A Machine Learning Tutorial for Veterinarians: Examples Using Canine Atopic Dermatitis

Nathan Bollig, DVM

Computation and Informatics in Biology and Medicine Postdoctoral Fellow
and
Ph.D. student, Computer Sciences

University of Wisconsin
4720 Medical Sciences Center
1300 University Avenue
Madison, Wisconsin 53706

2020 Virtual Talbot Veterinary Informatics Symposium
www.avinformatics.org

Outline

- Canine atopic dermatitis
- Introduction to machine learning
- Modeling classification tasks for canine atopic dermatitis
- Evaluating model performance
- Comparing machine learning algorithms
- Important takeaways

Outline

- Canine atopic dermatitis
- Introduction to machine learning
- Modeling classification tasks for canine atopic dermatitis
- Evaluating model performance
- Comparing machine learning algorithms
- Important takeaways

Canine atopic dermatitis (CAD, atopy)

- Common inflammatory skin disease in dogs
- Treated with allergen specific immunotherapy (ASIT), administered either subcutaneously or sublingually
- Although sublingual administration is effective in people, more evidence is needed to support efficacy of sublingual immunotherapy in dogs
- There are inconclusive results on risk factors for CAD in the United States

Outline

- Canine atopic dermatitis
- Introduction to machine learning
- Modeling classification tasks for canine atopic dermatitis
- Evaluating model performance
- Comparing machine learning algorithms
- Important takeaways

An impossibly simple problem

- Consider an overly simplistic (and incorrect) premise:
  - If a dog is greater than *t* years old, it will get CAD. You want the computer to display a message if a dog meets this condition.
- How to determine the threshold *t*?
- The simplicity here is in the feature representation and the premise

A classification task

Dog has disease or it doesn't = yes or no

This outcome is referred to as a **class label**

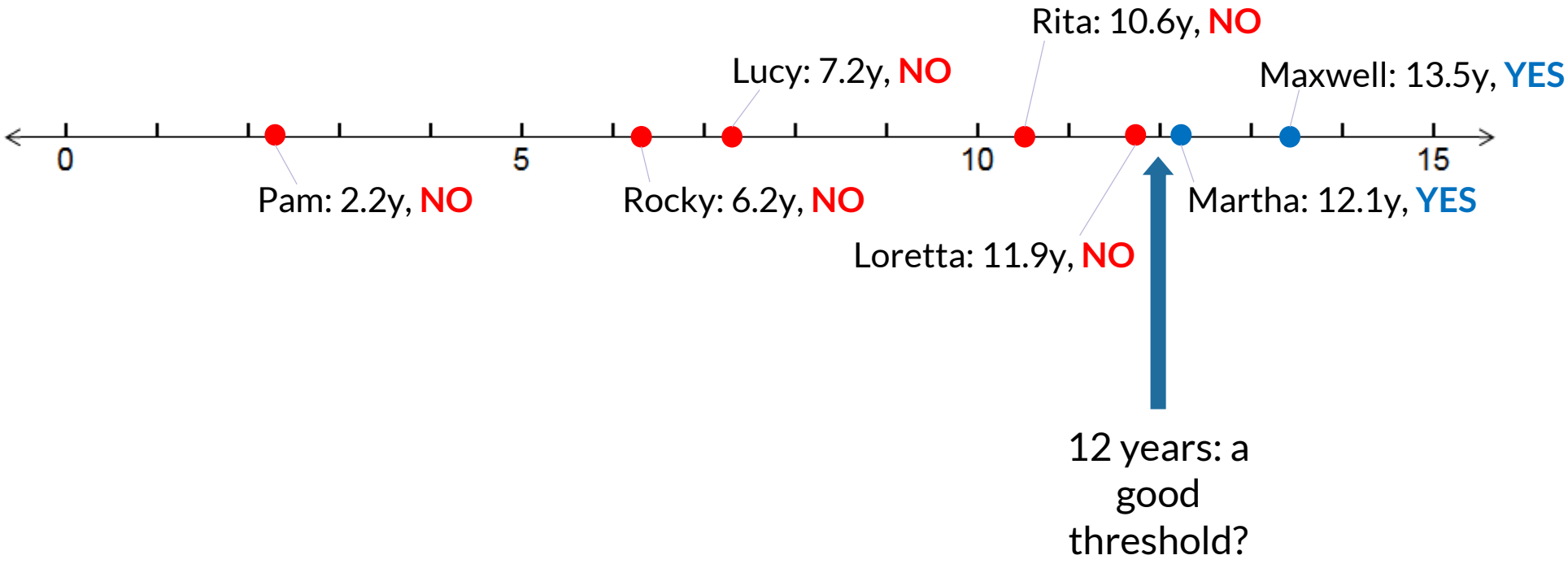
Method 1: Traditional Programming

Specify a classification rule (threshold value)

Method 2: Machine Learning

Learn the classification rule from examples

A classification task



- Once a threshold is determined, we have a **model** – a rule that we can use to classify dogs in the future

Accuracy of a ML model depends on…

- How dogs are represented (“feature representation”)
- Quality data
- Learning algorithm
  - If data cannot be cleanly separated into classes, then there would be different ways of finding the best threshold
  - Especially when there are more features, there are many types of learning algorithms we could use

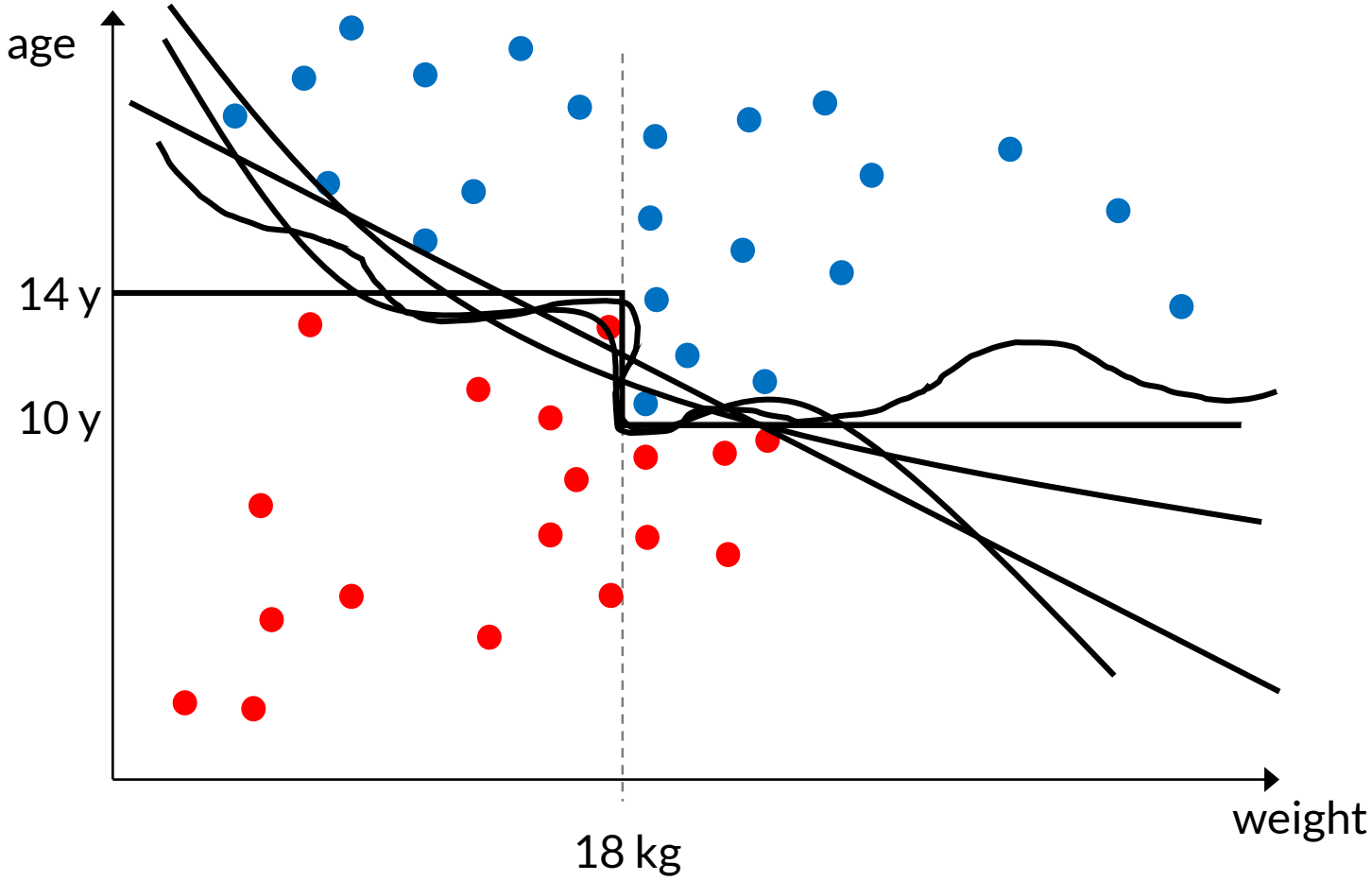
General Idea

- A machine learning model takes an input A and gives an output B
  - E.g. A = dog age in years, B = yes or no
- The **task** is well-defined, i.e. we know exactly what A is and what B can be
- Instead of implementing direct instructions for how to carry out a task, a machine learning program automatically **learns** with **experience**
  - "Learns": With respect to a given task, the program performs more accurately
  - "Experience" is **training data**
- A **learning algorithm** creates a model from data

Classification in 2 dimensions

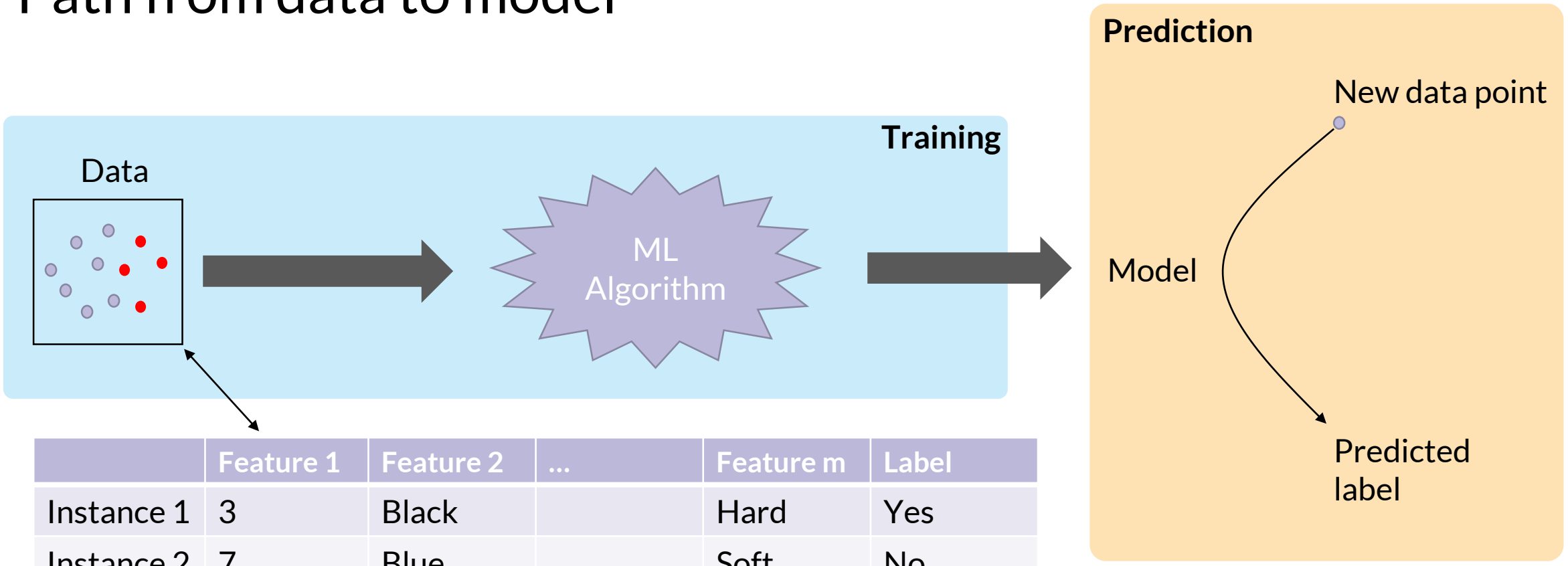
- Imagine the following:
  - If weight **≥ 18 kg**, then
    - If age is ≥ 10 y, then **YES** (has atopy).
    - If age is < 10 y, then **NO** (does not have atopy).
  - If weight **< 18 kg**, then
    - If age is ≥ 14 y, then **YES** (has atopy).
    - If age is < 14 y, then **NO** (does not have atopy).

Classification in 2 dimensions



If weight ≥ 18 kg, then
If age is ≥ 10 y, then **YES**
If age is < 10 y, then **NO**
If weight < 18 kg, then
If age is ≥ 14 y, then **YES**
If age is < 14 y, then **NO**

Path from data to model



	Feature 1	Feature 2	...	Feature m	Label
Instance 1	3	Black		Hard	Yes
Instance 2	7	Blue		Soft	No
...					
Instance n	17	Yellow		Fuzzy	No

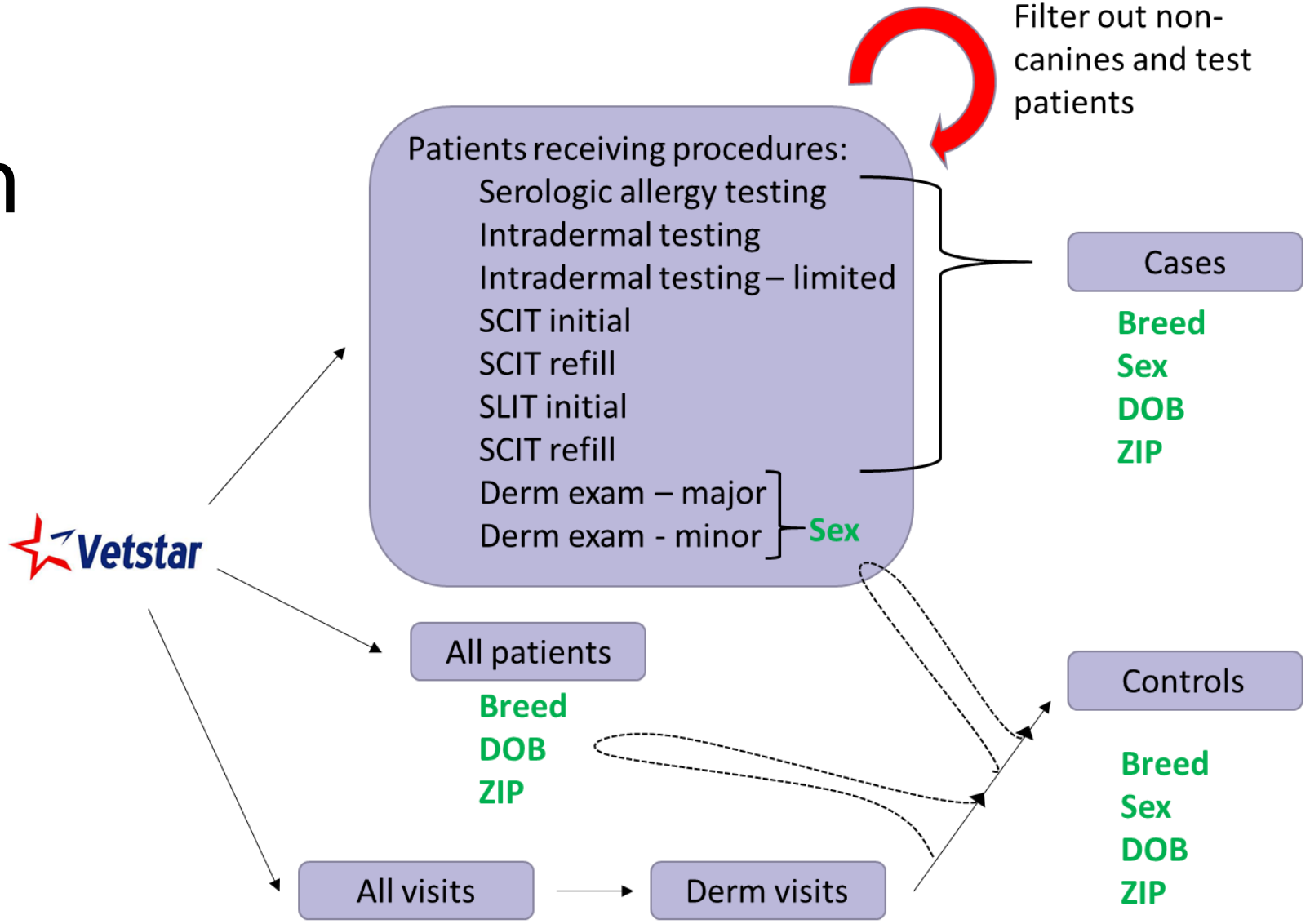
Important Questions

1. What is a machine learning algorithm? How does it create a model from the training data?
2. Why are there different machine learning algorithms, and how do you pick the best one?
3. Once a model is created, how do we measure its accuracy?

Outline

- Canine atopic dermatitis
- Introduction to machine learning
- Modeling classification tasks for canine atopic dermatitis
- Evaluating model performance
- Comparing machine learning algorithms
- Important takeaways

Data set construction



Two classification tasks

- Task 1: Fit a model to predict treatment success from factors that characterize the type of treatment.
- Task 2: Fit a model to predict case vs. control status from a set of possible risk factors.

Treatment success definition

- Patients treated with allergy shots were identified based on having received an initial allergy shot set
- Treatment success was then defined as positive (indicating “treatment success”) if and only if a patient received a refill set

CAD case and control definition

item	CAD Case Definition: 1 and (2 or 3 or 4 or 5 or 6)
1	Canine patient seen by dermatology
2	Clinical diagnosis is coded as atopic dermatitis
3	Patient has received intradermal testing (procedure 66076)
4	Patient has received intradermal testing - limited (procedure 66013)
5	Patient has received serologic allergy testing (procedure 03389)
6	Patient has been prescribed immunotherapy treatment (procedures noted below)

Controls were defined as a sample of canine dermatology patients not included in the case group

Dataset columns

Column Name	Description
breed_cat	Patient breed (Spaniel, Retriever, Shepherd, Pointer, Hound, Bulldog breed, Terrier, Setter, Northern breed, Poodle, Toy breed, Pinscher, Large breed, Spitz, Mixed breed, Other)
sex	Patient sex (female, male, neutered, spayed)
zip	ZIP code for patient address
RUCC	Rural-urban continuum codes (RUCC) characterizes county population numerically from 1 (largest) to 9 (smallest)
case	Case (1) or control (0)
dob	POSIX timestamp of patient date of birth
therapy	Patient therapy (allergy shot, sublingual, or none)
first_proc_date	POSIX timestamp of patient's first treatment date
first_proc_season	Season of patient's first season
age_days	Patient age at day of first treatment
age_cat	Ages are categorized as 1 ("young", less than 660 days) and 2 ("old", at least 660 days)
first_dvm_code	Numerical code representing attending DVM at first treatment
tx_success	Treatment success (1) or failure (0), where success is defined by patient returns > 0
returns	Number of return visits after initial treatment
dob_season	Season of patient's date of birth

Spreadsheet view of our dataset

	Feature 1	Feature 2	…	Feature m	Label
Patient 1	3	Black		Hard	?
Patient 2	7	Blue		Soft	?
…					
Patient n	17	Yellow		Fuzzy	?

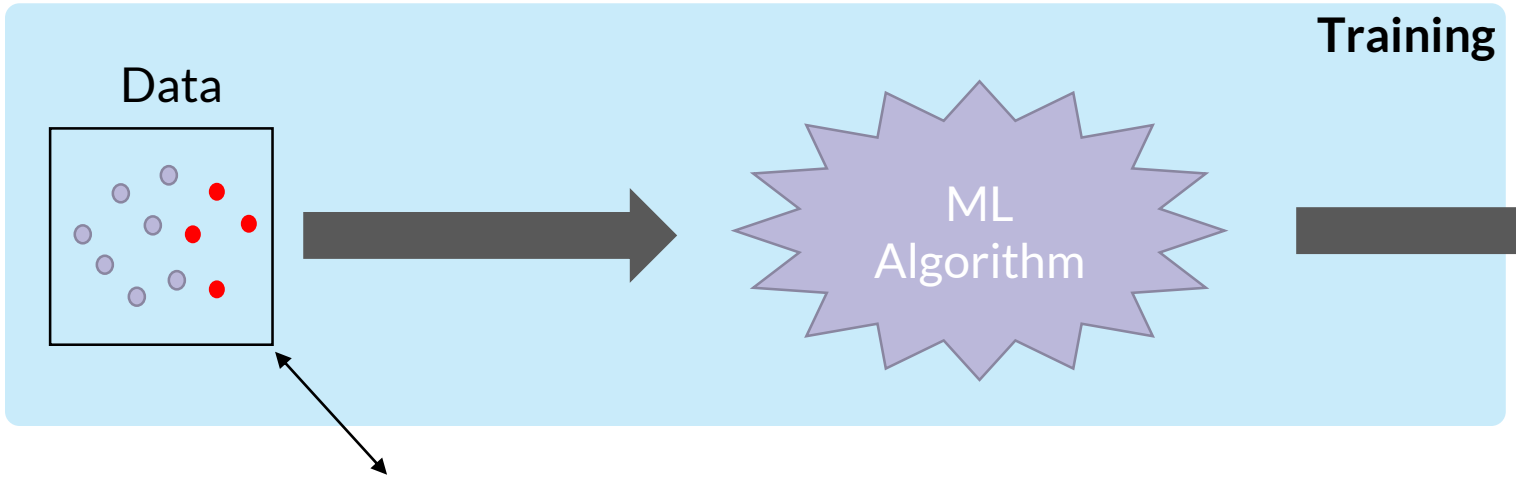
- Columns are potential features – whatever column is used as class label is not a feature, and some features may need to be omitted for an informative model
- As a basic concept, machine learning is a process that strives to fill in a column of a spreadsheet using the other columns of the spreadsheet

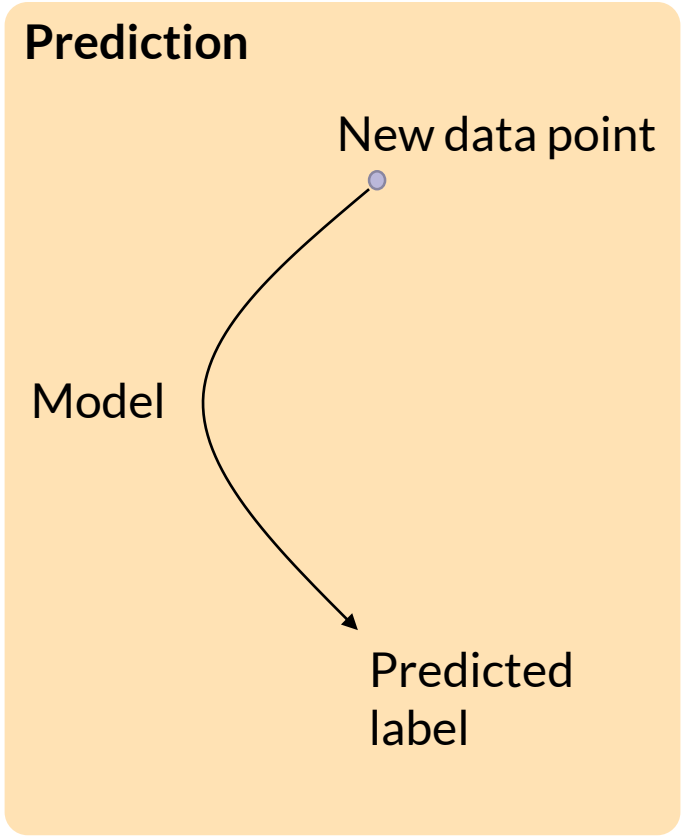
Outline

- Canine atopic dermatitis
- Introduction to machine learning
- Modeling classification tasks for canine atopic dermatitis
- Evaluating model performance
- Comparing machine learning algorithms
- Important takeaways

Once a model is created, how do we measure its accuracy?

Path from data to model





	Feature 1	Feature 2	...	Feature m	Label
Instance 1	3	Black		Hard	Yes
Instance 2	7	Blue		Soft	No
...					
Instance n	17	Yellow		Fuzzy	No

Photo by Elke Vogelsang

Testing a model requires data

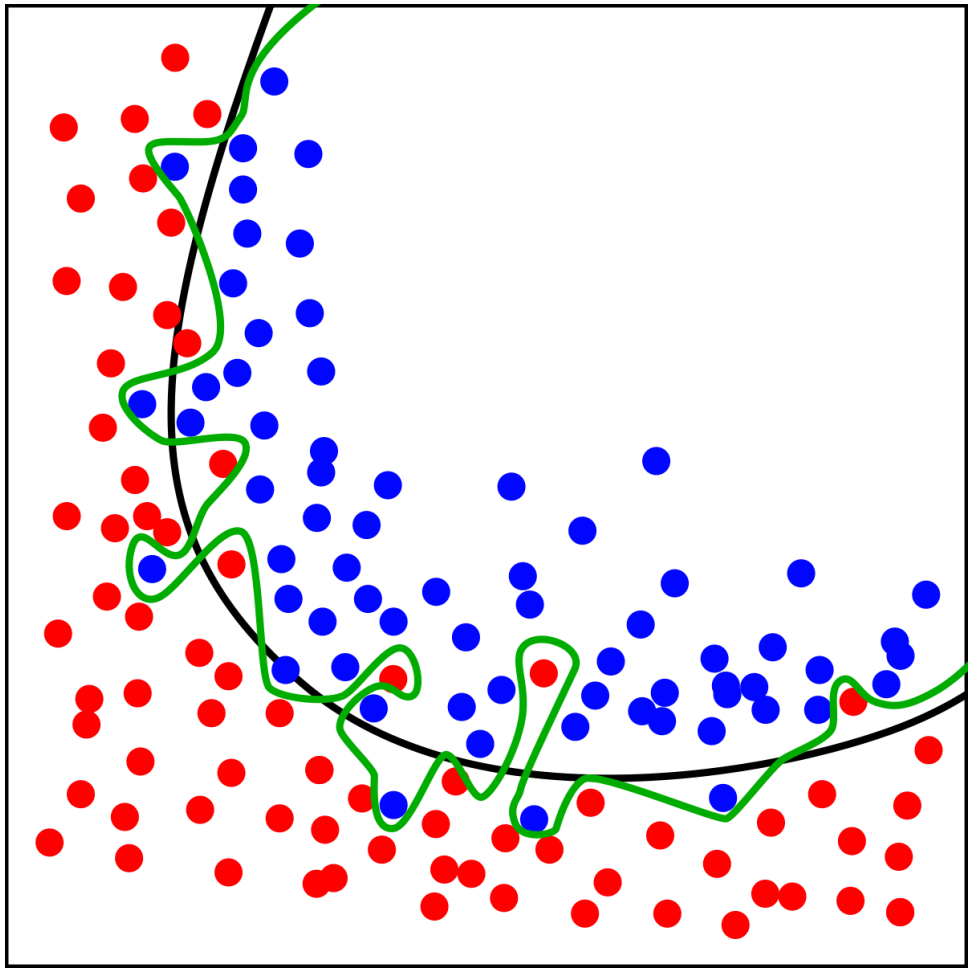
Suppose you have clinical records of 1,000 dogs including their case/control status. You want to train a machine learning model to predict case/control status. You also need to demonstrate that your model works by testing it on data, and you want to test it on the largest amount of data available. Which of the following would best achieve this?

A. Create a model using the 1,000 records and show that the model correctly classifies a large percentage of all of them.

B. Create a model using 800 records and show that the model correctly classifies a large percentage of the remaining 200 of them.

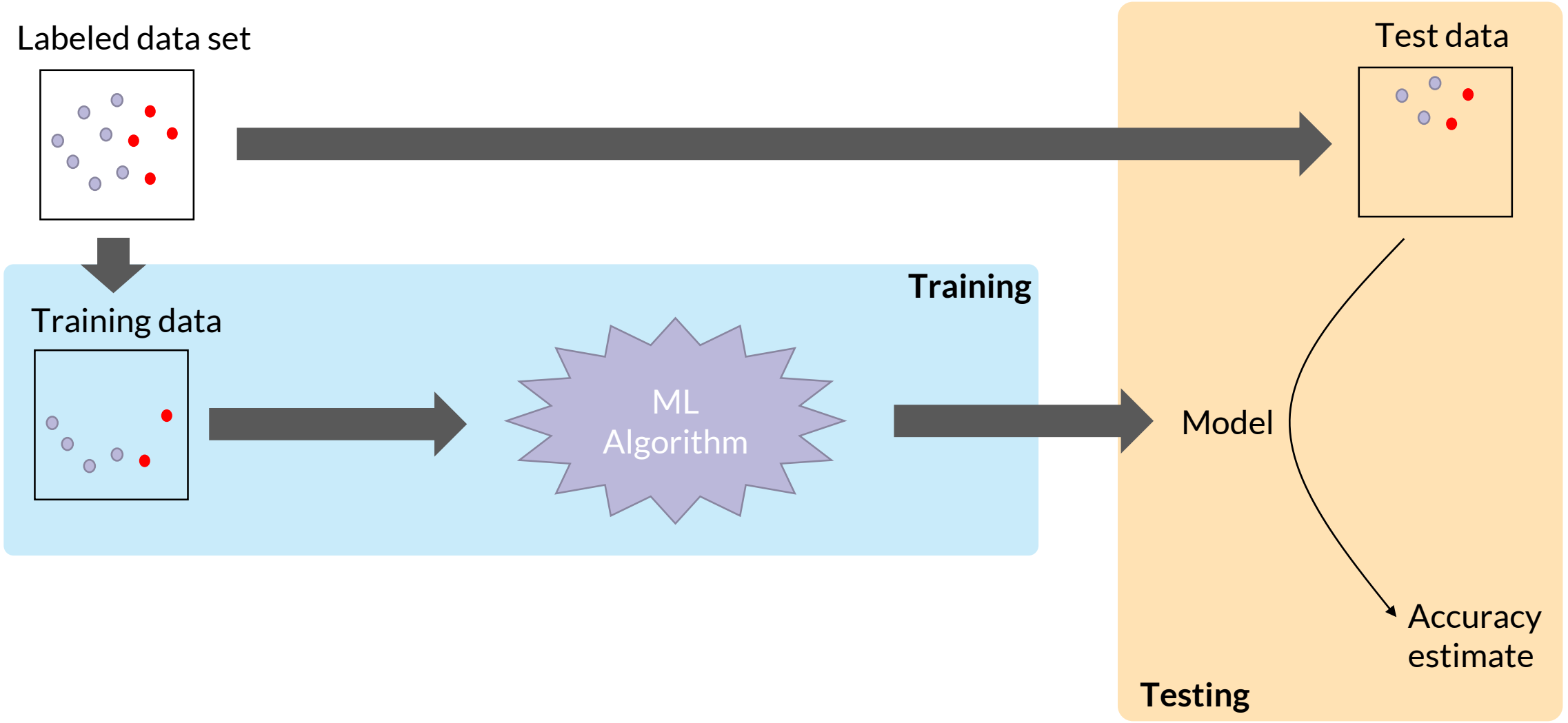
Evaluating a classification model

- Goal: Evaluate whether the model can generalize its experience beyond the specific training data.
- If the training accuracy is high, but test accuracy is low, this is called **overfitting**.
  - It happens a lot, and represents a situation where the performance on the training set is not very informative



https://en.wikipedia.org/wiki/Overfitting

Train/test split



Performance scores for binary classification models

- Raw Accuracy = correct classifications / total items in test set
- Could be misleading when there is **class skew**: What if atopy only was present in 1% of the population?
  - A model could hypothetically just predict that every dog does not have atopy and it would be correct 99% of the time
- There are several common performance stats used in place of accuracy:
  - recall, precision, F1 score, AUC of ROC curve or precision-recall curve

Outline

- Canine atopic dermatitis
- Introduction to machine learning
- Modeling classification tasks for canine atopic dermatitis
- Evaluating model performance
- Comparing machine learning algorithms
- Important takeaways

Why are there different learning algorithms? How to pick the best one?

No free lunch

"No system – even the universe itself – can give guarantees about prediction, control, or observation."

– David Wolpert

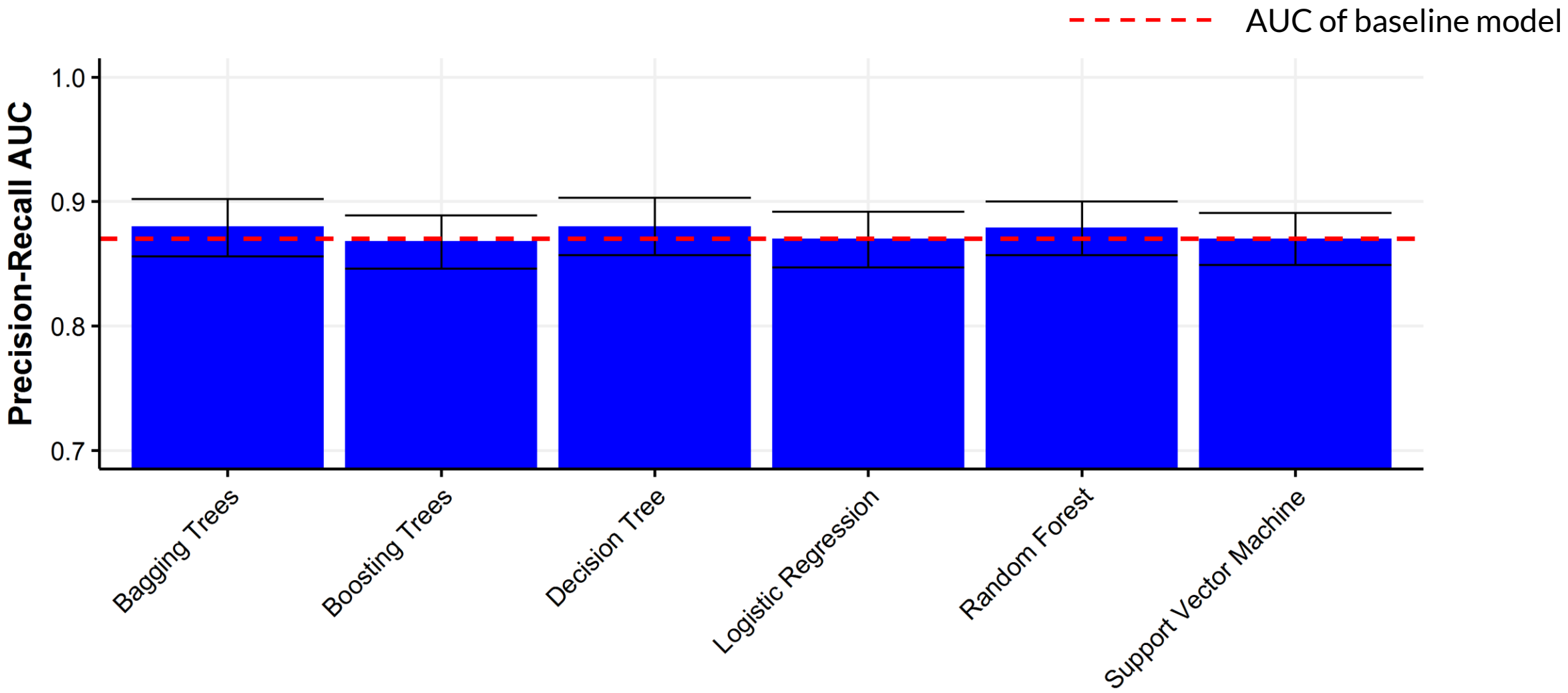
Implications of NFL theory:

- There is no universal learning algorithm
- You should empirically test and compare multiple learning algorithms, to see which performs best for the task at hand

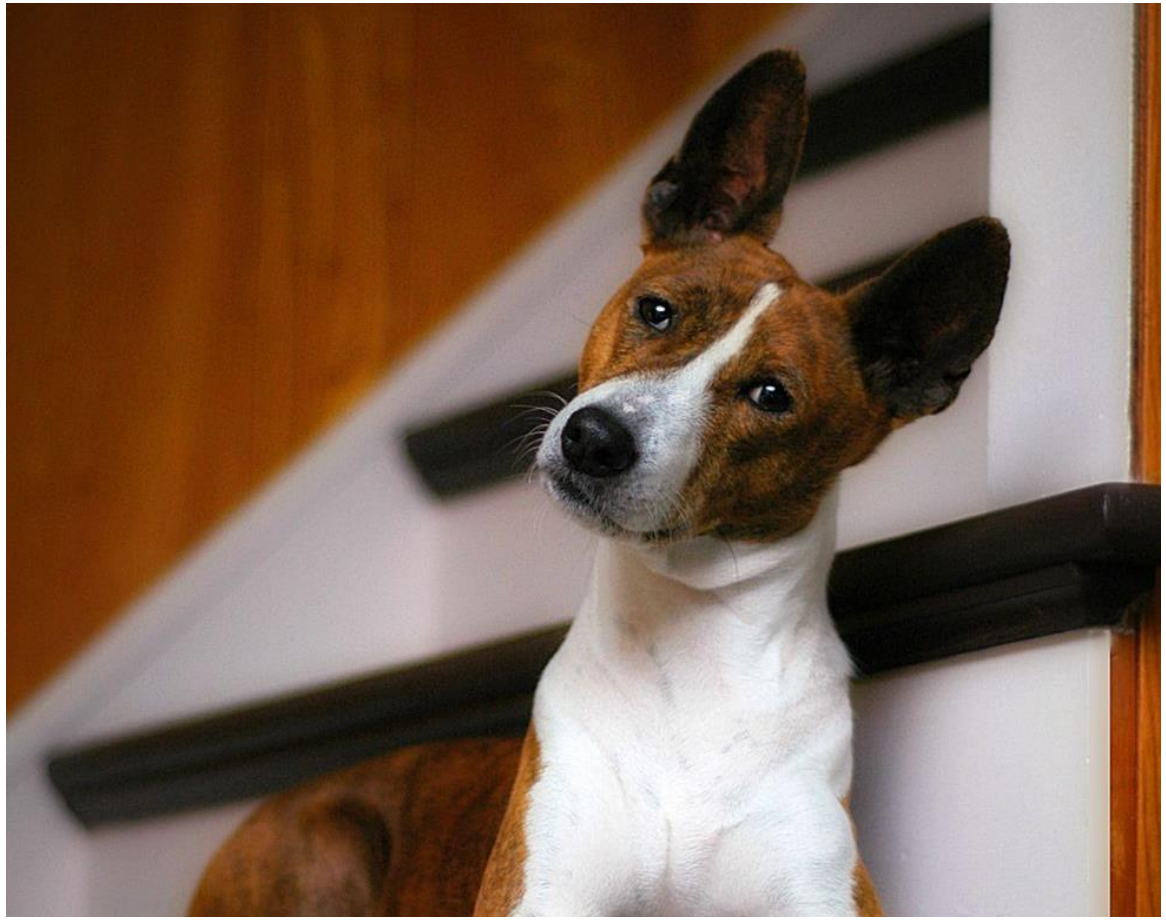
Task 1: Predicting Treatment Success

- 411 cases with available treatment success data
- Treatment success rate was 74%
- We will assess 6 different learning algorithms and report AUC of precision-recall curve
- Results are compared to a null **baseline classifier** that indiscriminately predicts positive

Task 1: Predicting Treatment Success

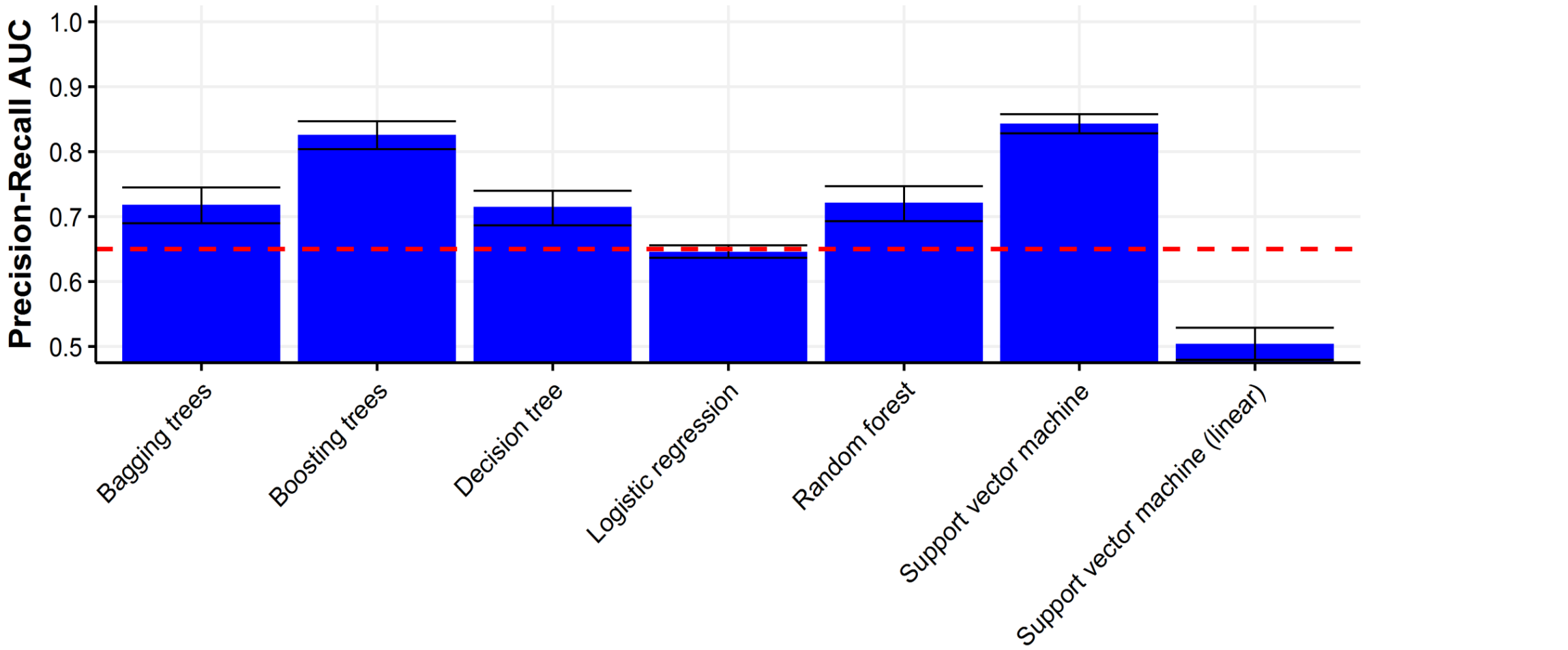


Task 2: Predicting Case/Control Status



- 2,249 cases and controls
- 657 cases -> 29.2% case prevalence
- Same approach

Task 2: Predicting Case/Control Status



Model Transparency

- We want to **trust** that the model makes decisions in a rational way
- **Feature importance** is a measure of how much the model relies on a particular feature when making a prediction
- Highly **transparent** models – feature importance data can be gathered during training or is evident in the description of the model itself


Human and Machine Learning pp 279-292 | Cite as

Trust and Transparency in Machine Learning-Based Clinical Decision Support

Authors | Authors and affiliations

Cosima Gretton


Machine intelligence in healthcare —perspectives on trustworthiness, explainability, usability, and transparency

Christine M. Cutillo, Karlie R. Sharma, Luca Foschini, Shinjini Kundu, Maxine Mackintosh, Kenneth D. Mandl & MI in Healthcare Workshop Working Group

npj Digital Medicine **3**, Article number: 47 (2020) | Cite this article

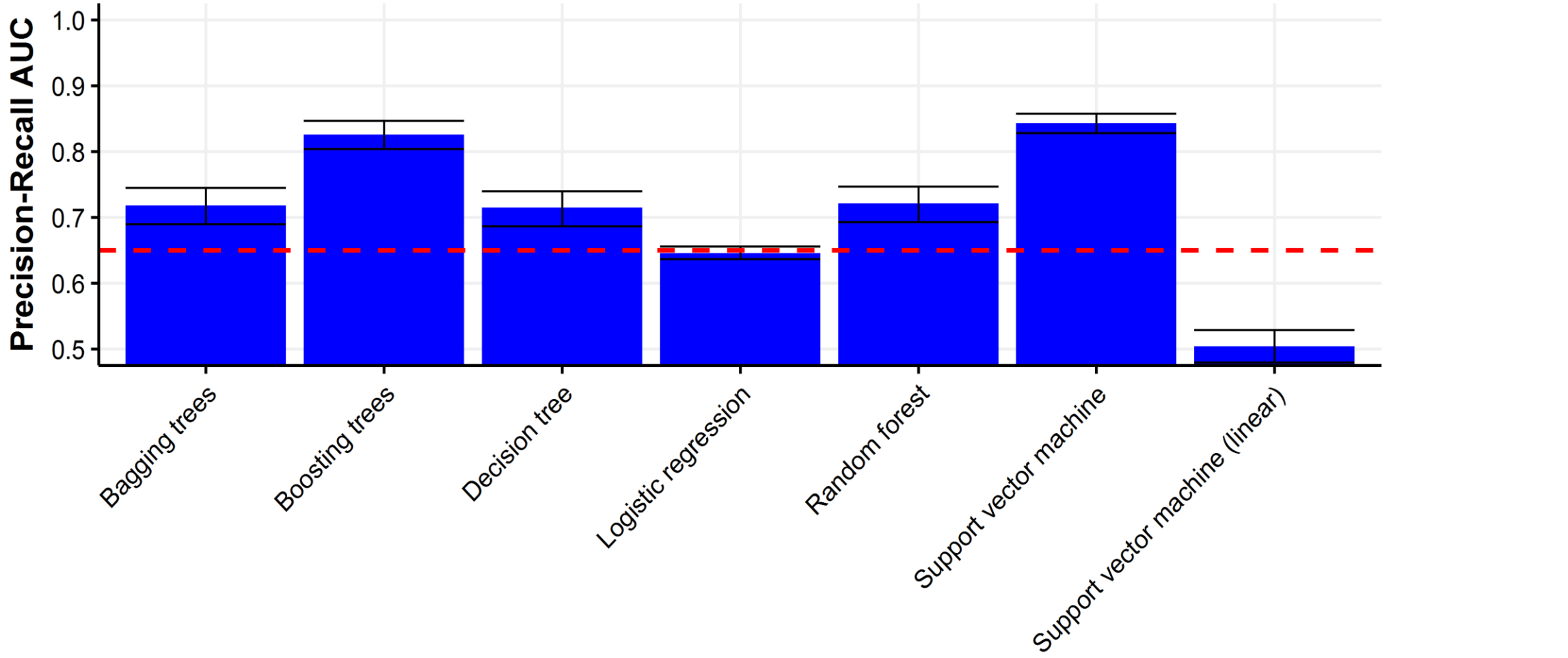

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

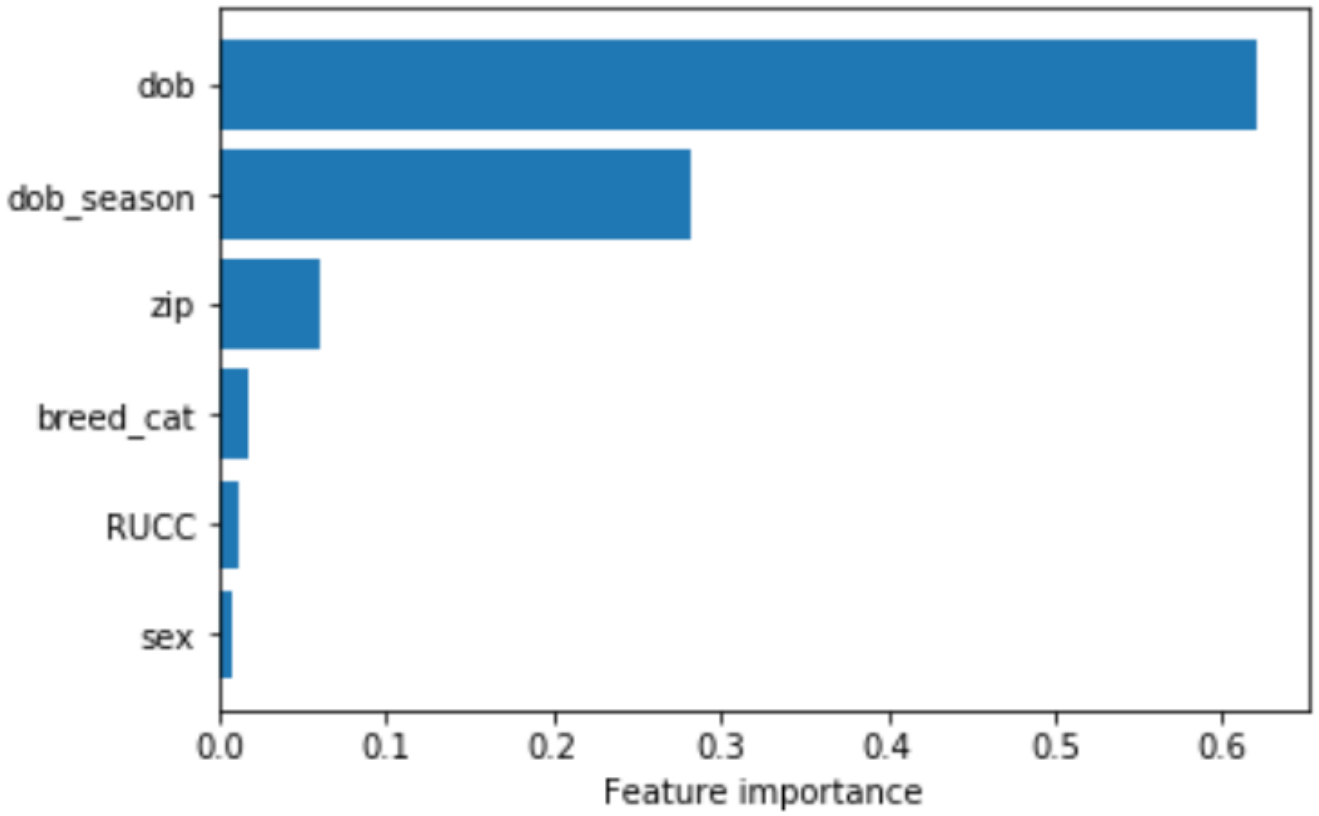
Cynthia Rudin

Nature Machine Intelligence **1**, 206–215(2019) | Cite this article

Task 2: Predicting Case/Control Status



Feature importance of GBT model (reduction in Gini impurity)



Is there a relationship between CAD case status and date of birth?

Outline

- Canine atopic dermatitis
- Introduction to machine learning
- Modeling classification tasks for canine atopic dermatitis
- Evaluating model performance
- Comparing machine learning algorithms
- Important takeaways

#1: Evaluation methodology is critical

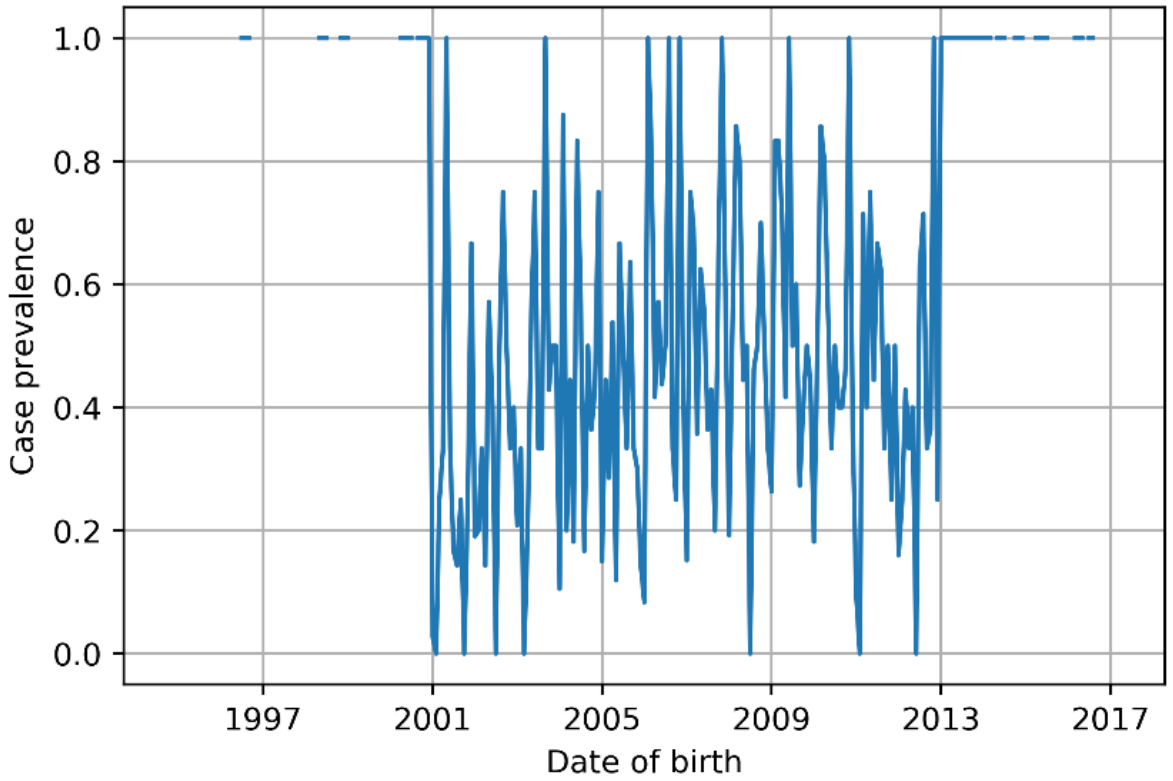
- Every machine learning model needs to be evaluated against a collection of independent data
- The training set should not include any information about the test data
- The statistic used should be appropriate for the situation (accuracy may not be best)

#2: Feature importance does not designate a linear relationship

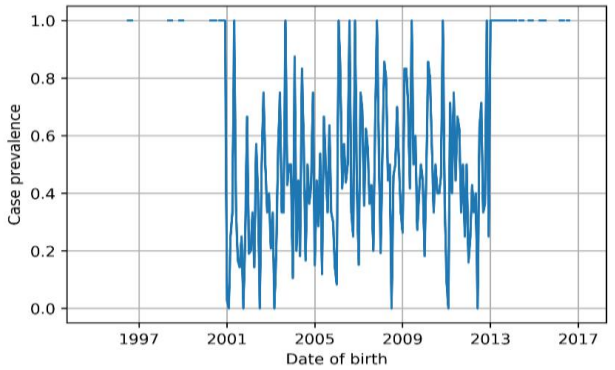
- Strong importance does not designate a linear relationship between the feature value and the log-odds of the case status, as in logistic regression
- We do not conclude that case frequency increases for animals born at a later time, only that there is some relevant signal in the date of birth feature
- Sometimes highly important features correlate negatively with the output, or there could be a non-linear relationship

#3: Do not accept results at face value

Future case prevalence for each cohort of patients born in the same month

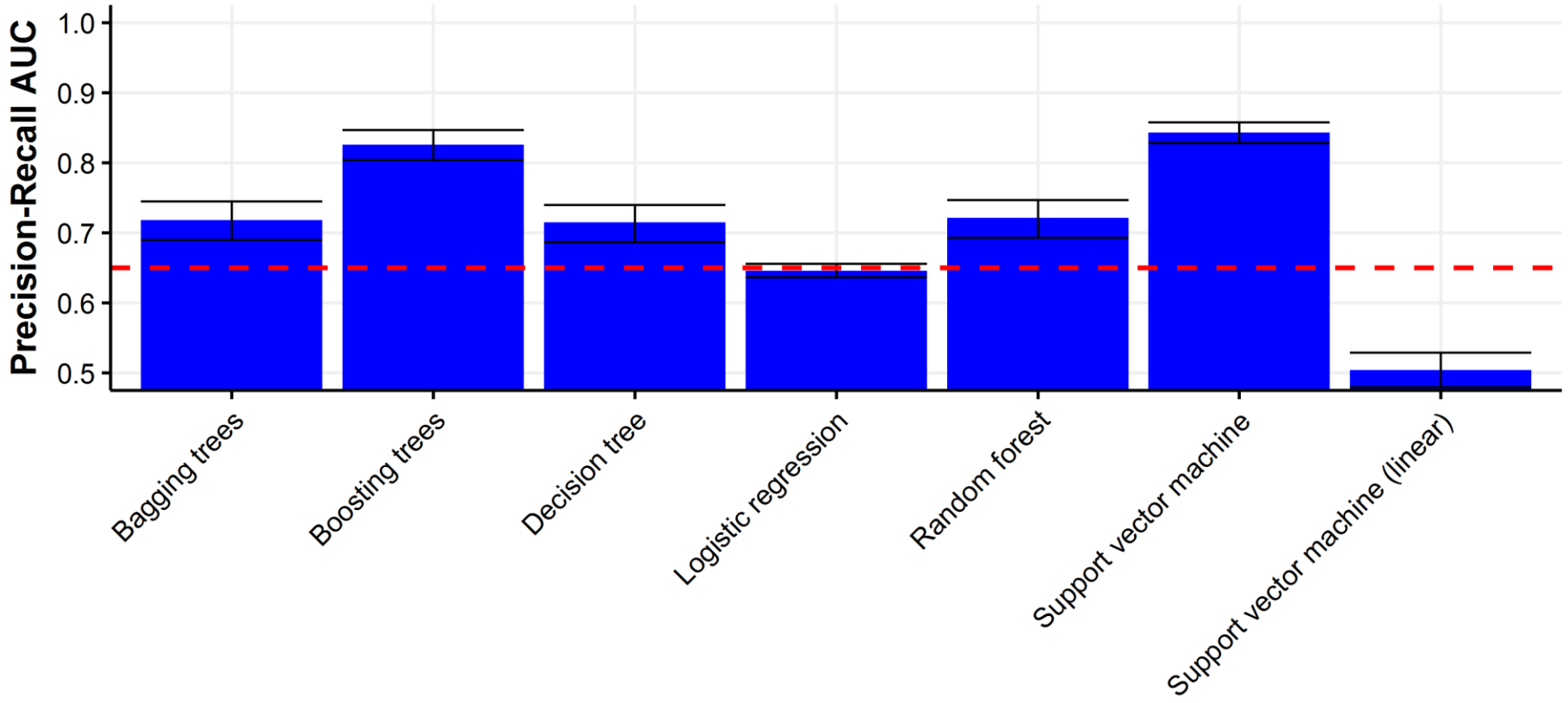


Did something go wrong?



- Did database queries used to extract controls utilize a smaller date range than those that were used to extract cases?
- Is there a bug in the code used to generate this figure?
- Is there some extraneous explanation for this pattern, such as changes in clinic population, marketing, or in the underlying database infrastructure?
- WHAT HAPPENED: The code used to assemble the data set contained a bug that inadvertently caused dates in the control group to be incorrectly converted.

#3: Do not accept results at face value… because the result dramatically changes when we run the experiment without the date of birth feature



But... there are many success stories

ELSEVIER

Domestic Animal Endocrinology
Volume 72, July 2020, 106396

Machine learning algorithm as a diagnostic tool for hypoadrenocorticism in dogs

K.L. Reagan [a], B.A. Reagan [b], C. Gilor [c]

https://doi.org/10.1016/j.domaniend.2019.106396

VetTag: improving automated veterinary diagnosis coding via large-scale language modeling

Yuhui Zhang, Allen Nie, Ashley Zehnder, Rodney L. Page & James Zou

npj Digital Medicine **2**, Article number: 35 (2019) | Cite this article

https://rdcu.be/b6fhU

Machine learning for syndromic surveillance using veterinary necropsy reports

Nathan Bollig, Lorelei Clarke, Elizabeth Elsmo, Mark Craven

https://doi.org/10.1371/journal.pone.0228105

ELSEVIER

Computers and Electronics in Agriculture
Volume 169, February 2020, 105163

PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases

Sarah Valentin [a, b, c], Elena Arsevska [a, c], Sylvain Falala [a], Jocelyn de Goër [d], Renaud Lancelot [a, c], Alizé Mercier [a, c], Julien Rabatel [c], Mathieu Roche [b, c]

https://doi.org/10.1016/j.compag.2019.105163

Bayesian and Machine Learning Models for Genomic Prediction of Anterior Cruciate Ligament Rupture in the Canine Model

Lauren A. Baker, Mehdi Momen, Kore Chan, Nathan Bollig, Fernando Brito Lopes, Guilherme J. M. Rosa, Rory J. Todhunter, Emily E. Binversie, Susannah J. Sample and Peter Muir

G3: GENES, GENOMES, GENETICS *August 1, 2020 vol. 10 no. 8 2619-2628; https://doi.org/10.1534/g3.120.401244*

#4: Machine learning cannot solve every problem

Important Question:

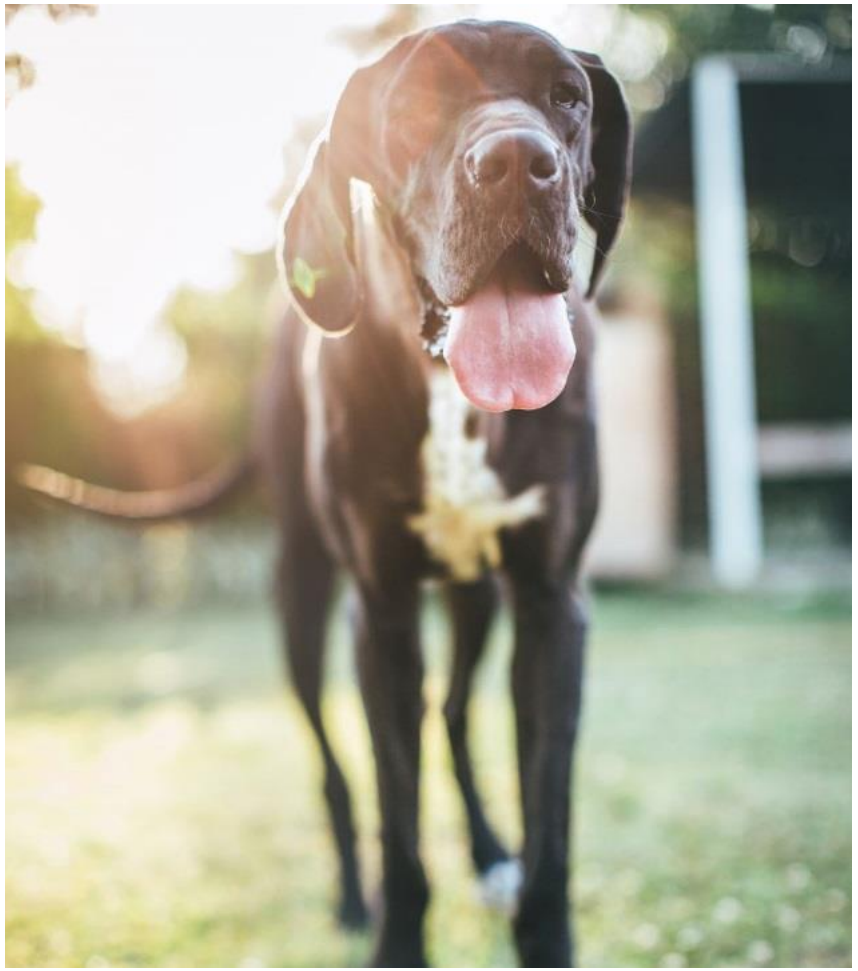
> **Does the training data carry signal relevant to the desired prediction task?**

What about predicting what you have for dinner based on an image of the sky above your house? Probably unsuccessful.

For machine learning to work, there must be a relationship between the inputs and the output

Machine learning could not solve the tasks posed in this tutorial

#5: Success of data science initiatives depends on quality data management



- When data is high quality, it becomes a rich resource that has the potential to change the quality of life of people and animals
- Structured data (like diagnosis codes) can make all the difference
  - Free-text can be very limiting
- Discrete problem list better than using patient returns as proxy for treatment success

#6: Be skeptical and communicate cautiously

- Evaluate each new development with an open mind
- Success of a ML system depends not just on the generic methodology, but on the (1) specific task and (2) quality/quantity of data

Questions?

- Thanks to Douglas DeBoer and Dörte Döpfer for their assistance with this work.
- Article and Code: https://github.com/nathanbollig/ML-for-veterinarians
- Contact: nbollig@wisc.edu

AVI Association for Veterinary Informatics

#BetterDataSavesPets

Education Opportunities:

Annual Talbot Symposium

VMX 2021

Virtual Education, TBA

Follow AVI on LinkedIn & Twitter:

@avinformatics

#Talbot20

#BetterDataSavesPets

Sources for images

- https://www.goodhousekeeping.com/life/pets/g4531/cutest-dog-breeds/
- https://www.sciencemag.org/news/2019/11/here-s-better-way-convert-dog-years-human-years-scientists-say
- https://mashable.com/video/automatic-dog-pet-scratch/
- https://www.plupetstore.com/3-dog-blogs-youll-want-read.html
- https://thebark.com/content/vet-advice-relief-your-dogs-itchy-skin
- http://www.vetstar.com/support/OLD_brochures/Vetstar%20Brochure.pdf
- https://www.bragmedallion.com/blog/authors-there-is-no-such-thing-as-a-free-lunch/
- https://suwalls.com/world/amazing-sunset-sky-above-the-forgotten-house-on-the-field
- https://tigerturf.com/in/how-to-build-a-synthetic-grass-tennis-court/
- https://www.insider.com/funny-dog-photos-faces-portraits-2020-5
- https://shopus.furbo.com/blogs/knowledge/heres-why-two-dogs-are-better-than-one
- https://icanhas.cheezburger.com/dogs/tag/measuring