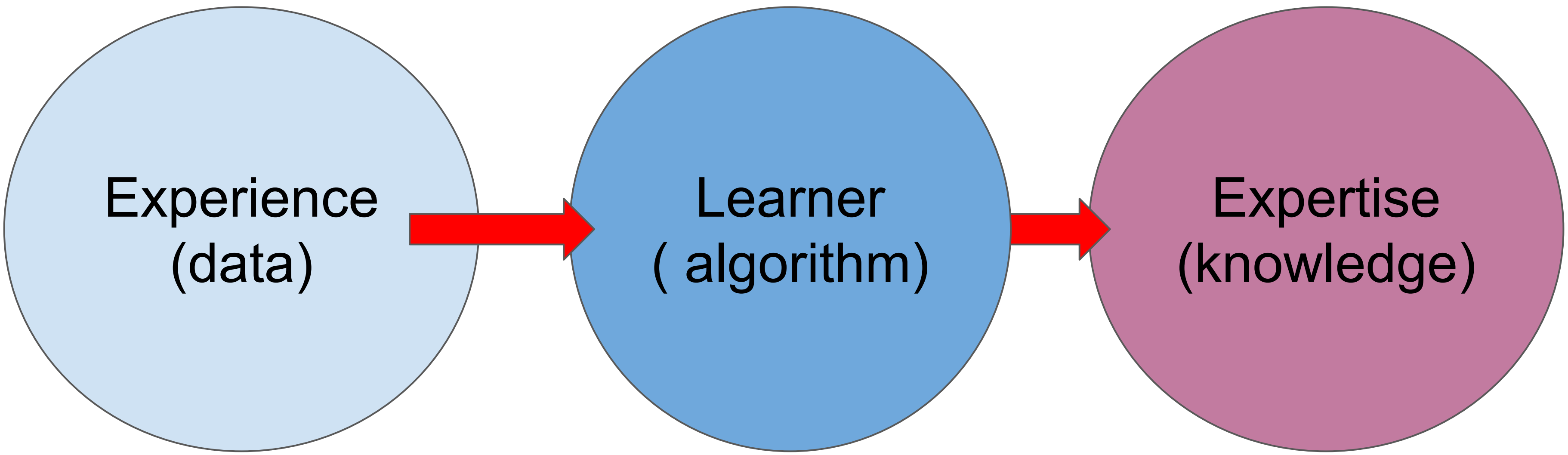
Machine Learning in Astronomy

LPNHE, Paris - 10 February 2020

Emille E. O. Ishida
Laboratoire de Physique de Clermont - Université Clermont-Auvergne
Clermont Ferrand, France

What is learning?

"Learning is the process of converting experience into expertise"



Shalev-Shwartz, S. and Ben-David, S., Understanding Machine Learning - from theory to algorithms, 2014, Cambridge University Press

Why do we need machines to learn?

- Tasks there are too complex to explicitly program
- Tasks dealing with too large data volumes
- Tasks which require flexibility

Shalev-Shwartz, S. and Ben-David, S., Understanding Machine Learning - from theory to algorithms, 2014, Cambridge University Press

Why do we need machines to learn?

- Tasks there are too complex to explicitly program

- Tasks dealing with too large data volumes

- Tasks which require flexibility

Shalev-Shwartz, S. and Ben-David, S., Understanding Machine Learning - from theory to algorithms, 2014, Cambridge University Press

Supervised Learning

Learn by example

Learn

Training sample

Features
+
Labels

Apply

Target sample

Feature
s

Machine Learning:

(a personal favorite)

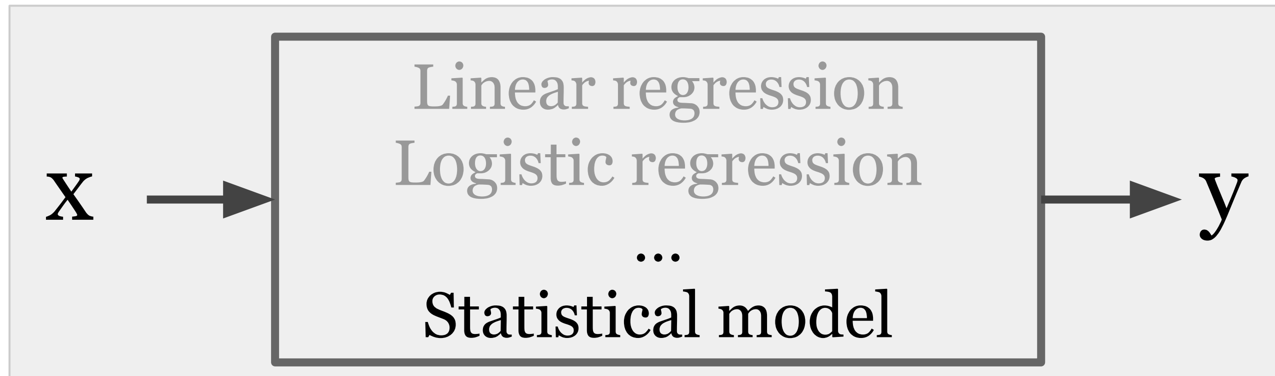
Supervised definition

Hypothesis:

Breiman, L., *Statistical Modeling: The Two Cultures*, Stat. Sci, Volume 16 (2001)

Hypothesis:

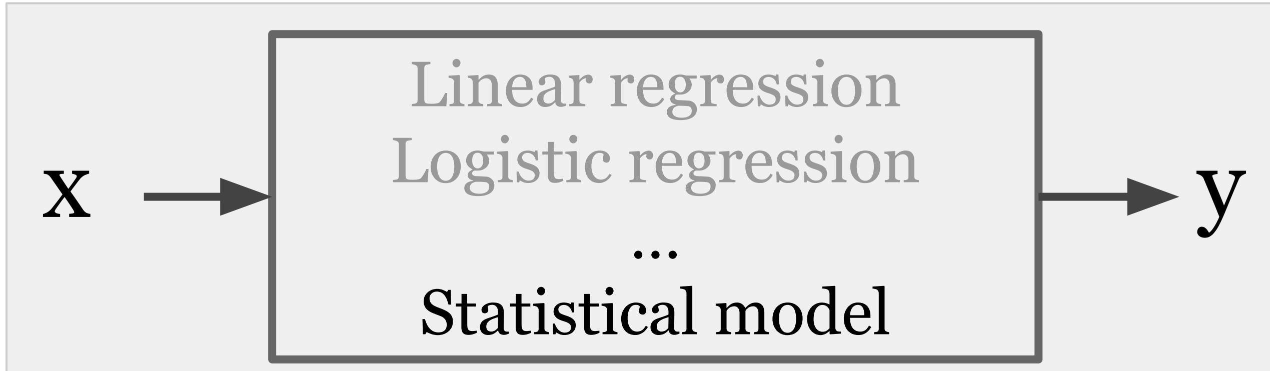
Physical modeling:



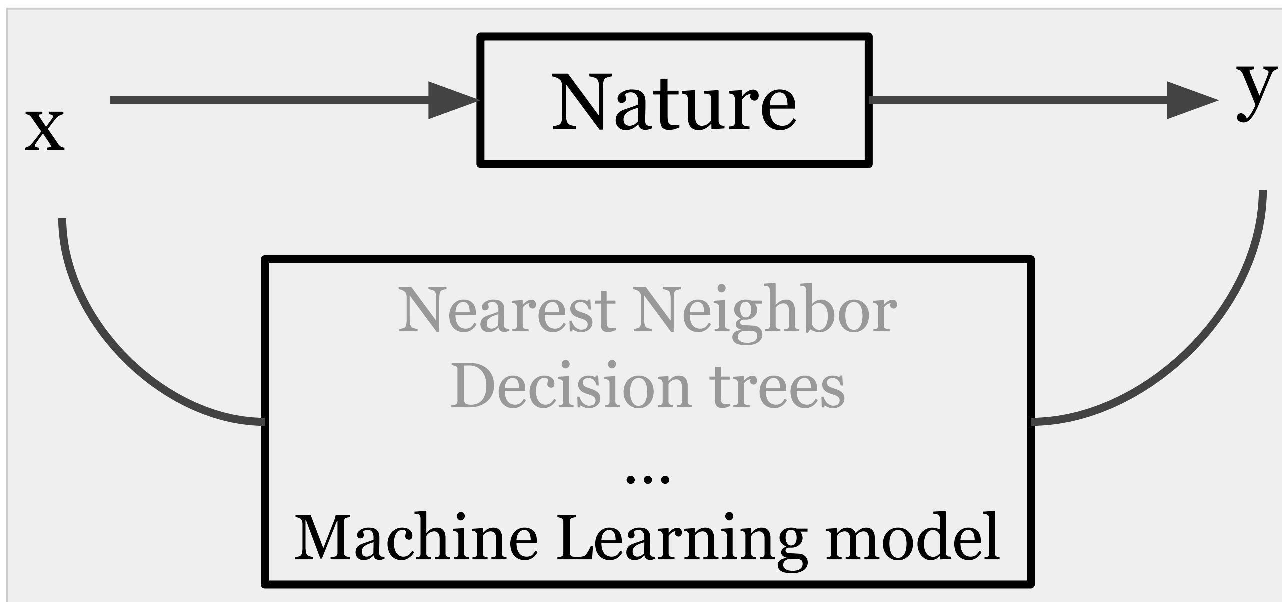
Breiman, L., *Statistical Modeling: The Two Cultures*, Stat. Sci, Volume 16 (2001)

Hypothesis:

Physical modeling:



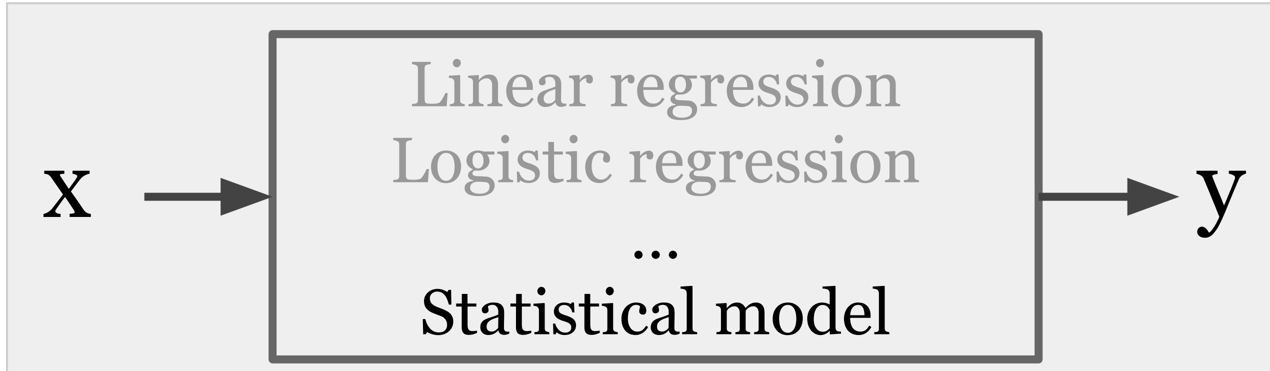
Algorithmic modeling:



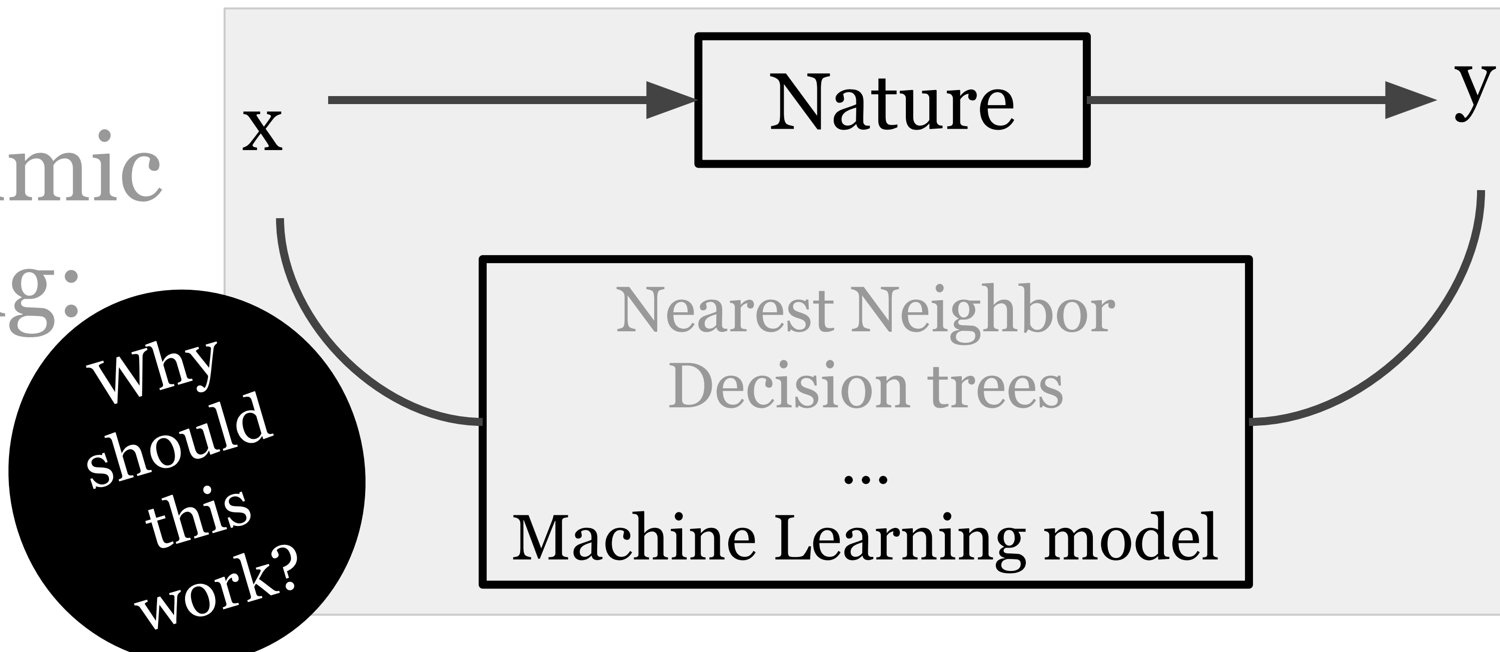
Breiman, L., *Statistical Modeling: The Two Cultures*, Stat. Sci, Volume 16 (2001)

Hypothesis:

Physical modeling:



Algorithmic modeling:

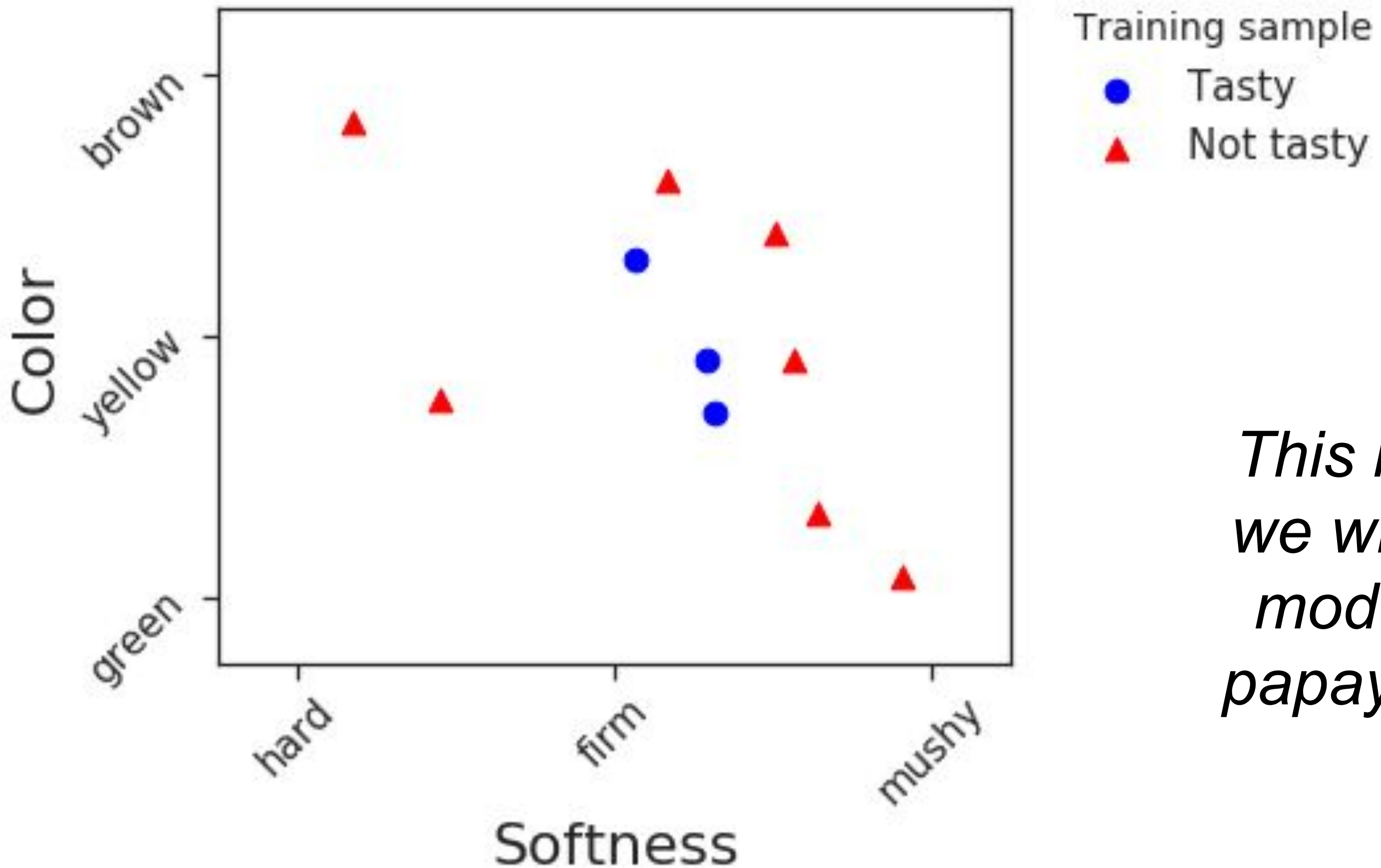


Breiman, L., *Statistical Modeling: The Two Cultures*, Stat. Sci, Volume 16 (2001)

A controlled example:

Papaya tasting

Binary classification

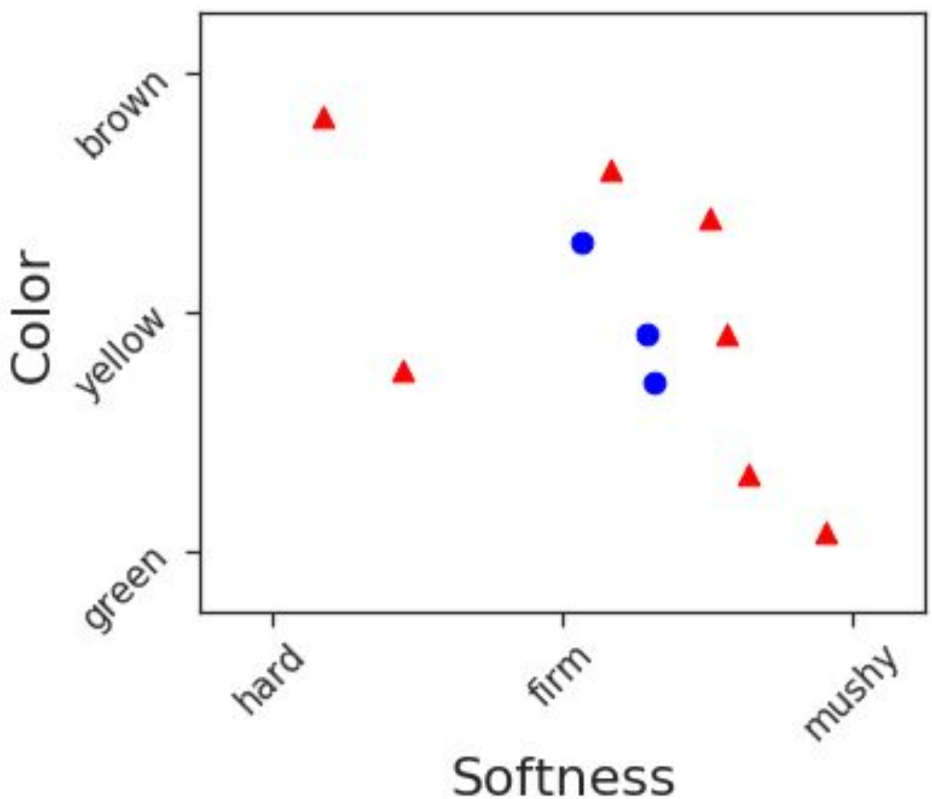


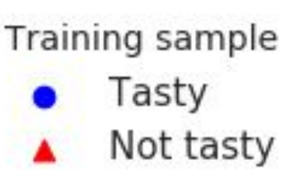
This is all the data we will input to the model about the papayas in the real world!

YouTube class on the papaya testing example:
https://www.youtube.com/watch?v=b5NlRg8SjZg&list=PLPW2keNyw-usgvmR7FTQ3ZRjfLs5jT4BO&index=2&t=0s

A controlled example:

Papaya tasting





X: set of all features,

x = [softness, color]

Y: set of possible labels,

y = [tasty, not tasty]

D: data generation model,

$D \Longrightarrow P(X)$

True Labelling function: y = f(x)

S: training sample: $[x_i, y_i]$, $i \in$ *training*

m: number of objects for training

h_S learner: $y_{est;i} = h_S(x_i)$

L metric: $L\,(y_{true;i} - y_{est;i})$, $i \in$ *training*

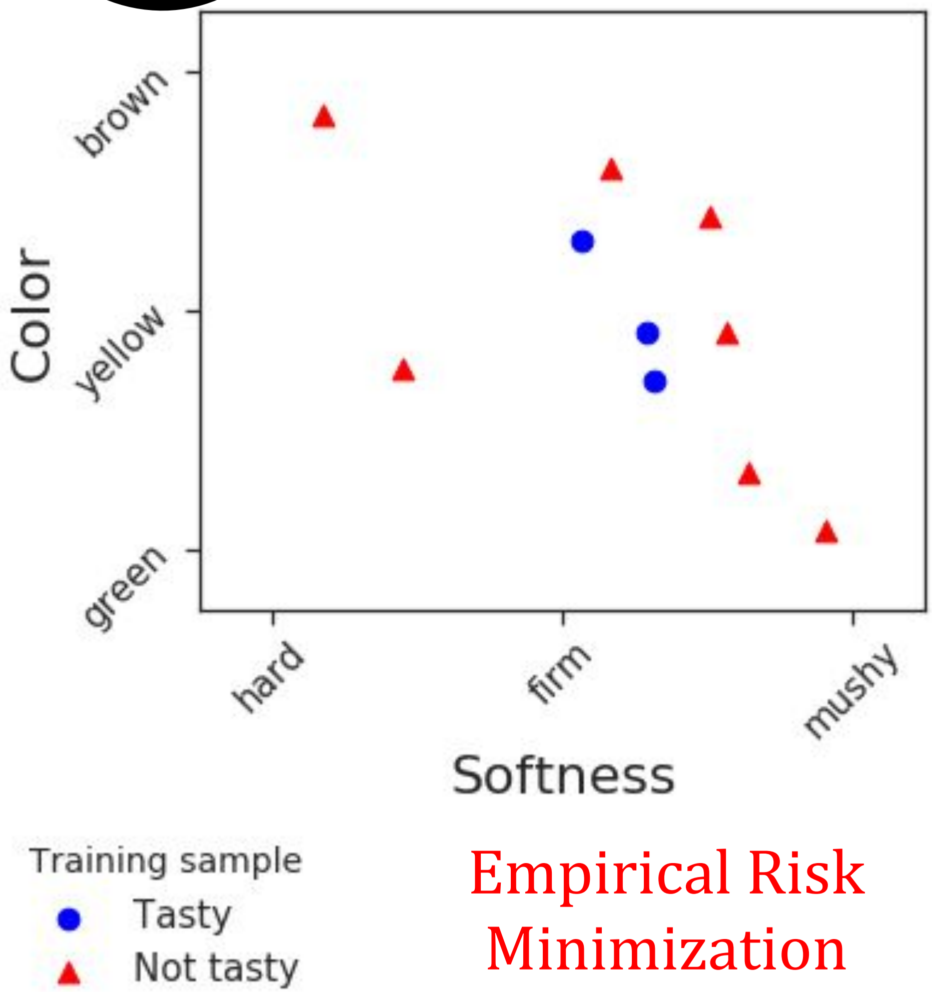
Empirical Risk Minimization (ERM)

$$L_{\mathcal{D}}(h_S) = \frac{|\{x \in \mathcal{D} : h_S(x) \neq f(x)\}|}{m}$$

A c ed example:

Why should this work?

Papaya tasting



X: set of all features,

x = [softness, color]

Y: set of possible labels,

y = [tasty, not tasty]

D: data generation model,

$D \Longrightarrow P(X)$

True Labelling function: $y = f(x)$

S: training sample: $[x_i, y_i]$, $i \in$ *training*

m: number of objects for training

h_S learner: $y_{est;i} = h_S(x_i)$

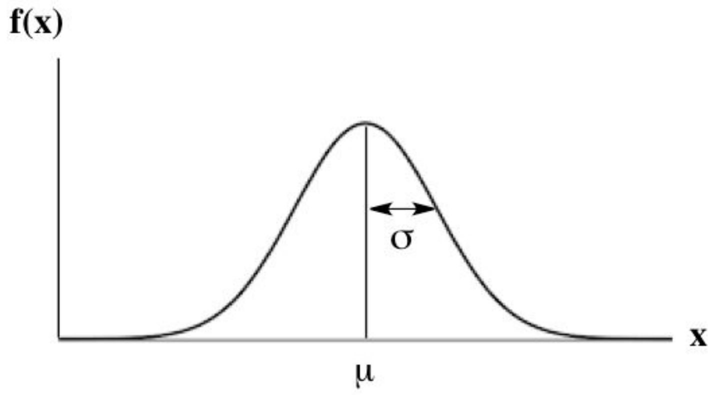
L metric: $L\,(y_{true;i} - y_{est;i})$, $i \in$ *training*

Empirical Risk Minimization (ERM)

$$L_{\mathcal{D}}(h_S) = \frac{|\{x \in \mathcal{D} : h_S(x) \neq f(x)\}|}{m}$$

Representativeness

Probability distribution, *P*



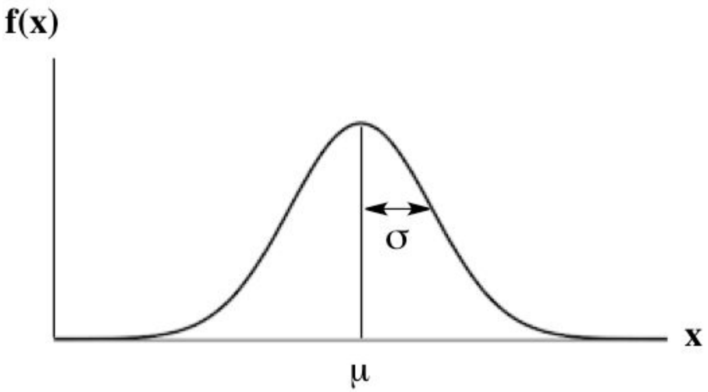
$$(\mu_P, \sigma_P)$$

Sample, S1

$$(\mu_{S_1}, \sigma_{S_1})$$

Representativeness

Probability distribution, *P*



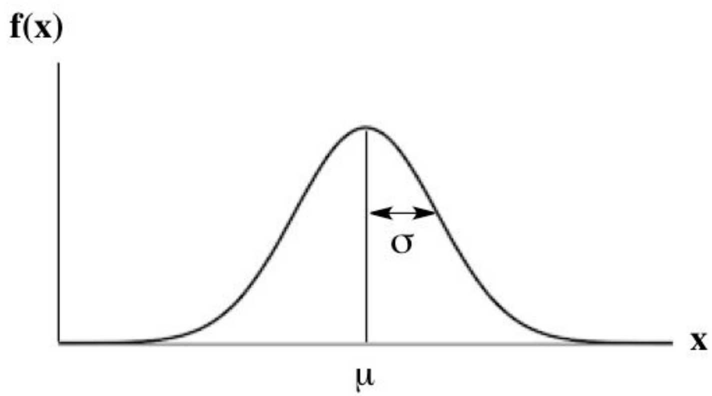
$$(\mu_P, \sigma_P)$$

Sample, S1

$$(\mu_{S_1}, \sigma_{S_1})$$

Representativeness

Probability distribution, *P*



$$(\mu_P, \sigma_P)$$

Sample, S1

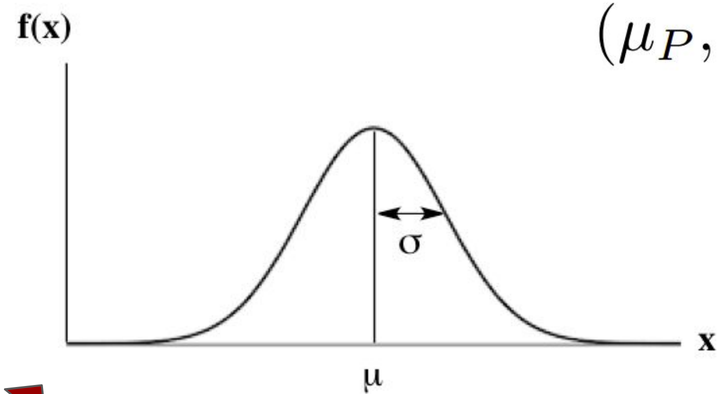
$$(\mu_{S_1}, \sigma_{S_1})$$

S_1 is *representative* of *P*

Representativeness

Probability distribution, *P*

(μ_P, σ_P)



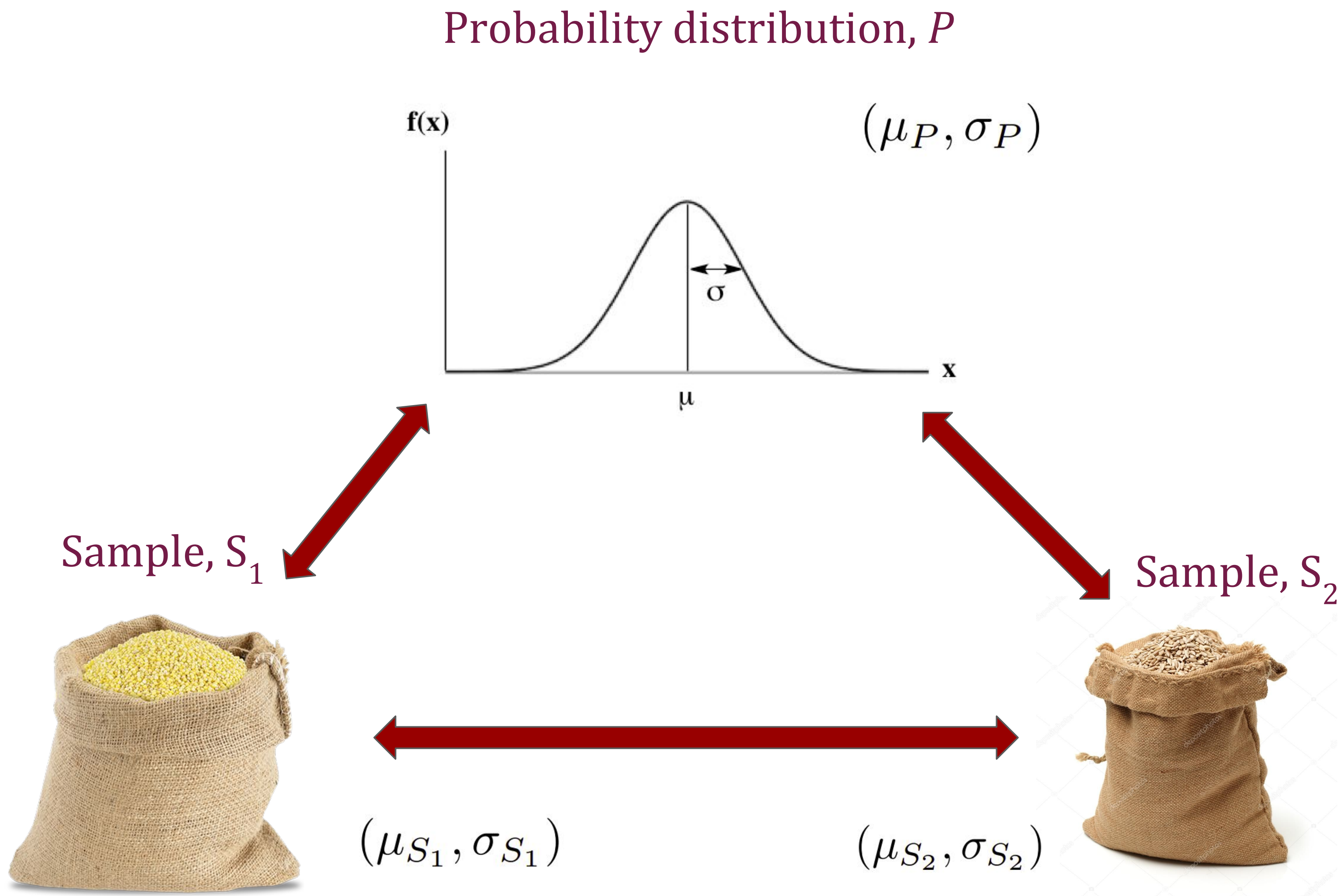
Sample, S_1

$(\mu_{S_1}, \sigma_{S_1})$

Sample, S_2

$(\mu_{S_2}, \sigma_{S_2})$

Representativeness



Supervised ML **model**

data	**training**, target
χ	set of all samples, x
Y	set of possible labels, y
h_{train}	learner: $y_{est;i} = h_{train}(x_i)$
L	Loss function

+ Representativeness between training and target

Data generation model:

$x_i \sim P_\chi$

$f \rightarrow$ true labeling function, $y_i = f(x_i)$

$L_{data,f}(h) \equiv P_{x\sim data}(h_{train}(x) \neq f(x))$

Shai and Shai, *Understanding ML: From Theory to Algorithms*, 2014, CUP

Supervised ML **model**

data	**training**, target
$\mathcal{X}$	set of all examples, x
Y	set of possible labels, y
h_{train}	learner: $y_{est;i} = h_{train}(x_i)$
L	Loss function

+ Representativeness between training and target

Machine Learning algorithm

Data generation model:

$x_i \sim P_X$

$f \rightarrow$ true labeling function, $y_i = f(x_i)$

$L_{data,f}(h) \equiv P_{x\sim data}(h_{train}(x) \neq f(x))$

Shai and Shai, *Understanding ML: From Theory to Algorithms*, 2014, CUP

Machine Learning algorithm

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. kNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boosting algorithms
    1. GBM
    2. XGBoost
    3. LightGBM
    4. CatBoost

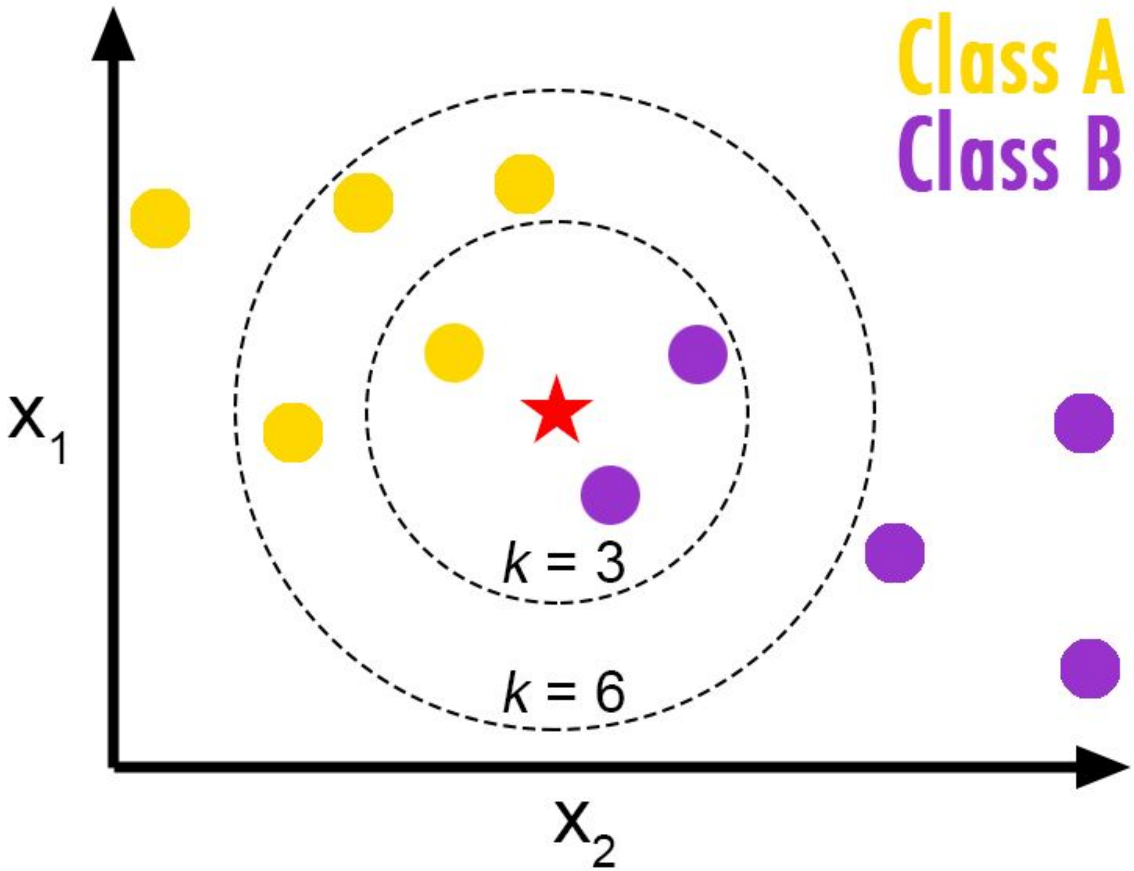
+ All things deep

https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

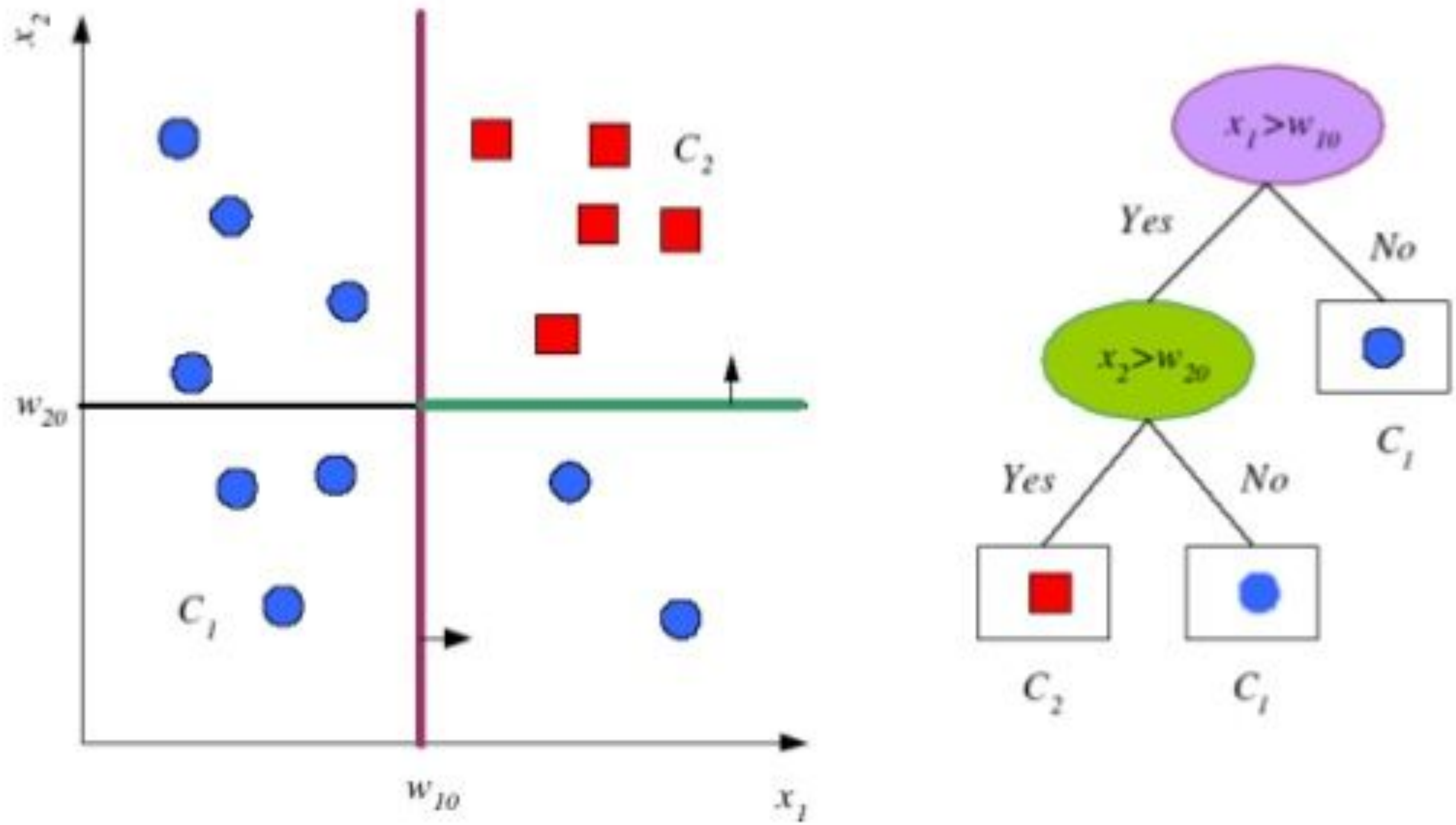
Example of supervised ML algorithm

k-Nearest Neighbor (kNN)

Distance based



Decision Trees

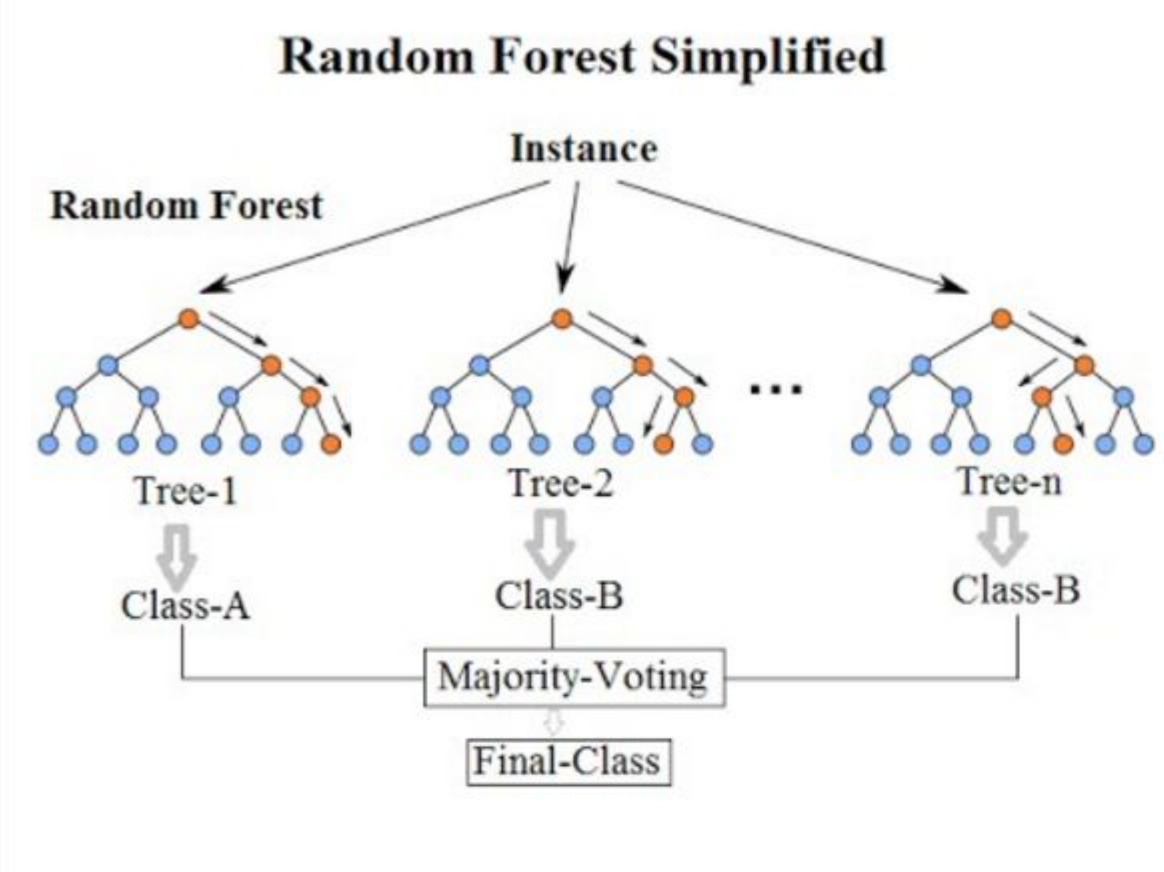


http://www.lewisgavin.co.uk/Machine-Learning-Decision-Tree/

Example of supervised ML algorithm

Random Forests

Ensemble method

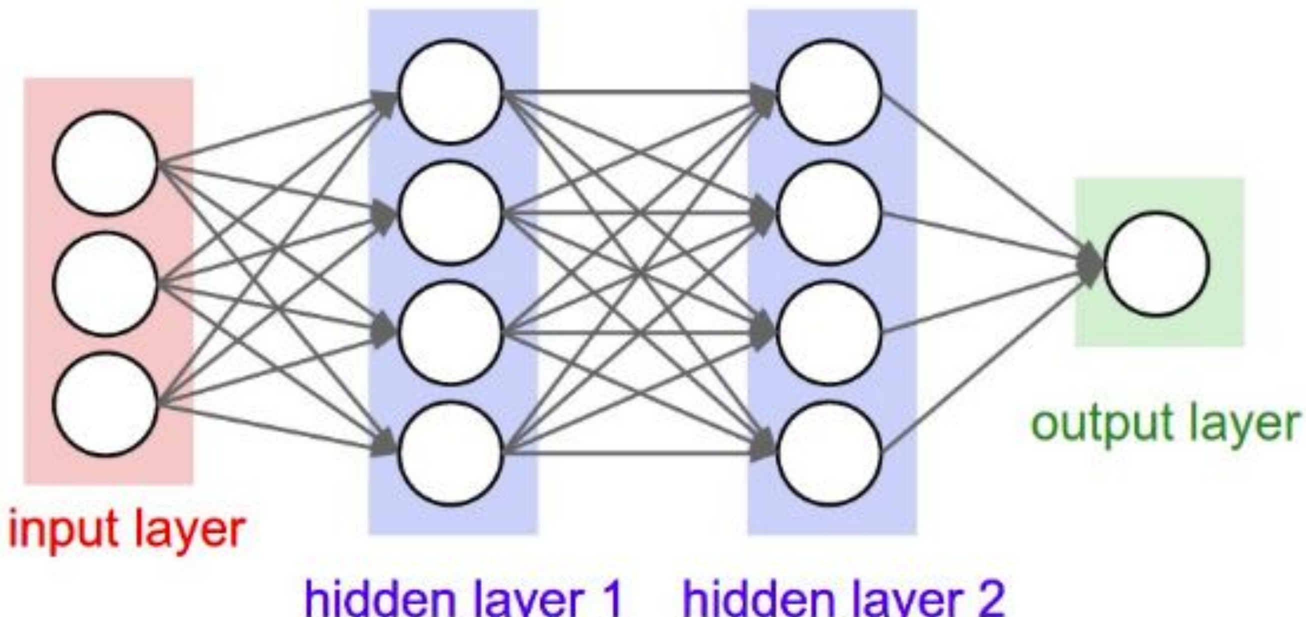


https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

Example of supervised ML algorithm:

Deep Neural Network

All layers internal to the network (not input or output layer) are considered hidden layers.



Slide by Alexandre Boucaud, ADA IX, 2018l

Supervised ML **model**

data **training**, target

+ Representativeness between training and target

χ set of all samples, x

Y set of possible labels, y

h_{train} learner: $y_{est;i} = h_{train}(x_i)$

L Loss function

Data generation model:

$x_i \sim P_\chi$

$f \rightarrow$ true labeling function, $y_i = f(x_i)$

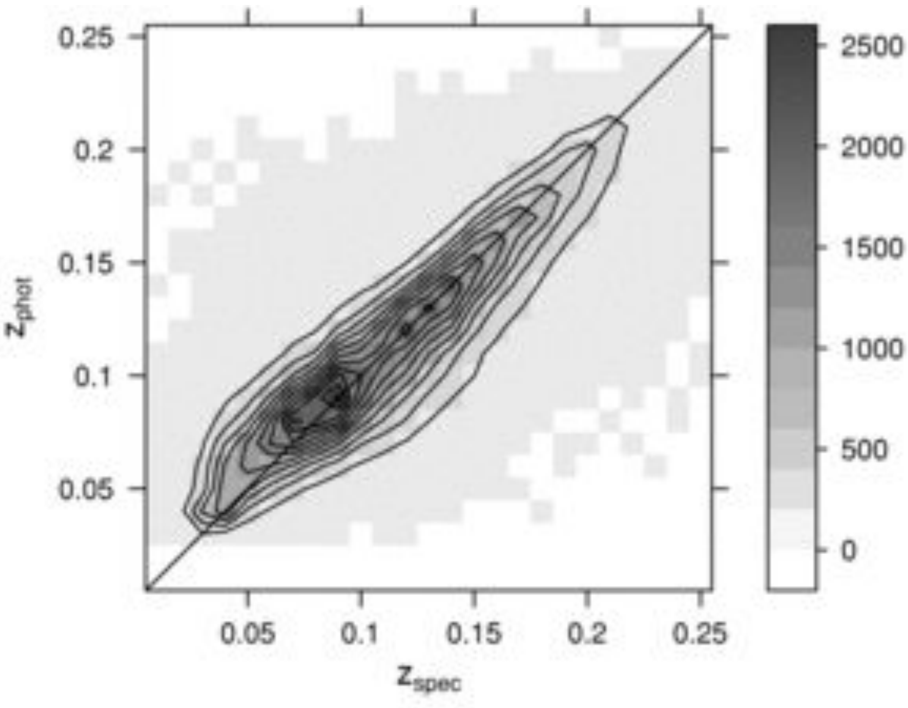
$$L_{data,f}(h) \equiv P_{x\sim data}(h_{train}(x) \neq f(x))$$

Shai and Shai, *Understanding ML: From Theory to Algorithms*, 2014, CUP

Classical use I:

Photometric Redshift

Random Forest



(a)

Neural Network

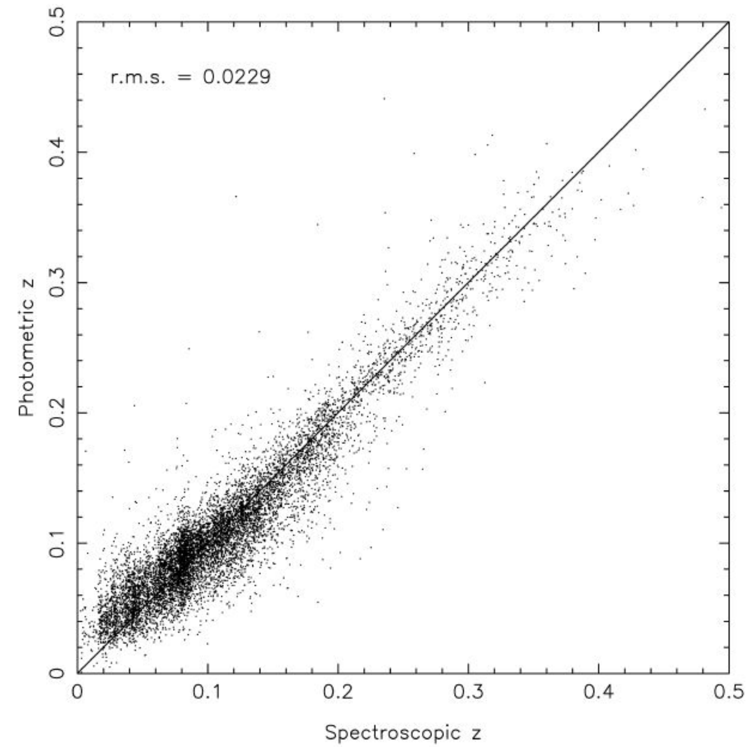


FIG. 2.— Spectroscopic *vs.* photometric redshifts for ANN*z* applied to 10,000 galaxies randomly selected from the SDSS EDR.

Carliles et al., 2010

Collister & Lahav, 2003

Symbolic Regression

Mathematical atoms:
+, - , x , / , pow

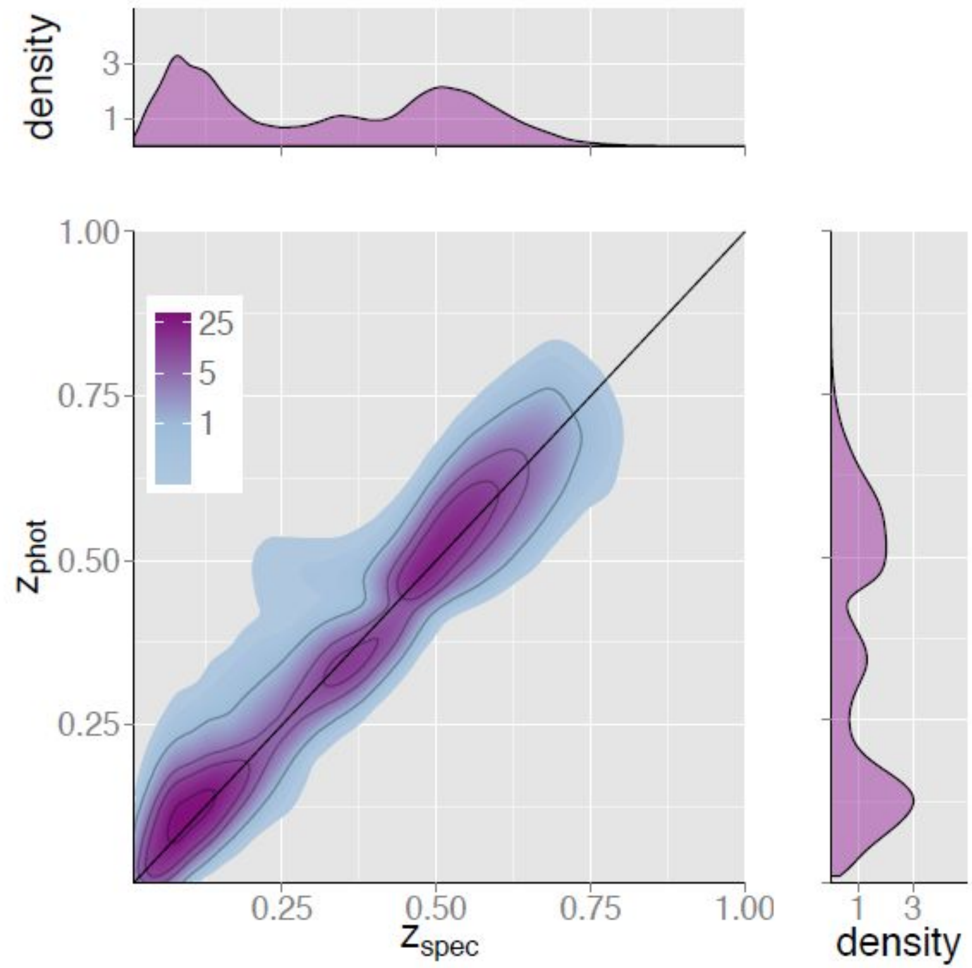
1 - Random construction of an analytical expression

2 - find the best parameters

3 - if result is better than previous keep it, otherwise discard it

Final expression:

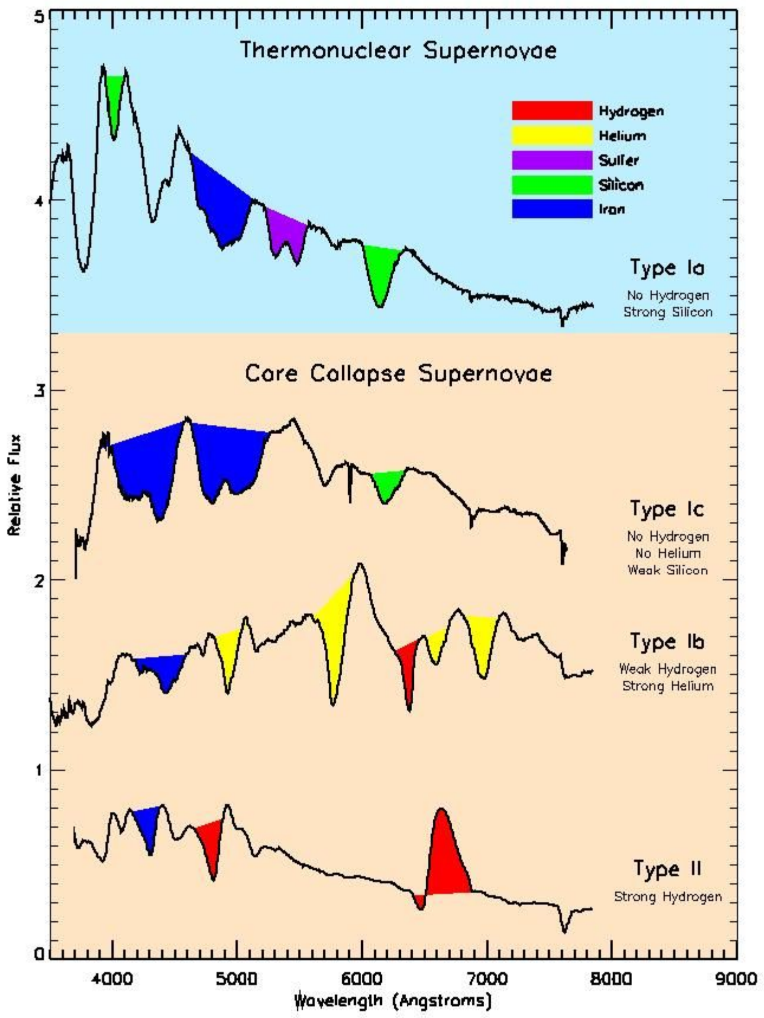
$$z_{\text{phot}} = \frac{0.4436r - 8.261}{24.4 + (g-r)^2(g-i)^2(r-i)^2 - g + 0.5152(r-i)}.$$

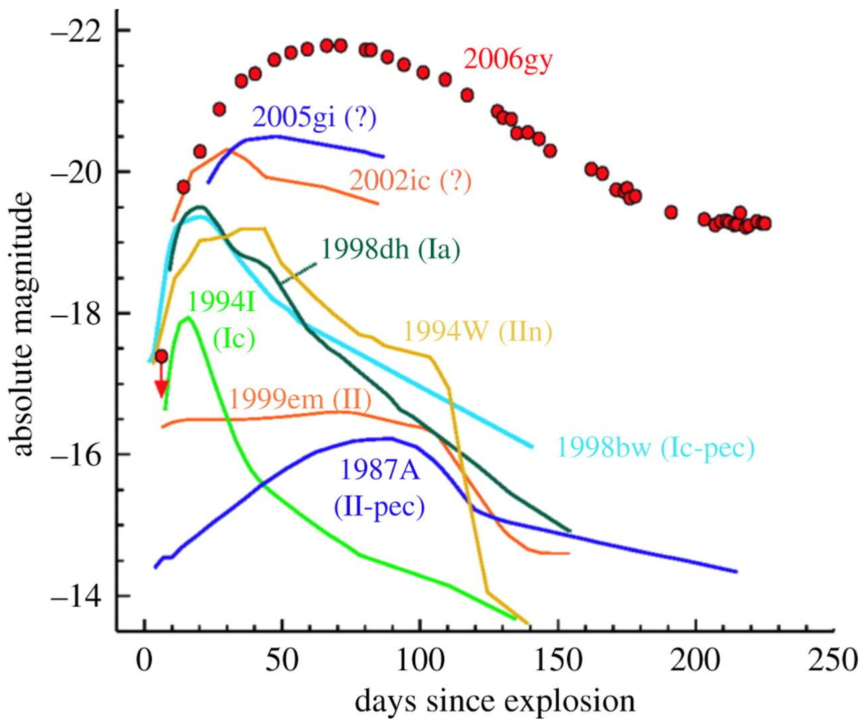


Pre-COIN paper:
Krone-Martins, ***Ishida*** *& de Souza, MNRASL 443 (2014)*

Classical use II:

SN Photometric classification



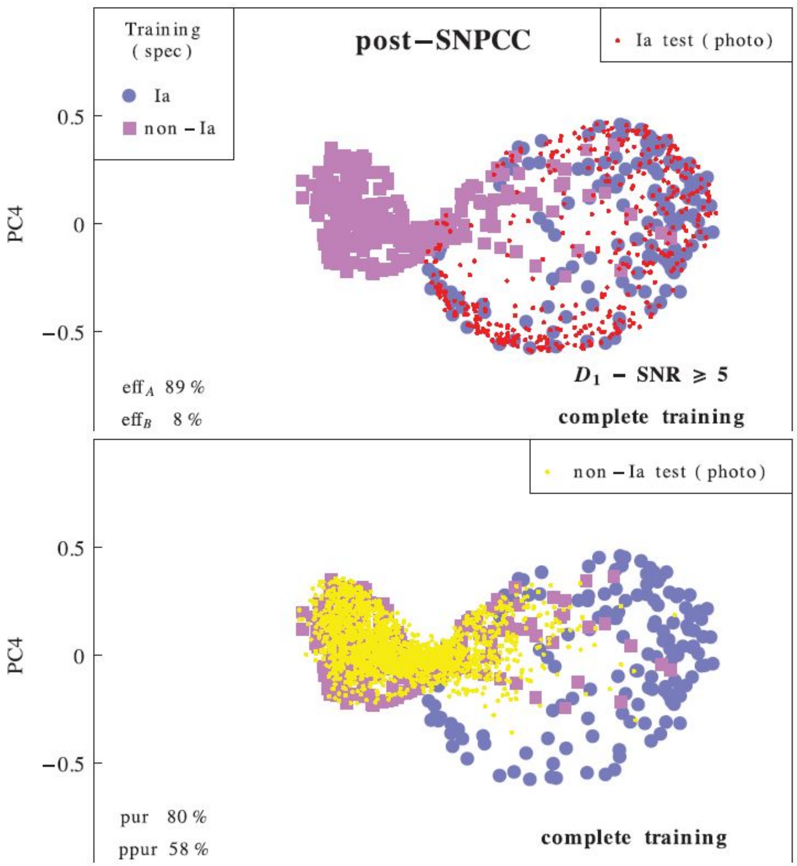


http://rsta.royalsocietypublishing.org/content/370/1960/774

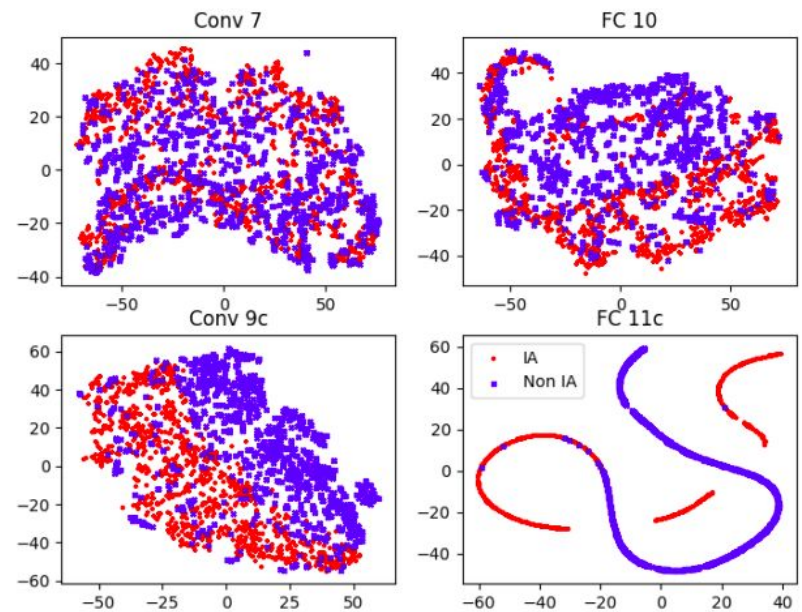
Classical use II:

SN Photometric classification

Nearest Neighbor



Deep Neural Network



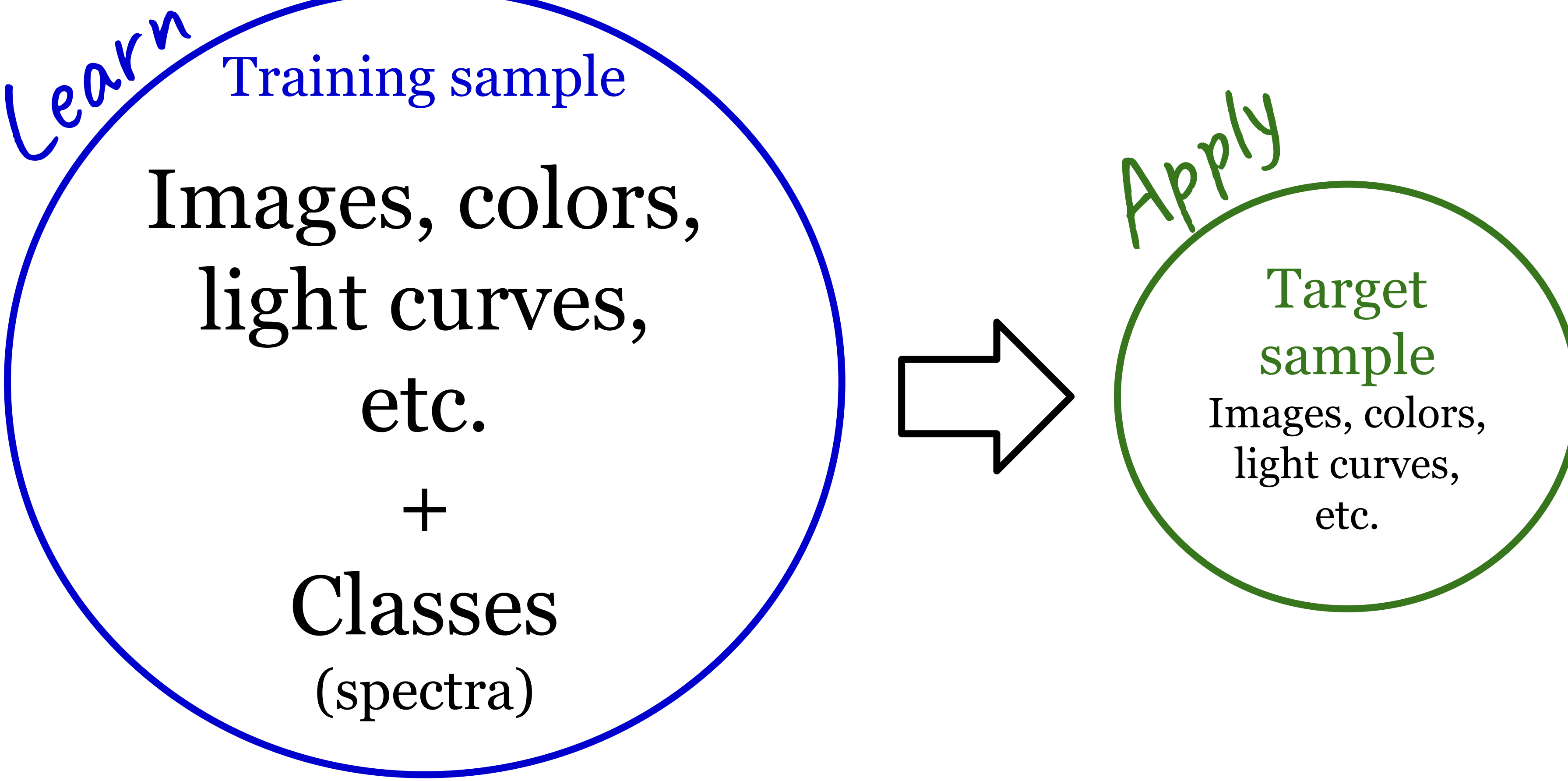
Pre-processing is super important!

Ishida *e de Souza, 2013*

Pasquet et al.,

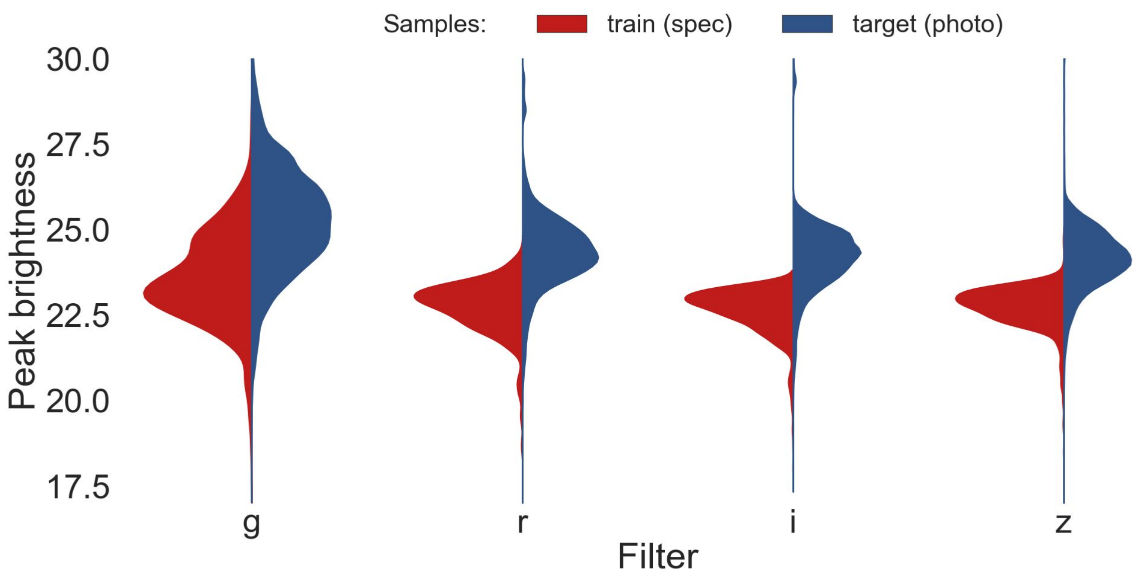
In astro, training means spectra

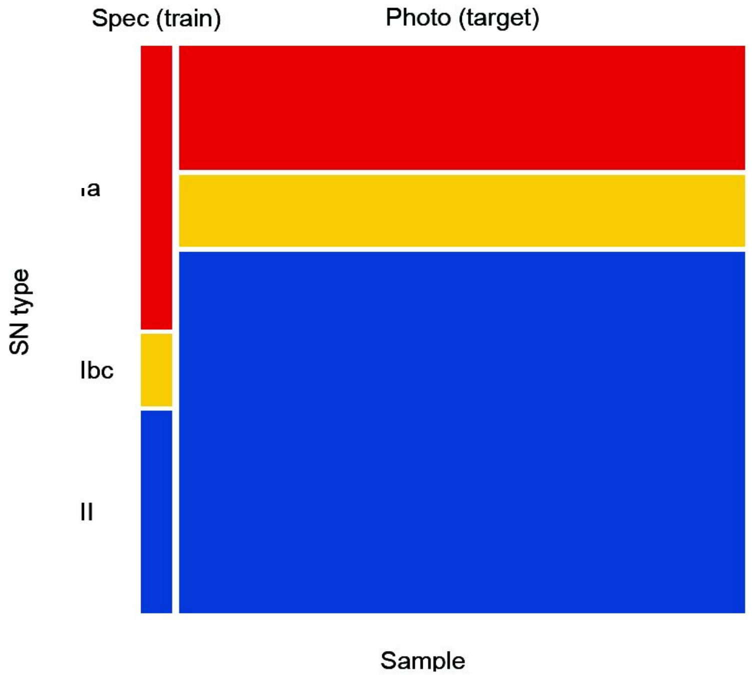
Ideal Supervised learning situation



Real astro-supervised learning situation

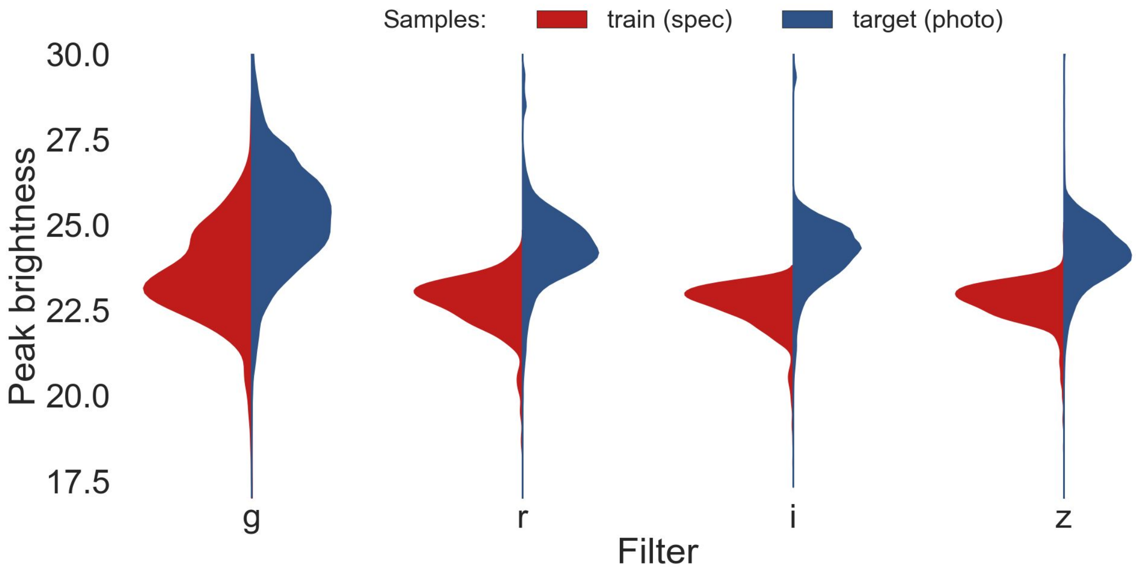
Representativeness

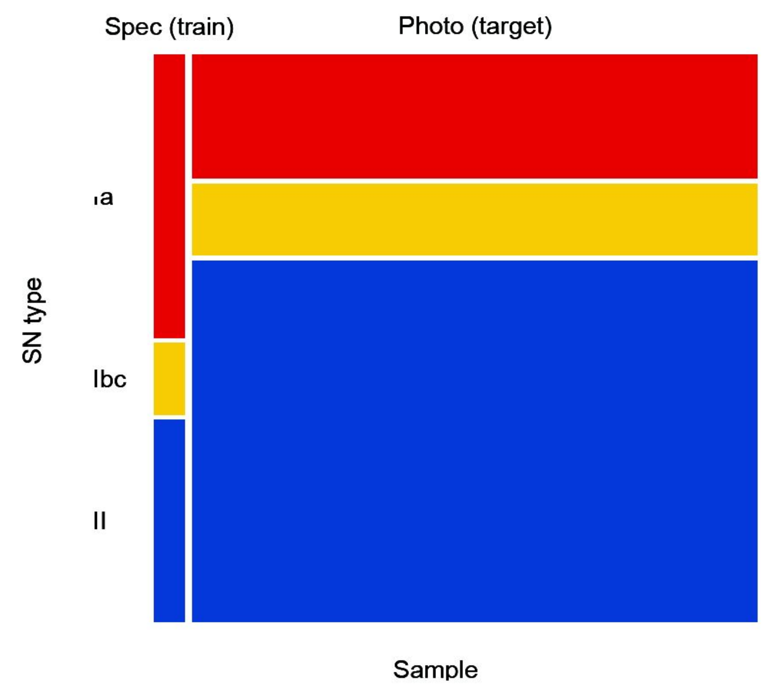




From COIN Residence Program #4, ***Ishida*** *et al., 2019, MNRAS, 483 (1), 2–18*

Representativeness



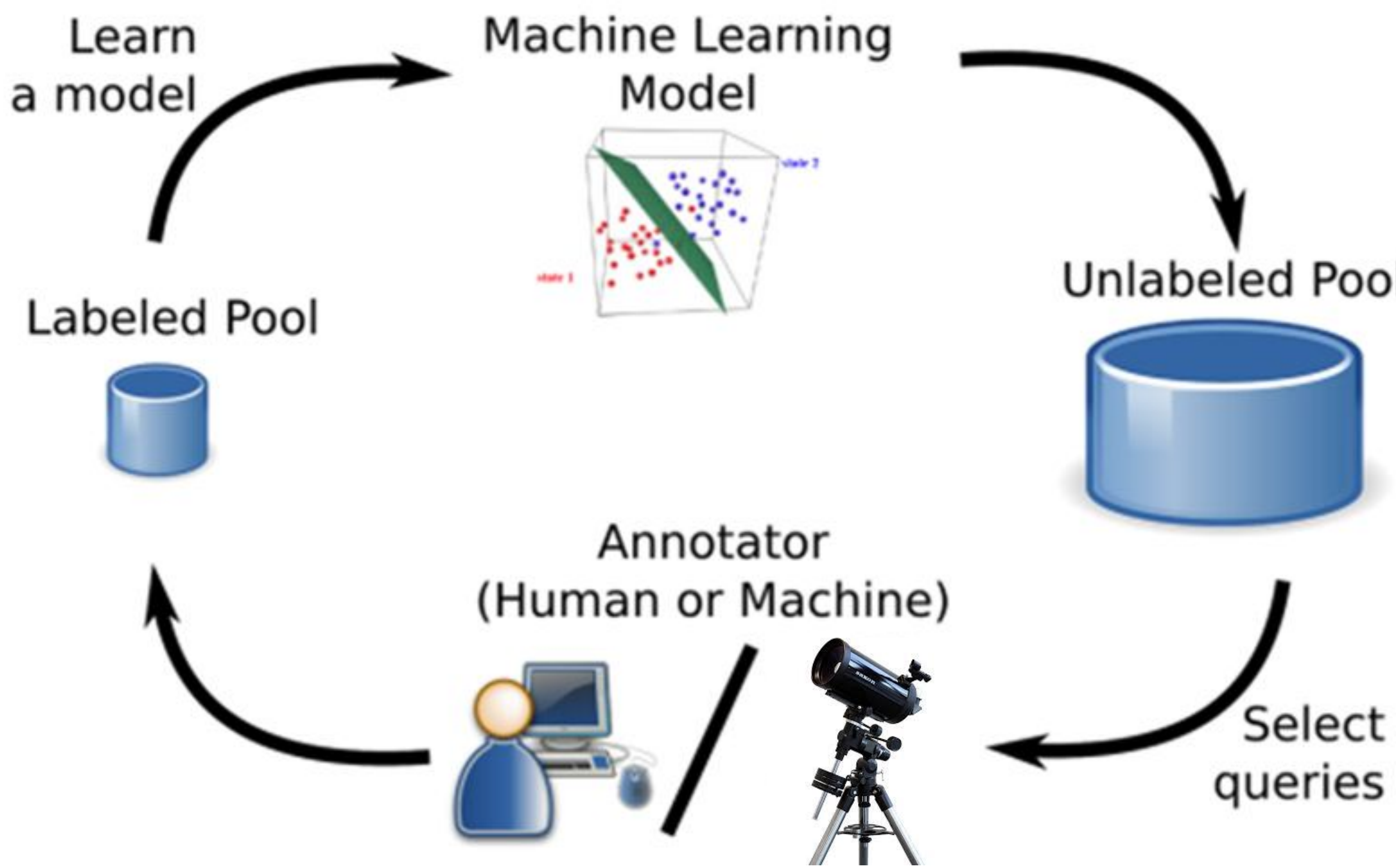


Can machines learn **better**, with **fewer** labelled examples, **if** they are carefully chosen?

From COIN Residence Program #4, ***Ishida*** *et al., 2019, MNRAS, 483 (1), 2–18*

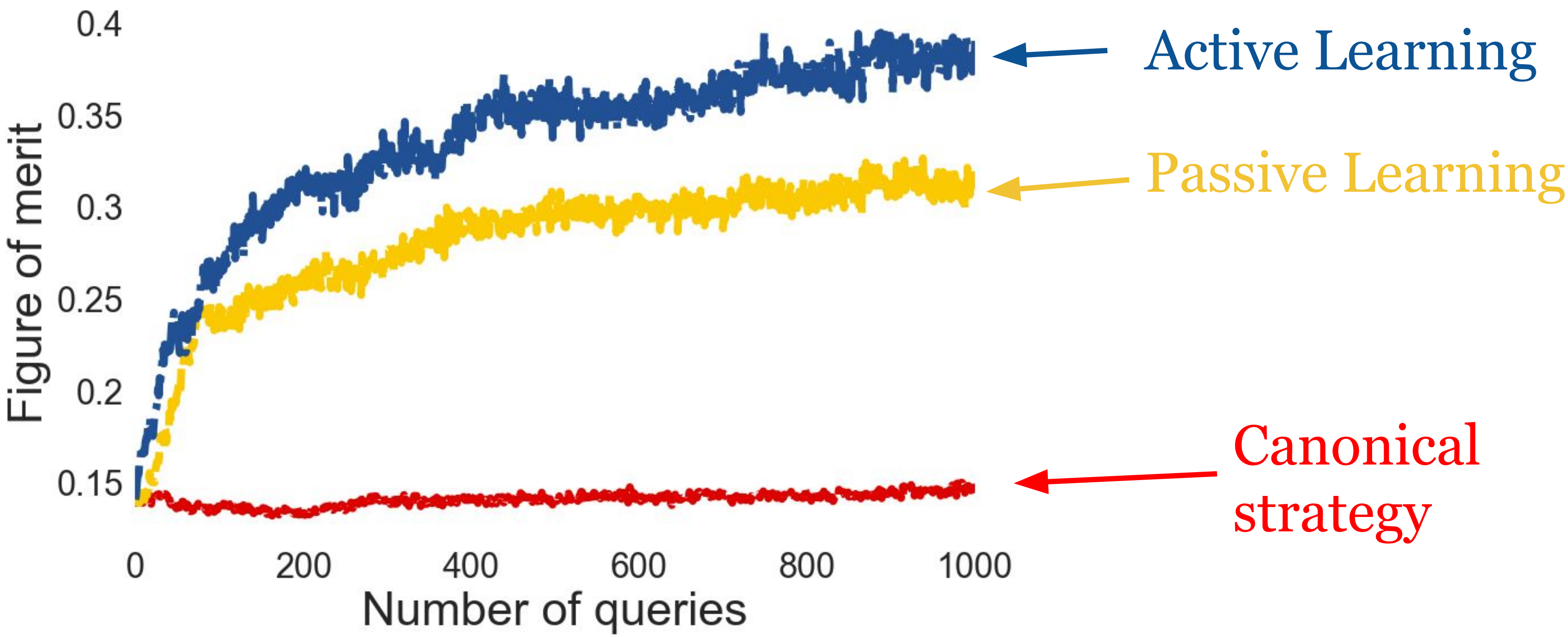
Active Learning

Optimal classification, minimum training



AL for SN classification

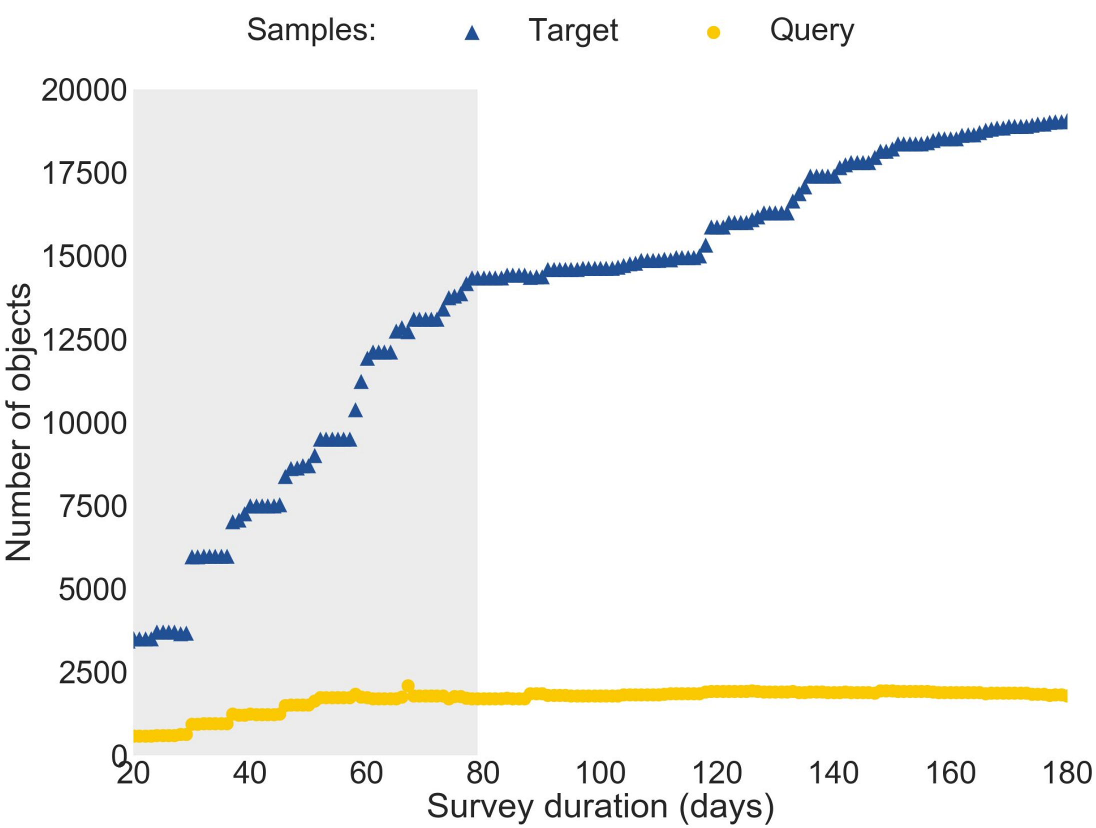
Static results



Ishida** et al., 2018 - arXiv:astro-ph/1804.03765 - from **CRP #4

Time Domain

Survey evolution



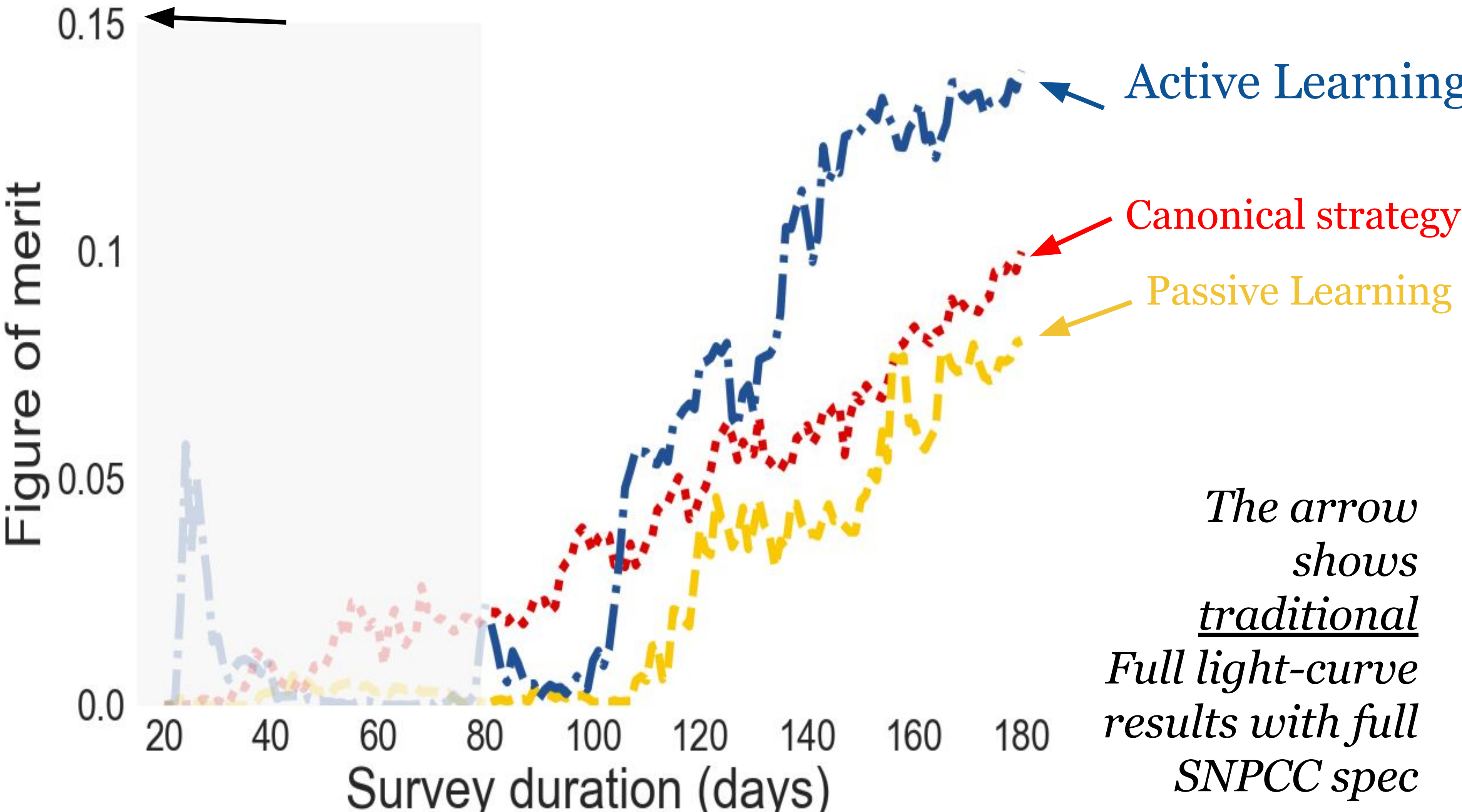
1. Feature extraction done daily **with available observed epochs until then.**

2. Query sample is also re-defined daily: objects with **r-mag < 24**

3. **No** need for an **initial training** sample

Ishida** et al., 2018 - arXiv:astro-ph/1804.03765 - from **CRP #4

Do we even need a training set?



The arrow shows <u>traditional</u> Full light-curve results with full SNPCC spec 1103 spectra

From COIN Residence Program #4, ***Ishida*** *et al., 2019, MNRAS, 483 (1), 2–18*

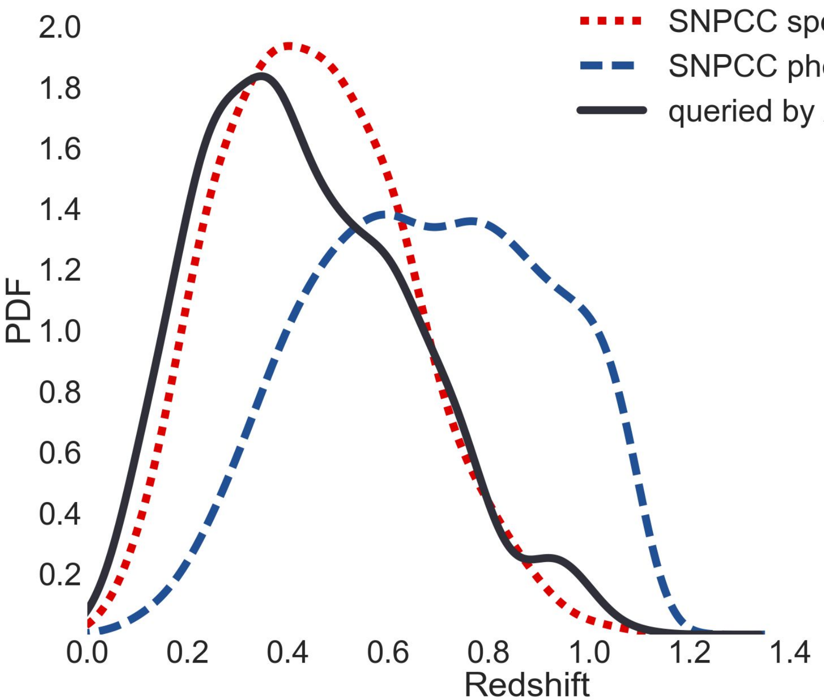
The queried sample

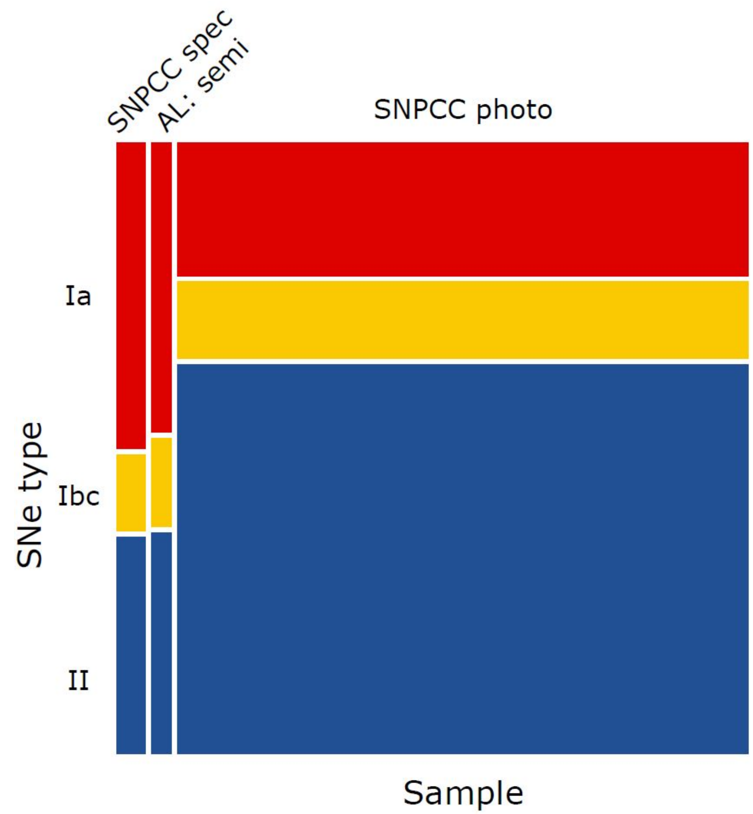
Partial LC, no training, time domain, batch

SNPCC spec:
1103 objects

Queried sample:
800 objects

Telescope time:
Queried/spec = 0.999





Ishida** et al., 2018 - arXiv:astro-ph/1804.03765 - from **CRP #4

Take home messages:

Astronomy needs
optimized training samples
for
Machine Learning
applications

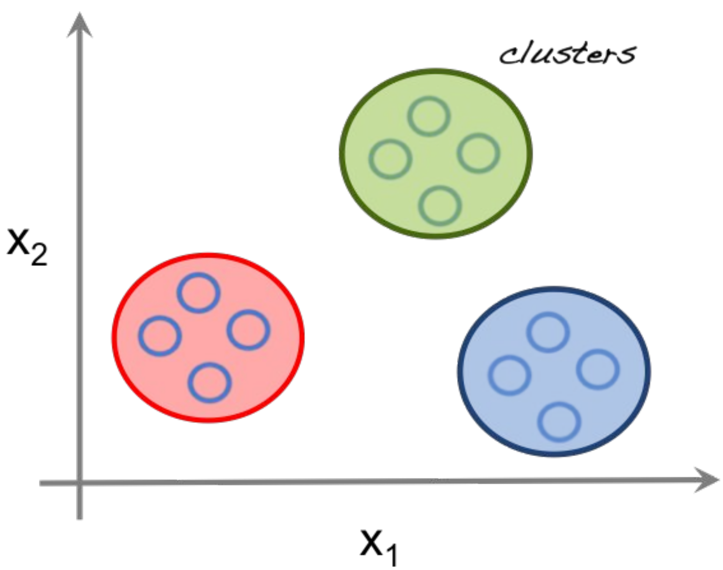
Given the volume and complexity of upcoming data
Machine Learning is not optional

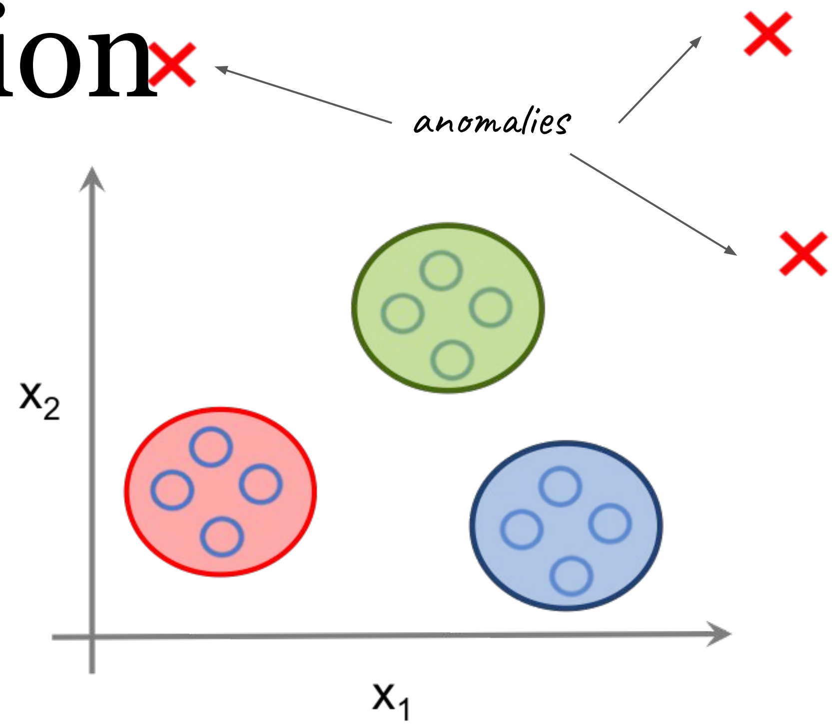
However, it should be used with parsimony …
Using off-the-shelve algorithms is not advisable!

THANK YOU

Extra slides

Clustering and Anomaly Detection





"An anomaly is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"

Hawkins, 1980

Active Anomaly Detection

A strategy

If yes: check next obj in the anomaly score board

If no: update the weights to accommodate the new information

Isolation Forest

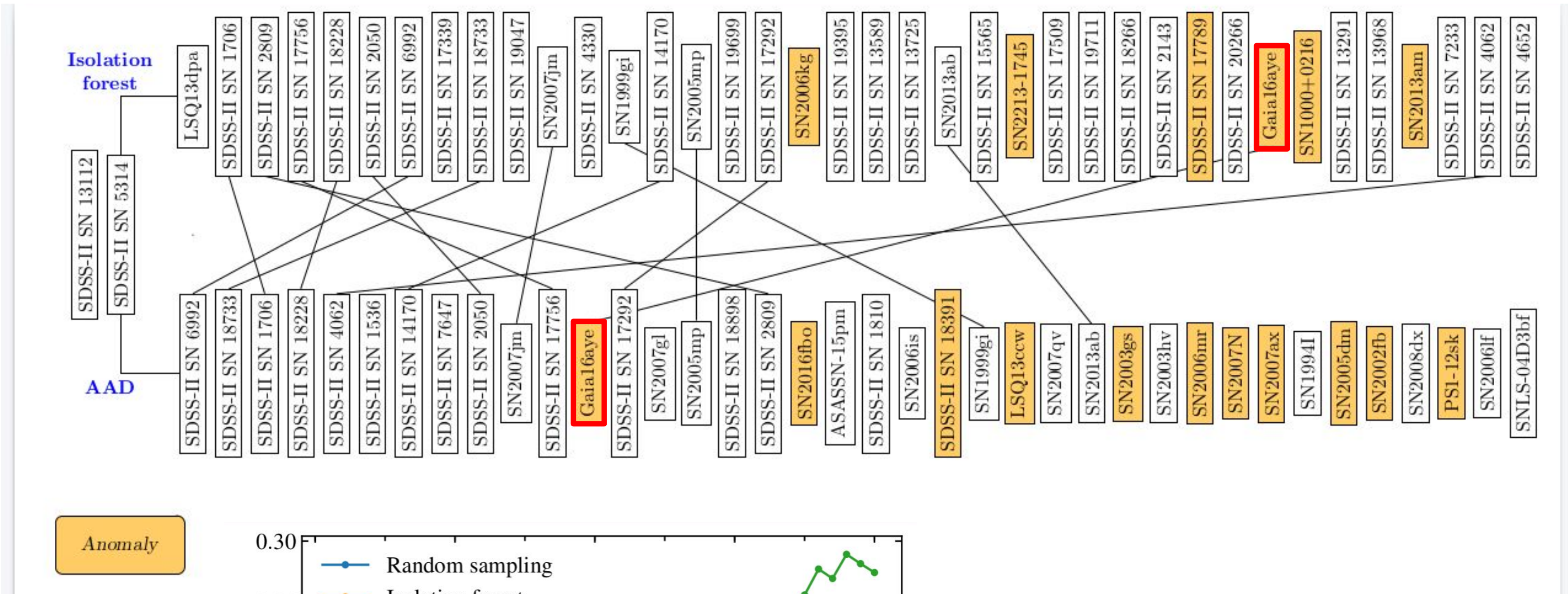
List of anomalies

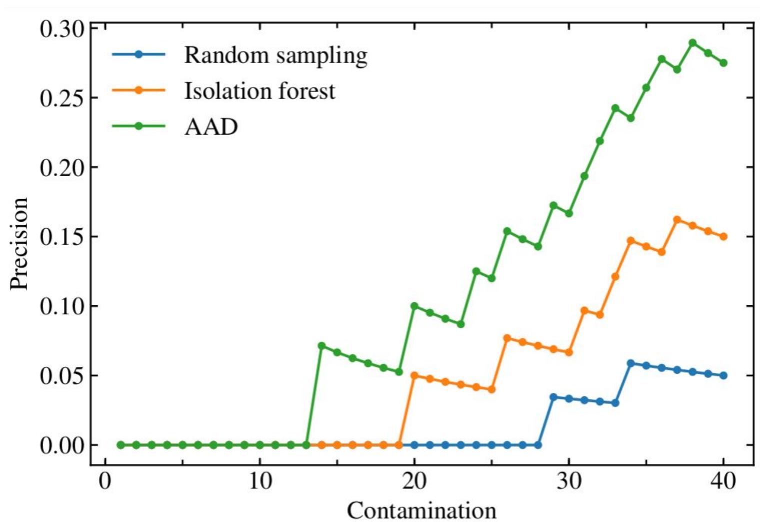
Spec. check object with highest anomaly score

Das, S., Wong, W-K., Dietterich, T., Fern, A. and Emmott, A. (2016). *Incorporating Expert Feedback into Active Anomaly Discovery* in the Proceedings of the IEEE International Conference on Data Mining

Active Anomaly Detection

In the open supernova catalog



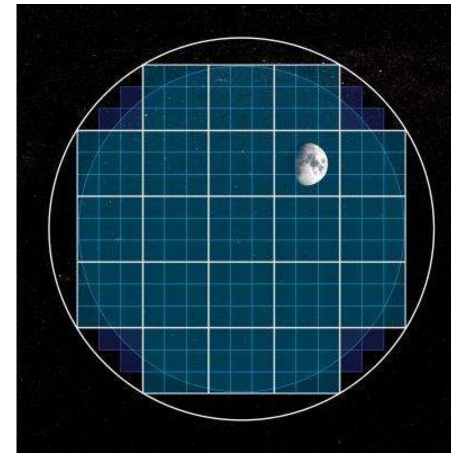


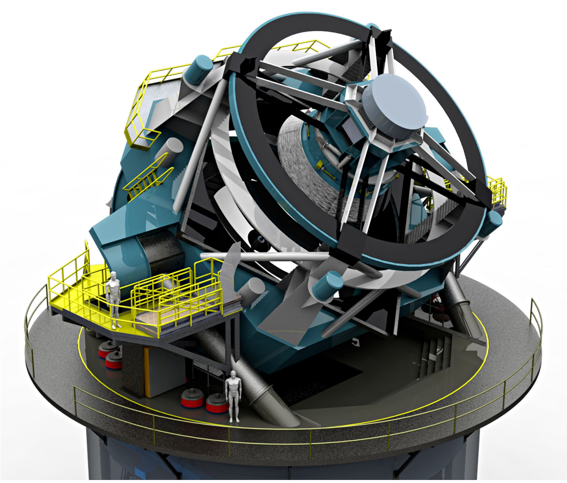
Ishida et al., 2019

The Large Synoptic Survey Telescope

Photometric obs:
~minute

Spectroscopic obs:
>= 1 hour (e.g. SDSS)
Multi-fiber spec.
Pointing is not trivial





Camera: 3**.2 Giga** pixels and 1.65m

Primary mirror: **8.4m**

Field of view: **3.5 deg,** 40x full moon

Data production :**15 TB/night**

(3yr LSST=internet today)

~10 million alerts/night

30.000 type Ia SN/yr (today ~1000)

Expected ~ **1000 spectra/yr** (~ 3%)

The LSST alert stream

5 **Confirmed sources**
New source!

4 **Spectroscopic Observations**
Organise follow-up

3 **Cross-Match**
Correlate to other experiments and catalogs

2 **Process Alerts**
Broker core algorithms

1 **LSST raw alert data**
Difference image analysis
~1TB/night expected.

Diagram by Julien Peloton, LAL

Fink: a community broker based on Active Learning and Spark

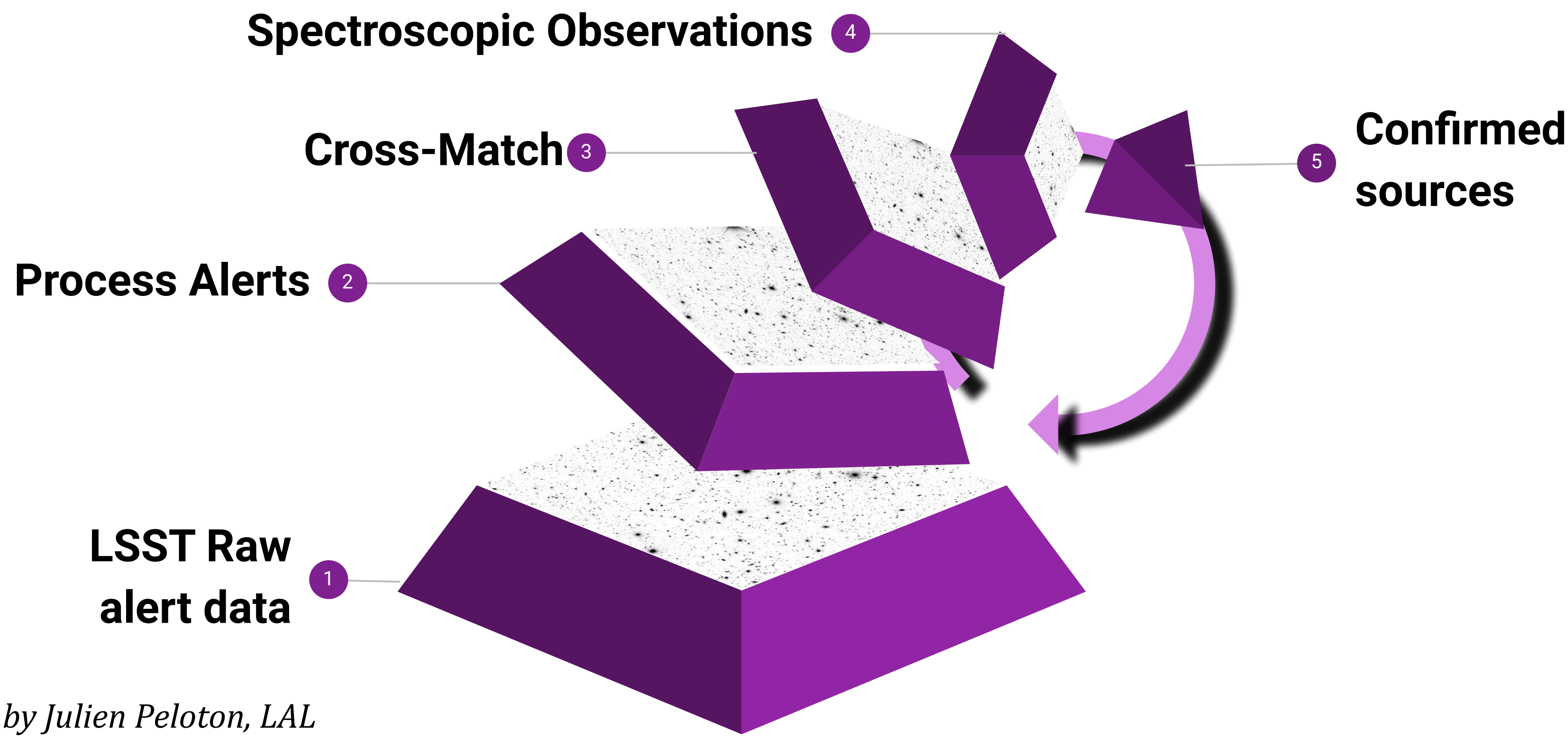
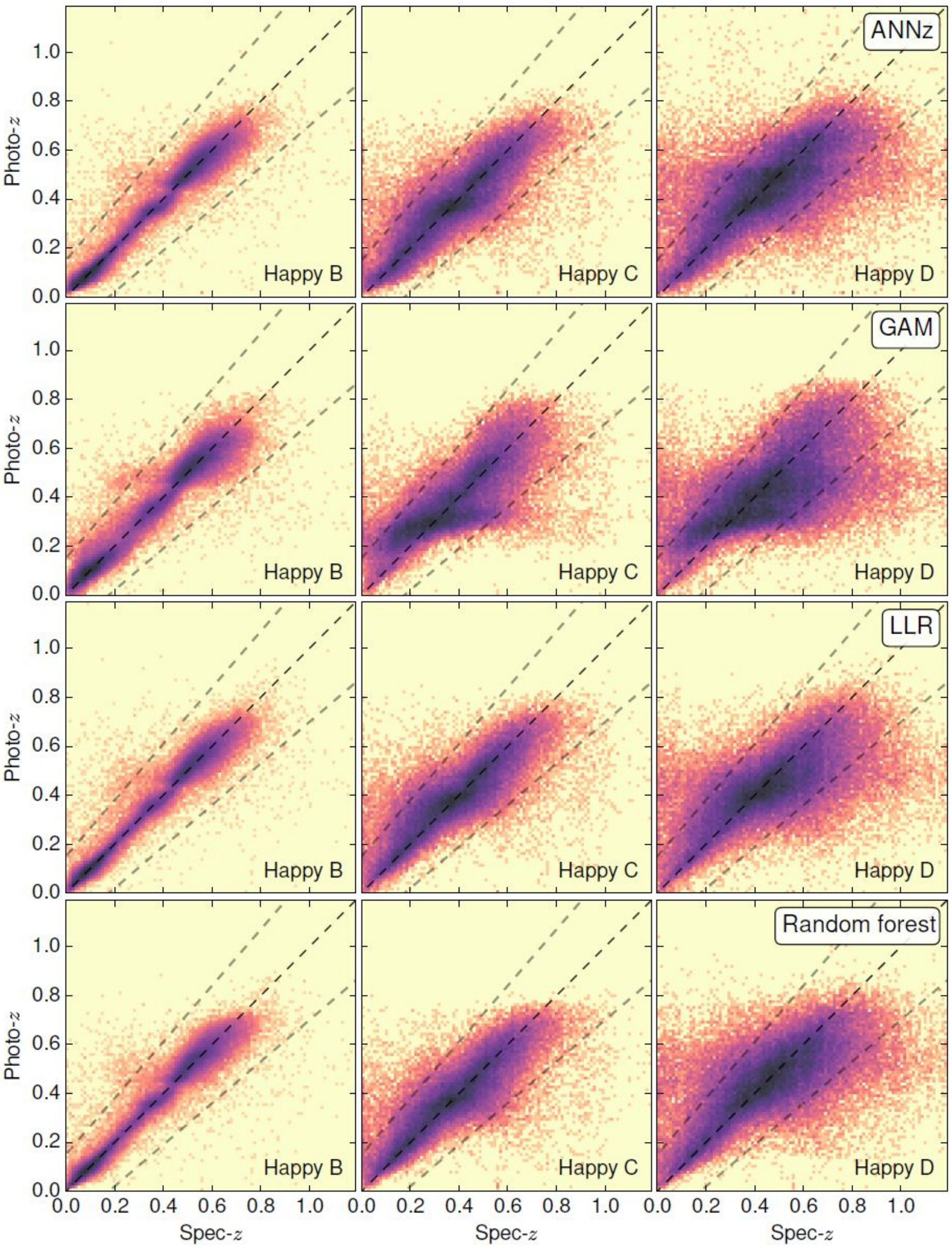


Diagram by Julien Peloton, LAL

https://fink-broker.readthedocs.io/en/latest/



Happy catalogue

The effect of coverage + photometric errors

Beck et al., astro-ph:1701.08748, MNRAS 2017