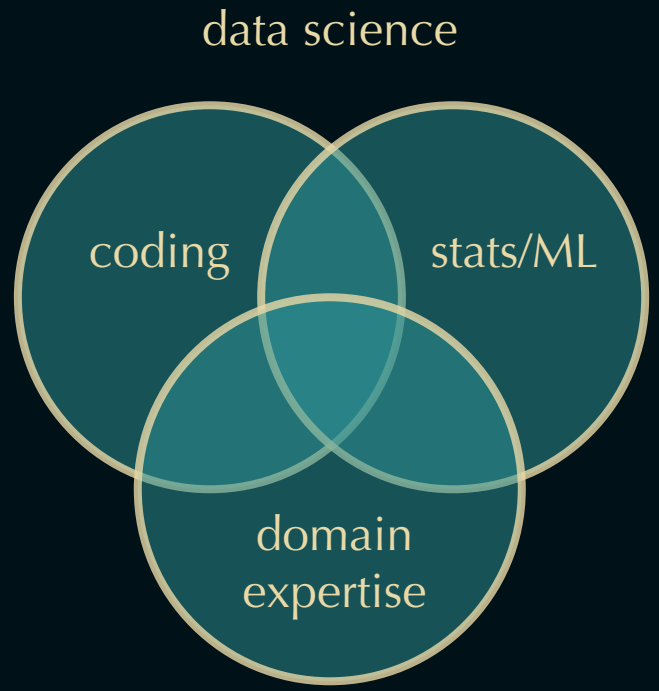
Intro to machine learning (ML) for EEB

About me

- evolutionary genetics, computer science
- departmental data scientist in EEB
  - collab with faculty on rotating basis
  - 1:1 consultations
  - ***intelligible*** (hopefully) workshops for technical training
    - ML, stats, coding, data processing/wrangling

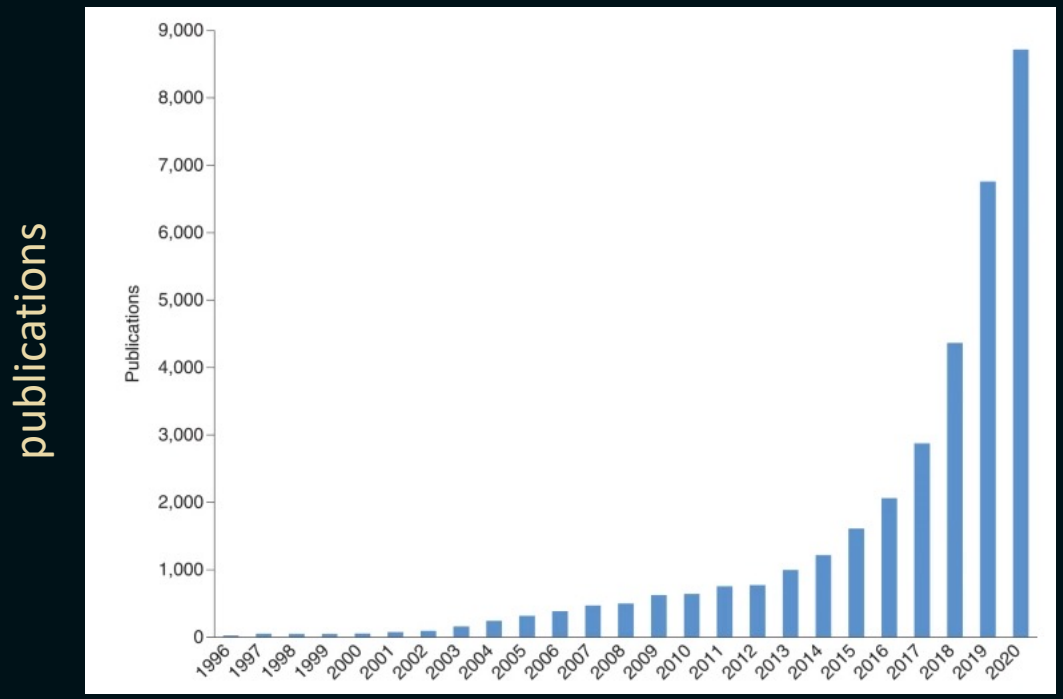


Workshop outline

- Day 1: Introduction
  - ***why*** learn and use ML
  - ***what*** is ML
    - definition
    - workflow: model training and selection
    - 3 popular models with EEB examples
- Day 2: Hands-on data analysis (python)
  - ***how*** to use ML
    - classify penguin species

Why learn and use ML

- ML use in biology increasing exponentially
- understanding ML can help
  - read literature
  - navigate collaborations
  - use/adapt the latest models

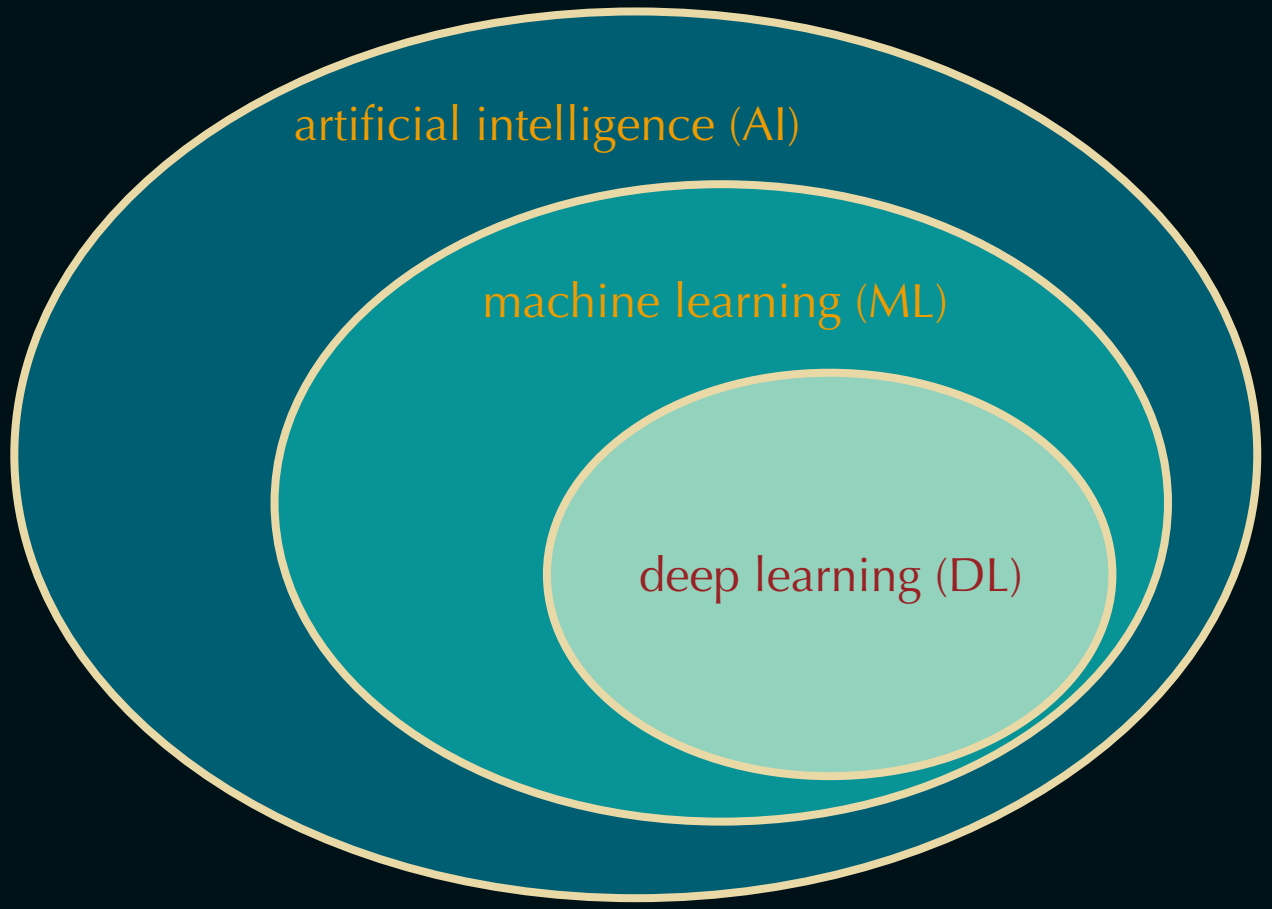


Walsh et al. 2021

- ML useful for
  - pattern recognition
  - predicting unknown quantities
  - making decisions based on data

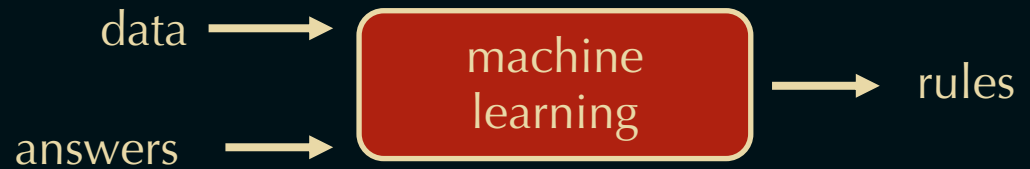
- don't necessarily need large datasets!

What is machine learning?



AI: automate an intellectual **task**

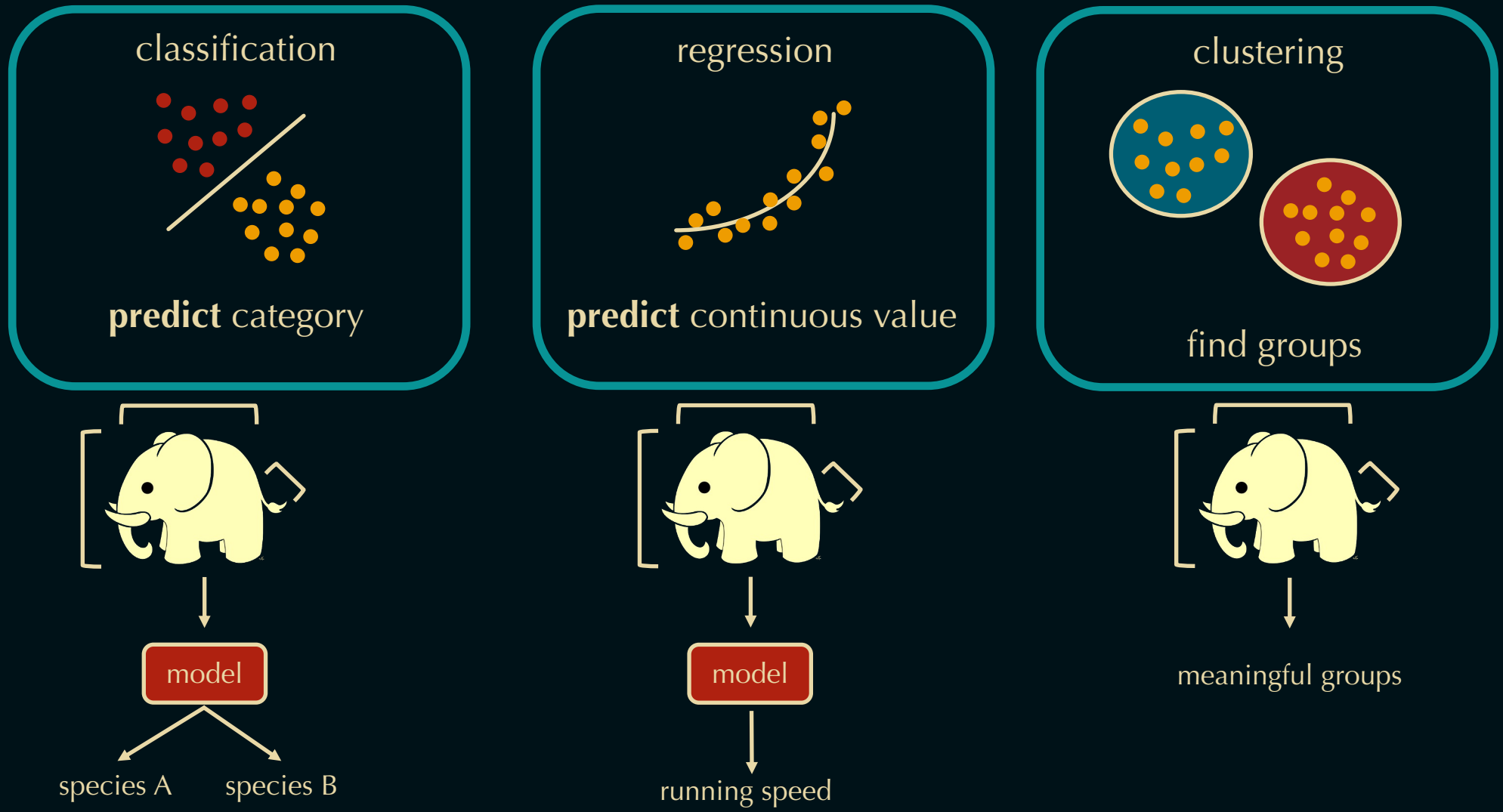
ML: use data to learn rules needed to perform a **task**



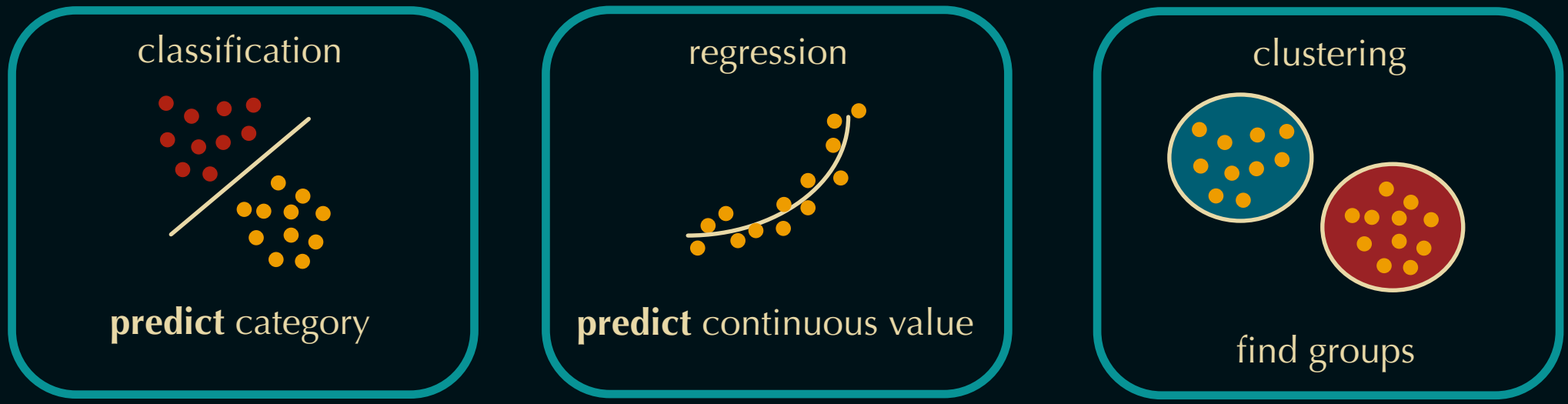
DL: powerful ML approach using neural networks

ML *tasks*

- 3 primary **tasks** in ML
- model **learns** how to do it best, using **data**



ML *tasks*



supervised

unsupervised

response y

features $x_1, x_2 \ldots x_f$

sample $x_1, x_2 \ldots x_f$

learn rules/function to **predict** response from features

learn groups from features

Examples of responses

sample $\left\{ \begin{array}{cc} y & \text{response} \\ x_1, x_2 \ldots x_f & \text{features} \end{array} \right.$

response y

- categorical
  - phenotype (disease status, resistance)
  - biological grouping (species, cell type)
  - ecological (predator/prey)
- continuous
  - phenotype (growth rate, body weight)
  - ecological parameters (species richness, population size, range)

Examples of features

sample $\left\{ \begin{array}{cc} y & \text{response} \\ x_1, x_2 \ldots x_f & \text{features} \end{array} \right.$

features $(x_1, x_2 \ldots x_f)$

- categorical
  - molecular biology (DNA mutation, motif)
  - environmental (habitat, soil type)
- continuous
  - pixel intensity
  - molecular biology (gene expression, protein or metabolite concentration)
  - environmental (precipitation, temperature, soil pH)
  - organismal (age, limb length, body temperature, respiratory rate)

Terminology in literature (just as reference)

sample $\left\{\right.$ y response

$x_1, x_2 \ldots x_f$ features

sample
- data point
- observation

y
- response
- target
- class (categorical)
- label (categorical)
- outcome
- dependent variable

$x_1, x_2 \ldots x_f$
- features
- predictors
- descriptors
- attributes
- covariates
- independent variables

learned parameters
- coefficients and intercept
- weights and bias

$y = \beta_0 + \beta_1 x_1 + \cdots$

$y = b + w_1 x_1 + \cdots$

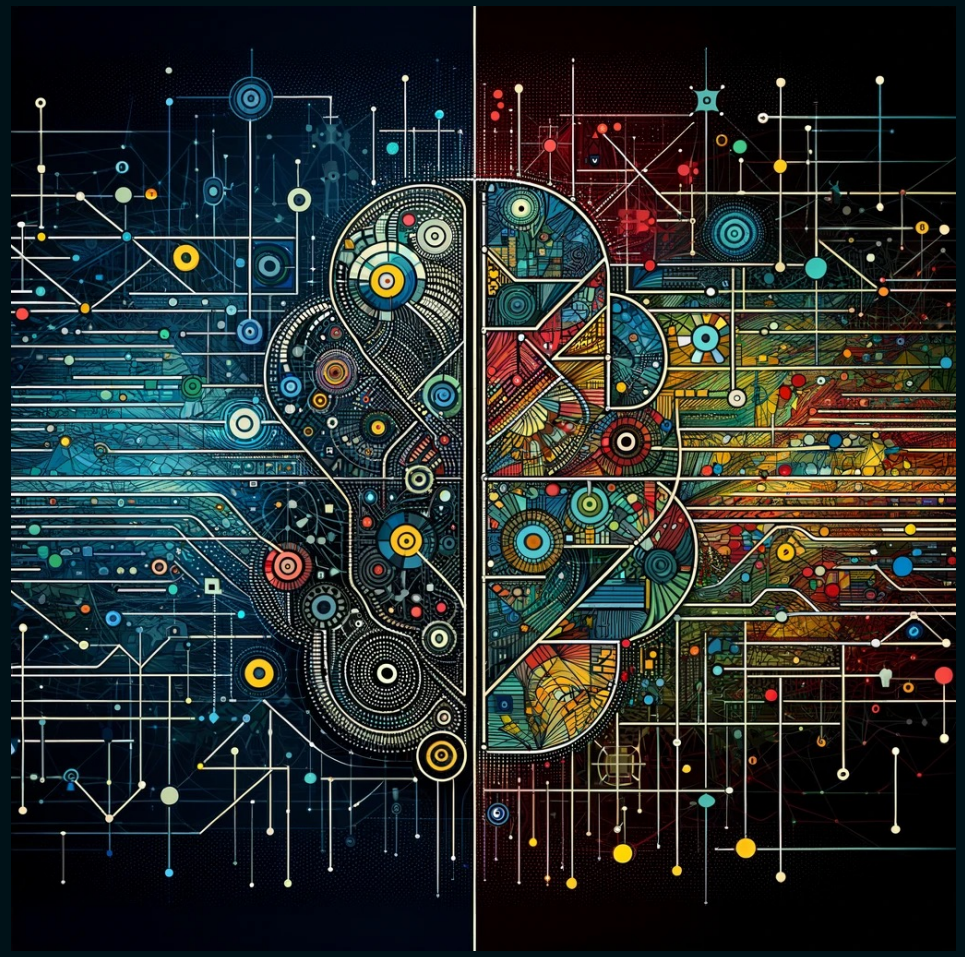
many terms in English, but the **math** is always the same!

What is ML summary

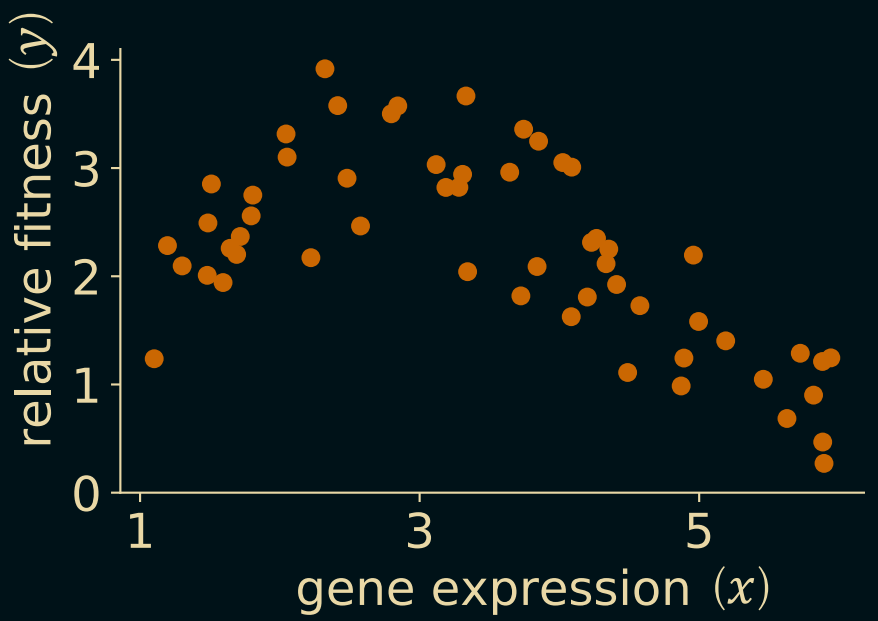
- learn rules to predict a **response** using **features**
  - categorical or continuous
- **tasks**
  - classification
  - regression
  - clustering

} *supervised*

ML workflow: model training and selection



Select model, evaluation metric



goal: regression to predict a continuous value

model: polynomial regression (b/c wiggly line)

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{linear regression}} + \beta_2 x^2 + \beta_3 x^3 + \cdots$$

evaluation metric: Mean Squared Error

$$MSE = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

(**lower** is better!)

straight line (simple)

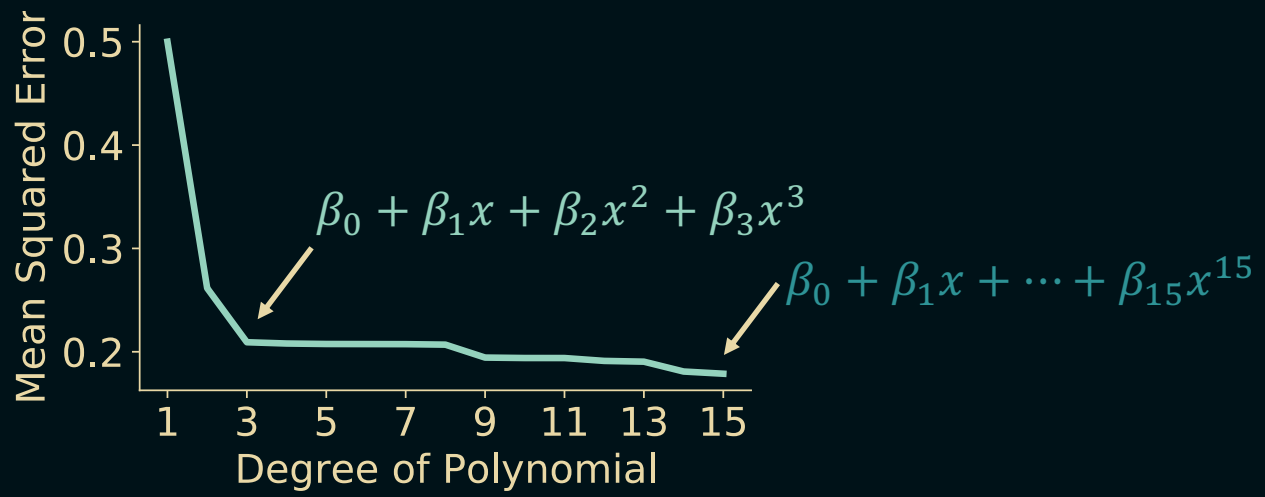
wiggly line (complex)

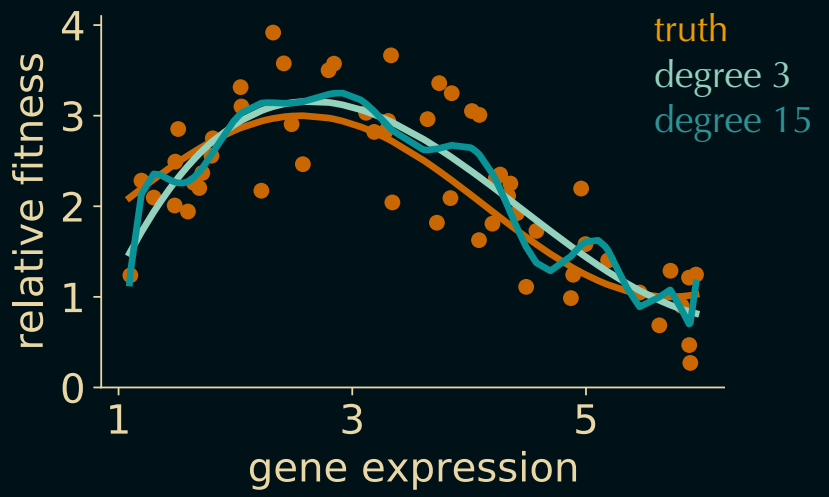
degree	
1	$y = \beta_0 + \beta_1 x$
2	$y = \beta_0 + \beta_1 x + \beta_2 x^2$
...	
15	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{15} x^{15}$

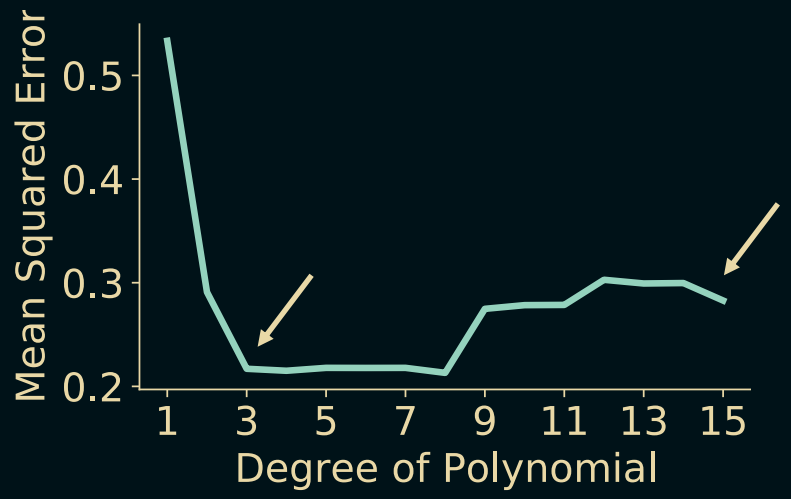
choose "hyperparameters":

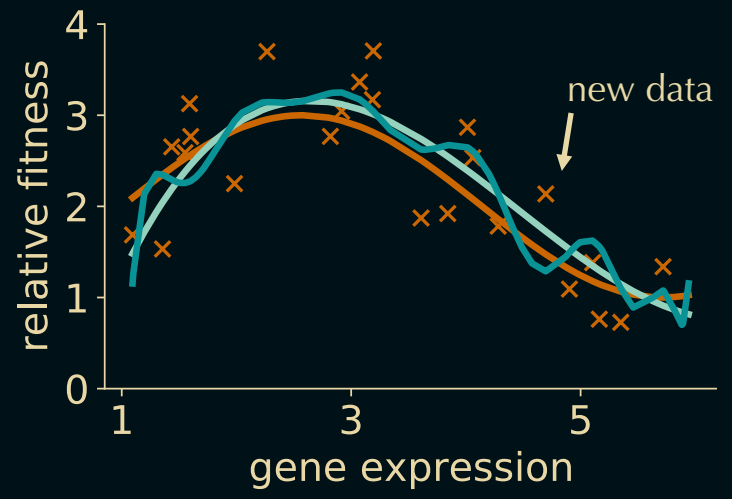
- control model complexity, learning process
- manually set beforehand
- what polynomial degree do we use?
  - let's use 1 through 15

Train and test model

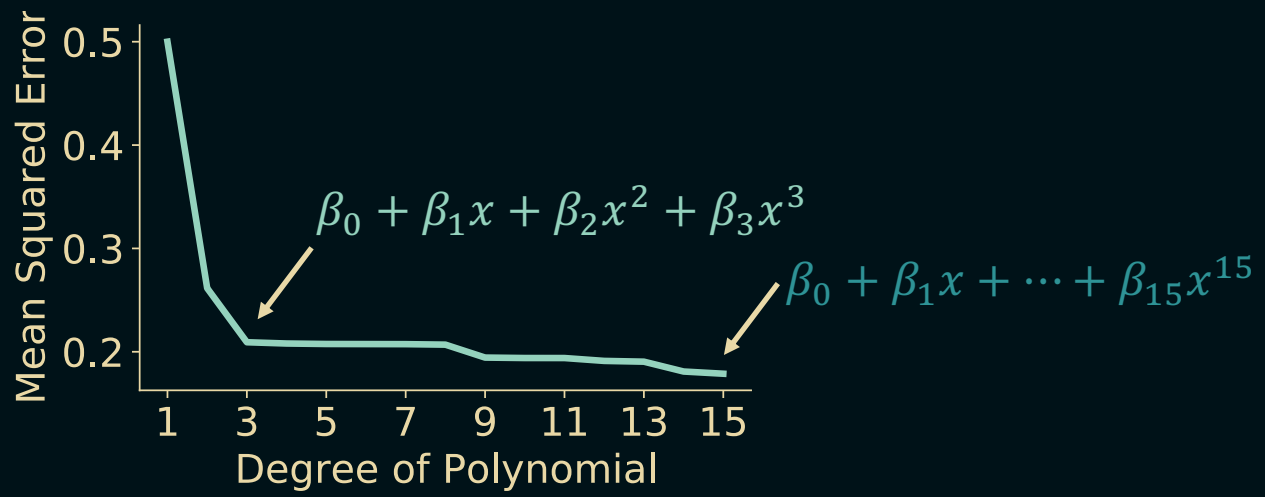


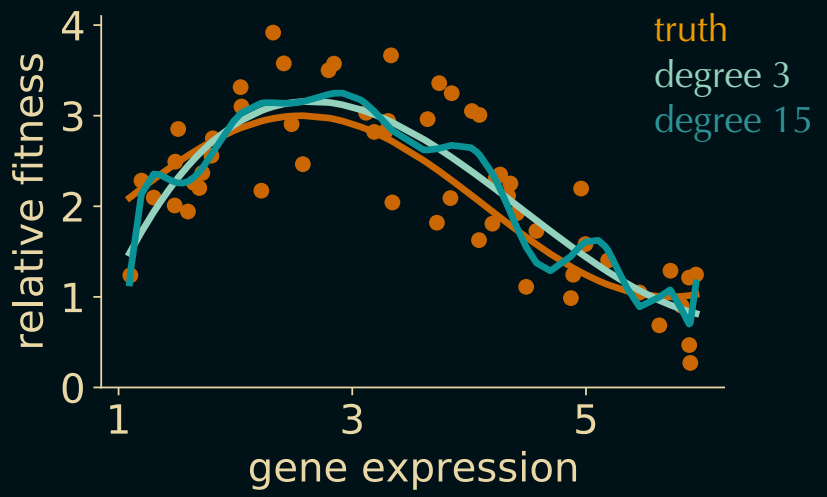


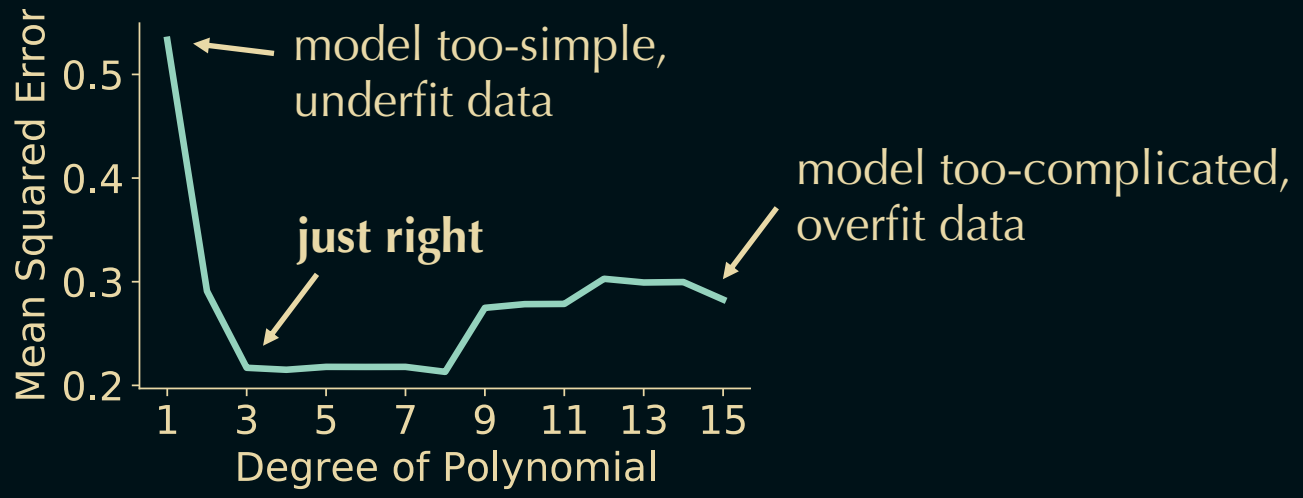


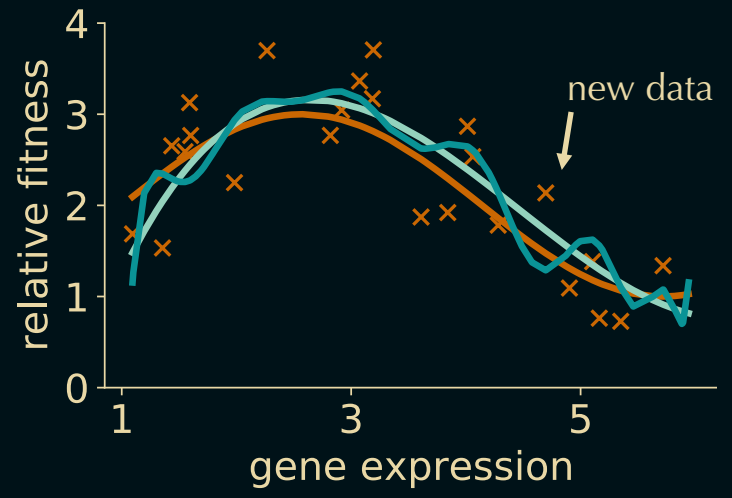


Train and test model







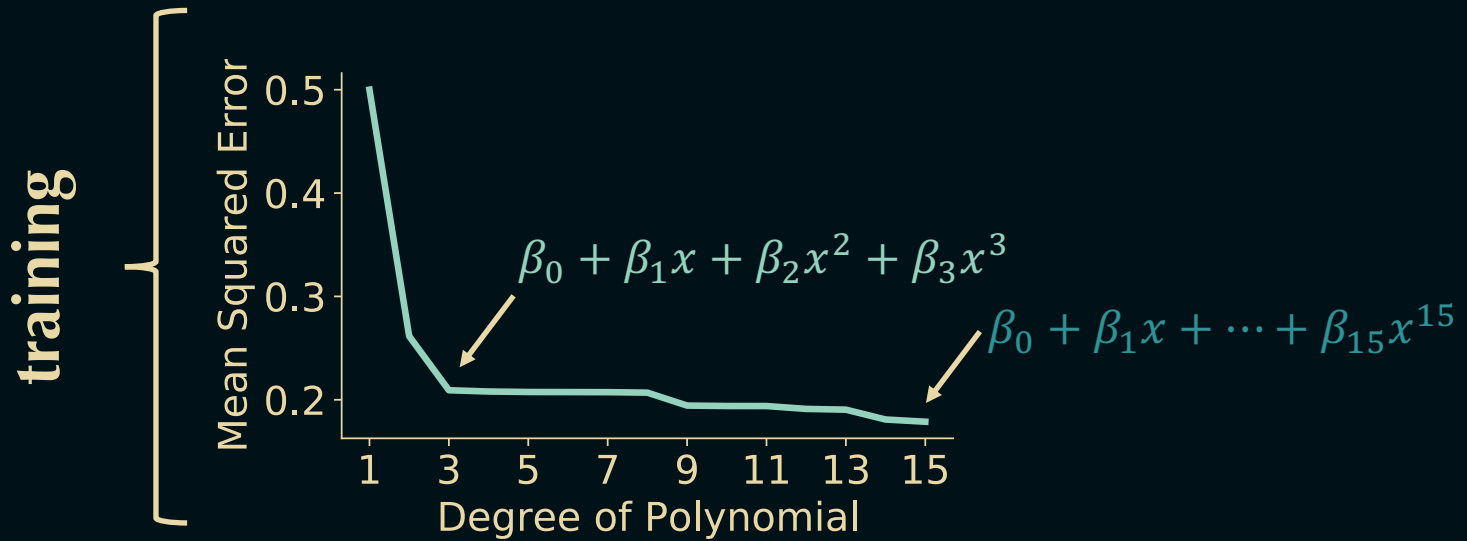


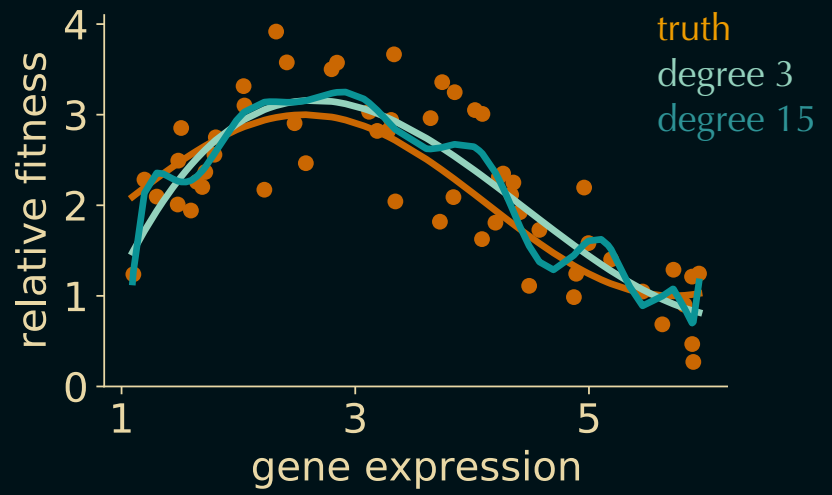
Split data into training and testing

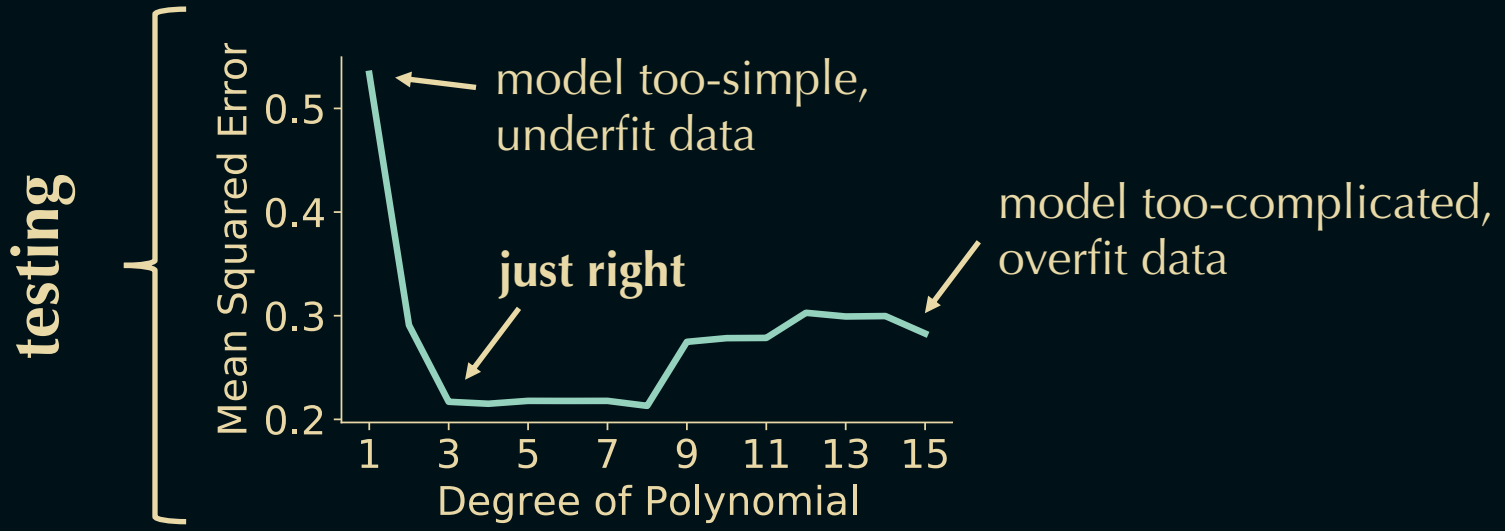
ML solution:

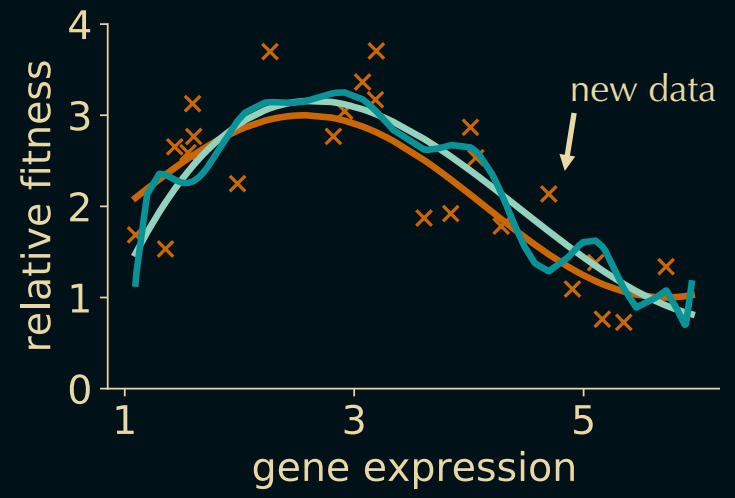
- **train** models on 70-80% of your data
- **test** models' predictions on remaining 20-30%
- choose *simplest* model with highest prediction accuracy!

Train and test model



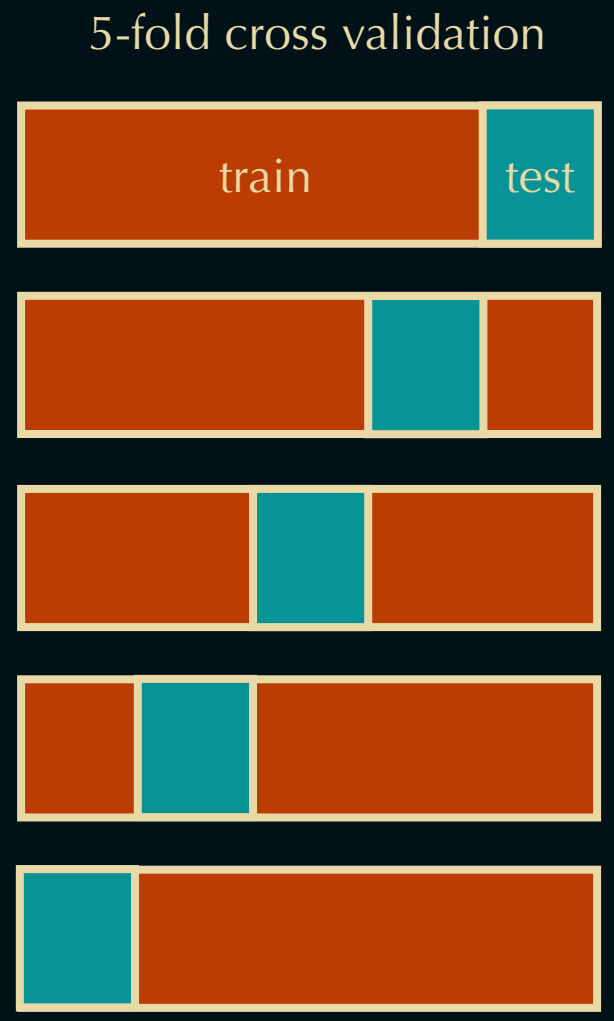






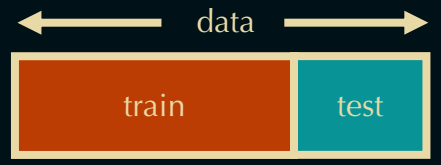
K-fold cross validation

- instead of using a fixed fraction of data for testing
  - divide data into training/testing sets *K* times (e.g. 5)
- yields distribution of accuracy scores, not just point estimate
- not just a hack, statistically valid

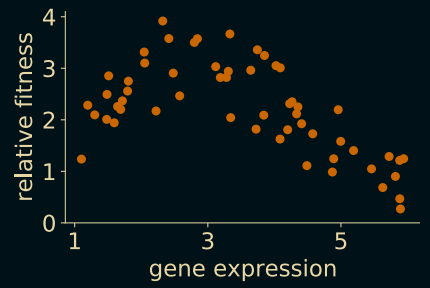


Model training and selection summary

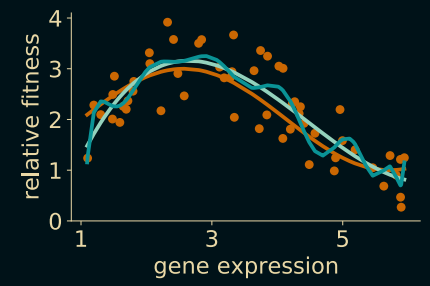
1. split data for training and testing



2. choose
   - base ML model: poly. reg.
   - evaluation metric: MSE
   - hyperparameters: poly. degree 1 – 15



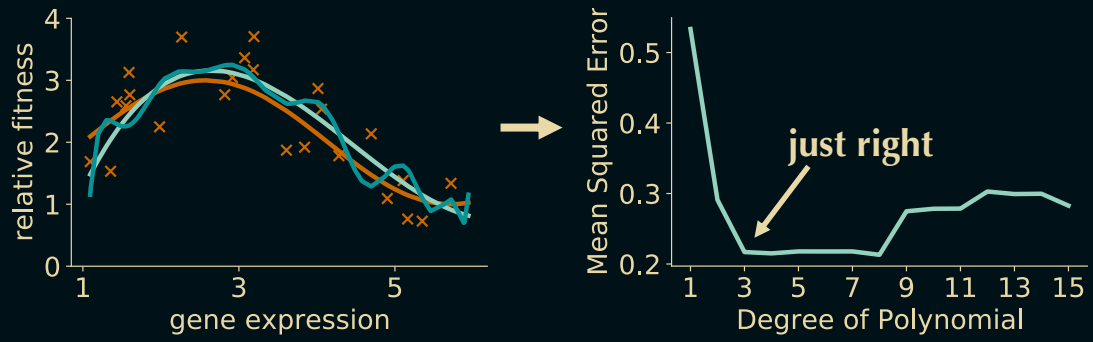
3. fit models on training data, one for each hyperparameter value



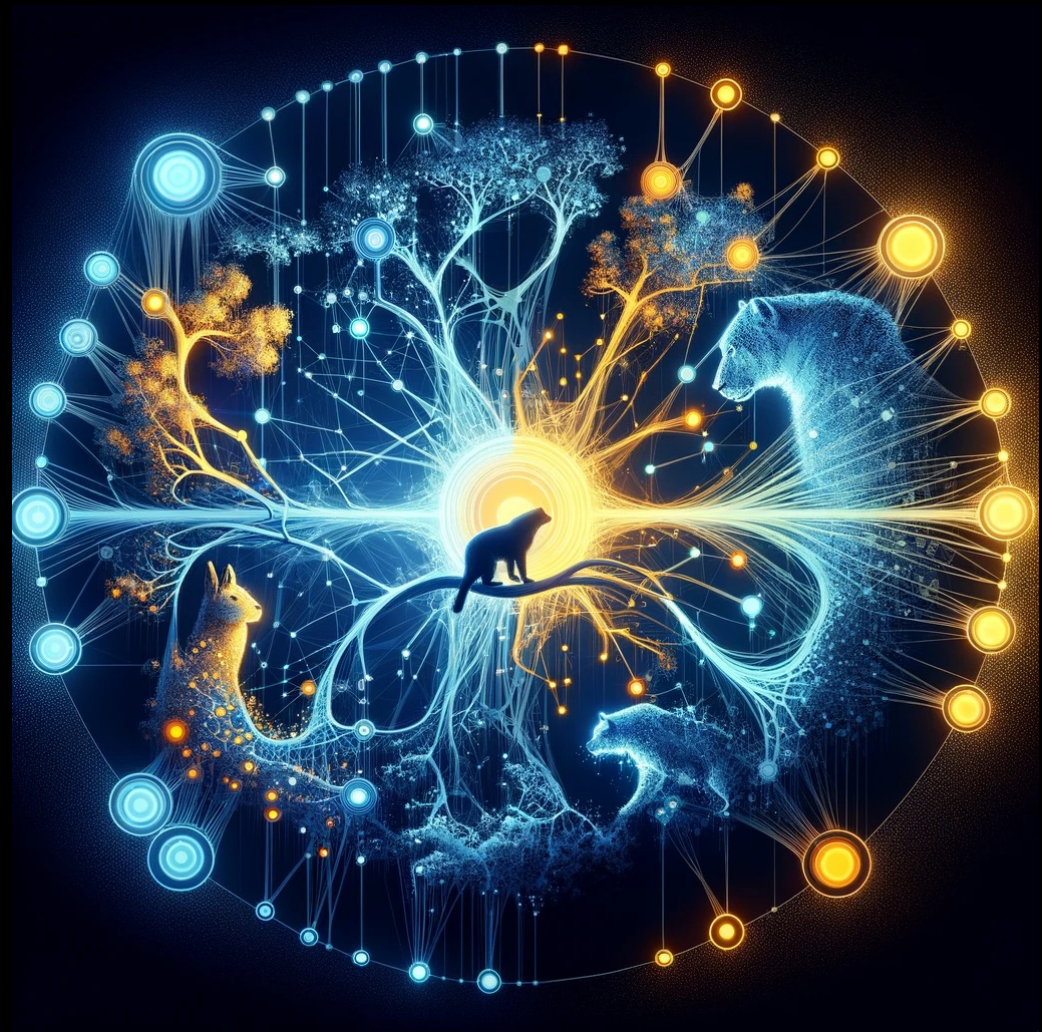
model selection emphasizes predictive power on new data

4. assess predictive performance with test data

5. select simplest model with highest accuracy

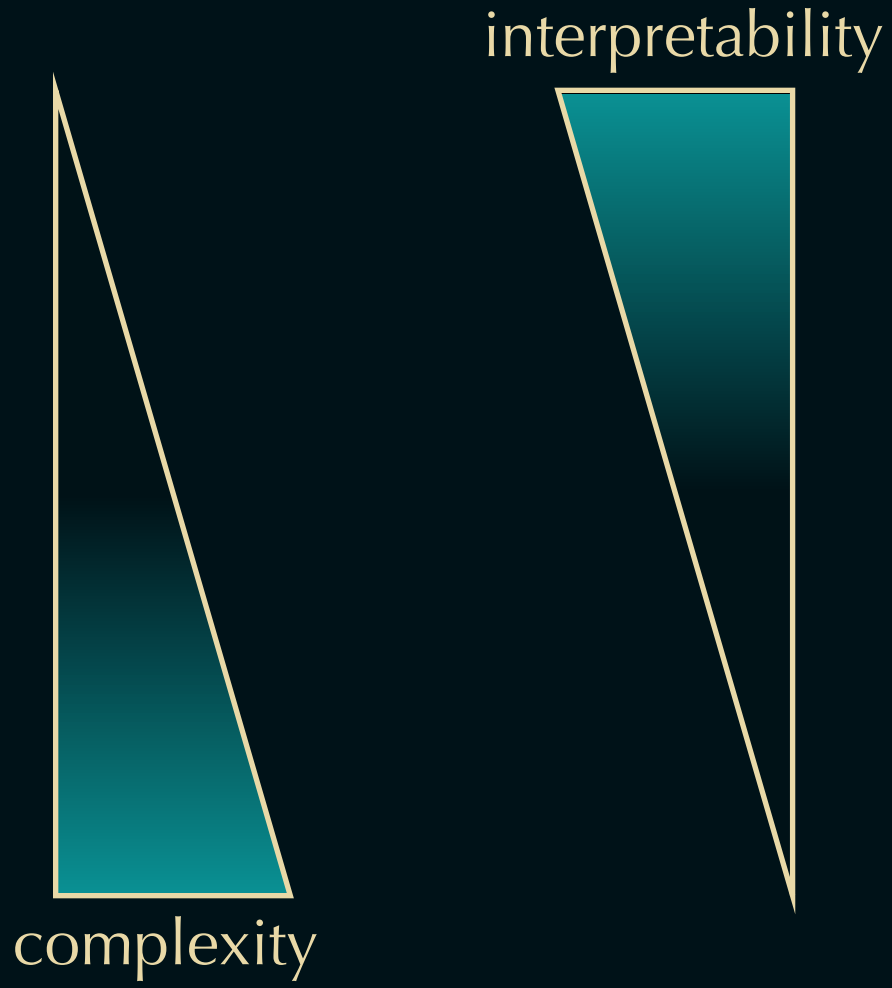


Popular ML models with EEB examples



Popular ML models

- **linear and logistic regression**
- k-nearest neighbors
- naïve Bayes
- support vector machines
- **decision trees, random forests**
- gradient boosting machines
- **neural networks**
  - **many architectures!**



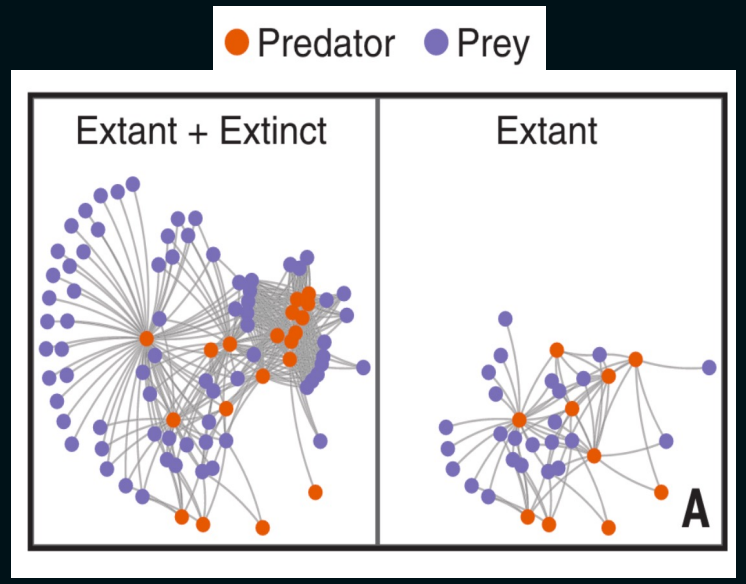
FAUNAL DECLINE

Collapse of terrestrial mammal food webs since the Late Pleistocene

Evan C. Fricke[1,2,3]*, Chia Hsieh[1], Owen Middleton[4], Daniel Gorczynski[1], Caroline D. Cappello[5], Oscar Sanisidro[6], John Rowan[7], Jens-Christian Svenning[8], Lydia Beaudrot[1]

ML to reconstruct food webs

predict interactions



what:

- predict predator-prey interactions from traits

why:

- hard to directly observe predator-prey interactions for some species (esp. extinct ones)
- model allows simulation of food webs with diff species compositions, through time

data:

- morphological, life-history, ecological traits from 6,234 extant and recently extinct mammals
  - **problem**: missing data!
  - **solution**: random forests for imputation

- predator-prey interactions for 1,112 species
  - **problem**: many interactions not observed!
  - **solution**: logistic regression and NN for **classification**

results:

- NN 90% accurate!

Data imputation model

sample $\left\{ \begin{array}{cc} y & \text{response} \\ x_1, x_2 \ldots x_f & \text{features} \end{array} \right.$

sample
- a single organism

response y
- some morphological/ecological trait in the dataset that wasn't measured for some samples

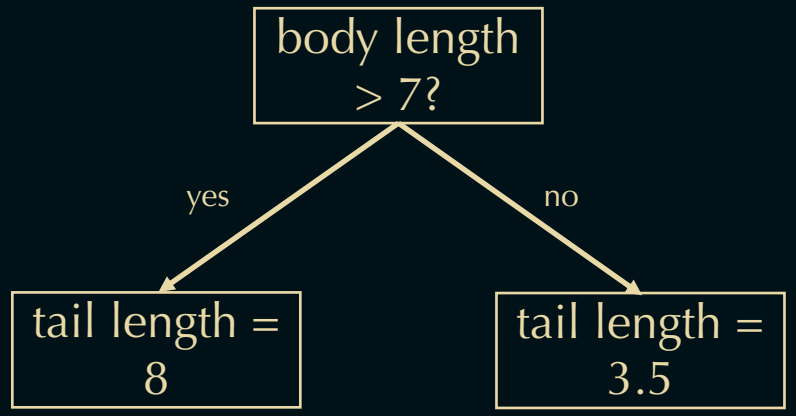
features $(x_1, x_2 \ldots x_f)$
- other morphological features that had data

model
- use decision trees to predict y given $(x_1, x_2 \ldots x_f)$

Decision trees: regression

- split data into smaller subsets that are as homogenous as possible
- within each subset, make a prediction
- **rules to learn**: what splits to make
- **hyperparameters**: how many splits to make
  - learn via training/testing

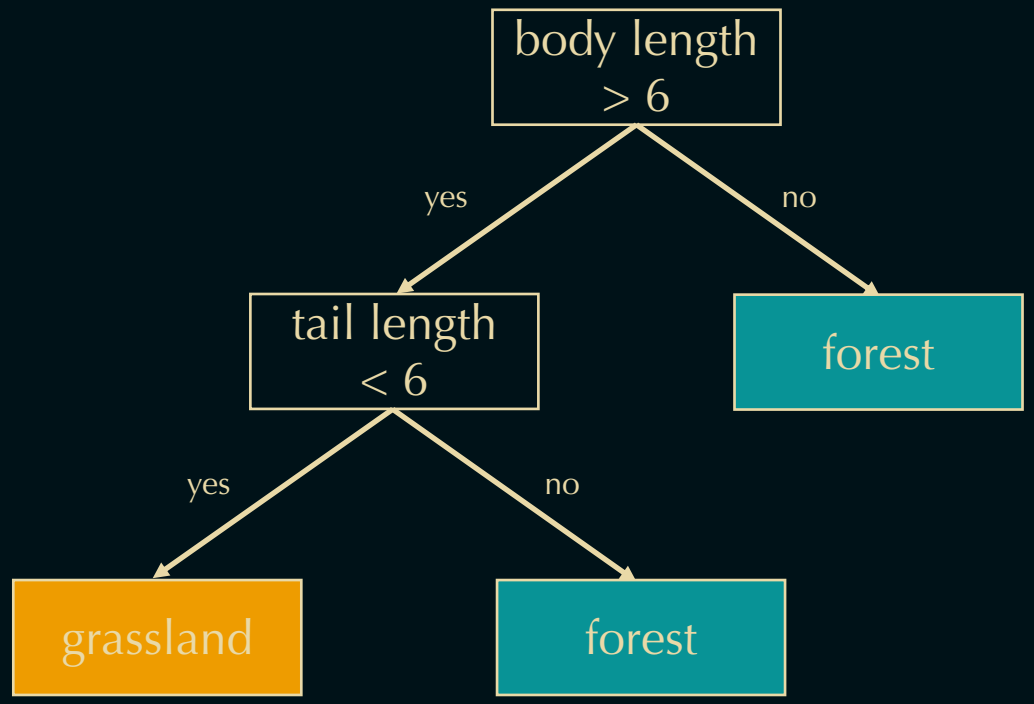
*predictions are means within each subset

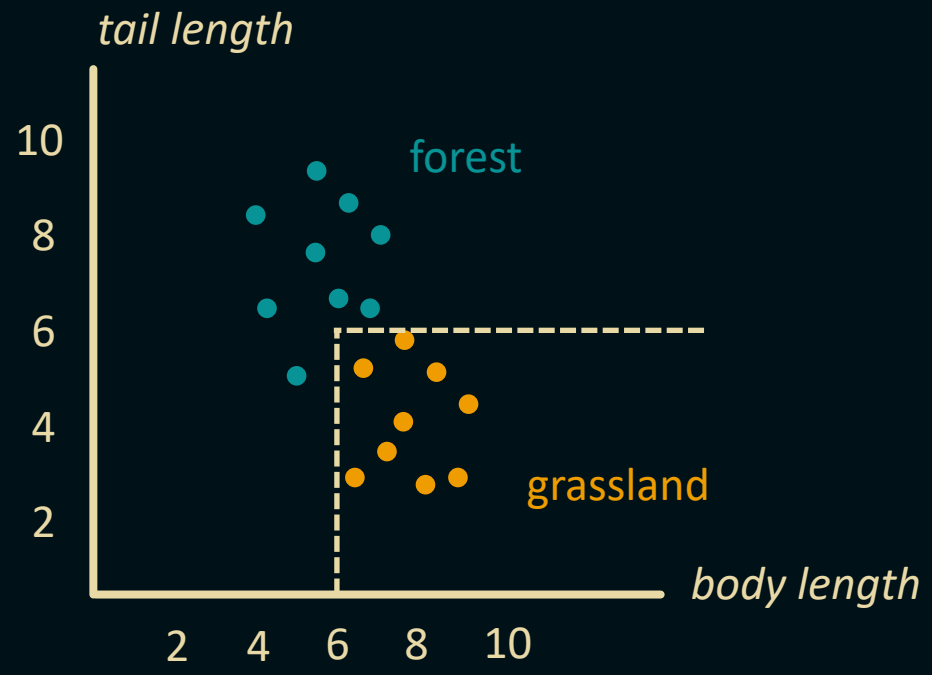


- typically we'd use **many** traits simultaneously
- complex relationships no problem!
- no assumptions about data distributions

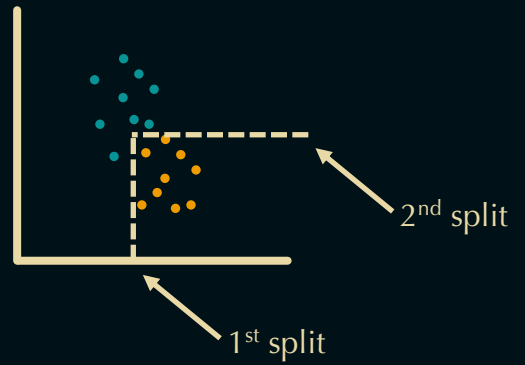
Decision trees: classification

- decision trees can also predict categorical variables, using other continuous **and** categorical variables

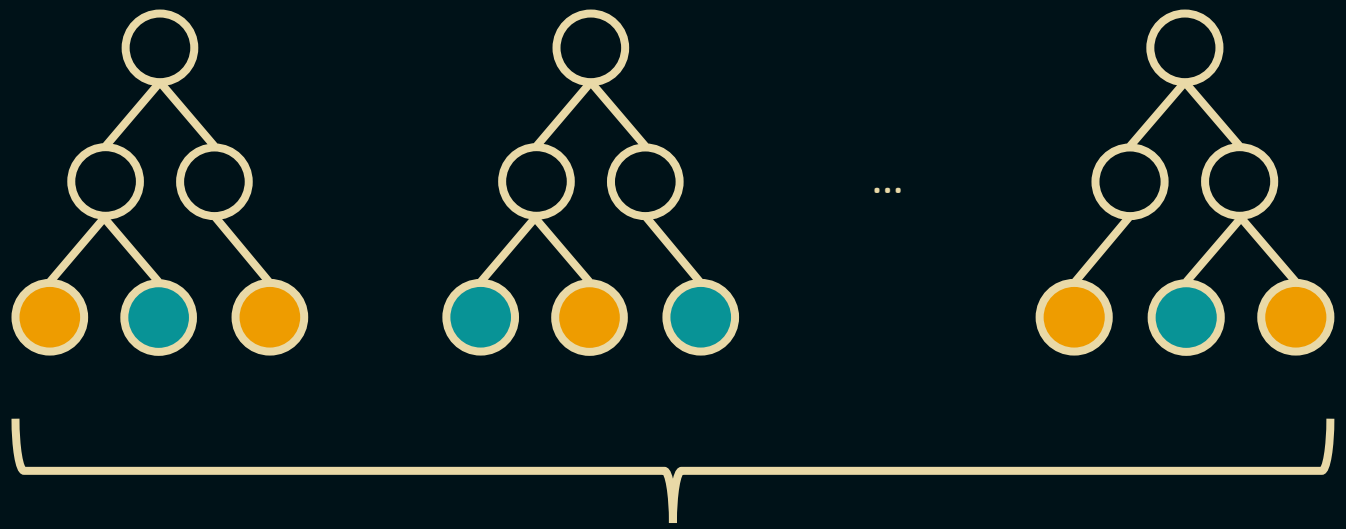




Random forest: ensemble of decision trees



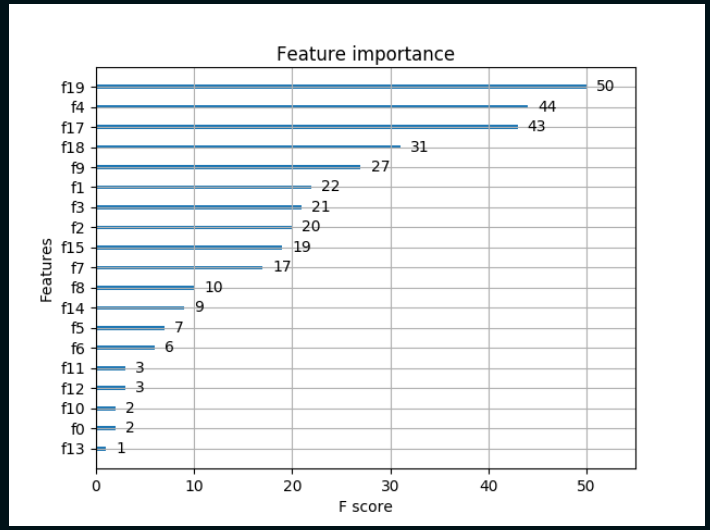
- "random forests" typically outperform a single, optimized decision tree
- ensemble of trees, each made by randomly sampling
  - samples (w/ replacement)
  - features (at each question/split)



- resample mammals
  - resample morph/eco features
    - learn 1st best split
  - resample morph/eco features
    - learn 2nd best split
  - etc.
- repeat N times, where N is # of trees in ensemble
- repeat H times, where H is number of hyperparameter values, # of data splits trees can make

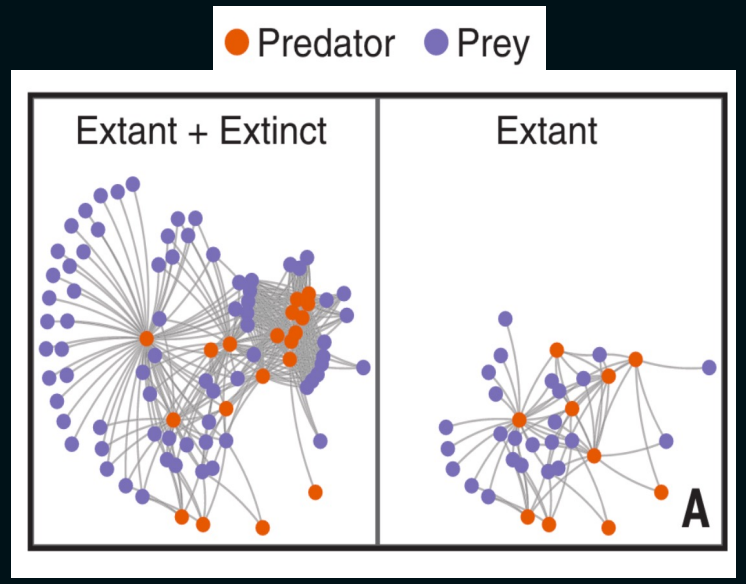
Nice aspects of random forests

- interpretable: feature importance scores show how much each feature contributed to prediction
- make no assumptions about how data are distributed
  - can complement or outperform (G)LMs or (G)LMMs



ML to reconstruct food webs

predict interactions



what:

- predict predator-prey interactions from traits

why:

- hard to directly observe predator-prey interactions for some species (esp. extinct ones)
- model allows simulation of food webs with diff species compositions, through time

data:

- morphological, life-history, ecological traits from 6,234 extant and recently extinct mammals
  - **problem**: missing data!
  - **solution**: random forests for imputation

- predator-prey interactions for 1,112 species
  - **problem**: many interactions not observed!
  - **solution**: logistic regression and NN for **classification**

Species interaction models

sample $\left\{ \begin{array}{cc} y & \text{response} \\ x_1, x_2 \ldots x_f & \text{features} \end{array} \right.$

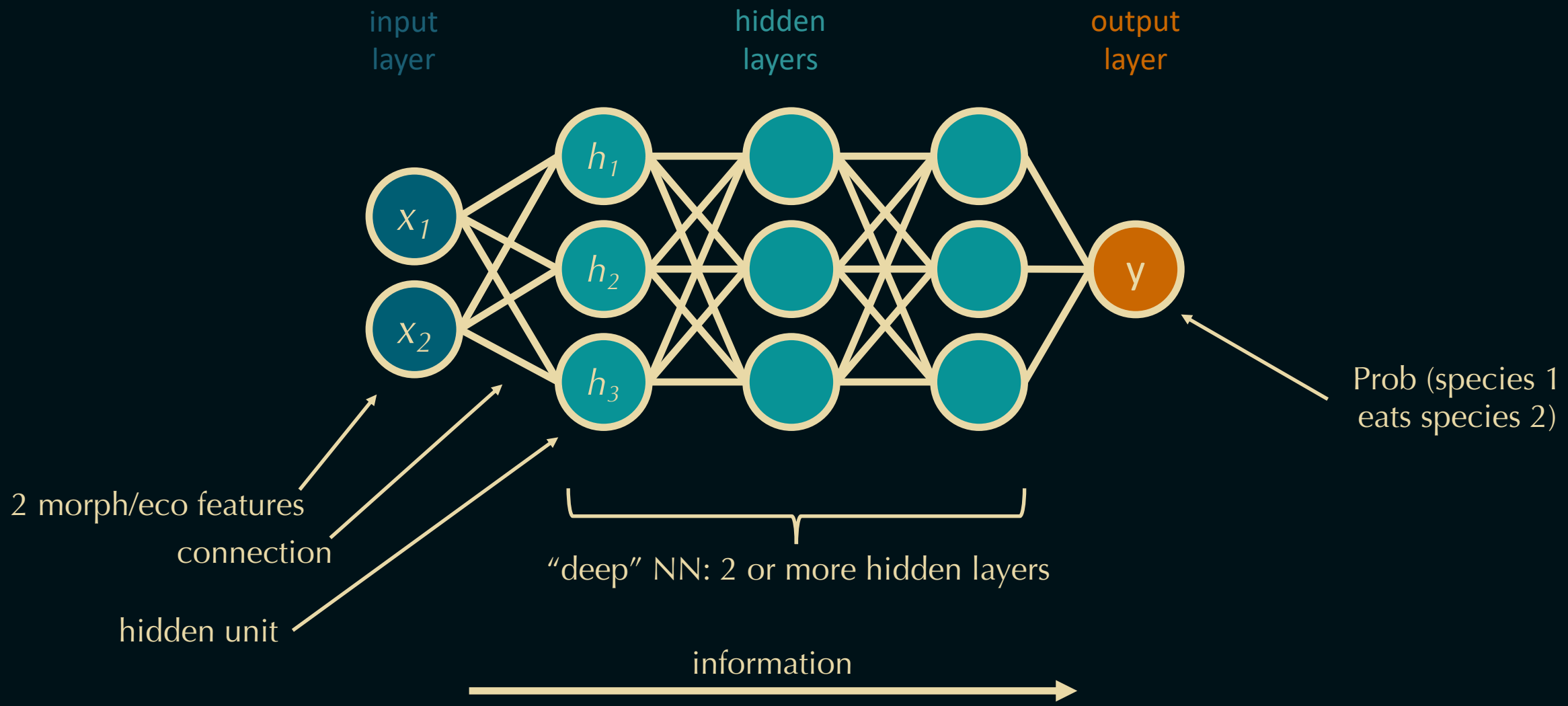
sample
- a pair of individual organisms (diff species)

response y
- probability species A eats species B (categorical)

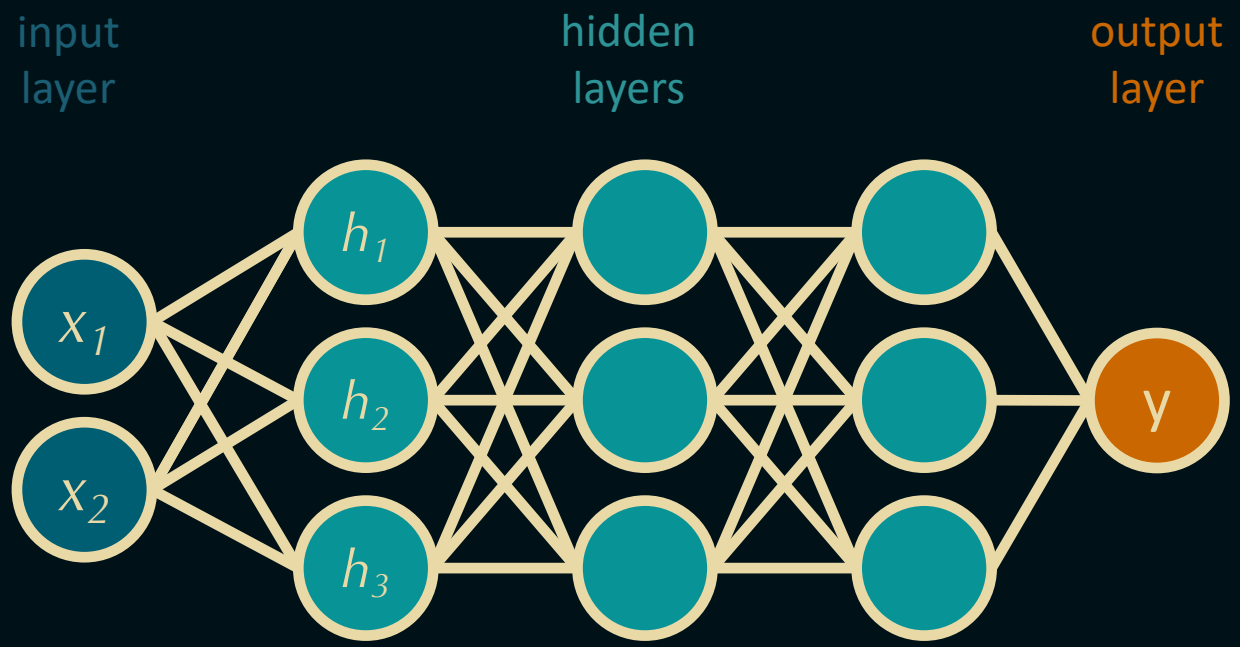
features $(x_1, x_2 \ldots x_f)$
- morphological, life-history, ecological traits **from both species** (continuous and categorical)

models
- use logistic regression and neural networks to predict y given $(x_1, x_2 \ldots x_f)$

Neural networks

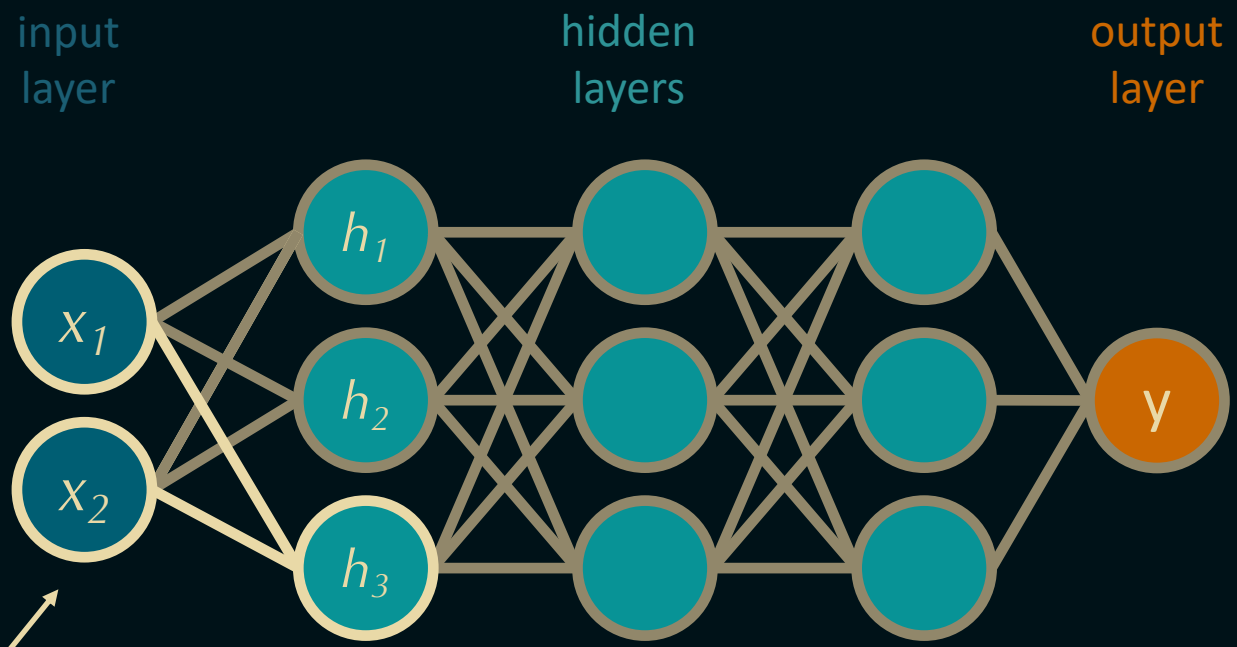


Neural networks



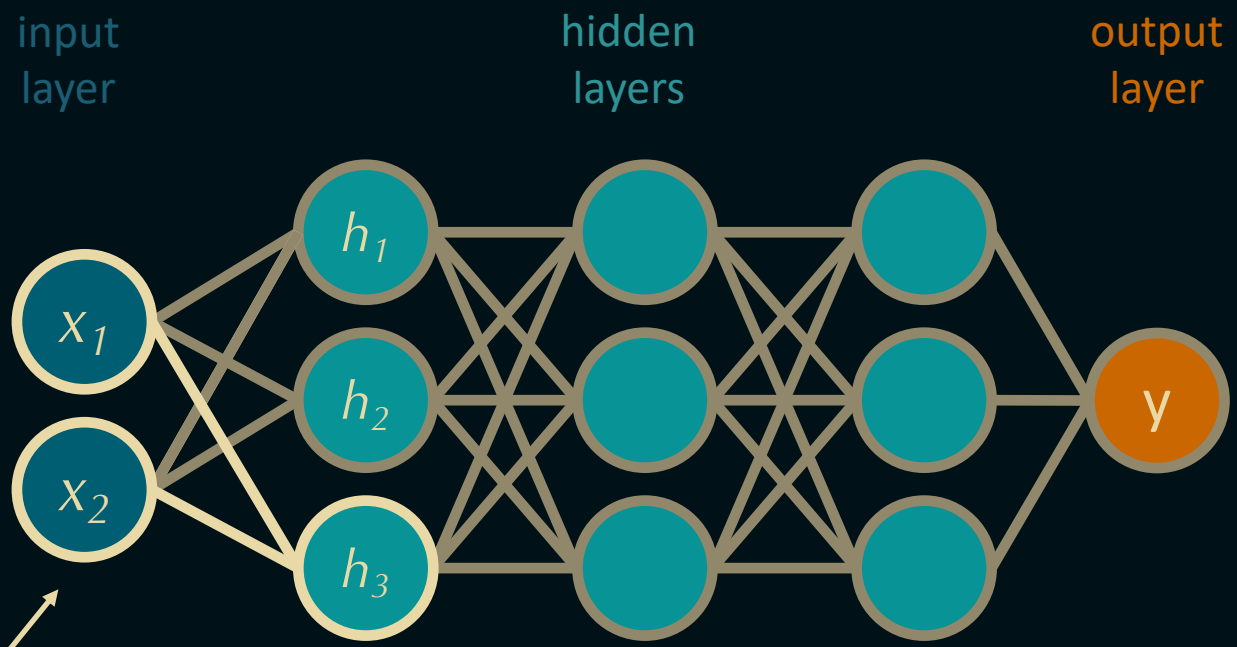
- inspired by brain
- NNs learn how to adjust connections based on input data, which can "turn off" some hidden units

Neural networks



2 morph/eco features

Neural networks

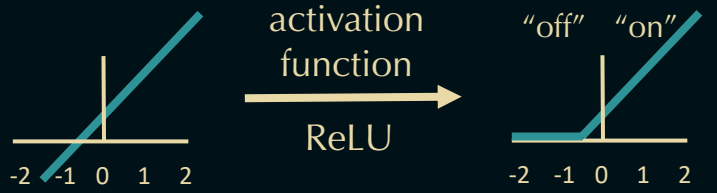


2 morph/eco features

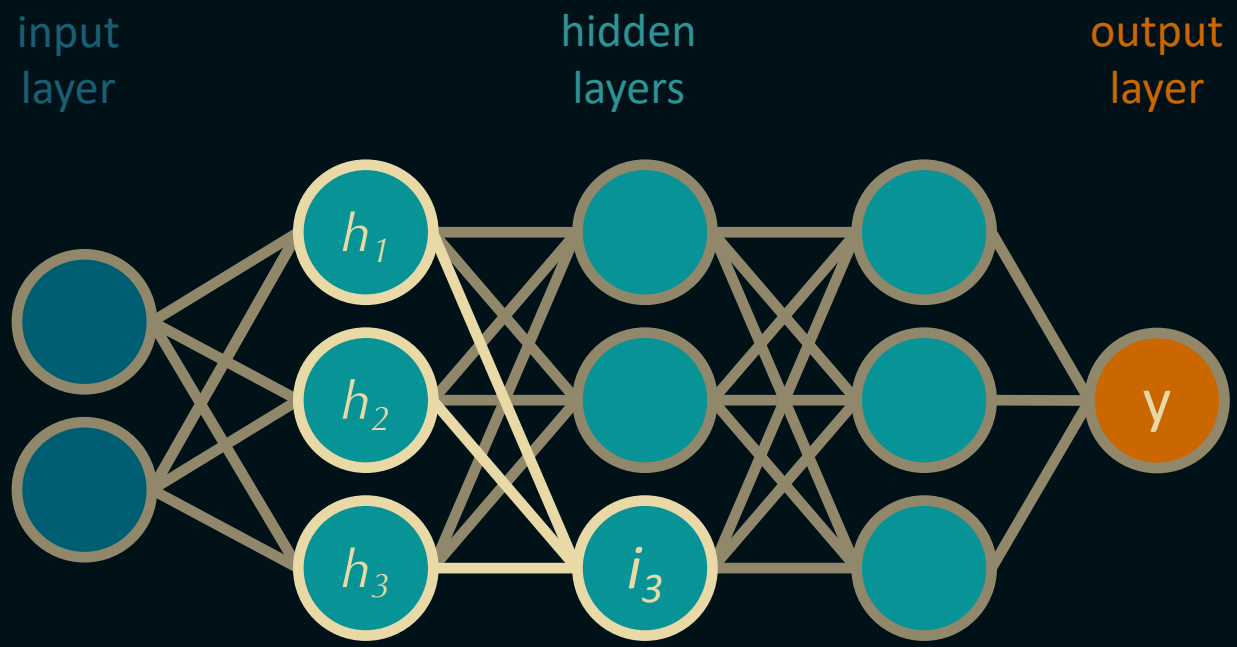
$$h_3 = ReLU(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

activation function

linear function



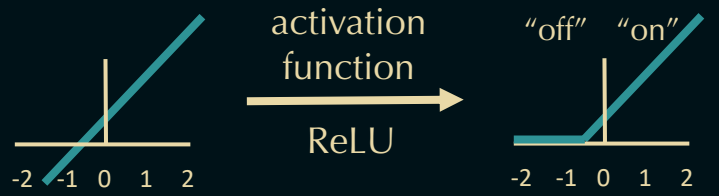
Neural networks



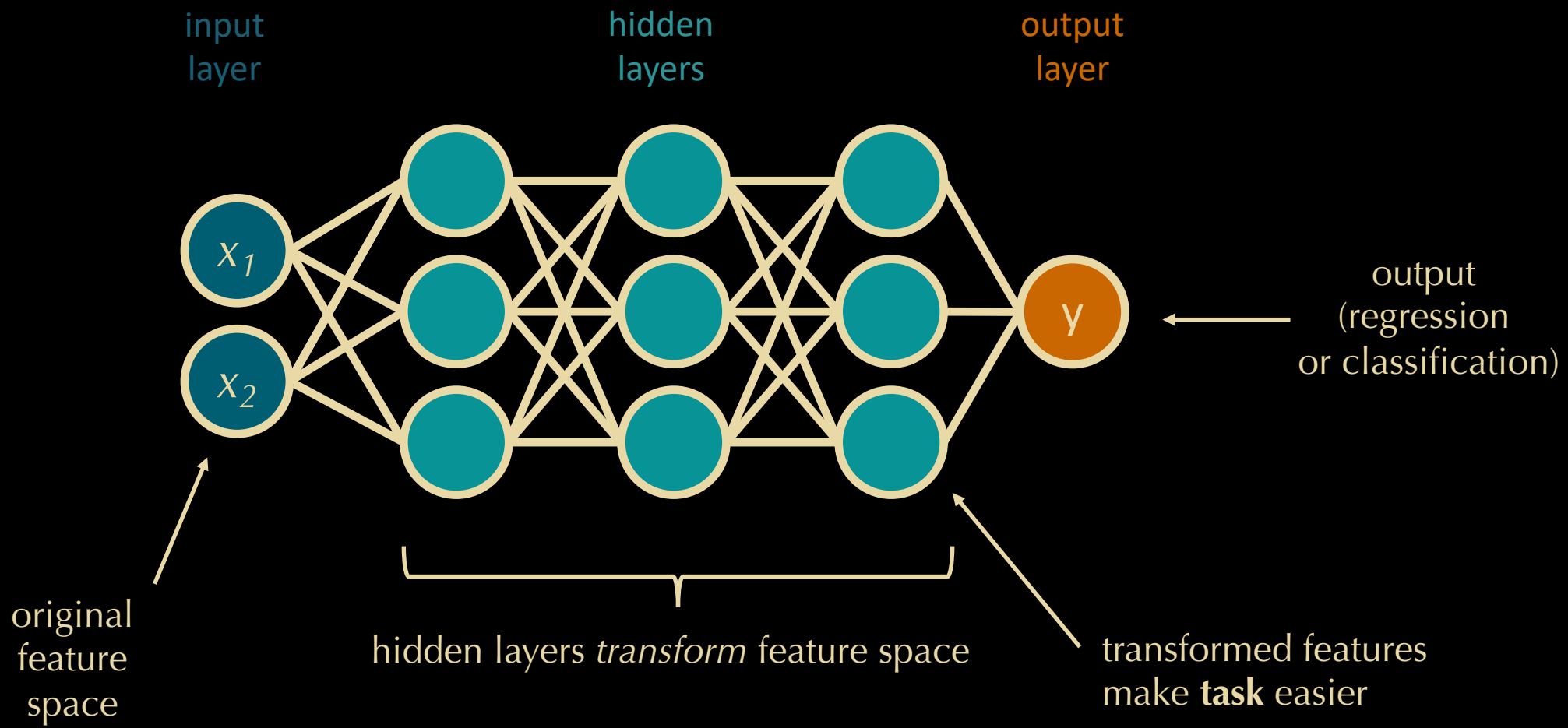
$$i_3 = ReLU(\beta_0 + \beta_1 h_1 + \beta_2 h_2 + \beta_3 h_3)$$

activation function

linear function

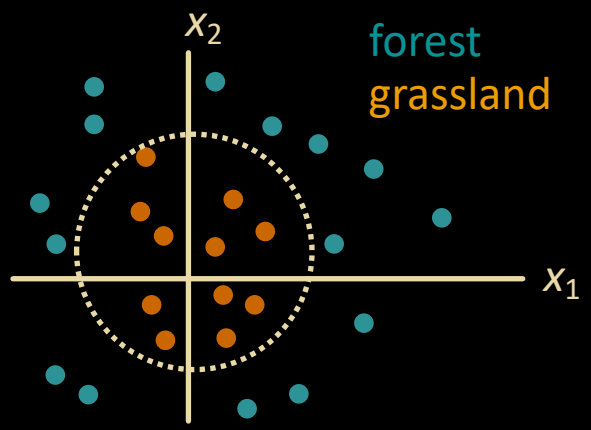


Neural networks “transform” feature space



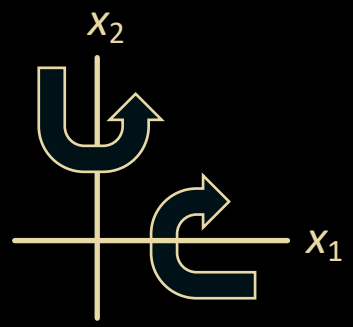
Transforming feature space

original feature “space”

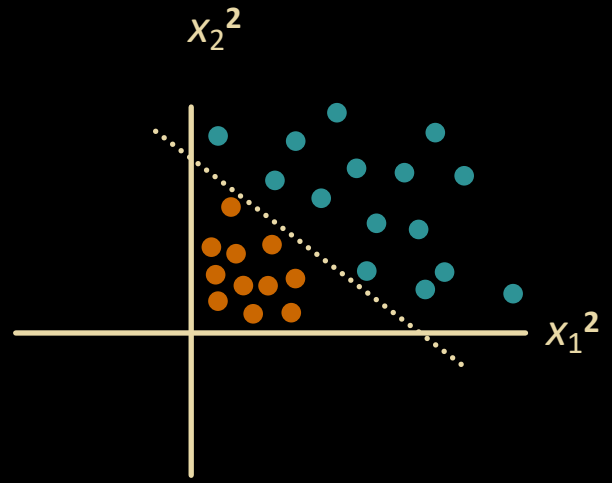


square feature values

fold feature space



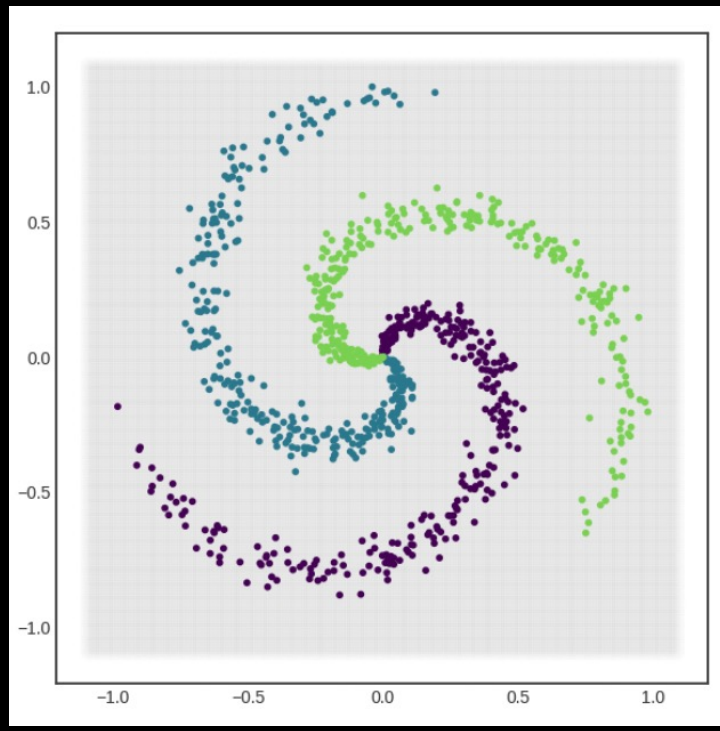
transformed feature “space”



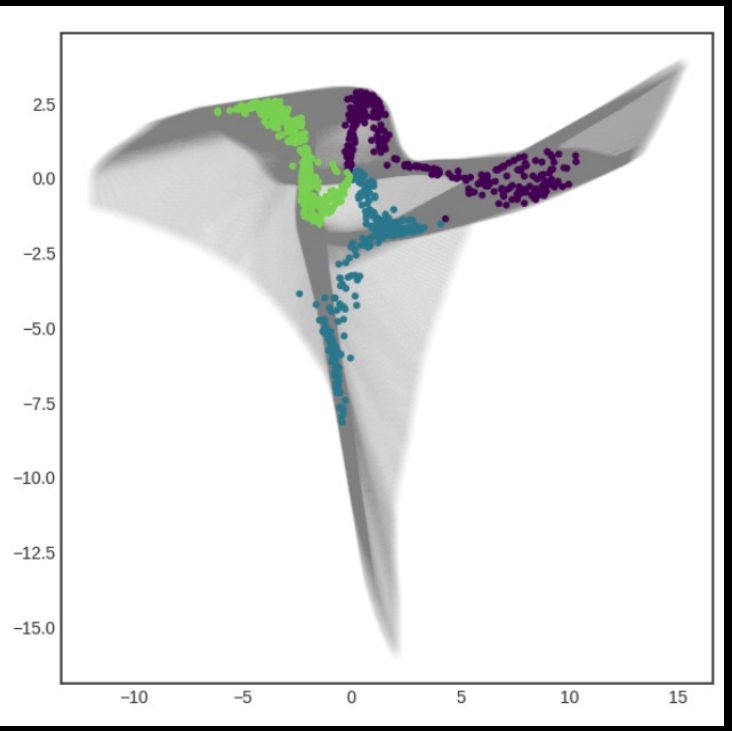
learning a linear decision boundary in transformed space
learns a circular one in the original feature space

NNs transform feature space

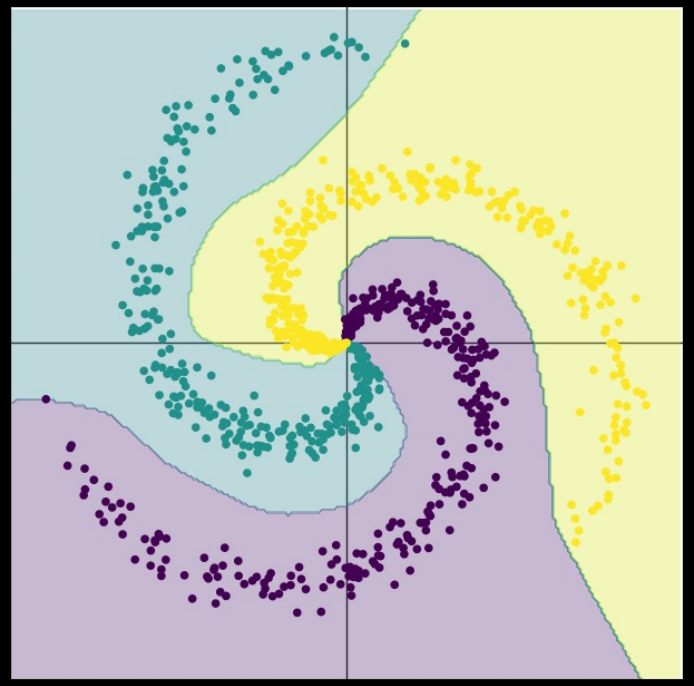
original feature “space”



transformed feature “space”

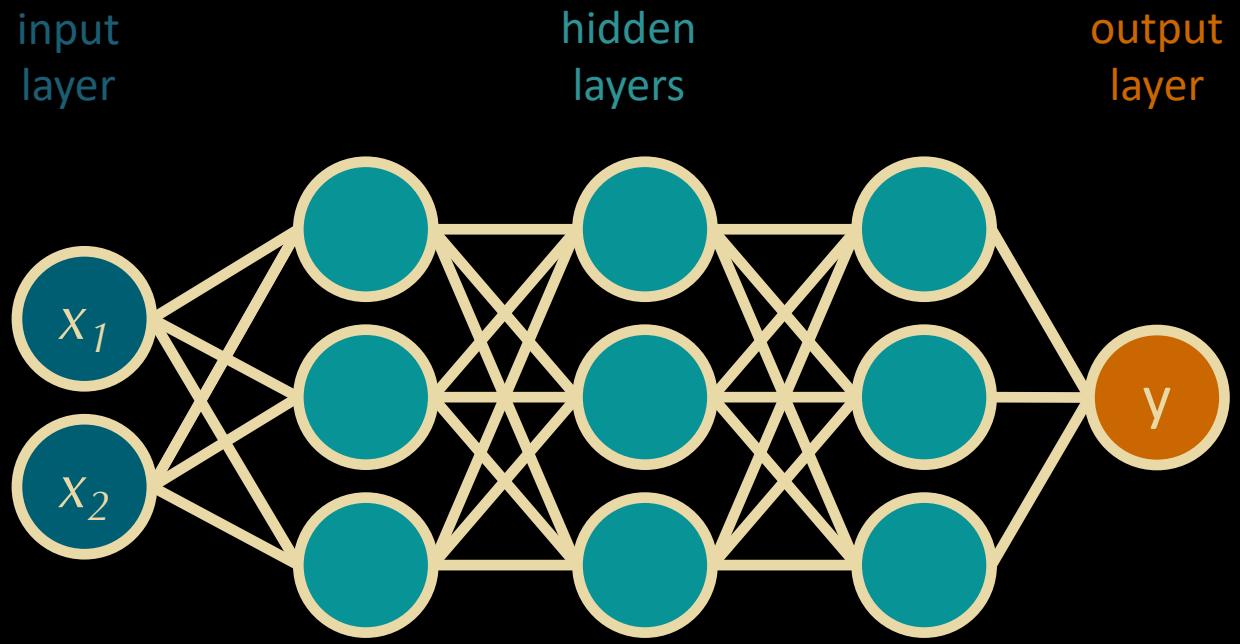


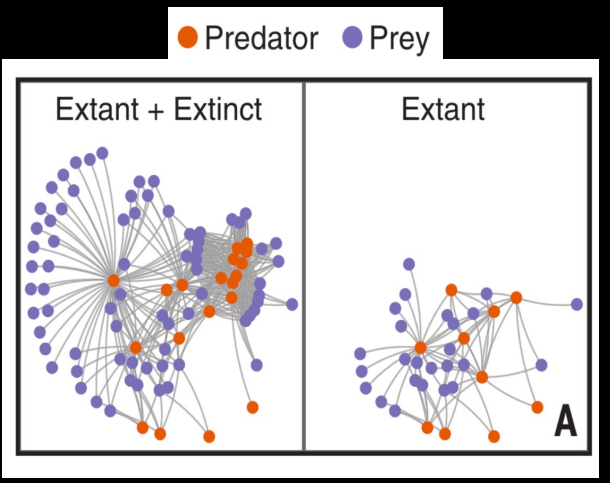
converting back to original space



note: deeply understanding this isn’t required to use NNs

Predator-prey prediction





- input morphological traits of species 1 and species 2, each as separate feature in input layer
- predict whether species 1 preys on species 2
- 90% accurate!

Two things that make NNs extremely powerful

1. **activation functions**
   - makes linear equations have non-linear behavior
2. **multiple hidden layers**
   - 1st layer learns from input, 2nd layer learns from the 1st layer
   - deeper layers synthesize knowledge in complex ways

NNs seem complicated, why would I use them?

- if interpretability extremely important, perhaps don't!
  - approaches do exist to identify what they learn
- if data complex and *prediction accuracy* extremely important, try them!
- NNs automatically perform 'feature selection'

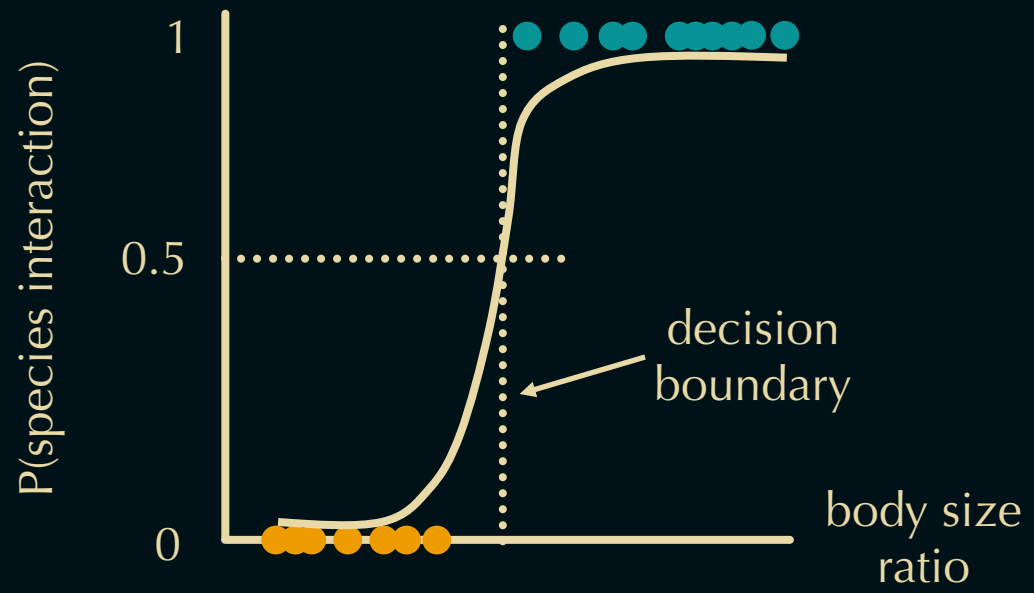
Logistic regression

- uses features to predict binary response
  - features used in simple linear formula
  - response is odds, logged
    - log used to make sure response b/t 0 and 1
- **rules to learn**: coefficients (β's) so that the log odds best reflect data
  - β's describe how each feature contributes to prediction

$$\log\left(\frac{p(interaction)}{p(no\ interaction)}\right) = \beta_0 + \beta_1 x_1$$

decision:

- interaction if $p(interaction) > p(no\ interaction)$
  - $\frac{p(interaction)}{p(no\ interaction)} > 1$
  - $\log\left(\frac{p(interaction)}{p(no\ interaction)}\right) > 0$



$$p(interaction) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}}$$

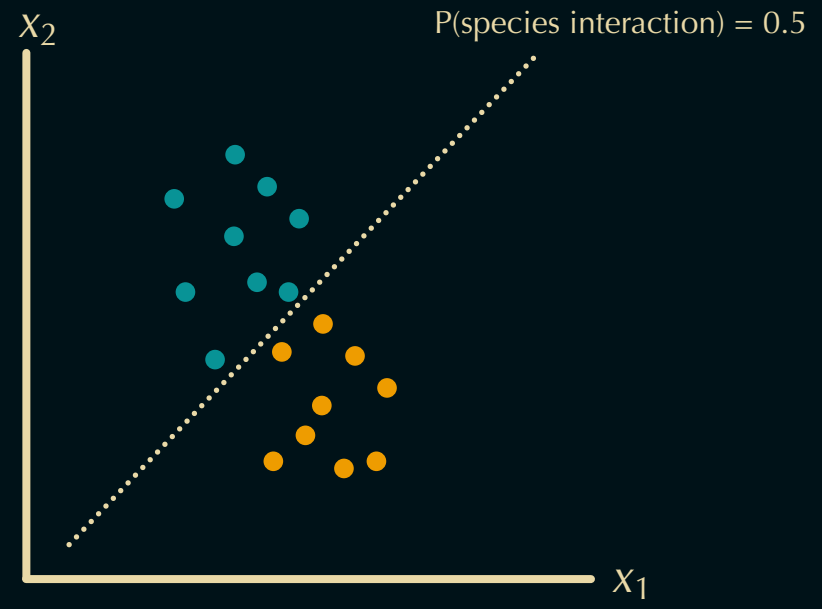
Logistic regression

- uses features to predict binary response
  - features used in simple linear formula
  - response is odds, logged
    - log used to make sure response b/t 0 and 1
- **rules to learn**: coefficients (β's) so that the log odds best reflect data
  - β's describe how each feature contributes to prediction

$$\log\left(\frac{p(interaction)}{p(no\ interaction)}\right) = \beta_0 + \beta_1 x_1$$

decision:

- interaction if $p(interaction) > p(no\ interaction)$
  - $\frac{p(interaction)}{p(no\ interaction)} > 1$
  - $\log\left(\frac{p(interaction)}{p(no\ interaction)}\right) > 0$



$$p(interaction) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}}$$

- we will use multiple features next time to classify penguin species!

Summary

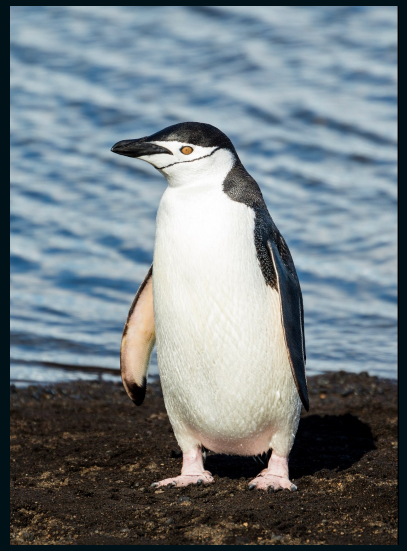
ML foundational principles

- 3 tasks
  - classification
  - regression
  - clustering
- focus on *prediction*
  - sometimes at the expense of interpretability
  - simpler, interpretable models available
- models built via *training* and *testing*

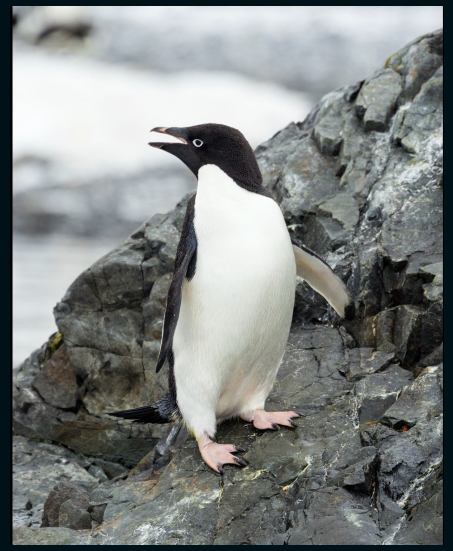
next session:

- classify penguin species using logistic regression and random forests

Chinstrap



Adelie



Gentoo

Color palette