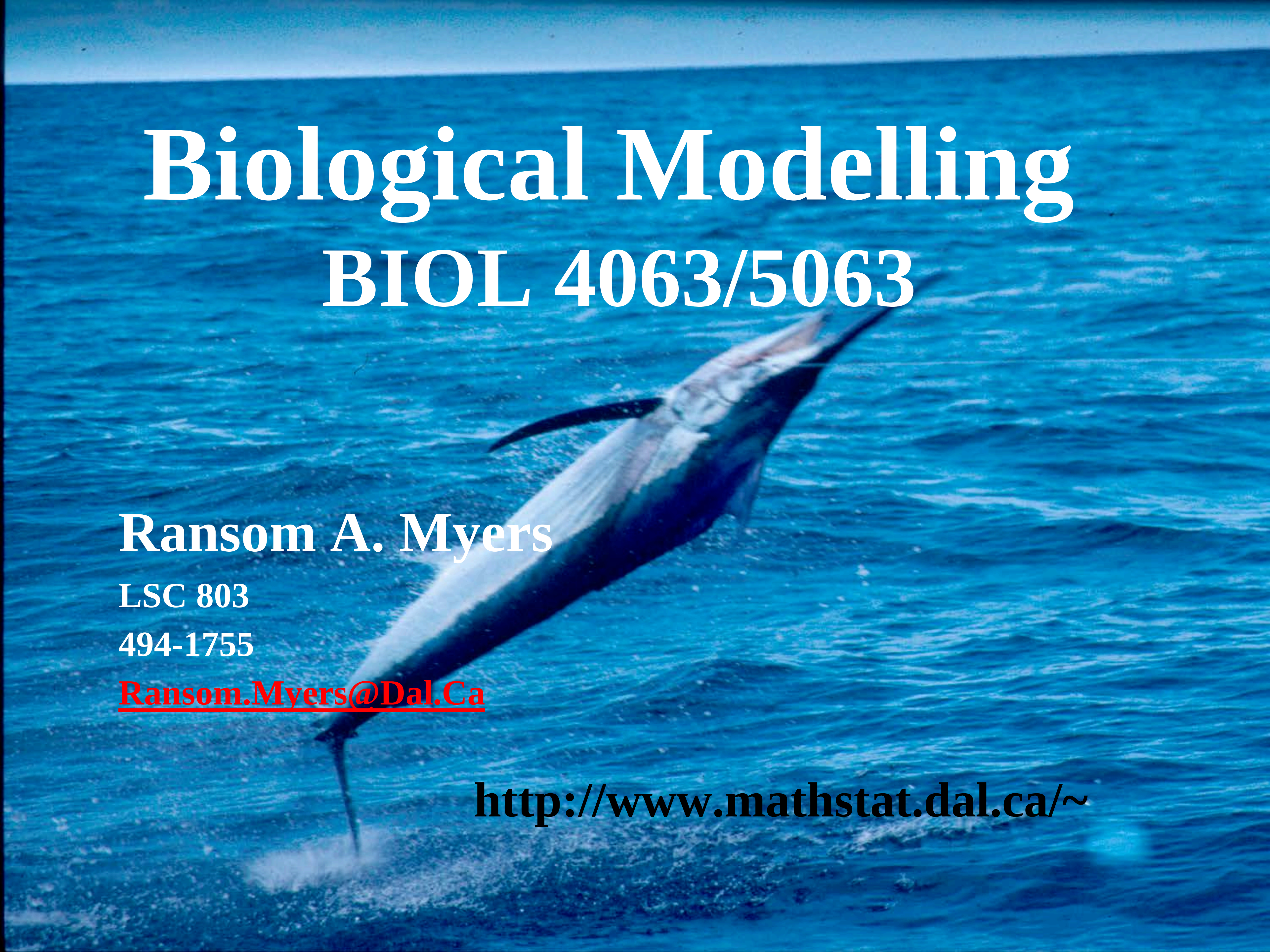


Biological Modelling

BIOL 4063/5063

Ransom A. Myers

LSC 803

494-1755

Ransom.Myers@Dal.Ca

http://www.mathstat.dal.ca/~

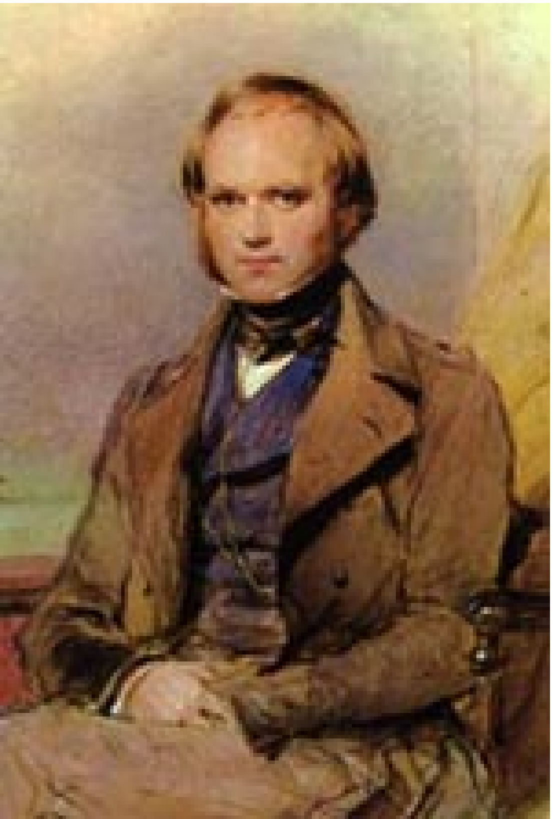
BIOL 4063/5063 – Biological Modelling

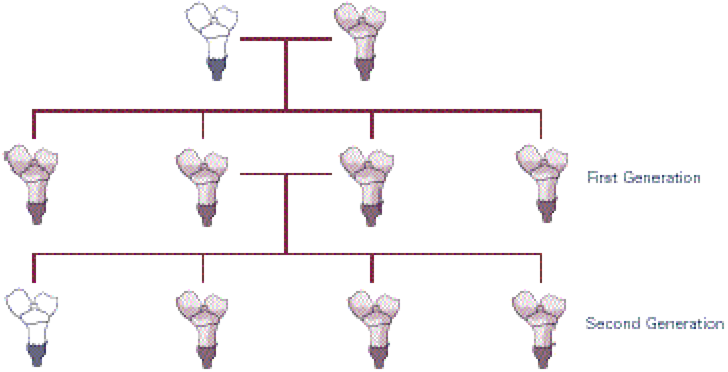
- Lectures Tues & Thurs, LSC 200
- No required text; readings will be available on reserve and lecture notes will sometimes be online.
- Grading : Projects, 60%

  midterm exam, 20%

  final exam, 20%

Modelling is as central to biology as it is to physics
Medelian genetics was believed to be incompatible with natural selection until dynamic models were used to explain the behaviour of genes, and statistical methods were developed to interpret the data.



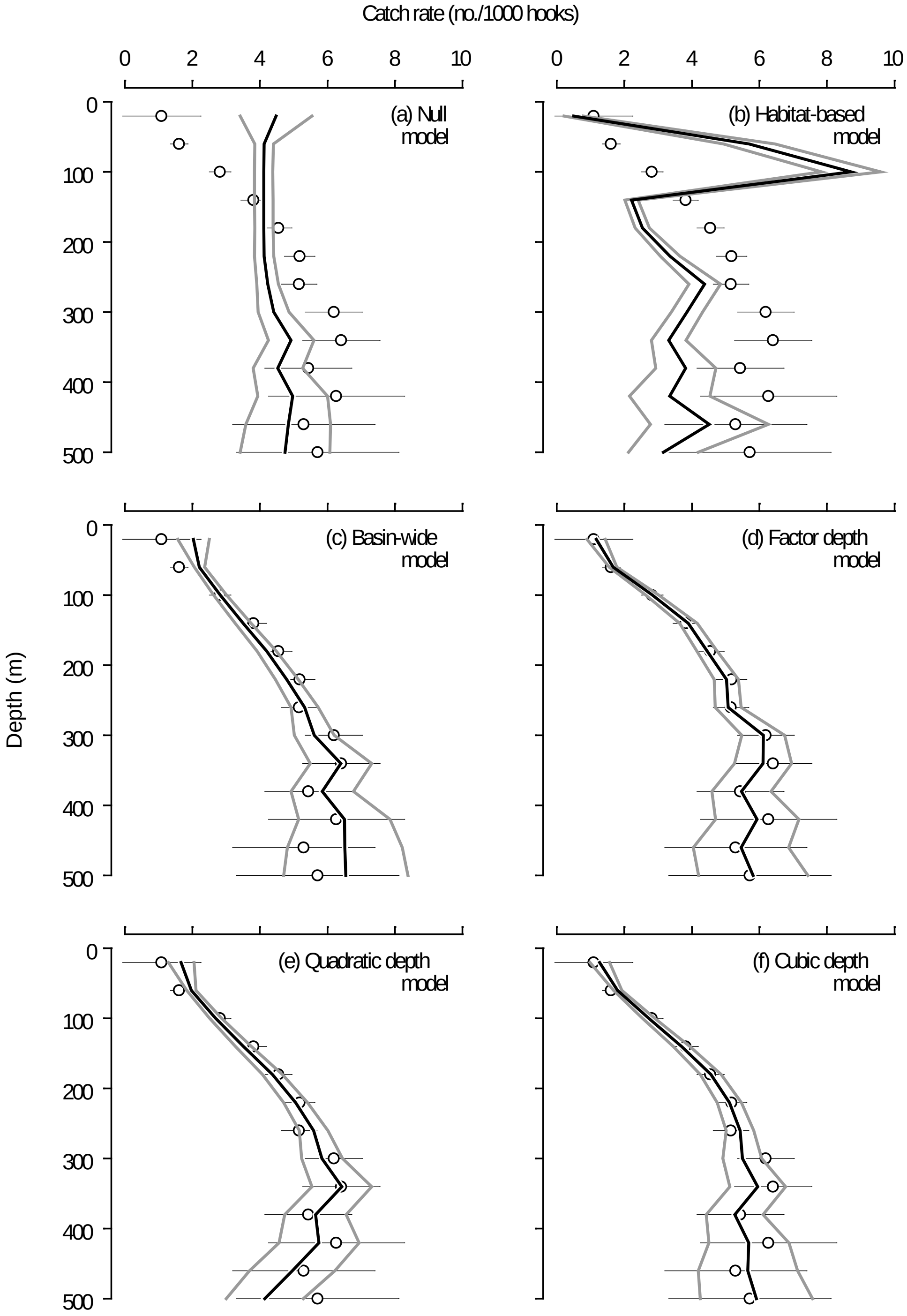


Gregor Mendel

Each student should come and talk with one of us.

Projects

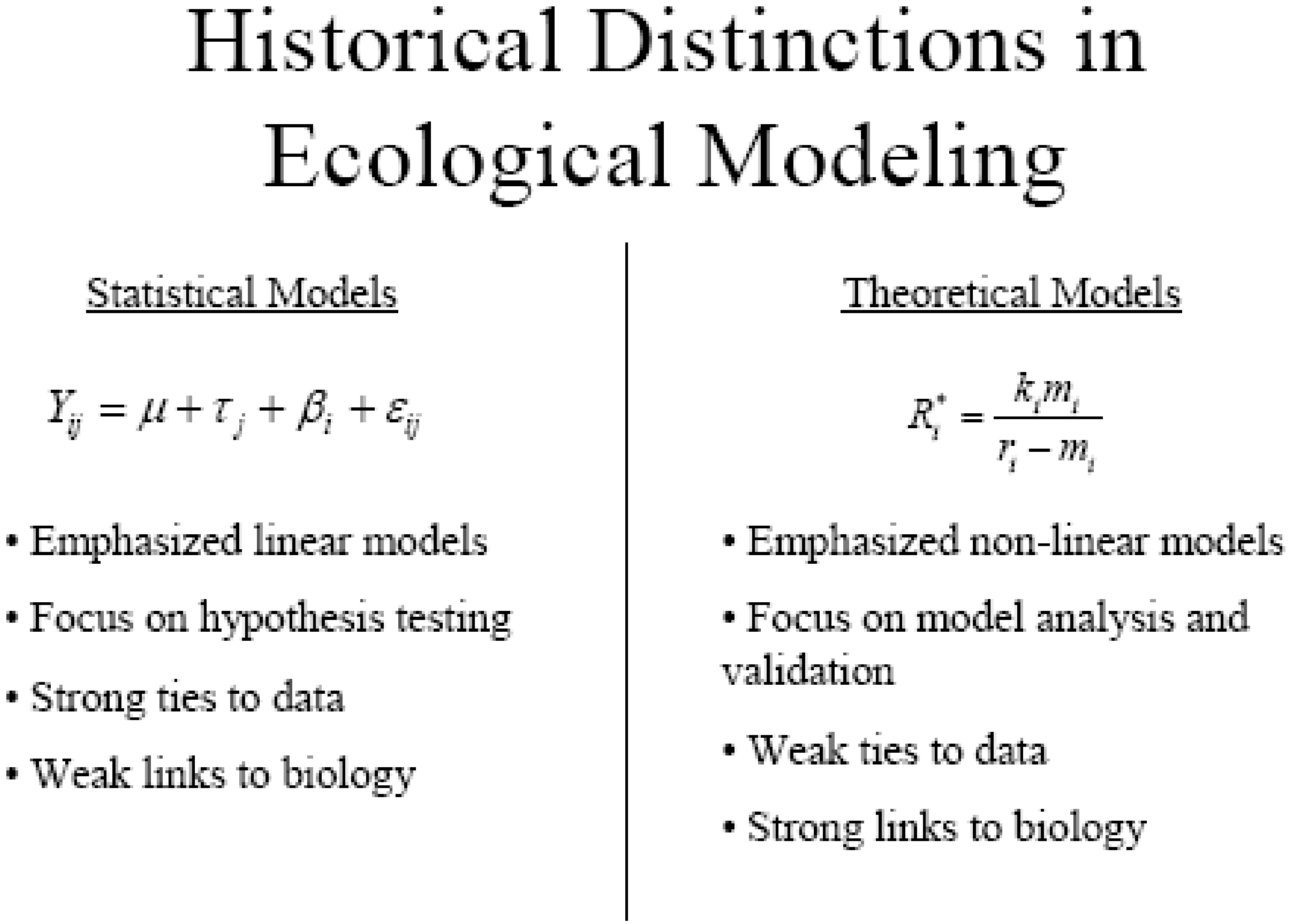
- Each student will develop a modelling project, write it up, and present it to the class.
- This may be used in your thesis.
- Graduate students should aim for publication (this has happened with several of my students).



Example of a class project that we are after.
This one is in press in a journal (Fisheries Oceanography).

Labs:

- We will use R (a free statistical language that is widely used).
- BUGS (also free, can be used to fit complex models).



Historical Distinctions in Ecological Modeling

Statistical Models

$$Y_{ij} = \mu + \tau_j + \beta_i + \varepsilon_{ij}$$

- Emphasized linear models
- Focus on hypothesis testing
- Strong ties to data
- Weak links to biology

Theoretical Models

$$R_i^* = \frac{k_i m_i}{r_i - m_i}$$

- Emphasized non-linear models
- Focus on model analysis and validation
- Weak ties to data
- Strong links to biology

The purpose is to join the two.

Four ways to write up a model.

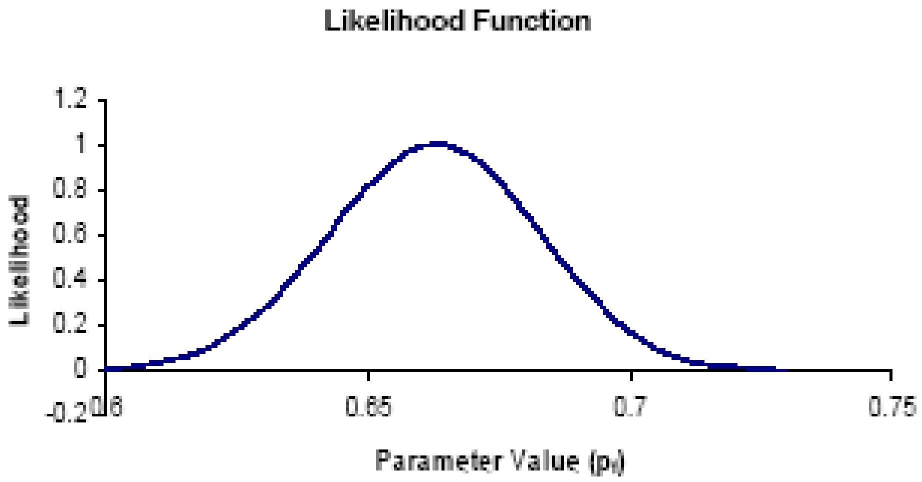
- English – Likelihood is the probability of observing the data given a parameter.

- Analysis (Equations)

$$\mathcal{L}(\theta \mid x_1, x_2, x_3 \ldots x_n) = \prod_{i=1}^{n} cg(x_i \mid \theta)$$

$$\mathcal{L}(p_f \mid data) = \prod_{i=1}^{4} p_f^{\,x_i}(1-p_f)^{n_i - x_i}$$

- Figures



- Simulations
- sum(dbinom(46:54, 100, 0.5))
- plot (k, dbinom(k, n, pi/10, log=TRUE), type='l', ylab="log density", main = "dbinom(*, log=TRUE) is better than log(dbinom(*))")

Lecture Next Thursday: I will begin to explain what likelihood methods are.

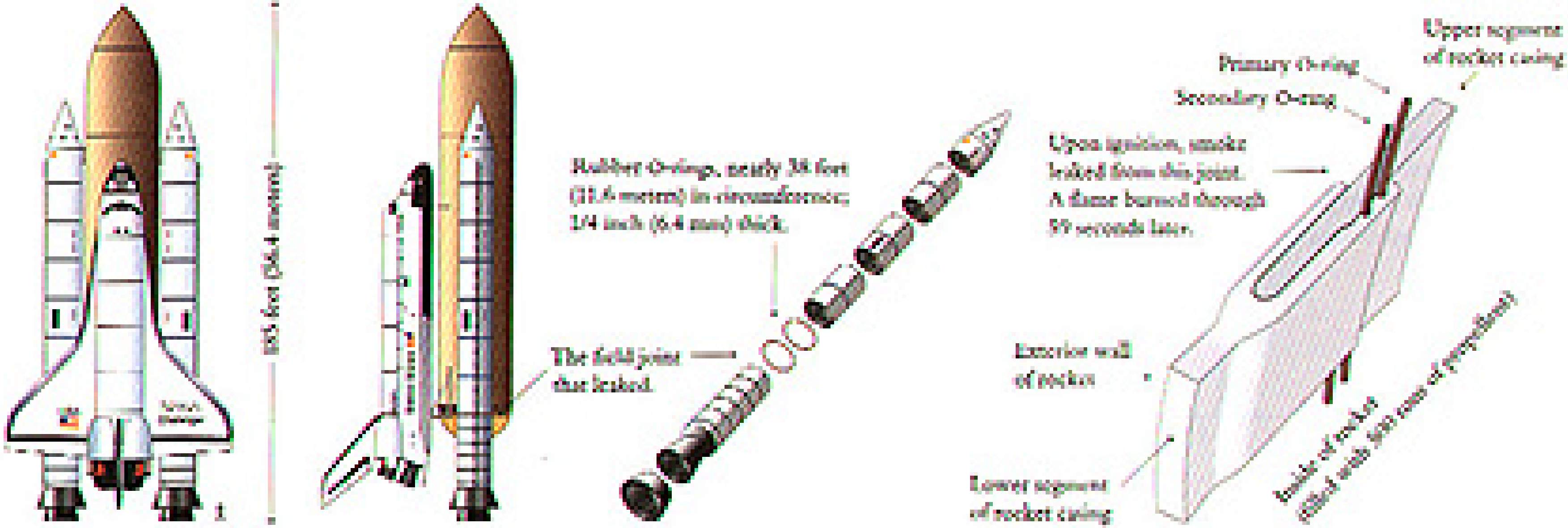
Statistics Covered:

- Graphical Methods
- Likelihood and Bayesian Methods
- Generalized linear models
- Mixed effects models
- Meta-analysis
- Fitting dynamical models to data
- State space models

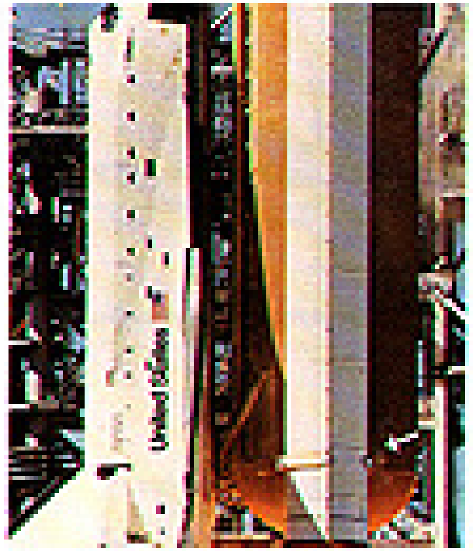
Difference between physics and biology

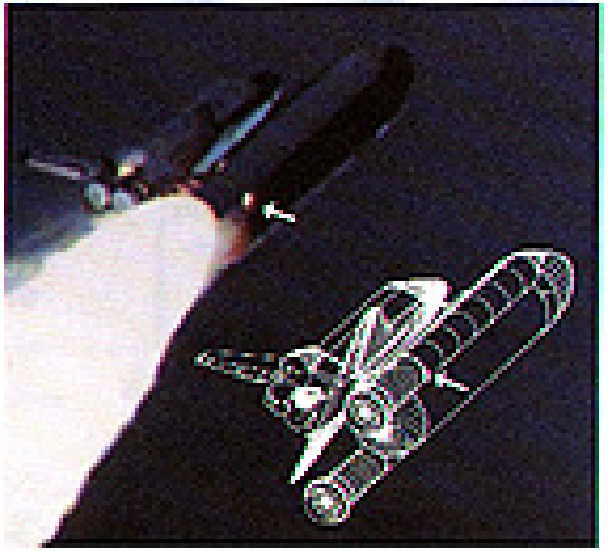
- Physics has laws
- Biology has lawyers looking for loopholes.

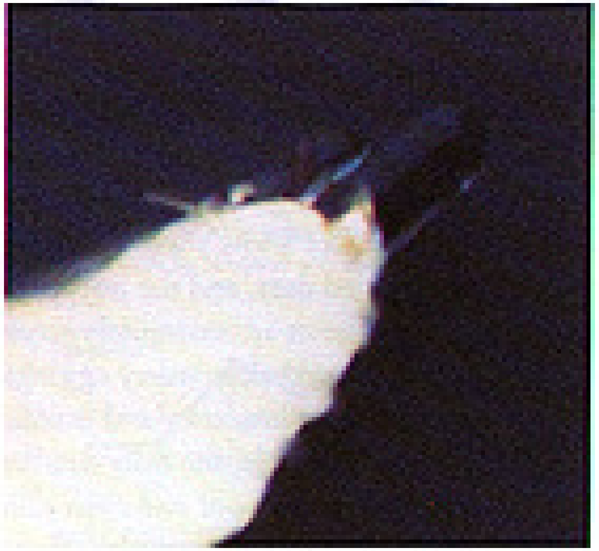
Why is it important to learn mathematical modelling?


185 feet (56.4 meters)
The field joint that leaked
Rubber O-rings, nearly 38 feet (11.6 meters) in circumference; 1/4 inch (6.4 mm) thick
Upper segment of rocket casing
Primary O-ring
Secondary O-ring
Upon ignition, smoke leaked from this joint. A flame burned through 59 seconds later.
Exterior wall of rocket
Lower segment of rocket casing
Inside of rocket (filled with fuel core of propellant)

The shuttle consists of an orbiter (which carries the crew and has powerful engines in the back), a large liquid-fuel tank for the orbiter engines, and 2 solid-fuel booster rockets mounted on the sides of the central tank. Segments of the booster rockets are shipped to the launch site, where they are assembled to make the solid-fuel rockets. Where these segments mate, each joint is sealed by two rubber O-rings as shown above. In the case of the Challenger accident, one of these joints leaked, and a torch-like flame burned through the side of the booster rocket.

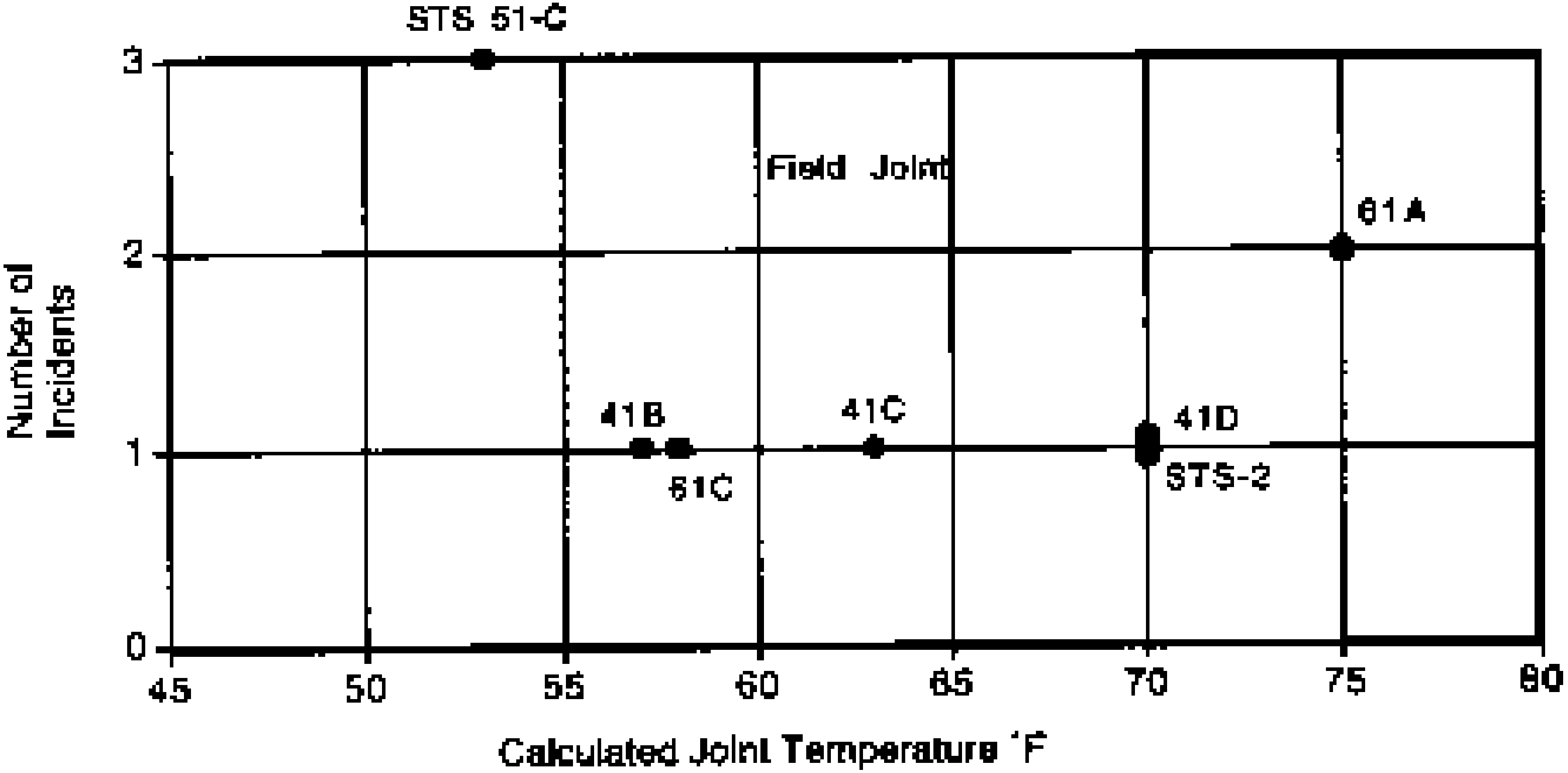




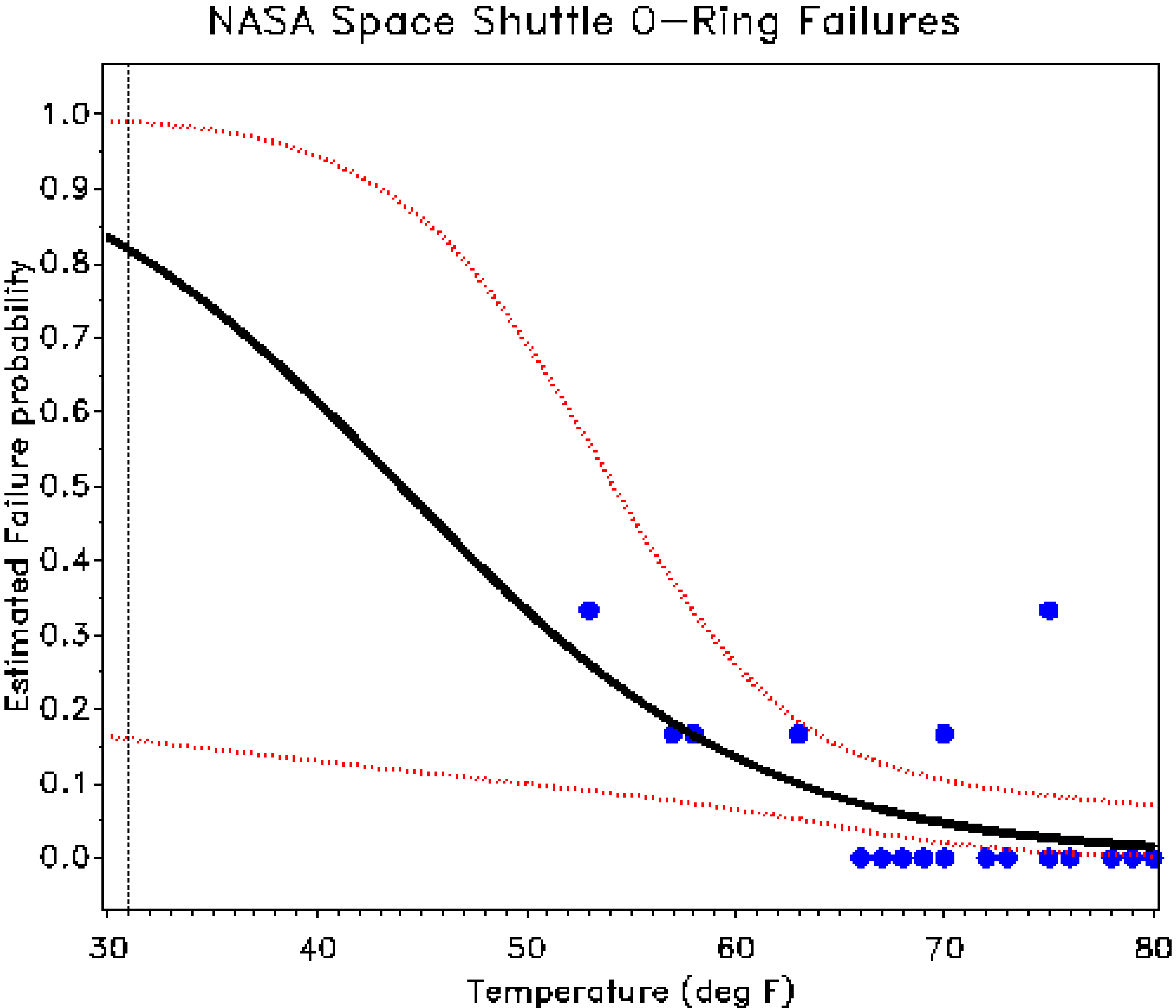


The Space Shuttle Challenger exploded shortly after take-off in January 1986. Subsequent investigation determined that the cause was failure of the O-ring seals used to isolate the fuel supply from burning gases.

NASA staff had analysed the data on the relation between ambient temperature and number of O-ring failures (out of 6), but they had excluded observations where no O-rings failed, believing that they were uninformative. Unfortunately, those observations had occurred when the launch temperature was relatively warm (65-80 degF).



There's not much data at low temperatures (the confidence band is quite wide), but the predicted probability of failure is uncomfortably high. Would you take a ride on Challenger when the weather is cold?



Look at the data

- Download the package DAAG
- Print out “orings”

	Temperature	Erosion	Blowby	Total
1	53	3	2	5
2	57	1	0	1
3	58	1	0	1
4	63	1	0	1
5	66	0	0	0
6	67	0	0	0
7	67	0	0	0
8	67	0	0	0
9	68	0	0	0
10	69	0	0	0
11	70	1	0	1
12	70	0	0	0
13	70	1	0	1
14	70	0	0	0
15	72	0	0	0
16	73	0	0	0
17	75	0	0	0
18	75	0	2	1
19	76	0	0	0
20	76	0	0	0
21	78	0	0	0
22	79	0	0	0
23	81	0	0	0

Program

```
oldpar <- par(mfrow=c(1,2))
plot(Total~Temperature, data = orings[c(1,2,4,11,13,18),])
# the observations included in the pre-launch charts
plot(Total~Temperature, data = orings)
par(oldpar)
```

Lessons for biologists:

- Don’t ignore zeros! Many students who go into biology appear to be genetically predisposed to ignore zeros.
- Distinguish between zeros, missing values, and censored values.
- The right statistical model, in this case a generalized linear model with a logit link and binomial error or a log link with a negative binomial error.

- Data + Theory
- Modelling (i.e. Information)
- Estimation (i.e Evidence)
- Action

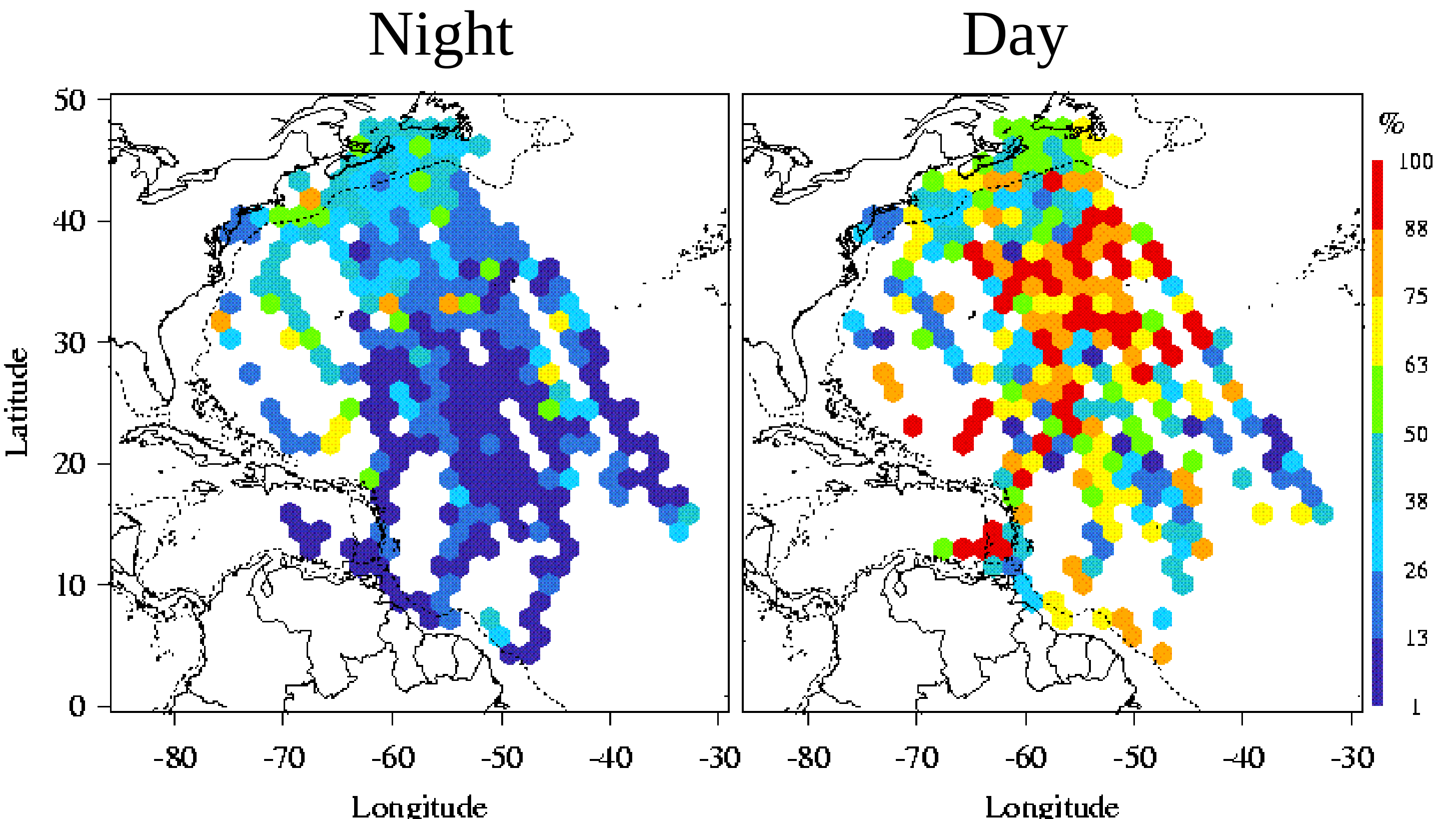
- Like good writing, good graphical displays of data communicate ideas with clarity, precision, and efficiency.
- Like poor writing, bad graphical displays distort or obscure the data, make it harder to understand or compare, or otherwise thwart the communicative effect which the graph should convey.

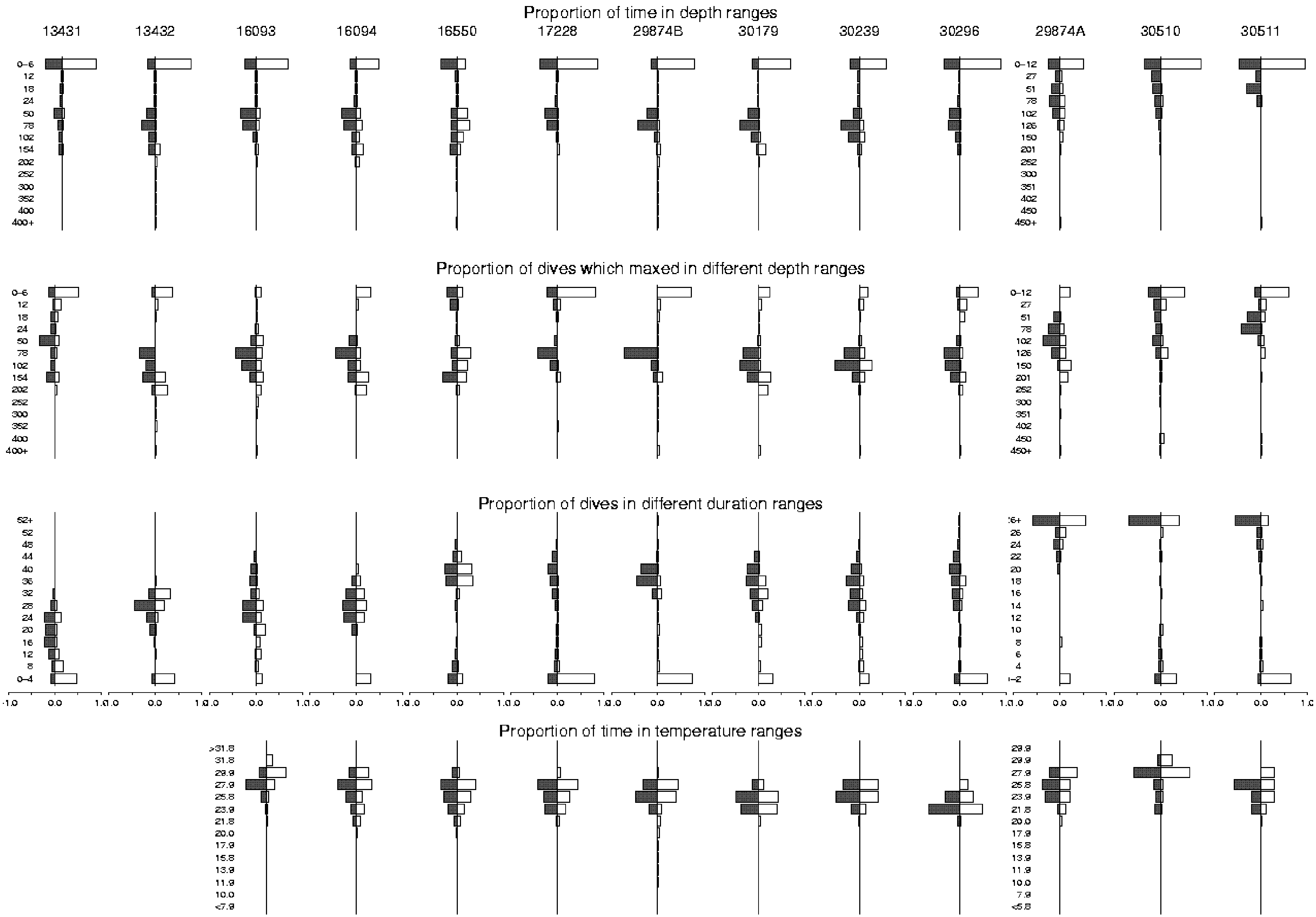
- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of a data set

When ever thinking about a new problem, it is worth thinking very hard about an optimal presentation of the data:

Leatherbacks change their behaviour as the migrate south

Proportion of time at surface




Proportion of time in depth ranges
13431
13432
16093
16094
16550
17228
29874B
30179
30239
30296
29874A
30510
30511
0–6
12
18
24
50
78
102
154
202
252
300
352
400
400+
0–12
27
51
78
102
126
150
201
252
300
351
402
450
450+
Proportion of dives which maxed in different depth ranges
Proportion of dives in different duration ranges
52+
52
48
44
40
36
32
28
24
20
16
12
8
0–4
36+
26
24
22
20
18
16
14
12
10
8
6
4
1–2
-1.0
0.0
1.0
Proportion of time in temperature ranges
>31.8
31.8
29.9
27.9
25.8
23.9
21.8
20.0
17.9
15.8
13.9
11.9
10.0
<7.9
29.9
29.9
27.9
25.8
23.9
21.8
20.0
17.9
15.8
13.9
11.9
10.0
7.9
<5.8

Determining the Effects of Exploitation on Shark Populations Using Fishery Dependent Data

Julia Baum

Dan Kehler

Ransom A. Myers

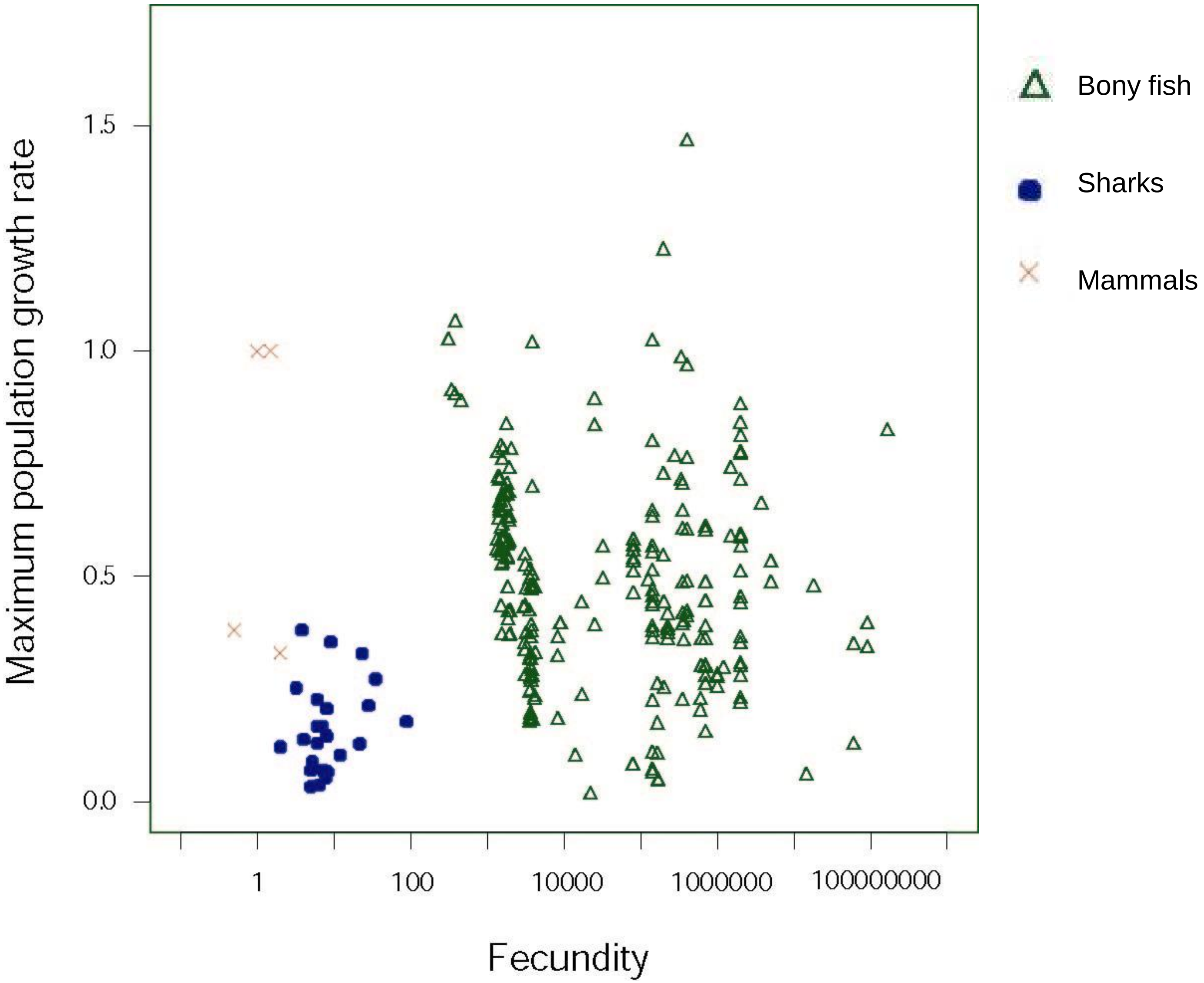
DALHOUSIE
University

Life history of sharks…

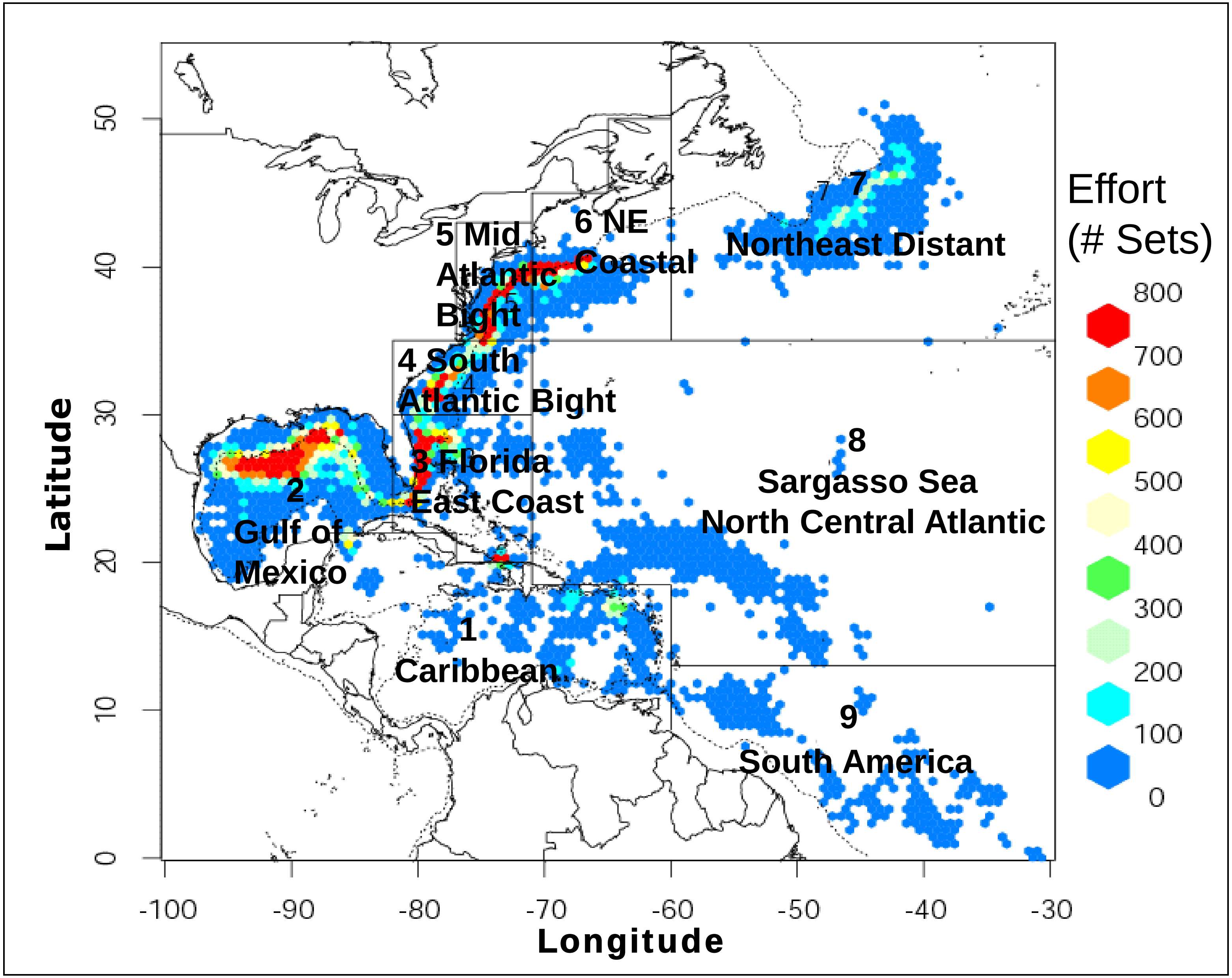
- slow growth rates
- late age at maturity
- low fecundity

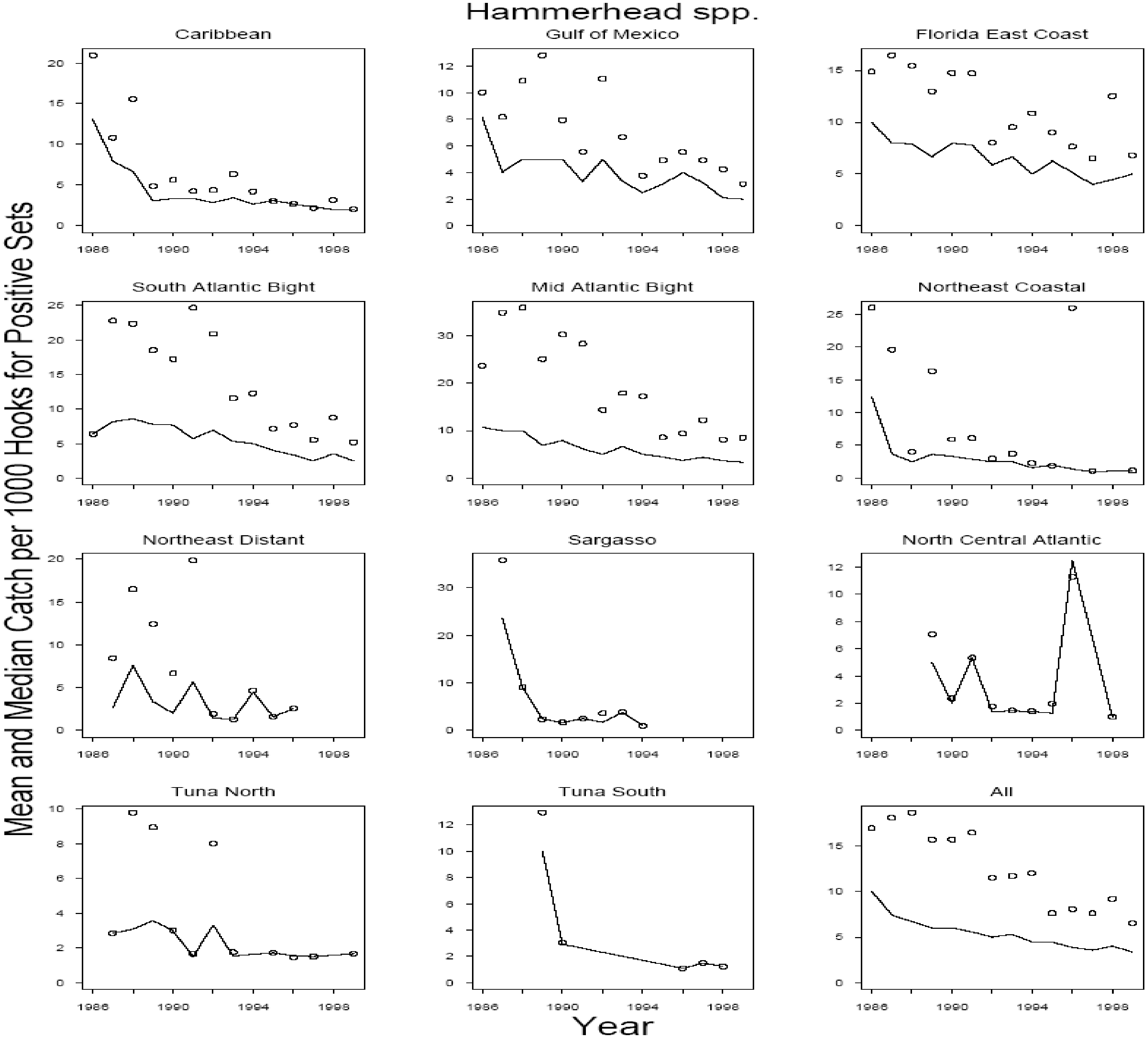
= Vulnerable to Overexploitation

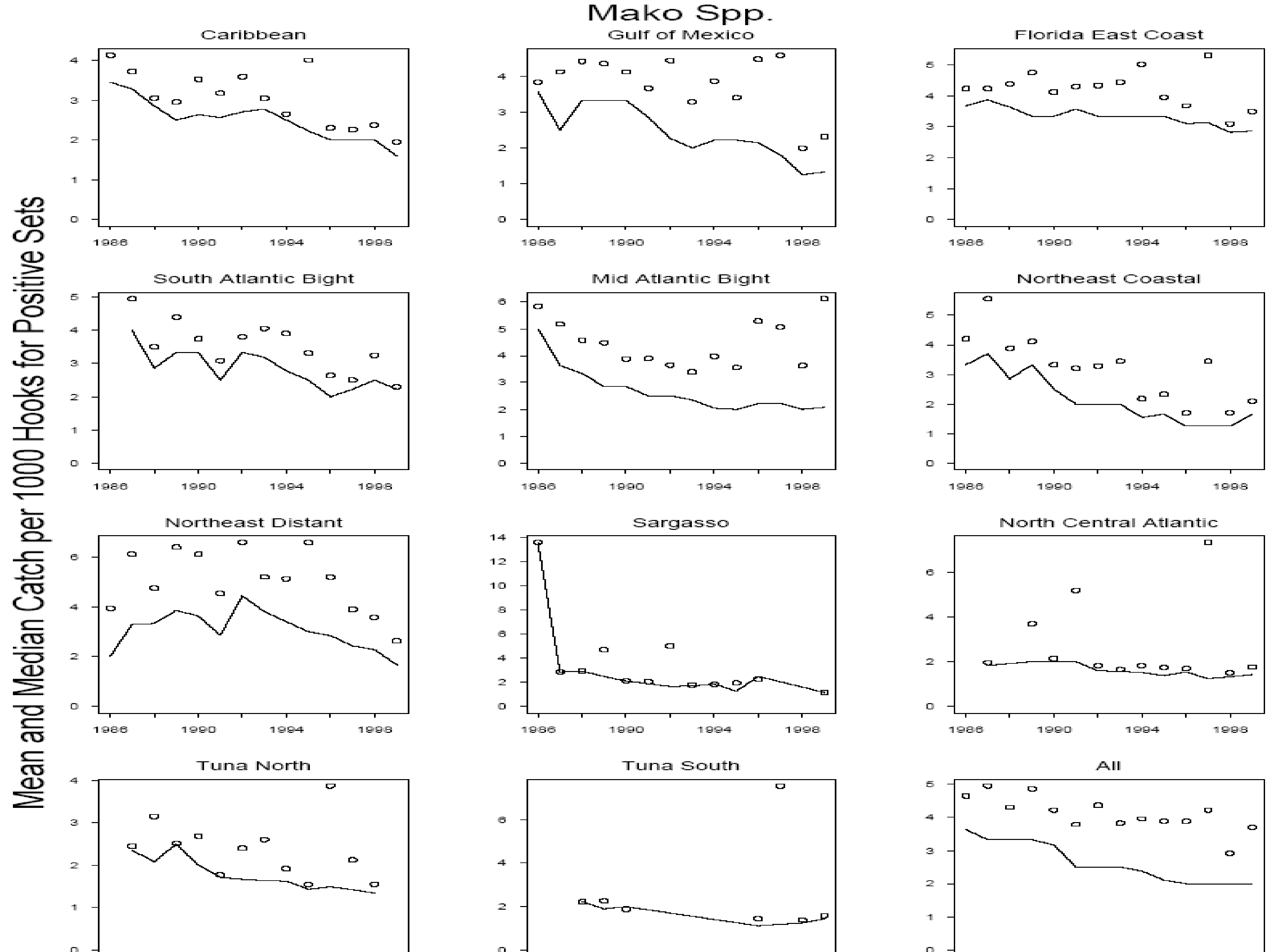
Life history of sharks...



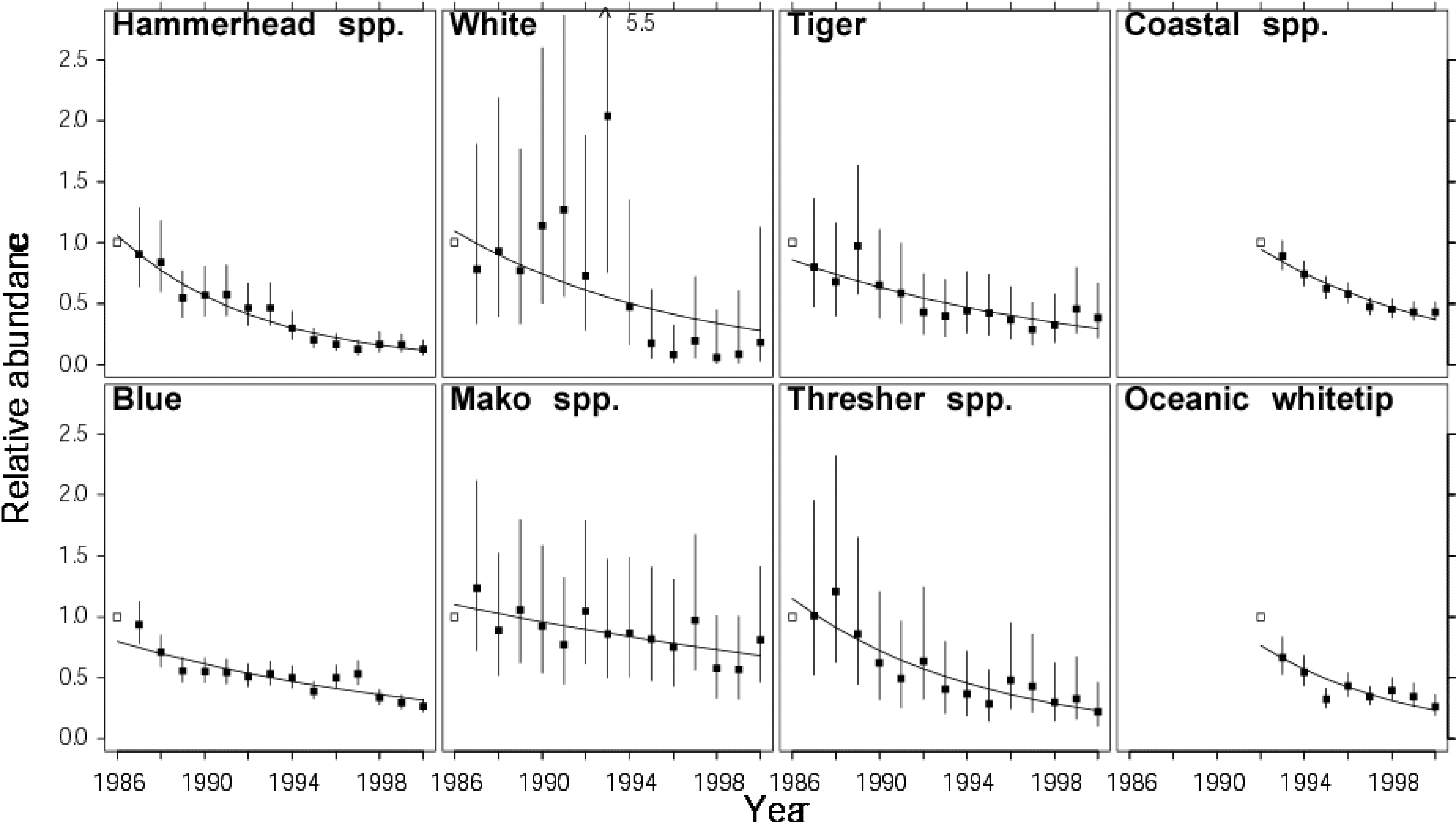
U.S. Atlantic Pelagic Longline Sets: 1986-2000



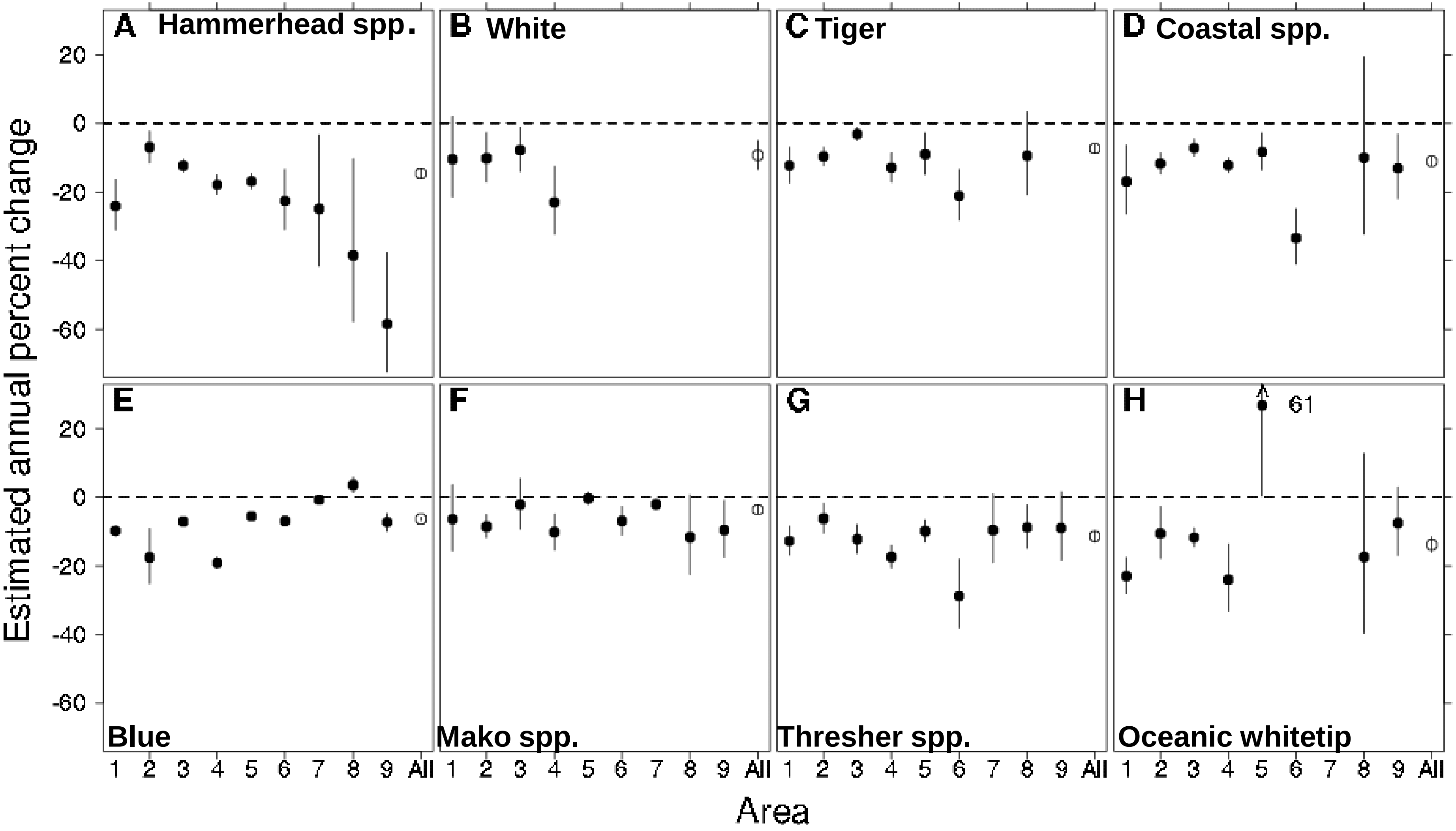

Hammerhead spp.
Caribbean
Gulf of Mexico
Florida East Coast
South Atlantic Bight
Mid Atlantic Bight
Northeast Coastal
Northeast Distant
Sargasso
North Central Atlantic
Tuna North
Tuna South
All
Mean and Median Catch per 1000 Hooks for Positive Sets
Year
1986
1990
1994
1998
0
5
10
15
20
25
30
2
4
6
8
12


Mako Spp.
Mean and Median Catch per 1000 Hooks for Positive Sets
Caribbean
0
1
2
3
4
1986
1990
1994
1998
Gulf of Mexico
0
1
2
3
4
1986
1990
1994
1998
Florida East Coast
0
1
2
3
4
5
1986
1990
1994
1998
South Atlantic Bight
0
1
2
3
4
5
1986
1990
1994
1998
Mid Atlantic Bight
0
1
2
3
4
5
6
1986
1990
1994
1998
Northeast Coastal
0
1
2
3
4
5
1986
1990
1994
1998
Northeast Distant
0
2
4
6
1986
1990
1994
1998
Sargasso
0
2
4
6
8
10
12
14
1986
1990
1994
1998
North Central Atlantic
0
2
4
6
1986
1990
1994
1998
Tuna North
0
1
2
3
4
Tuna South
0
2
4
6
All
0
1
2
3
4
5

Results


Hammerhead spp.
White
5.5
Tiger
Coastal spp.
Blue
Mako spp.
Thresher spp.
Oceanic whitetip
Relative abundane
0.0
0.5
1.0
1.5
2.0
2.5
1986
1990
1994
1998
Year

1 Caribbean
2 Gulf of Mexico
3 Florida
4 S Atlantic Bight
5 Mid Atlantic Bight
6 NE Coastal
7 NE Distant
8 Sargasso
9 S America



Data Analysis

- Assume catch follows negative binomial distribution
- Analyse positives only → zero-truncated distribution

$$f(y_T) = \frac{\dfrac{\Gamma(y+\theta)}{\Gamma(y)}^{y_T}\left(\dfrac{\mu}{\theta+\mu}\right)^{y_T}\left(\dfrac{\theta}{\theta+\mu}\right)^{\theta}}{1-\left(\dfrac{\theta}{\theta+\mu}\right)^{\theta}}$$

Data Analysis

Parameter estimation: Generalized linear models:

TNB with fixed θ is a one-parameter exponential family of distributions

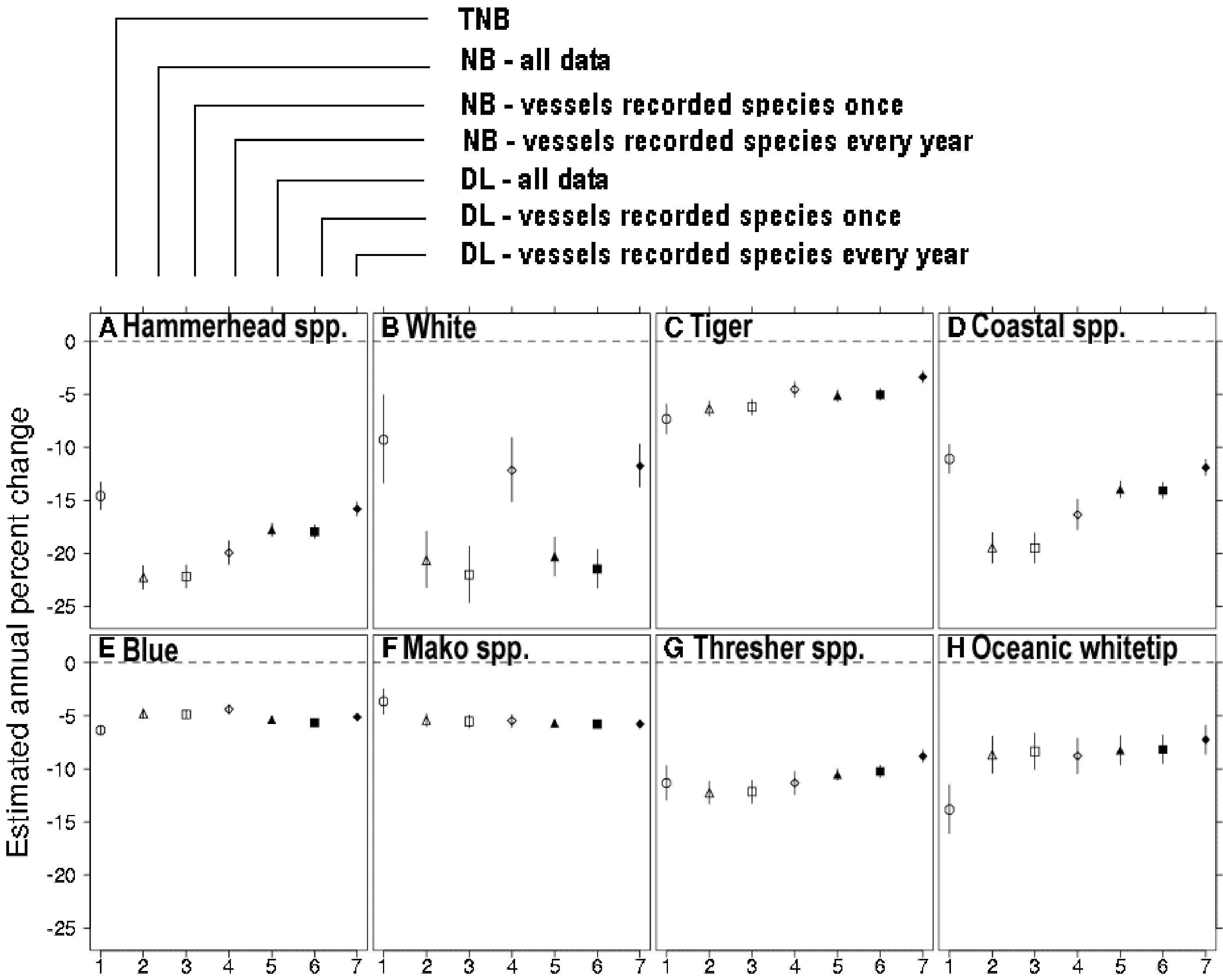
Base model

Main effects: area, season, light sticks, temperature and year

Interactions: area*season, area*light

Multiple tests for robustness

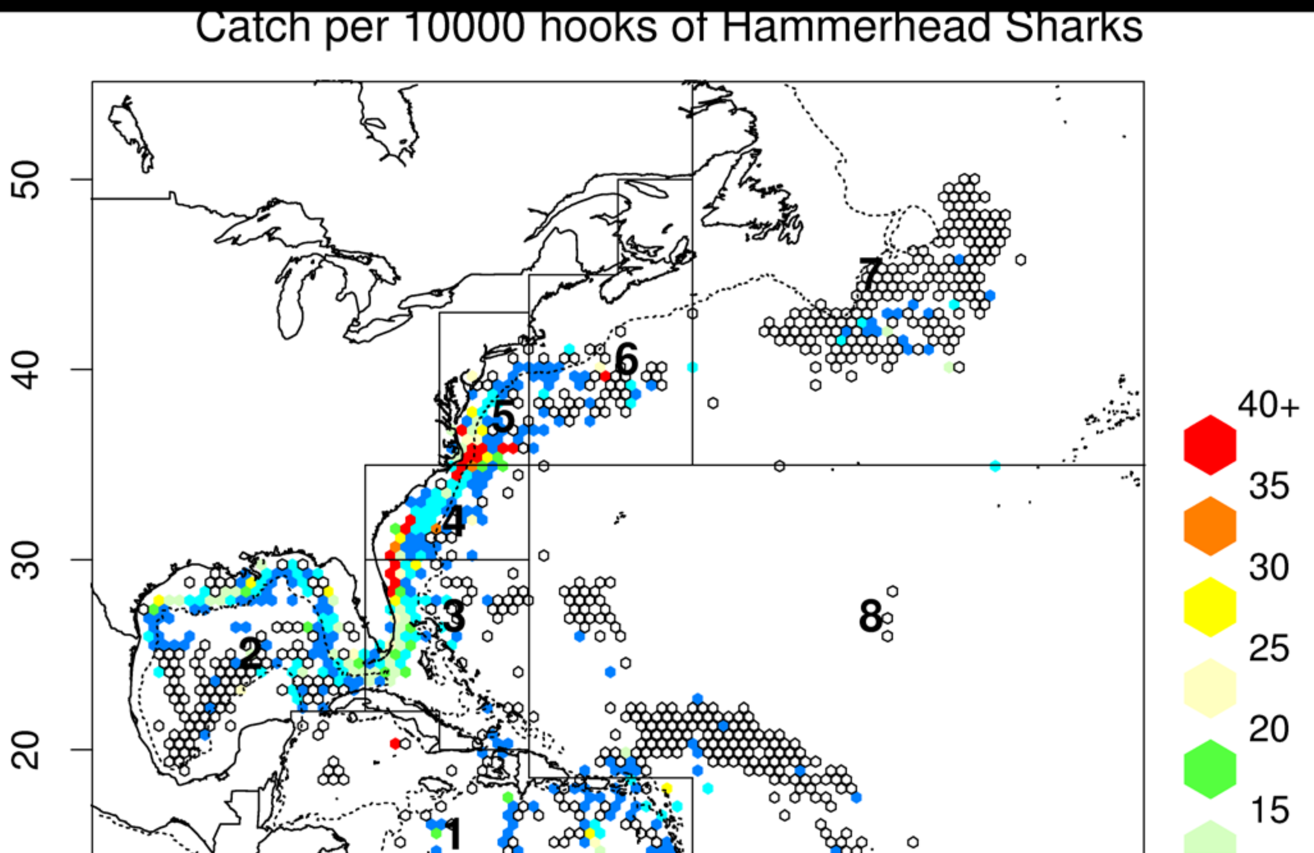
E



Hammerhead sharks

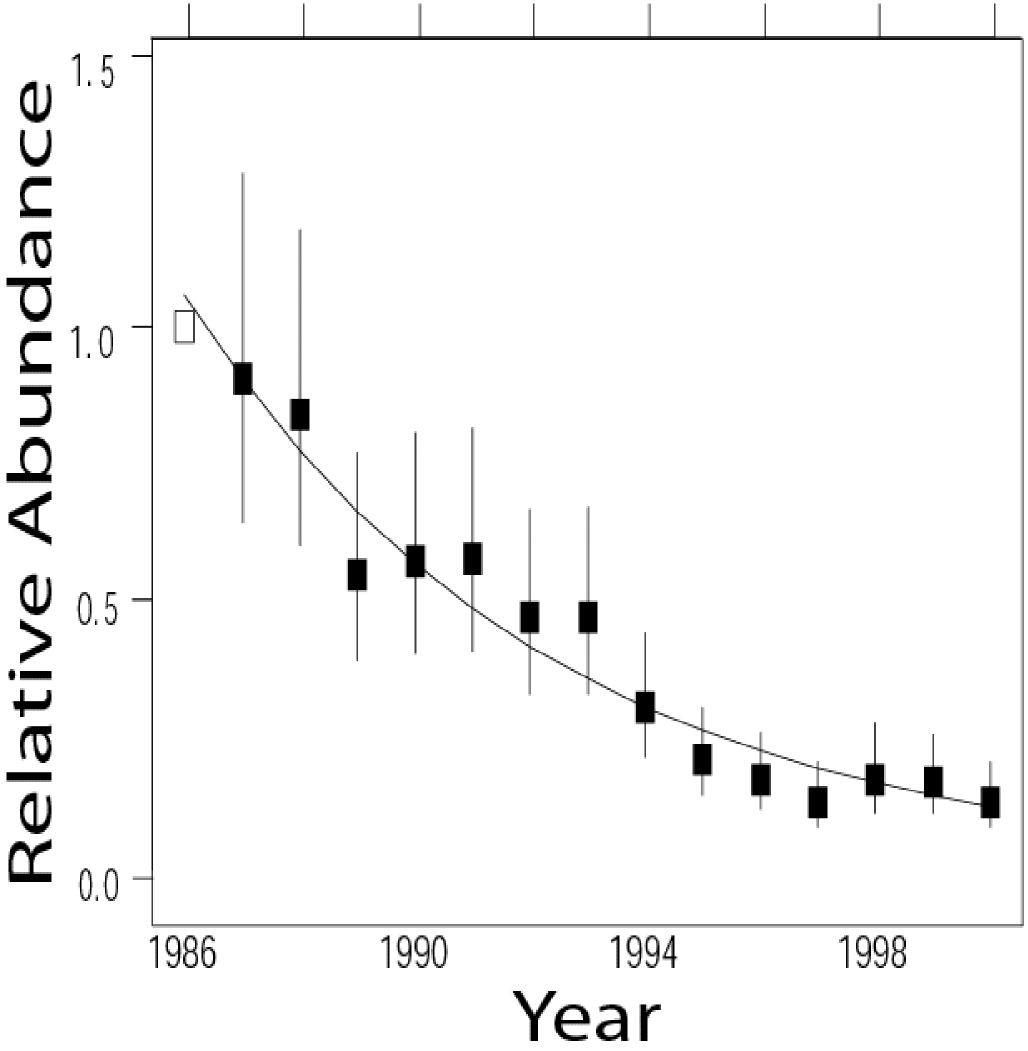
[redacted]

[redacted]



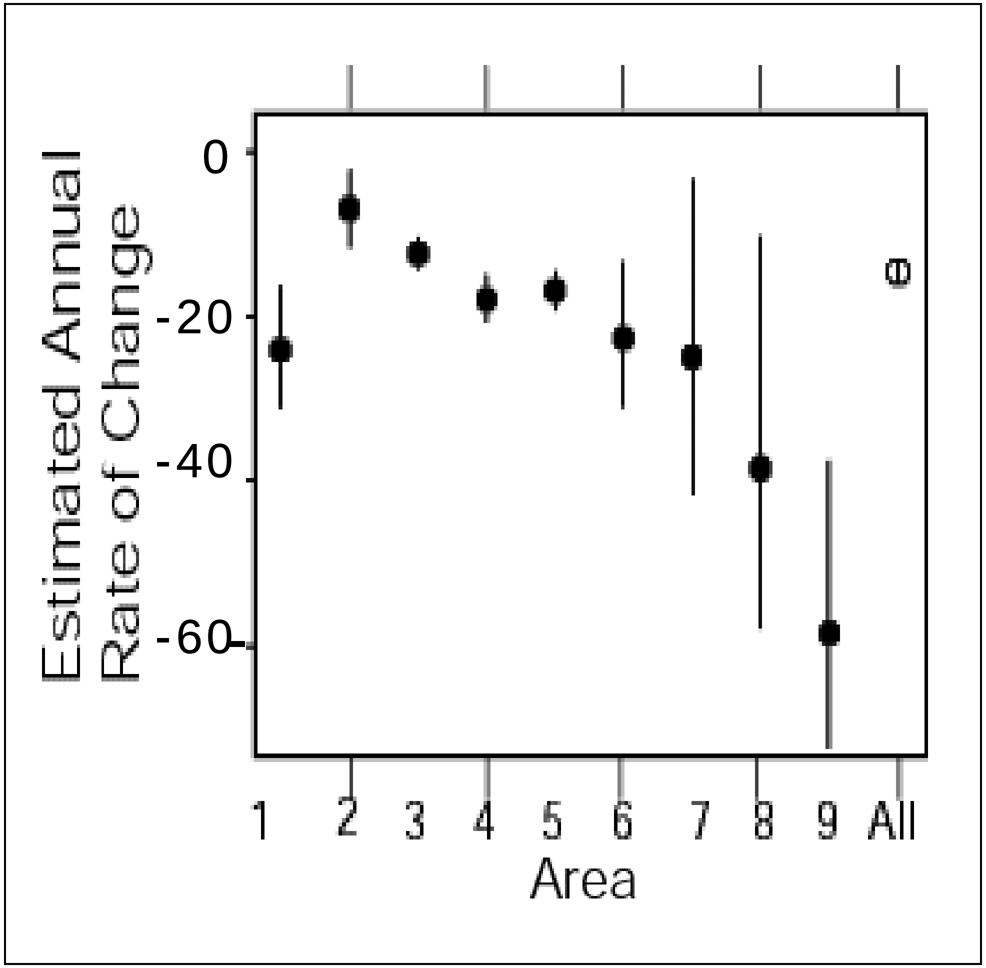
Hammerhead sharks

Sphyrna lewini



Hammerhead sharks

Sphyrna lewini



1 Caribbean

2 Gulf of Mexico

3 Florida

4 S Atlantic Bight

5 Mid Atlantic Bight

6 NE Coastal

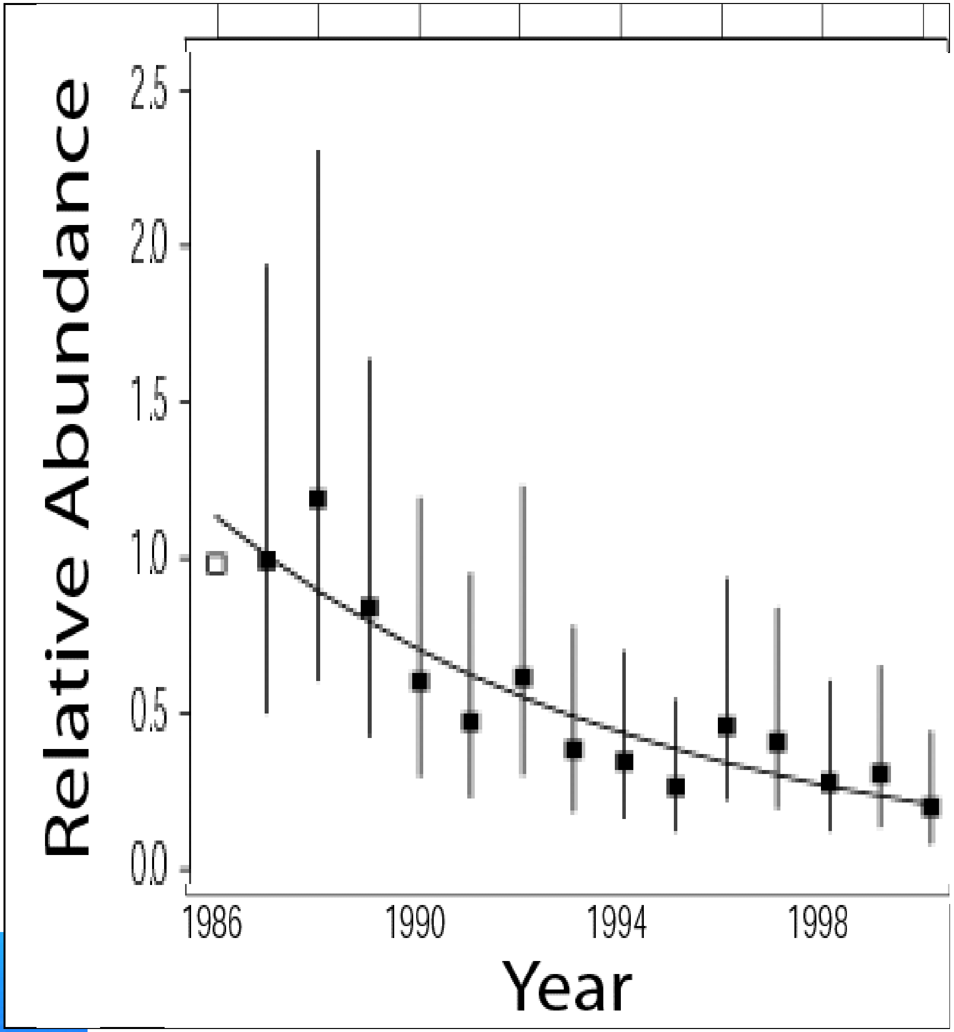
7 NE Distant

8 Sargasso

9 S America

Thresher sharks

Alopias spp.

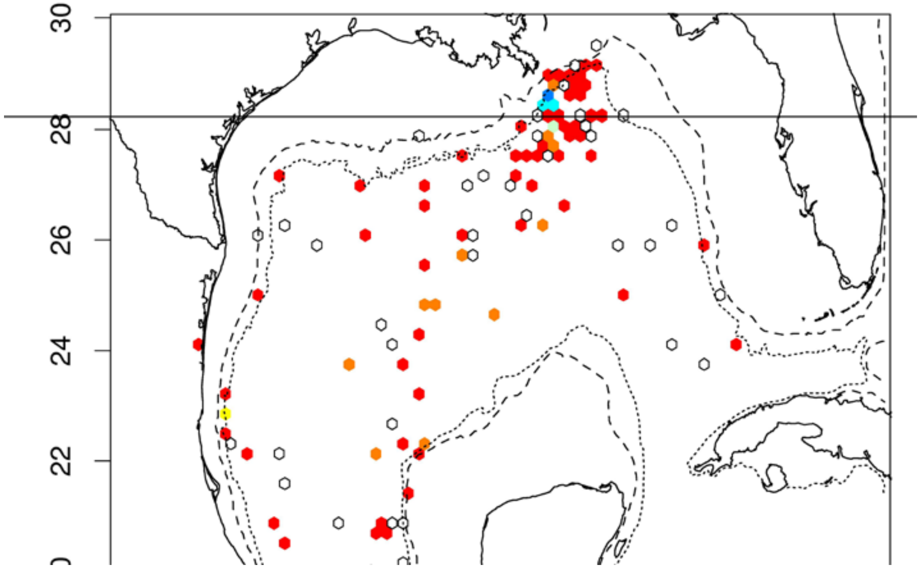


Repeat analysis using
all other independent data

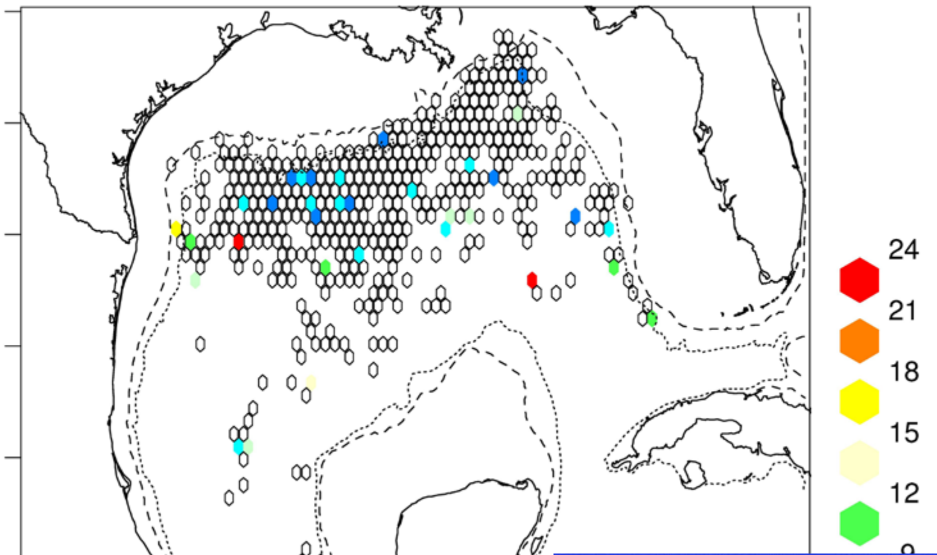
Loss of sharks in the Gulf of Mexico

300 fold decline – no one noticed

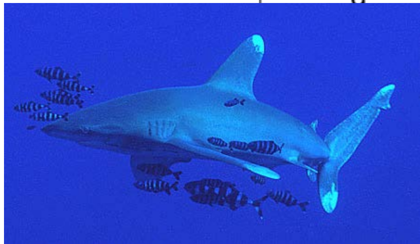
[redacted]



1950's



1990's



Oceanic Whitetip captures per 10,000 hooks

Baum and Myers, 2004 Ecology Letters

With training, “experts” can ignore the most obvious of data:

1872 - Man's head and leg and dolphin in stomach

1872 – 8 Great White Sharks reported caught

1888 - Woman's body and lamb in stomach

1894 - Preserved at Zagreb Nat. Hist. Mus.

1926 - Woman's shoes, laundry in stomach

1946 - Pig of 10 kg in stomach

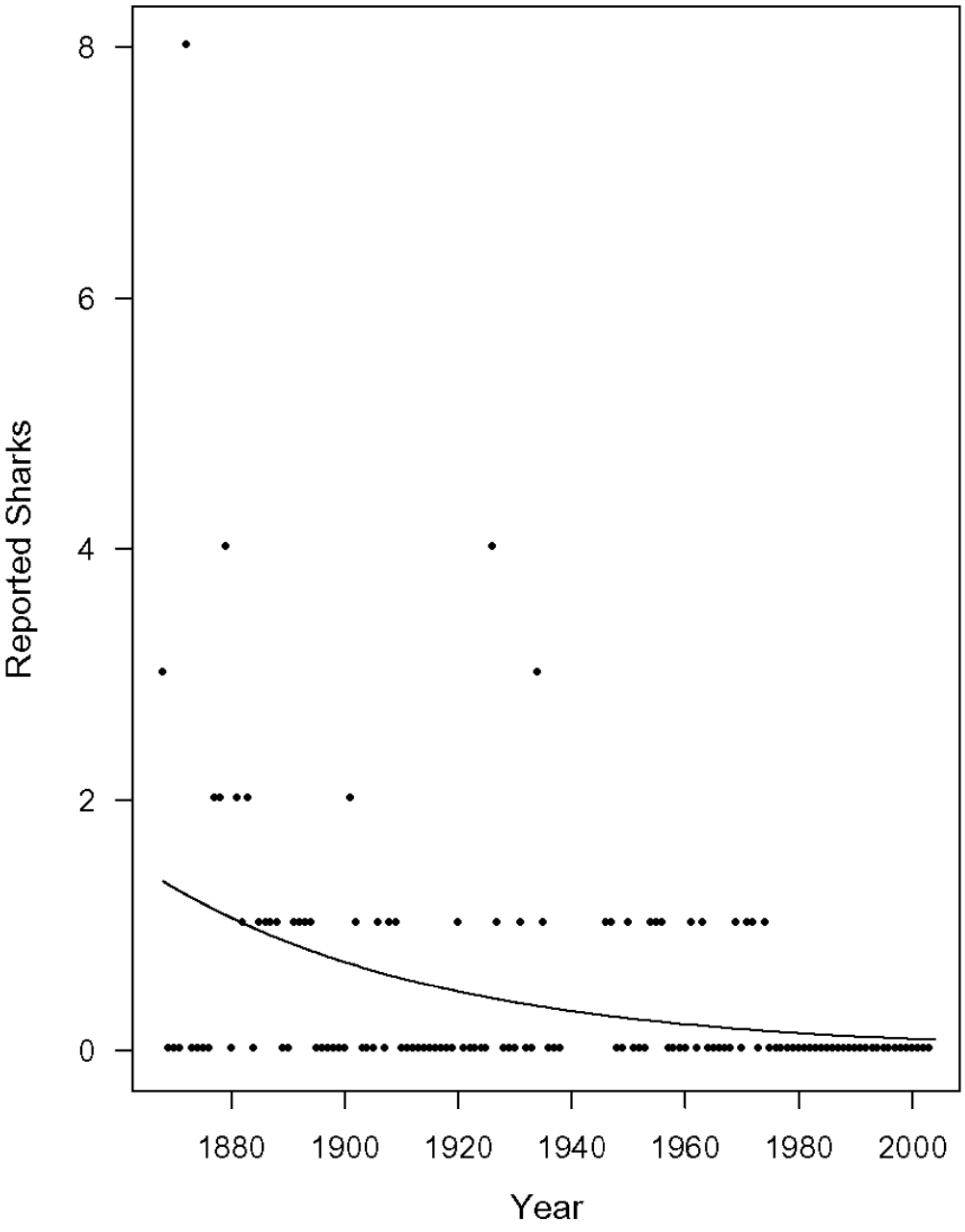
1950 - Encounter during eating a dead calf

1954 - Attack on boat

1975+ -No sightings.

Soldo and Jardas, Periodicum Bologorum, 2002

Newspaper reports of sharks in Croatia



Plotting the data and model output

It is impossible to underestimate the importance of plotting data.

Florence Nightingale made huge impacts because she knew enough math to understand problems, and she was a genus in the art of plotting data.

Florence Nightingale made huge impacts because she could plot data

- In 1840, Florence Nightingale begged her parents "to let her study mathematics instead of doing worsted work and practicing quadrilles."

Florence Nightingale invented new ways to look at data to convince the British Military of their stupidity

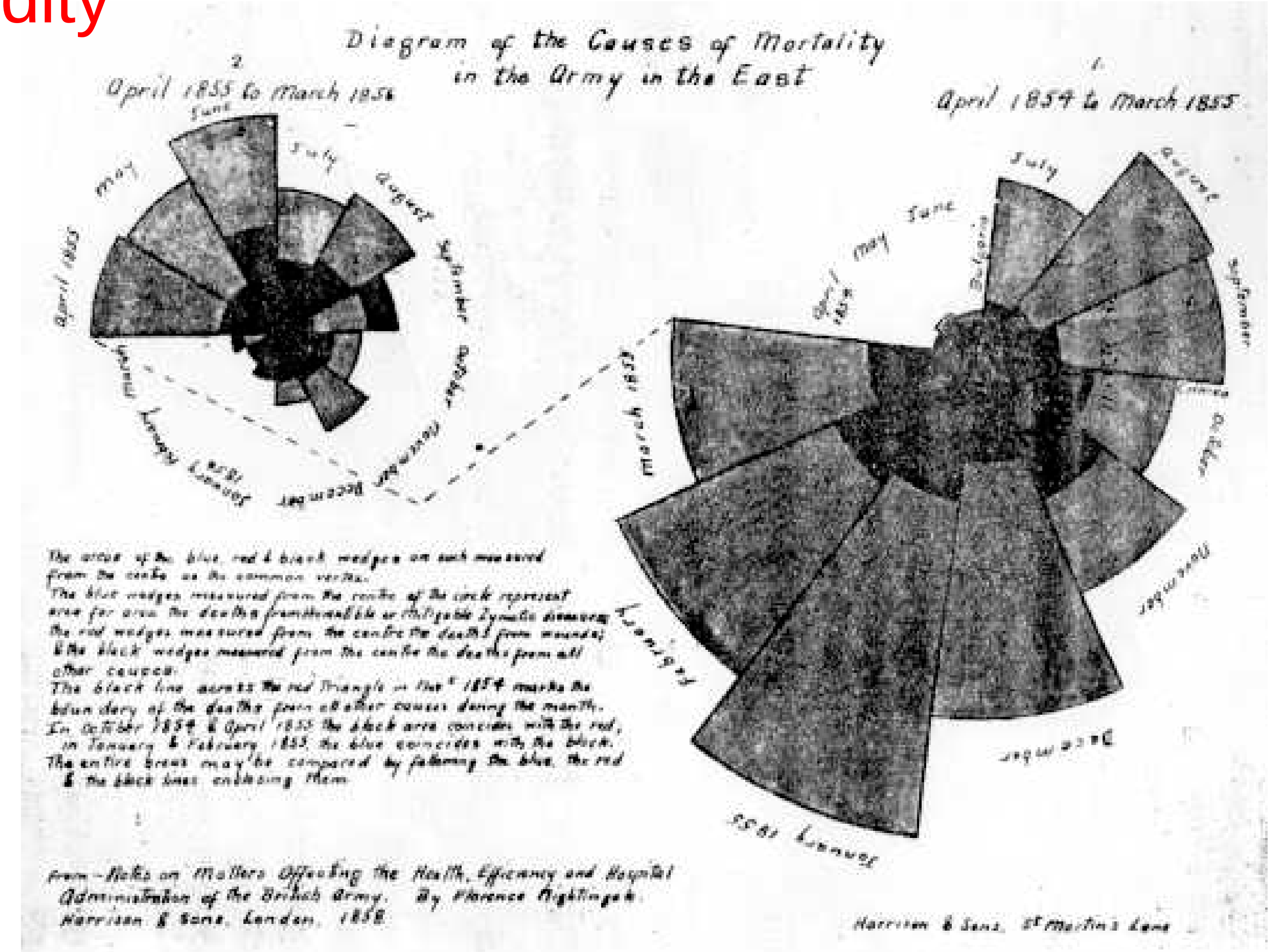


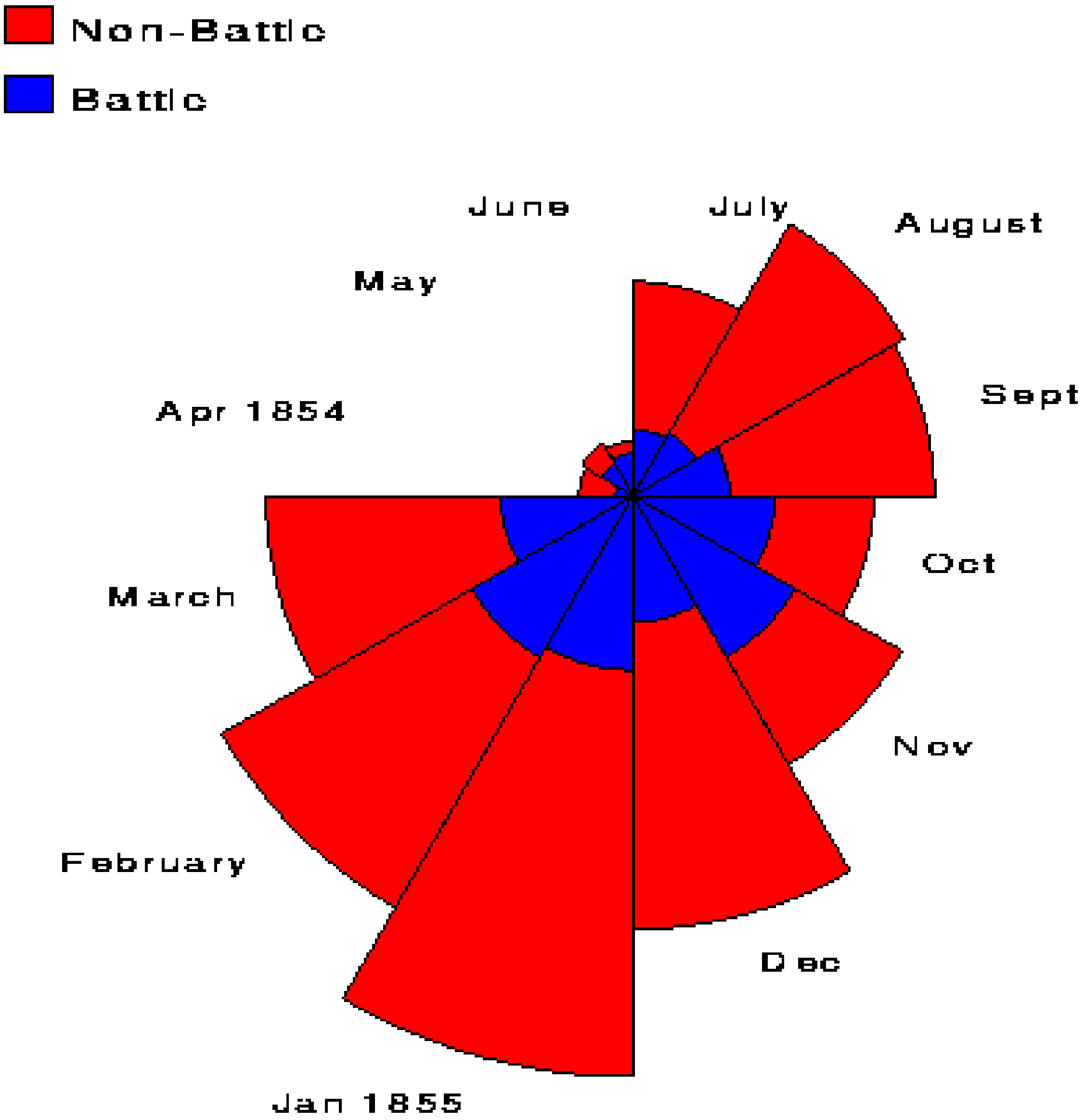
Diagram of the Causes of Mortality
in the Army in the East

The areas of the blue, red & black wedges are each measured
from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent
area for area the deaths from Preventible or Mitigable Zymotic diseases;
the red wedges measured from the centre the deaths from wounds;
& the black wedges measured from the centre the deaths from all
other causes.
The black line across the red Triangle in Nov.r 1854 marks the
boundary of the deaths from all other causes during the month.
In October 1854 & April 1855 the black area coincides with the red,
in January & February 1855 the blue coincides with the black.
The entire areas may be compared by following the blue, the red
& the black lines enclosing them.

From — Notes on Matters Affecting the Health, Efficiency and Hospital
Administration of the British Army. By Florence Nightingale.
Harrison & Sons. London. 1858.

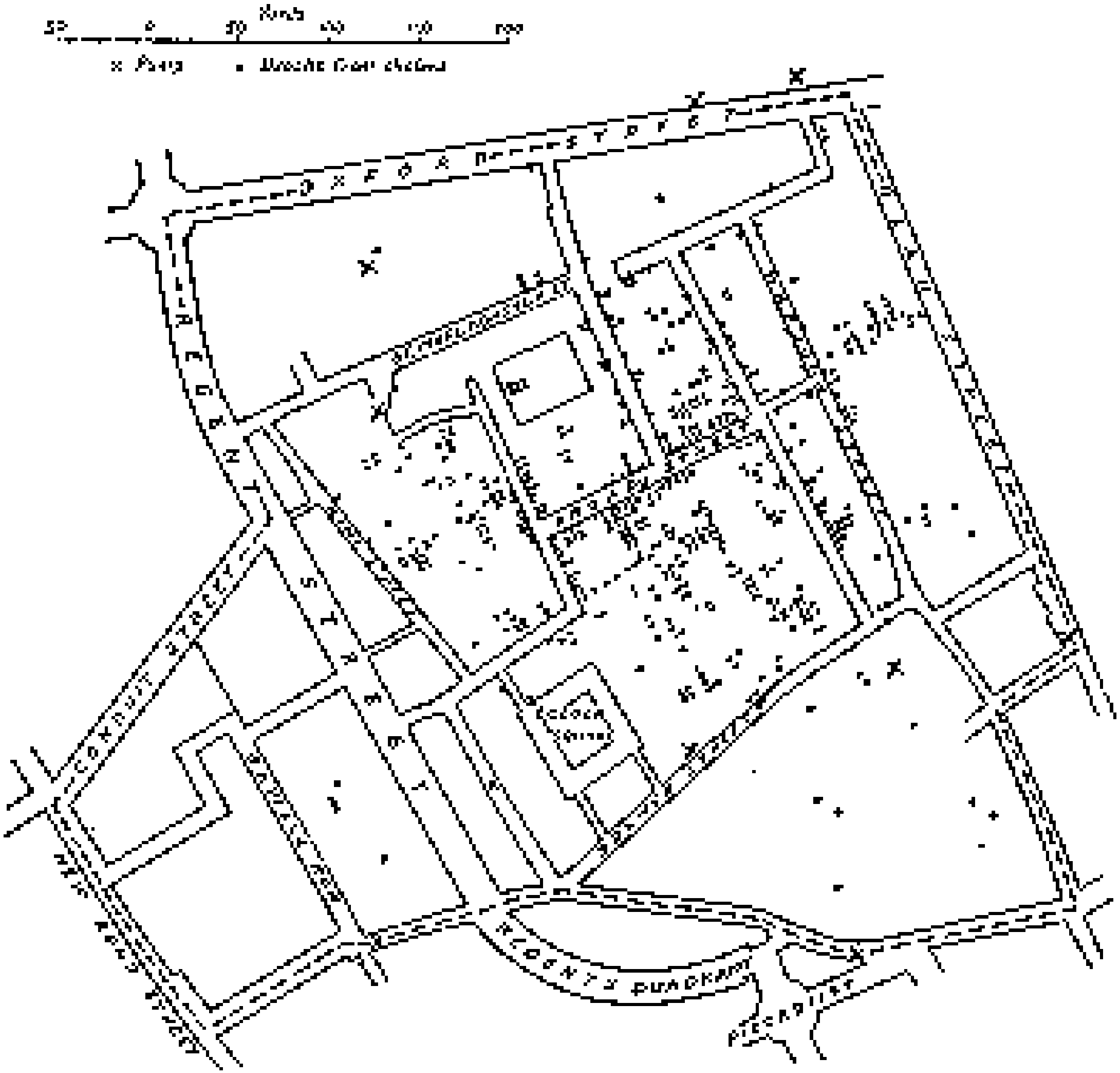
Harrison & Sons, St Martin's Lane

Causes of Mortality in the Army in the East
April, 1854 to March 1855



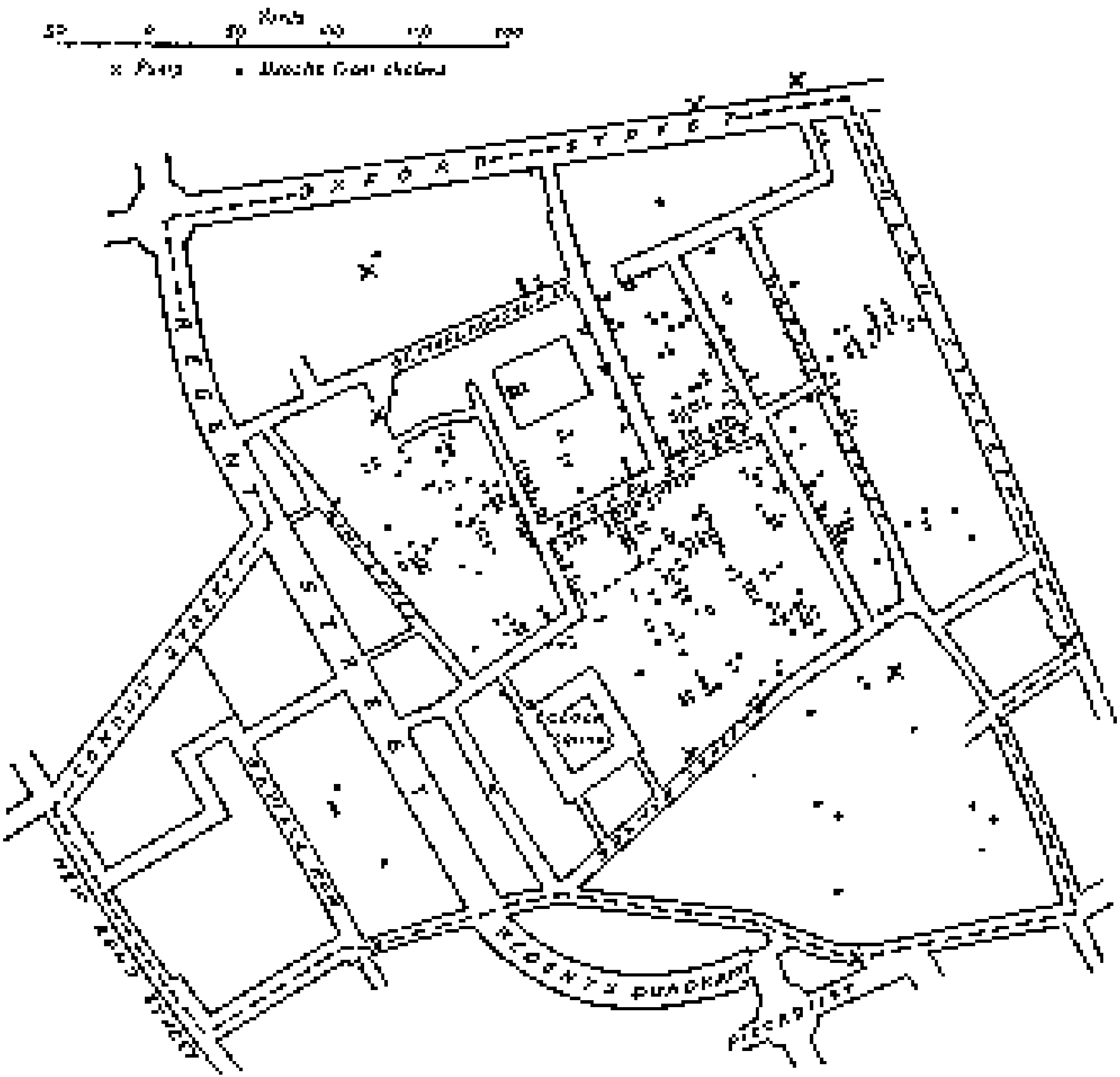
From: F. Nightingale, "Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army", 1858

The most influential plot in epidemiology:

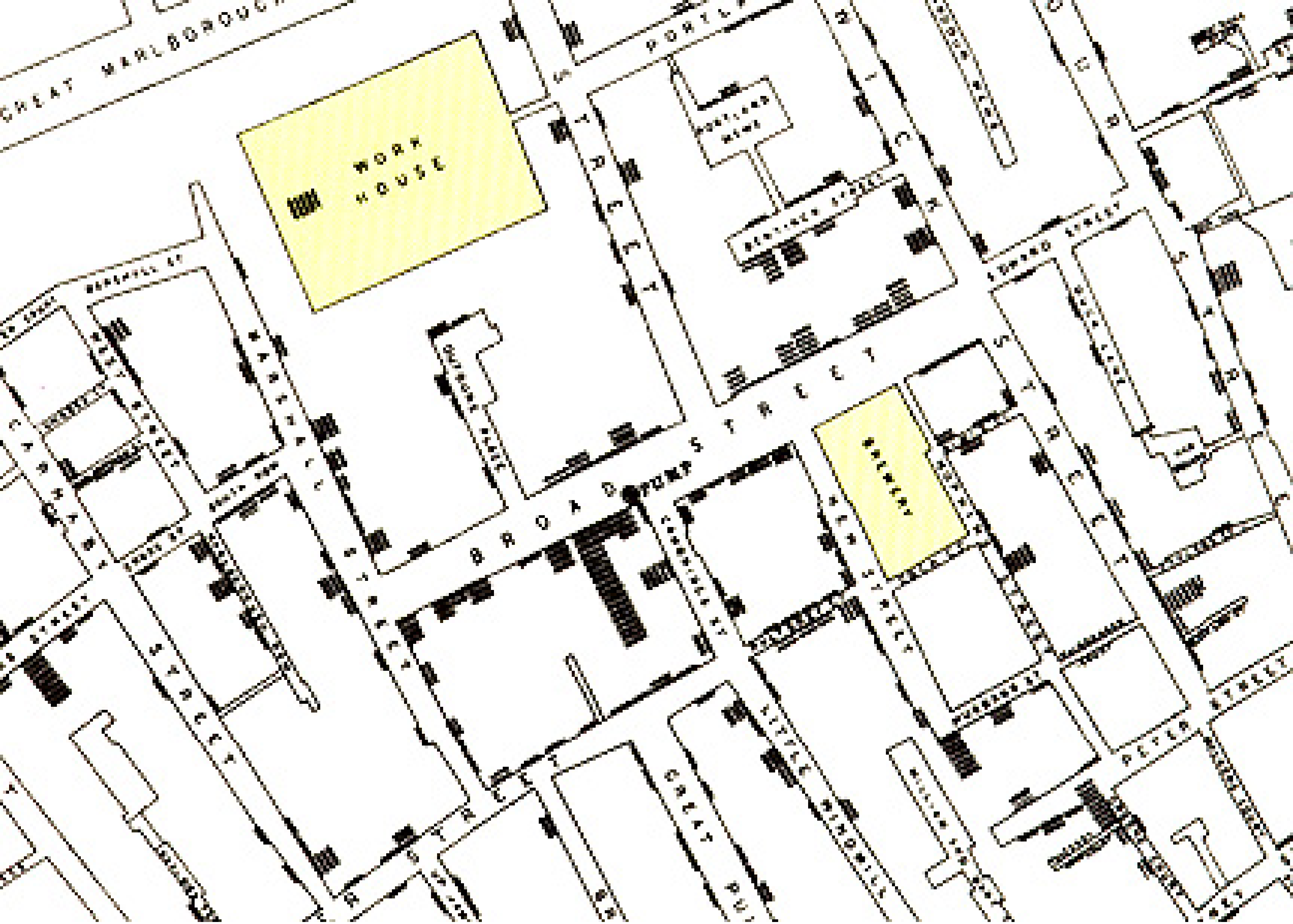


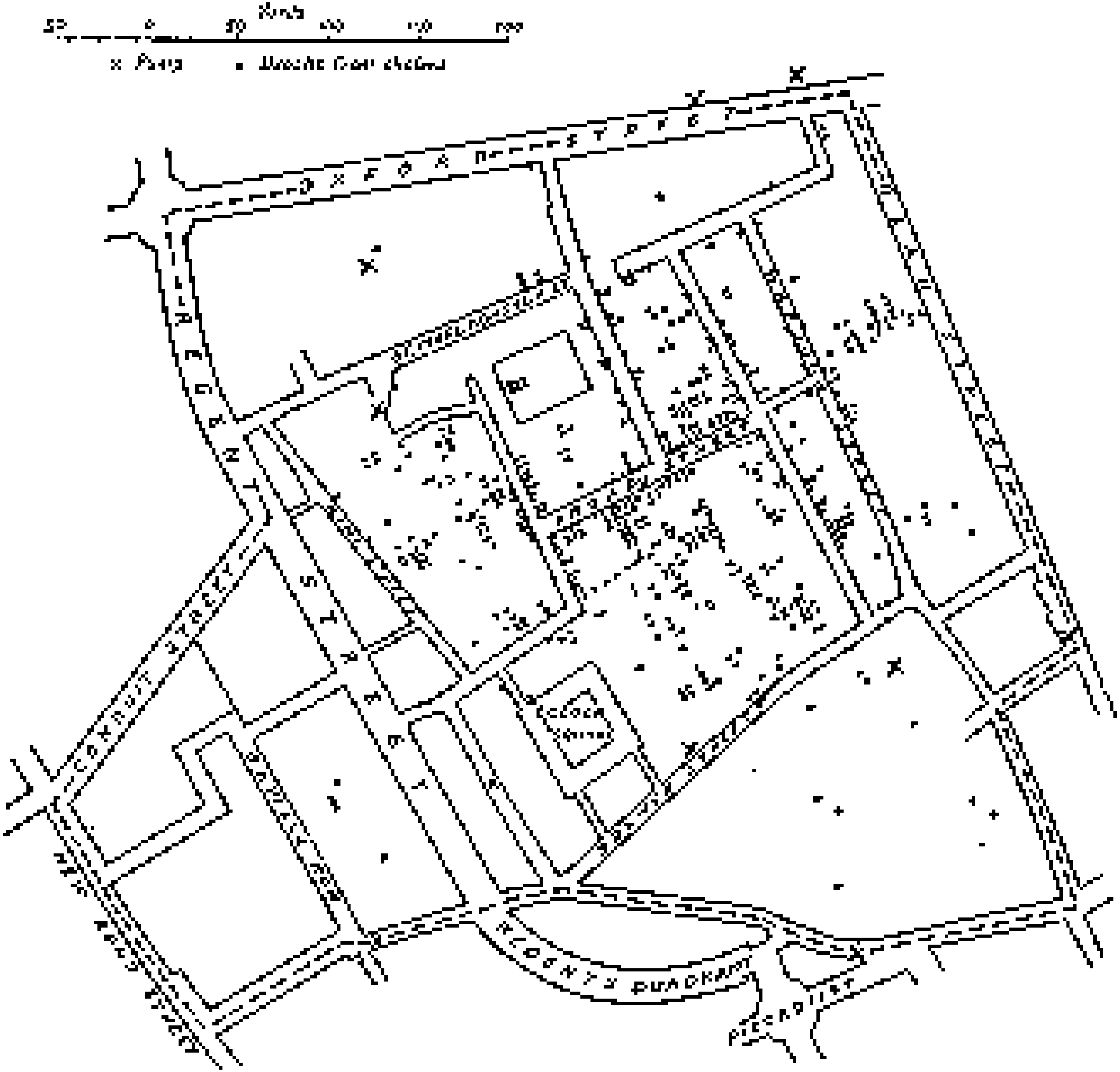
The 1854 London Cholera Epidemic.

- One of the first uses of a map to display epidemiological data was this dot chart (from Tufte, 1983, p. 24) by Dr. John Snow (1855) showing deaths from cholera (dots) in relation to the locations of public water pumps. Tufte says, "Snow observed that cholera occurred almost entirely among those who lived near (and drank from) the Broad Street water pump. He had the handle of the contaminated pump removed, ending the neighborhood epidemic which had taken more than 500 lives."



Yards
50
0
50
100
150
200
x Pump
■ Deaths from cholera
OXFORD STREET
REGENT STREET
CONDUIT STREET
GOLDEN SQUARE
REGENTS QUADRANT
PICCADILLY


GREAT MARLBOROUGH
WORK HOUSE
PORTLAND
PORTLAND MEWS
BROAD STREET
PUMP
BREWERY
MARSHALL
CAMBRIDGE ST
NEW STREET
LITTLE WINDMILL
GREAT PULTENEY
PETER STREET


Yards
50
0
50
100
150
200
x Pump
■ Deaths from cholera
OXFORD STREET
REGENT STREET
CONDUIT STREET
GOLDEN SQUARE
REGENTS QUADRANT
PICCADILLY

Next step of plotting data: look at residuals

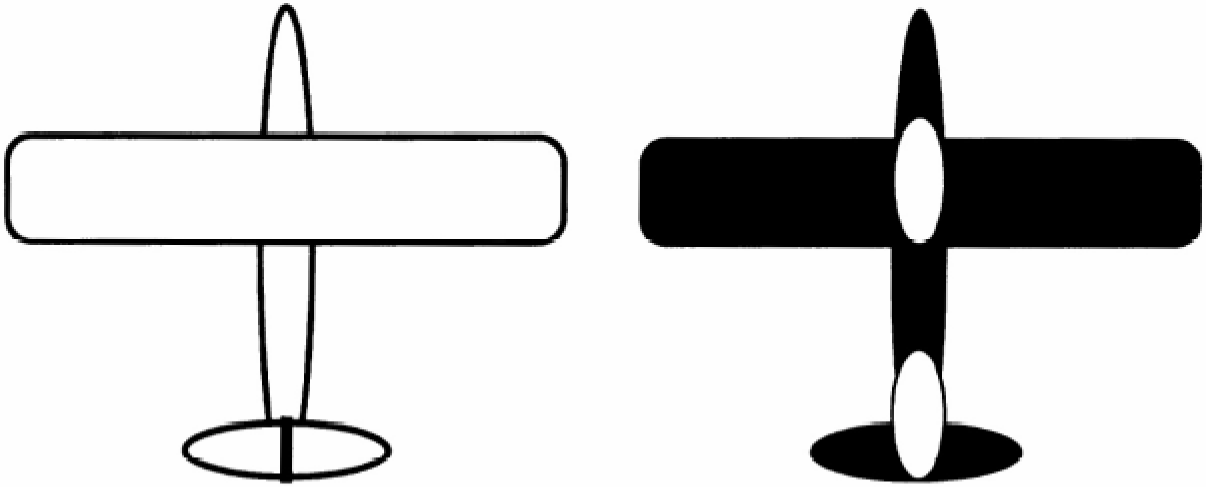
- Dr. Snow looked at other cases of cholera away from the Broad Street pump, and found that they either visited the area regularly, or worked in the area.
- Exceptions, e.g. residuals, made the results stronger.

Abraham Wald
1902-1950

Father of

Decision Theory &
Sequential Analysis

During WWII and later in Korea and Vietnam, the U.S. Navy and Air Force studied bullet-hole patterns on returning aircraft to determine where to reinforce the aircraft against ground fire. Abraham Wald (a statistician at U.S. Center for Naval Analyses) worked on this problem from 1941. Wald dryly noted better information would have been obtained from the planes that hadn't returned. He nevertheless managed to construct statistical models which gave a useful insight into the vulnerability of different parts of the aircraft.



An outline of a plane.

A depiction of a plane with shading indicating where returning planes had been shot.

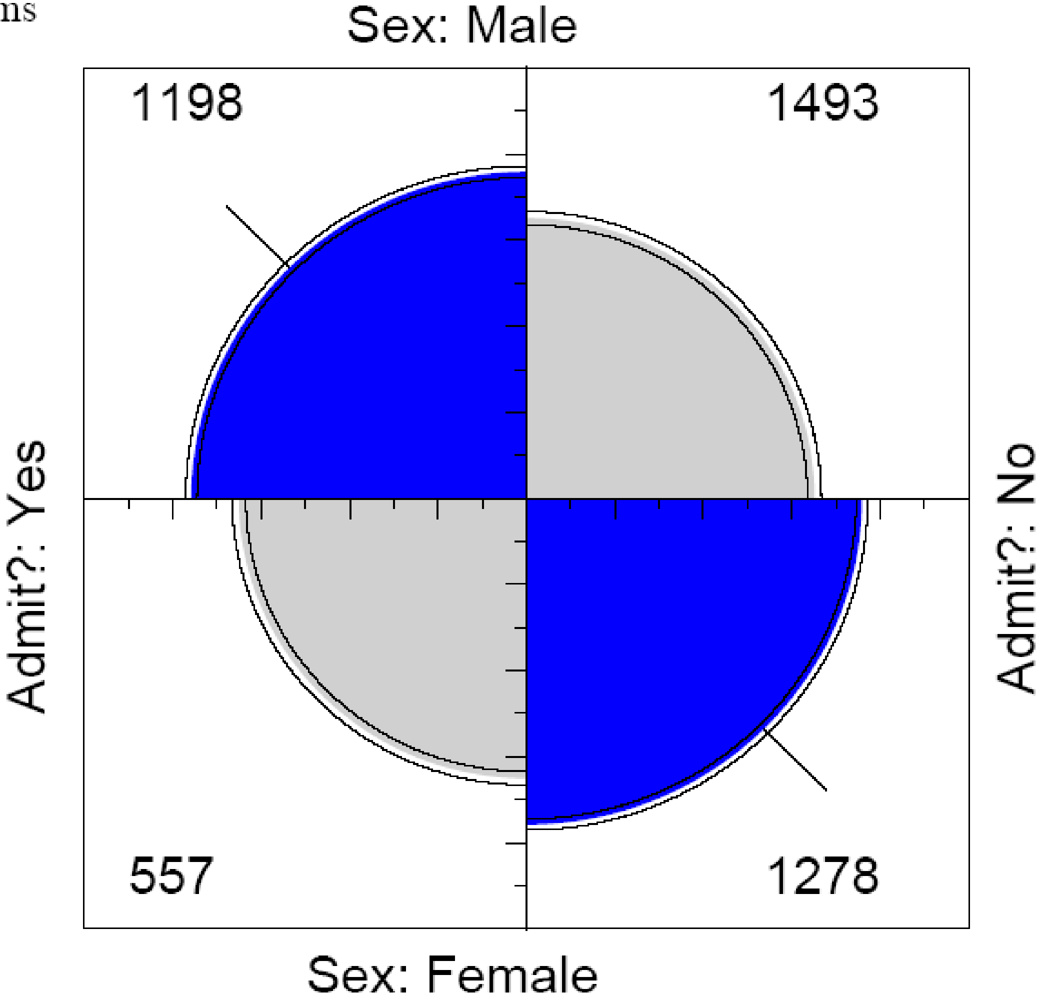
Figure 6. A schematic representation of Abraham Wald's ingenious scheme to investigate where to armor aircraft.

Table 2: Admissions to Berkeley graduate programs

	Admitted	Rejected	Total
Males	1198	1493	2691
Females	557	1278	1855
Total	1755	2771	4526

Table 2: Admissions to Berkeley graduate programs

	Admitted	Rejected	Total
Males	1198	1493	2691
Females	557	1278	1855
Total	1755	2771	4526



(1198/1493)/(557/1278)
= 1.84

Figure 5 shows the aggregate data from Table 2. The sample odds ratio, Odds (Admit|Male) / (Admit|Female) is 1.84 indicating that males were almost twice as likely to be admitted. The confidence rings in the figure do not overlap, showing that this association is highly significant. Does this constitute evidence for gender bias in admission?

ns

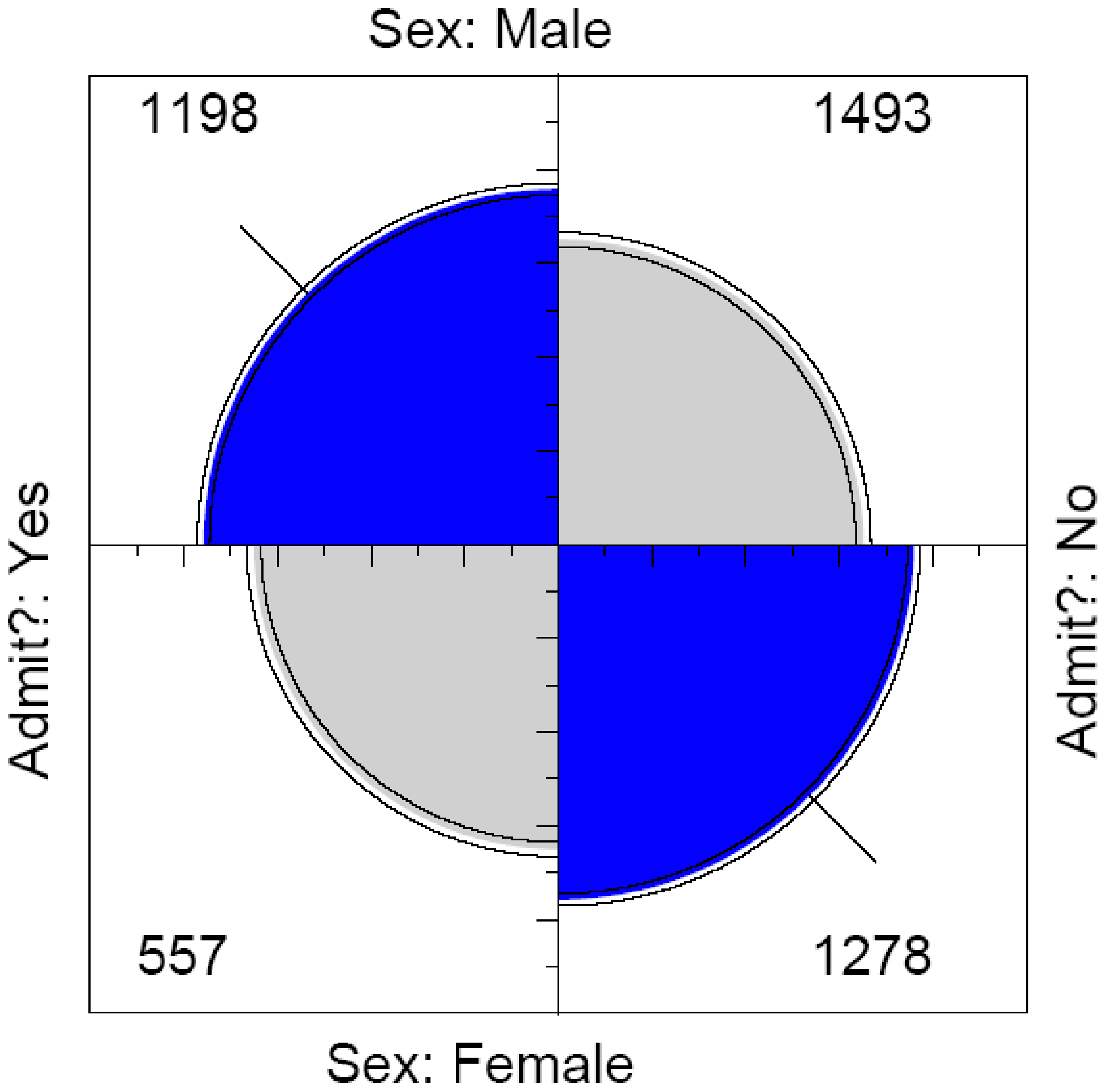


Figure 5: Fourfold display for Berkeley admissions data, margins equated

Table 2: Admissions to Berkeley graduate programs

	Admitted	Rejected	Total
Males	1198	1493	2691
Females	557	1278	1855
Total	1755	2771	4526

Table 2 is an example of a 2 × 2 table. For such data, the *odds ratio*, $\theta = n_{11}n_{22}/n_{12}n_{21}$, is a natural measure of the strength of association between the two variables.

The ***fourfold display*** depicts these frequencies by quarter circles, whose radius is proportional to $\sqrt{n_{ij}}$, so the area is proportional to the cell count (Fienberg, 1975, Friendly, 1994a,c). The cell frequencies are usually scaled to equate the marginal totals, and so that the ratio of diagonally opposite segments depicts the odds ratio. Confidence rings for the observed θ allow a visual test of the hypothesis $H_0 : \theta = 1$ corresponding to no association. They have the property that the rings for adjacent quadrants overlap *iff* the observed counts are consistent with the null hypothesis.

The admissions data shown in Figure 5 came from the six largest at Berkeley. To determine the source of the apparent sex bias in favor of males, we make a new plot, Figure 6, stratified by department.

Surprisingly, Figure 6 shows that, for five of the six departments, the odds of admission is approximately the same for both men and women applicants. Department A appears to differs from the others, with women approximately 2.86 $(= (313/19)/(512/89))$ times as likely to gain admission.

The resolution of this contradiction can be found in the large differences in admission rates among departments. Men and women apply to different departments differentially, and in these data women happen to apply in larger numbers to departments that have a low acceptance rate. The aggregate results are misleading because they falsely assume men and women are equally likely to apply in each field.

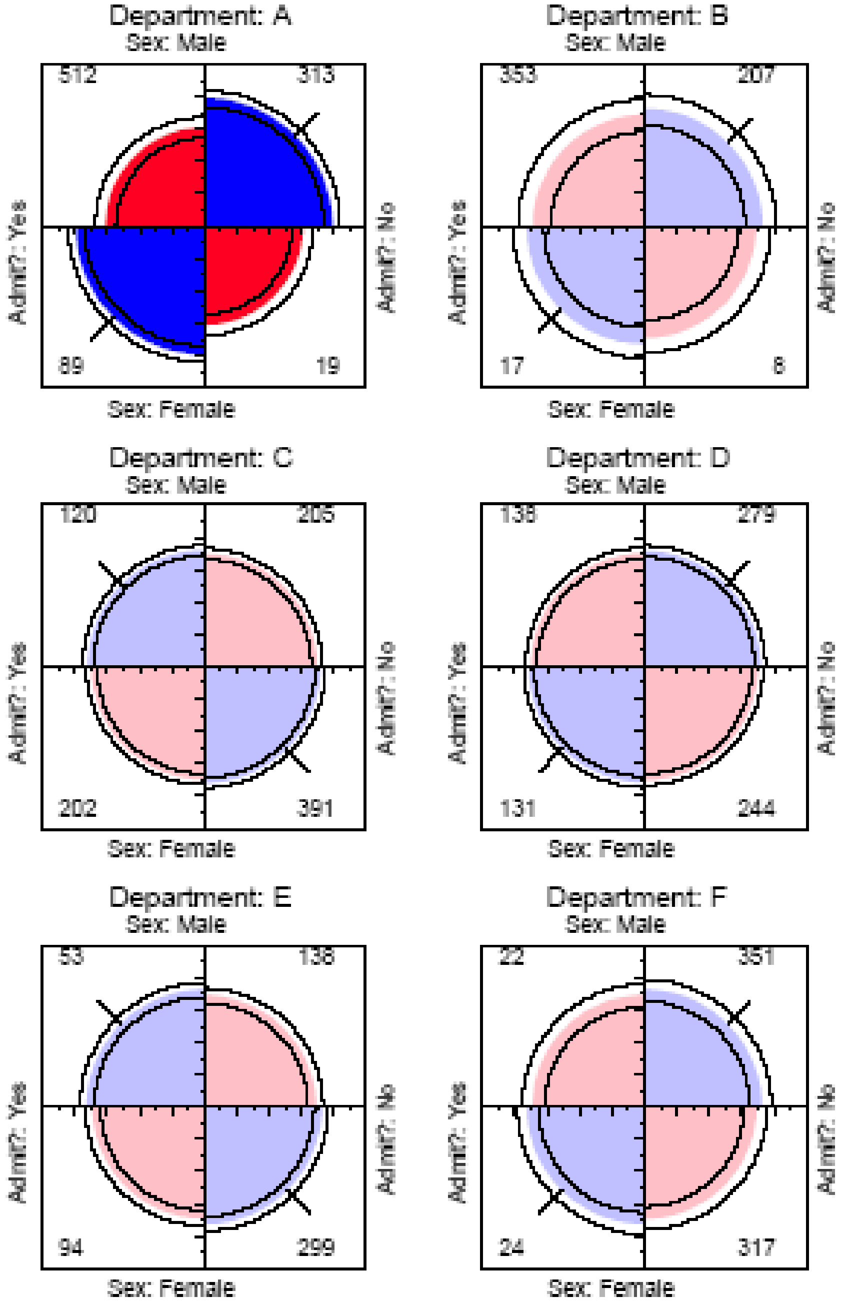


Figure 6: Fourfold display for Berkeley admissions data, by department

To make these plots:

- Download the vcd package.
- data(UCBAdmissions)
- ## Use the Berkeley admission data.
- x <- aperm(UCBAdmissions, c(2, 1, 3))
- dimnames(x)[[2]] <- c("Yes", "No")
- names(dimnames(x)) <- c("Sex", "Admit?", "Department")
- ftable(x)
- ## Fourfold display of data aggregated over departments, with
- ## frequencies standardized to equate the margins for admission
- ## and sex.
- `fourfold(margin.table(x, c(1, 2)))

To make these plots:

- ## Fourfold display of x, with frequencies in each table
- ## standardized to equate the margins for admission and sex.
- fourfold(x)
- cotabplot(x, panel = cotab_fourfold)
- ## Fourfold display of x, with frequencies in each table
- ## standardized to equate the margins for admission. but not for sex.
- fourfold(x, margin = 2)

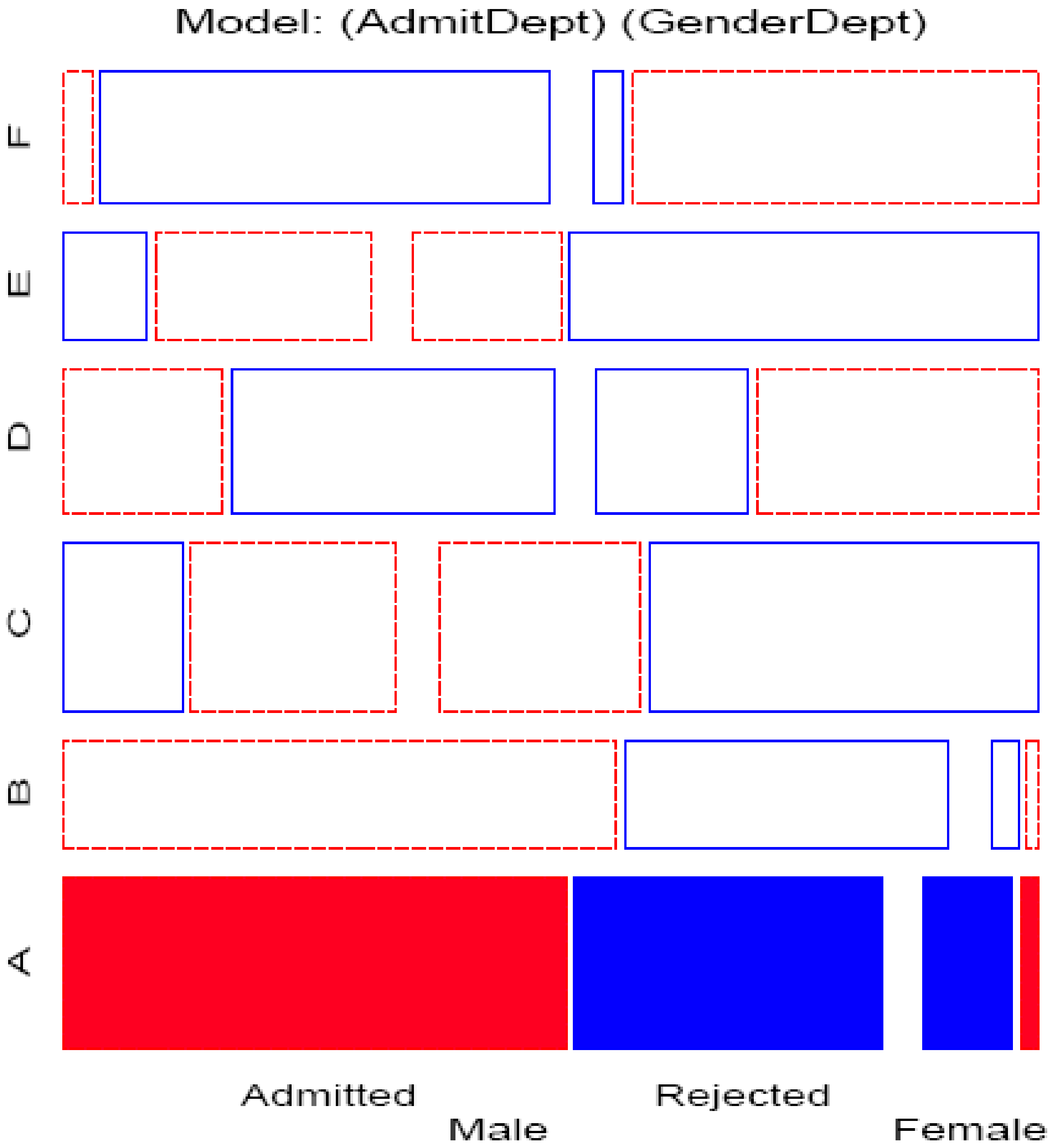


Figure 7: Three-way mosaic plot for Berkeley data: Conditional independence

Another data set: the Titanic Survival Rates

- data(Titanic)
- doubledecker(Titanic)
- doubledecker(Titanic, depvar = "Survived")
- doubledecker(Survived ~ ., data = Titanic)

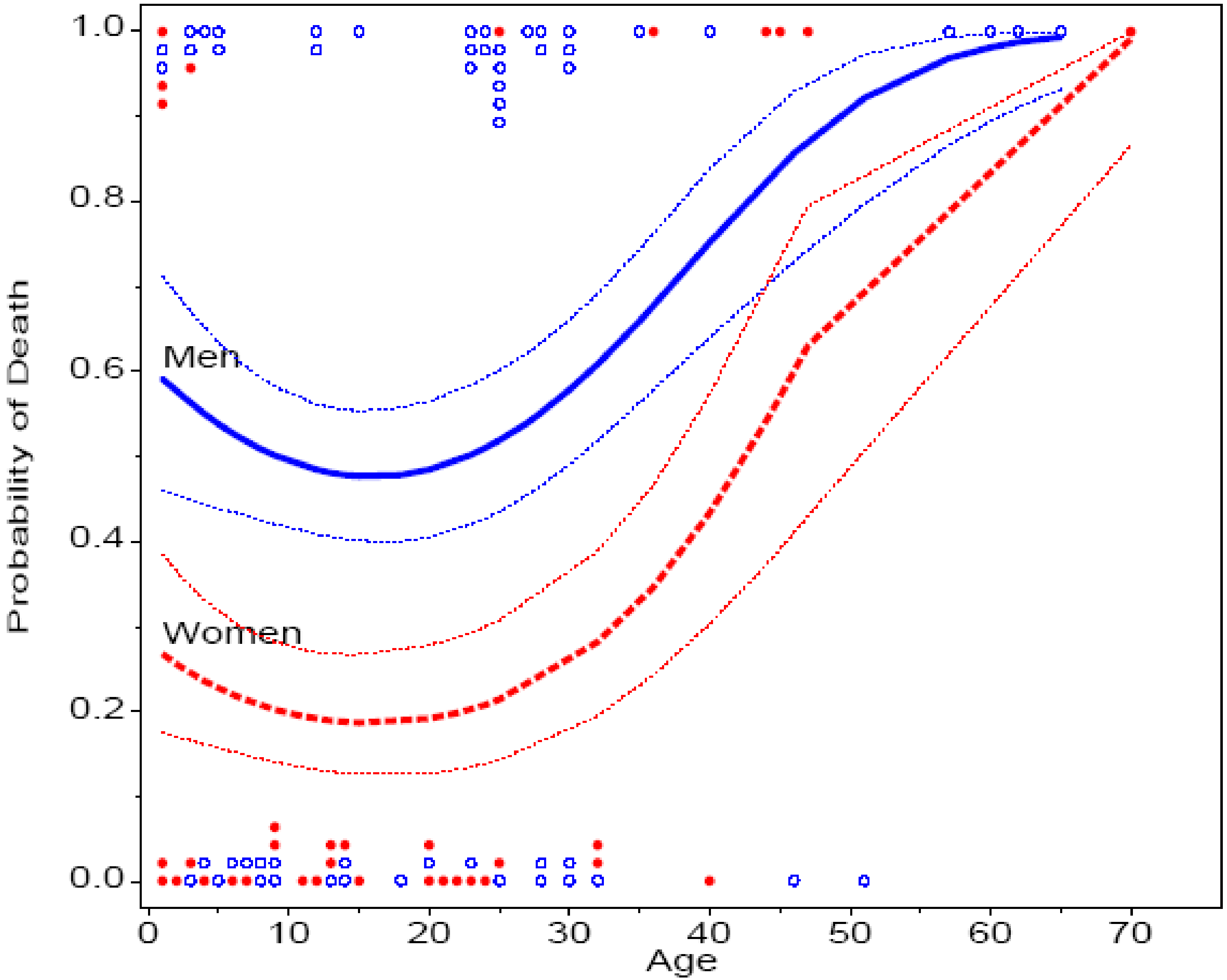
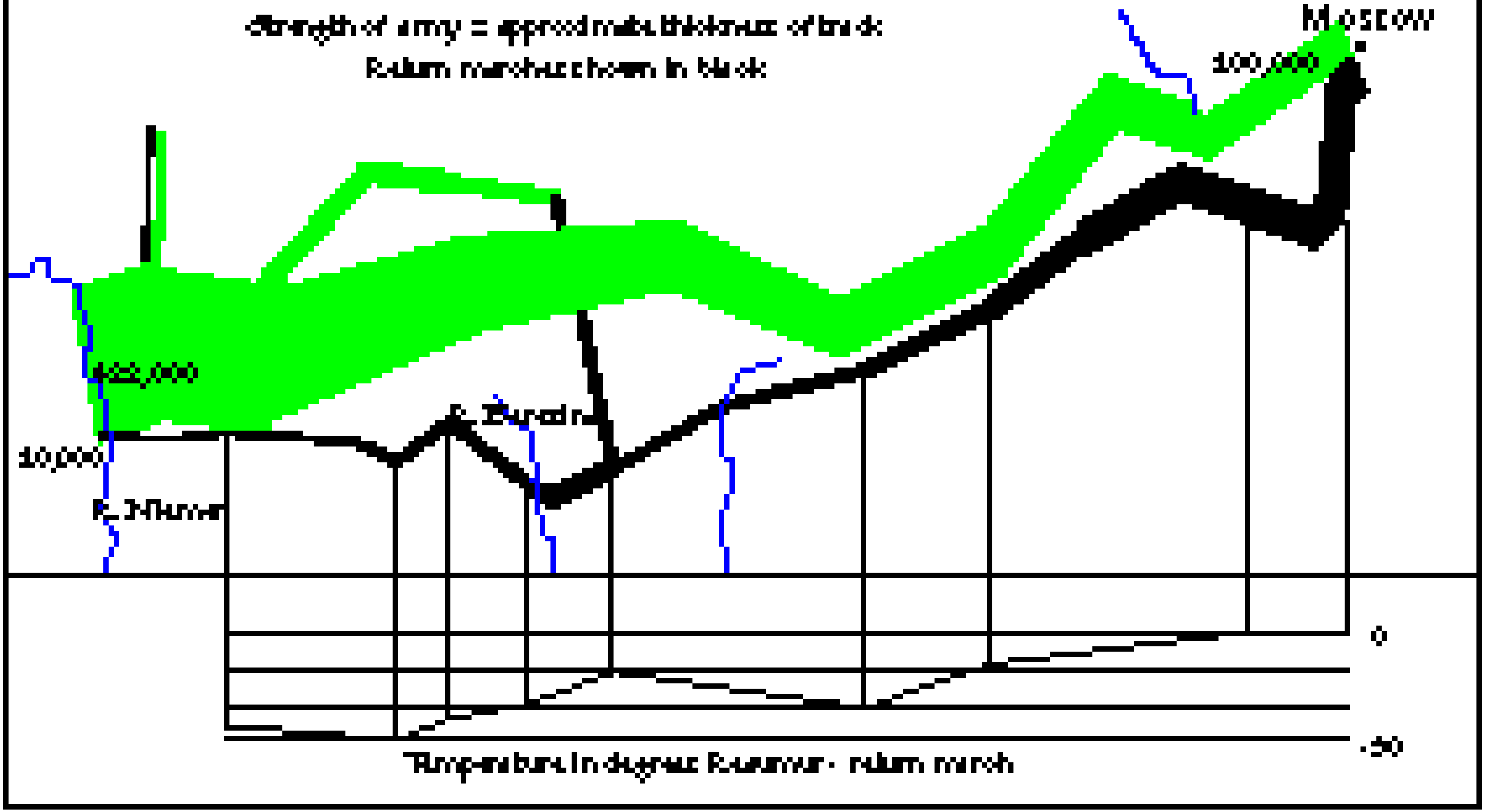
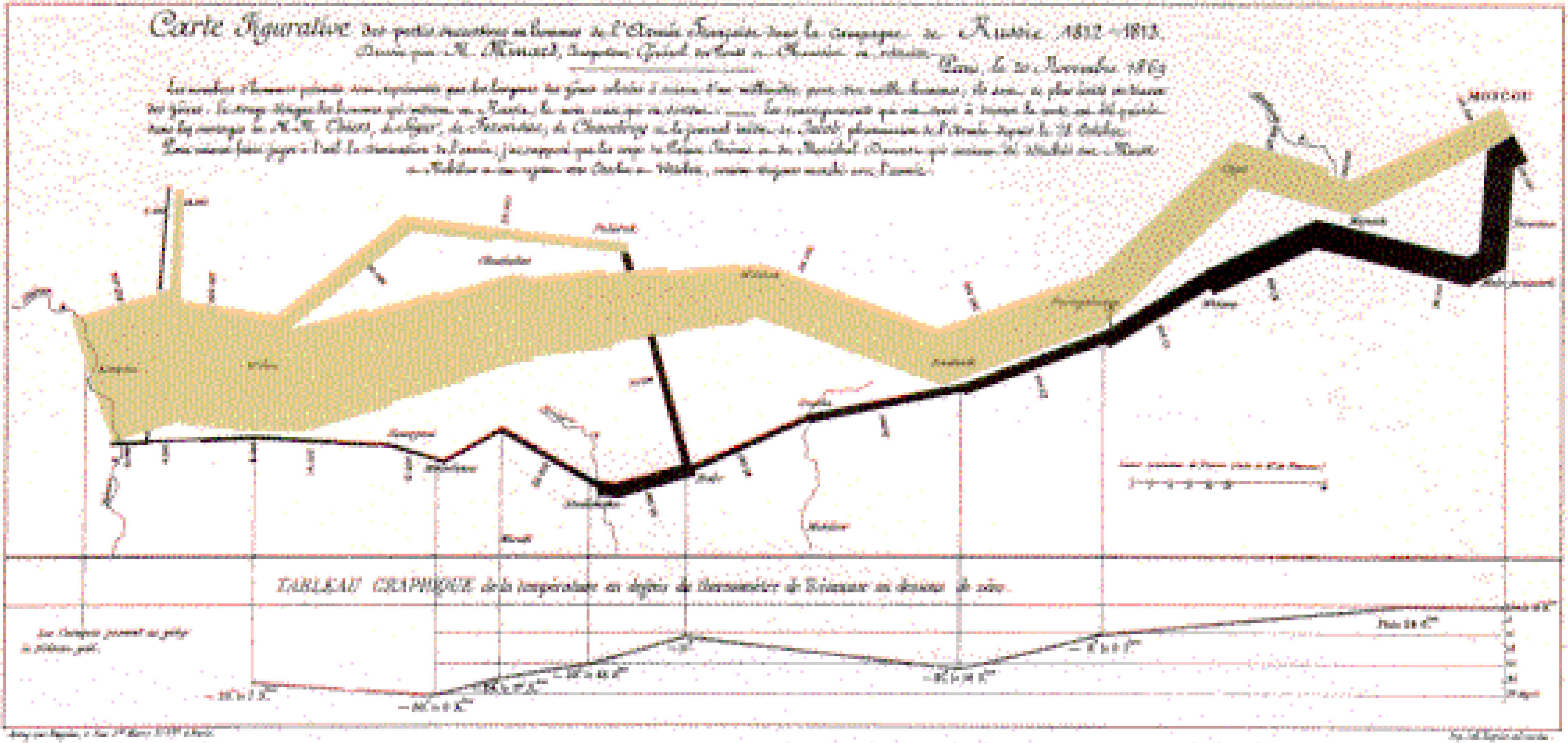
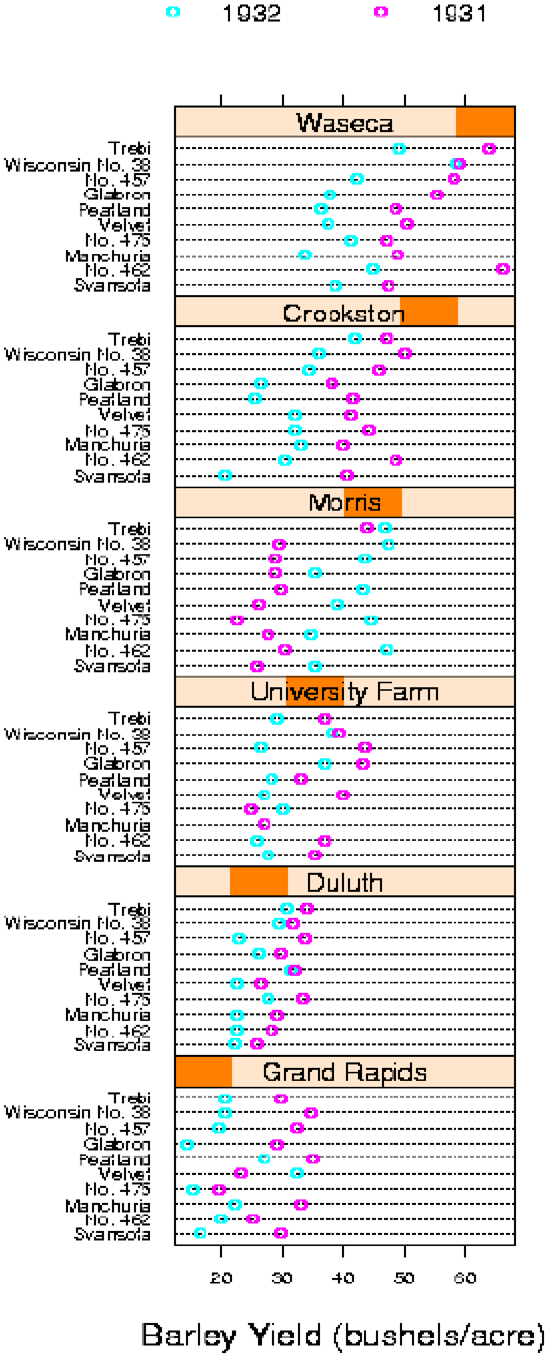


Figure 20: Donner Party, fitted logistic model, Pr(Death) $\sim$ Age + Age^2 + Male

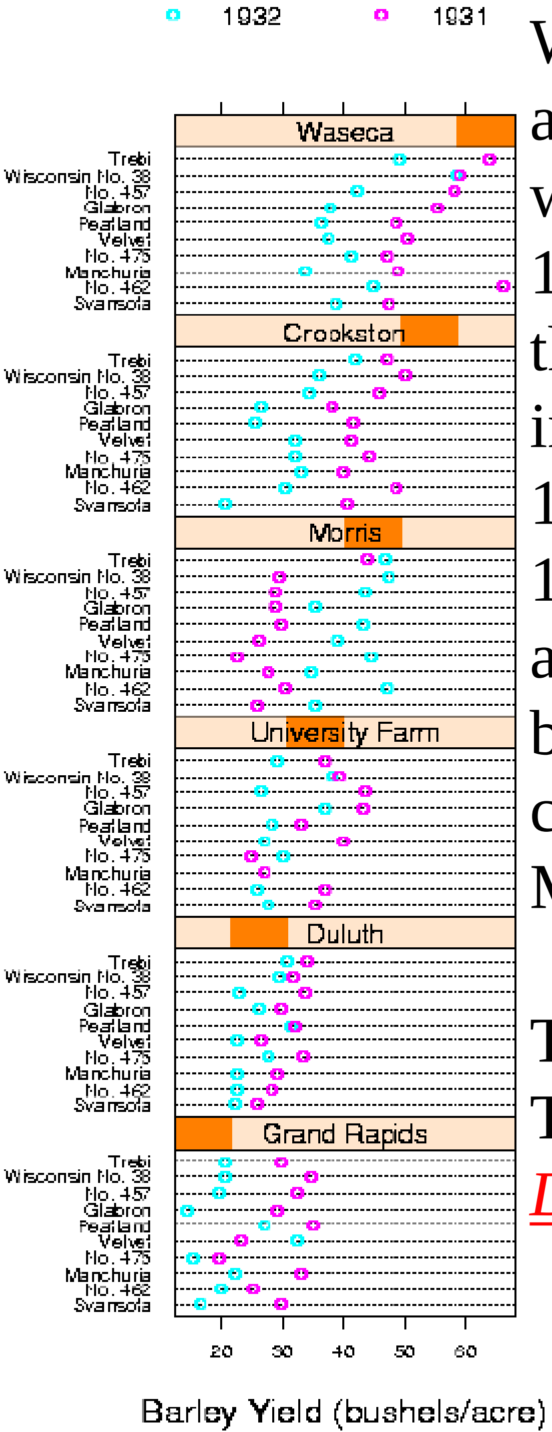
Napoleon's Russian Campaign of 1812




Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869
MOSCOU
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

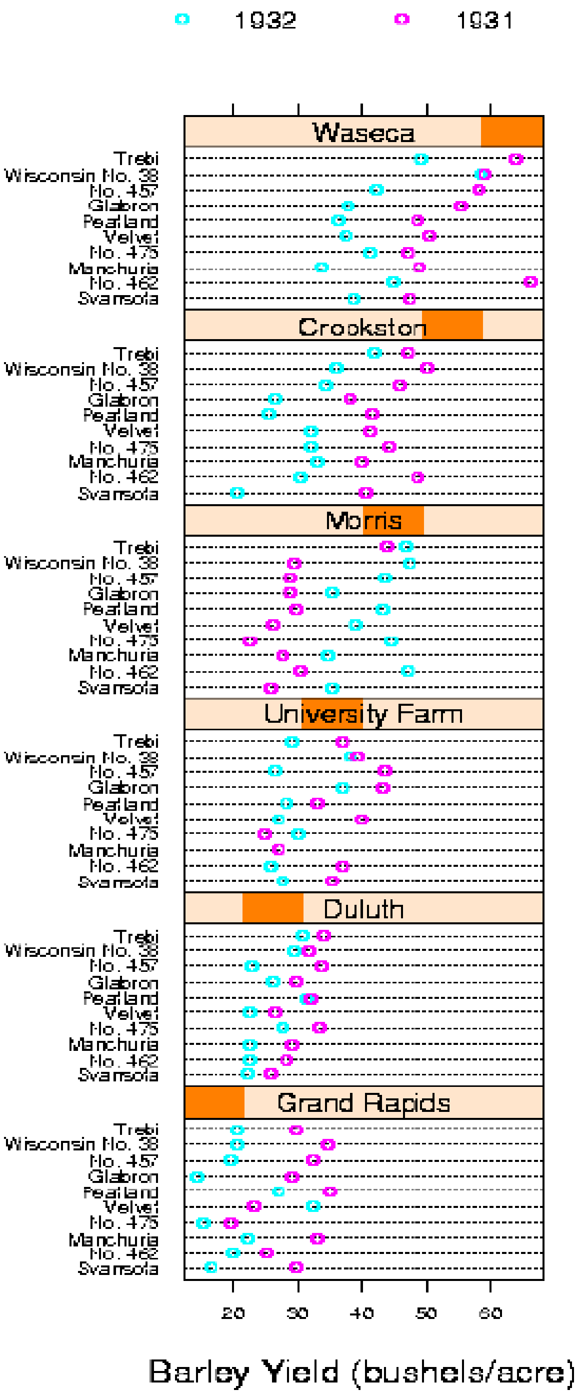


The figure is a Trellis display of data from an agricultural field trial of barley yields at six sites in Minnesota; ten varieties of barley were grown in each of two years. The data were presented by R. A. Fisher in *The Design of Experiments* and analyzed subsequently by many others.



William Cleveland's display of these data shows an apparent surprise missed by previous investigators, which occurs at the Morris site: For all other sites, 1931 produced a significantly higher overall yield than 1932. The reverse is true at Morris. But most importantly, the amount by which 1932 exceeds 1931 at Morris is similar to the amounts by which 1931 exceeds 1932 at the other sites. More displays, a statistical modeling of the data, and some background checks on the experiment led to the conclusion that the data are in error -- the years for Morris were inadvertently reversed.

The background of the data, and analysis with Trellis are described in more detail in *The Visual Design and Control of Trellis Display*



The graph uses *main effect ordering* to arrange the 6 sites and 10 barley varieties from bottom to top according to increasing values of the median yields (collapsed over other factors). This greatly aids perception of trends in the data and makes the Morris data stand out as unusual.

To run this analysis:

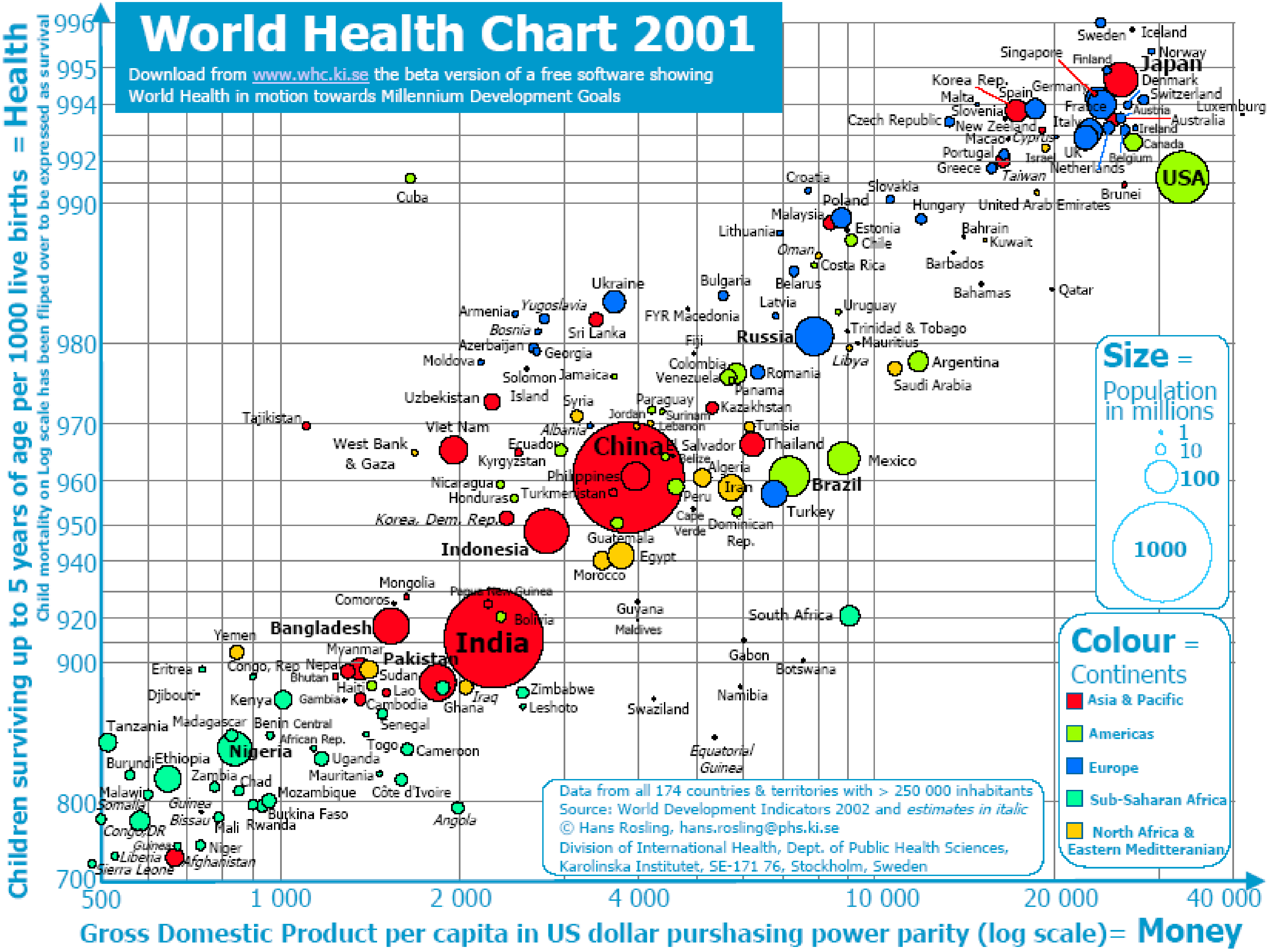
- Load lattice package
- “barley” is a data frame with 120 observations on the following 4 variables.
- **yield** Yield (averaged across three blocks) in bushels/acre.
- **variety** Factor with levels "Svansota", "No. 462", "Manchuria", "No. 475", "Velvet", "Peatland", "Glabron", "No. 457", "Wisconsin No. 38", "Trebi".
- **year** Factor with levels 1932, 1931
- **site** Factor with 6 levels: "Grand Rapids", "Duluth", "University Farm", "Morris", "Crookston", "Waseca"

Dotplot:

- dotplot(variety ~ yield | site, data = barley, groups = year, key = simpleKey(levels(barley$year), space = "right"), xlab = "Barley Yield (bushels/acre) ", aspect=0.3, layout = c(3,2), ylab=NULL)
- \

World Health Chart 2001

Download from www.whc.ki.se the beta version of a free software showing World Health in motion towards Millennium Development Goals



Data from all 174 countries & territories with > 250 000 inhabitants
Source: World Development Indicators 2002 and *estimates in italic*
© Hans Rosling, hans.rosling@phs.ki.se
Division of International Health, Dept. of Public Health Sciences,
Karolinska Institutet, SE-171 76, Stockholm, Sweden

Size = Population in millions

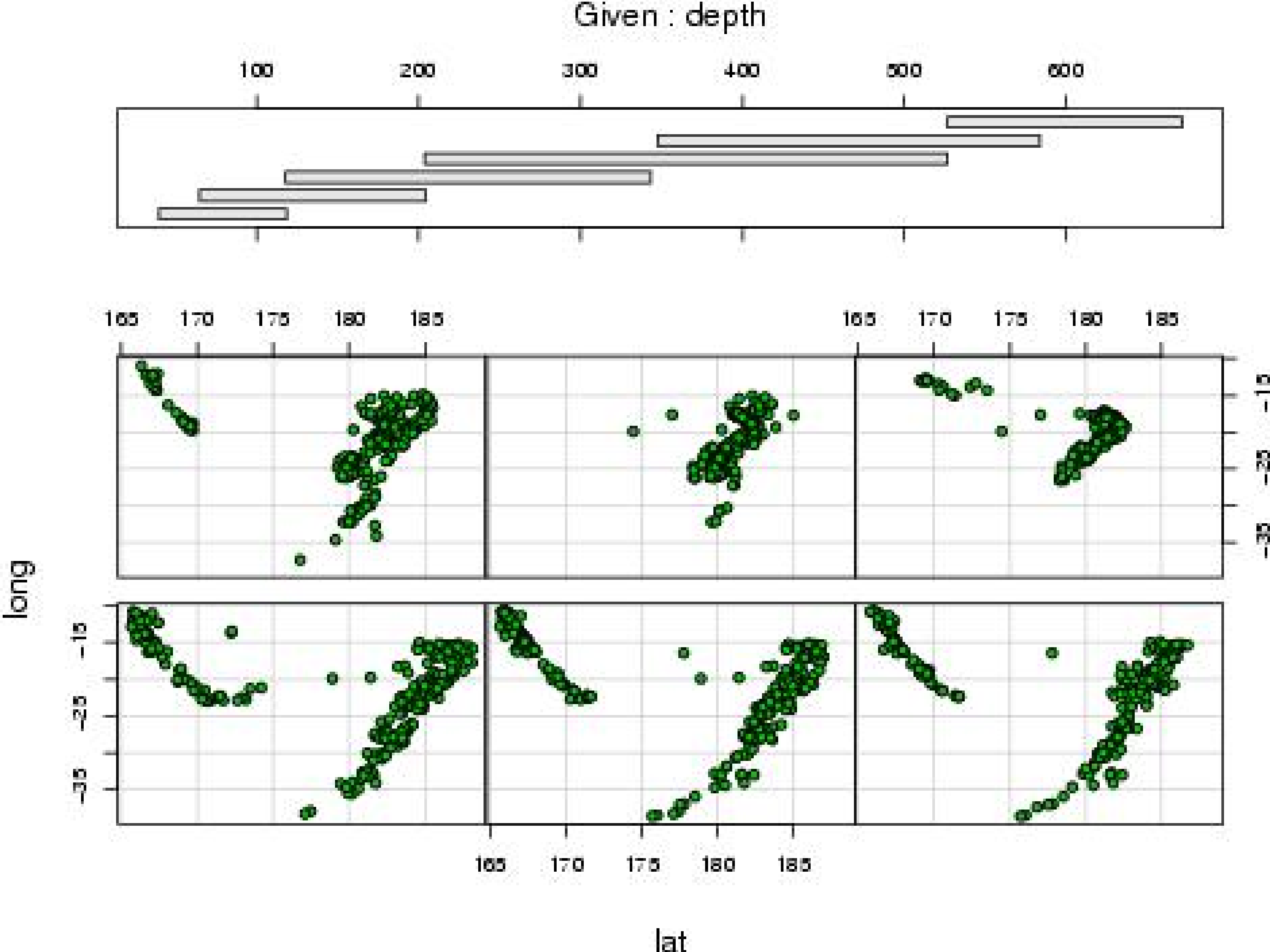
- 1
- 10
- 100
- 1000

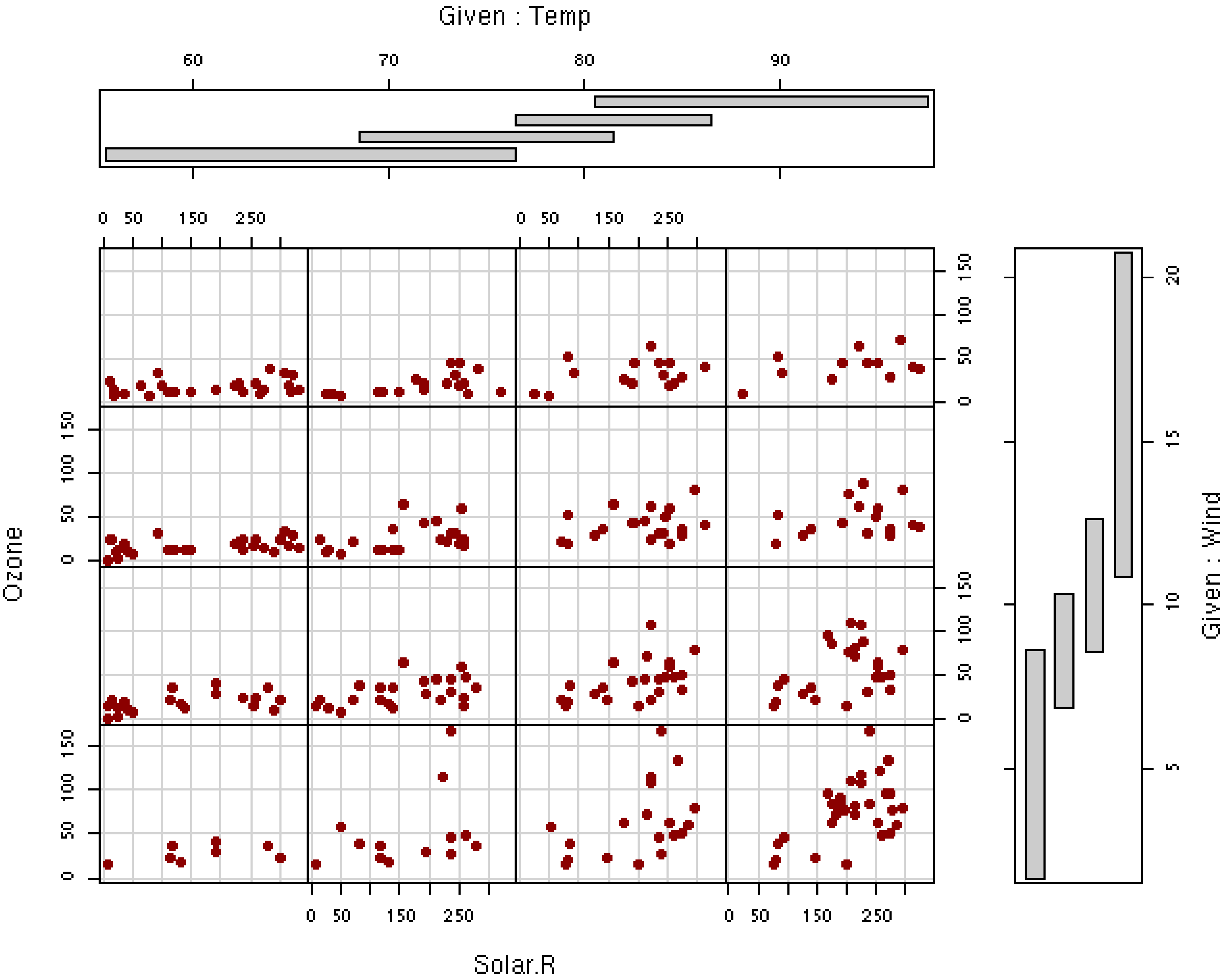
Colour = Continents

- 🟥 Asia & Pacific
- 🟩 Americas
- 🟦 Europe
- 🟩 Sub-Saharan Africa
- 🟨 North Africa & Eastern Meditteranian

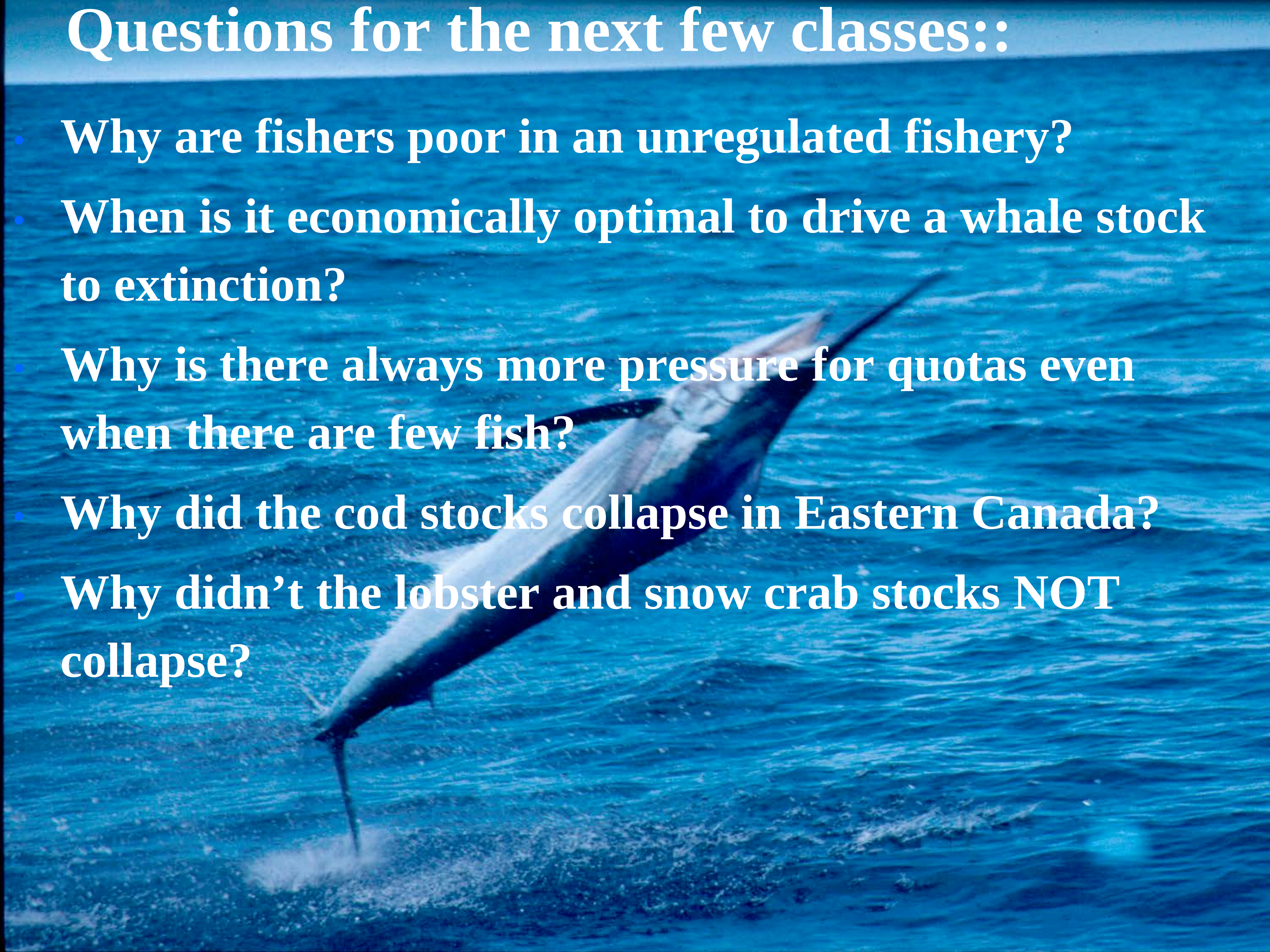
Tufte's Rules:

- Tone down secondary elements of a picture: *layer* the figure to produce a visual hierarchy.
- Replace coded labels in the figure by direct ones.
- Produce emphasis by using the smallest possible effective distinctions.
- Eliminate all unnecessary parts of a figure.
- Use *small multiples*: numerous repetitions of a single figure with slight variations.
- Make the graphics carry a story.


Given : depth
100
200
300
400
500
600
165
170
175
180
185
-15
-25
-35
long
lat


Given : Temp
60
70
80
90
0
50
150
250
Ozone
Solar.R
Given : Wind
5
10
15
20
100

Questions for the next few classes::



- Why are fishers poor in an unregulated fishery?
- When is it economically optimal to drive a whale stock to extinction?
- Why is there always more pressure for quotas even when there are few fish?
- Why did the cod stocks collapse in Eastern Canada?
- Why didn’t the lobster and snow crab stocks NOT collapse?