Something about machine learning

Viviana Acquaviva
(CUNY)
vacquaviva@citytech.cuny.edu

What is machine learning?

THE ART OF TEACHING A MACHINE TO MAKE DECISIONS

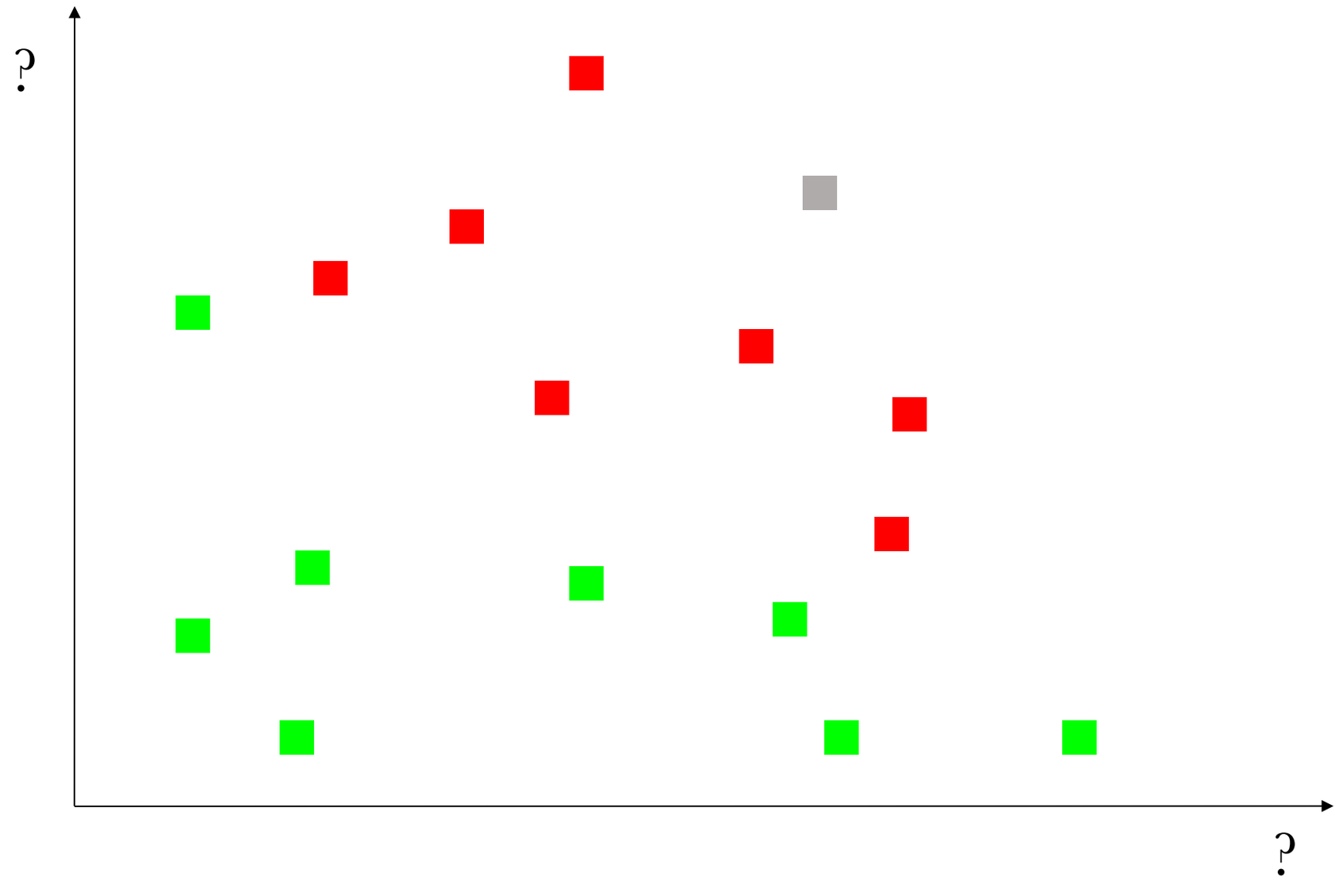
Recognize

Group together

Predict

Simplify

Our brain machine learns



(of course)
ML is not the only way

I CAN ALSO WRITE A FORMULA
(MAKE A MODEL)
TO PREDICT COLOR BASED ON COORDINATES

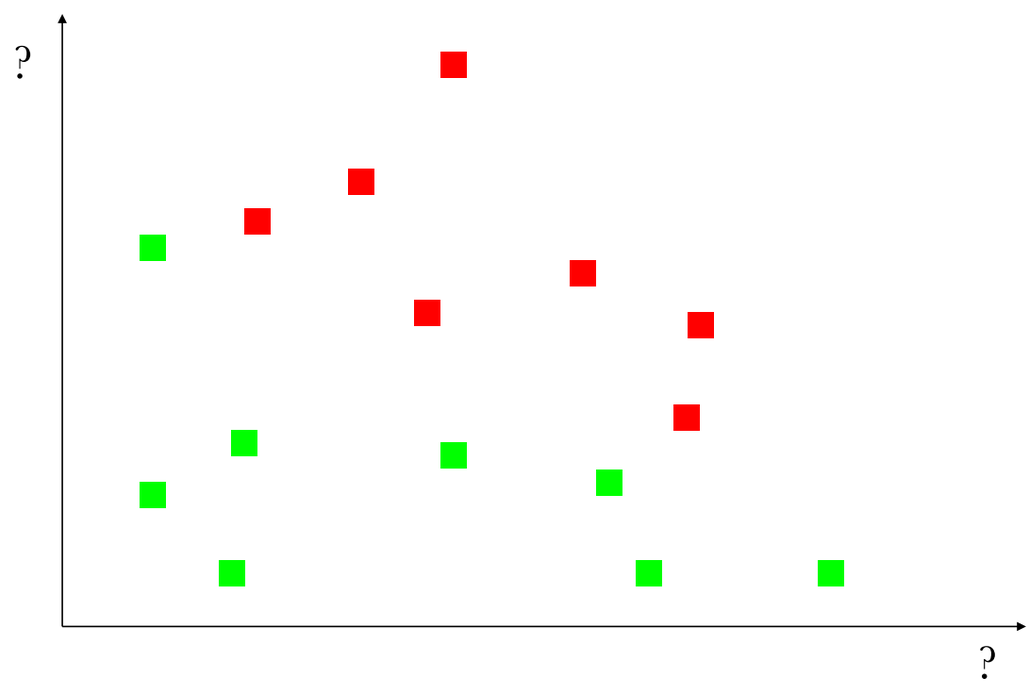
MACHINE LEARNING ALGORITHMS
PROVIDE AN "IMPLICIT MODEL" IN THE
FORM OF A PATH TO A DECISION
AND PERHAPS RESEMBLE MORE THE WAY
WE (HUMANS) SOLVE PROBLEMS

MACHINE LEARNING JARGON

Features are observable quantities known for all objects

Label is the target property that we want to predict

SUPERVISED ML ASSUMES THAT WE HAVE A SET OF OBJECTS WITH KNOWN LABELS, called the LEARNING SET



Performance is limited by size and quality of the learning set.

INPUT	OUTPUT
1	3
2	3
3	?

INPUT	OUTPUT
one	3
two	3
three	5
four	4
five	4
six	?

Data representation (sometimes called feature engineering) and determining whether you have enough data to create a good model are crucial.

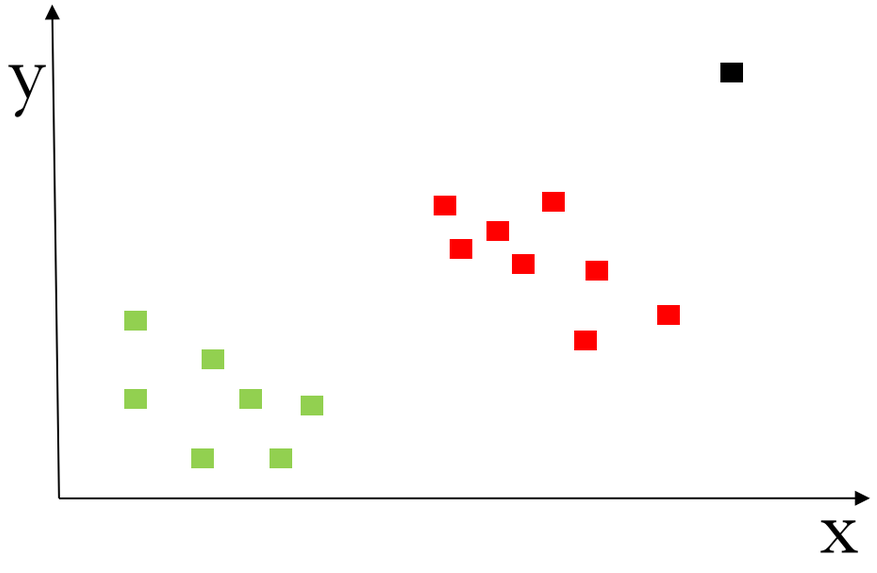
UNSUPERVISED MACHINE LEARNING

No labeled examples

MAP

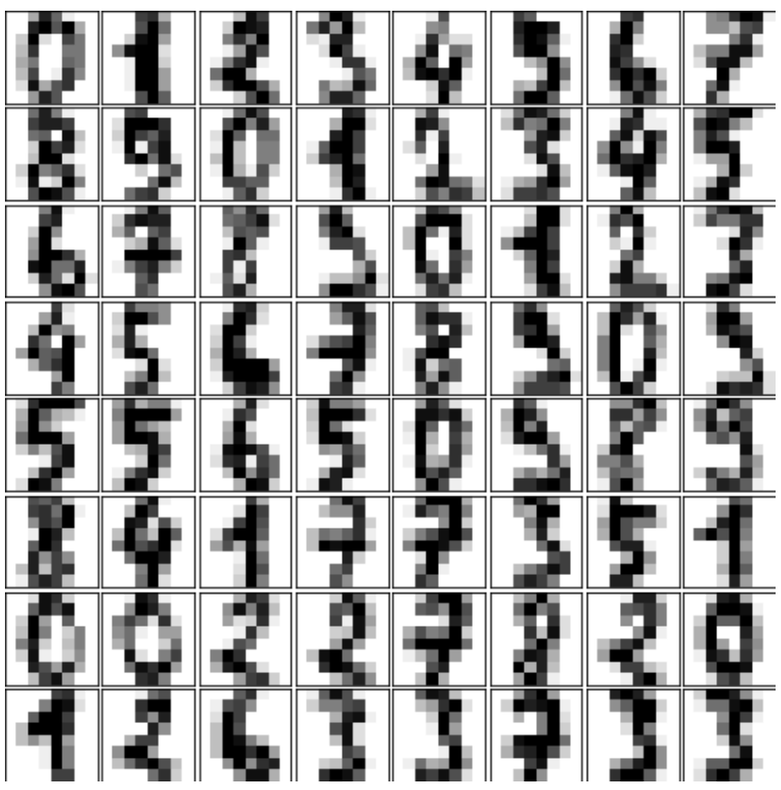
Useful to group together similar objects, find outliers, or find more efficient representations of data

Can be combined with human input or limited labels to understand the groups



Regression vs. classification

Usually we talk about classification when the target is a discrete variable (or class). For example in this image recognition problem:



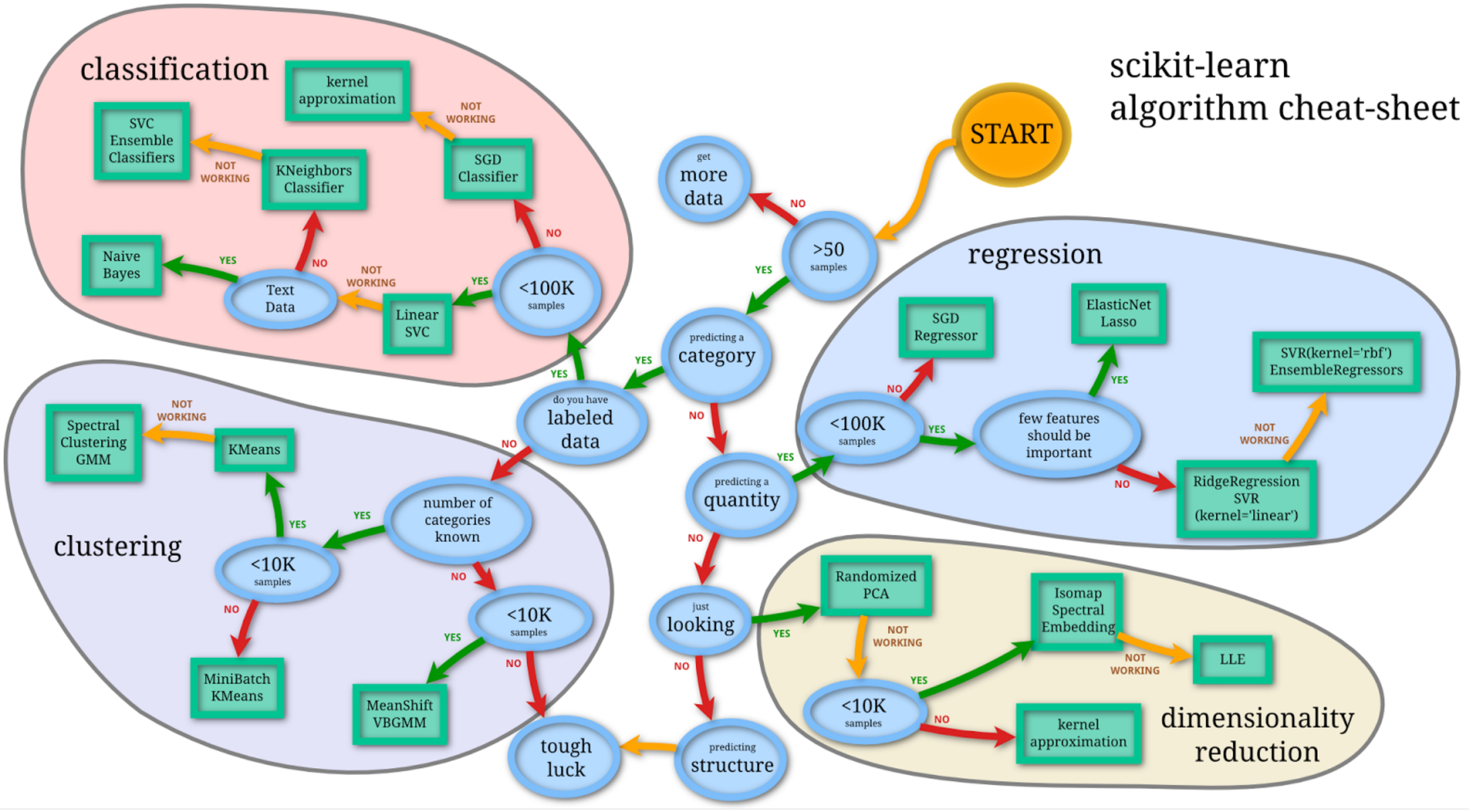
There are a finite (10) numbers of possible outcomes.

Regression vs. classification

- Vice versa, if we are trying to predict, say, the probability that it will rain in half a hour based on the current weather conditions, the outcome (target) is a continuous variable that can have all values between 0 and 1.

What if I was trying to decide whether or not I should bring an umbrella?

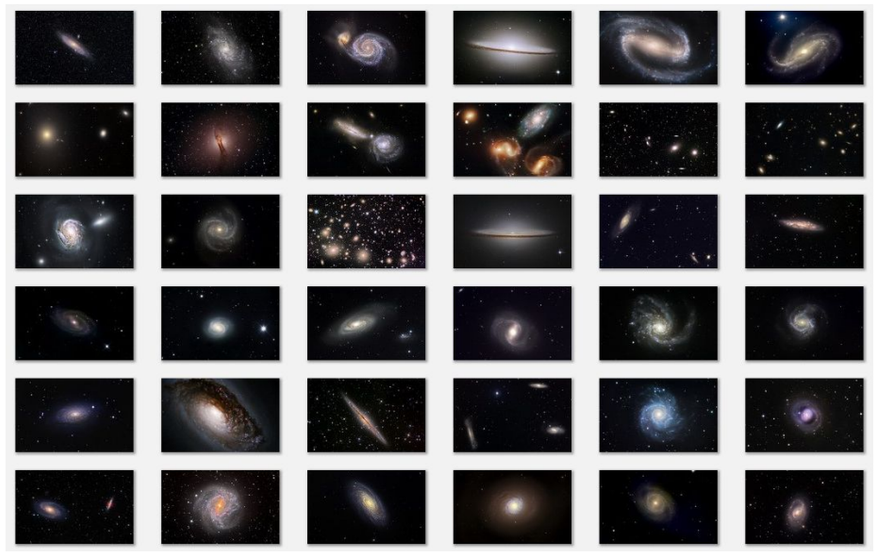
ALGORITHMS ABOUND



What can we do with them?

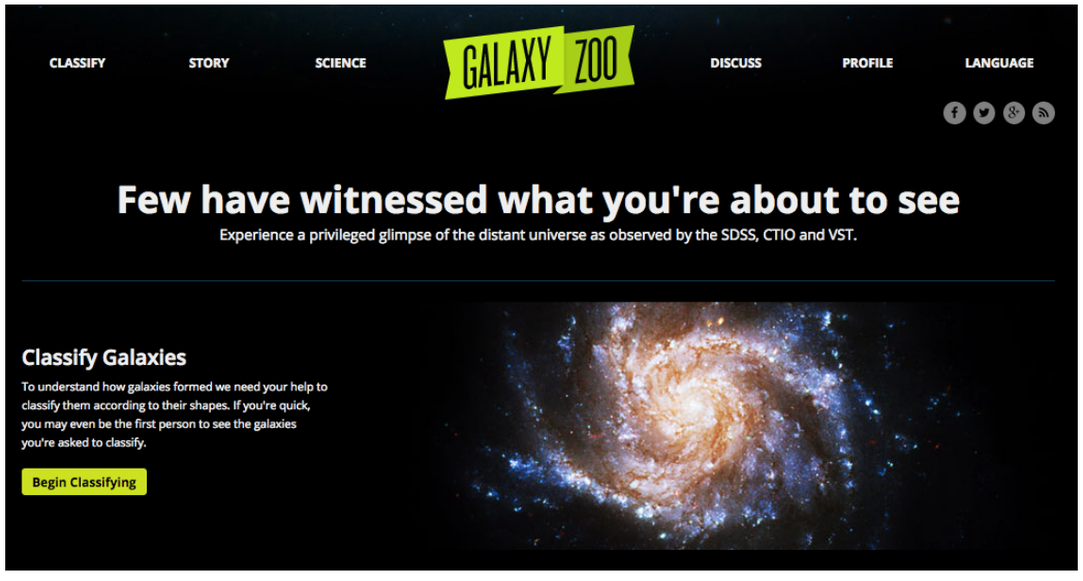
1. Save time.

Galaxy morphology



Trained humans are the best classifiers.
But what to do when
you have millions of objects?

Citizen science

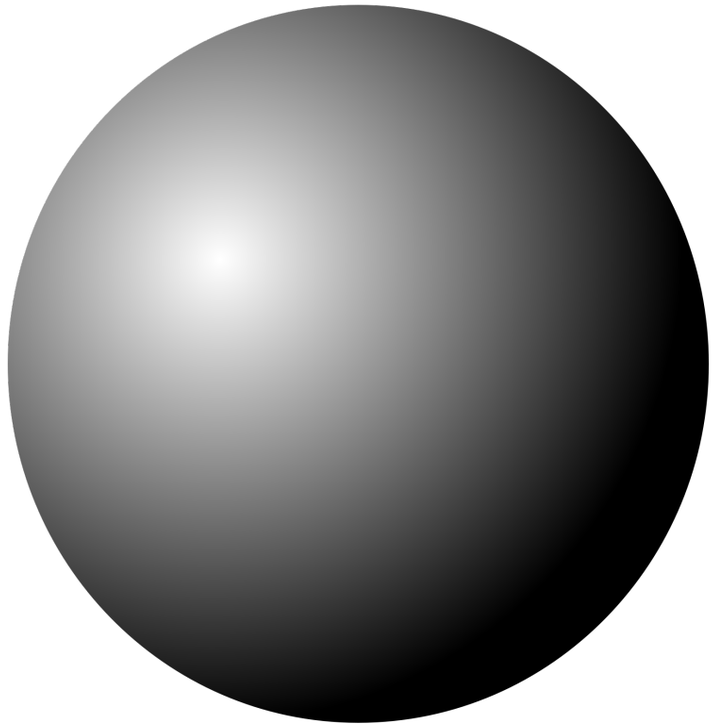


Automated Classification via Machine Learning

(supervised/unsupervised approach, see e.g. Hocking et al 2017)

2. Provide an alternative to simplistic models.

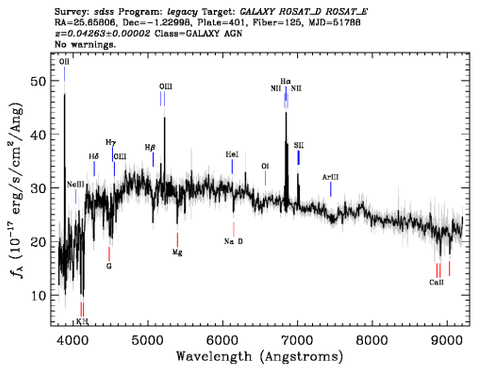
≠

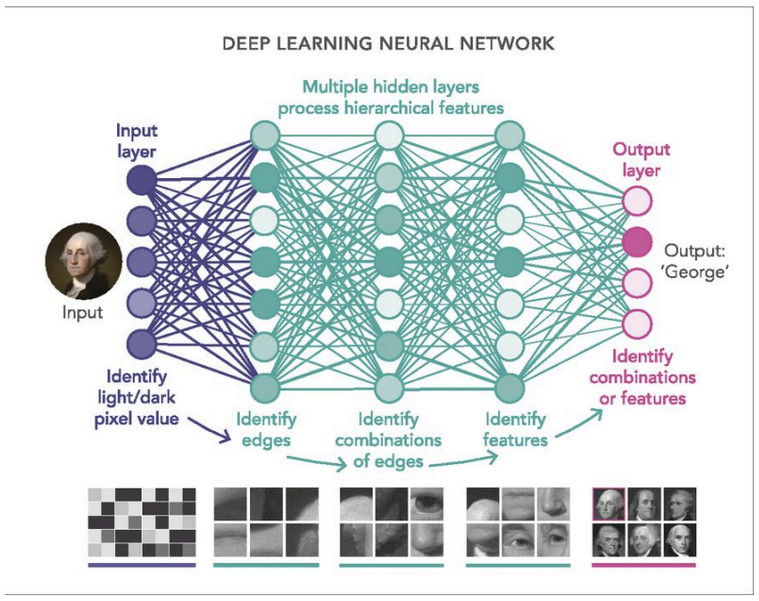


2. Provide an alternative to simplistic models.

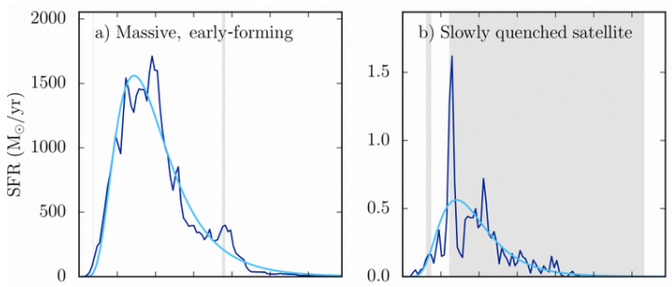
Example: Galaxy Photometric Redshifts or SFH.

Capture complex relationships between input and output, parametrization-free and likelihood-free (input can be a simulator)



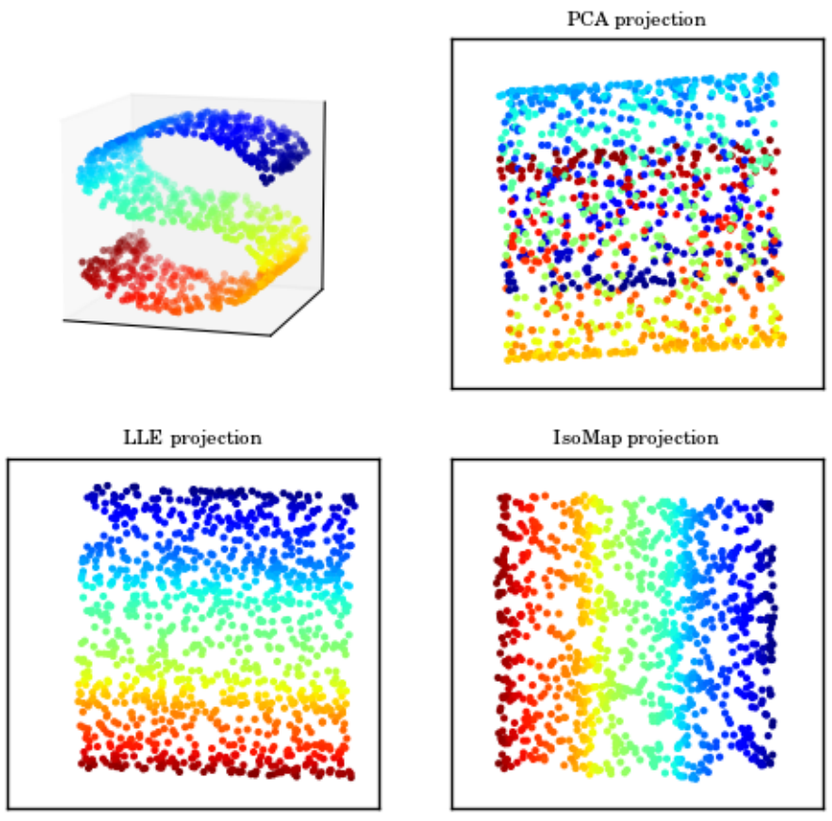


star formation histories (from Illustris)



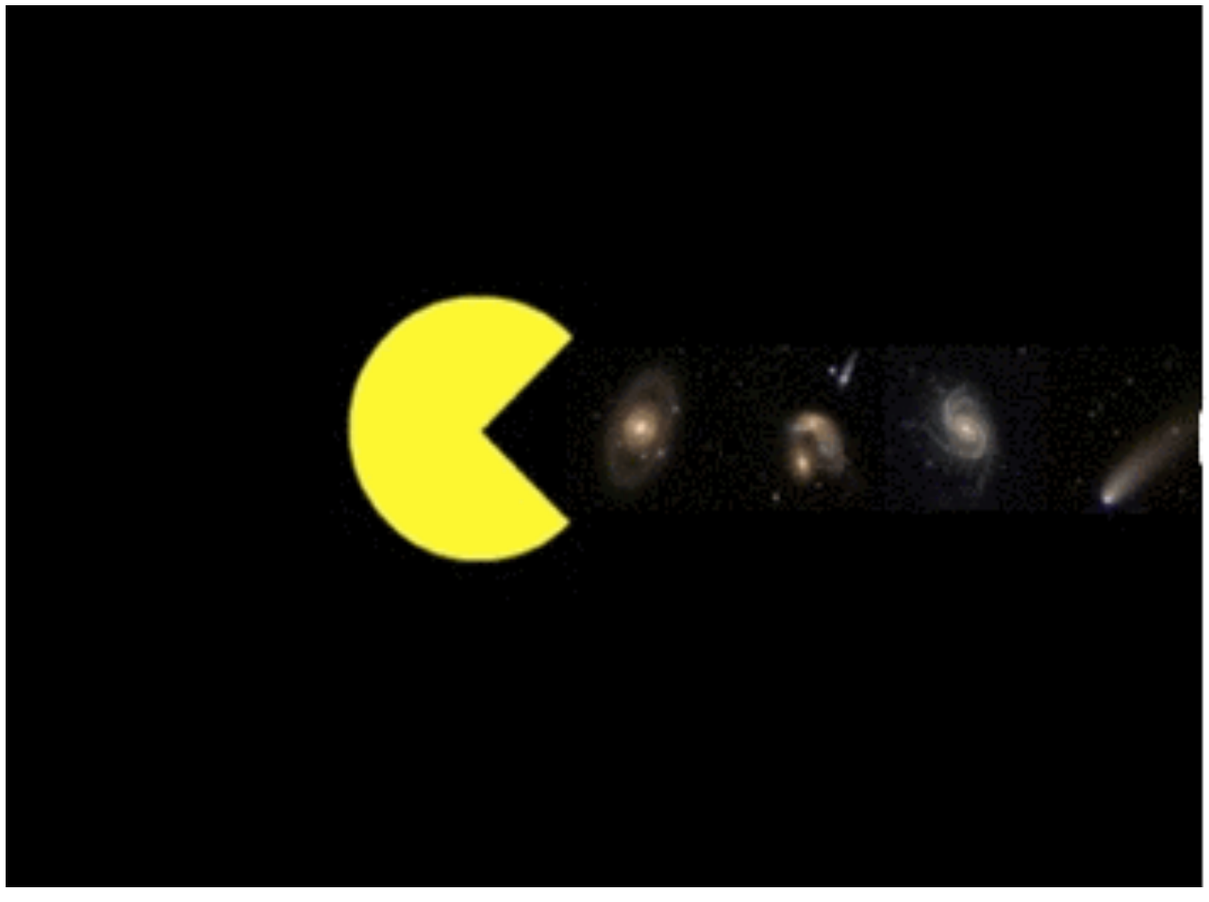
3. Make problems more tractable, e.g.:

Via dimensionality reduction



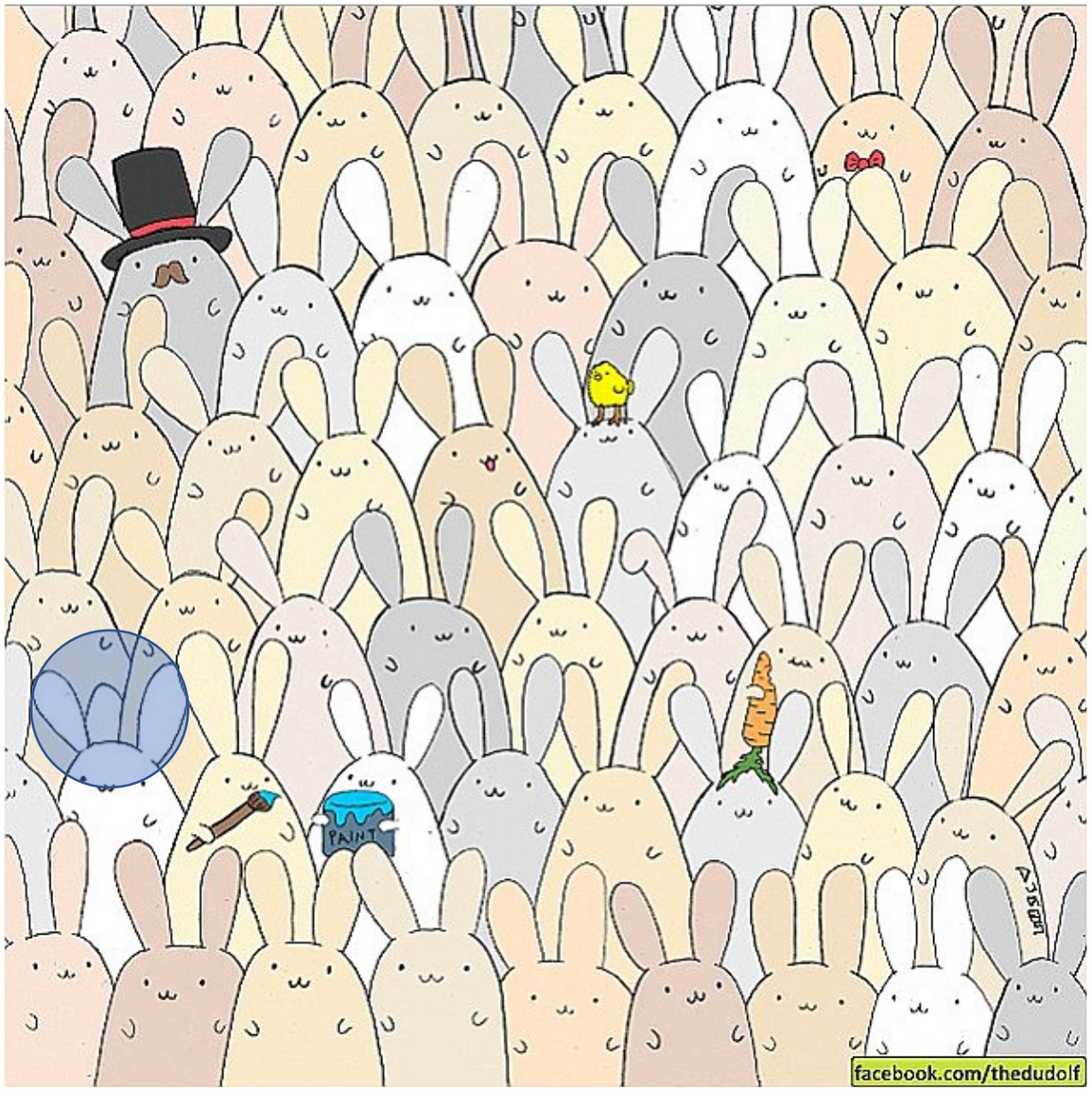
3D -> 2D
With various degrees of information loss

By working with higher level data



Model based inference:
P(model | image) = mess ☹

4. Allow serendipitous and data-driven discoveries

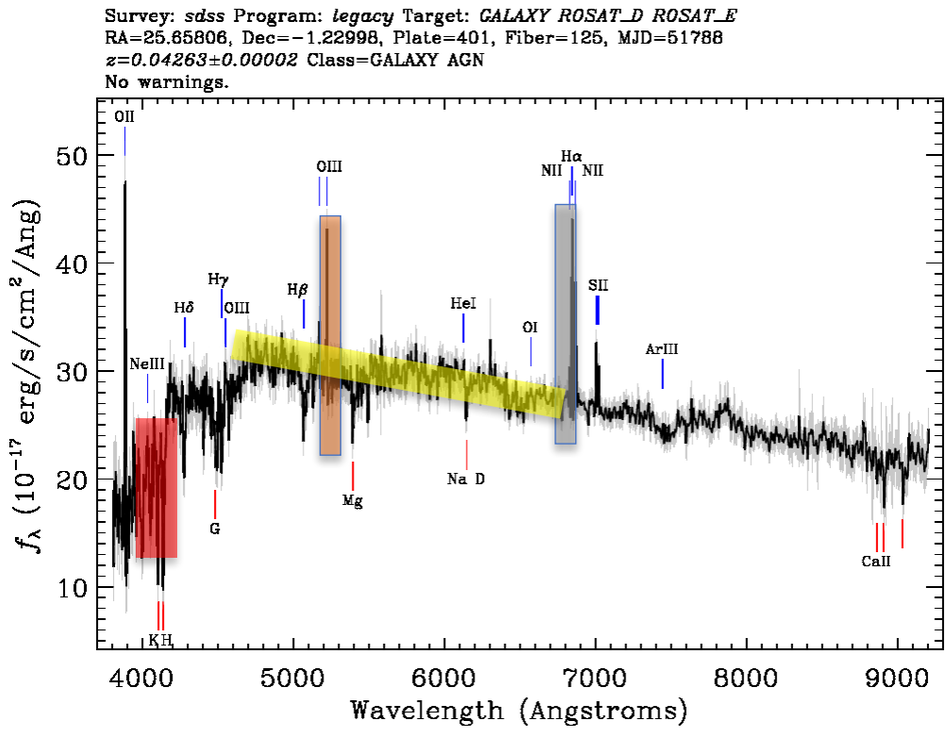


Outliers might be very interesting objects. In model fitting they are often discarded. Data mining techniques can tell you about new categories of objects.

5. Provide data-driven representations (feature generation)

1. better results

2. new intuitions



Machine Learning vs Model Fitting

ML

- Data-driven (only as good as the data)
- Likelihood-free (can use sims as input)
- Usually generalizes poorly (model derived using some data can't be applied blindly to different data)
- Answers questions; interpretation is possible but might be non-trivial
- Fast(er)
- More robust/accommodating of mixed, missing, and high-level data
- Allows serendipitous discoveries

MF

- Intuition or model-driven (only as good as the scientist :))
- Generalizes well if model (physics) is well understood
- Builds subject matter knowledge; easier to interpret
- Might be computationally intensive
- Dealing with heterogenous data often a pain in the neck
- Leads to loss of information if models are too simplistic

Synergy is often the best strategy

Questions?

Let’s play a fun game ☺

Let's answer these questions

Is this supervised or unsupervised ML?

Is this a classification or regression problem?

What could be useful features (data) to collect?