
# Machine Learning

Daniela Huppenkothen

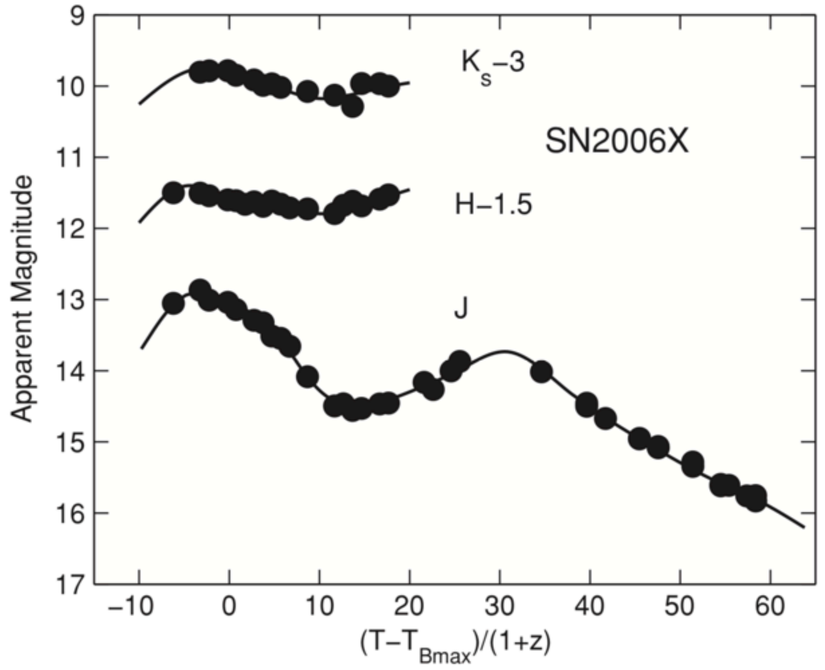
*DIRAC Institute, University of Washington*


https://github.com/dhuppenkothen/cargese2018_tutorials


# Yesterday ...

# SN Ia Light Curves

**39 SN**
**=**
**347 parameters**

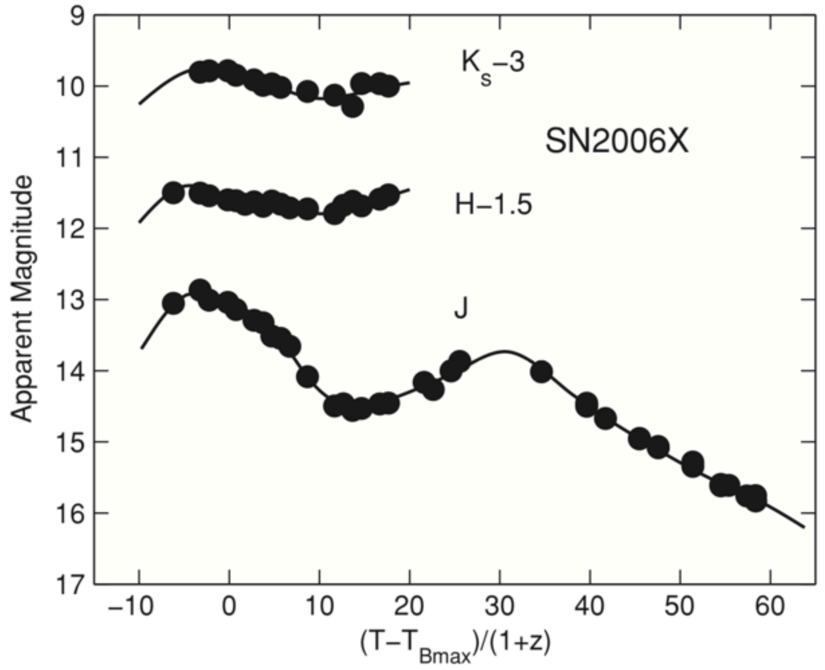



Mandel et al (2009)

**3900 SN**
**=**
**>30,000 parameters**

# SN Ia Light Curves

39 SN
=
347 parameters

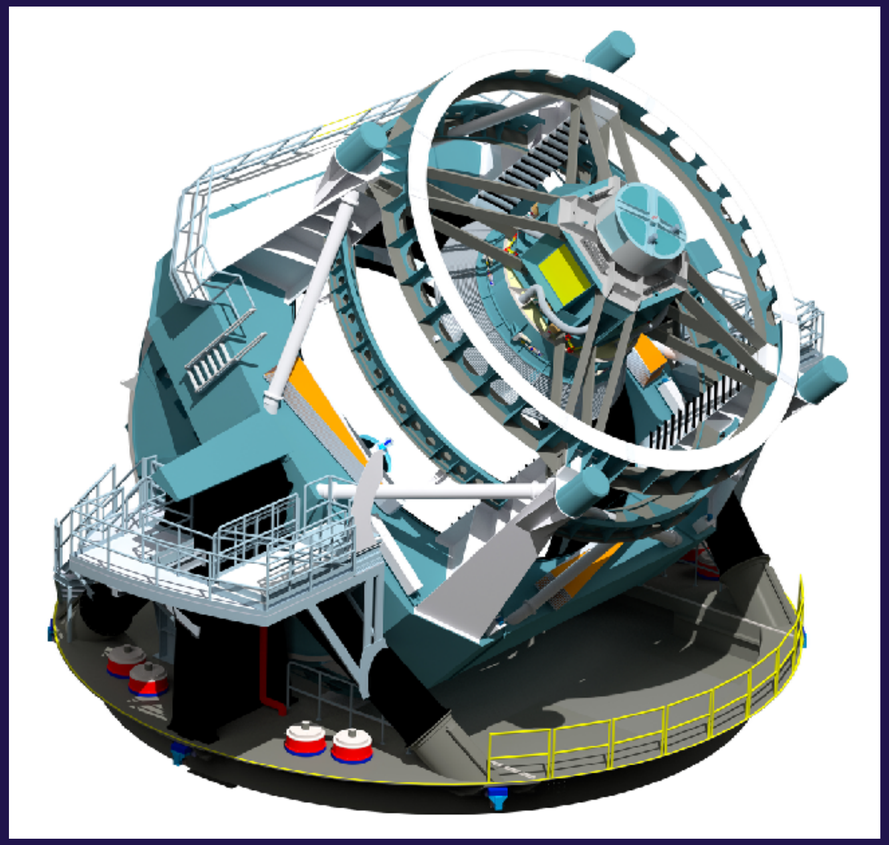



Mandel et al (2009)

3900 SN
=
>30,000
parameters

😱😳😱

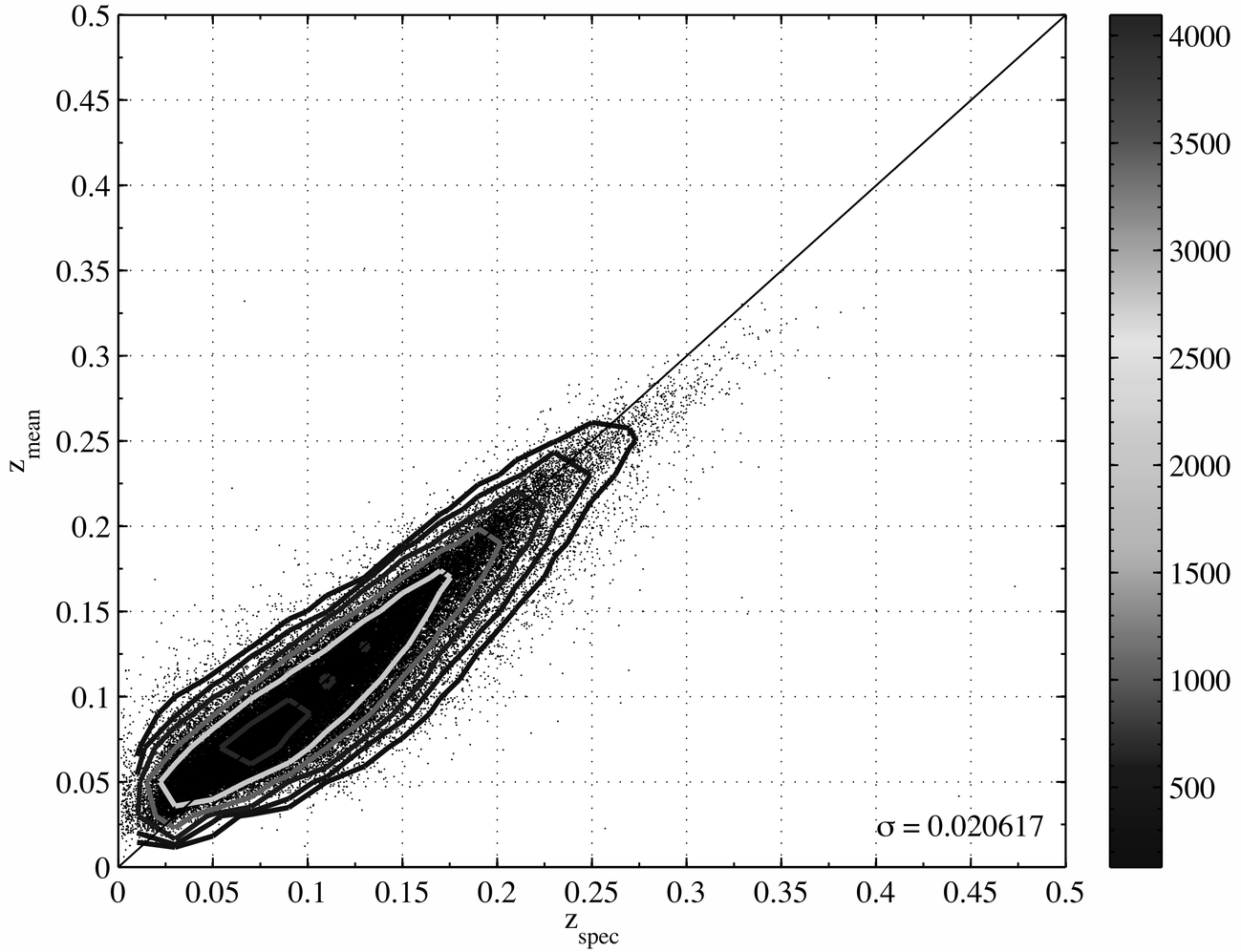
# Large Synoptic Survey Telescope



Credit: The LSST Corporation

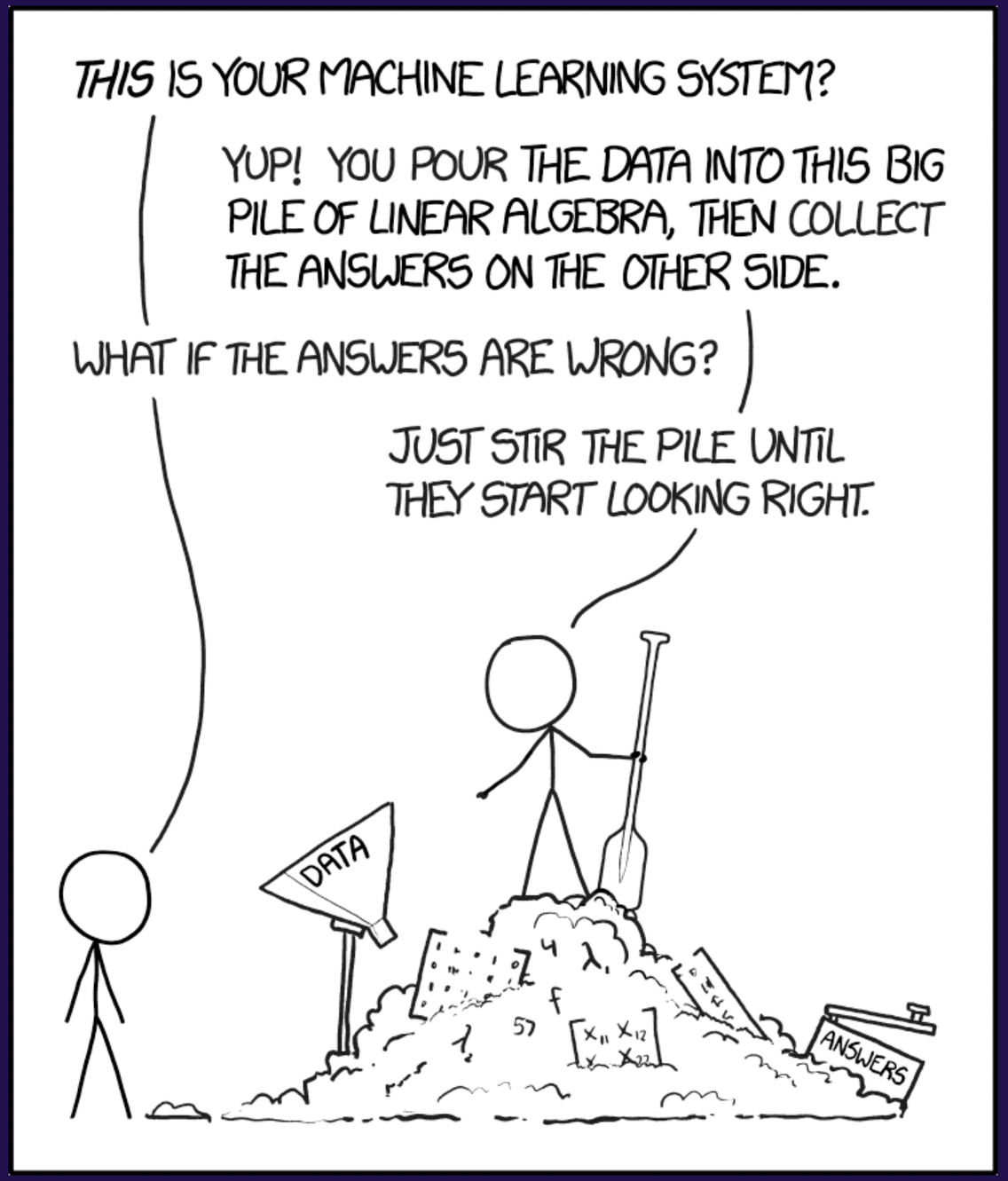
~ 40 billion sources!

# Photometric Redshifts




Ball et al (2008)

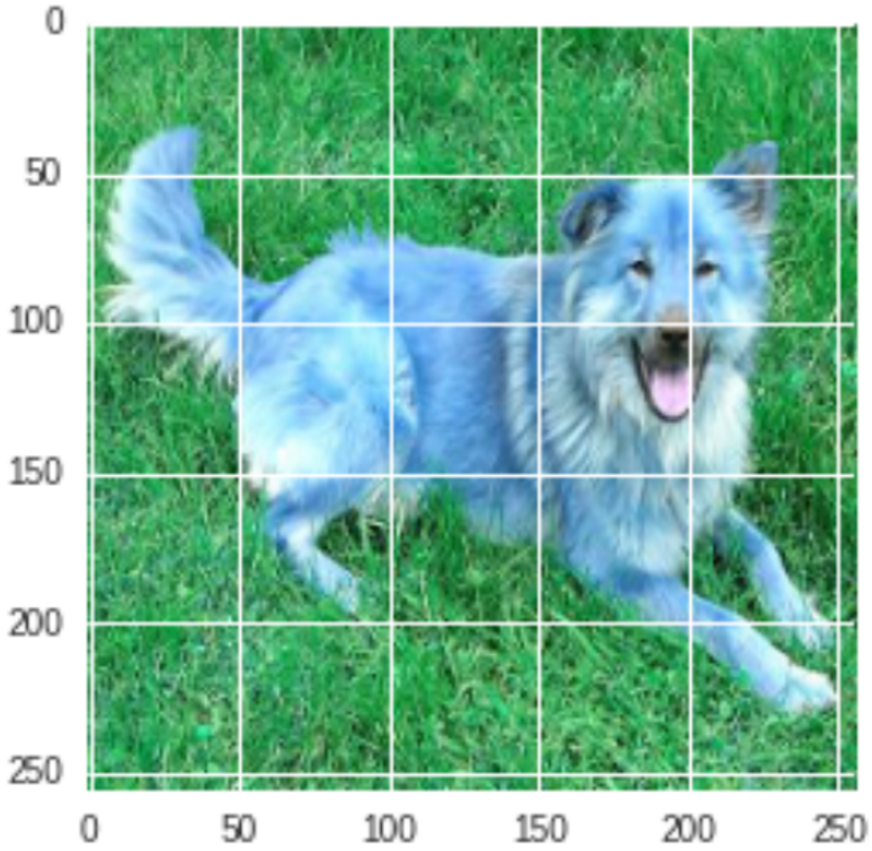
# Can we teach a computer to learn an empirical function?


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

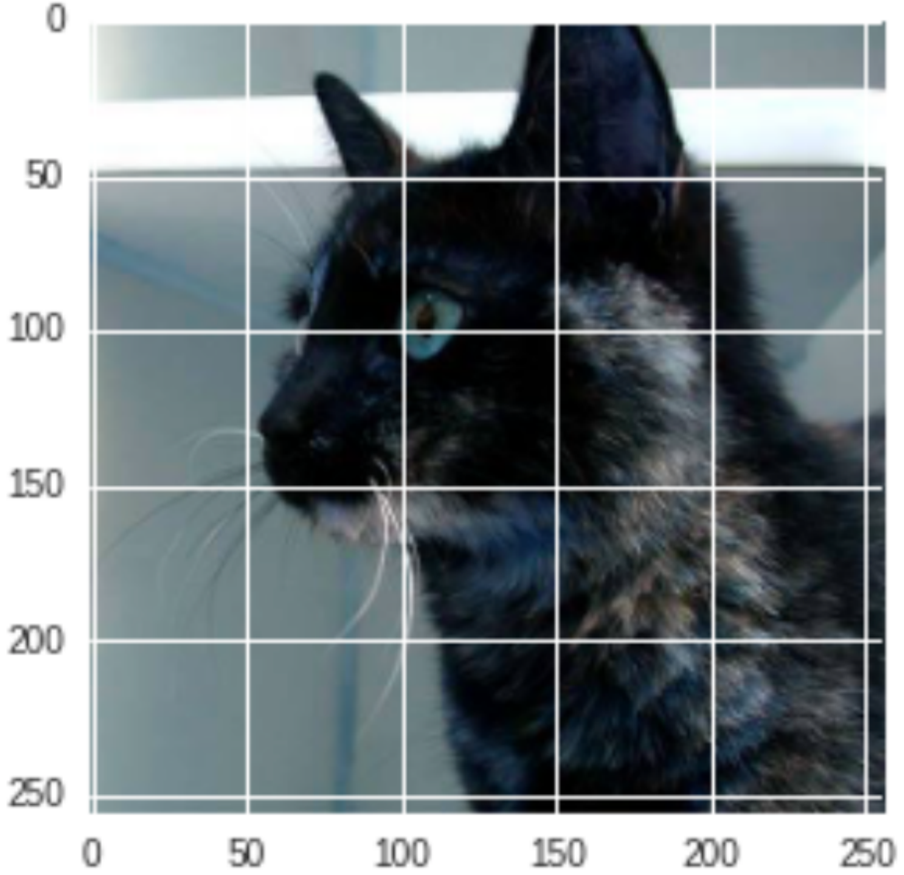
http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

# Let’s go to a simpler problem …

I am 100.00% sure this is a Dog



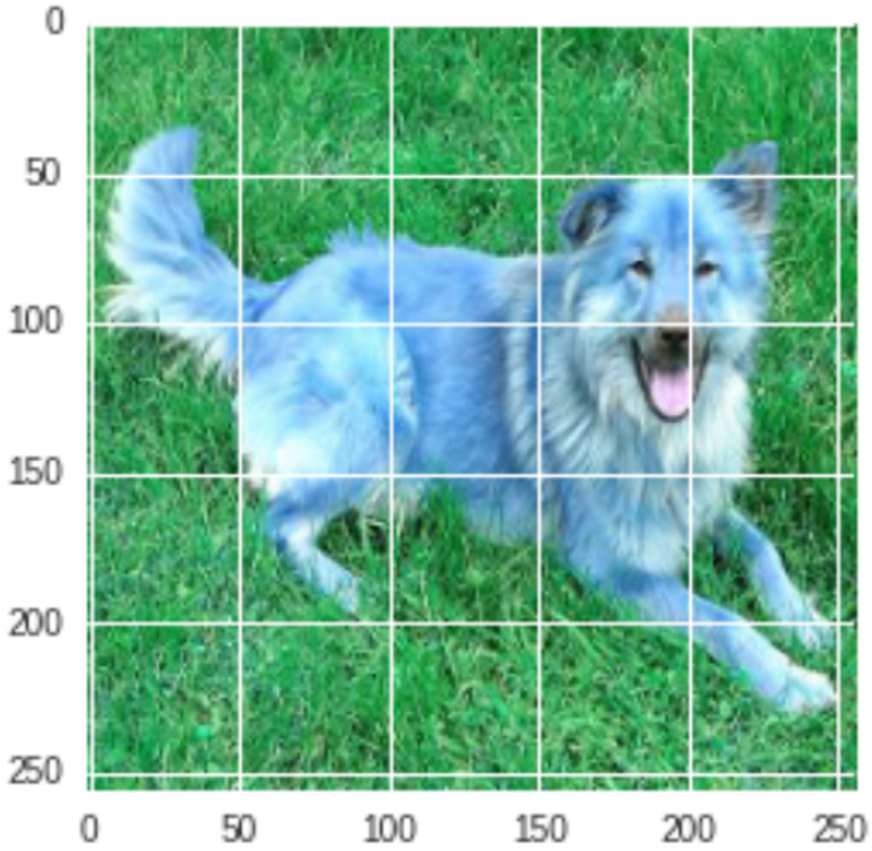

I am 99.27% sure this is a Cat



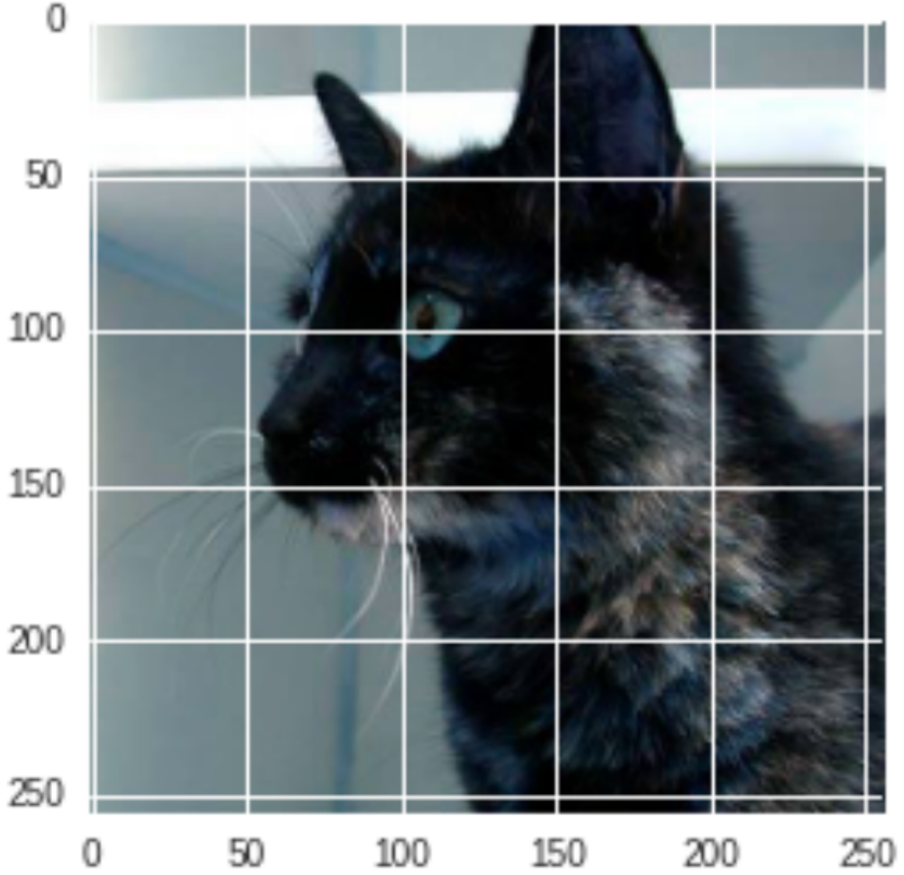

https://www.kaggle.com/jeffd23/catdognet-keras-convnet-starter

# Problem definition: “Distinguish a photo from a cat from a photo of a dog.”

I am 100.00% sure this is a Dog




I am 99.27% sure this is a Cat




https://www.kaggle.com/jeffd23/catdognet-keras-convnet-starter

# Need to translate abstract problem into algorithm …

# An even simpler problem ...

[redacted]

Aug 14

to

**Be careful with this message.** Similar messages were used to steal people's personal information. Unless you trust the sender, don't click links or reply with personal information. Learn more

Hello Dear
My name is Marie Michele. I wish to invest in your country which came into existence after a careful and comprehensive study and analysis of your country's development.
I am seeking that company or individual who have the ability to assist me for the transfer of inheritance fund, invest and management. I am 20years old. I really need your assistance,my family members poison my father which resulted to his death few year ago,and want to seat on the inheritance fund my Father left for me with the Security company, since the death of my parents they has been looking for me.

I am now in hiding and the documents of inheritance fund is with me. Please help me to have this transferred to your country and I will come to join you, I will be waiting for your reply to my email, Is only you i well trust with all my heart,i hope you will help,

Regards

...


# Rule-Based Spam Filter

“You have won”: -1,000
“plotting”: +1,000


# Rule-Based Spam Filter

“You have won”: -1,000
“plotting”: +1,000

## But: what should the scores be?


# Rule-Based Spam Filter

“You have won”: -1,000
“plotting”: +1,000

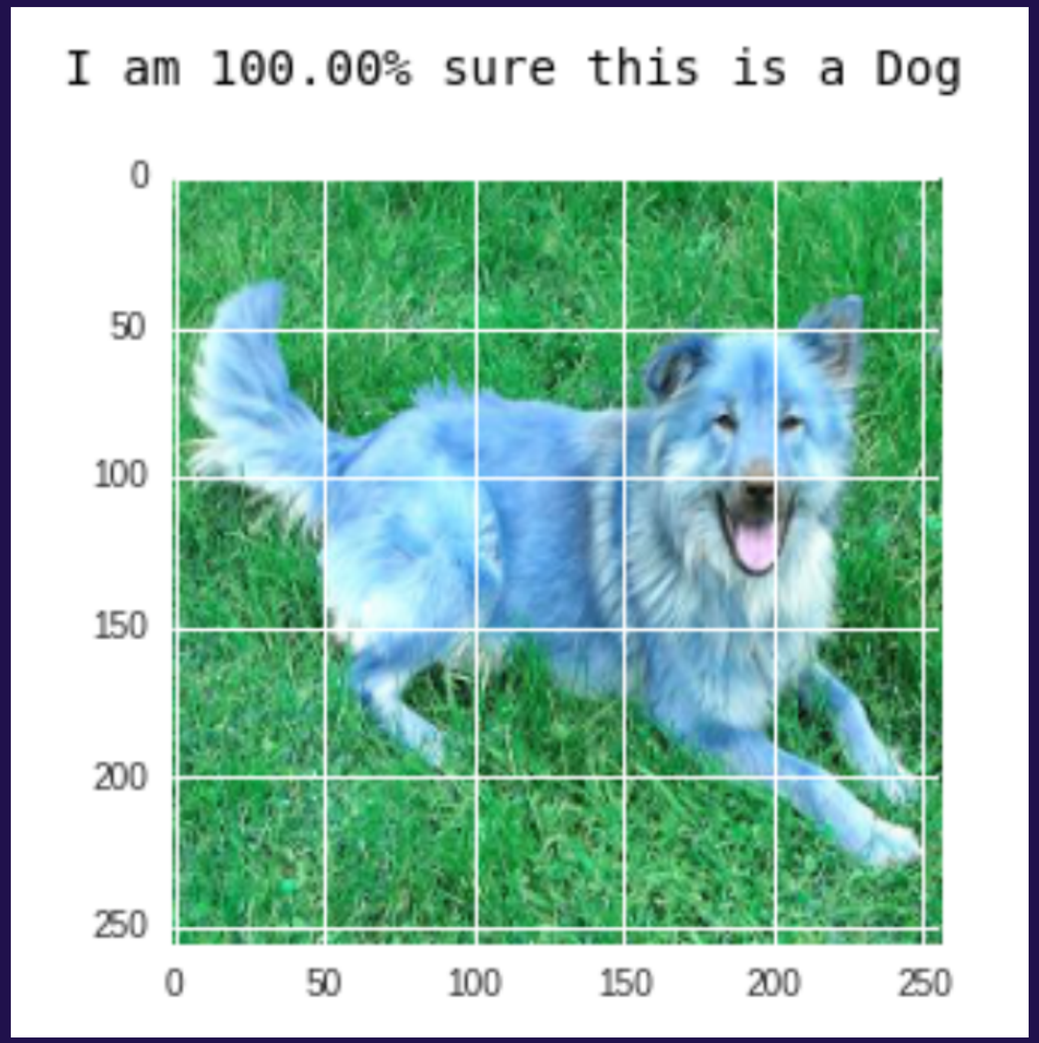
**But: what should the scores be?**

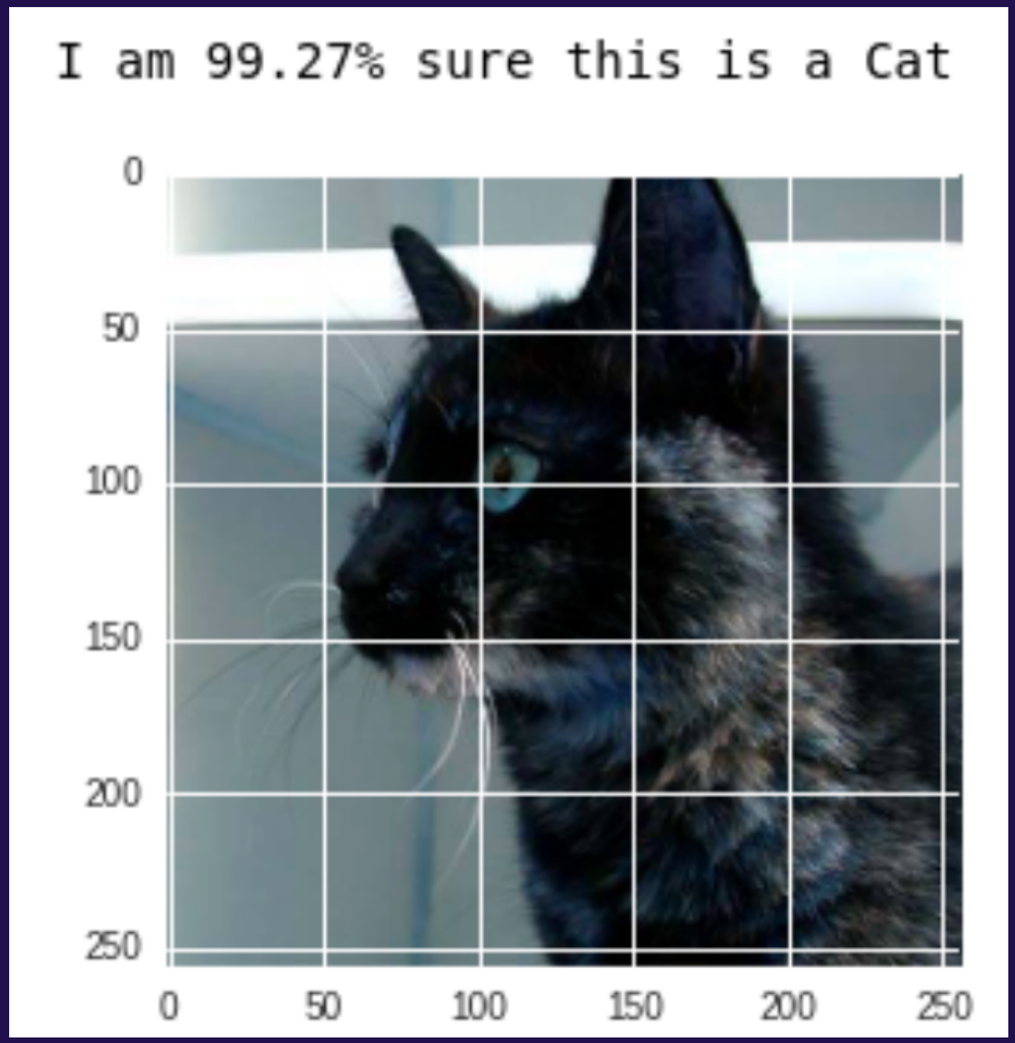
**What about combinations of phrases?**


# A better mathematical description

- Create a feature vector with one-hot encoding
- Each word gets a weight, do inner product of weights * features
- Optimize weights using a known data set of spam/non-spam e-mails
- Note: could also include combinations of words (n-grams)

# What about images?




https://www.kaggle.com/jeffd23/catdognet-keras-convnet-starter

# Key question: what do we want to learn from our model?

# Key question: what do we want to learn from our model?

predict things

# Key question: what do we want to learn from our model?

predict things

learn (physical) parameters

# Key question: what do we want to learn from our model?

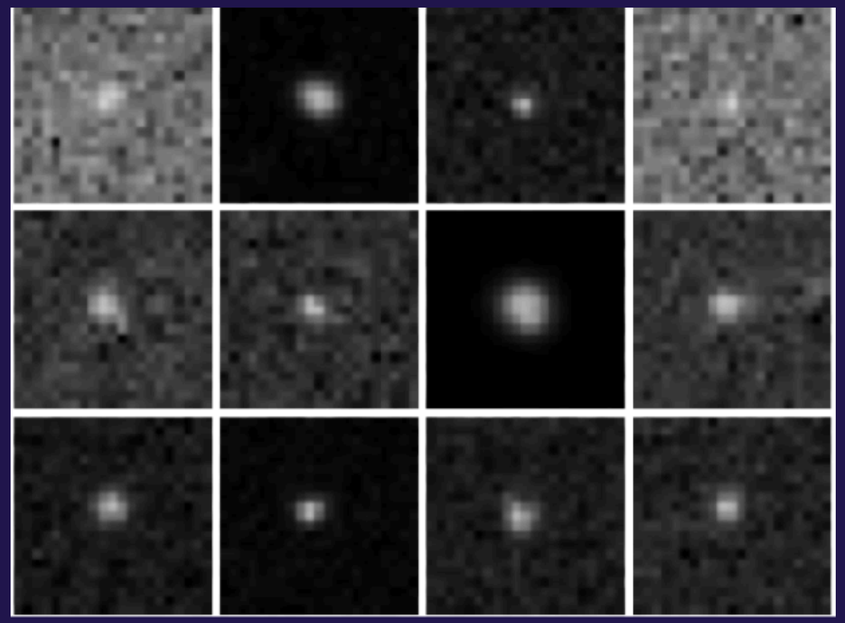
predict things

emulate a system

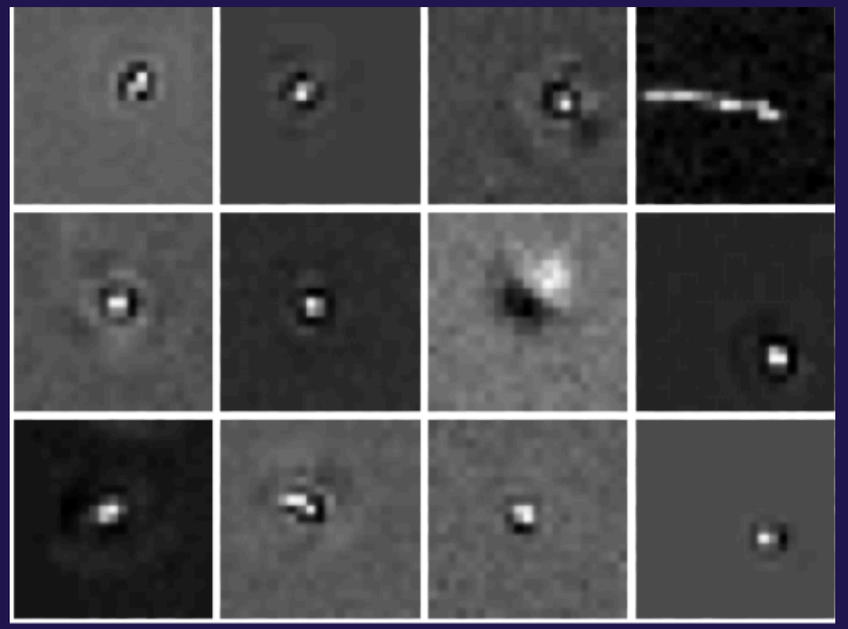
learn (physical) parameters

# is it real or bogus?

## real



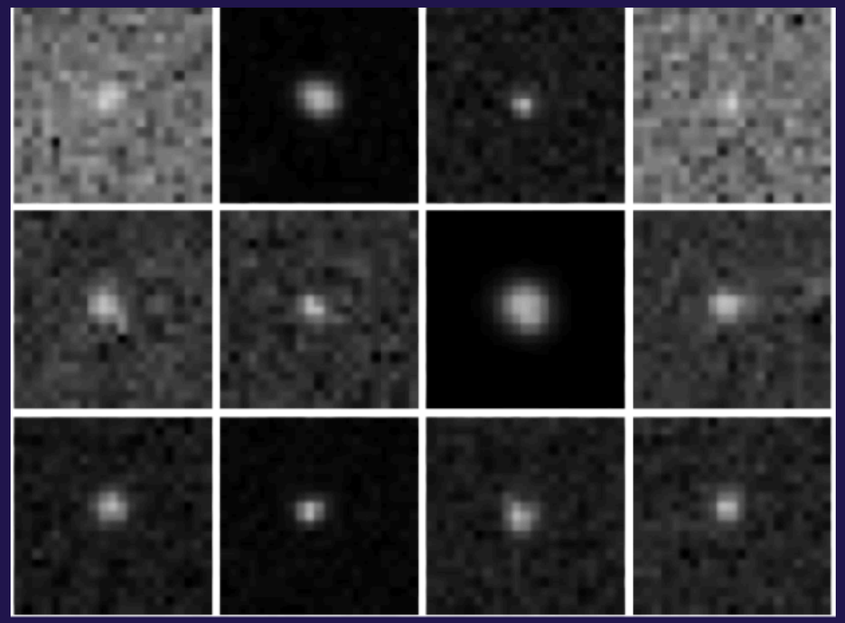
## bogus



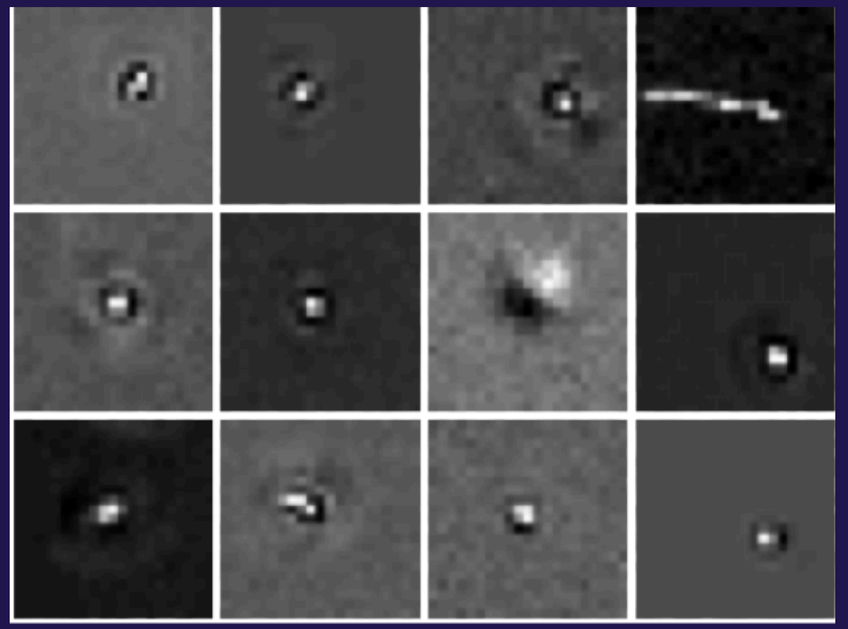
e.g. Brink et al (2012): Random Forests
Wright et al (2015): Support Vector Machines, Neural Networks, Random Forests
Masci et al (2016): Random Forests
Gieseke et al (2017): Convolutional Neural Networks
Sedaghat + Mahabal (2018): Convolutional Neural Networks

# is it real or bogus?

real



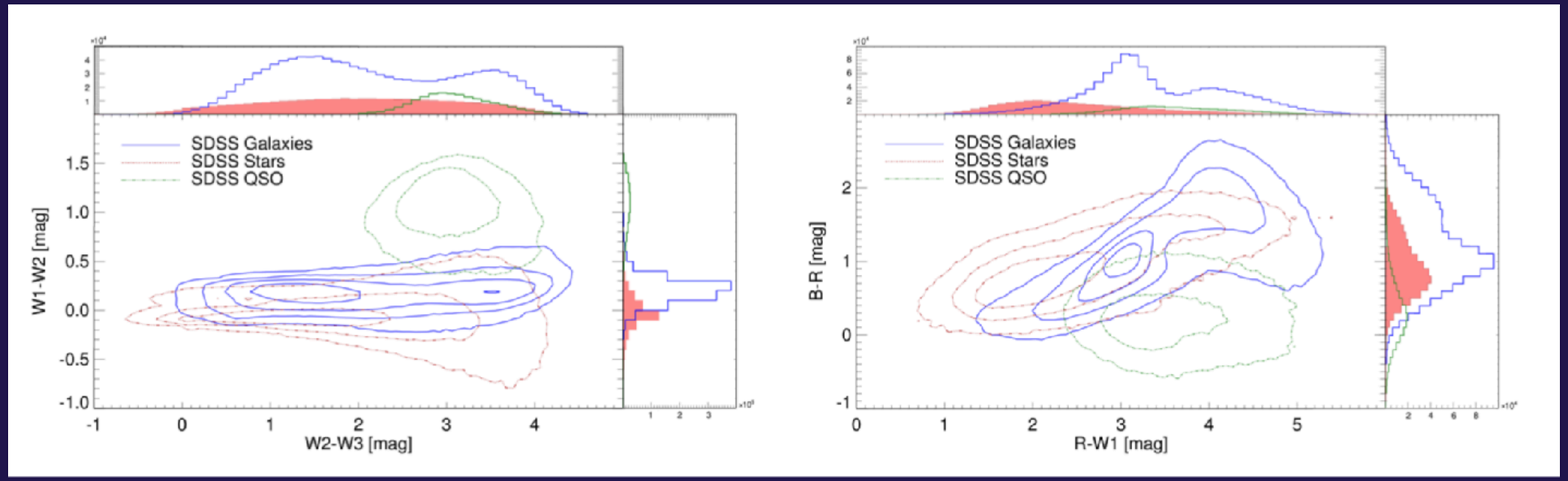
bogus



e.g. Brink et al (2012): Random Forests
Wright et al (2015): Support Vector Machines, Neural Networks, Random Forests
Masci et al (2016): Random Forests
Gieseke et al (2017): Convolutional Neural Networks
Sedaghat + Mahabal (2018): Convolutional Neural Networks

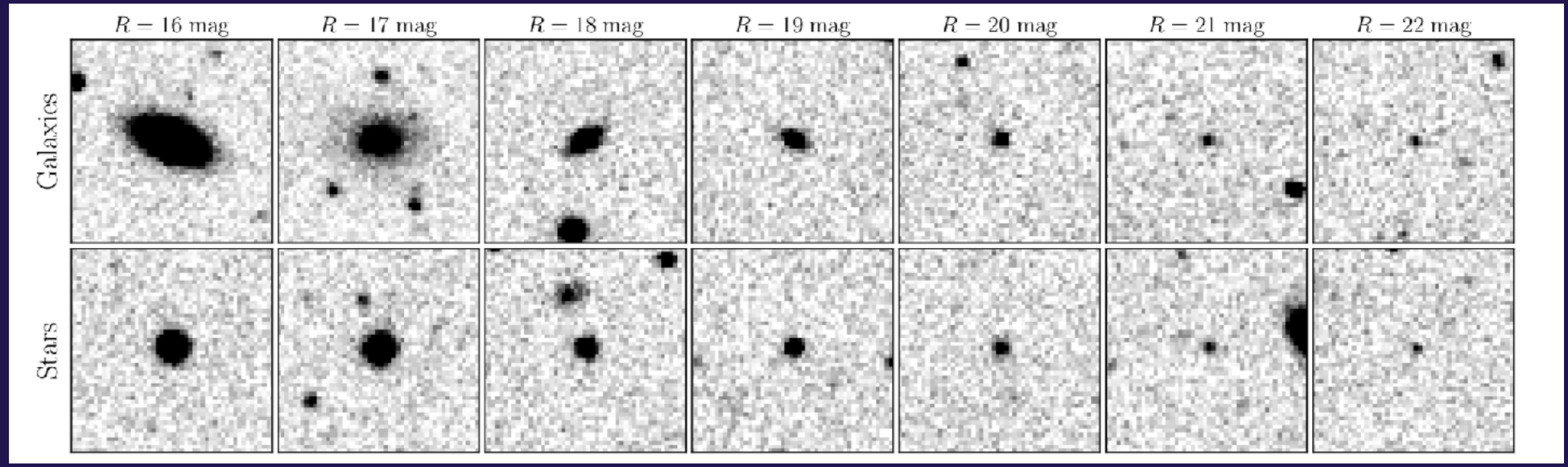
**many bogus versus few real events**

# Is it a star or a galaxy?




Krakowski et al (2016): support vector machines on SuperCOSMOS x WISE data

# Is it a star or a galaxy?



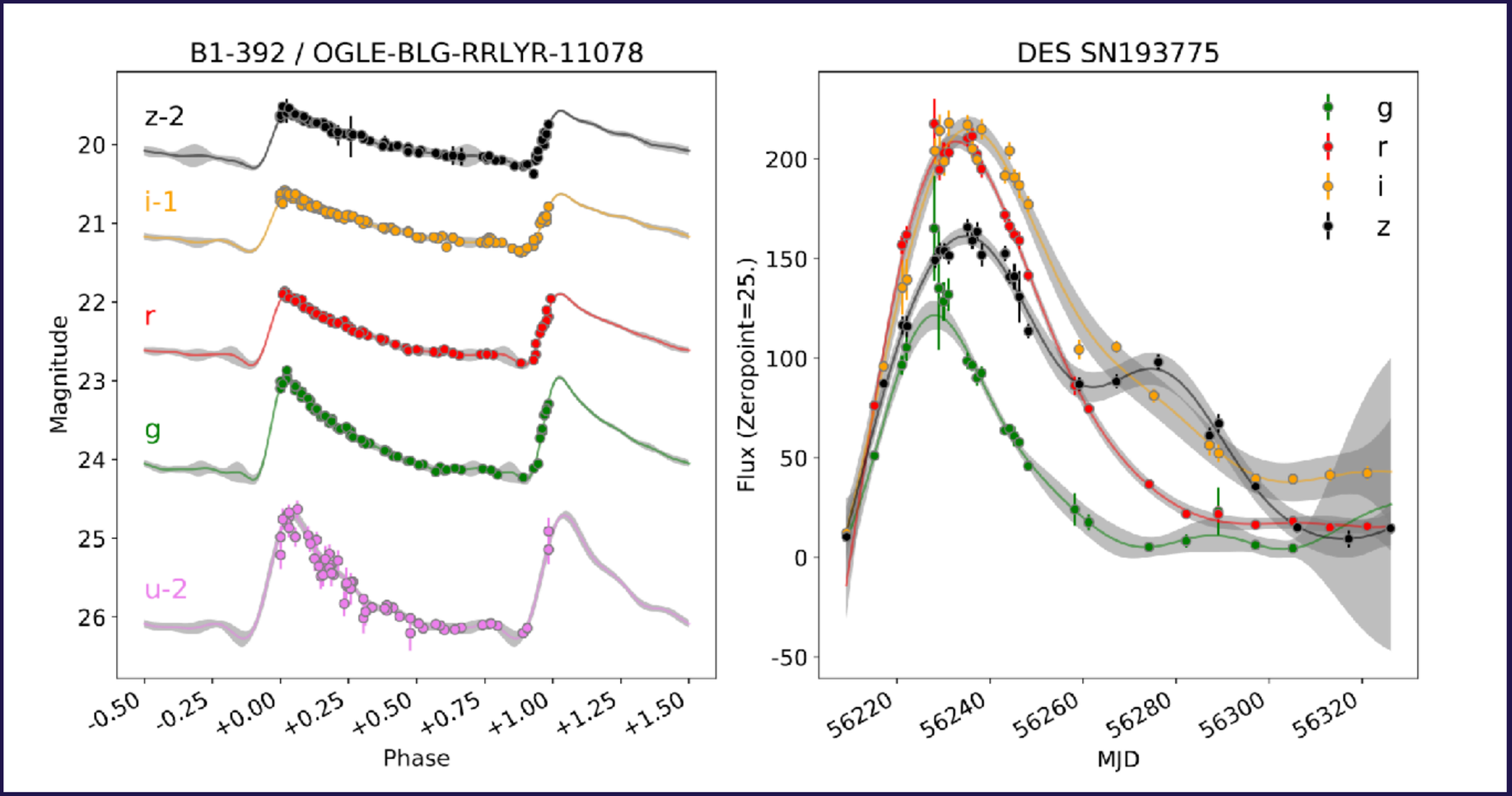

Miller et al (2017): Random Forests, iPTF

# think carefully about biases!

# follow up sources

e.g. “I want to follow up all the interesting supernovae with my spectrograph”

# follow up sources




Narayan et al (2018): Gaussian Process (ANTARES); see also Lochner et al (2016): Boosted Decision Trees

# find rare things

e.g. “I want to find supermassive binary black holes .”

# population studies

e.g. “I want to study all RR Lyrae stars in our galaxy to learn something about stellar evolution.”

# Supervised Learning

e.g. is this a star or a galaxy (classification)?
e.g. What’s the value of this photometric redshift (regression)?

# Supervised Learning

e.g. is this a star or a galaxy (classification)?
e.g. What's the value of this photometric redshift (regression)?

**requires ground-truth training data!**

# Unsupervised Learning

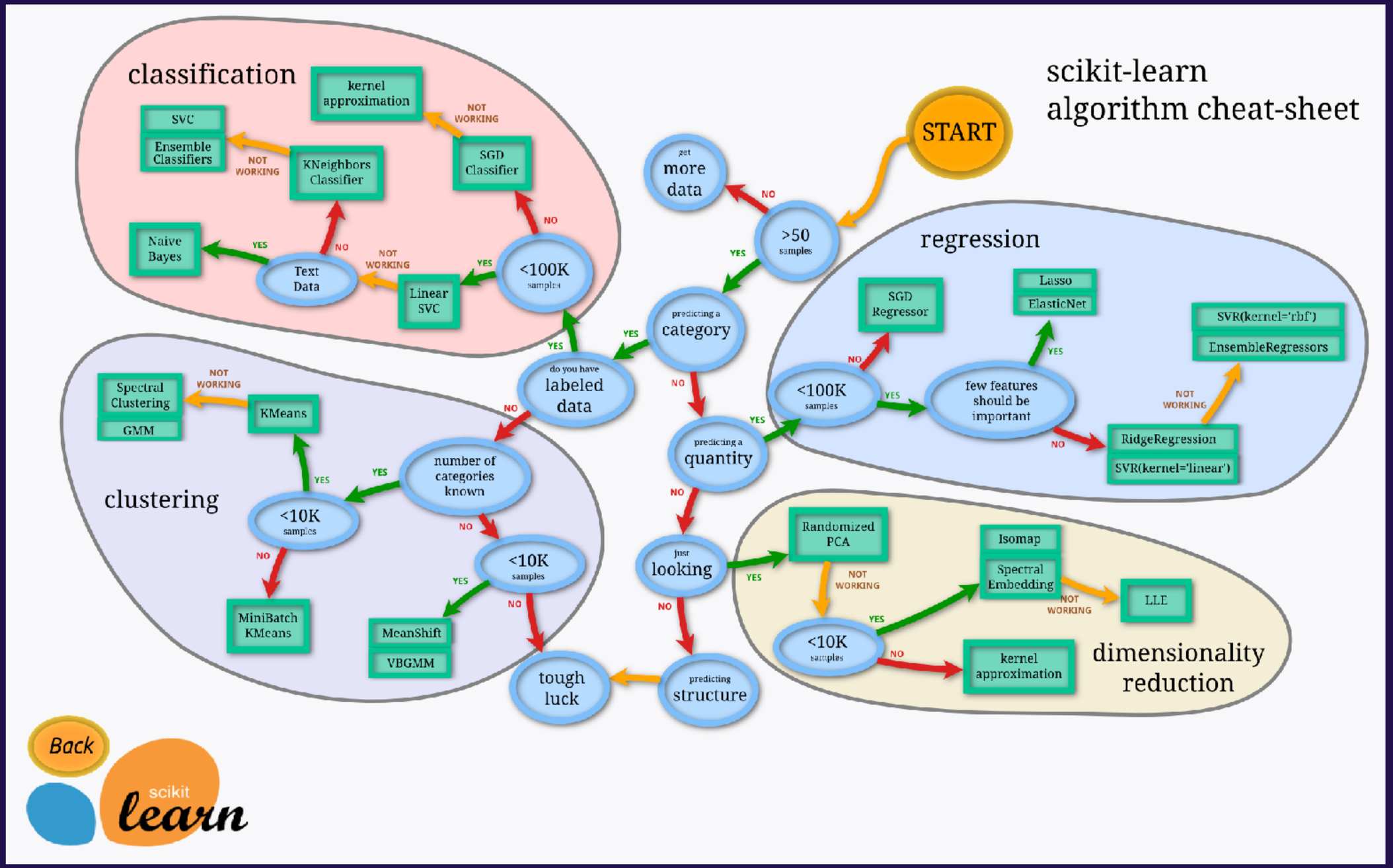
e.g. how many types of variable stars are there in my survey (clustering)?

# Unsupervised Learning

e.g. how many types of variable stars are there in my survey (clustering)?

**requires NO ground-truth training data, but often much more difficult to get to work!**

# Also: reinforcement learning, active learning, transfer learning …



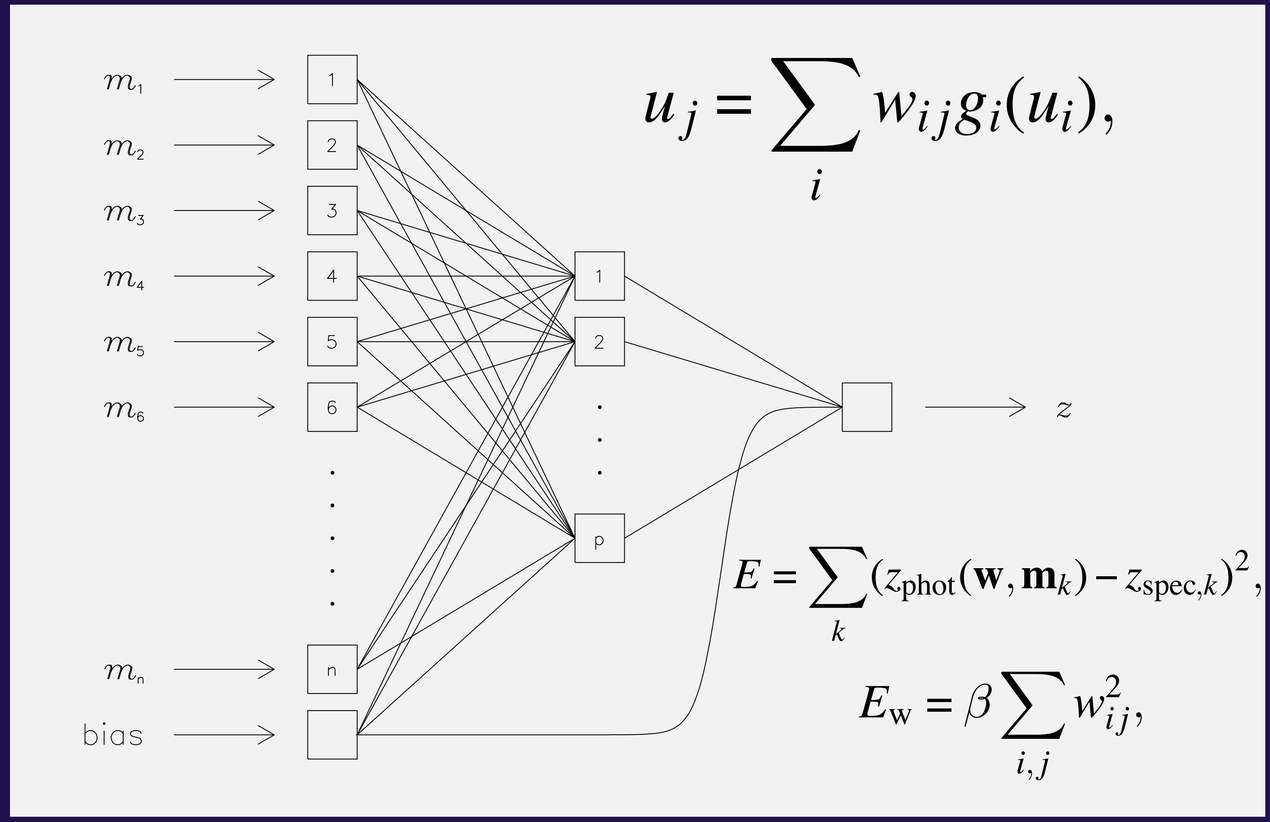

Credit: Andreas Müller

# Some terminology …

# linear model = model linear in the parameters

$$z = f(x) = w_0 + w_1 x_1 + \ldots + w_n x_n$$

linear model



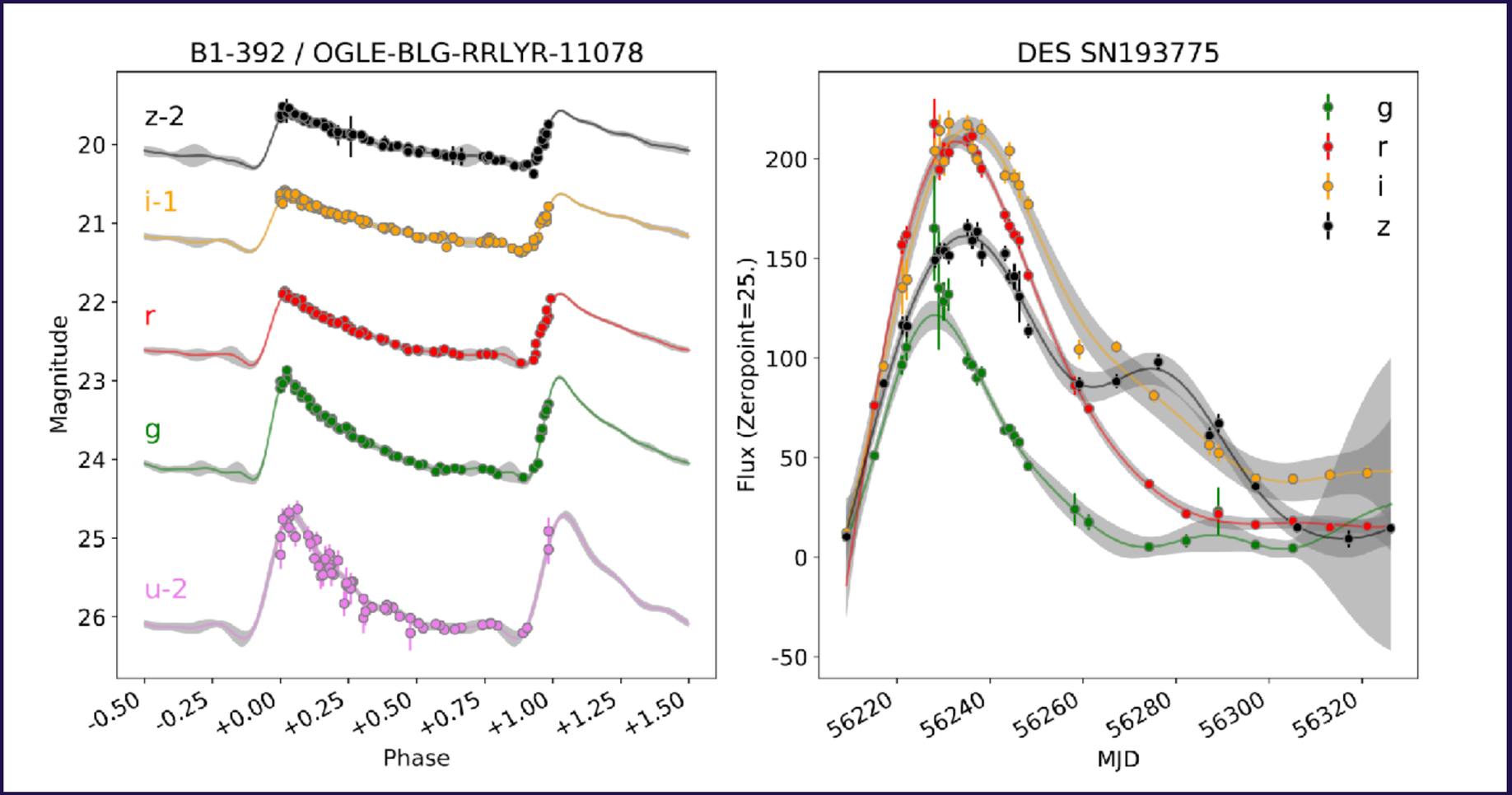

non-linear model

# Non-linear models are far more flexible!

# Non-linear models are far more flexible!

(but often also harder to interpret)

**generative model**: a model that can generate data




Narayan et al (2018): Gaussian Process (ANTARES); see also Lochner et al (2016): Boosted Decision Trees

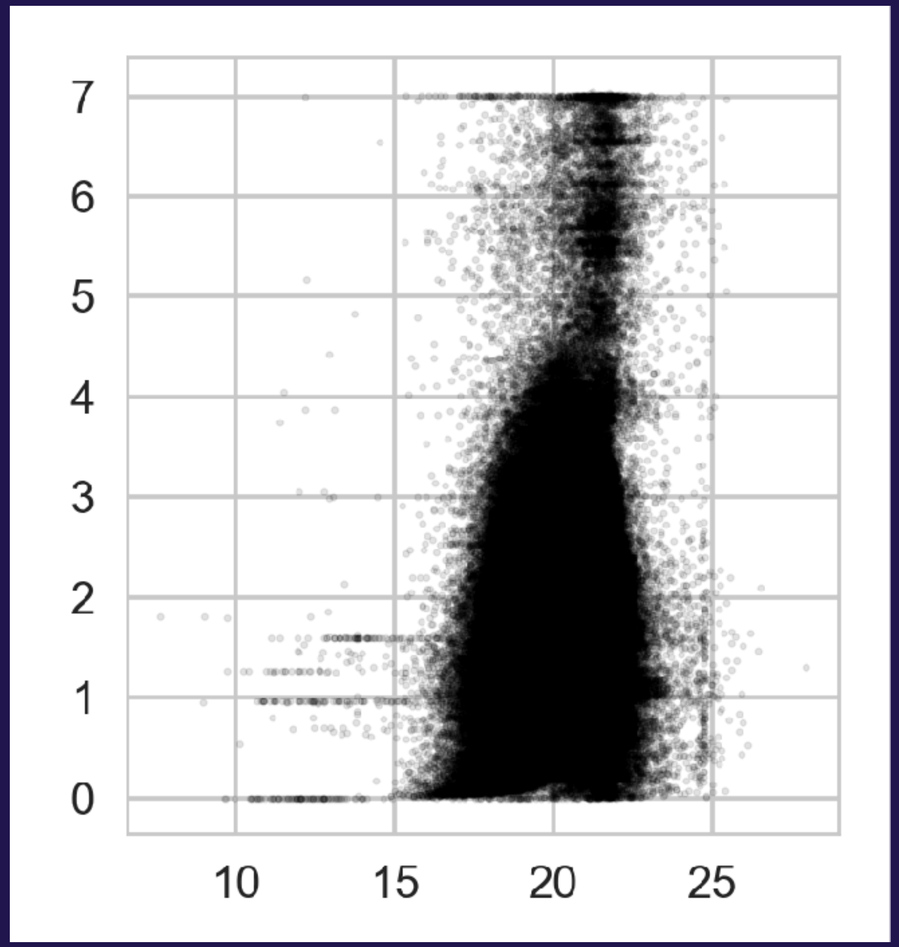
# How to get started …

# Advice: Start simple + supervised

# Define Your Problem

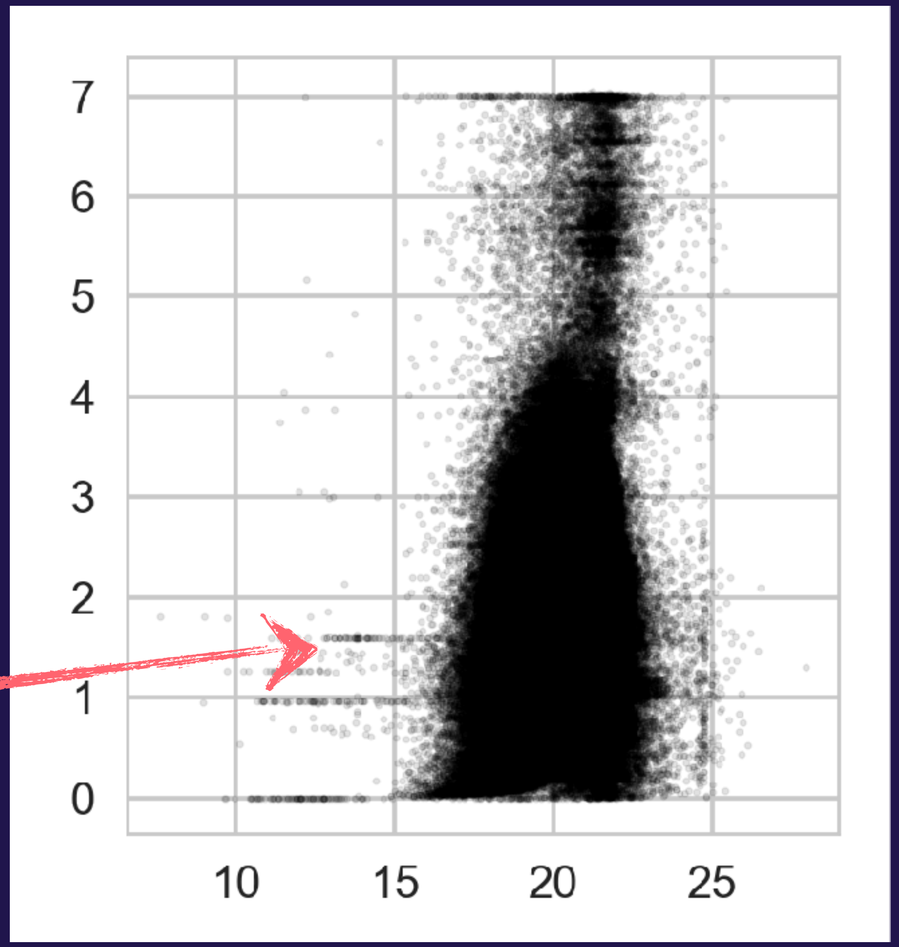
- what is the ultimate goal of your ML procedure?
- Prediction vs. inference vs. emulation?
- What will you do with the results?

# Preprocessing (1): Sanity-check your data!



# Preprocessing (1): Sanity-check your data!

???

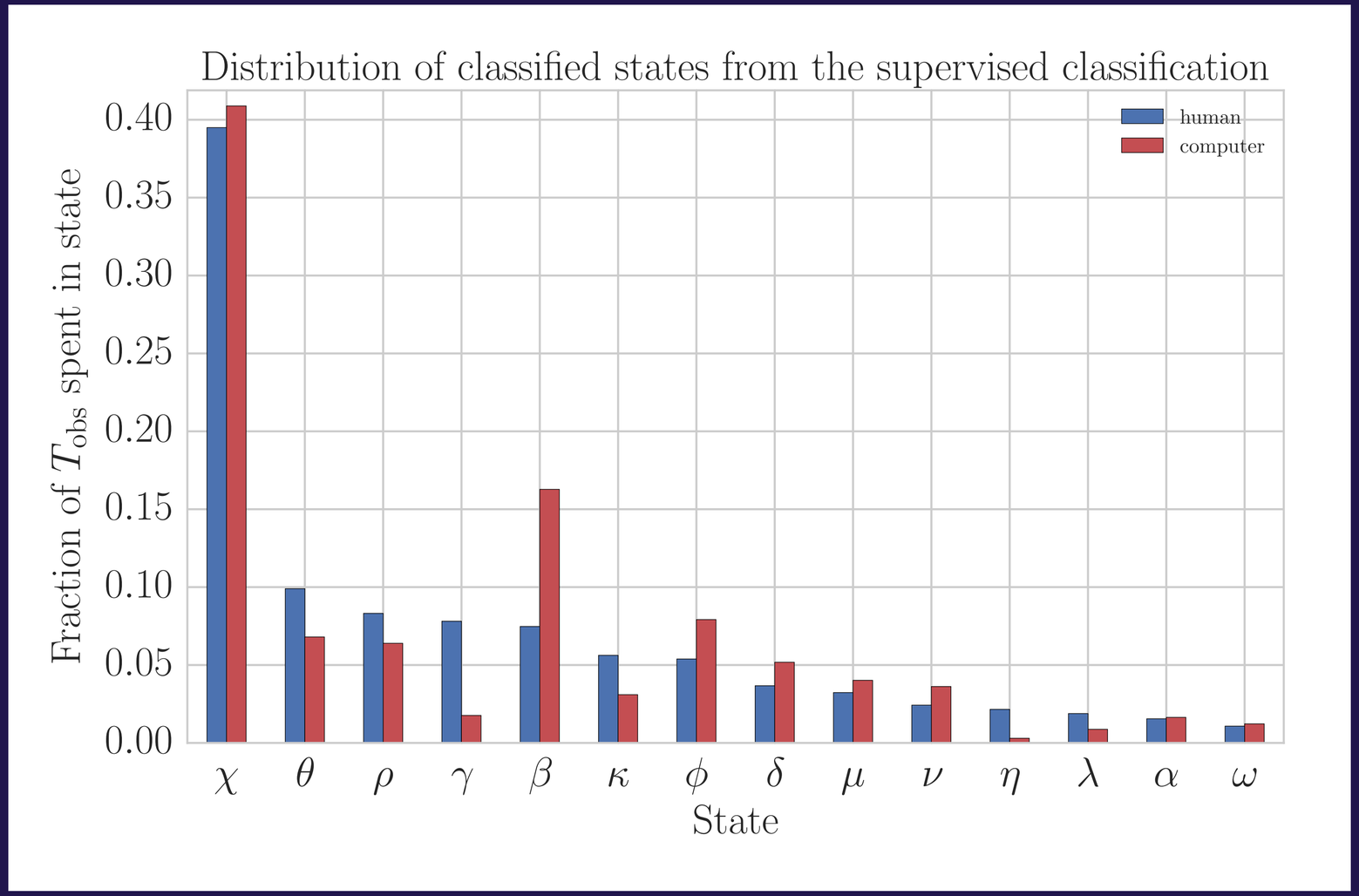


# Preprocessing (2): The Art of Squishing Data Into Vectors

- e.g. one-hot encoding
- pooling things into the same vectors
- feature engineering via summary statistics

**Use your domain knowledge!**

# Preprocessing (3): Dealing with Imbalanced Data Sets

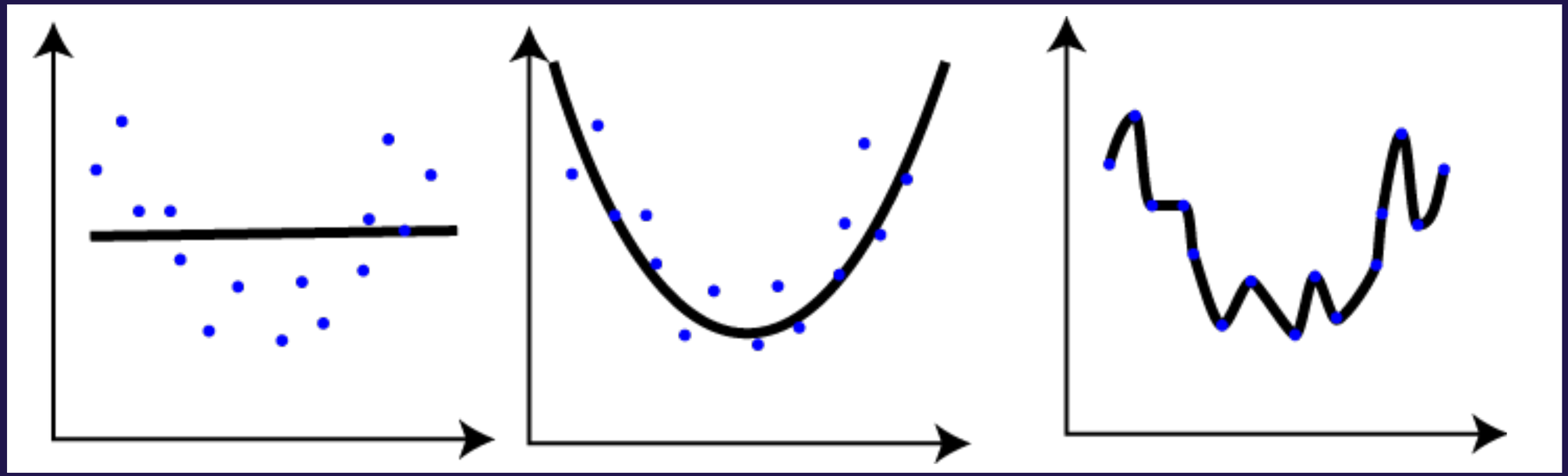



Huppenkothen et al, 2017

# Machine Learning (1): Pick a Simple Supervised Algorithm

- test your feature engineering
- figure out which features are important
- figure out which scoring function might work for your problem
- figure out what kind of biases/systematic trends there are in your data

# Machine Learning (2): Model Selection + Hyper-parameter Estimation



1) avoid overfitting (prediction)
2) decide between (physical) models (inference)

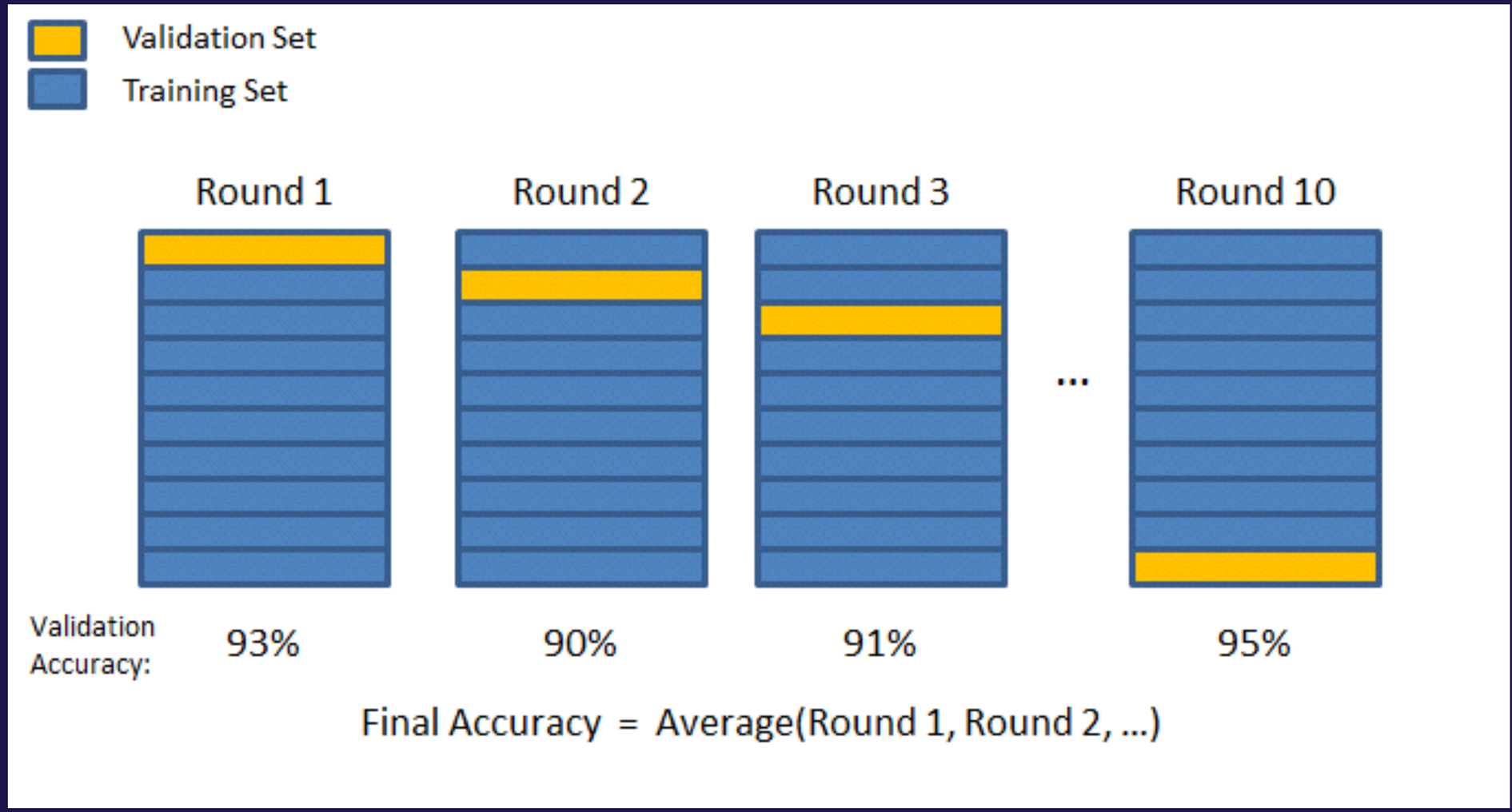
# Nomenclature

**training set**: a data set to train your algorithm on

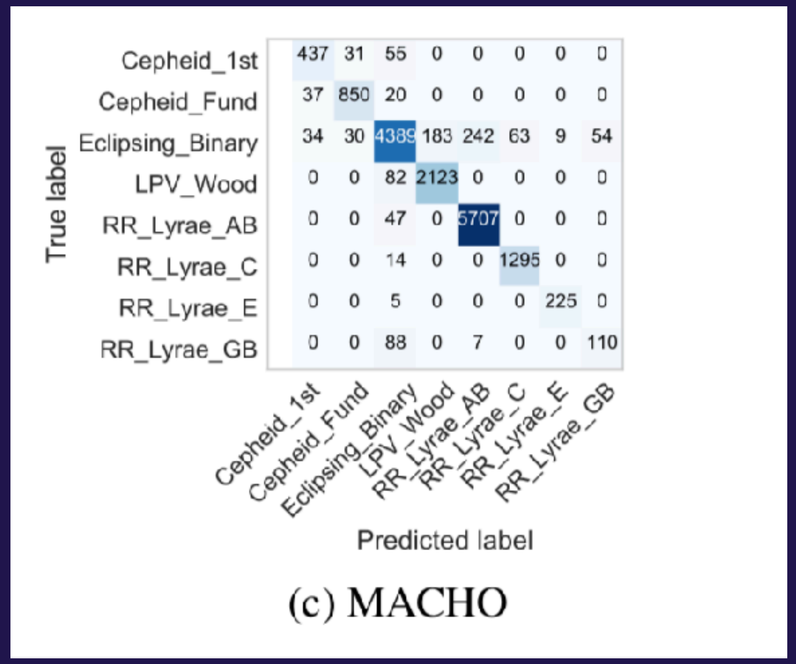
**validation set**: a data set to use for comparing the performance of different models

**test set**: a data set reserved to compute the error estimate of the final chosen model

# Hold-out + k-fold cross-validation



# Machine Learning (2): Critique Your Analysis




Naul et al (2017)

# … repeat (and get more complex/unsupervised if it’s needed)

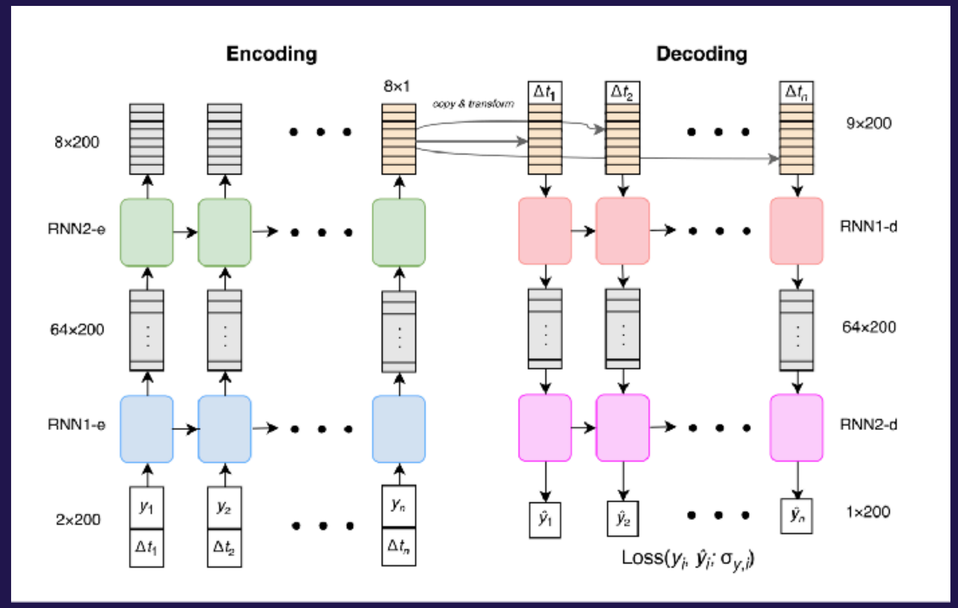
# Pitfalls + Questions

# Features or No Features, That Is The Question!

| Feature | Description |
|---|---|
| amplitude | Half the difference between the maximum and the minimum magnitude |
| beyond1std | Percentage of points beyond one st. dev. from the weighted mean |
| flux_percentile_ratio_mid20 | Ratio of flux percentiles (60th - 40th) over (95th - 5th) |
| flux_percentile_ratio_mid35 | Ratio of flux percentiles (67.5th - 32.5th) over (95th - 5th) |
| flux_percentile_ratio_mid50 | Ratio of flux percentiles (75th - 25th) over (95th - 5th) |
| flux_percentile_ratio_mid65 | Ratio of flux percentiles (82.5th - 17.5th) over (95th - 5th) |
| flux_percentile_ratio_mid80 | Ratio of flux percentiles (90th - 10th) over (95th - 5th) |
| linear_trend | Slope of a linear fit to the light curve fluxes |
| max_slope | Maximum absolute flux slope between two consecutive observations |
| median_absolute_deviation | Median discrepancy of the fluxes from the median flux |
| median_buffer_range_percentage | Percentage of fluxes within 20% of the amplitude from the median |
| pair_slope_trend | Percentage of all pairs of consecutive flux measurements that have positive slope |
| percent_amplitude | Largest percentage difference between either the max or min magnitude and the median |
| percent_difference_flux_percentile | Diff. between the 2nd & 98th flux percentiles, converted to magnitude[a] |
| QSO | Quasar variability metric in Butler & Bloom (2010) |
| non_QSO | Non-quasar variability metric in Butler & Bloom (2010) |
| skew | Skew of the fluxes |
| small_kurtosis | Kurtosis of the fluxes, reliable down to a small number of epochs |
| std | Standard deviation of the fluxes |
| stetson_j | Welch-Stetson variability index J[b] |
| stetson_k | Welch-Stetson variability index K[b] |

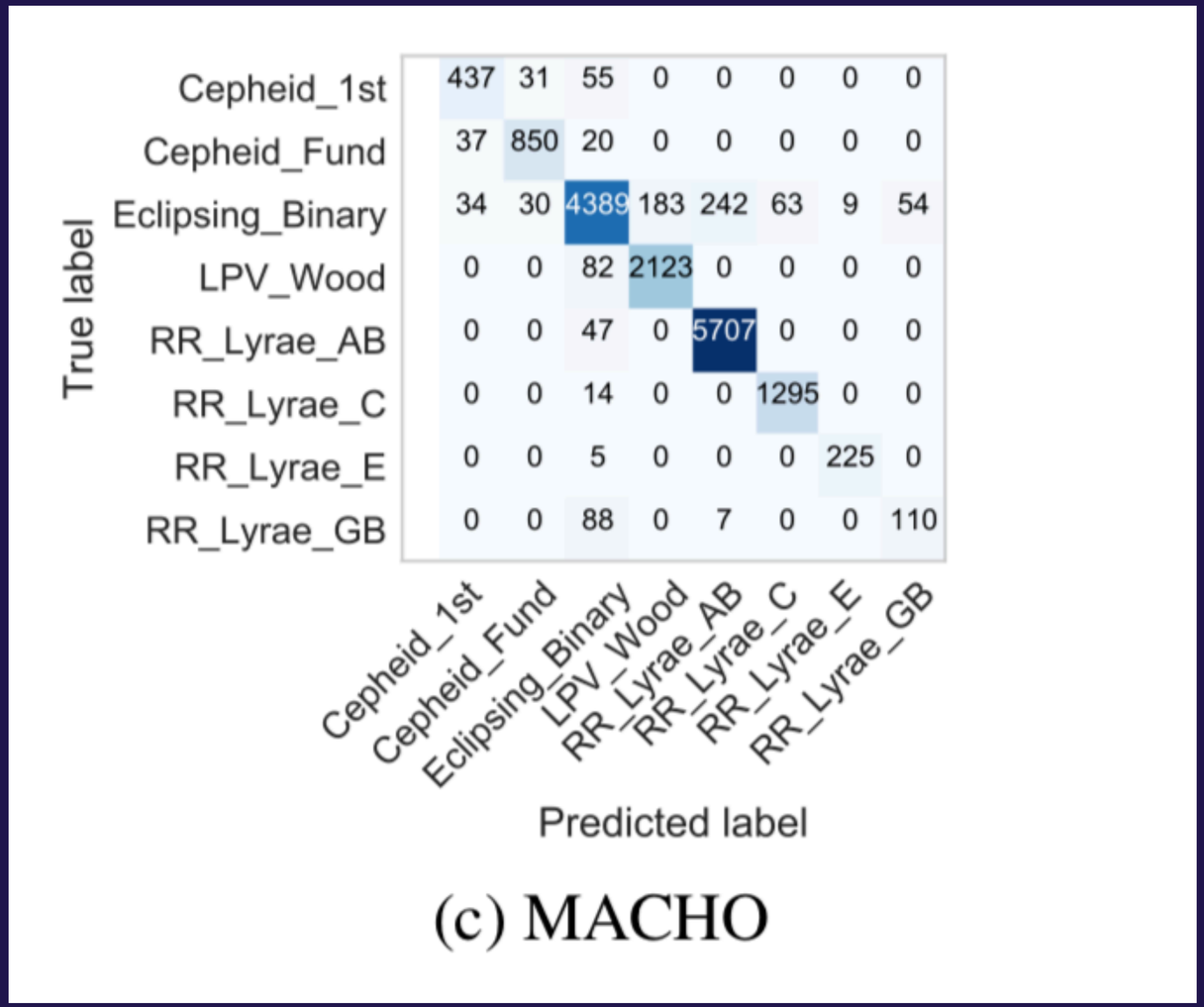
a Ivezic (2005)

Richards et al (2011)




Naul et al (2017)

# Imbalanced Classes




Naul et al (2017)

# Finding Training Data Can Be Hard!

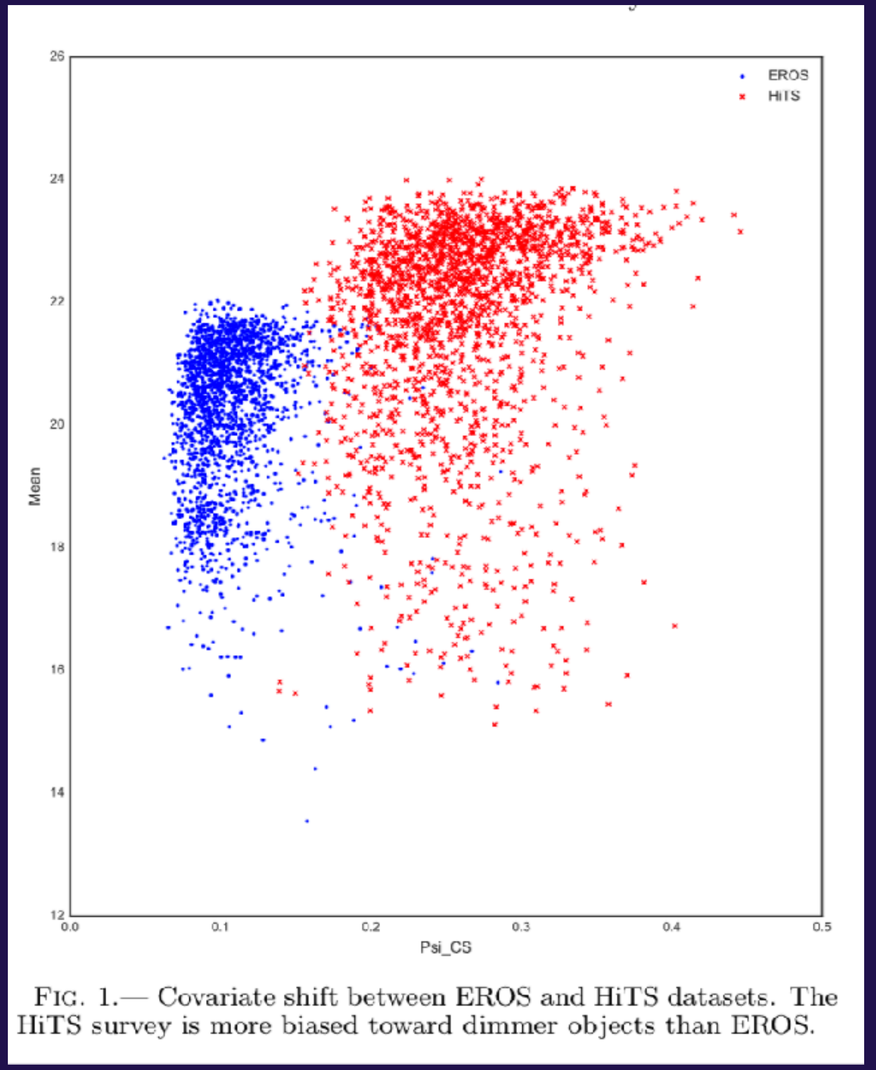
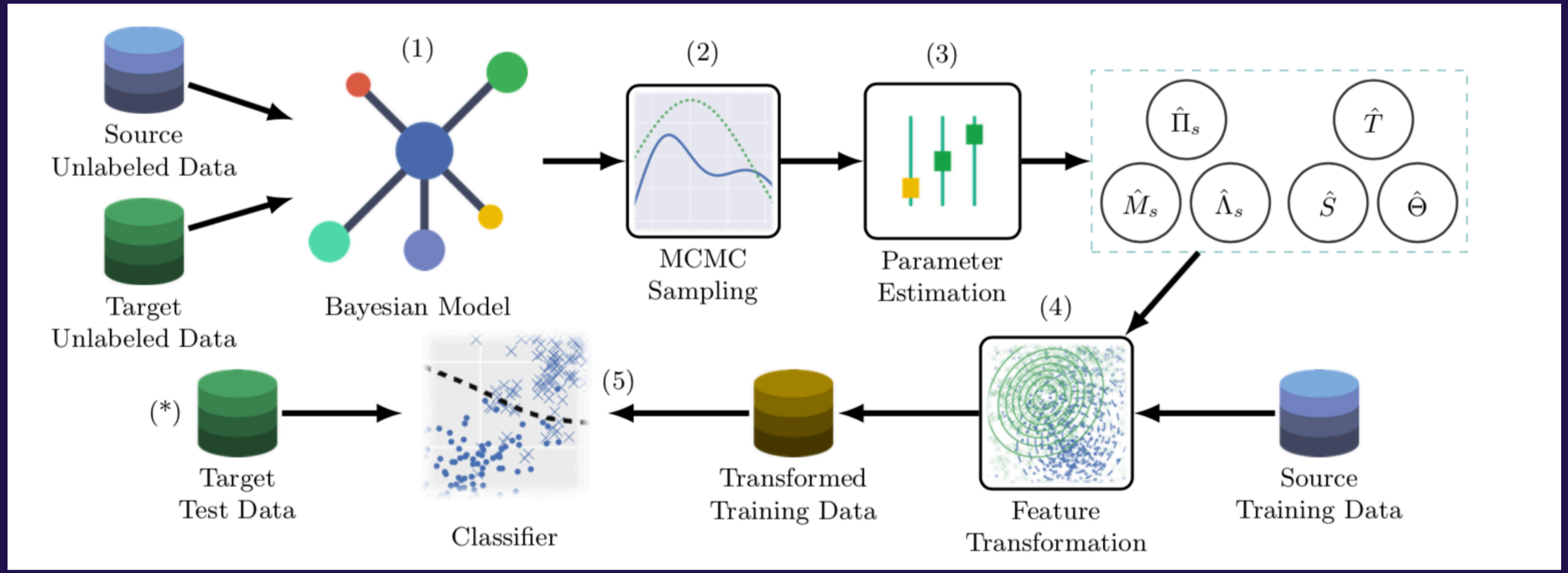



FIG. 1.— Covariate shift between EROS and HiTS datasets. The HiTS survey is more biased toward dimmer objects than EROS.

Benavente et al (2017)

# Domain Adaptation




Benavente et al (2017)

# Biased Training Data

# GALAXY ZOO: MORPHOLOGICAL CLASSIFICATION AND CITIZEN SCIENCE

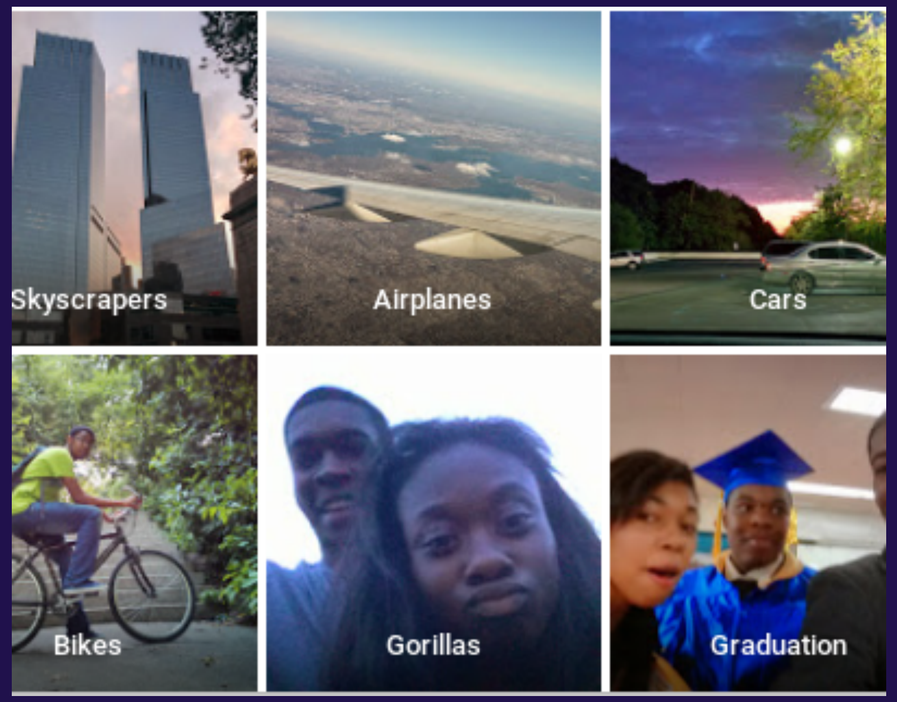
L. Fortson[1], K. Masters[2], R. Nichol[2], K. Borne[3], E. Edmondson[2], C. Lintott[4,5], J. Raddick[6], K. Schawinski[7], J. Wallin[8]

*to be published in Advances in Machine Learning and Data Mining for Astronomy*

The results of the mirror image bias testing are discussed extensively in Land et al. (2008). They showed a significant bias in favour of anti-clockwise direction arms (in both the original and mirrored images). The interpretation of this bias could be due to psychological effects (possibly related to the preference for right handedness amongst the population), or possibly site design (it being easier to click the anti-clockwise button for example). However, once this bias was corrected for, the data could still be used (see below).

Most problems with machine learning outcomes are either due to a biased training set or a faulty interpretation

# Example 1: Biased Training Data

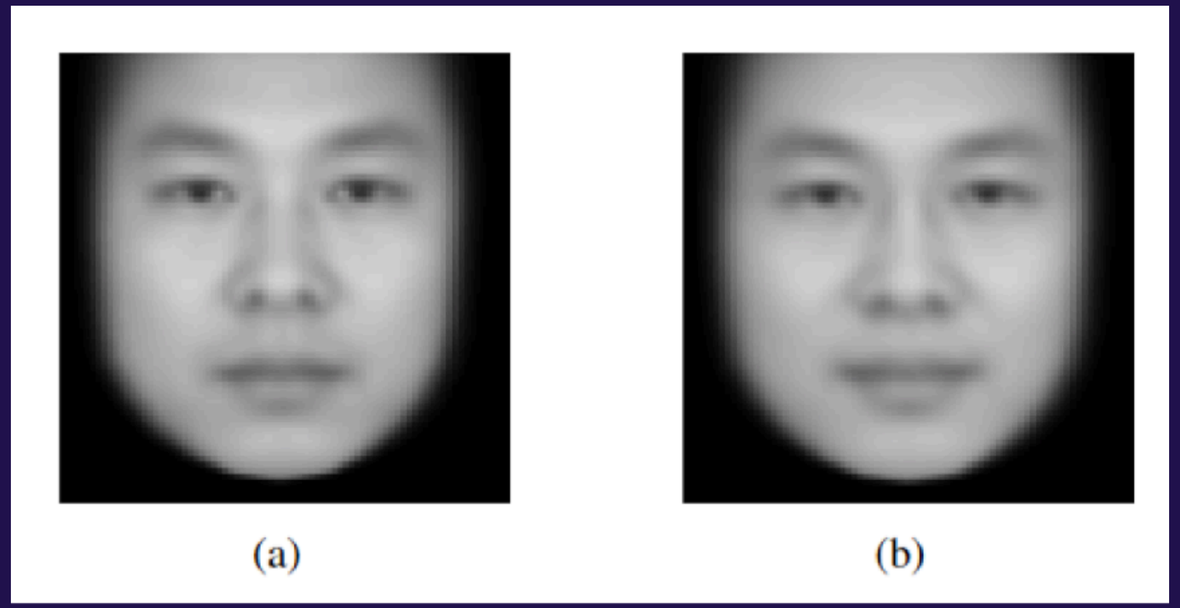



**ImageNet: Out of ~500 images of humans, only 2 depict African Americans**

http://approximatelycorrect.com/2016/11/07/the-foundations-of-algorithmic-bias/

# Example 2: Interpretation

**Wu + Zhang (2016): Automated Inference on Criminality using Face Images**




https://callingbullshit.org/case_studies/case_study_ml_sexual_orientation.html

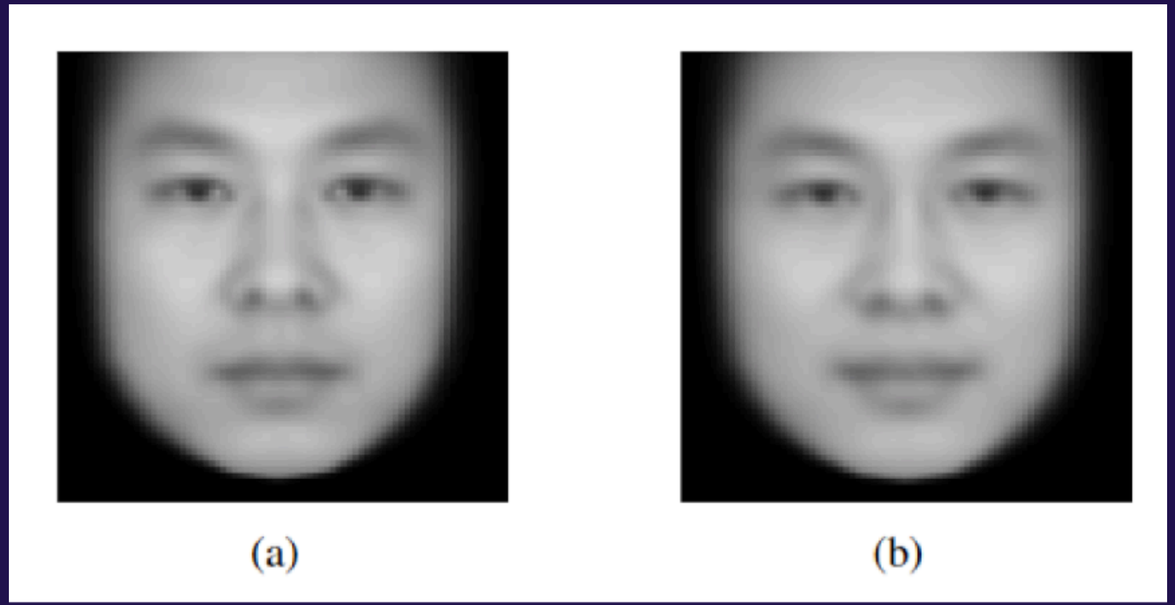
https://callingbullshit.org/case_studies/case_study_ml_sexual_orientation.html

## Wu + Zhang (2016): Automated Inference on Criminality using Face Images

> “Unlike a human examiner/judge, a computer vision algorithm or classifier has absolutely no subjective baggages [sic], having no emotions, no biases whatsoever due to past experience, race, religion, political doctrine, gender, age, etc., no mental fatigue, no preconditioning of a bad sleep or meal. The automated inference on criminality eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together.”

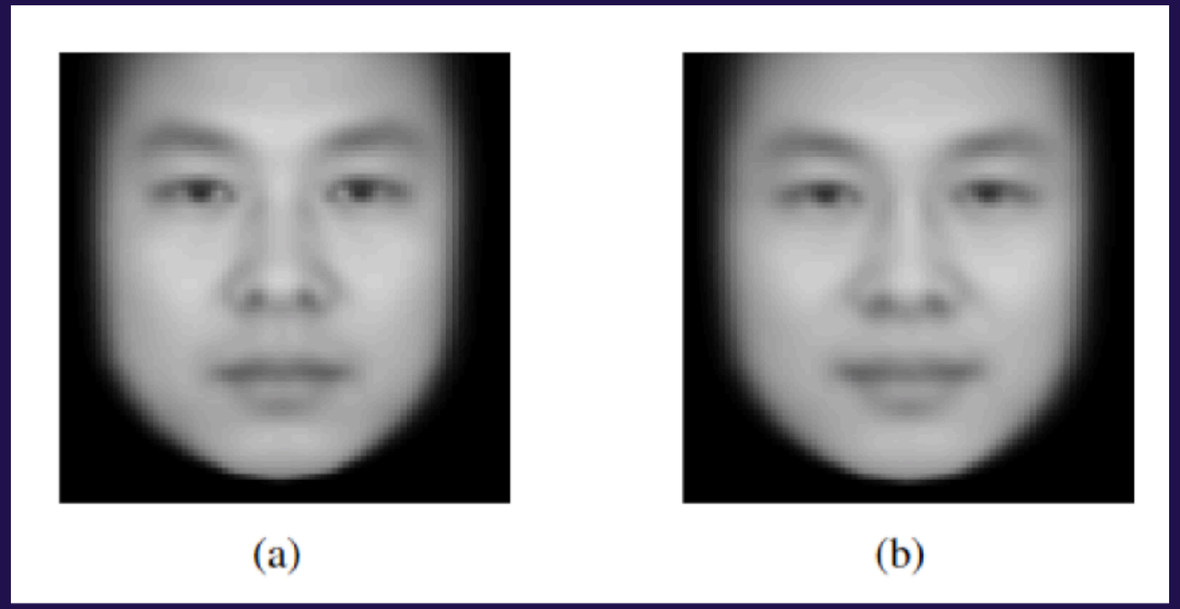
https://callingbullshit.org/case_studies/case_study_ml_sexual_orientation.html

# Wu + Zhang (2016): Automated Inference on Criminality using Face Images



https://callingbullshit.org/case_studies/case_study_ml_sexual_orientation.html

## Wu + Zhang (2016): Automated Inference on Criminality using Face Images



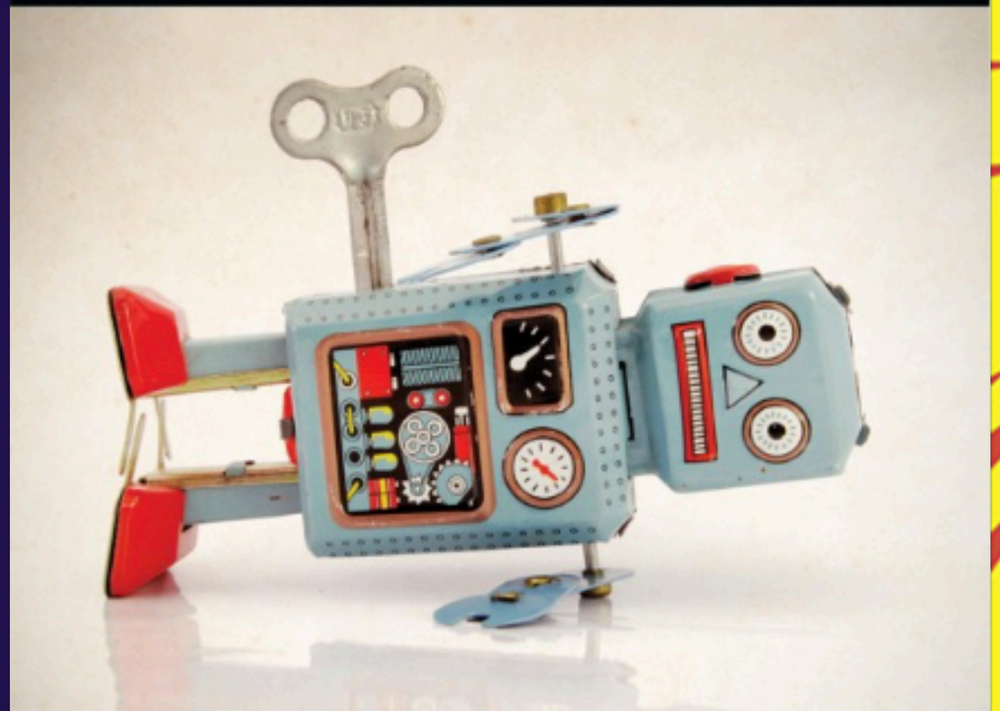

**biased training data: algorithm reproduces facial expression, not facial features**

DiRAC

Meredith Broussard

# Artificial Unintelligence

HOW COMPUTERS MISUNDERSTAND THE WORLD



# WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

# Conclusions

- machine learning presents us with new and important tools to make sense of large survey data sets
- astronomy data is idiosyncratic: out-of-the-box solutions will likely not work
- start building your tools now
- inference with (traditional) machine learning models is hard, but new developments can help!
- other fields are developing relevant methods, but communication is difficult

Tiana_Athriel

dhuppenk@uw.edu

# And now it’s your turn!

1) Group Assignment
2) Your mission statement: go to: https://github.com/dhuppenkothen/cargese2018_tutorials/blob/master/tutorial2/EXERCISE.md
3) Get started (like yesterday):
    1) your laptop
    2) Binder
    3) Colab

Click to add text

# Click to add title

Click to add subtitle

# Click to add title

Click to add subtitle

# Click to add title

Click to add text

Click to add text

# Click to add title

Click to add text

Click to add text

# Click to add title

Click to add text

# Click to add title


# Click to add title

Click to add subtitle

Click to add text

# Click to add title

Click to add subtitle

Click to add text


Click to add text

# xx %

# xx %

# Click to add title

# Click to add title

# Click to add title

Click to add subtitle