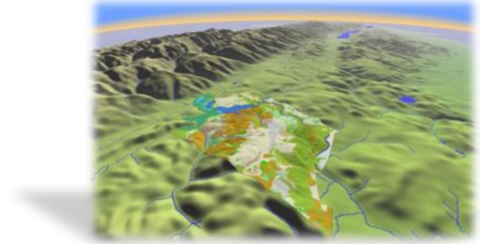
NICHOLAS SCHOOL OF THE ENVIRONMENT AND EARTH SCIENCES
DUKE UNIVERSITY



ENVIRON 761: Habitat Classification & Species Distribution Modeling

Instructor: John Fay
(Adapted from Dean Urban)

Habitat classification and modeling

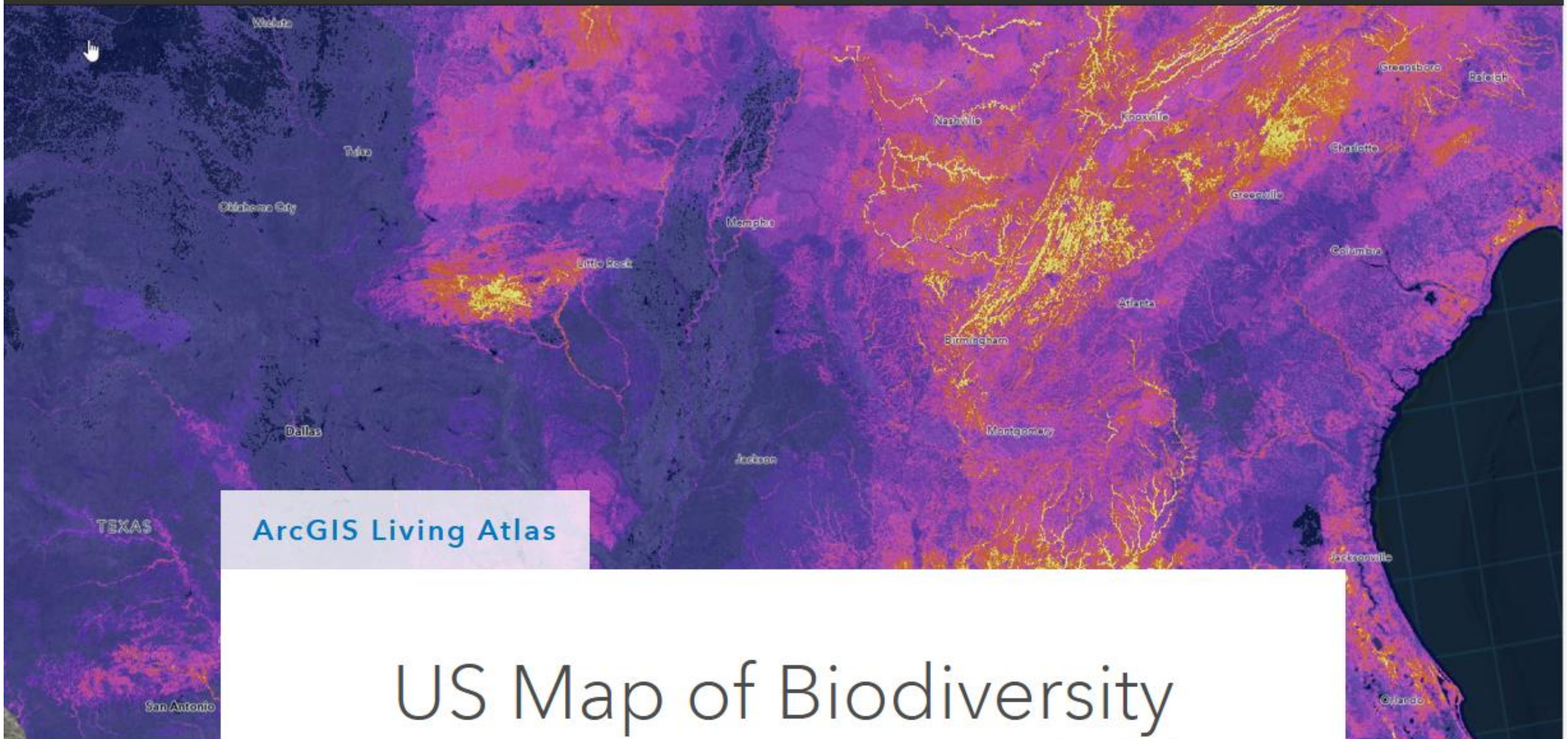
Habitat models underpin most of natural resource management

- Wildlife management
- Conservation planning
- Assessing future scenarios (climate!)

Habitat classification and modeling

ArcGIS Blog

Overview Topics



ArcGIS Living Atlas

US Map of Biodiversity Importance now available in ArcGIS Living Atlas

Habitat classification

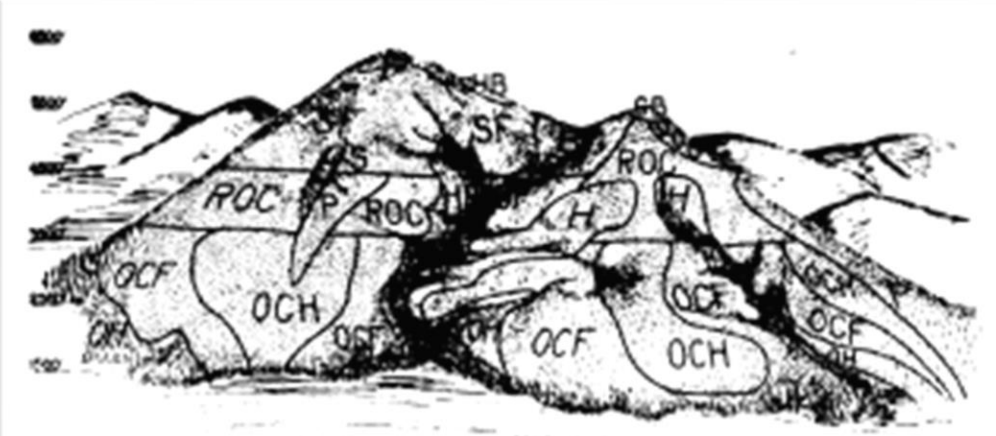


FIG. 21. Topographie disposition of vegetation types. View of idealized mountain and valley, looking east, with 6500-ft peak bearing subalpine forest on left, lower 5500-ft peak covered up to summit bald with deciduous forest on right. Vegetation types:

BG—Beech Gap
CF—Cove Forest
F—Fraser Fir Forest
GB—Grassy Bald
H—Hemlock Forest
HB—Heath Bald
OCF—Chestnut Oak-Chestnut Forest
OCH—Chestnut Oak-Chestnut Heath
OH—Oak-Hickory Forest
P—Pine Forest and Pine Heath
ROC—Red Oak-Chestnut Forest
S—Spruce Forest
SF—Spruce-Fir Forest
WOC—White Oak-Chestnut Forest

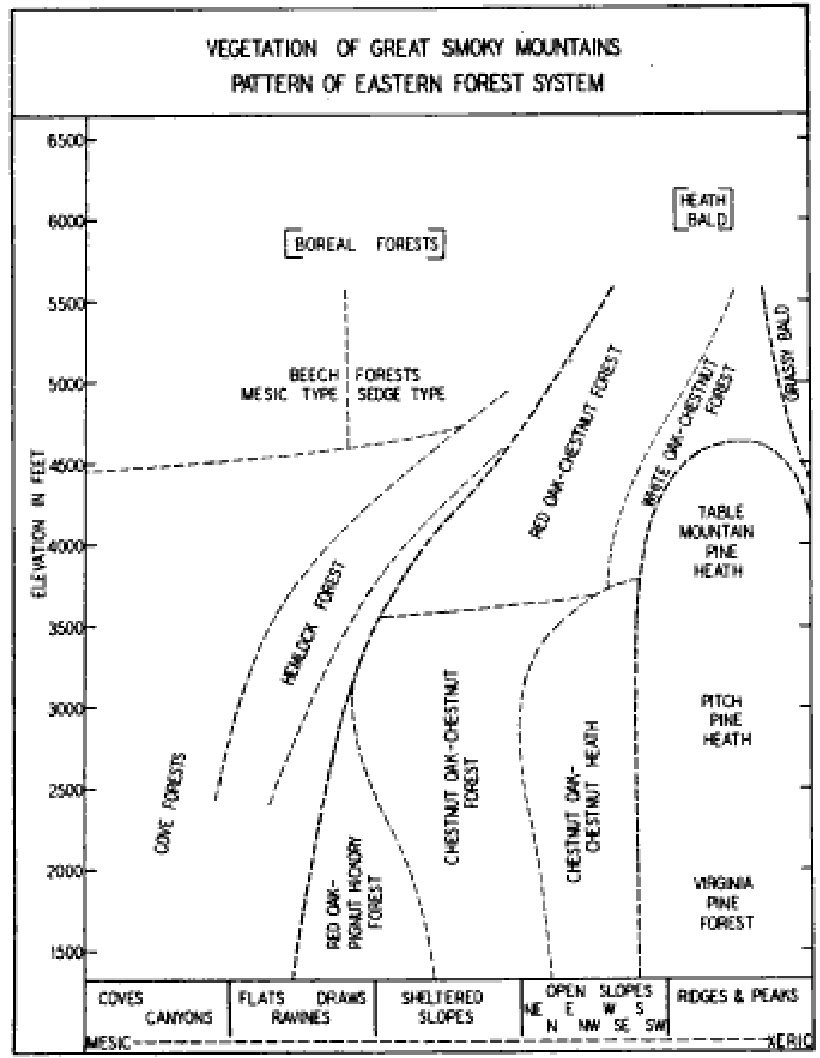
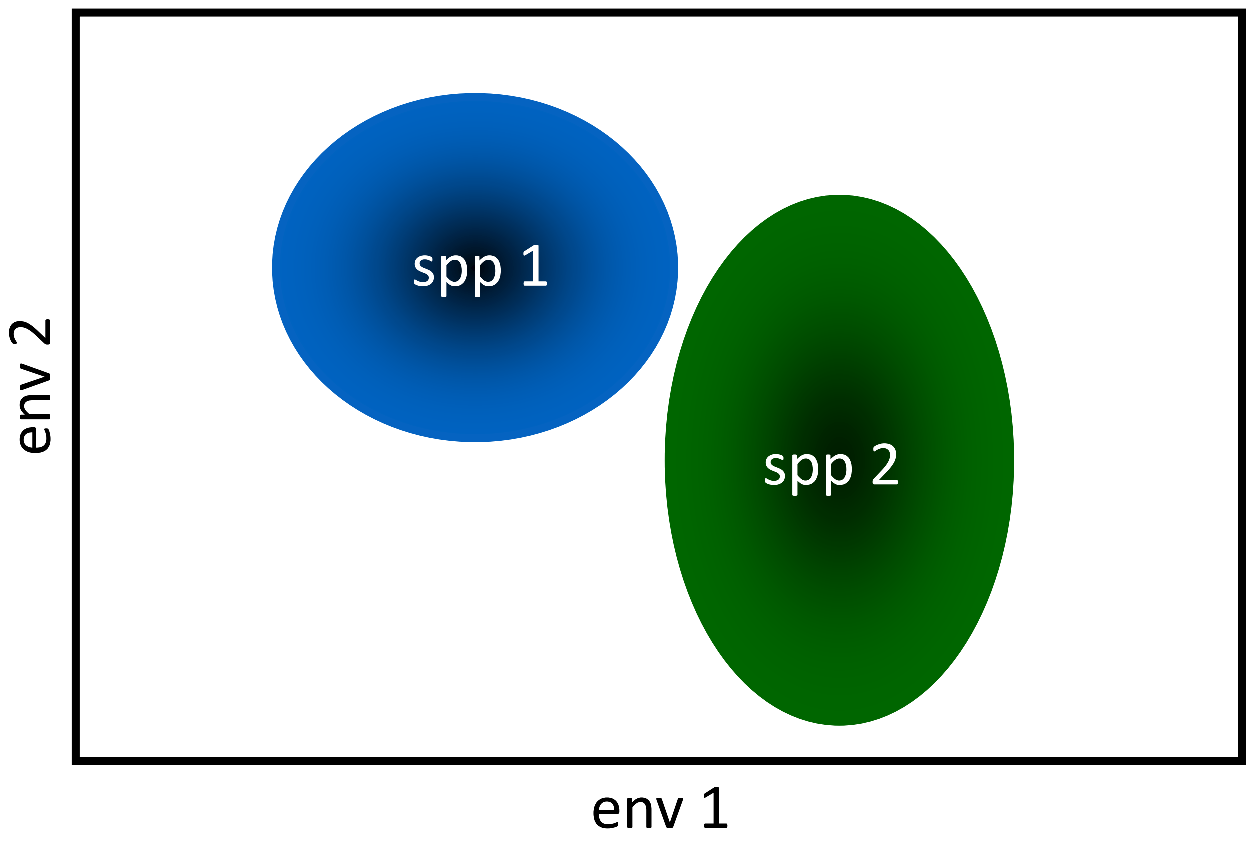


FIG. 19. (Vegetation of Great Smoky Mountains, pattern of Eastern Forest System.)

the Hutchinsonian niche

Issues:
- max overlap?
- packing?
- relevant axes?



Three interconnected models

Austin (2002, 2007):

- **Ecological model**
  - What we expect, and why
- **Data model** ⟵ **GIS**
  - What we measure, and why
- **Statistical model**
  - How we “fit” ecology to data

https://doi.org/10.1016/S0304-3800(02)00205-3
https://doi.org/10.1016/j.ecolmodel.2006.07.005

Ecological models: scaling

- **Fine scale**: *community ecology*
  - Ecology is about niche theory

- **Landscape scale**:
  - Ecology is about area, edge, isolation, ...

- **Larger scales**: *biogeography*
  - Ecology is about evolutionary history, ...

Data models: variables

- **Field studies**:
  - Choose variables based on ecology

- **Landscapes**:
  - Geospatial data in a GIS, especially **biophysical proxies** (select variables based on conceptual model)

  - Beware spatial resolution!

Data models: coding

Cols = variables
(Species = 0/1)

Rows = samples

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Species	Sample	X	Y	Elevation	Slope	Sun	SlpTAsp	TCI	AWC	pH	SOM	Depth
2	1	5	277140	3883680	969	16.72356	149	-7.33113	59	19.7	5.2	4.5	152.4
3	1	9	277260	3883680	972	22.11129	196	-20.9066	59	17.6	5.5	3	152.4
4	1	6	277200	3883590	932	18.45642	167	-12.5873	75	19.7	5.2	4.5	152.4
5	1	10	277260	3883560	929	20.828	194	-20.0212	65	19.7	5.2	4.5	152.4
6	1	7	277230	3883500	907	14.30291	168	-11.9954	81	19.7	5.2	4.5	152.4
7	1	3	277260	3883500	910	17.43931	182	-16.053	70	19.7	5.2	4.5	152.4
8	1	20	276240	3883470	1019	24.02984	180	-17.2856	57	19.7	5.2	4.5	152.4
9	1	1	276660	3883470	997	15.36952	153	-7.91589	74	17.6	5.5	3	152.4
10	1	2	276870	3883470	998	18.66144	113	1.95065	52	17.6	5.5	3	152.4
11	1	11	277440	3883470	944	21.19362	195	-20.2676	57	19.7	5.2	4.5	152.4
12	1	18	276240	3883440	1006	22.54033	170	-14.1851	61	19.7	5.2	4.5	152.4
13	1	25	276630	3883440	994	14.53876	124	-0.5074	73	18	5.2	10.5	165.1
14	1	32	277020	3883440	938	22.38417	90	7.287575	59	19.7	5.2	4.5	152.4
15	1	16	276060	3883410	1030	12.95884	126	-0.67821	61	19.7	5.2	4.5	152.4
16	1	28	276690	3883410	984	9.462322	151	-6.80663	87	19.7	5.2	4.5	152.4
17	1	33	277170	3883410	889	17.9598	152	-8.43162	75	17.6	5.5	3	152.4
18	1	31	277260	3883410	884	15.98069	171	-12.9287	81	19.7	5.2	4.5	152.4
19	1	39	277500	3883410	952	17.69631	188	-17.629	51	17.6	5.5	3	152.4
20	1	17	276150	3883380	1006	16.41644	104	4.524986	65	18	5.2	10.5	165.1
21	1	23	276390	3883380	992	20.09799	193	-19.501	60	19.7	5.2	4.5	152.4
22	1	22	276600	3883380	992	10.72416	121	0.934672	63	17.6	5.5	3	152.4
23	1	24	276390	3883350	983	18.41883	188	-17.9468	61	19.7	5.2	4.5	152.4
24	0	34	277050	3883350	904	25.65023	91	6.205355	61	19.7	5.2	4.5	152.4
25	0	37	277260	3883350	869	12.3342	170	-11.9678	93	19.7	5.2	4.5	152.4
26	0	15	275970	3883290	1027	16.03195	93	7.526546	72	18	5.2	10.5	165.1
27	0	30	276780	3883290	970	8.735691	154	-7.40829	70	19.7	5.2	4.5	152.4
28	0	35	277200	3883290	857	10.90502	141	-4.43547	95	17.6	5.5	3	152.4
29	0	41	277560	3883290	936	21.56378	199	-21.5506	57	19.7	5.2	4.5	152.4
30	0	43	277620	3883290	953	20.96671	198	-20.9635	51	19.7	5.2	4.5	152.4

sample

Ready NUM

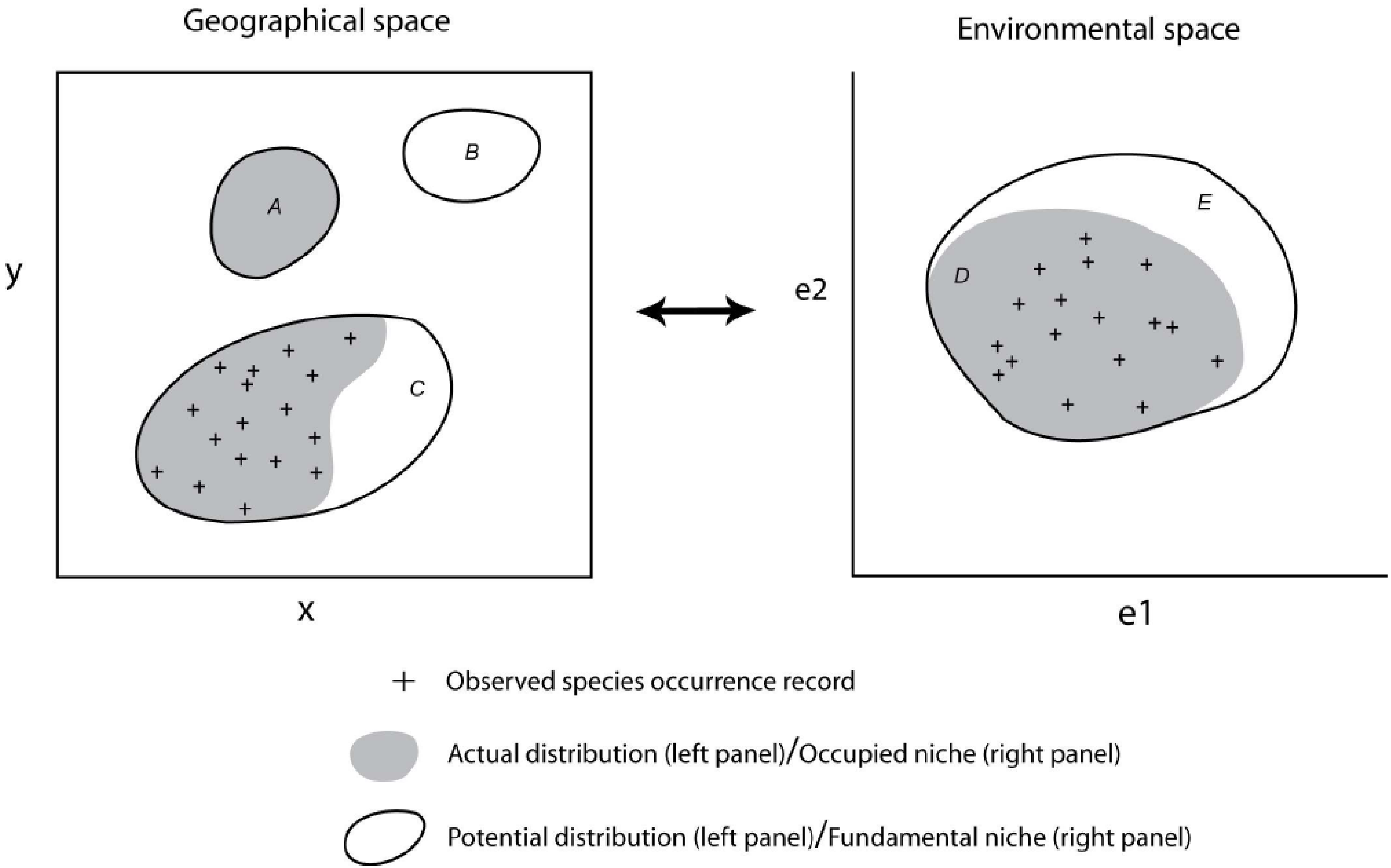
Data spaces and translations

- Field data, map data are in *geographic space*

- Statistics translate these into *parameter space*

- Often, we will want to back-translate the statistics into a map (the locations are what's interesting)

Data spaces and translations



Pearson 2008

Statistical models: preamble

Caveats:

- Once the data are coded, the statistics are blind to ecology
- The onus in on the investigator to put the ecology back on completion, for interpretation

Statistical models

Data models: coding

Cols = variables
(Species = 0/1)

Rows = samples

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Species	Sample	X	Y	Elevation	Slope	Sun	SlpTAsp	TCI	AWC	pH	SOM	Depth
2	1	5	277140	3883680	969	16.72356	149	-7.33113	59	19.7	5.2	4.5	152.4
3	1	9	277260	3883680	972	22.11129	196	-20.9066	59	17.6	5.5	3	152.4
4	1	6	277200	3883590	932	18.45642	167	-12.5873	75	19.7	5.2	4.5	152.4
5	1	10	277260	3883560	929	20.828	194	-20.0212	65	19.7	5.2	4.5	152.4
6	1	7	277230	3883500	907	14.30291	168	-11.9954	81	19.7	5.2	4.5	152.4
7	1	3	277260	3883500	910	17.43931	182	-16.053	70	19.7	5.2	4.5	152.4
8	1	20	276240	3883470	1019	24.02984	180	-17.2856	57	19.7	5.2	4.5	152.4
9	1	1	276660	3883470	997	15.36952	153	-7.91589	74	17.6	5.5	3	152.4
10	1	2	276870	3883470	998	18.66144	113	1.95065	52	17.6	5.5	3	152.4
11	1	11	277440	3883470	944	21.19362	195	-20.2676	57	19.7	5.2	4.5	152.4
12	1	18	276240	3883440	1006	22.54033	170	-14.1851	61	19.7	5.2	4.5	152.4
13	1	25	276630	3883440	994	14.53876	124	-0.5074	73	18	5.2	10.5	165.1
14	1	32	277020	3883440	938	22.38417	90	7.287575	59	19.7	5.2	4.5	152.4
15	1	16	276060	3883410	1030	12.95884	126	-0.67821	61	19.7	5.2	4.5	152.4
16	1	28	276690	3883410	984	9.462322	151	-6.80663	87	19.7	5.2	4.5	152.4
17	1	33	277170	3883410	889	17.9598	152	-8.43162	75	17.6	5.5	3	152.4
18	1	31	277260	3883410	884	15.98069	171	-12.9287	81	19.7	5.2	4.5	152.4
19	1	39	277500	3883410	952	17.69631	188	-17.629	51	17.6	5.5	3	152.4
20	1	17	276150	3883380	1006	16.41644	104	4.524986	65	18	5.2	10.5	165.1
21	1	23	276390	3883380	992	20.09799	193	-19.501	60	19.7	5.2	4.5	152.4
22	1	22	276600	3883380	992	10.72416	121	0.934672	63	17.6	5.5	3	152.4
23	1	24	276390	3883350	983	18.41883	188	-17.9468	61	19.7	5.2	4.5	152.4
24	0	34	277050	3883350	904	25.65023	91	6.205355	61	19.7	5.2	4.5	152.4
25	0	37	277260	3883350	869	12.3342	170	-11.9678	93	19.7	5.2	4.5	152.4
26	0	15	275970	3883290	1027	16.03195	93	7.526546	72	18	5.2	10.5	165.1
27	0	30	276780	3883290	970	8.735691	154	-7.40829	70	19.7	5.2	4.5	152.4
28	0	35	277200	3883290	857	10.90502	141	-4.43547	95	17.6	5.5	3	152.4
29	0	41	277560	3883290	936	21.56378	199	-21.5506	57	19.7	5.2	4.5	152.4
30	0	43	277620	3883290	953	20.96671	198	-20.9635	51	19.7	5.2	4.5	152.4

sample

Ready NUM

Data models: observations

Kinds of locational observations:

1. Where you saw species X ("*habitat*")
2. Where you looked but didn't see it ("*nonhabitat*")
3. Where it *might* have occurred ("*available habitat*")

→ All statistical models proceed from some combination of these data

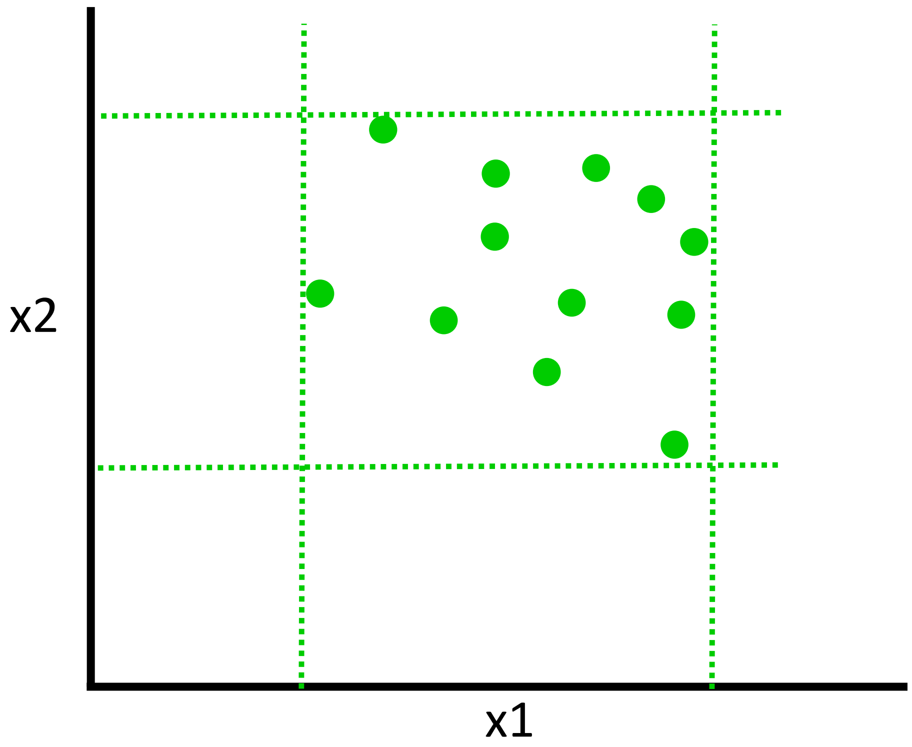
Statistical models: logic

- *"Habitat"* cf *"nonhabitat"* – are these 2 samples different on the predictors?

- *"Habitat"* cf *"available habitat"* – is this sample different from a random draw of what *might* have been observed?

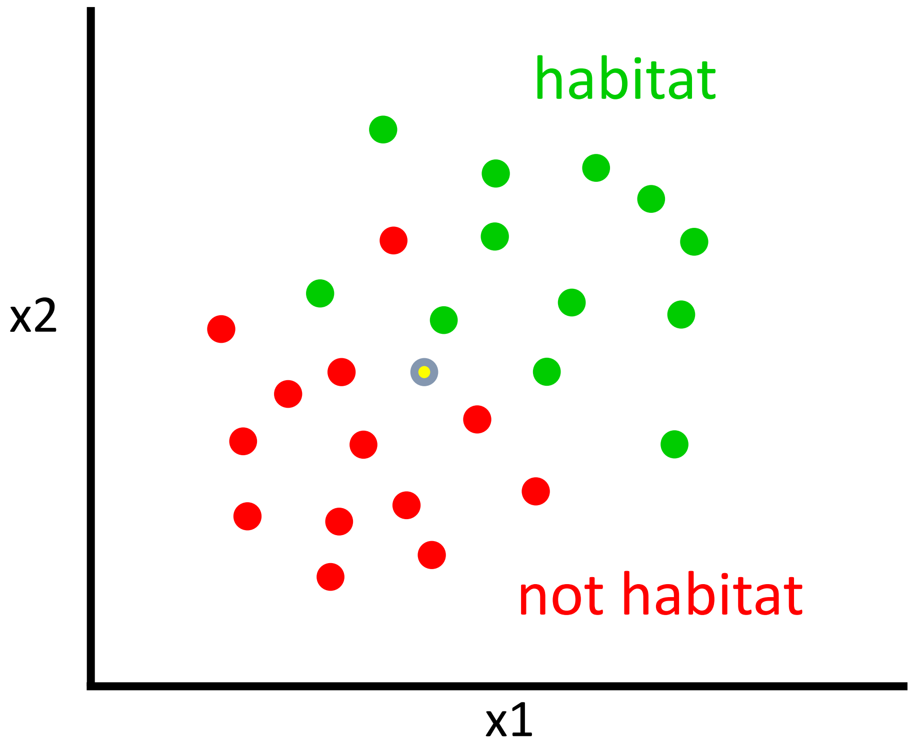
- 1-sample *"habitat"* – show me all the places that *look* like "habitat"

Generative models: “envelopes”



- Define limits in terms of lower and upper bounds (or some arbitrary confidence ellipse)
- Simple and easy!

Discriminative models: logic



- Q: what function of X1 and X2 best separates the 2 groups?
- A: provided by several alternative statistical methods

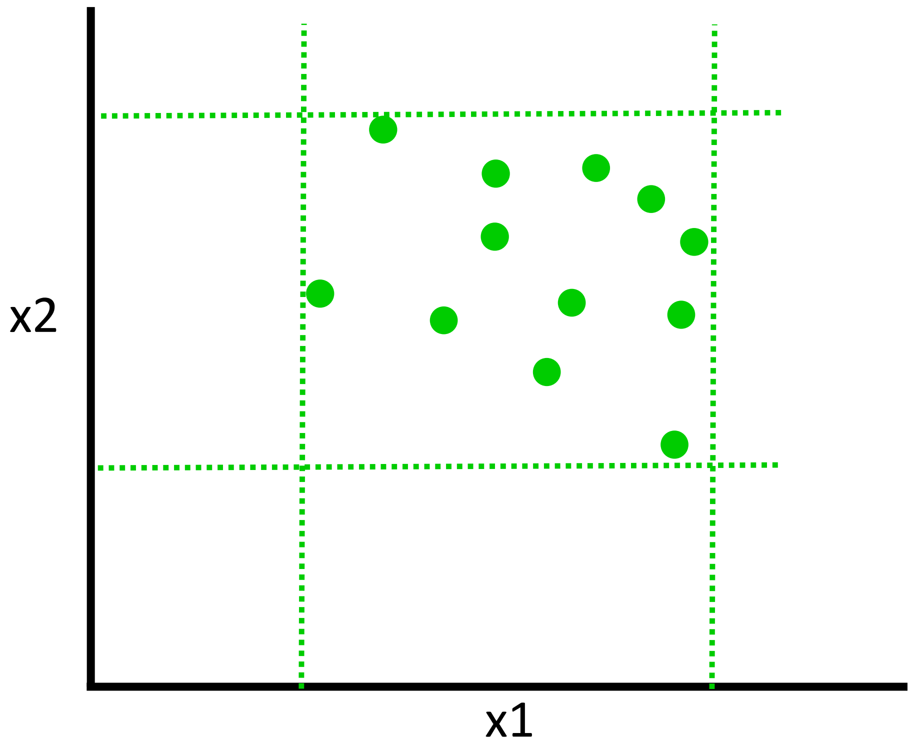
Models: tour guide

- There are multiple approaches to this task—each represented by a few techniques

- For each:
  - What does it do?
  - Advantages and disadvantages
  - Current status (popularity)

- Relationships among techniques

Statistical models: (1) “envelopes”

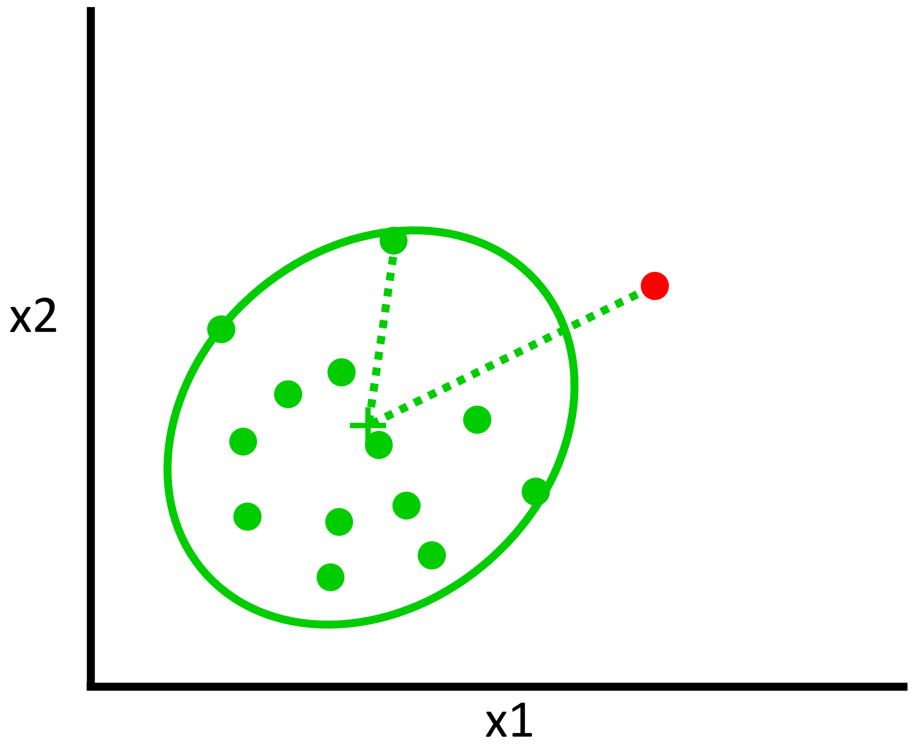


- Define limits in terms of lower and upper bounds (or some arbitrary confidence ellipse)

Envelopes: summary

- *Advantages:*
  - Simple (especially in GIS)
  - Can use any data (or none!)
- *Disadvantages:*
  - Poor leverage statistically (presence only)
- *Status:*
  - Common and popular
  - Fancy extensions (GARP, DOMAIN, …)

Envelopes: Mahalanobis D^2



- D^2 = (squared) distance from group centroid (accounting for any correlation among the *x*'s)

→ How much does this sample look like "habitat"?

Envelopes: Mahalanobis D2

- Advantages:
  - Requires only “habitat” data
  - Can be “tuned” to application
- Disadvantages:
  - Requires ratio-scale data
  - Hard to interpret variables
- Status:
  - Resurgence in mapping applications
  - The “classifier” in supervised methods

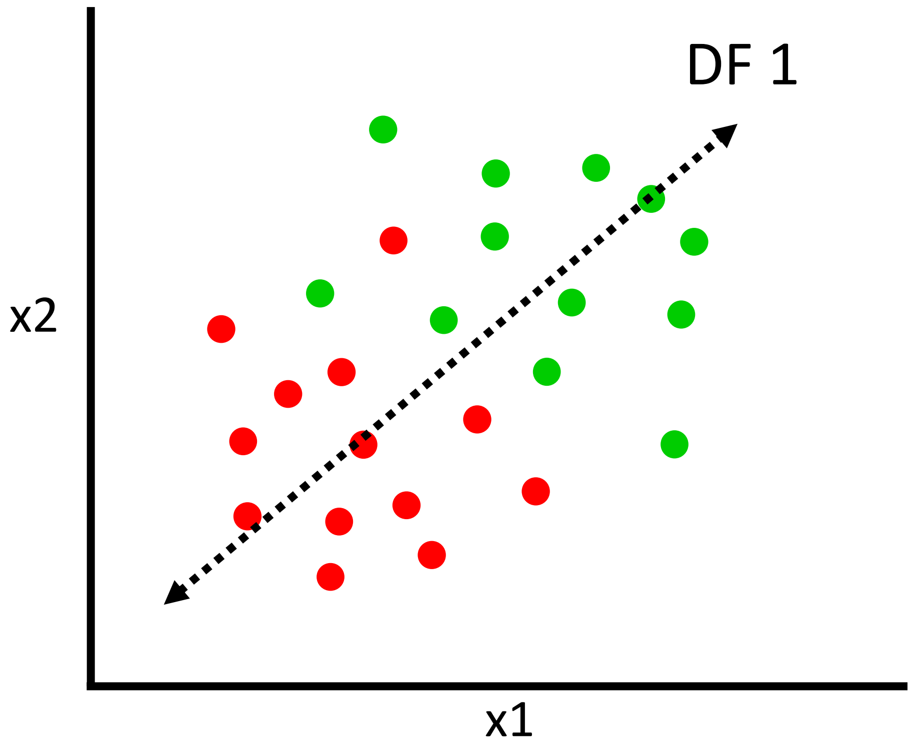
Statistical Models

1. Envelope Models
2. Discriminant Function Analysis (DFA)

Statistical models: (2) DFA

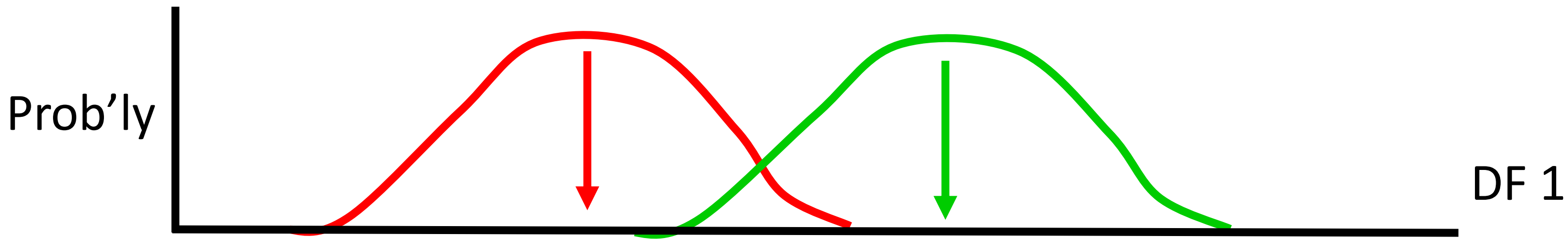
- Discriminant functions analysis
  - Finds the best linear function of the original (predictor) variables that separates the 2 groups
  - Maximizes among-group to within-group variability on this function

DFA: logic



- DF 1 maximally separates the groups
- Note (here) neither X1 nor X2 can separate the groups by itself

DFA: interpretation



- DFA tests separation of the group means
- Correlations between DFs and original variables provide for interpretation
- Classification is based on a (new) sample's proximity to each group mean

DFA: summary

- *Advantages:*
  - Does what we want!
- *Disadvantages:*
  - assumes multi-normality
  - the variables are ratio scale
  - the functions are linear
- *Status*:
  - new versions (robust, quadratic, flexible)

Statistical Models

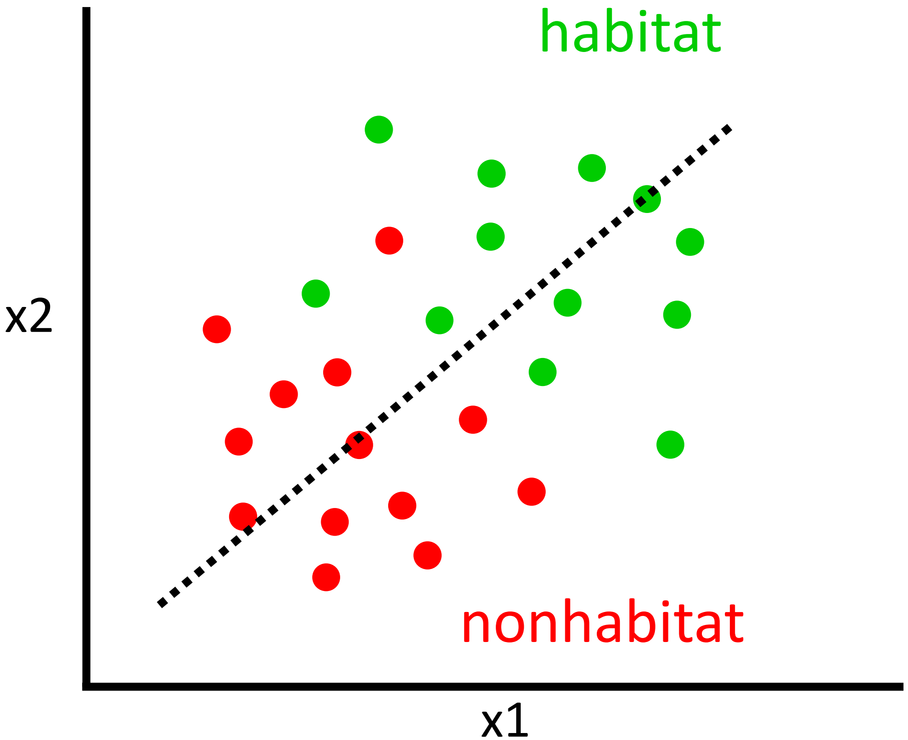
1. Envelope Models
2. Discriminant Function Analysis (DFA)
3. Generalized Linear Models (GLM)

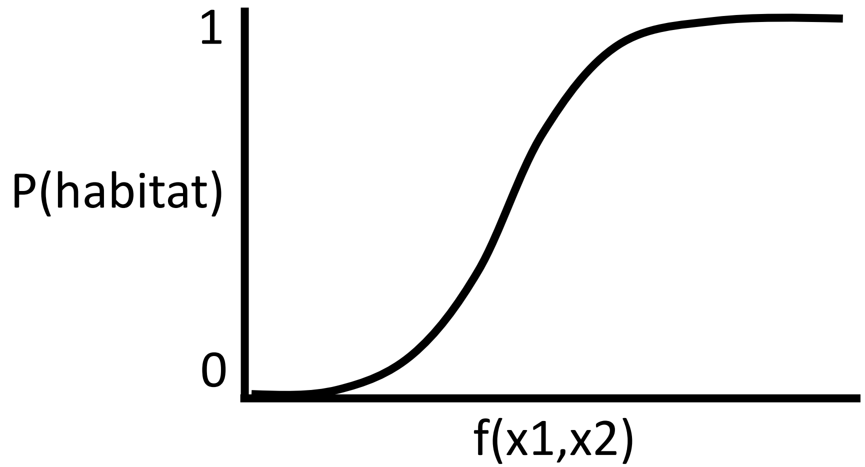
Statistical models: (3) GLMs

- Linear model:
  - $Y = b0 + b1x1 + b2x2 + \ldots + \text{error}$

- Generalized linear model:
  - $U = b0 + b1x1 + b2x2 + \ldots + \text{error}$
  - Y = link function of U
  - Link function maps the linear term to the distribution of the data

GLMs: Logistic regression





habitat = 1; not = 0

GLMs: Logistic regression

Logit model:

$$P(\text{habitat}) = e^u/(1+e^u)$$

where...

$$u = f(x1, x2, \ldots)$$

so...

$$\ln[P(\text{habitat})/P(\text{not})] = u$$

GLMs: summary

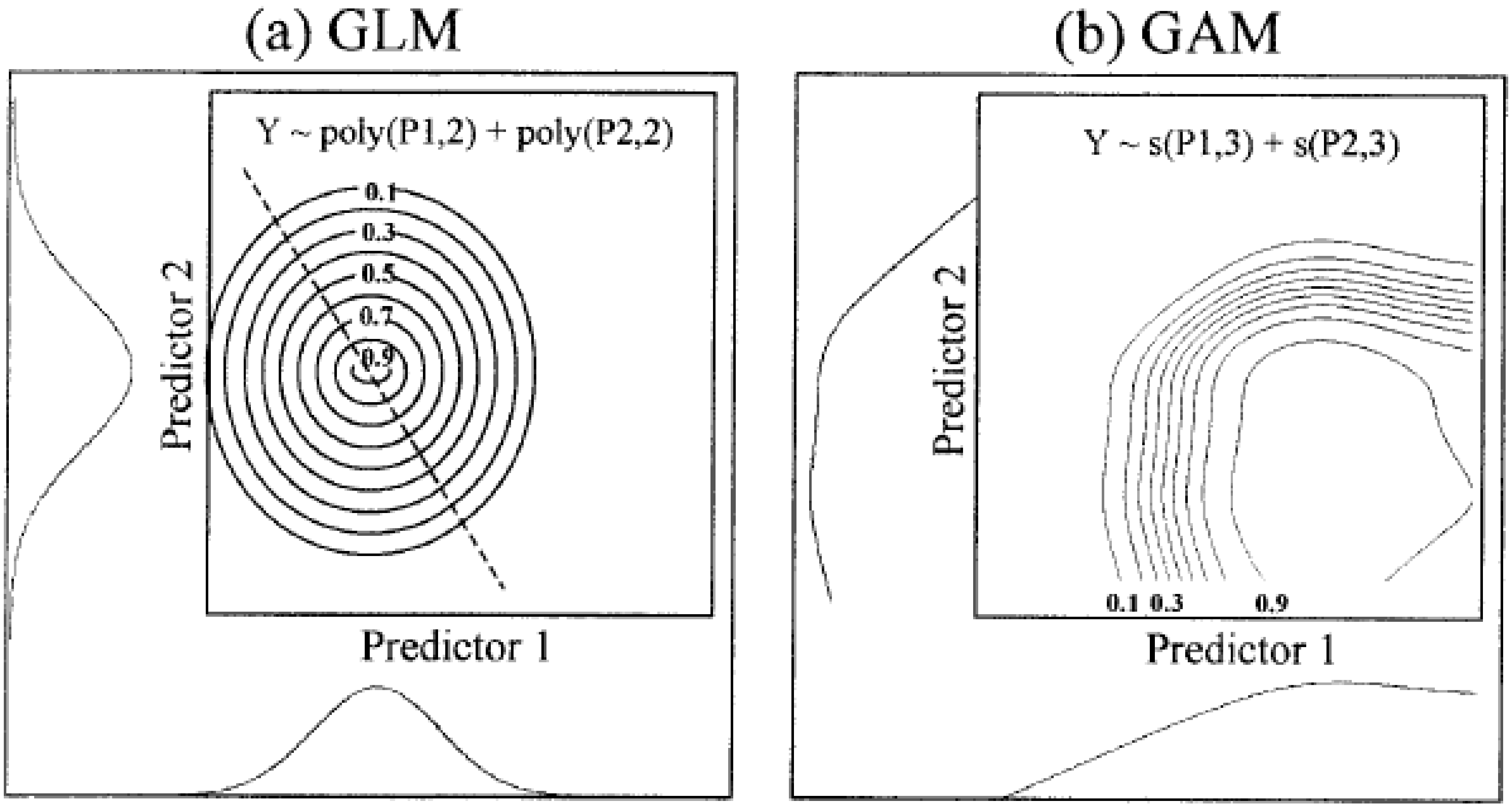
- Advantages:
  - Lots of distributions and link functions
  - Can use mixed data types
  - Can be “tuned” as a predictor
- Disadvantages:
  - (it’s still a regression)
- Status:
  - the workhorse model

GLMs: extensions

Extensions to the basic GLM …

- GAM:
  - b’s become smoothing functions
- GLMM, GEE (mixed models):
  - Spatial structure (distributions) OK
- MARS
  - Multivariate adaptive GLMs

GLM vs GAM



A. Guisan, N.E. Zimmermann / Ecological Modelling 135 (2000) 147–186

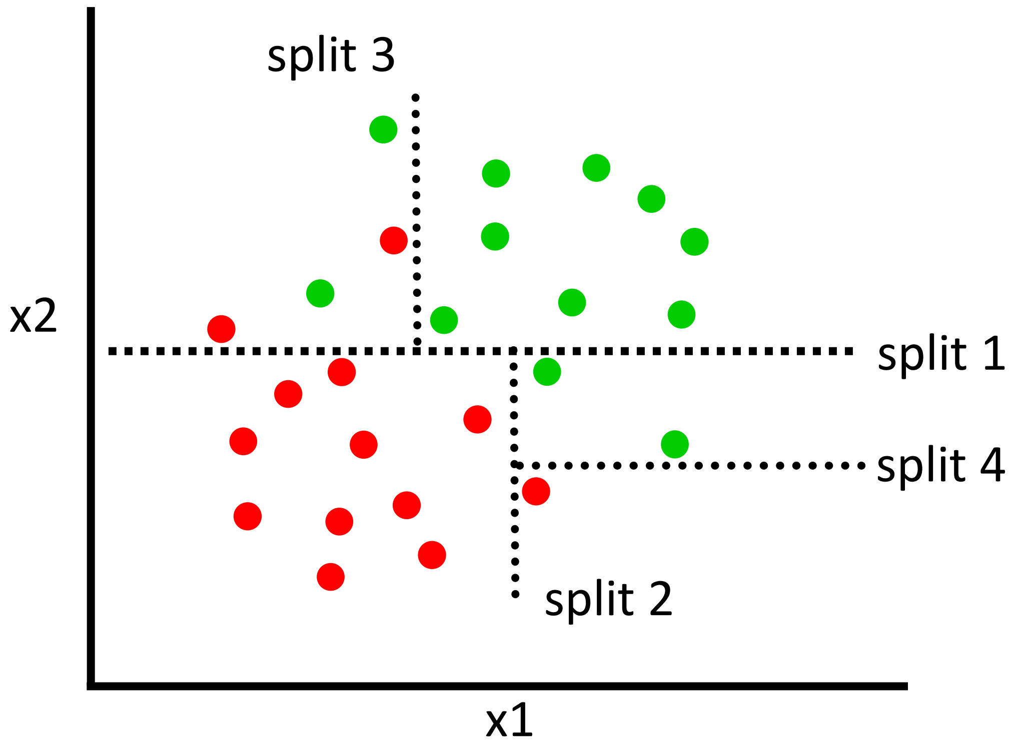
Statistical Models

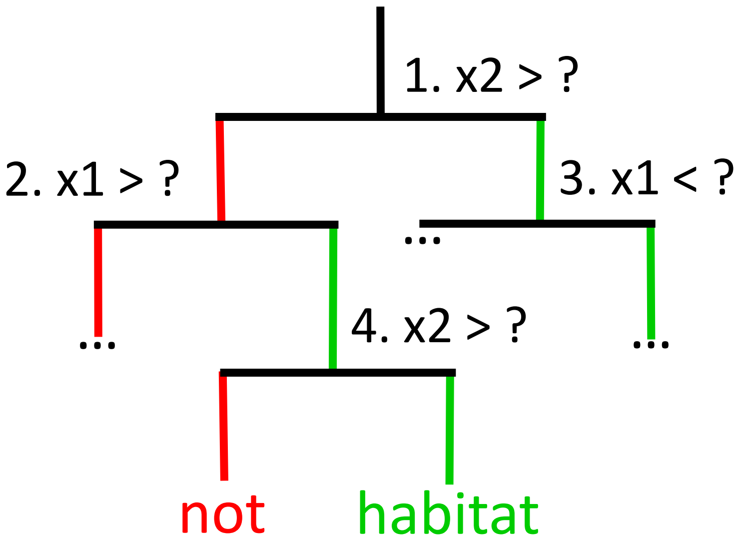
1. Envelope Models
2. Discriminant Function Analysis (DFA)
3. Generalized Linear Models (GLM)
4. Classification & Regression Trees (CART)

Statistical models: (4) CART

- Consider: “sugar pine is found at middle elevations on mesic slopes; also at lower elevations on NE slopes of in pockets of deep soil, or at higher elevations on SW slopes …”
- Need a model that can handle compensatory, substitutable settings: a classification (or regression) tree

CART: logic


split 3
x2
split 1
split 4
split 2
x1


1. x2 > ?
2. x1 > ?
3. x1 < ?
...
4. x2 > ?
...
...
not
habitat

CART: summary

- *Advantages:*
  - Can handle complex complementary or substitutable cases
  - Can use any data types
  - Provides intuitive decision tree

- Disadvantages:
  - Over-fitting (unstable)

- *Status:*
  - Extensions are popular

CART: extensions

Extensions to CART:

- "Bagged" trees
  - Resampled, then averaged

- "Boosted" trees
  - Resampled and re-weighted; averaged

- Random forests
  - Resampled observations & predictors; averaged (1000's of trees)

Statistical Models

1. Envelope Models
2. Discriminant Function Analysis (DFA)
3. Generalized Linear Models (GLM)
4. Classification & Regression Trees (CART)
5. Maximum Entropy (MaxEnt)

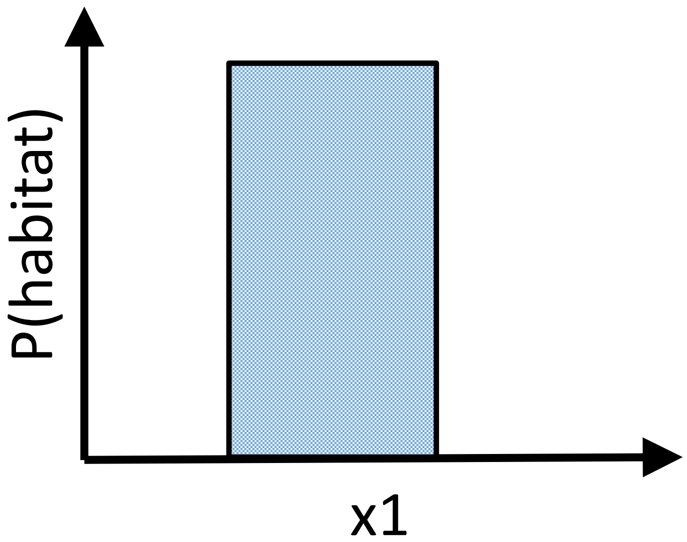
Statistics: (5) Maximum entropy

- Goal: find a distribution function that describes the data as closely as possible (an “envelope” model)

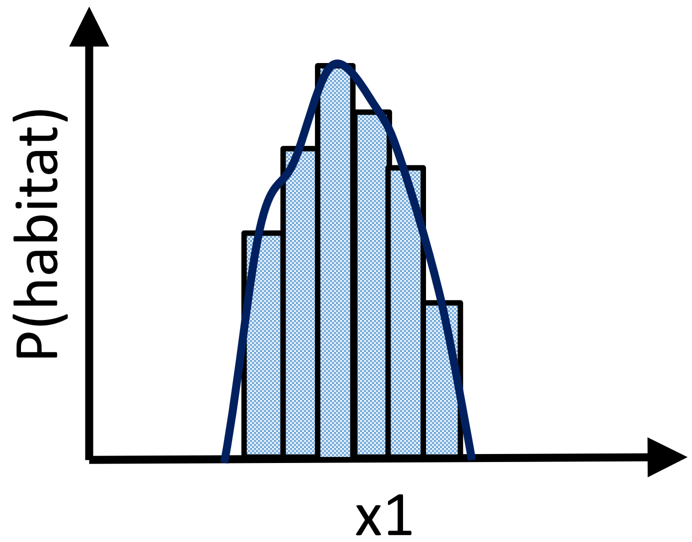
- Theory: the function that does this is the one with maximum entropy while also meeting specified constraints

Maximum entropy: logic

Envelope model:

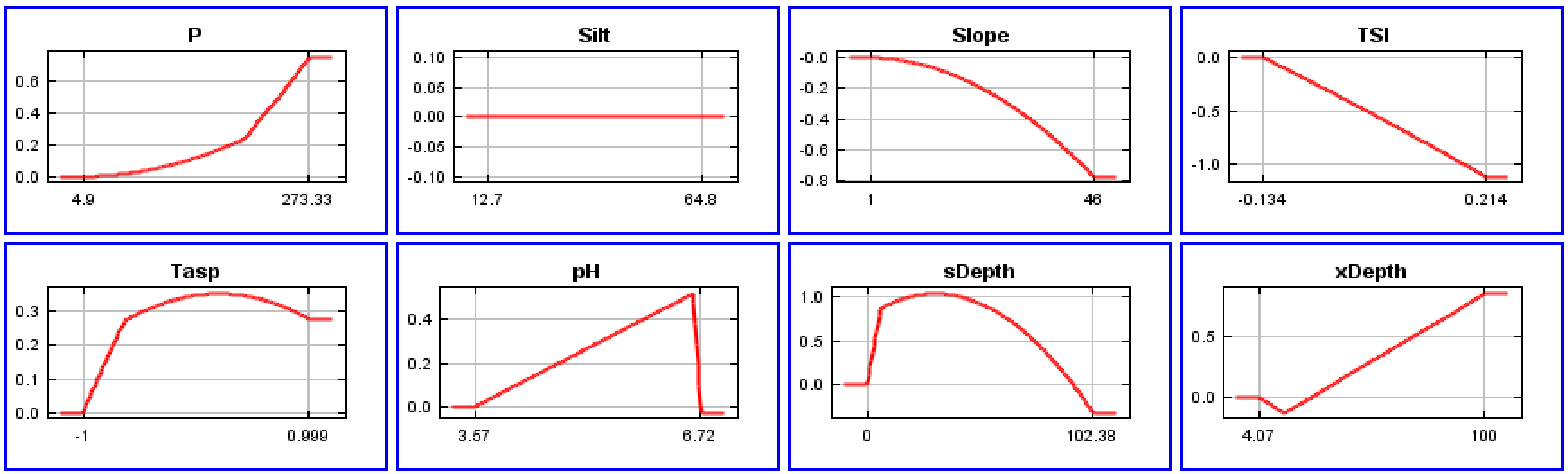


Maxent model:



Maxent: estimation (cf GAM)

- Examples of maxent features: piecewise “features” of the variables (categorical, linear, quadratic, threshold, hinges, interactions)



Maxent: estimation (maxent)

- **Estimation via maximum entropy:**
  - Fitted distribution (the “model”) should be consistent with the data but not assume anything beyond this
  - Fit is to minimize distributional difference between the presences and the background of what is available
  - Solution is machine-learning

Maxent: interpretation

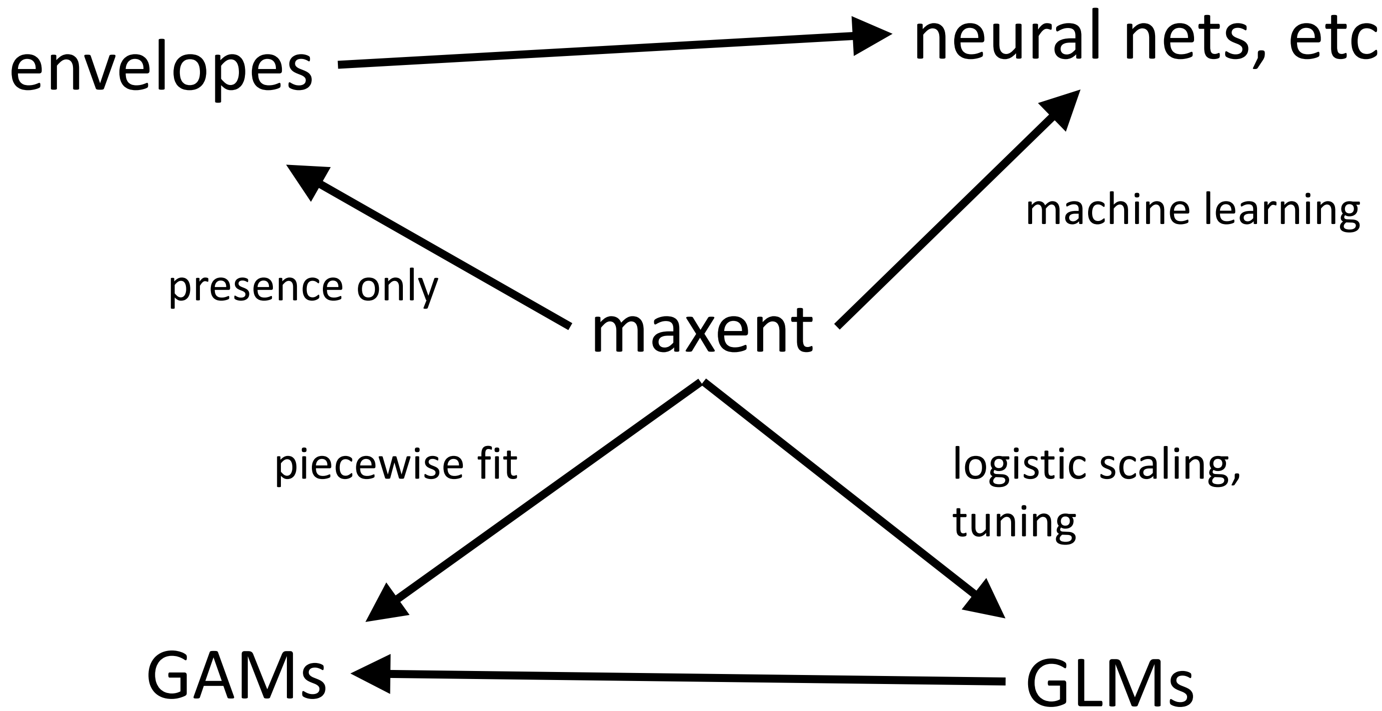
- The maxent software package:

- Presence-only model (not really)

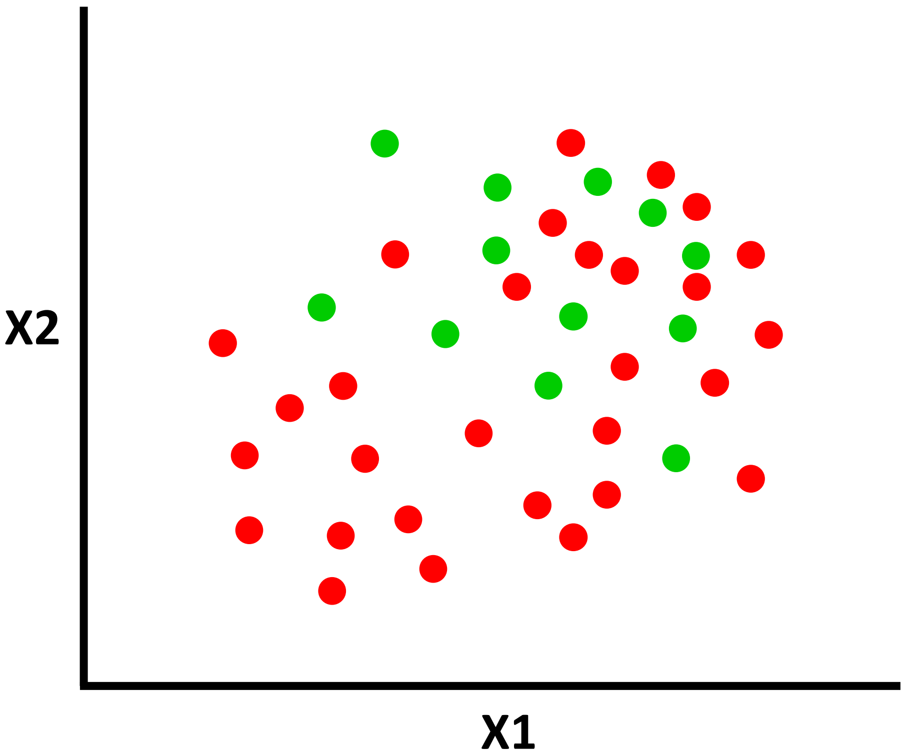
- A machine-learning solution

- A user-friendly interface (!)
  - Optimize true positives vs area
  - Tuning possible
  - Rescaled to look like a GLM
  - Lots of interpretative aids!

Models: connections



Statistics: Applications



- *Generative* models (envelopes, maxent) often perform better than *discriminative* models for rare species
- Models with flexible fits (CART, maxent) often perform better than global, linear models

Statistical models: reminders

- In ecological applications, models often perform very differently
  - Try a few models and compare/average

- The statistical tools are blind to ecology:
  - Implications of assumptions often must be accounted in model interpretation