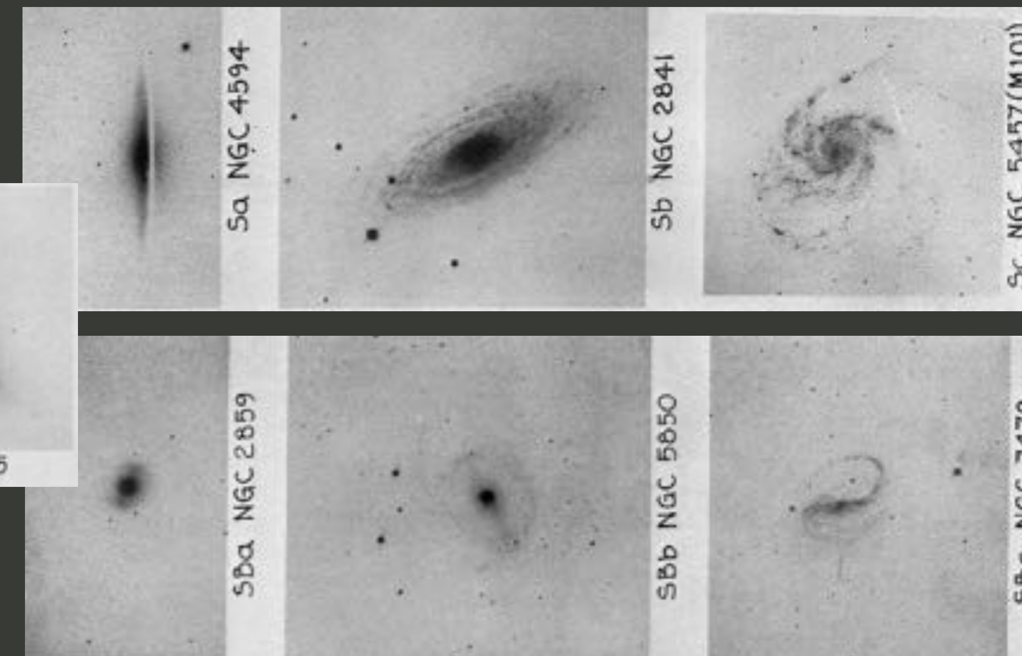
(c) CS U of Toronto

Adam A Miller
CIERA/Adler Planetarium

LSSTC DSFP Session 4
17 Sep 2017

Classification





Hubble 26

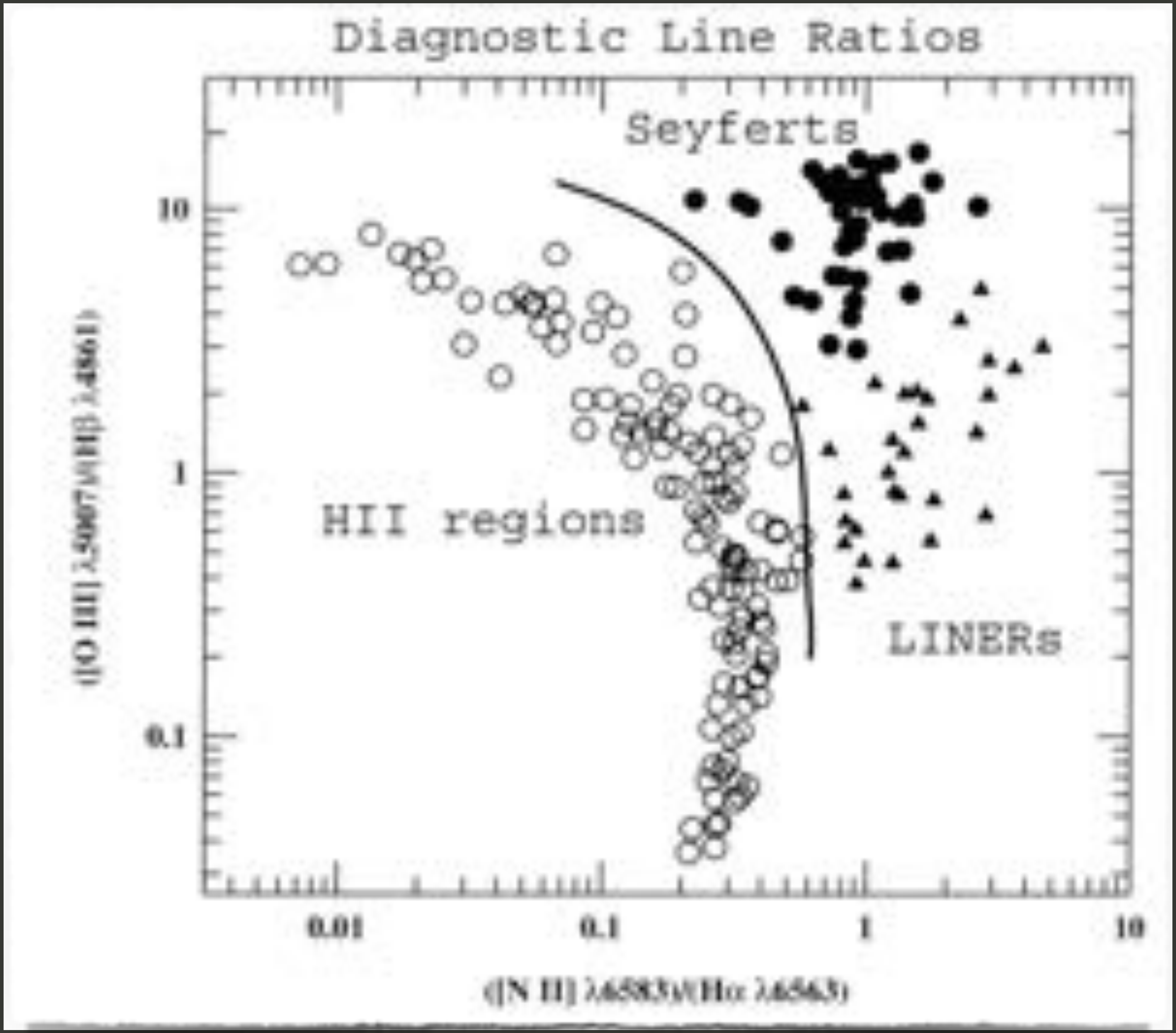
Fundamental problem for (nearly) all subfields of astronomy
- a lot of astro is essentially taxonomy

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

Example - BPT diagram   Baldwin+81

Classification

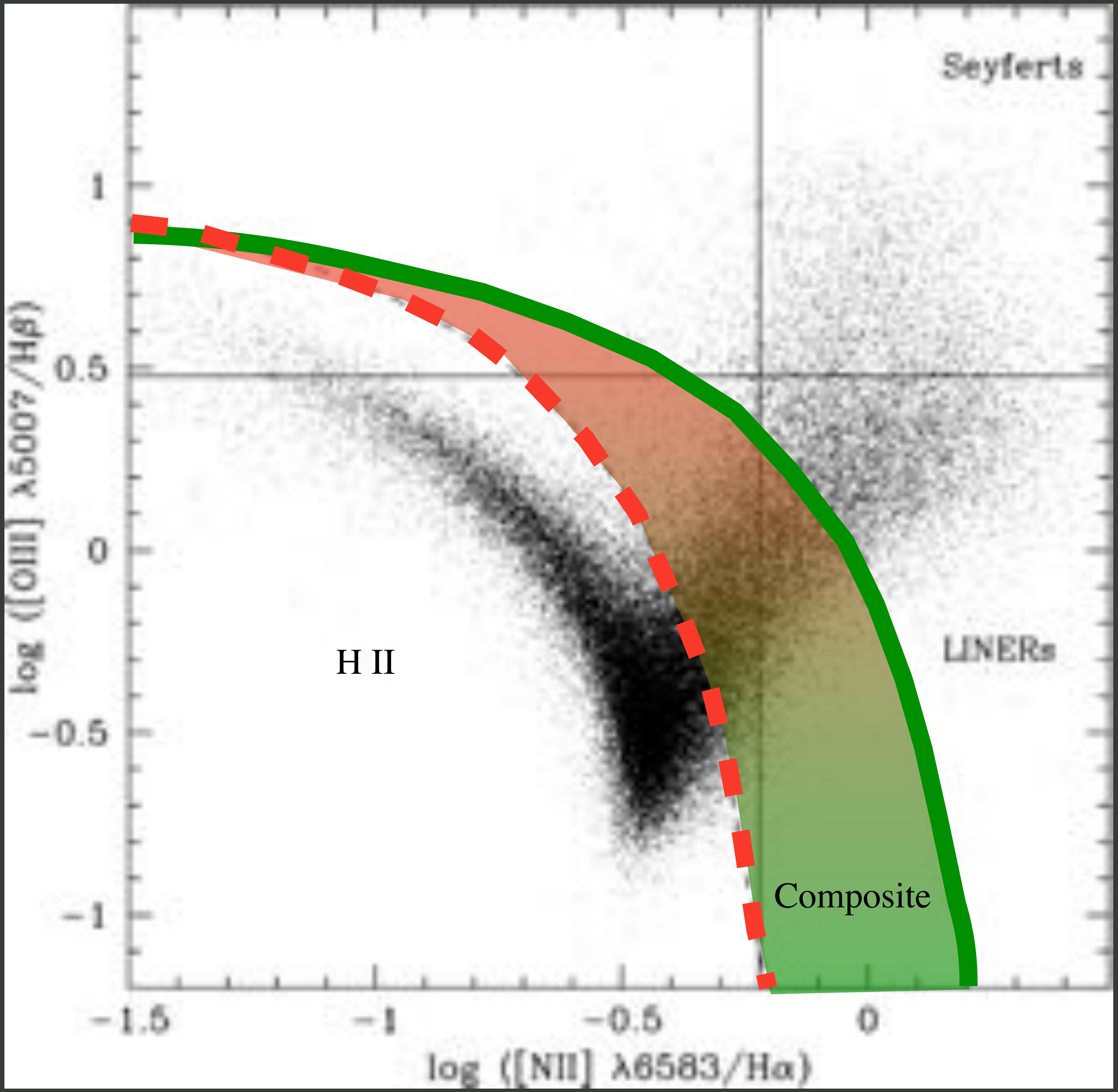
BPT circa 1987



credit: Mark Whittle

Classification

BPT circa SDSS



continuous distribution

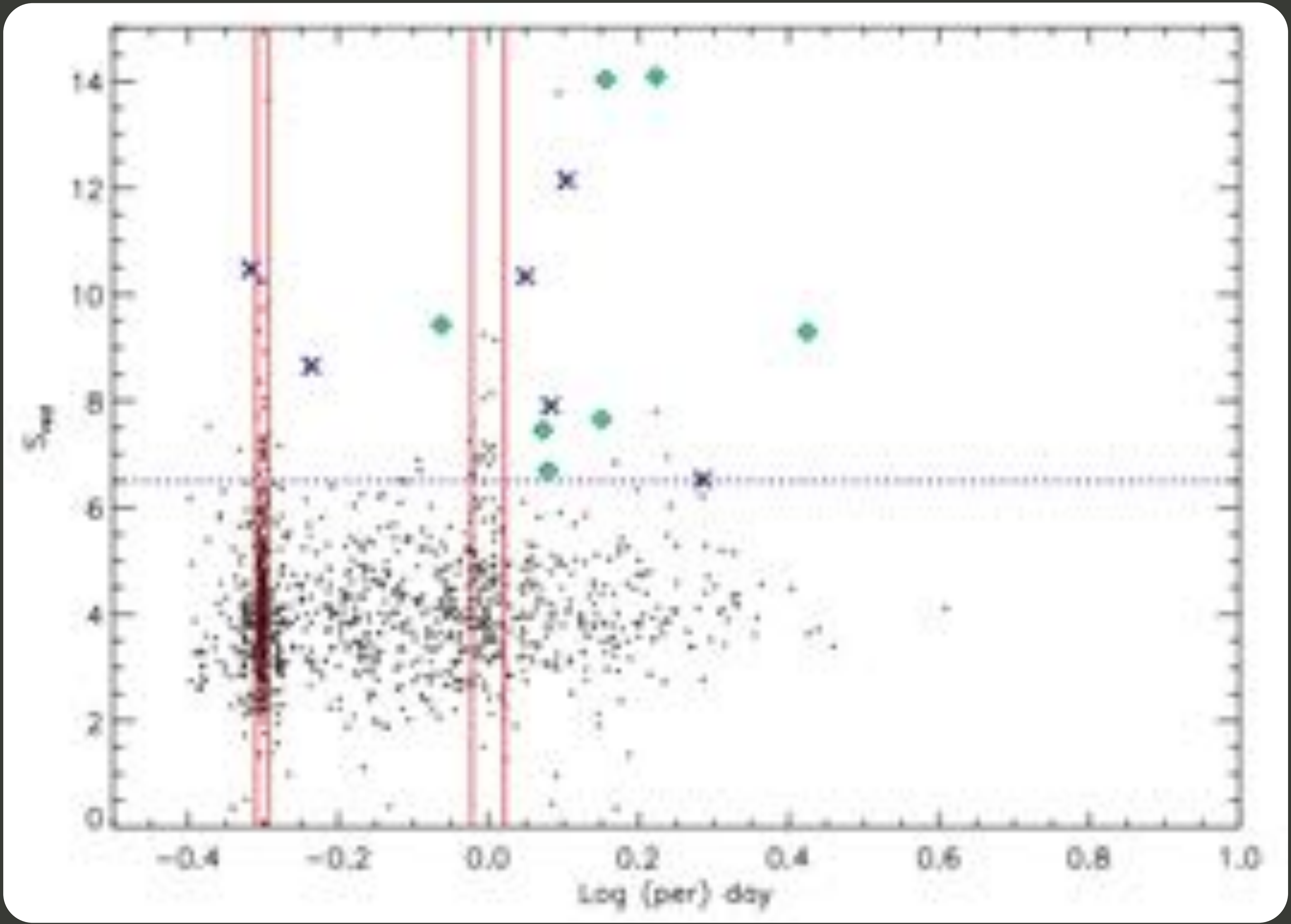
different class bounds

new (ill-defined) classes

Kauffmann+03

Classification

I’m guilty too



Miller+08

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

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

Classification

Machine Learning

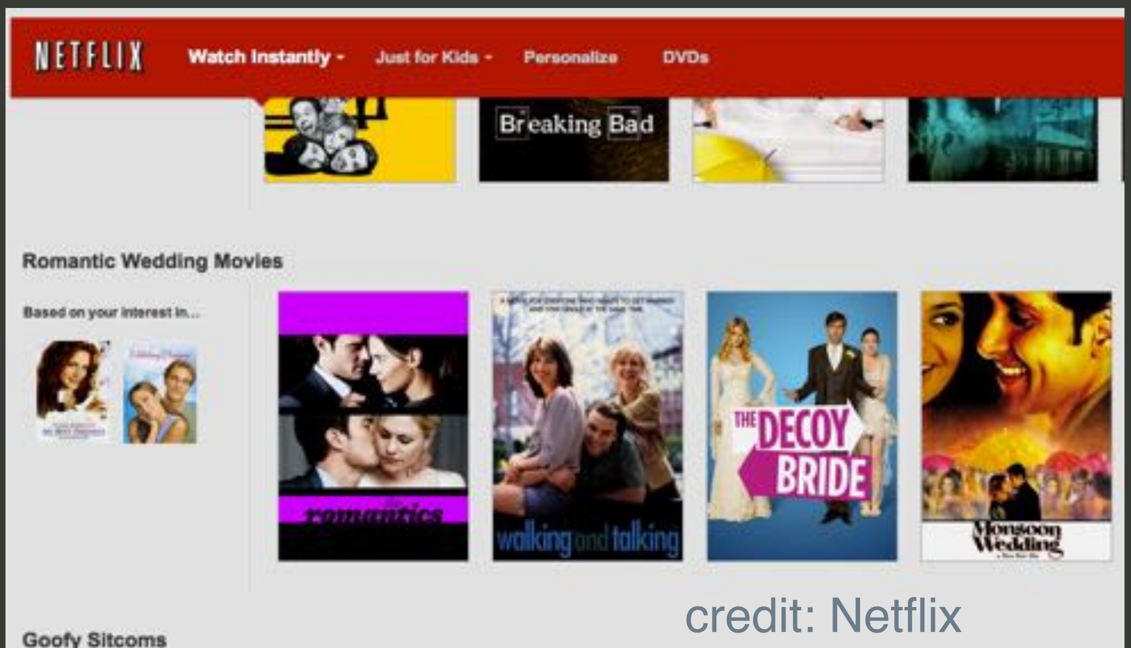
(aka - data mining, clustering, pattern recognition, AI (sorta) etc)

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

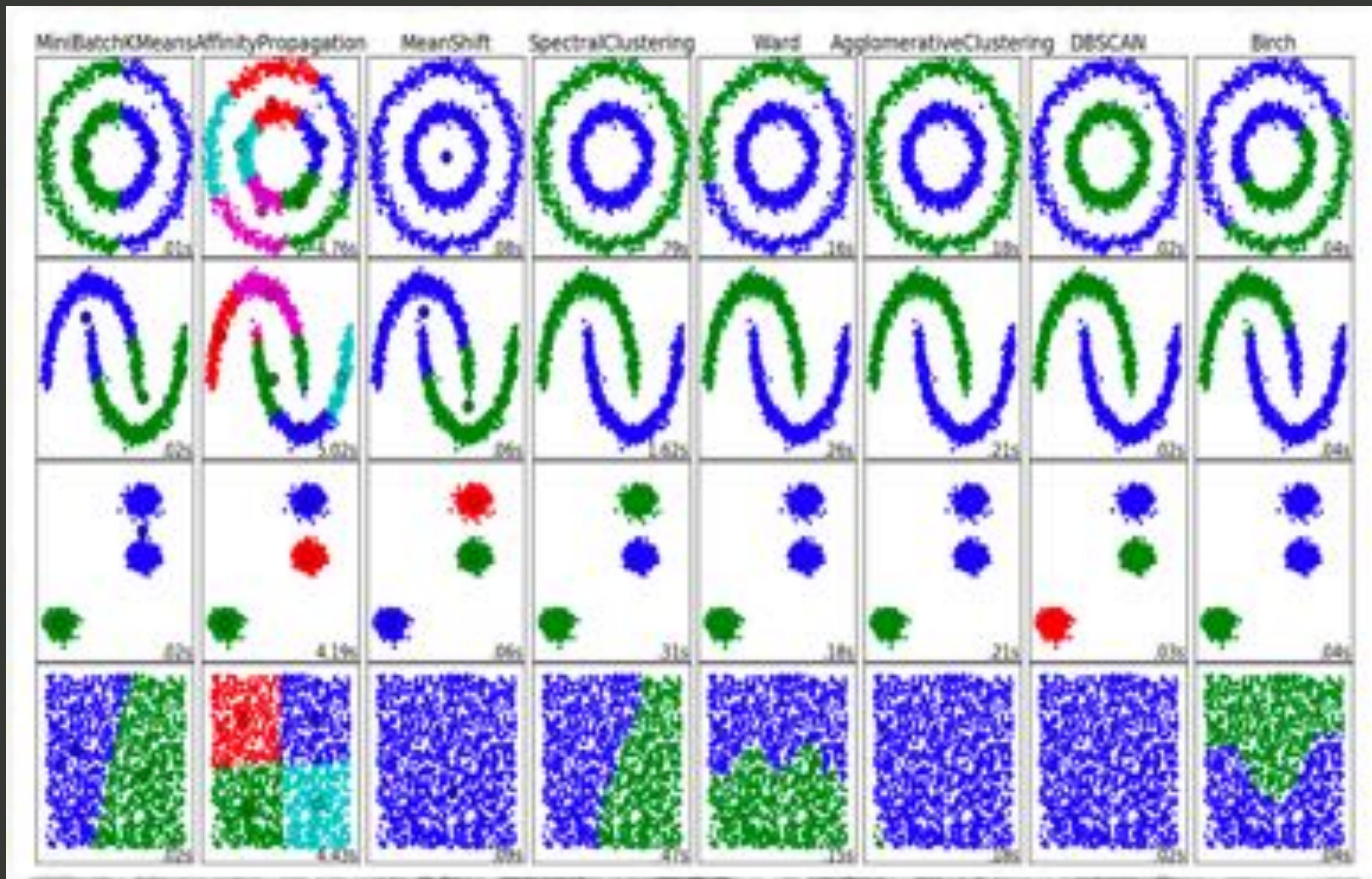
Can optimize accuracy or MSE
results still vary from algorithm to algorithm

Useful for classification & regression

Classification

Machine Learning

Unsupervised



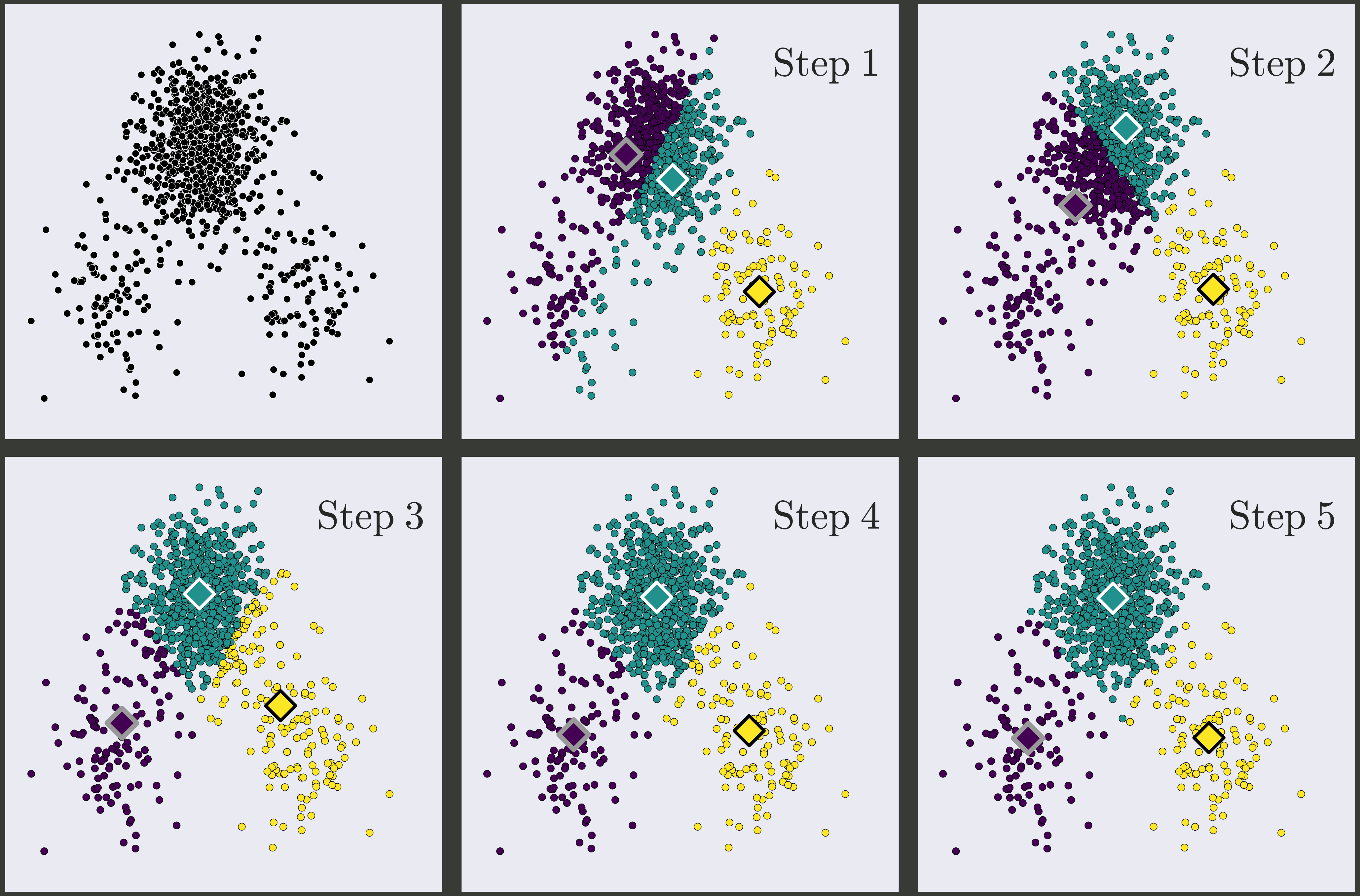
credit: scikit-learn

Classification

Machine Learning

Unsupervised

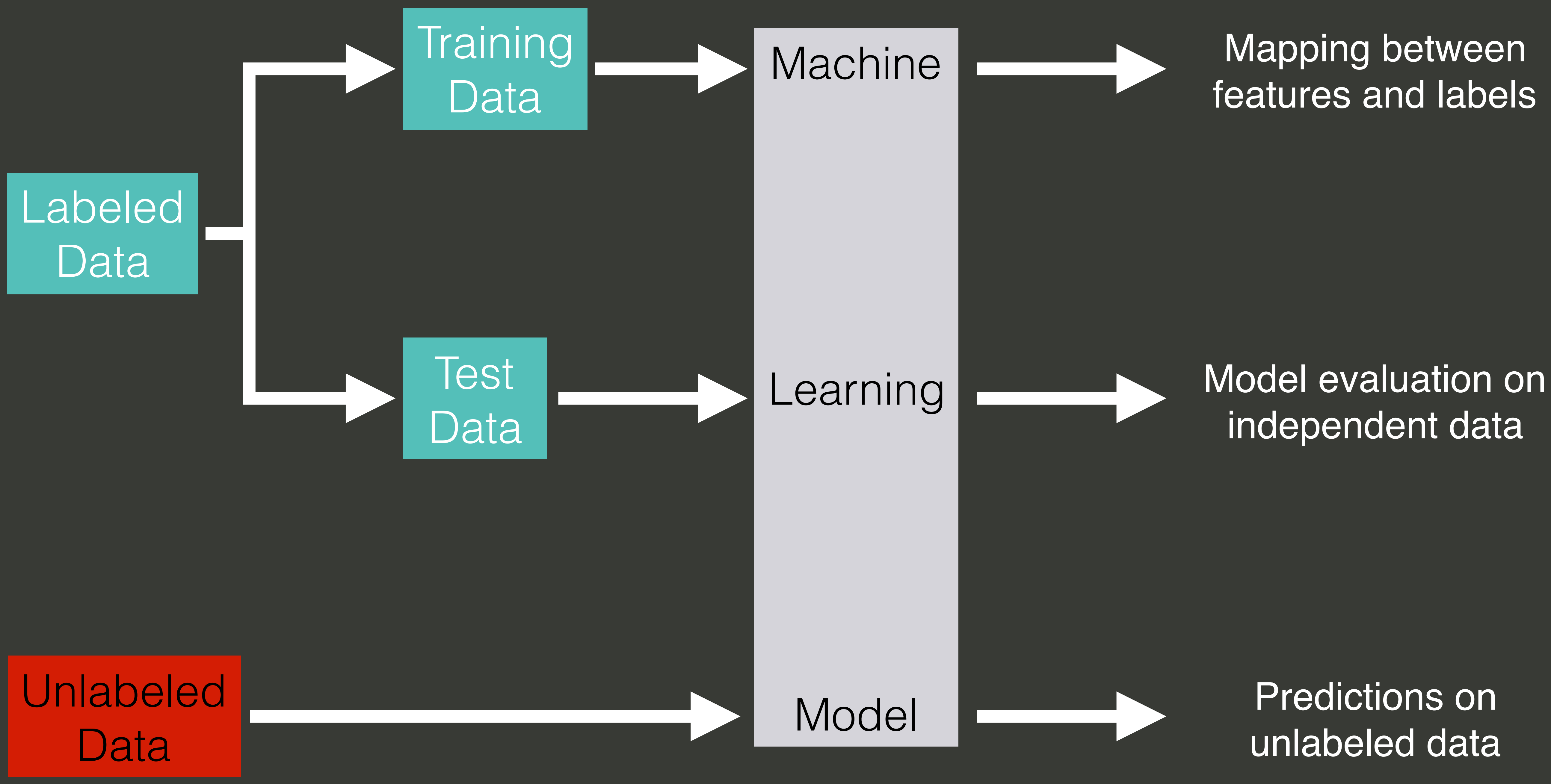
Famous algorithm: **K-means**



Classification

Machine Learning

Supervised

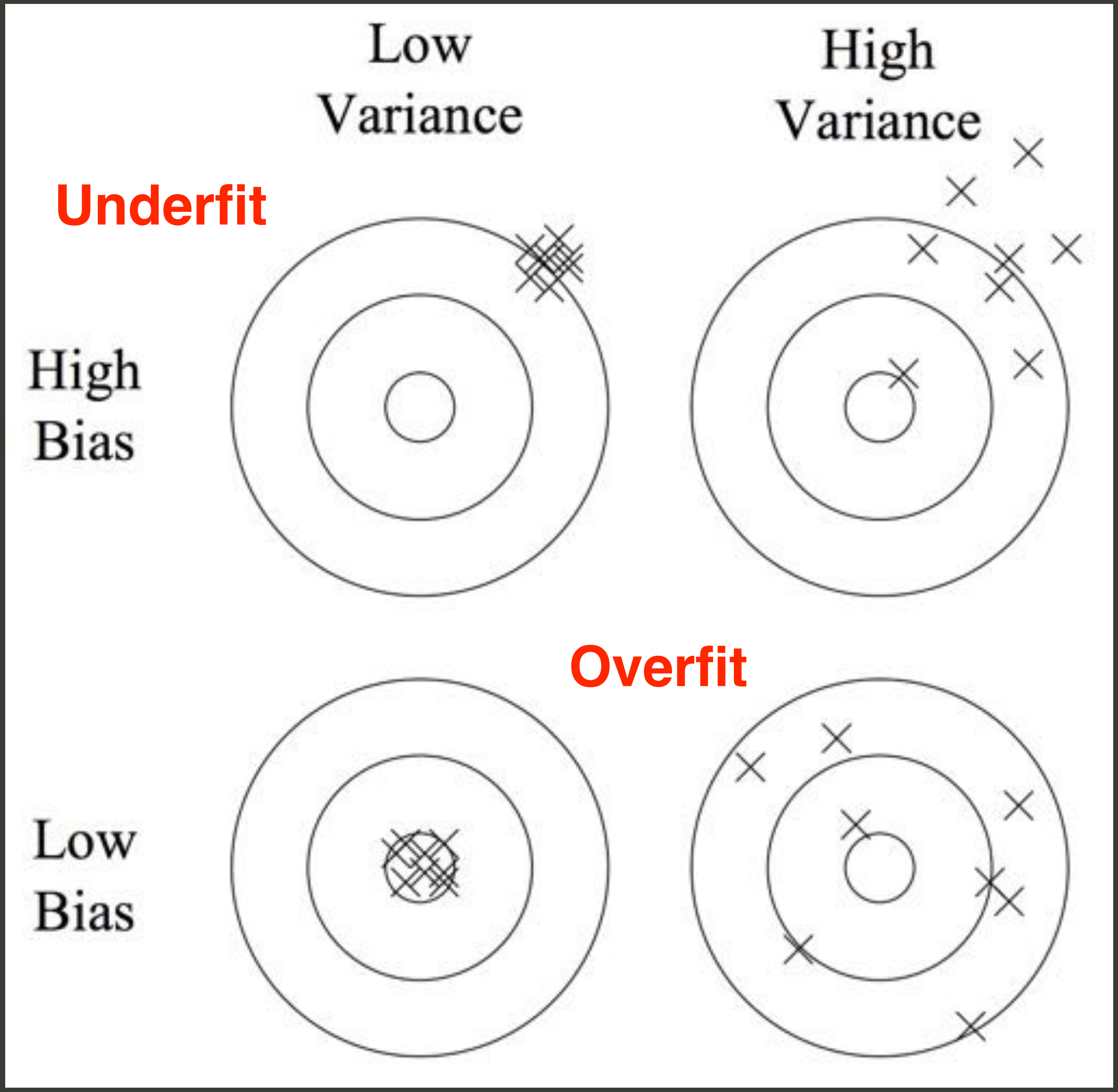


Classification

Machine Learning

Supervised

Goal: optimal trade off between bias and variance



credit: Arjun Krishnan

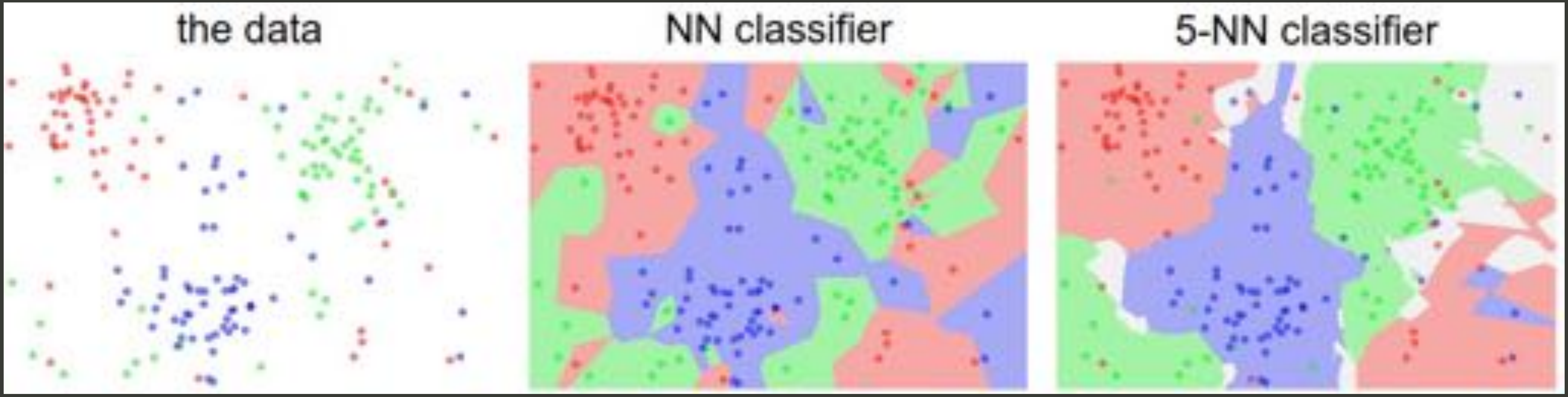
Classification

Machine Learning

Supervised

Famous algorithm: ***k*-nearest neighbors**

User specifies k ➤ k closest training set sources determine final classification



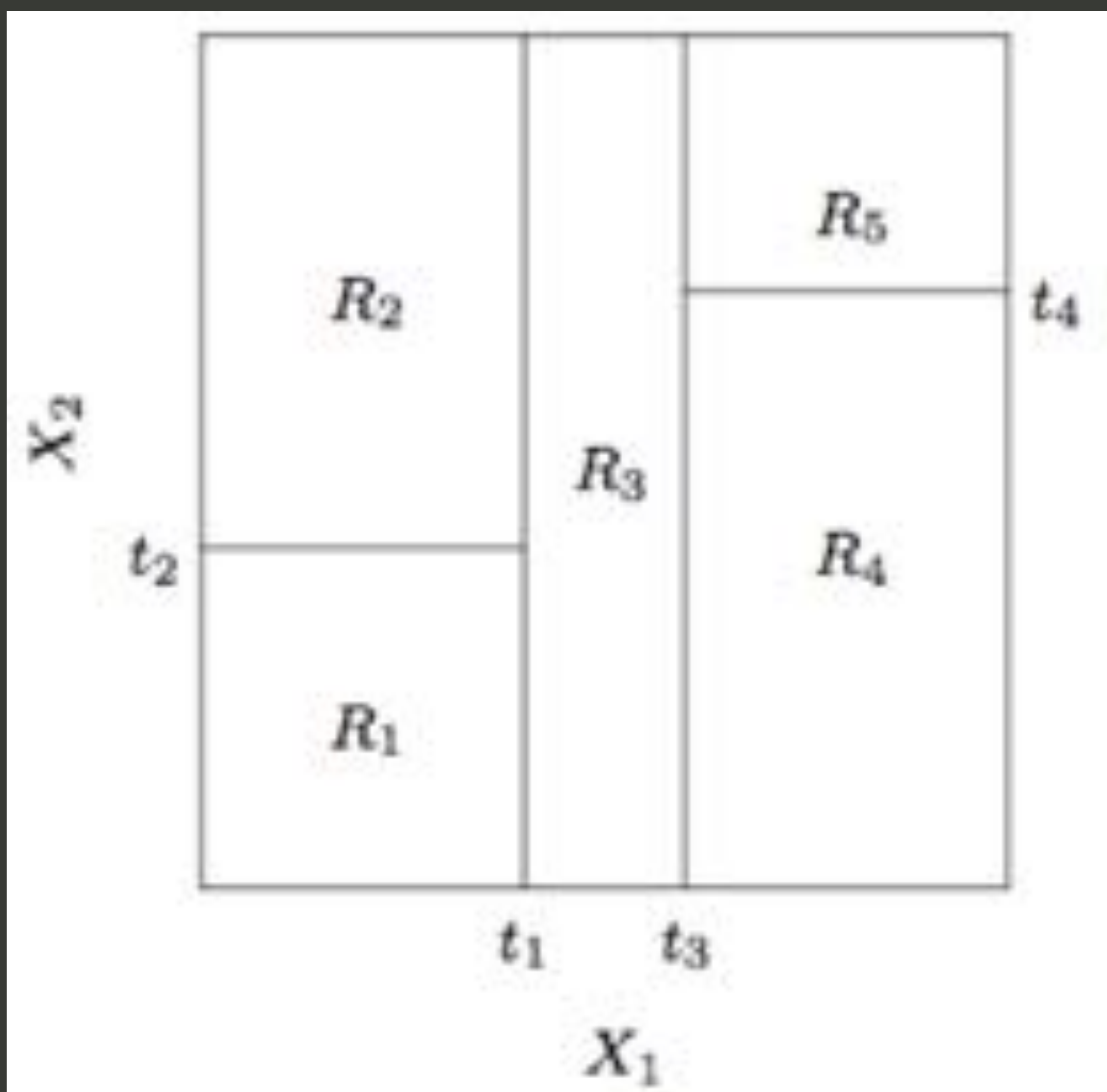
credit: http://cs231n.github.io/classification/

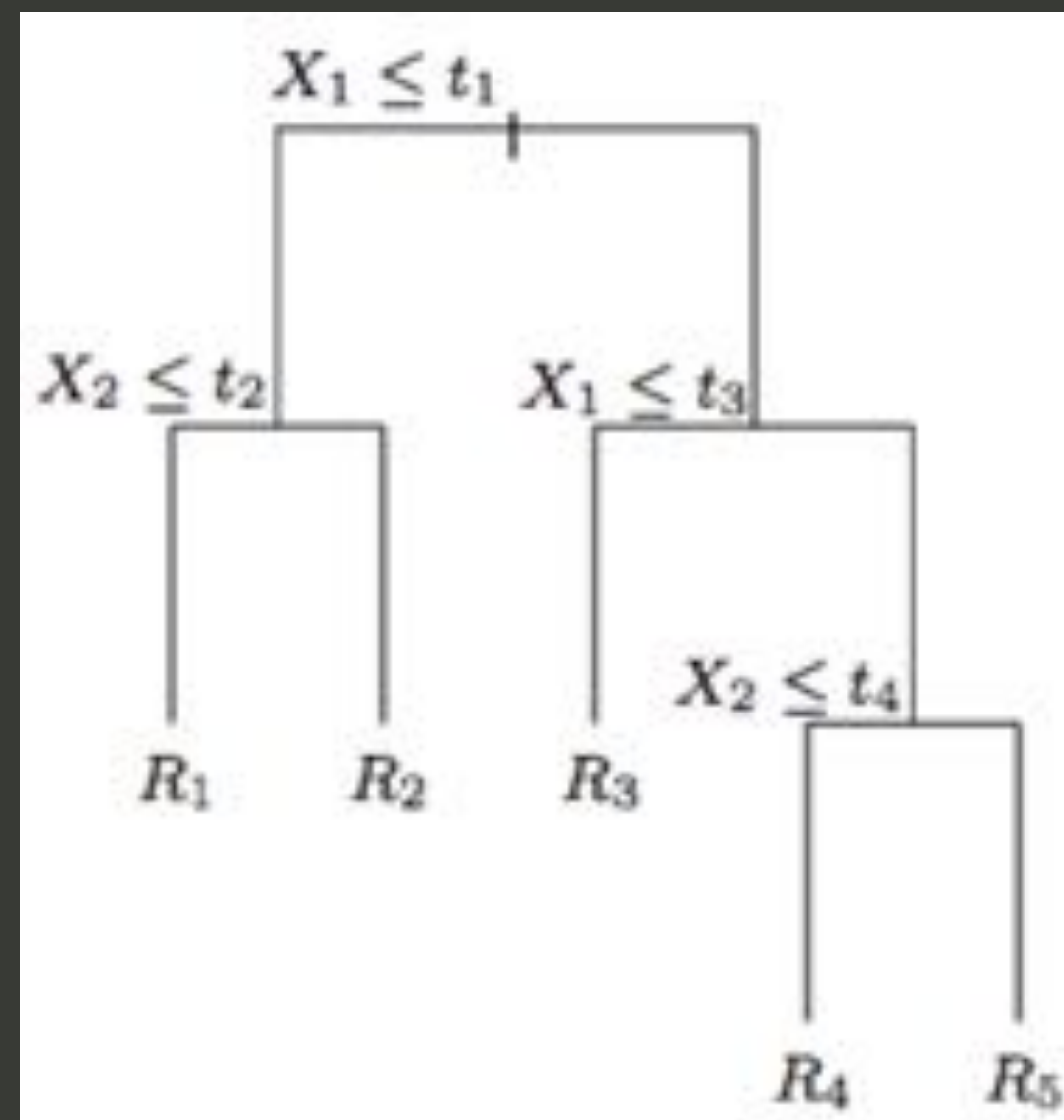
Classification

Machine Learning

Supervised

Famous algorithm: **Decision Tree**





Hastie, Tibshirani, Friedman 09

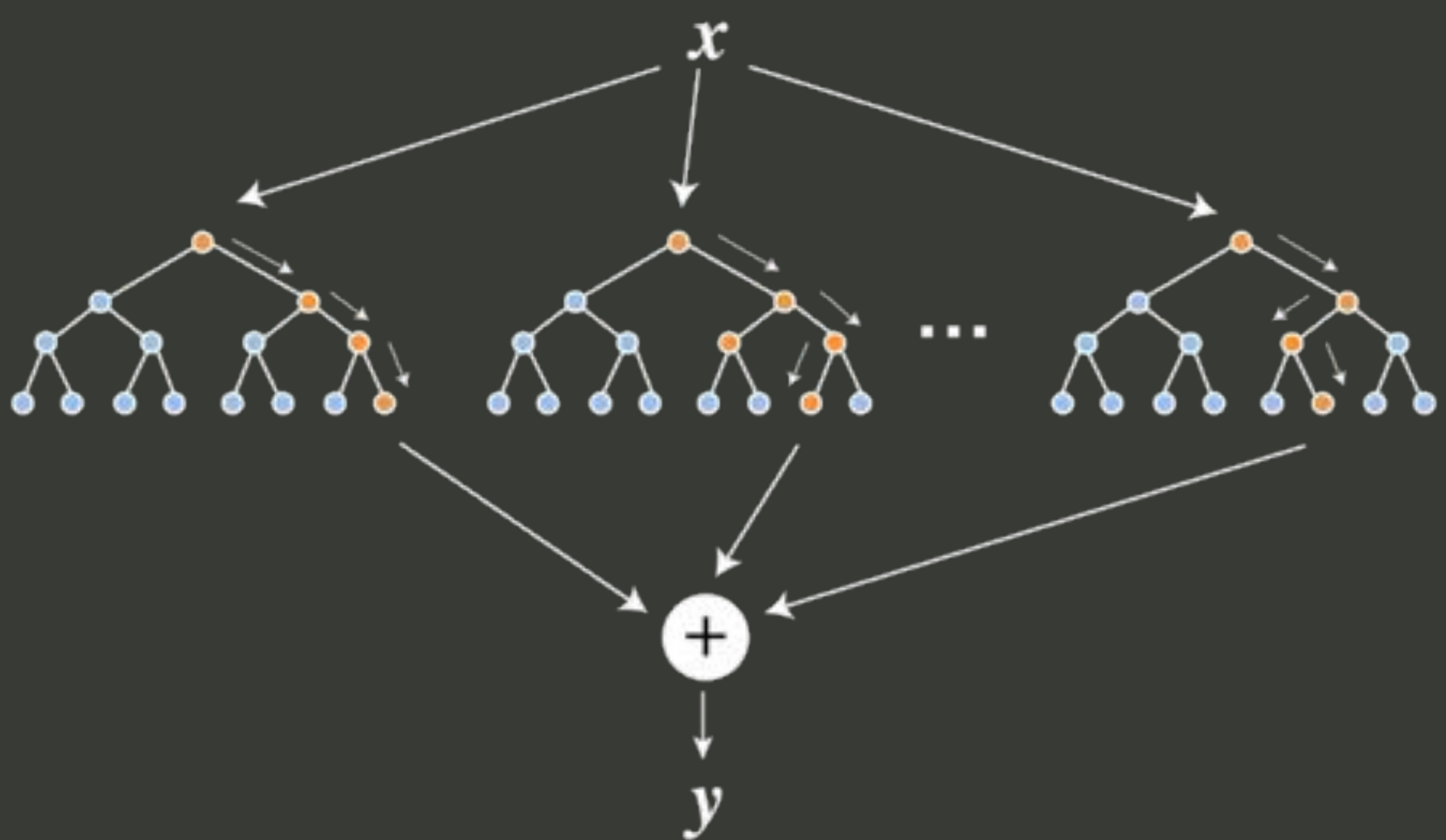
Classification

Machine Learning

Supervised

Famous algorithm: **Random Forest**

Aggregates results from a collection of multiple decision trees
Use bagging (bootstrap w/ replacement) for each tree
Select only a random subset of features for split at each node
Average of de-correlated trees reduces variance relative to single tree



credit: http://kazoo04.hatenablog.com/entry/2013/12/04/175402

sklearn Makes ML "Easy"

4 lines to construct a complex model

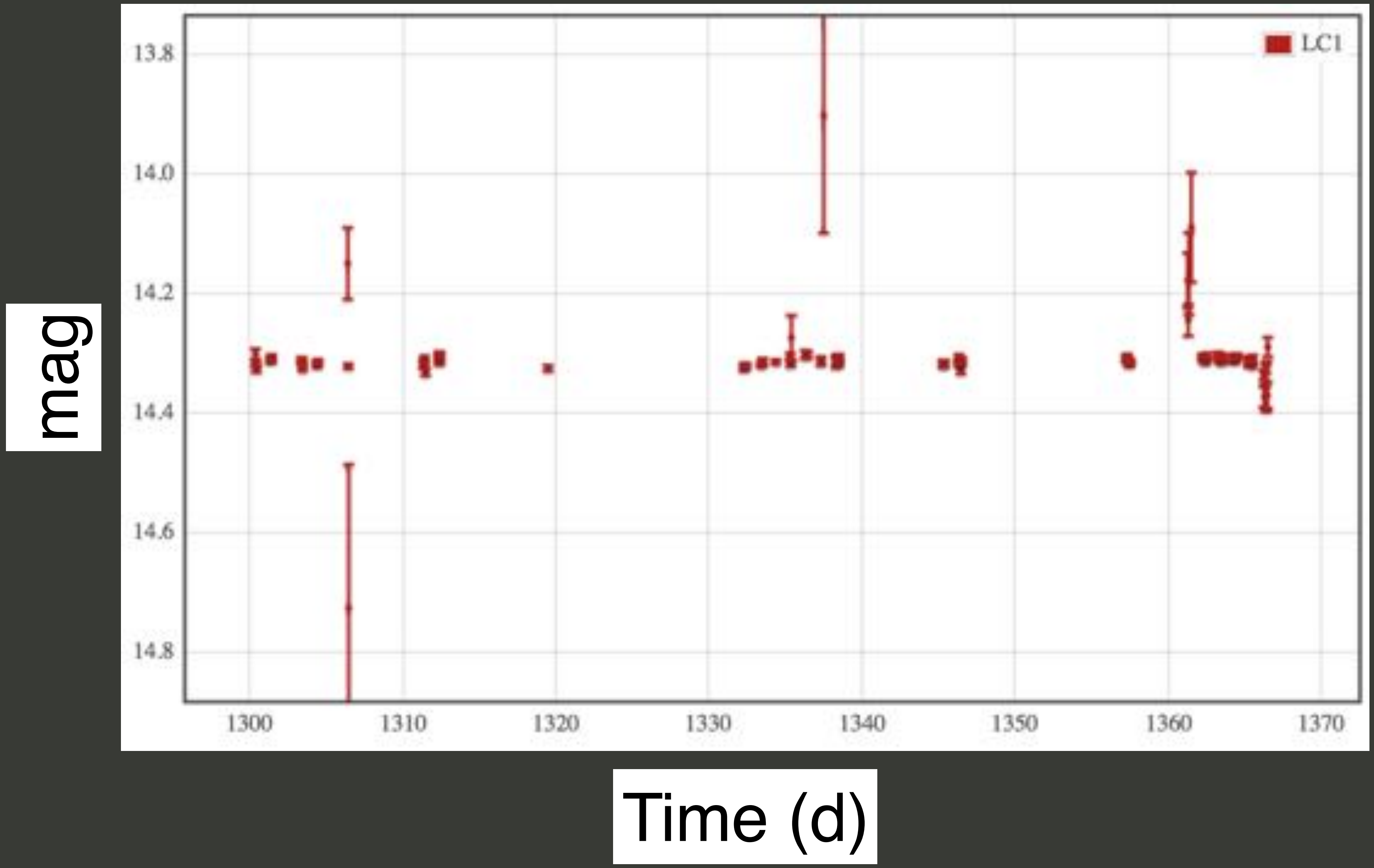
```
from sklearn import datasets
from sklearn.ensemble import RandomForestClassifier
iris = datasets.load_iris()
RFclf = RandomForestClassifier().fit(iris.data, iris.target)
```

sklearn is so easy, it's actually DANGEROUS

Living Dangerously

Crappy Data

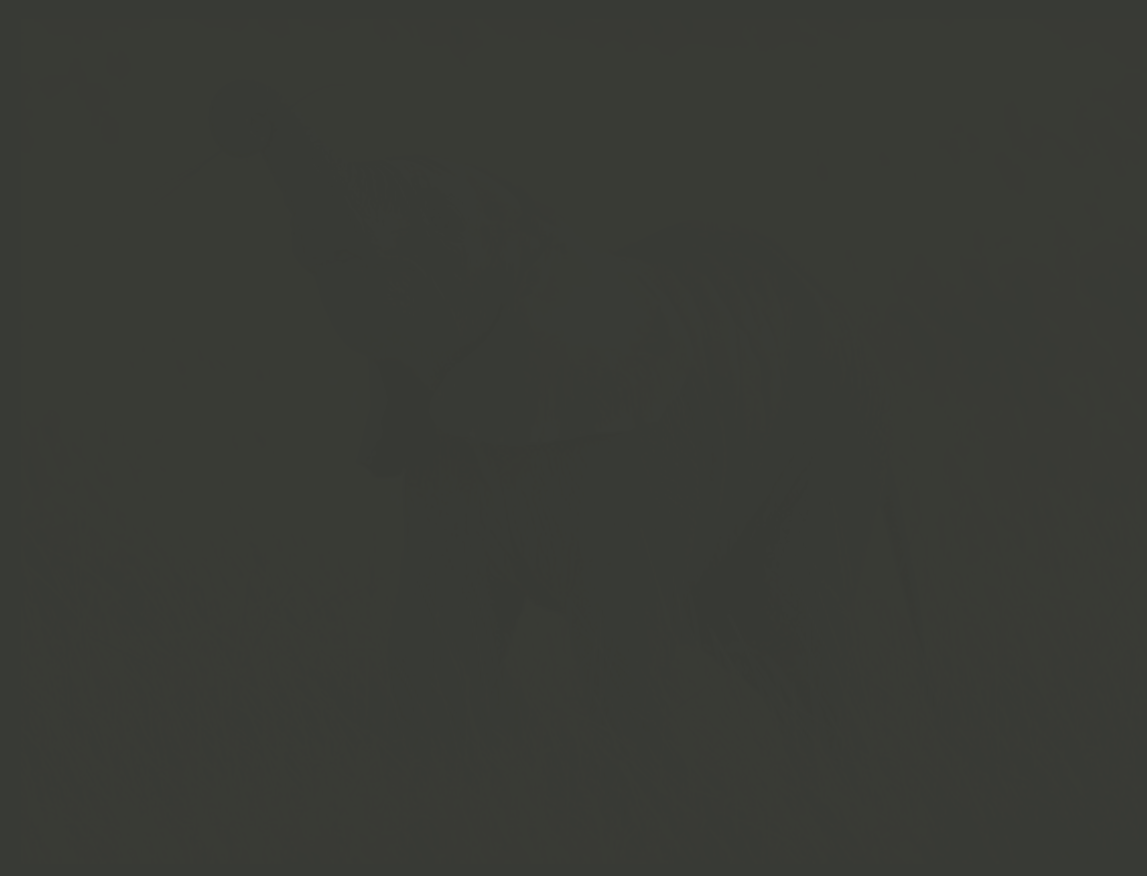
Heteroskedastic Errors



Living Dangerously

Crappy Data

Faint Objects



Living Dangerously

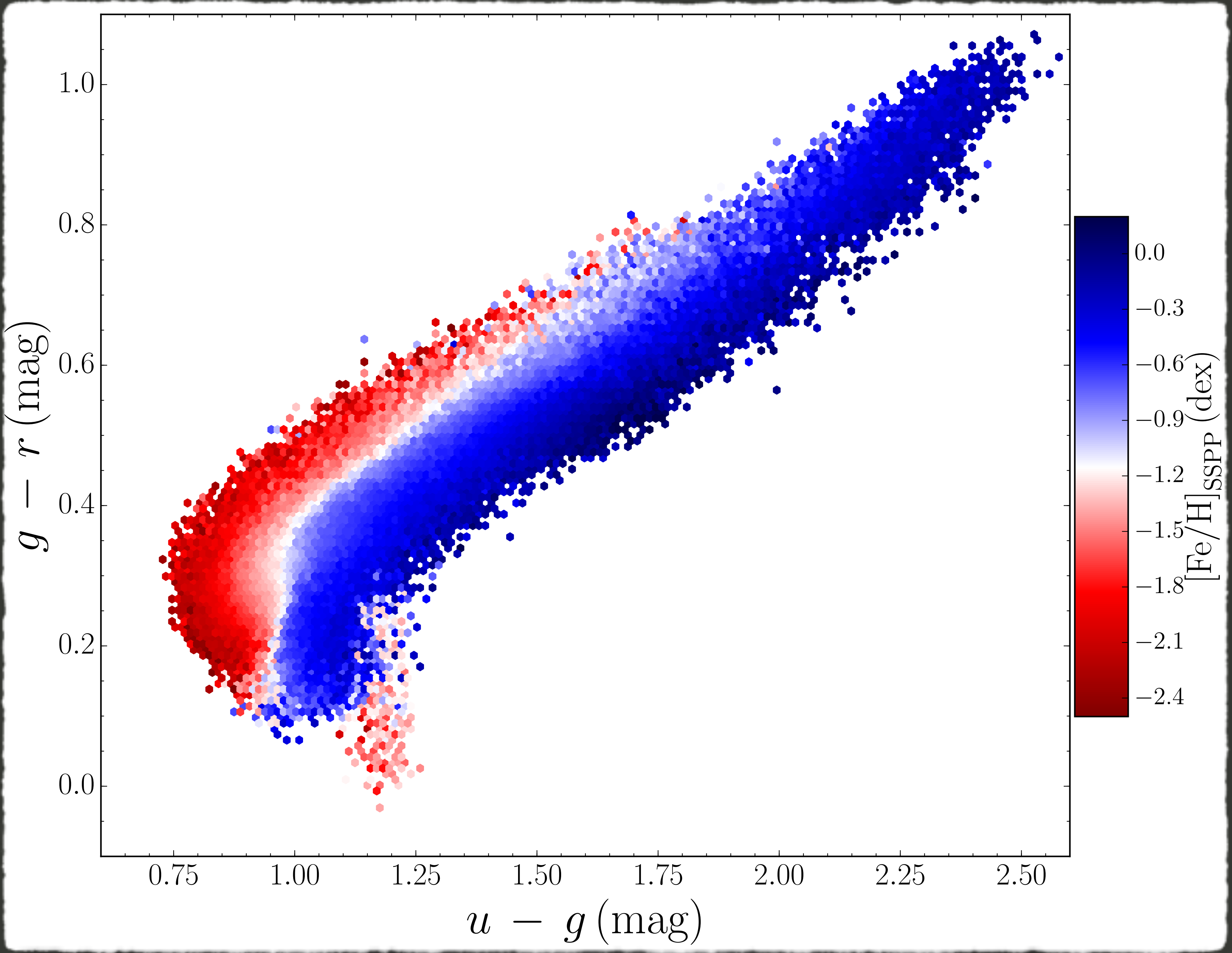
Crappy Data

Faint Objects

Living Dangerously

Crappy Data

Identify EMP stars with machine-learning

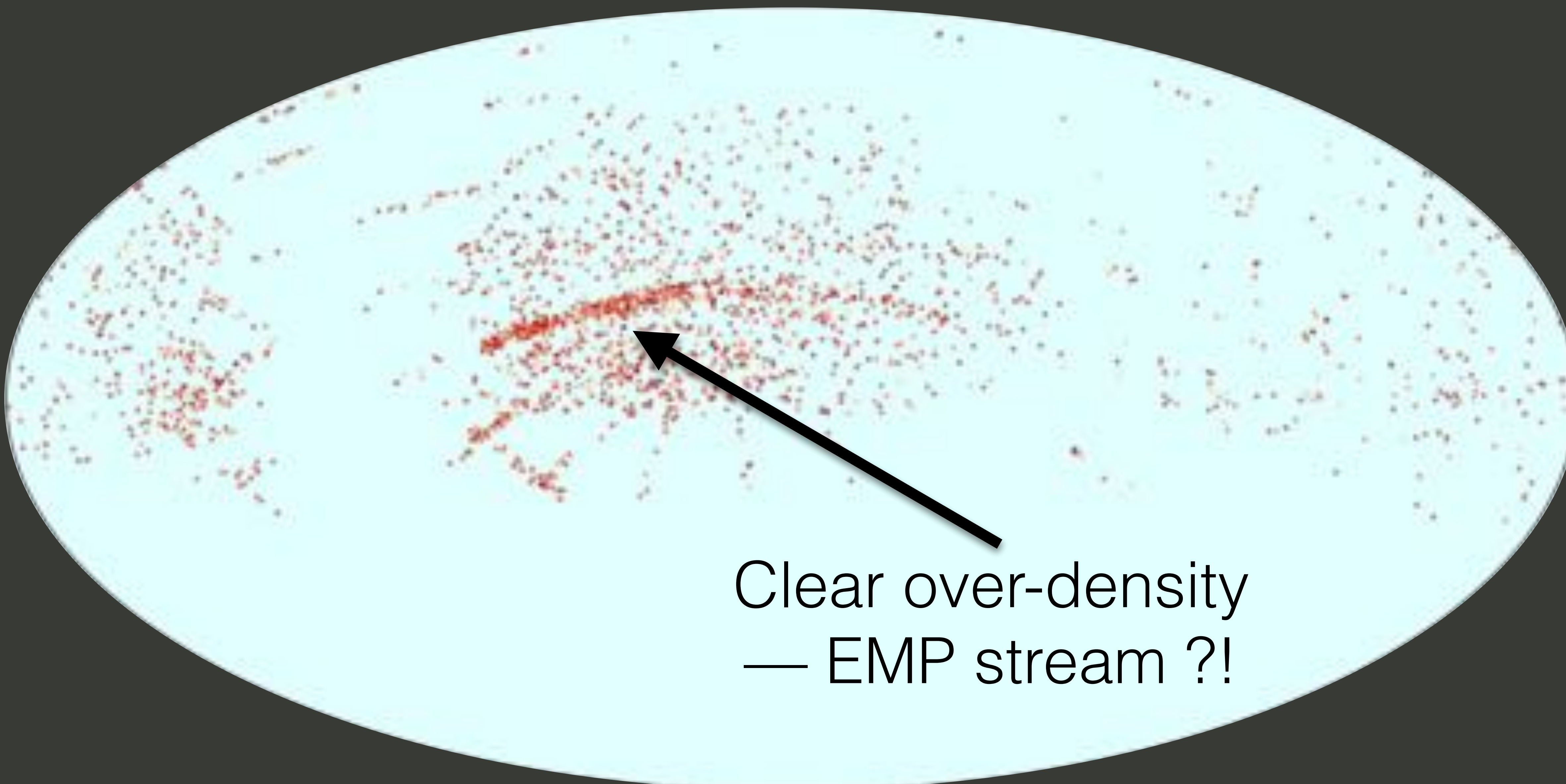


Miller 15

Living Dangerously

Crappy Data

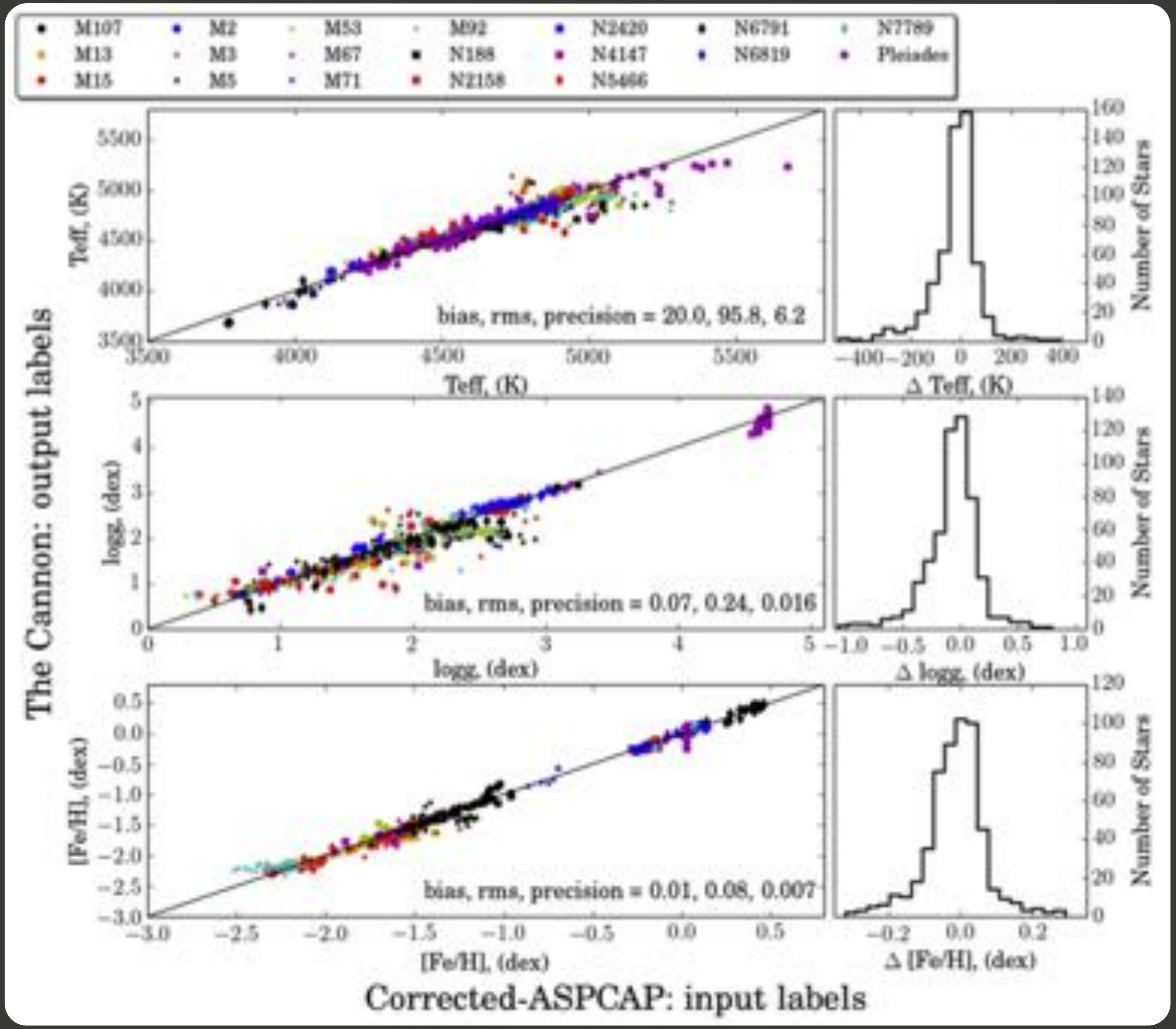
Identify EMP stars with machine-learning



Miller+15

Concepts Worth "Stealing" From ML

Evaluate algorithms with independent test sets

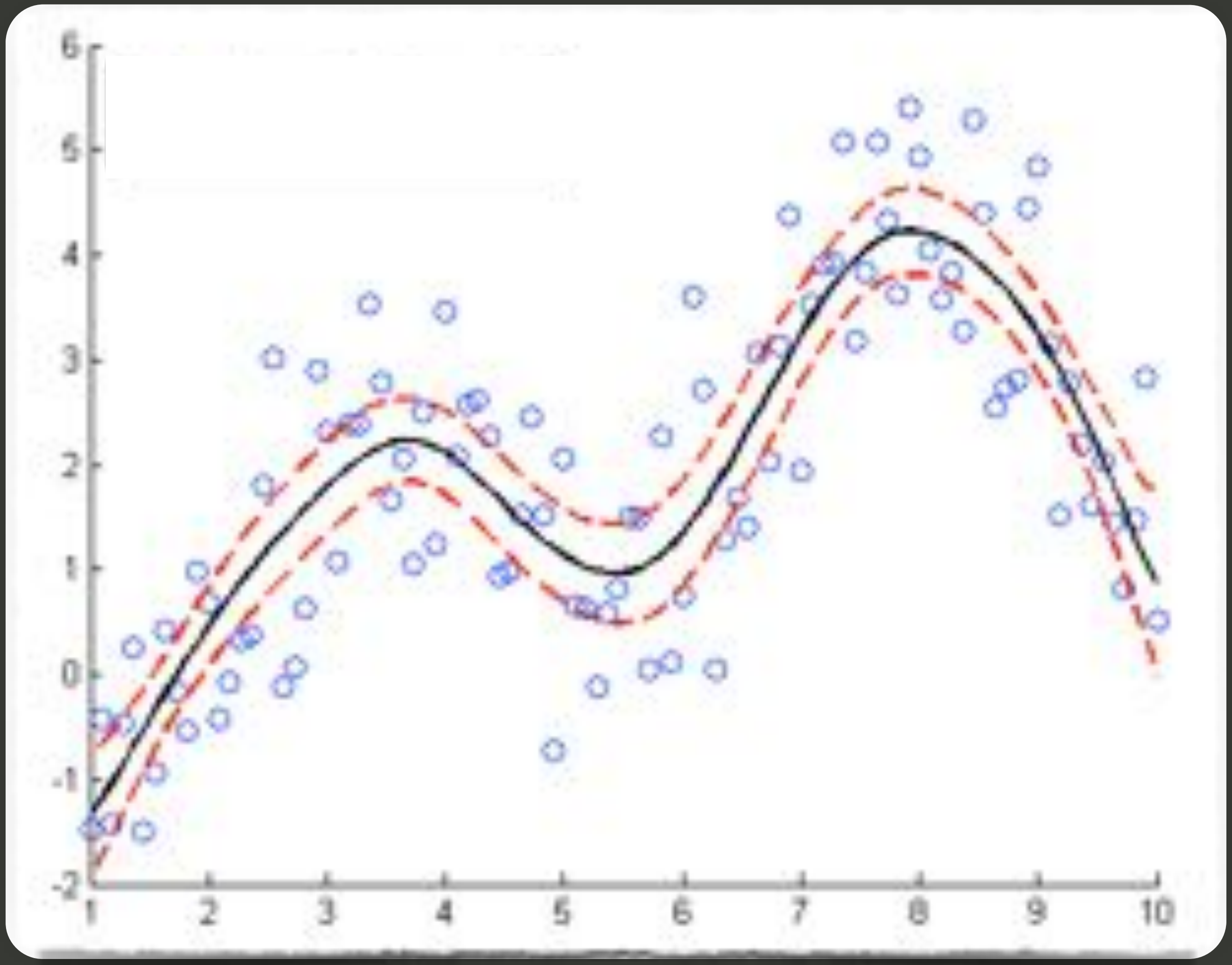


Ness+15

Concepts Worth “Stealing” From ML

Evaluate algorithms with independent test sets

Embrace flexibility, allow data to drive models



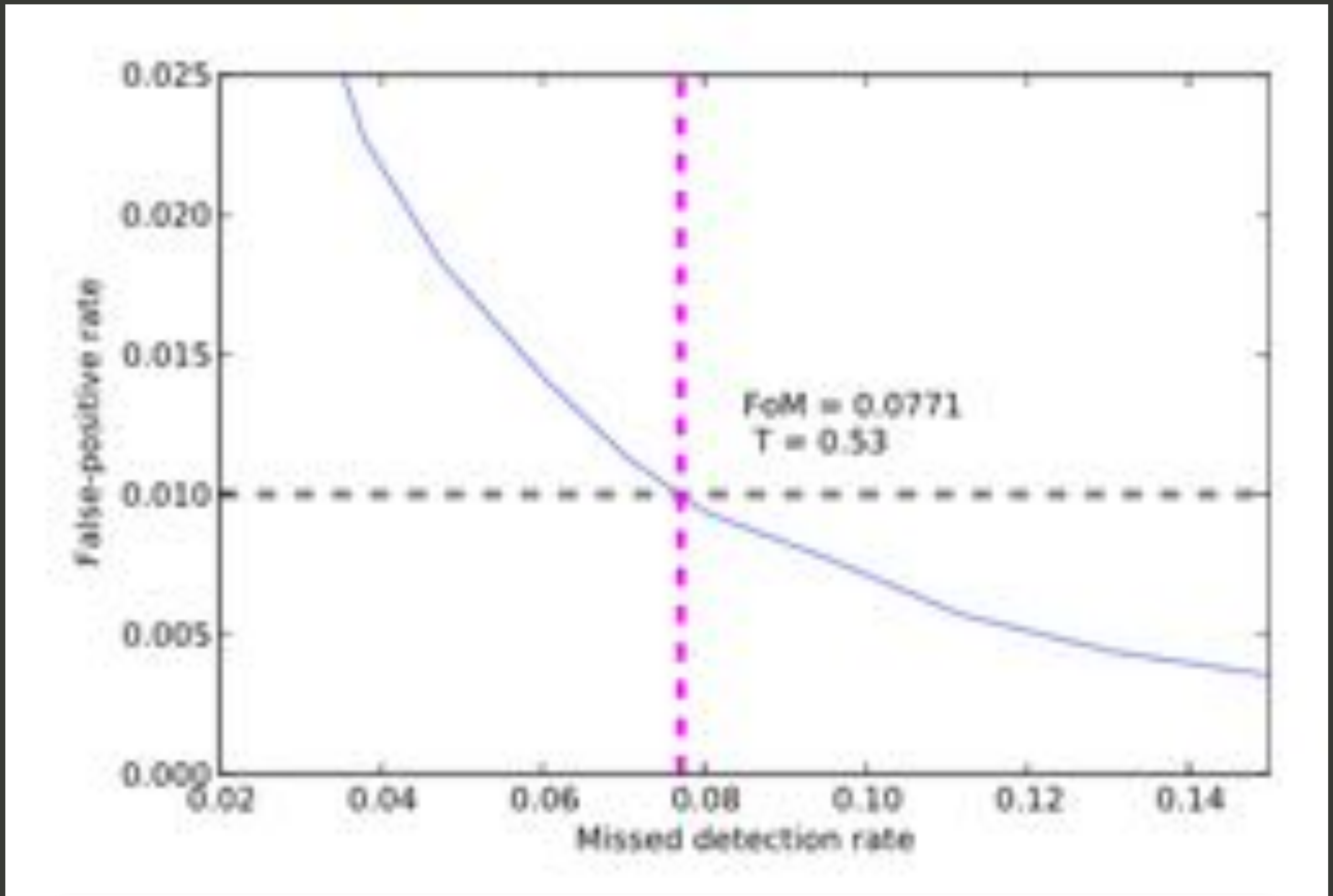
credit: blogs.mathworks.com

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

Conclusions

Data-driven solutions are a necessity for ever-growing wide-field surveys (ZTF, LSST, etc)

- 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